UNIVERSITÀ DEGLI STUDI DI CATANIA

Dipartimento di Matematica e Informatica
Dottorato di Ricerca in Informatica XXVIII Ciclo

DOTT. SALVATORE ALAIMO

From diagnosis to therapy: algorithmic methodologies for precision medicine

PH.D. THESIS

Tutor:
Ch.mo Prof. Alfredo Pulvirenti
Co-Tutor:
Ch.mo Prof. Alfredo Ferro

ANNO ACCADEMICO 2015-2016

Tutor: Ch.mo Prof. Alfredo Pulvirenti
Co-Tutor: Ch.mo Prof. Alfredo Ferro

Dott. Salvatore Alaimo

From diagnosis to therapy: algorithmic methodologies for precision medicine

Abstract

In recent years, it has become established the idea of a novel medicine where a patient is the center around which multidisciplinary teams (made up of physicians, statisticians and bioinformaticians) sew targeted treatments. Precision medicine involves the use of detailed patient-specific molecular information for diagnosing, categorizing and guiding treatment of a disease, with the main purpose of improving the clinical outcome compared to a more classical approach. In precision medicine it is supposed that the cause of a disease is at least partially attributable to specific genetic or epigenetic characteristics of a patient. Therefore, identifying these specificities helps building the best treatment for each individual. Next-generation sequencing techniques are massively employed, giving the ability to quickly and at relatively low cost analyze whole genomes, epigenomes and transcriptomes. This ability is clinically important since the prediction of treatment effectiveness is usually affected by many factors. A fundamental function in this new medicine is played by bioinformatics. It has a crucial role in every aspect of precision medicine, such as the accurate classification of patients, the prediction of new therapies based on current knowledge, the identification of possible outcomes of a disease or therapy, and the enrichment of current knowledge on pathogenic processes or on pharmaceuticals.

The aim of this thesis is the development of an integrated framework, based on synergistically operating tools, models and algorithms, which help to fill some of the major gaps in each step of the production of highly customized therapies, overcoming, if possible, the limitations of currently employed techniques, defining a new standard for precision medicine informatics.

Contents

List of Figures

List of Tables

Acknowledgments

I wish to thank all those who have made this thesis possible with their help, encouragement and support. I owe them a great debt of gratitude. First and foremost, Prof. Alfredo Ferro, he welcomed me in a wonderful group of people who work togheter with great enthusiasm, helping in times of need. I also thank Prof. Alfredo Pulvirenti who first introduced me to this wonderful group, and encouraged and followed my work over the years, providing a wise guidance and support. I must also thank Prof. Rosalba Giugno that over the years has always provided valuable tips and important ideas. I am extremely grateful to all three of them, without their support this work would not have existed.

I also wish to thank all my colleagues who have endured me, and spent some of their time helping my work.

Finally I would like to thank my family, to whom I dedicate this work, for being always a pillar to trust throughout my life, giving me a very important support that allowed the completion of this work.

1

Introduction

In recent years, with the increasing knowledge on genetic diseases, such as cancer, it has become established the idea of a novel medicine where a patient is the center around which multidisciplinary teams (made up of physicians, statisticians and bioinformaticians) sew targeted treatments. Precision medicine involves the use of detailed patient-specific molecular information for diagnosing, categorizing and guiding treatment of a disease, with the main purpose of improving the clinical outcome compared to a more classical approach [1]. In precision medicine it is supposed that the cause of a disease is at least partially attributable to specific genetic or epigenetic characteristics of a patient. Therefore, identifying these specificities helps building the best treatment for each individual. However, at present precision treatments rely mainly on the identification of specific biomarkers, that is molecular events, which are connected in some way with the response to the treatment but may be completely unrelated to the disease [2–4].

In precision medicine, Next-generation Sequencing (NGS) techniques are massively employed. This implies the ability to quickly and at relatively low cost analyze the whole genome, epigenome and transcriptome of a single sample. This ability is clinically important since the prediction of treatment effectiveness is usually affected by many factors.

A fundamental function in this novel medicine is played by bioinformatics. It has a crucial role in every aspect of precision medicine, such as the accurate classification of patients, the prediction of new therapies based on current knowledge, the identification of possible outcomes of a disease or therapy, and the enrichment of current knowledge on pathogenic processes or on pharmaceuticals. Indeed, precision medicine heavily depends on the ability to collect, manage, and process complex information [5]. All data collected from each individual patient data must be integrated and summarized in order to simplify the final decision-making process of the therapies.

In the medical field, a large *volume* of patient data is scattered across a wide *variety* of databases that increase in size at an extraordinary *velocity*. The extraction of the hidden *value* of this data is confronting us with specific challenges at the technical level, in the implementation of computational infrastructures, at organizational and managerial level, in the collection and storage of data, and at scientific level, creating sophisticated algorithmic models for extracting as much value as possible from the data. All this must provide support for the real-time therapeutic decision process, allowing physicians to propose therapies tailored to each patient in the shortest time possible. For that bioinformatics is simultaneously the most important enabler and detractor to the application of personalized medicine, due to the many challenges that must be addressed to make it reality [6]:

- the development of systems that enable data integration, traceability and sharing;
- the development of bioinformatics pipelines to extract the most relevant biological information from large amounts of data;

- the reproducibility of results.

Many recent publication pointed out the key role of bioinformatics in precision medicine [7]. In addition, many tools have been developed to promote the sharing and analysis of genomic data in translational research [8].

Nevertheless, these tools have some major shortcomings. The aim of this thesis is the development of an integrated framework, based on inter-operating tools, models and algorithms, which help to fill some of the major gaps in each step of the production of highly customized therapies, overcoming, if possible, the limitations of currently employed techniques. A summarized workflow is illustrated in Figure 1.1.

In order to predict possible personalized therapies a correct classification of patients, based on their unique biological characteristics, is needed [9]. To date, this process is accomplished through the use of biomarkers, which as mentioned previously may not be at all related to the pathology. One class of techniques, pathway analysis, summarizes genetic and transcriptomic data of a patient in order to obtain a functional assessment of his biological processes [10–16]. This can be used to realize a novel class of functional biomarkers that not only identify the state of each patient, but also provide synthesized information on his physiological processes [17]. Despite this proves promising, currently available methods are not accurate enough and present significant limitations [10]. For this purpose, we developed a new approach, which, by introducing further biological knowledge in current models, tries to obtain more accurate results, making possible its use in the classification of disease states [17].

The correct classification of patients is a step around which we can build a personalized therapy. In order to properly construct a treatment, the knowledge of the inner workings of each drug is necessary to properly select the most appropriate ones, based on current treatment standards and patient data. However this stage is currently limited due to incomplete knowledge on drugs mode of action [18–20] . This led to the development of a novel class of chemoinformatics techniques dealing with the predictions of drug-target interactions (DTI), therefore increasing

their understanding [21–29]. Such methods are complex but a class of techniques, the recommender systems, due to their simplicity are easily applied in this context [30]. Nevertheless, at present, no biological information, other than current known interactions, is introduced in the model. For this reason we decided to fill this important gap by devising a recommendation technique that introduce biological information in the model, making it more accurate and useful [31]. This allowed us to fill another major gap, the lack of an algorithmic methodology for the prediction of drug combinations. A drug combination consists of putting together two or more drugs whose synergistic action potentiates their final effect or reduces side effects. This practice is, therefore, crucial in the realization of personalized therapies. However, no automated methodology was available to support such activity. We have, therefore, taken advantage of this deficiency to develop a novel methodology that automatically predicts drug combinations, helping further investigations [32].

The methods outlined above are fundamental to the correct classification of patients and, therefore, personalization of therapies. However, a lot of information on the functioning of some complex biological processes is still missing. For example, long non-coding RNAs (lncRNA), long RNA molecules correlated with the onset and progression of pathological phenomena [33–37], have mostly unclear functions [38–40]. At time, only one algorithmic model was available for the prediction of associations between lncRNAs and diseases [41]. The model used a recommendation algorithm, in conjunction with known associations, to predict novel ones. Such method is not sufficiently accurate due to their reduced number. For this reason, we chose to employ the larger knowledge on lncRNAs-target-disease interactions to build a novel recommendation model, which uses tripartite network to predict associations [42]. The model using a greater number of known interactions can achieve better results, facilitating experimentations. Another biological process of fundamental importance is RNA editing [43–47], a post-translational modification of RNA nucleotides whose malfunction can lead to serious consequences [48, 49]. At present little is known about the phenomenon and reliable methods to predict putative editing

sites are needed. In this sense, a prediction model based on logistic regression has been designed to accurately predict possible sites subject to editing [50]. Finally, as part of more general process of novel hypotheses production for further experimental testing, we developed a methodology, which analyzes current knowledge in the form of scientific publications, identifying new latent hypotheses directing the experimental process.

Despite the above algorithms are necessary to enhance existing biological knowledge and enable the development of customized therapies, confirmation of their hypotheses always requires an accurate and expensive phase of experimental validation. Considering the huge number of predictions that these methods usually generate, it is necessary to be able to simulate some of these assumptions in silico, thus reducing the experimental part only in the most promising ones. At present many methods have been proposed to perform simulations [51–55]. Such methodologies typically predict the concentration of molecular elements (eg. RNA, proteins) assuming a certain interaction network, or pathway, between them. However, their results are not always reliable due to the incompleteness of our knowledge of biological processes. Therefore, we decided to develop a simulation algorithm that estimate the activity of biological processes, representing them by the state (activated or inhibited) of their endpoints. This led to the production of a randomized algorithm that uses synthetic expression values and pathway analysis to obtain high precise simulations on the state of pathways representing biological processes, filling in part the inaccuracy of currently available methods.

Some results illustrated in this thesis have been published in several peer-reviewed journals in the field of bioinformatics:

- Salvatore Alaimo, Alfredo Pulvirenti, Rosalba Giugno, and Alfredo Ferro. Drug–target interaction prediction through domain-tuned network-based inference. *Bioinformatics*, 29(16):2004–2008, 2013 [31];

- Salvatore Alaimo, Rosalba Giugno, and Alfredo Pulvirenti. ncpred: ncrna-disease association prediction through tripartite network-based inference. *Frontiers in bioengineering*

and biotechnology, 2, 2014 [42];

- Salvatore Alaimo, Vincenzo Bonnici, Damiano Cancemi, Alfredo Ferro, Rosalba Giugno, and Alfredo Pulvirenti. Dt-web: a web-based application for drug-target interaction and drug combination prediction through domain-tuned network-based inference. *BMC systems biology*, 9(Suppl 3):S4, 2015 [32];

- Giovanni Nigita, Salvatore Alaimo, Alfredo Ferro, Rosalba Giugno, and Alfredo Pulvirenti. Knowledge in the investigation of a-to-i rna editing signals. *Frontiers in bioengineering and biotechnology*, 3, 2015 [50];

- Federica Eduati, Lara M Mangravite, Tao Wang, Hao Tang, J Christopher Bare, Ruili Huang, Thea Norman, Mike Kellen, Michael P Menden, Jichen Yang, et al. Prediction of human population responses to toxic compounds by a collaborative competition. *Nature biotechnology* 2015 [56];

- Salvatore Alaimo, Rosalba Giugno, Mario Acunzo, Dario Veneziano, Alfredo Ferro, and Alfredo Pulvirenti. Post-transcriptional knowledge in pathway analysis increases the accuracy of phenotypes classification. *arXiv preprint arXiv:1510.08237*, 2015 [17].

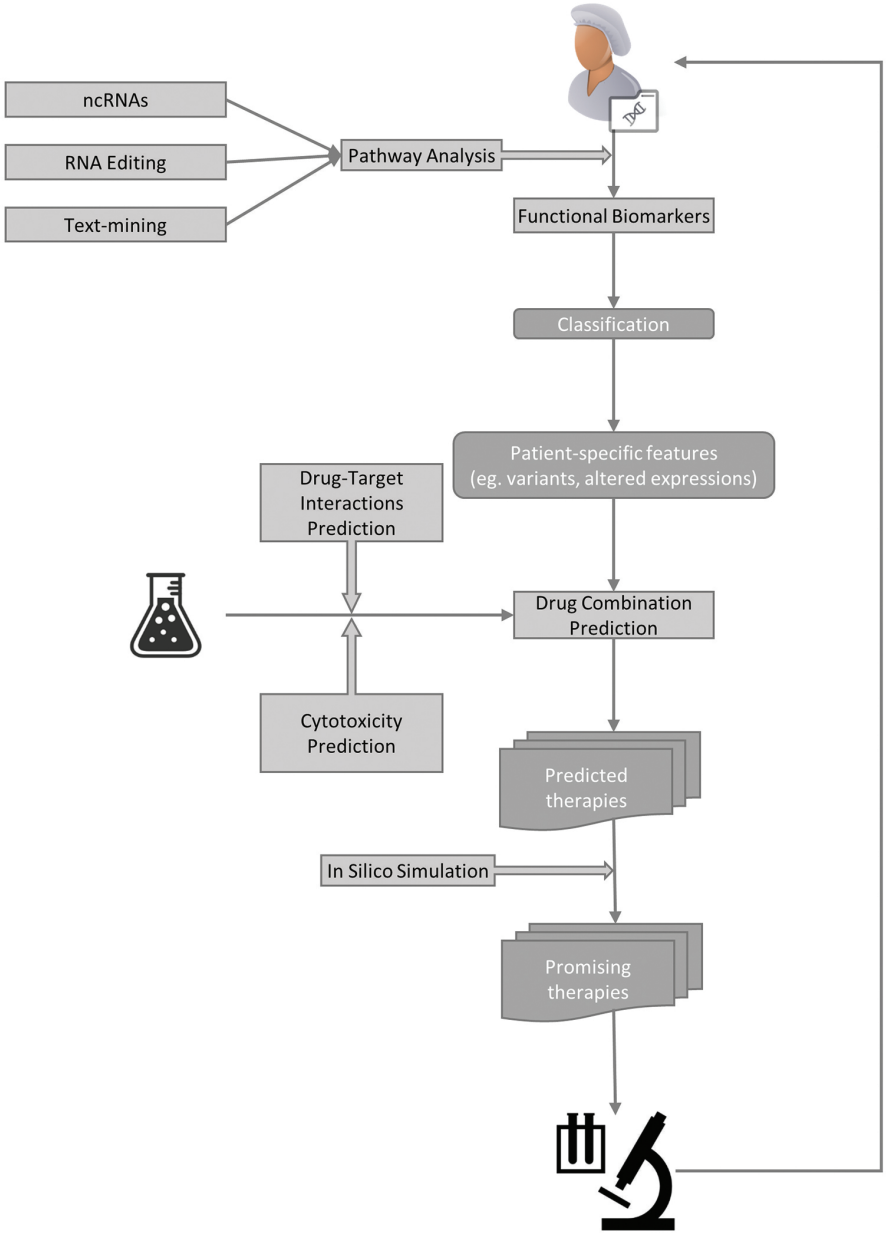


Figure 1.1: Overview of the workflow presented in this thesis. Starting from patient clinical, genetic and transcriptomic data, through pathway analysis we determine functional biomarkers, which are used to classify the patient with respect to a reference and determine his specific characteristics. These characteristics are used together with the predicted and validated information about drugs to define possible therapies, which are subsequently filtered by an in silico simulation. The most promising ones are validated in laboratory. The treatment thus generated can be administered to the patient, and the process continues until his remission. Pathway analysis is further empowered with information coming from text-mining or predictions of non-coding RNAs interaction, or RNA editing events.

2

Prerequisites

2.1 Mathematical and Algorithmic Prerequisites

This section is a brief introduction to probability, classification and recommendation algorithms.

2.1.1 Fundamentals of Probability and Statistics

In this section I will illustrate the fundamental concepts of probability and statistics. The first part will describe the basics of probability theory and the concept of random variables. Later, the basic concepts of statistics (such as mean, variance) and a brief description on the normal and hypergeometric distributions will be presented.

Fundamentals of Probability The probability is a concept used everyday. Taking a card from a deck of playing cards, flipping a coin represent classic examples. The process of tossing a

coin is called an *experiment*. The set of all possible experiments is defined *population* or *sample space*. The probability theory deals with experiments which may have different sets of outcomes. For simplicity we will consider only events whose sample space is **finite**. Any subset of a sample space is called an **event**. An event containing exactly one element is called **elementary event**.

The elements contained in an event is the subset of possible outcomes of an experiment. The degree of certainty that an event contains the actual result of the experiment is indicated with a real number between 0 and 1 called the **probability** of the event. A value of 0 means the certainty that the event does not contain the result of the experiment, 1 means that we are sure of the opposite. Intermediate values represent the different degrees of belief. A more formal definition of probability is as follows:

Definition 2.1. Suppose we have a sample space $\Omega = \{e_1, e_2, \ldots, e_n\}$ containing n distinct elements, and let $P(\Omega)$ denote the power set of Ω, or the set of events. A function $p : P(\Omega) \to [0, 1]$ that assigns a real number to each event $E \subseteq \Omega$ is called a **probability function** on the events of Ω if it satisfies the following conditions:

1. $0 \leq p(\{e_i\}) \leq 1 \quad for\ 1 \leq i \leq n$
2. $\sum_{i=1}^{n} p(\{e_i\}) = 1$
3. $\forall E \subseteq \Omega : \ \ p(E) = \sum_{e \in E} p(\{e_i\})$
4. $p(\{\}) = 0$

Definition 2.2. The pair (Ω, p) is called a **probability space**.

In probability theory the **principle of indifference** assumes fundamental importance. It states that the outcomes are to be considered equiprobable if there is no reason to expect one over the other. This implies that if there are n elements in a sample space, each will have a probability equal to $1/n$.

Axiom 2.1. Given a probability space (Ω, p), some axioms are:

1. $p(\Omega) = 1$,

2. $0 \leq p(E) \leq 1 \quad \forall E \subseteq \Omega$,

3. $\forall E, F \subseteq \Omega$ such that $E \cap F = \emptyset,\ p(E \cup F) = p(E) + p(F)$,

4. $\forall E, F \subseteq \Omega$ such that $E \cap F \neq \emptyset,\ p(E \cup F) = p(E) + p(F) - p(E \cup F)$.

Suppose we want to determine the probability of an event E assuming that another event F happened. This quantity is called the **conditional probability** of E given F and is denoted by $p(E|F)$. It can be estimated as:

$$p(E|F) = \frac{p(E \cap F)}{p(F)}. \tag{2.1}$$

We say that two events E and F are **independent** if $p(E|F) = p(E)$. This implies that if two events are independent then $p(E \cap F) = p(E)\,p(F)$. If two events are independent once we know the outcome of a third event, we are in the presence of conditional independence. More formally we say that two events E and F are **conditionally independent** given G if $p(E|F \cap G) = p(E|G)$.

Now, given two events E and F such that $p(E) \neq 0$ and $p(F) \neq 0$, the **Bayes Theorem** affirms that:

$$p(E|F) = \frac{p(F|E)\,p(E)}{p(F)}. \tag{2.2}$$

Random Variables Given a probability space (Ω, p), we call **random variable** X a function whose domain is Ω. The range of values taken by X is called the **space** of X. When dealing with random variables, we denote with $X = x$ the event $\{e \in \Omega | X(e) = x\}$. We call the values of $p(X = x)$[*] for all values x of X the **probability distribution** (or **probability mass function** or **pmf**) of the random variable X. When we are referring to the probability distribution of X, we write $p(X)$. Let X and Y be two random variables defined on the same sample space Ω. We

[*]For simplicity, we will use the notation $p(x)$ instead of $p(X = x)$.

call $p(X = x, Y = y)$ the **joint probability distribution** of X and Y. Then the following two equations holds:

$$p(X = x) = \sum_{y} p(X = x, Y = y), \tag{2.3}$$

$$p(X = x) = \sum_{y} p(X = x|Y = y)\, p(Y = y). \tag{2.4}$$

In equation 2.3, the probability distribution $p(X = x)$ is also called the **marginal probability distribution** of X relative to the joint distribution $p(X = x, Y = y)$.

Given a set of n random variables $X_1, \ldots, X_n$ defined on the same sample space Ω, we call **chain rule** the following equation:

$$p(x_1, \ldots, x_n) = p(x_n|x_{n-1}, x_{n-2}, \ldots, x_1) \times \ldots \times p(x_2|x_1) \times p(x_1). \tag{2.5}$$

Another special function associated with any discrete random variable is called **cumulative distribution function** (or **cdf**). This function is defined for any real number x as the probability that the random variable will be less than or equal to x. The function notation F is often used to represent the cdf of a random variable. By definition, we can write

$$F_X(x) = p(X \leq x).$$

The above arguments assumes that the set of possible values of a random variable is finite and countable. If the set of possible values is uncountably infinite (i.e. real numbers), we speak instead of **continuous random variables**. Unlike their discrete counterparts, the associated probability function, called **probability density function** (or **pdf**), is a continuous function that describes the relative likelihood for the random variable to take on a given value. Given a continuous random variable X, its probability density function is typically indicated by $f_X(x)$. The area defined by the function in a precise range of values corresponds to the probability of

observing such values. More formally:

$$p\left(a \leq X \leq b\right) = \int_a^b f_x\left(x\right) dx.$$

This implies that $p\left(a\right) = \int_a^a f_x\left(x\right) dx = 0$ for each x in the space of the variable.

In order for the area of a *pdf* to correspond with a probability, some restrictions are necessary:

1. $f_X\left(x\right) \geq 0 \quad \forall x$
2. $\int_{-\infty}^{+\infty} f_x\left(x\right) dx = 1$

The cumulative distribution function for a continuous random variable is the continuous function defined by

$$F_X\left(x\right) = p\left(X \leq x\right),$$

for any real number x. Since the function F is defined as a probability, it only returns values in the range $[0, 1]$. Plugging any real number x into $F_X\left(x\right)$ returns the probability that the random variable X will have a value less than or equal to the number x. Note that $F_X\left(x\right)$ and $f_X\left(x\right)$ are closely related by this definition:

$$F_X\left(x\right) = \int_{t=-\infty}^{t=x} f_X\left(t\right) dt,$$

where t has been used as the dummy variable of integration to be perfectly correct from a calculus standpoint. From the other direction, we have

$$f_X\left(x\right) = \frac{dF_X\left(x\right)}{dx}$$

and so it is clear that given one of these functions, we can determine the other.

Now, suppose we have two random variables X and Y defined on the same probability space Ω. If for all values x of X and y of Y, the events $X = x$ and $Y = y$ are inde-

pendent, $p(X = x|Y = y) = p(X = x)$, then we say that X and Y are **independent**, and write $I(X, Y)$. If we have a third random variable Z, and whenever $p(z) \neq 0$ the events $X = x$ and $Y = y$ are conditionally independent given $Z = z$, $p(X = x|Y = y \cap Z = z) = p(X = x|Z = z)$, then we say that X and Y are **conditionally independent** given Z, and write $I(X, Y|Z)$.

Fundamentals of Statistics A *pmf* gives all the information we need about a discrete random variable, while a *pdf* does the same for a continuous random variable. We can calculate the probability of any event we want from them. However, to summarize the distribution more concisely, two common calculations can be made for a random variable: the **expected value** (or **mean**), and the **variance**.

The mean of a random variable is defined as the average value that would be observed for the random variable if it could be observed over and over again an infinite number of times. It is defined as

$$E[X] = \sum x \cdot p(X = x)$$

for a discrete random variable, and

$$E[X] = \int x \cdot f(x)\, dx$$

for a continuous random variable. The sum or integral is taken over all possible values of X. In either case, it is reasonable to interpret this calculation as taking the weighted average of the possible values of X, where the weights are the probabilities.

The variance of a random variable is a measure of how spread out its possible values are. A random variable will have a small variance if its possible values all fall in a small, tight range with high probability. It will have a large variance if its possible values are very spread out over a wide range and there is a reasonable chance than any of those values could be observed. The variance

calculation is

$$V[X] = \sum x^2 p(X = x) - E[X].$$

For a continuous random variable, we just replace summation with integration. Another equivalent formula for the variance which is sometimes easier to deal with is

$$V[X] = E\left[X^2\right] - (E[X])^2$$

A calculation related to the variance is called the standard deviation of a random variable. It is simply defined as the positive square root of the variance. This is actually used more often in practice than the variance because the units of the standard deviation calculation are the same as the units of the original variable, while the units of the variance are the square of the units of the original variable.

$$\sigma_X = \sqrt{V[X]}$$

Suppose now we have two random variables X and Y. Their relationship can be obtained through a measure called **covariance**, or $Cov(X, Y)$. It is defined as

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y].$$

The covariance itself does not convey much meaning concerning the relationship between X and Y. To accomplish this we compute the **correlation coefficient** from it. Suppose we have two discrete numeric random variables X and Y. Then the correlation coefficient $\rho(X, Y)$ of X and Y is given by:

$$\rho(X, Y) = \frac{Cov(X, Y)}{V[X]V[Y]}.$$

The correlation coefficient is always between -1 and $+1$. A value greater than 0 indicates the variables are positively correlated, and a value less than 0 indicates they are negatively correlated. By positively correlated we mean as one increases the other increases, while by negatively corre-

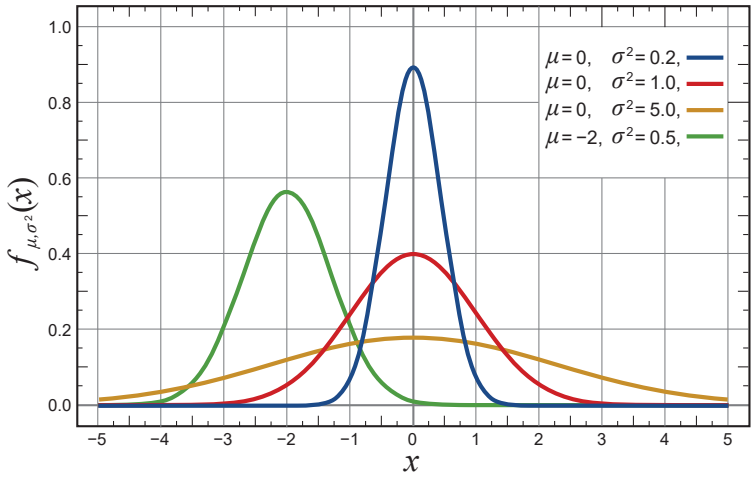


Figure 2.1: Examples of normal distributions with different values of μ and σ^2. In red the standard normal distribution is highlited.

lated we mean as one increases the other decreases. Finally if the two variables are independent then their correlation coefficient will be 0.

The Hypergeometric Distribution The hypergeometric distribution is a discrete probability distribution that describes the probability of k successes in n draws, without replacement, from a finite population of size N that contains exactly K successes, wherein each draw is either a success or a failure.

A random variable X follows the hypergeometric distribution if its *pmf* is given by

$$p\left(X = k\right) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

The *pmf* is positive when $\max\left(0, n + K - N\right) \leq k \leq \min\left(K, n\right)$.

The Normal Distribution The normal distribution (Figure 2.1) is a very common continuous probability distribution often used in the natural and social sciences to represent real-valued random variables whose distributions are not known [57].

The importance of the normal distribution is given by the central limit theorem that, in its

most general form, states the averages of random variables independently drawn from independent distributions converge in distribution to the normal. Since physical quantities are expected to be the sum of many independent processes often they have distributions that are nearly normal [58].

The probability density function of the normal distribution is:

$$f\left(x|\mu,\sigma\right) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Here, μ is the expectation of the distribution. The parameter σ is its standard deviation with its variance then σ^2. A random variable with a Gaussian distribution is said to be **normally distributed**. When $\mu = 0$ and $\sigma = 1$, the distribution is called **standard normal distribution** (Figure 2.1), $N\left(0,1\right)$. The normal distribution is also often denoted by $N\left(\mu,\sigma^2\right)$. Thus when a random variable X is distributed normally with mean μ and variance σ^2, we write $X \sim N\left(\mu,\sigma^2\right)$.

The distribution has the following properties:

- It is symmetric around its mean, which is at the same time its median, and it divides the data in half [59].
- It is unimodal: its first derivative is positive for $x < \mu$, negative for $x > \mu$, and zero only at $x = \mu$ [59].
- Its density has two inflection points (where the second derivative of f is zero and changes sign), located one standard deviation away from the mean [59].

2.1.2 Introduction to classification

Data classification has applications in a wide variety of fields. The problem attempts to learn relationship between sets of feature variables and targets variable. Many practical problems can

be expressed as associations between feature and target variables, this provides a broad range of applicability. The problem can be stated as follows:

Definition 2.3. Given a set of training data points along with associated training labels, determine the class label for an unlabeled test instance.

Classification algorithms typically contain two phases:

1. **Training Phase**: a model is constructed from the training data.

2. **Testing Phase**: the model is used to assign a label to an unlabeled test instance.

In some cases the training phase is omitted entirely, and the classification is performed directly from the relationship between a test instance and the set of training instances. Examples of such a scenario are the nearest neighbor classifiers, where the class of an instance is assigned by looking at its neighborhood.

The classification problem segments unseen test instances into groups, as defined by the class label. While the segmentation of examples into groups is also done by clustering, there is a key difference between the two problems. In the case of clustering, the segmentation is done by using similarities between feature variables, with no prior knowledge. In classification, segmentation is done on the basis of a training data set. As a result, the classification problem is referred to as **supervised learning**, while clustering as **unsupervised learning**.

Logistic Regression Typically data classification algorithms use statistical inference techniques to find the best class to which assign one instance, giving at the same time its assignment probability $p\left(Y|X\right)$, where Y is the class to which the instance is assigned and X are its features. Classification algorithms that probabilistically classify instances are called **probabilistic methods**. **Logistic regression** is a probabilistic approach used to determine $p\left(Y|X\right)$, when Y is a discrete value and $X = \{X_1, \ldots, X_d\}$ is a vector containing both discrete and continuous variables. For simplicity, here we will describe only the case in which $Y \in \{0, 1\}$. Formally, the logistic

regression model is defined as:

$$p\left(Y=1|X\right)=g\left(\theta^{T}X\right), \tag{2.6}$$

where

$$g\left(z\right)=\frac{1}{1+e^{-z}} \tag{2.7}$$

and

$$\theta^{T}X=\theta_{0}+\sum_{i=1}^{d}\theta_{i}X_{i}. \tag{2.8}$$

$g\left(z\right)$ is called the **logistic function**. As the sum of probabilities must be equal to 1, $p\left(Y=0|X\right)$ can be calculated as $1-p\left(Y=1|X\right)$. Because logistic regression predicts probabilities, rather than just classes, we can use the log-likelihood as:

$$\log\left(\frac{p\left(Y=1|X\right)}{1-p\left(Y=1|X\right)}\right)=\log\frac{g\left(\theta^{T}X\right)}{1-g\left(\theta^{T}X\right)}=\theta^{T}X. \tag{2.9}$$

When classifying using logistic regression, the training phase consists in determining the parameters θ which minimizes classification error (eg. rate of incorrectly classified data points). Therefore, this parameter vector can be used to classify unlabeled instances.

Nearest Shrunken Centroids Algorithm A different approach to classification is the **nearest shrunken centroid** methodology [60]. It represents the data points in a class through a centroid, to which de-noising procedure is applied to improve the predictive value of the method.

Let x_{ij} be the value of i-th feature ($i=1,2,\ldots,p$) of sample j ($j=1,2,\ldots,n$). Suppose, also, to have K classes and to indicate with C_k and n_k respectively the set and the number of samples in class k. The i-th component of class k centroid is calculated as the average value of the i-th feature for all elements of class k, $\bar{x}_{ik}=\sum_{j\in C_k}x_{ij}/n_k$, while the overall centroid is the mean of the i-th feature for all values, $\bar{x}_i=\sum_{j=1}^{n}x_{ij}/n$.

The method consists in shrinking class centroids toward overall centroid after a standardizing

by the standard deviation of each feature of a class. Standardization has the effect of giving greater weight to those components whose intra-class values are more stable. Let

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k \cdot (s_i + s_0)}, \tag{2.10}$$

where

$$s_i^2 = \frac{1}{n-K} \sum_k \sum_{j \in C_k} (x_{ik} - \bar{x}_{ik})^2, \tag{2.11}$$

$m_k = \sqrt{1/n_k - 1/n}$, and s_0 is equal to the median value of s_i over the set of features.

Thus d_{ik} can be interpreted as a *t statistic* that compares class k to the overall centroid for feature i. We can therefore rewrite equation 2.10 as:

$$\bar{x}_{ik} = \bar{x}_i + m_k \, (s_i + s_0) \, d_{ik}, \tag{2.12}$$

The method shrinks d_{ik} toward zero, giving d'_{ik} and yielding **shrunken centroids** or **prototypes**:

$$\bar{x}'_{ik} = \bar{x}_i + m_k \, (s_i + s_0) \, d'_{ik}. \tag{2.13}$$

Such a shrinkage is obtained through a process of *soft thresholding* as:

$$d'_{ik} = sign \, (d_{ik}) \, (|d_{ik}| - \Delta)_+ \, , \tag{2.14}$$

where $t_+ = t$ if $t > 0$ and zero otherwise, and Δ is set so as to reduce classification error. The method just described has the property to eliminate many irrelevant components from class prediction by increasing Δ parameter.

Suppose now we have an unclassified sample $x^\star = \left\{x_1^\star, x_2^\star, \ldots, x_p^\star\right\}$. Authors define a

discriminant score for class k as:

$$\delta_k\left(x^\star\right) = \sum_{i=1}^{p} \frac{\left(x_i^\star - \bar{x}'_{ik}\right)^2}{\left(s_i + s_0\right)^2} - 2\log \pi_k, \tag{2.15}$$

where $\pi_k = {}^{n_k}/_{n}$ is the class prior probability, which gives the overall frequency of class k. A sample is then classified as the class t which minimizes $\delta_t\left(x^\star\right)$.

2.1.3 Introduction to recommendation systems

Fundamentals of recommendation systems Recommendation algorithms are a class of systems for information filtering whose main objective is the prediction of users' preferences for some objects. In recent years, they have become commonly used and applied in various fields. Their main application lies in e-commerce in the form of web-based software. However, they have been successfully employed in other areas related, for example, to bioinformatics [30, 61].

A recommendation system consists of users and objects. Each user collects some objects, for which he can also express a degree of preference. The purpose of the algorithm is to infer the user's preferences and provide scores to objects not yet owned, so that the ones, which most likely will appeal the user, will be rated higher than the others.

In a recommendation system, we denote the set of objects as $O = \{o_1, o_2, \ldots, o_n\}$ and the set of users as $U = \{u_1, u_2, \ldots, u_m\}$. The whole system can be fully described by a sparse matrix $T = \{t_{ij}\}_{n\times m}$ called utility matrix. In such a matrix, t_{ij} has a value if and only if the user u_j has collected and provided feedback on the object o_i. In the event that users can only collect objects without providing any rating, the system can be described by a bipartite graph $G\left(O, U, E\right)$ where $E = \{e_{ij} : o_i \in O, u_j \in U\}$ is the set of edges. Each edge indicates that a user has collected an object. This graph can be described in a more compact form by means of an adjacency matrix $A = \{a_{ij}\}_{n\times m}$, where $a_{ij} = 1$ if u_j collected o_i, and $a_{ij} = 0$ otherwise. A reasonable assumption in this case is that the objects collected by a user corresponds to his preferences, and the recommendation algorithm aims to predict users' views on other items.

Up to now, the algorithm mostly applied in this context is collaborative filtering (CF) [62, 63]. It is based on a similarity measure between users. Consequently, the prediction for a particular user is computed employing information provided by similar ones. A Pearson-like evaluation is typically employed to evaluate similarity between two users:

$$s_{ij} = \frac{\sum_{l=1}^{n} a_{li} a_{lj}}{\min\left\{k\left(u_i\right), k\left(u_j\right)\right\}}, \tag{2.16}$$

where $k\left(u_i\right)$ is the number of items collected by the user u_i. For any user-object pair $(u_i - o_j)$, if not already collected ($a_{ij} = 0$), a predicted score v_{ij} can be computed as:

$$v_{ij} = \frac{\sum_{l=1, l\neq i}^{m} s_{li} a_{jl}}{\sum_{l=1, l\neq i}^{m} s_{li}}. \tag{2.17}$$

Two factors influence positively v_{ij}: objects collected from a large number of users, objects collected frequently from users very similar to u_i. The latter correspond to the most significant predictions. All items are then sorted in descending order using their prediction score, and only those at the top will be recommended.

Evaluating recommendation systems Verifying the reliability of a recommender system result is typically a complex phase. A basic evaluation strategy considers it as a classification algorithm that distinguishes, for each user, liked objects from un-liked ones. We can then apply traditional metrics such as mean squared error or receiver operating characteristic curves to evaluate results. Another strategy is to define new metrics specifically designed to assess performances of a recommendation system [64].

In common between the two approaches is the application of a k-fold cross-validation to obtain a more accurate estimate of methods reliability. The set of all user-object preferences is randomly partitioned into k disjoint subsets. One is selected as a test set, and the recommendation algorithm is applied to the others. Evaluation metrics are, then, computed, using the test

set as a reference. The process is repeated until all the partitions have been selected as test set, and the results of each metric are averaged in order to obtain an unbiased estimate of the quality of the methodology.

Four metrics have been specifically developed to assess the quality of a recommender algorithm. Two measure performances in terms of predictions accuracy, by measuring the capability of recovering interactions in the test set, while two measures recommendation diversity:

Recovery of deleted links, r. An accurate method typically will place potentially preferable objects higher than non-preferable ones. Assuming that a user has collected only liked items, the pairs present in the test set, in principle, should have a higher score than the others. Therefore, by applying the recommendation algorithm and computing the sorted set of predictions for a user u_j, we can compute a relative rank for an uncollected object o_i, whose position in the list is p, as:

$$r_{ij} = \frac{p}{o - k_j}. \tag{2.18}$$

Such a rank should be smaller if the pair $u_j - o_i$ is part of the test set. The recovery (r) corresponds to the average of such relative ranking for all user-object pairs in the test set. The lower its value, the greater is the ability of the algorithm to recover deleted interactions, and therefore to achieve accurate results.

Precision and recall enhancement, $e_P(L)$ and $e_R(L)$. Typically, only the highest portion of the recommendation list of a user is employed for further purposes, which is why a more practical measure of the reliability of a recommendation system may consider only the Top-L predictions. For a user u_i, let D_i be the number of deleted object for user u_i, and $d_i(L)$ the ones predicted in the Top-L places. An average of the ratios $d_i(L)/L$ and $d_i(L)/D_i$ for all users with at least one object in the test set, correspond, respectively, to the precision $P(L)$ and recall $R(L)$ for the recommendation process [63, 65].

We can get a better perspective by considering these values with respect to random model.

Let $P_{rand}(L)$ and $R_{rand}(L)$ be, respectively, the precision and the recall of a recommendation algorithm that randomly assign scores to user-object pairs. If the user u_i has a total of D_i objects in the test set, then $P^i_{rand}(L) = {}^{D_i}/_{(o-k_i)} \approx {}^{D_i}/_{o}$, since the total number of objects is much greater than the number of collected ones. Averaging for all users, we obtain $P_{rand}(L) = {}^{D}/_{ou}$, where D is the size of the test set. By contrast, the average number of deleted objects in the Top-L positions is given by ${}^{L \cdot D_i}/_{(o-k_i)} \approx {}^{L \cdot D_i}/_{o}$ and, therefore, $R_{rand}(L) = {}^{L}/_{o}$. We can now define precision and recall enhancement as:

$$e_P(L) = \frac{P(L)}{P_{rand}(L)} = P(L) \cdot \frac{ou}{D}, \tag{2.19}$$

$$e_R(L) = \frac{R(L)}{R_{rand}(L)} = R(L) \cdot \frac{o}{L}, \tag{2.20}$$

A high value of precision enhancement indicates that the fraction of relevant predictions made by the algorithm is substantially higher than a completely random one. A high recall enhancement indicates that the percentage of correct predictions is significantly higher than the null model.

Personalization, $h(L)$. A first measure of diversity to consider when evaluating a recommendation algorithm is the uniqueness of the predictions made for different users, namely the inter-user diversity. Given two users u_i and u_j, a measure of inter-list distance can be computed as:

$$h_{ij}(L) = 1 - \frac{q_{ij}(L)}{L}, \tag{2.21}$$

where $q_{ij}(L)$ is the number of common Top-L predictions between the two users. It follows immediately that this distance has a value 0 if the two users have the same prediction, 1 in the case of completely different lists. The average distance calculated for all possible pairs of users corresponds to the personalization metric. Higher, or lower, values correspond, respectively, to a greater, or lesser, diversity of recommendations.

Surprisal/novelty, $I(L)$. Evaluating the ability of a recommendation system to generate novel and unexpected predictions is a key measure. In this context, we define as unpredictability of results, the ability to suggest items for which it is very unlikely that a user may already know them. To measure this, we use the concept of *self-information* or *surprisal* [66], which determines how unexpected is an object with respect to its global popularity. Given an object o_j, the probability that a user has collected it is given by $k(j)/m$. Its self-information is therefore $I_j = \log_2 (m/k(j))$. The average of such values for the Top-L predictions of a user u_i correspond to its self-information, $I_i(L)$. By averaging for all users, we get a measure of the global surprisal $I(L)$.

In classical applications, a value L equal to 30 is chosen a priori. In any case, no variations in the relative performances of the algorithms can be observed by varying L, as long as its value is significantly smaller than the number of objects in the system.

2.1.4 Text Annotators: TAGME

A typical paradigm used to simplify the representation in information retrieval (IR) problems is the bag-of-words model. A text is represented as a multi-set of his words, ignoring grammar and arrangement, but keeping the multiplicity. Recently, much work has been done to overcome such simplistic model with the aim to improve the search of information within textual data. In this sense, the identification of sequences of words (**spots, anchors**) in an input text and their annotation with unambiguous entities that describe their meaning is a fundamental step. Two main approaches have been used to annotate texts with such information.

The first approach extends the traditional topic-based vector-space model, which consists of a space with d dimensions where each one is a key topic, with additional dimensions from an external knowledge base, such as Wikipedia or the entire Web. This approach extends bag-of-words model with more concepts, but a big problem is posed by the contamination with unrelated concepts.

To overcome these limitations, a second approach tries to annotate only text fragments that contain salient concepts without using a vector-space model. The basic idea is to identify in the text sequences of short and meaningful terms and connect them to a concept unambiguously obtained from a catalog. The catalog can be either a limited set of concepts or a knowledge base such as Wikipedia. *TAGME* [67, 68] leverages this idea to obtain a fast and accurate annotation of a short text.

To speed up the processing of texts, *TAGME* indexes some information taken from Wikipedia.

- **Anchor dictionary**: an index of all anchors present in the Wikipedia pages, augmented with the titles of redirect pages plus some variants of the page titles.
- **Page catalog**: an index of all Wikipedia pages, without disambiguation pages, list pages, and redirect pages.
- **In-link graph**: a directed graph whose nodes are the pages in the Page Catalog, and whose edges are the links among these pages.

TAGME uses these data structures to annotate a short text via three main steps: (i) anchor parsing, (ii) disambiguation and (iii) pruning.

Anchor parsing The algorithm first receives a short text in input, tokenizes it, and then detects the anchors by querying the Anchor dictionary for sequences of up to 6 words. **Anchor disambiguation** Then the algorithm cross-references each of these anchors with one pertinent sense drawn from the Page catalog. This is done by means of a simple and fast *disambiguation score* that take into account the sparseness of the anchors and the possible lack of unambiguous anchors in short texts. **Anchor pruning** The disambiguation phase produces a set of candidate annotations, one per anchor detected in the input text. This set is pruned in order to discard the un-meaningful annotations. These annotations are detected by means of a scoring function that takes into account two features: the link probability of an anchor and the coherence between a

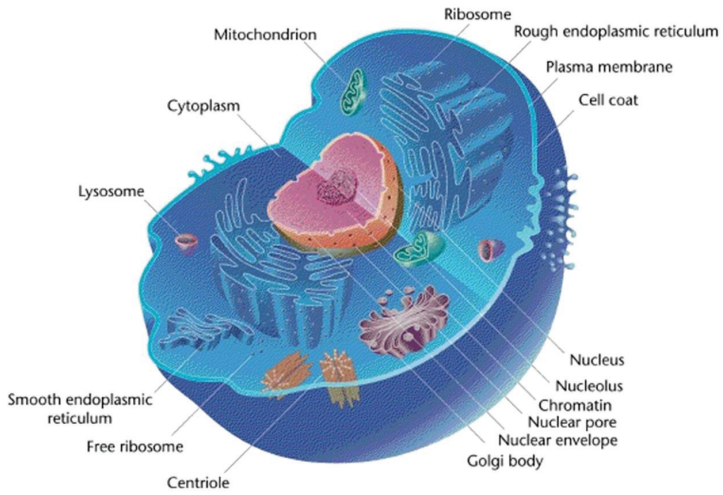


Figure 2.2: A typical animal cell with its organelles. Eukaryotic cells are typically much larger than prokaryotic ones. They have a variety of internal membrane-bound structures, called organelles, and a cytoskeleton composed of microtubules, microfilaments, and intermediate filaments, which play an important role in defining the cell's organization and shape. The most notable organelles are: the nucleus, surrounded by a double membrane with pores, that contains all the genetic material of the cell; the ribosomes involved in making protein; the mitochondria which are the powerhouse of the cell as they provide energy to the cell in the form of ATP. Image courtesy of [69].

candidate annotation and the others. The resulting annotations are returned by the algorithm along with scores that identify their reliability.

2.2 Biological Prerequisites

This section is a brief introduction to the necessary background in biology and genetics.

2.2.1 Genes, proteins and the central dogma of biology

The cell is the basic structural and functional unit of all organisms. A single cell is the lowest form of life thought to be possible. Most organisms consist of more than one cell, each of which becomes specialised into particular functions towards the cause of the entire organism (such as liver cells, skin cells, etc). Cells possess many membrane-bound structures inside them, called **organelles**. Figure 2.2 shows an overview of the various structures inside an animal cell.

In the cell, which is surrounded by an outer membrane, we find the **nucleus**. It contains **DNA** (Figure 2.3), a biomolecule where all instructions for building **Proteins** are deposited. Proteins

are the main actuators of a cell. Portions of DNA that encode for proteins are also called **genes**. The production of a protein occurs in a two step process. First the DNA is **transcribed** into **RNA**, a molecule that is responsible to transfer the instructions to the **ribosomes**. Subsequently, in a process called **translation**, an RNA molecule is used as a mold by ribosomes to assemble a protein. An active gene is said to be **expressed**, and this process is called **gene expression**. The **central dogma of molecular biology** defines precisely the flow of genetic information in a biological system (Figure 2.4). Citing Crick, who stated it first in 1956 [70, 71]:

> The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid. – *Francis Crick*

The central dogma has also been described as **DNA makes RNA and RNA makes protein** [72] a positive statement which was originally termed *the sequence hypothesis* by Crick. However, this simplification does not make it clear that the central dogma as stated by Crick does not preclude the reverse flow of information from RNA to DNA, only ruling out the flow from protein to RNA or DNA (Figure 2.4).

DNA The *DNA* (**deoxyribonucleic acid**) is the molecule that carries most of the genetic instructions used in the development, functioning and reproduction of all living organisms. A typical *DNA* molecule consists of two paired filaments (*strands*) that coil forming a double helix structure (Figure 2.5) [75]. Each *DNA* strand is composed of simpler structures called **nucleotides**. Each nucleotide consists of a nitrogenous base, a pentose sugar (called **deoxyribose**), and at least one phosphate group. The nitrogenous bases, which characterize each nucleotide, are of four types: **cytosine** (C), **guanine** (G), **adenine** (A), and **thymine** (T). Each nucleotide is linked to the next via a chain of covalent bonds between its sugar and the phosphate group of the next one, forming the so-called **backbone** of the *DNA* molecule. The nitrogenous bases of a strand form hydrogen bonds with those of the adjacent one by following precise rules of pairing, called

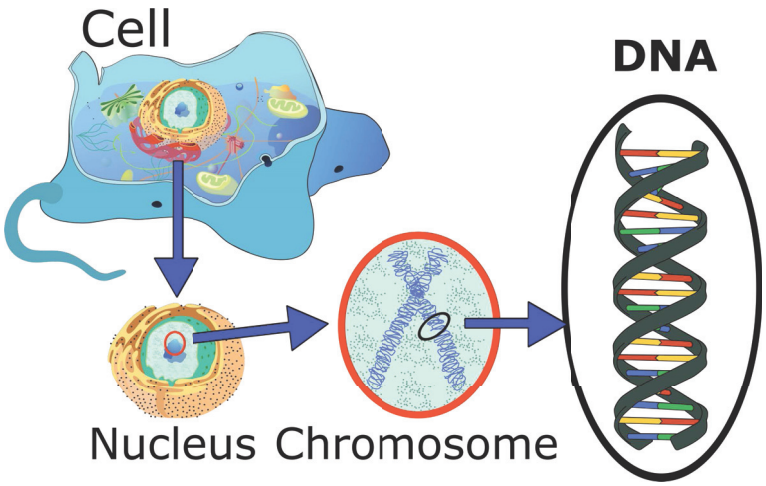


Figure 2.3: The DNA in an eukaryotic cell. DNA usually occurs as *linear chromosomes* in eukaryotes, and *circular chromosomes* in prokaryotes. The set of chromosomes in a cell makes up its **genome**; the human genome has approximately *3 billion base pairs* of DNA arranged into *46 chromosomes*[73]. Image courtesy of [74].

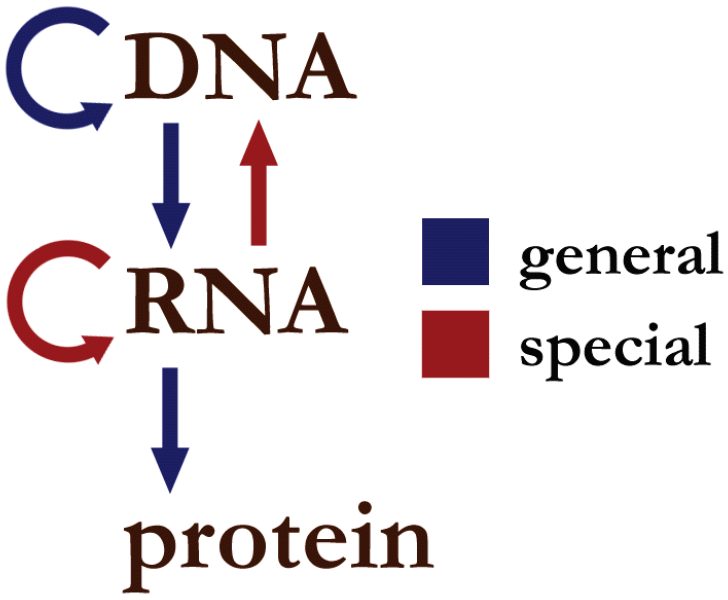


Figure 2.4: Information flow in biological systems.

canonical pairings or **Watson-Crick pairings**. The rules of pairing states that *A* binds with *T* and *C* with *G*. The latter pairing is more stable because of the greater strength of the bond.

The main task of a *DNA* molecule is *storing biological information*. For this reason, the backbone is resistant to cuts, and both strands of the molecule, thanks to the rules of base pairing, contain a copy of the same information.

In the cell nucleus, the *DNA* is organized into long structures called **chromosomes** (Figure 2.6). Typically eukaryotic cells contain *two copies* of the same linear chromosome (**diploid cells**), each coming from one of the two parents. During cell replication (**mitosis**), the chromosomes are duplicated in a copying process, which gives each cell its own set of genetic material. Some organelles, such as mitochondria and chloroplasts have their own *DNA* inside. Special structures called **histones** (Figure 2.7) are used to compact and organize the genetic material. *DNA*, packaged inside the nucleus by histones, forms a unit called **nucleosome**. Packed and condensed *DNA* forms the **chromatin**.

In prokaryotes, the genome is typically stored in a *single circular* chromosome, although small circular chromosomes (**plasmids**) are used as a complement, and are transferable between species.

DNA was isolated for the first time by *Friedrich Miescher* in *1869* and its molecular structure was identified by *James Watson* and *Francis Crick* in *1953* [75].

RNA *RNA* (**ribonucleic acid**) is a molecule involved in the encoding, decoding, adjustment, and expression of the information contained in the *DNA* of an organism (Figure 2.8). Like *DNA*, *RNA* is formed by a chain of nucleotides. Some *RNA* molecules also have an active role within the cell helping to catalyze biological reactions, or by identifying and communicating cellular signals.

The chemical structure of *RNA* is very similar to that of *DNA*, but some major differences are observable:

- *RNA* is a single stranded molecule typically much shorter than *DNA*. However, it can fold back on itself to form very complex structures (**secondary structure**) [78], which unlike

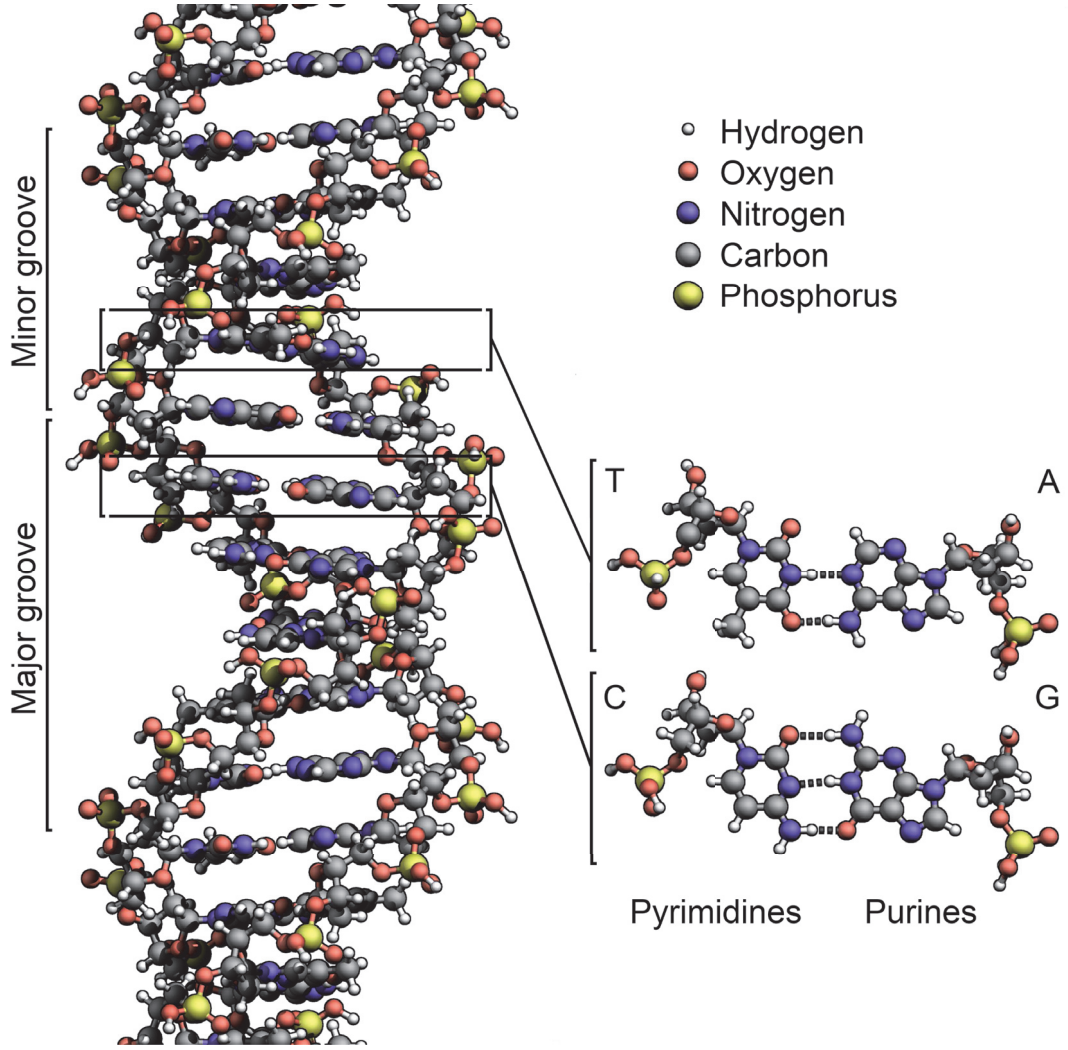


Figure 2.5: The structure of the DNA double helix. The atoms in the structure are colour-coded by element and the detailed structure of two base pairs are shown in the bottom right. Image courtesy of [76].

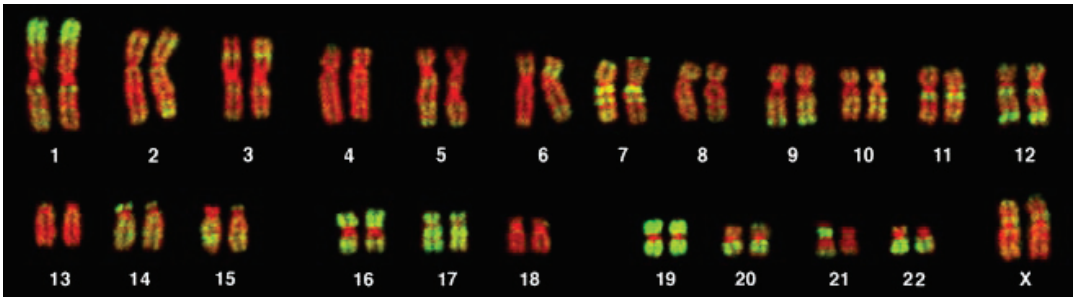


Figure 2.6: Fluorescent microscopy image of a human female karyotype, showing 23 pairs of chromosomes. Marked in yellow, we can see some genes. The largest chromosomes are around 10 times the size of the smallest. Chromosomes typically have a linear form. In this particular image the cell is preparing to divide so for each chromosome a copy was produced forming the classic X-shape. Image from Bolzer et al. [77].

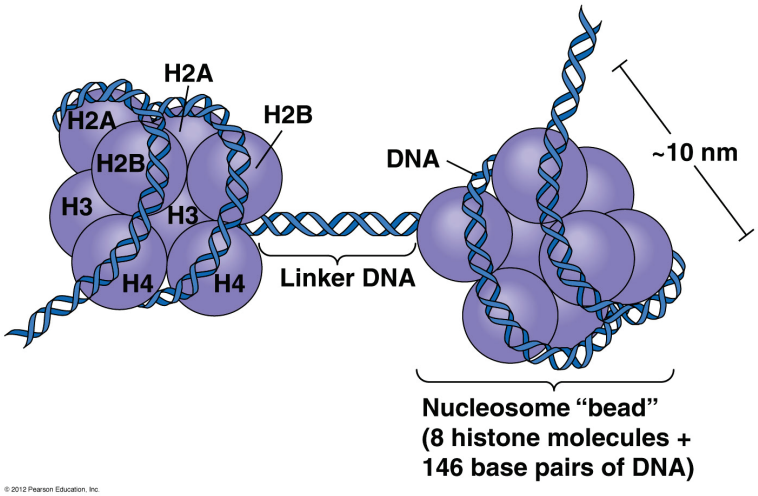


Figure 2.7: A nucleosome that is, a complex of histones with the DNA molecule coiled around them.

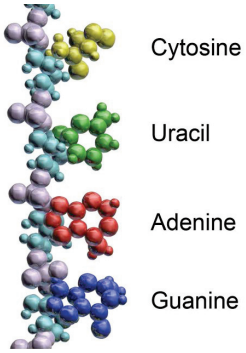


Figure 2.8: Bases in an RNA molecule.

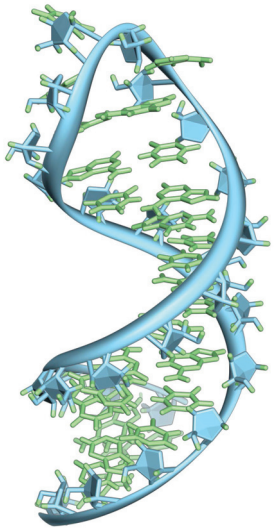


Figure 2.9: An example of *RNA* secondary structure. Highlighted are the bases (green) and the ribose-phosphate backbone (blue). Image courtesy of [80].

DNA can also carry out very complex functions [79] (Figure 2.9).

- The sugar that forms the nucleotides of *RNA* is a ribose, instead of a deoxyribose in *DNA*. This makes *RNA* less stable and easier to degrade.
- Thymine is replaced by **uracil** (*U*) of which it is a modified form.

In one cell there are many types of *RNA*, each processed differently, depending on the use made of it.

Genes A **gene** is a *DNA* region encoding a functional *RNA* or protein, and it is the molecular unit of heredity [81, 82] (Figure 2.10). The transmission of a gene to the offspring of an organism is the basis of the inheritance of phenotypic traits. Some traits are immediately visible, such as eye color, others are not, such as blood type. Genes can acquire mutations in the sequence, called **alleles**.

The set of genes in an eukaryotic organism or cell is known as its **genome**. It can be stored in one or more **chromosomes** (Figure 2.6). A chromosome is a long double-stranded *DNA* filament

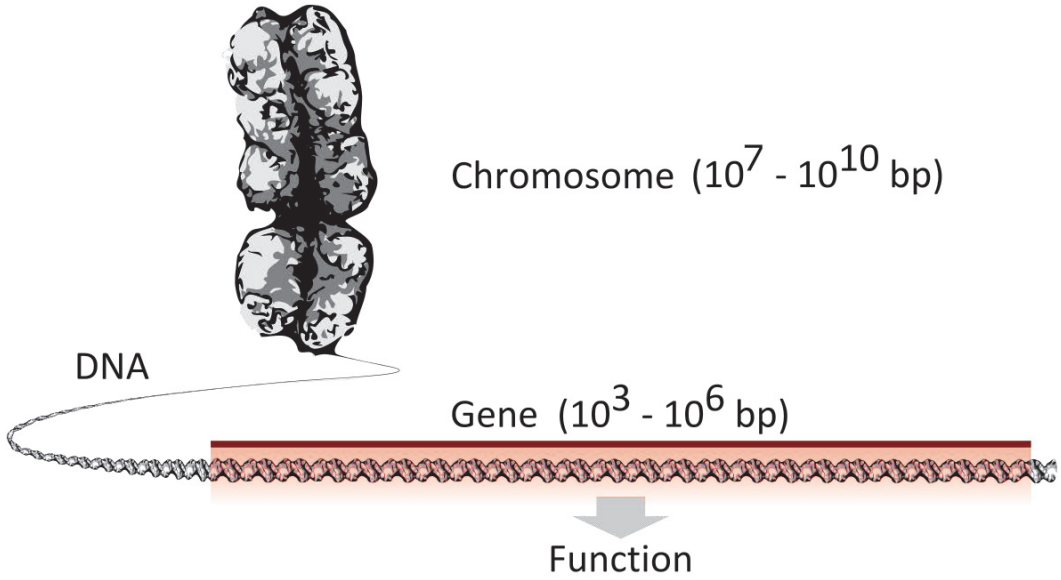


Figure 2.10: A gene is a segment of *DNA* that encodes function. A chromosome consists of a long strand of *DNA* containing many genes. A human chromosome can have up to 500 million base pairs of *DNA* with thousands of genes. Image courtesy of [83].

in which thousands genes are packed. The region of chromosome in which a specific gene is located is called its **locus**. Each locus contains an allele of the gene. Changes to the structure of chromatin affect gene activity, making genes more or less accessible to cell machinery.

Eukaryotes typically have many regions of *DNA* that have no apparent function (**junk DNA**) although recently many of such regions have been associated with the processes of gene activity regulation. In humans, only 2% of the genome is constituted by genes [84]. Differently from eukaryotes, prokaryotes chromosomes, probably due to the shorter genome, are typically extremely dense in genes, containing only a few regions that have no apparent function.

Proteins **Proteins** are large molecules that consist of one or more chains of **amino acids** (or **residues**). Proteins are the functional units of living organisms, and are responsible for conducting the majority of biological functions. A protein mainly differ from another in amino acidic sequence, which is typically folded to form a specific three-dimensional shape that determines its activity (Figure 2.11). A linear chain of amino acids is called a **polypeptide**. A protein consists

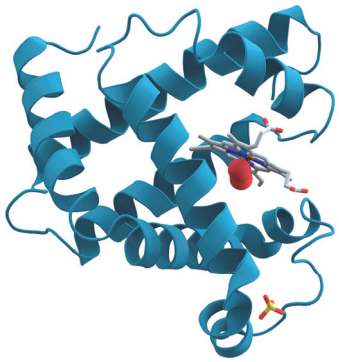


Figure 2.11: A representation of the 3D structure of the protein myoglobin. This protein was the first to have its structure solved by X-ray crystallography. Towards the right-center among the coils, a prosthetic group called a heme group (shown in gray) with a bound oxygen molecule (red).

of a long polypeptide chain.

The sequence of amino acids that makes up a protein is defined by the sequence of a particular gene. Typically there are 20 amino acids (A complete list is available in figure 2.12). Shortly after or even during synthesis, the residues in a protein are often chemically modified by post-translational modification, which alters its function. Sometimes proteins have non-peptide groups attached, which can be called **prosthetic groups** or **cofactors** (Figure 2.11). Proteins mainly work together to achieve a particular function, forming protein **complexes**.

Once formed, proteins only exist for a certain period of time, and are then degraded and recycled by the cell machinery through the process of protein turnover. A protein's lifespan is measured in terms of its half-life and covers a wide range (from minutes to years). Abnormal or mis-folded proteins, typically, are degraded more rapidly either due to being targeted for destruction or due to being unstable.

Transcription and Translation The protein production process, or expression of a gene, is composed of three phases: **transcription**, **mRNA maturation**, and **translation** (Figure 2.13). However, before talking about such a process, more details on the structure of a gene must be given.

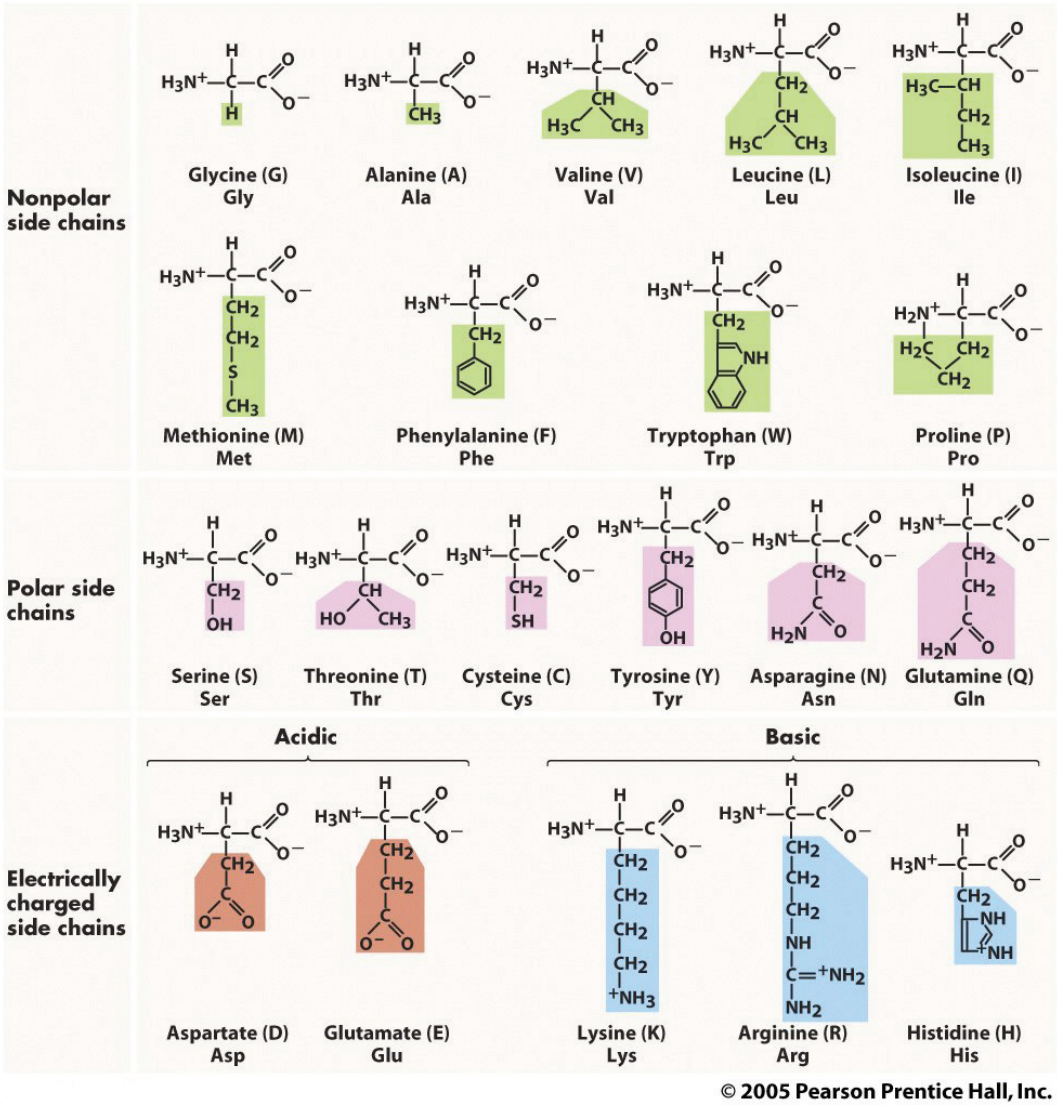


Figure 2.12: A complete list of the 20 amino acids grouped by their chemical properties.

A gene consists of many parts (see Figure 2.13), of which the portion used to produce a protein is typically small. A gene is enclosed between a **promoter** and a **terminator**. The promoter is the region in which the enzymes involved in the transcription bind to begin the process. The terminator is the portion in which the *RNA* synthesis process ends. An *RNA*, which provides instructions to build a protein, is called **messenger RNA** or **mRNA**. Downstream and upstream of a gene special sequences called **enhancer**, or **silencer**, may be present. They are used to facilitate or prevent the transcription process when bound to special proteins called **Transcription Factors**. Upstream and downstream of the region that contains the instructions for protein production there are two portions which are also transcribed, but not translated, called $5'$ **untranslated region** ($5'$ **UTR**) and $3'$ **untranslated region** ($3'$ **UTR**). The first has an important role in the regulation of the protein production process. The latter also has the task of regulating the translation efficiency, the stability of the messenger, and its location. Finally, the region coding for the protein is composed of portions that contain instructions, called **exons**, and unused portions, called **introns**, which may also contain other elements, such as other genes.

As mentioned previously, the DNA consists of two complementary and antiparallel strands. We denote each end of a strand with the symbols $5'$ and $3'$, respectively. Therefore a strand will have direction $5' \rightarrow 3'$, while the other $3' \rightarrow 5'$. A gene is typically present in one strand, or the **coding strand**. The opposite one is called the **template strand**, because it is used as a prototype for mRNA transcription.

The transcription is done in the nucleus by a particular family of molecular machinery called **RNA polymerase**. RNA polymerase binds to a promoter to begin the RNA production process, which continues by adding one nucleotide at a time up to the terminator. In this phase thymine are converted into uracils.

The RNA thus obtained is processed in a maturing phase. Initially the $5'$ UTR region is amended by adding molecules that serve to stabilize it ($5'$ **capping**). So the $3'$ UTR is cut in the terminal portion removing some nucleotides, and adding a long tail of adenine (**poly-A tail**) in

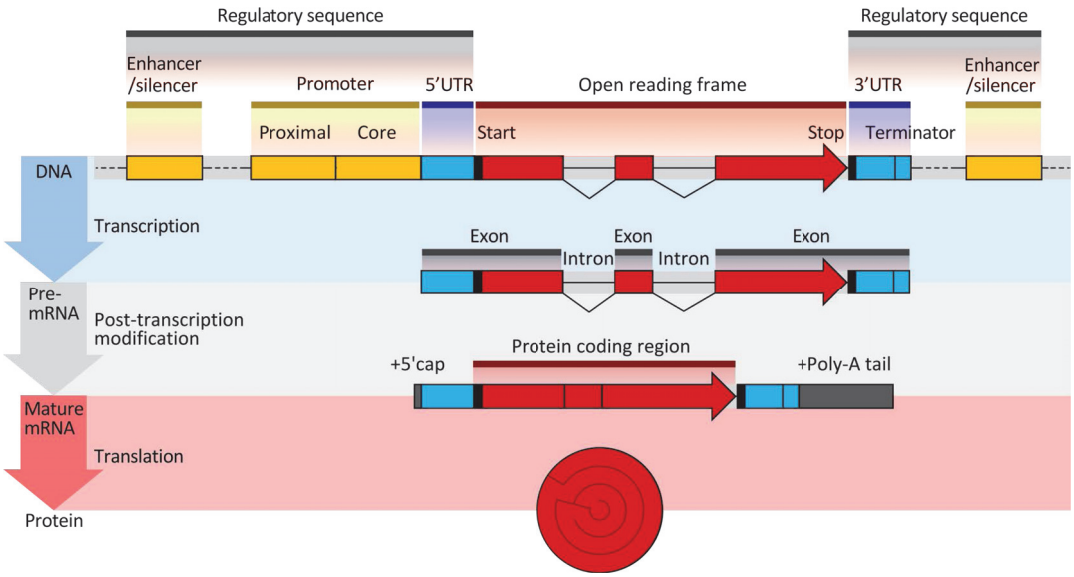


Figure 2.13: The structure of an eukaryotic protein-coding gene. Regulatory sequence controls when and where expression occurs for the protein coding region (red). Promoter and enhancer regions (yellow) regulate the transcription of the gene into a pre-mRNA which is modified to add a $5'$ cap and poly-A tail (grey) and remove introns. The mRNA $5'$ and $3'$ untranslated regions (blue) regulate translation into the final protein product.

a process called **polyadenylation**. Finally, introns are removed in the messenger (**RNA splicing**), and the mature mRNA is transported into the cytoplasm where the translation process begins.

Translation, operated by ribosomes, uses mRNA as a mold to produce a polypeptidic sequence. A ribosome is composed of two subunits (large and small). When the ribosome is not acting the two subunits are detached. The translation process itself is accomplished by particular RNA molecules called **transfer RNA** (or **tRNA**). Each tRNA associates a precise amino acid to a specific sequence of three nucleotides, called **codon**. The code that is thus formed is called the **genetic code** (Figure 2.14) and is universal to all organisms, with some small variations.

In short, the translation is made up of four stages:

1. **Initiation**: The ribosome is assembled around an mRNA and the first tRNA is attached to the start codon.

2. **Elongation**: The tRNA transfers an amino acid to the tRNA corresponding to the next codon.

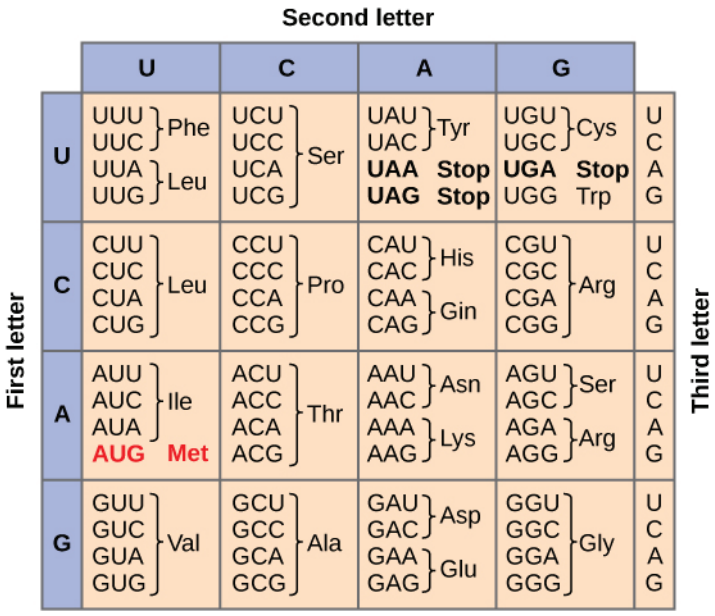


Figure 2.14: This figure shows the genetic code for translating each nucleotide triplet in mRNA into an amino acid or a termination signal in a nascent protein. Highlighted in red is the start codon that encodes always for methionine.

3. **Translocation**: The ribosome moves to the next codon of the mRNA to continue the process, creating a chain of amino acids.

4. **Termination**: When a stop codon is reached, the ribosome releases the polypeptide.

2.2.2 Non-Coding RNAs

A **non-coding RNA** (*ncRNA*) is an *RNA* molecule that is not translated into a protein. The types of *ncRNAs* are very abundant and functionally important. These include the **transfer RNA**, which we discussed in the previous section, the **microRNA**, core post-transcriptional regulators of gene expression, and **long ncRNAs**, whose function is not yet fully understood, although today some of them were associated with the onset or progression of diseases.

The first *ncRNA* to be discovered was a *tRNA* in yeast, and its structure was published in 1965 [85]. To date, the number of known *ncRNA* has exceeded one hundred thousand in human.

The biological roles assumed by these molecules are varied, and affect processes of fundamental

importance for the cell. ncRNAs have been associated to processes such as translation of proteins, *RNA* splicing, cell replication, regulation of gene expression, genome defense, and chromosomal structure maintenance.

Long non-coding RNAs

Long non-coding RNA (*lncRNA*) are *RNA* molecules typically longer than 200 nucleotides. A recent study showed that only one fifth of human genome transcription activity is linked to protein-coding genes. Many studies also suggest that the brain and the central nervous system express many more *lncRNA* than any other type of tissue.

lncRNAs have been associated with many cellular processes, although still the way in which they act are not completely clear. Some *lncRNAs* regulate the transcription process by serving as activators of transcription factors, or by remodeling the chromatin structure and, therefore, making areas of *DNA* more accessible for transcription. By pairing with *mRNA* molecules, some *lncRNAs* can post-transcriptionally adjust expression, by masking certain parts of *mRNA* unaccessible for cellular machinery, causing, therefore, its degradation.

Many other features and roles in various diseases have been associated with these molecules. However the knowledge on *lncRNAs* is still incomplete and numerous studies are underway to determine their exact number and function.

Micro RNAs

MicroRNAs (*miRNAs*) are small non-coding RNA molecules of about 22 nucleotides found in plants, animals, and some viruses. Their main function is the regulation of gene expression by post-transcriptional silencing [86, 87]. They act through base pairing with the complementary sequence of a *mRNA* molecule (Figure 2.15). As a result of such pairing, the molecule can be silenced by one of the following processes: cutting the *mRNA* in pieces, destabilization of the molecule by shortening of the poly-A tail, reduction in the efficiency of translation process [88, 89]. Classical characteristic of *DNA* regions that encodes for *miRNA* is their hairpin shape

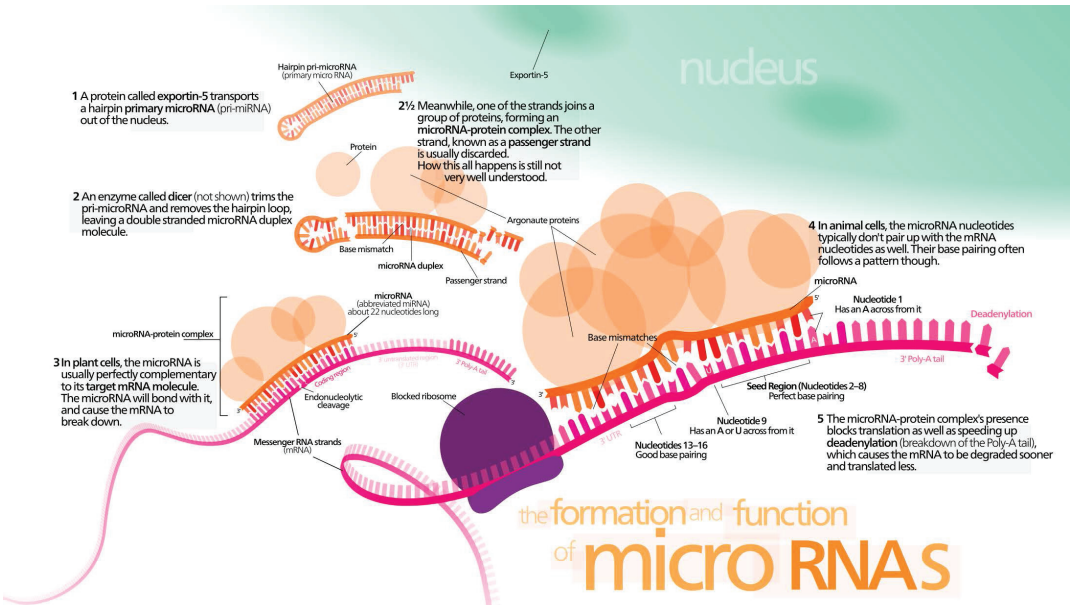


Figure 2.15: Diagram of *miRNA* action with *mRNA*. Image courtesy of [90].

after the transcription [87]. The human genome encodes about 1,800 *miRNAs*. An aberrated expression of these molecules was related to the onset of many disease states, and *miRNA*-based therapies are being studied.

MiRNAs are produced by specific genes or by introns of other genes. Sometimes a *miRNA* gene is transcribed together with its target, in a form of co-regulation [91]. Other *miRNAs* have instead a common promoter and arise from a single unit containing more hairpins.

Once transcribed by RNA polymerase, the **primary miRNA** (*pri-miRNA*) is processed by the **Drosha** enzyme to free the hairpin from a *pri-miRNA*, which can contain more than one [92]. The transcript obtained is called **precursor miRNA**, or *pre-miRNA* (Figure 2.16). The precursor is, therefore, exported from the nucleus through the nucleocytoplasmic protein **Exportin-5** [93]. In the cytoplasm the hairpin is cut by the **Dicer** enzyme, and the *miRNA* is released [94]. Such matured *miRNA*, along with Dicer and other protein, forms the **RNA-induced silencing complex**, or *RISC* [95]. Members of the **Ago** protein family are critical to the function of *RISC*. They are capable of binding with the *miRNA* and orient the interaction with its target [96]. The

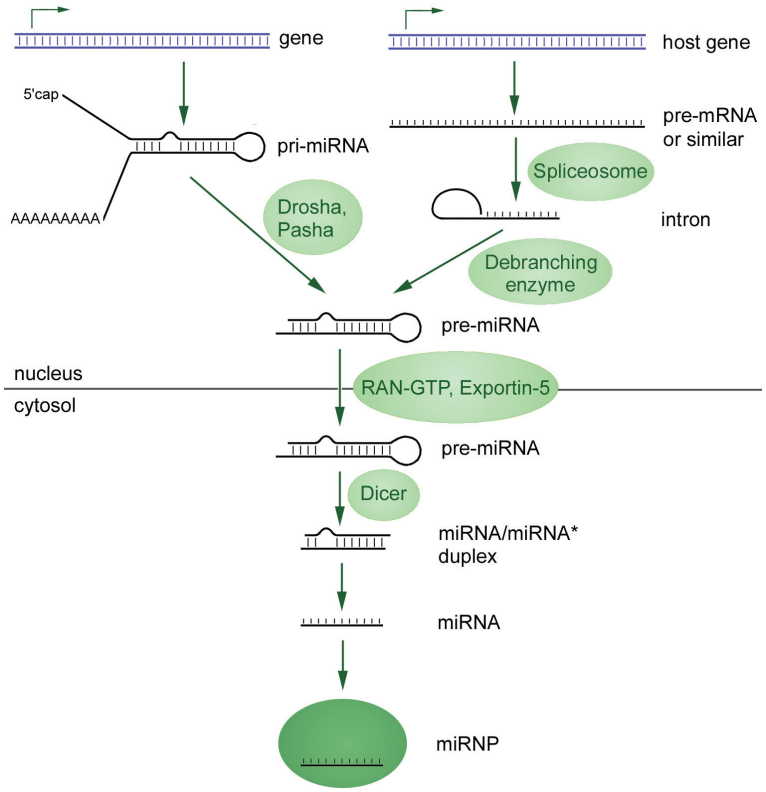


Figure 2.16: Overview of microRNA processing in animals, from transcription to the formation of the effector complex. There are two pathways, one for microRNAs from independent genes and one for intronic microRNAs. Image courtesy of [97].

human genome encodes *eight* kinds of Ago proteins.

2.2.3 RNA Editing

RNA editing is a type of post-transcriptional modification, taking place in eukaryotes, which alters the sequence of primary RNA transcripts by deleting, inserting, or modifying residues. Despite the discovery of several distinct types of RNA editing over the years, **adenosine-to-inosine** (*A-to-I*) RNA editing is now considered the most predominant in mammalians [47]. Through the deamination process, adenosine (A) is converted into inosine (I), which in turn is interpreted as guanosine (G) by both splicing and translation machineries [43]. Enzymes members of the **adenosine deaminase acting on RNA** (*ADAR*) family catalyze this biological

phenomenon only on *double-stranded RNA* structures [44, 46, 47]. The activity of RNA editing is higher in mammalian brain than in other tissues [98], hinting that editing may play a crucial role in the central nervous system [45]. Therefore, malfunctions of RNA editing machineries could lead to serious consequences [48, 49].

Adenosine-to-inosine RNA sites abundantly occur in *intronic regions* as well as in 3′ *UTR*s. RNA editing events can modify RNA molecules in several cellular contexts causing: the creation and/or destruction of splicing sites [43]; the modulation of gene expression pathways [99] during translation [47]; the gain or loss of miRNA binding sites during mRNA targeting [45, 100]. As it has been reported in the last few years, RNA editing sites can be found in non-coding RNA molecules, especially within pri-miRNA [101, 102], lncRNA [103], and precursor-tRNA [104], the latter deaminated by adenosine deaminases acting on tRNA (ADAT) enzymes.

It is possible to distinguish two forms of A-to-I RNA editing, promiscuous and specific. **Promiscuous A-to-I editing** occurs within longer duplexes of hundreds of nucleotides, as in the case of stem–loops that are formed by the pairing of repetitive elements (e.g., Alu elements). In those cases, up to 60% of adenosines could be edited [99, 105]. **Specific A-to-I RNA editing** occurs in short and/or unstable duplex RNA regions [106], in which at least 10% of their adenosines selectively could undergo deamination. A-to-I RNA editing events in small non-coding RNAs, such as microRNAs, are perfect examples of specific editing [47].

One of the main challenges in the study of the RNA editing phenomenon is certainly RNA editing occurrence. The detection of editing sites in RNA molecules in particular cellular conditions is very difficult considering that RNA editing is a dynamic spatial–temporal process. Despite the enormous efforts made in recent years, the real biological function underlying such a phenomenon, as well as ADAR's substrate features still remain unknown.

2.2.4 Epigenetics

Epigenetics studies the factors that modify gene expression in response to environmental and internal phenomena without causing changes to the DNA sequence (Figure 2.17) [107]. Classic example of epigenetic modifications is **DNA methylation**, which changes how genes are expressed, without altering the DNA sequence. Such changes can last for the entire life of the cell, or be spread to other generations without involving any change in the organism's genome.

DNA methylation is a process by which a *methyl group* is added to DNA. The methylation, by modifying how the DNA works, typically suppress the transcription of a gene. Two of the four nucleotides can be methylated (C and A). DNA methylation can permanently alter the expression of genes in cells, while the cells divide and differentiate from embryonic stem cells in specific tissues. The resulting change is normally permanent and unidirectional. However, DNA methylation can be removed either passively, by dilution as a result of cell division, or with an active process, by hydroxylation of the methyl groups that need to be removed [108, 109]. In mammals, DNA methylation is typically performed by two classes of enzyme activities: maintenance methylation and *de novo* methylation. Aberrations in the methylation process are associated with various pathologies. In cancer for example, very often a hyper-methylation is detectable for tumor suppressor genes, and a hypo-methylation for oncogenes [110].

2.2.5 Pathways

A **pathway** is a representation in the form of a graph of actions among molecules in a cell that lead to a certain product, or change in phenotype. A pathway can describe the processes that lead to the assembly of new molecules, such as fat or protein. They may also describe how, in response to external events, the cell changes the activity of its genes. Pathways play a major role in advanced genomic studies. Crucial elements in a pathway are its **endpoints**. They correspond to those molecules that directly affect the phenotype, depending on the current knowledge of the phenomenon.

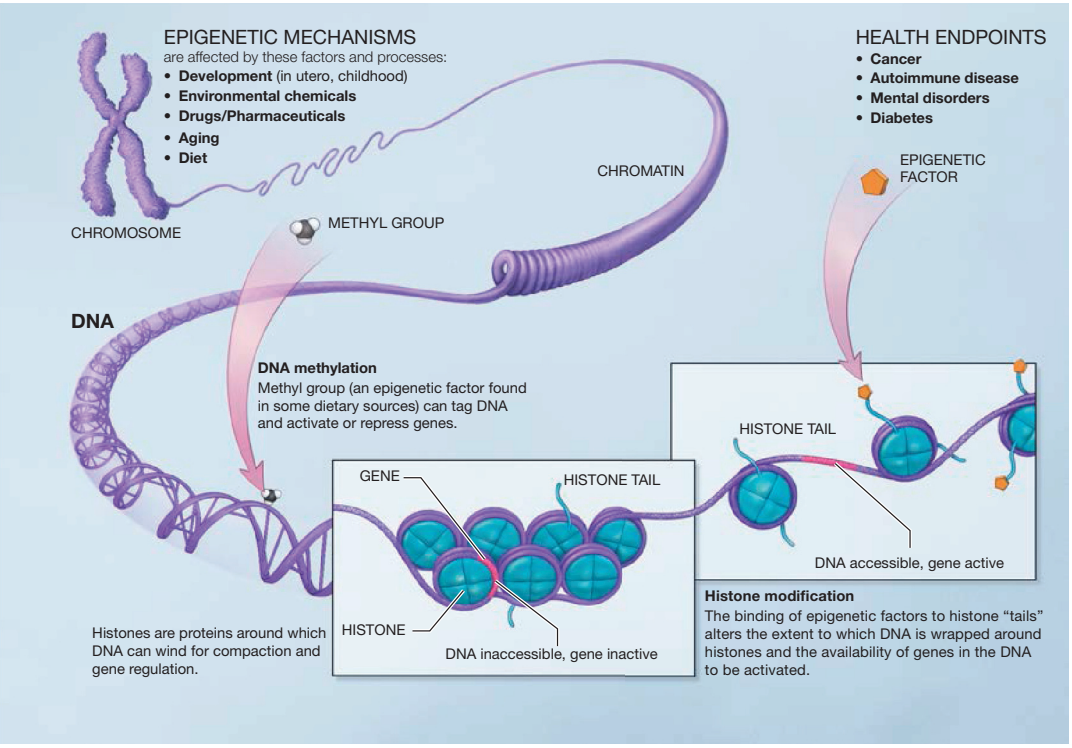


Figure 2.17: Epigenetic mechanisms are affected by several factors and processes including development in utero and in childhood, environmental chemicals, drugs and pharmaceuticals, aging, and diet. DNA methylation is what occurs when methyl groups, an epigenetic factor found in some dietary sources, can tag DNA and activate or repress genes. Histones are proteins around which DNA can wind for compaction and gene regulation. Histone modification occurs when the binding of epigenetic factors to histone "tails" alters the extent to which DNA is wrapped around histones and the availability of genes in the DNA to be activated. All of these factors and processes can have an effect on people's health and influence their health possibly resulting in cancer, autoimmune disease, mental disorders, or diabetes among other illnesses. Image courtesy of [111].

The most common types of pathway are:

- **Metabolic Pathway**: illustrate the chains of chemical reactions, possibly catalyzed by enzymes, that lead to the synthesis of molecules;
- **Genetic Pathway**: a collection of genes that interact with each other and with other substances in the cell in order to govern gene expression and protein activity;
- **Signal transduction pathway**: indicate the chain of reactions that occur within the cell when an external signal or molecule interacts with it.

2.2.6 DNA Sequencing: From Sanger to NGS

DNA sequencing is the process by which the sequence of nucleotides in a *DNA* molecule is determined. The knowledge of the *DNA* sequence is essential for every basic biological research and in medical diagnosis and preparation of highly precise therapies. Nowadays, sequencing is obtained with highly parallelized and performing machines that are able in a few days to sequence an entire human genome.

The foundations for *DNA* sequencing have been laid by the work of **Frederick Sanger** in 1955. He completed the sequence of all the *amino acids* in insulin, providing the first experimental evidence that biological entities were composed by specific molecular patterns, rather than a mixture random elements. This discovery allowed *Crick* in *October* 1954 to speculate that it was precisely the arrangement of nucleotides in *DNA* to determine a protein *amino acid* sequence.

The first method to determine *DNA* sequence was designed in 1970 at *Cornell University* [112–114]. The foundations laid by this work allowed the definition in 1977 of the methodology that today is still considered the gold standard for sequencing: the **Sanger method** [115].

In the *Sanger method* defective nucleotides (*ddNTPs*) linked to a fluorescent are exploited to determine, which nucleotide is present at a certain position of a DNA molecule (Figure 2.18). The sample to be sequenced is placed in a solution containing: regular nucleotides (*dNTPs*),

ddNTPs, and a *DNA polymerase*. The ratio of *dNTPs* and *ddNTPs* is typically of 100 to 1. At one end of the DNA sequence a *primer* is placed, so that once activated, the DNA polymerase can begin to add complementary nucleotides to the sequence reconstructing the opposite strand. In such a process, when the DNA polymerase incorporates a *ddNTP* the reaction will cease. In the solution thousands of copies of the same DNA filament are introduced. When for each of them the reaction has stopped, through a process called *capillary electrophoresis*, each one is separated according to its length, and by means of a laser light the nucleotide in each position can be then estimated. The Sanger method is highly accurate and can read DNA segments up to $1Kb$ (1 Kilobase = 1000 bases). Despite this, the phases of preparation and processing require a lot of time and its costs are high. For example, the human genome project took about 13 years to complete and a funding of abut $3 billion.

The high cost and low efficiency of Sanger methodology, coupled with the necessity of having the sequences of many genomes because of the new direction taken by biomedical and biological research, have prompted the development of new methods for high throughput sequencing. Technologies that parallelize the sequencing process, producing thousands or millions of sequences simultaneously (reads) were developed. The strategy applied in Sanger method was no longer sufficient to meet the current demands.

The first technology of **next generation sequencing** (*NGS*) was developed in the $1990s$ at *Lynx Therapeutics*. Their method *MPSS* (*massively parallel signature sequencing*) used a complex approach to sequence four nucleotides at a time. Their extremely complex methodology led in 2004 to the development of easier and less expensive technologies. Either way the properties of the output of *MPSS* have become typical of all NGS technologies, such as the hundreds of thousands of small reads. This common feature led to the development of a series of new computational tools which allow the analysis of such massive data. It is no longer possible to exclude computer analysis from biological experiments, because of the large quantity of generated data.

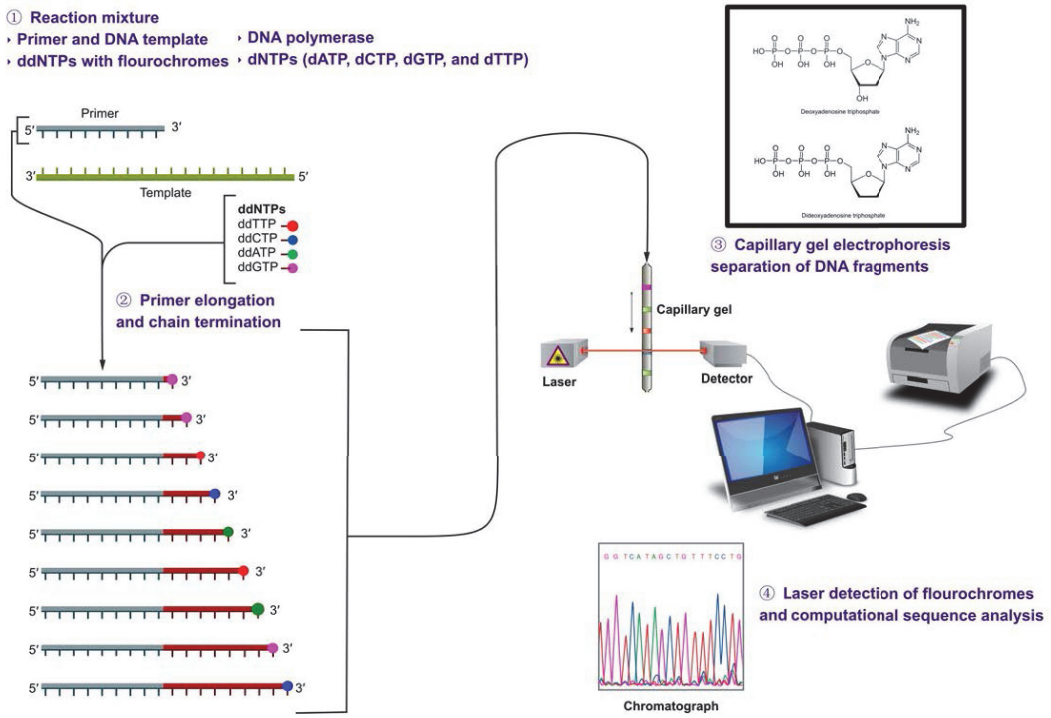


Figure 2.18: The Sanger (chain-termination) method for DNA sequencing. (1) A primer is annealed to a sequence, (2) Reagents are added to the primer and template, including: *DNA polymerase*, *dNTPs*, and a small amount of all four **dideoxynucleotides** (*ddNTPs*) labeled with fluorophores. During primer elongation, the random insertion of a *ddNTP* instead of a *dNTP* terminates synthesis of the chain because DNA polymerase cannot react with the missing hydroxyl. This produces all possible lengths of chains. (3) The products are separated on a single lane capillary gel, where the resulting bands are read by a imaging system. (4) This produces several hundred thousand nucleotides a day, data which require storage and subsequent computational analysis. Image courtesy of [116].

Nowadays there are many different strategies for NGS using various techniques (see figure 2.19). However, the common goal is to achieve sequencing of entire genomes at less than $\$1,000$.

Extracting RNA expressions

Quantifying RNA is a key tool in the analysis of biological and pathological processes. Determining the expression of an RNA involves understanding its activity, and the differences that emerge in the presence of pathological phenomena. Such information can be used not only to study the onset of diseases, but also for establishing new forms of objective laboratory tests, or to create more precise therapies with fewer side effects.

Among the techniques available to date to sample gene expression there are **microarrays** (Figure 2.20). A microarray is a collection of small DNA strands (or **probes**) attached to a solid surface, or **chip**. Scientists used such chips to be able to measure the level of expression of various RNA simultaneously. Each probe has a specific sequence of DNA of about 10-12 moles.

The principle behind microarray is hybridization between two strands of *DNA*, or the property of complementary nucleic acids to pair naturally forming hydrogen bonds between pairs of complementary bases. A greater number of complementary base pairs implies a greater strength in the bond. After washing the surface of a microarray, only the strongest bonds will remain intact. Through the use of dyes the spots where sequences are hybridized can be identified, and by comparing the color intensity between two different conditions an estimate of the expression can be determined. In order to be able to use microarrays with *RNA* molecules, they must first be converted in **coding DNA** (*cDNA*) by a process of reverse transcriptase.

Despite microarrays are widely used and relatively cheap, they present some substantial defects. The processes of synthesis, purification, and storage of the solutions necessary for manufacturing microarrays are extremely complex and expensive. In addition, in presence of very similar RNA families, the technique is quite imprecise due to the fact that molecules can hybridize in

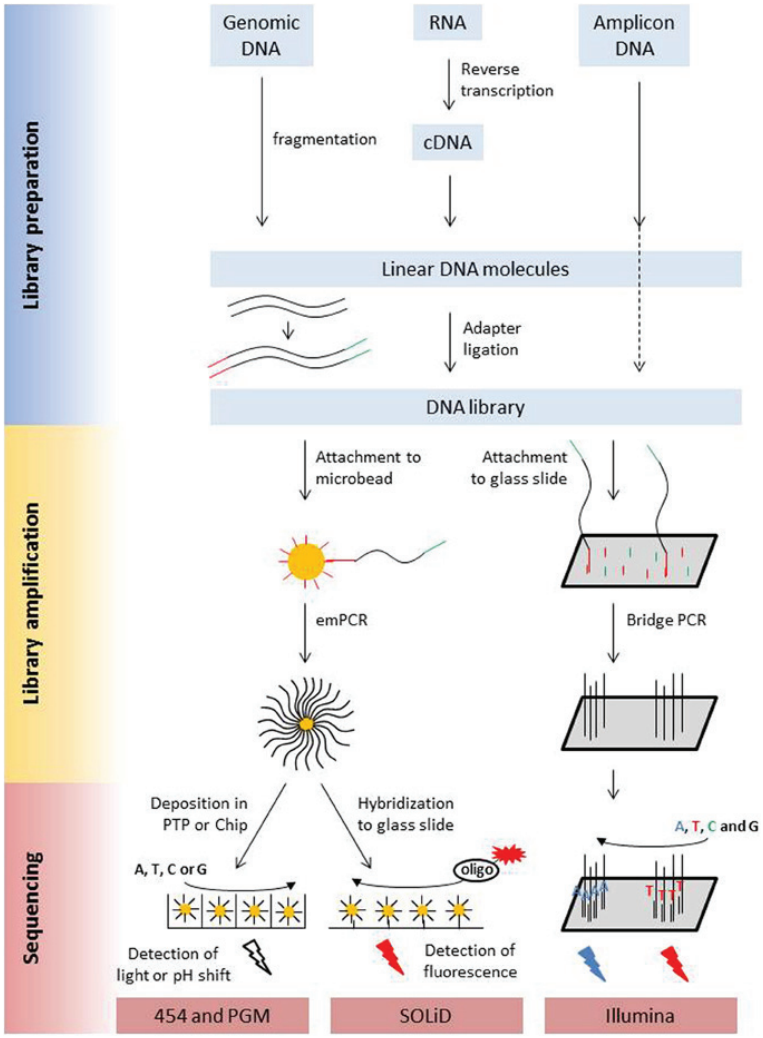


Figure 2.19: Schematic presentation of the library preparation and sequencing process of the most commonly used next generation sequencing platforms. All different types of starting molecules are converted into double stranded DNA molecules that are flanked by adapters. Adapters are sequencing platform specific and enable the binding of the library molecules to surfaces, either beads or a flow cell, where they are amplified prior to sequencing. Clonal amplicons are spatially separated on the glass slides, chips, or picotiterplate. Sequencing is either a sequencing by ligation process with fluorescently labeled oligonucleotides of known sequence (SOLiD) or a sequencing by synthesis process. During Illumina sequencing, four differently labeled nucleotides are flushed over the flow cell in multiple cycles, depending on the desired read length. During 454 and Ion PGM sequencing unlabeled nucleotides are flushed in a sequential order over the flow cell. Incorporation is detected via a coupled light reaction (454) or the detection of proton release during nucleotide incorporation. Image from Knief [117].

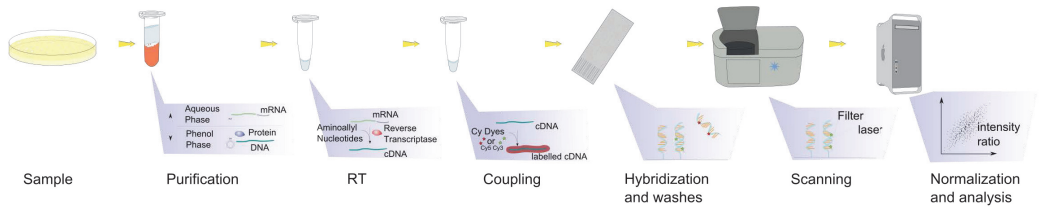


Figure 2.20: The steps required in a microarray experiment. The sample is prepared, purified by extracting *RNA*, which is subsequently converted into *DNA* by a process of reverse transcriptase. Dyes are added to the DNA molecules thus obtained and, therefore, placed in a microarray chip where the hybridization process will naturally occur with the probes in its surface. The chip is then placed into the scanning apparatus, which, via a laser light, will scan the surface, and after a normalization procedure, expression values will be available.

spots designed for other *RNAs* of the same family. This is known as **cross-hybridization** problem. Scanning the microarray to determine the intensity of colors can also introduce biases in presence of overlapping spots or poorly expressed *RNAs*, where the intensity may be insufficient, causing a failure in expression detecting. All these issues and the reduction of costs of NGS technologies led to the emergence of such techniques also for the detection of gene expression.

NGS applied for the analysis of *RNAs*, also called **RNA-Seq** (Figure 2.21), present a considerable number of advantages compared to microarrays. First they are able to capture virtually any transcript, also currently unknown. Any poorly expressed transcripts can also be recognized and properly quantified. RNA families can be separated, and the costs of the technology are continuously decreasing. RNA-Seq has also uncovered post-transcriptional alterations of RNA before unknown. Nevertheless, in order to properly analyze the results obtained by RNA-Seq, new tools for the quantification of expression values from NGS reads needed to be developed. These instruments, even though very effective, do not have the same precision than older technologies. This implies the need to validate results with more accurate techniques once post-NGS analysis have been carried out.

2.2.7 Drugs

A drug is a chemical compound used to diagnose, treat, cure, or prevent diseases. Pharmacotherapy is one of the most important processes in the medical field. Continuous developments in

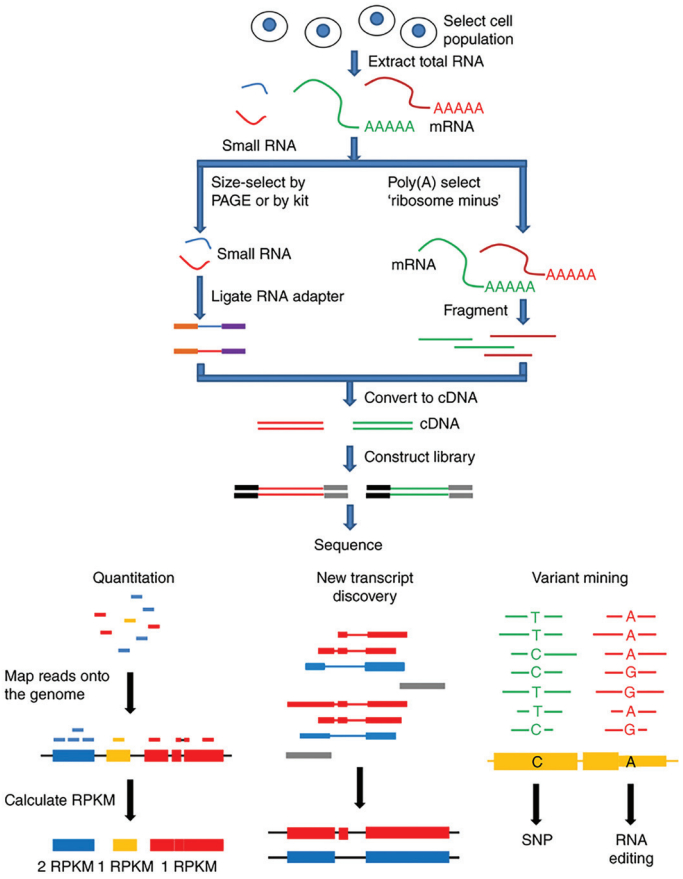


Figure 2.21: Schematic diagram of RNA-seq analysis. Total RNA is extracted from $300,000$ cells to 3 million cells. mRNA is fragmented into a uniform size distribution. The cDNA is then built into a library and sequenced. Mapping programs align reads to the reference genome. Gene expression can be quantified as absolute read counts or normalized values. If RNA-seq data sets are deep enough and the reads are long enough to map splice junctions, the mapped reads can be assembled into transcripts. The sequences of the reads can be mined by comparing the transcriptome reads with the reference genome to identify nucleotide variants that are either genomic variants or candidates for RNA editing. **RPKM** (*Reads Per Kilobase per Million mapped reads*) is a method of quantifying gene expression from RNA sequencing data by normalizing for total read length and the number of sequencing reads. Image from Zeng and Mortazavi [118].

the science of drugs have allowed the creation of molecules ever more precise and with fewer side effects. However, the process of discovering and developing new drugs is very complex, expensive, and requires extremely long time.

Traditional medicines are typically small molecules usually derived from chemical synthesis. New types of drug therapies include recombinant proteins, vaccines, gene therapies and cell therapies (eg stem cells).

Drugs act by interacting with cellular products, by modifying the activity of genes, by altering the reactivity of DNA, or by modifying the activity of enzymes. However not all interactions of a drug are known. This implies the occurrence of possible side effects, or the ability to use the molecule for purposes originally unexpected.

2.3 Fundamentals on the analysis of biological data

This section is a brief introduction to the analysis of biological data.

2.3.1 Fundamentals on enrichment analysis

The enrichment analysis is a technique used in order to assign a meaning to a group of some biological elements, typically genes. In the past, each gene was studied individually, and a function, or role in a biological process, were associated after extensive experiments. Such a knowledge, gained through the years, has allowed the construction of large databases such as the **Gene Ontology**[119], cataloging our experience in a very precise and standard format. This gave use the opportunity to develop a number of novel techniques that automate the discovery process of biological function and role, using statistical techniques to determine unexpected attributes (Figure 2.22).

A typical enrichment procedure starts from the results arising from some experiment. The elements found in such assessment are first grouped together in some way, for example by using prior knowledge, then a statistical test is applied to determine which group that shows a signifi-

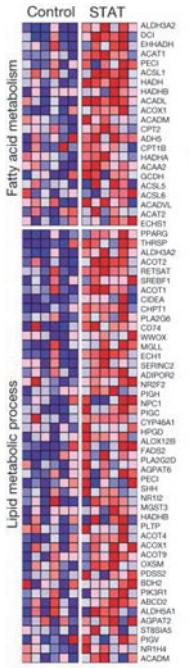


Figure 2.22: Example of enrichment analysis. The results of an experiment were analyzed and grouped into clusters. Two clusters showed a statistically significant association for two different biological processes. Image from Cho et al. [120].

cant over-representation of some biological characteristic, and such feature is a starting point for future investigation.

The Gene Ontology Among the most pressing problems in modern biology is the lack of a universal terminology standard. Many similar terms differ from species to species, and sometimes also from laboratory to laboratory. The Gene Ontology (GO) is one of the most important bioinformatic initiative which objective is to unify and standardize the representation of attributes (terminologies) associated with genes and their products of all sorts[121], in order to simplify the process of communication and sharing data. The project aims to maintain and develop a controlled vocabulary, annotate genes and their products, and provide tools for a simplified access. GO is part of the wider Open Biomedical Ontologies (OBO) project, which aims to unify the terminology in the biomedical and clinical field.

The terms in GO are organized in a hierarchy so that the high level terms are more general, and therefore assigned to more genes. The terms descendants are connected to their relatives by their relationship type, typically **is a** or **part of**. For example, the nucleus is *part of* a cell, while a neuron *is a* cell. These relationships form a directed acyclic graph (DAG), where each term can have one or more parents, and zero or more children (Figure 2.23). Users can then choose the level of specificity that they want to capture by selecting a level on the DAG.

The GO covers three different domains:

- **Cellular components**: parts of a cell or its external environment;
- **Molecular functions**: the primary activity of a gene product at the molecular level, such as binding or catalysis
- **Biological processes**: operations or sets of molecular events with a defined beginning and end, relevant to the functioning of cells, tissues, organs and organisms.

Each term in the ontology has a name that can be a word or a phrase, a unique identifier, a source (as a citation), and a domain where it belongs. There can also be synonymous terms. Finally,

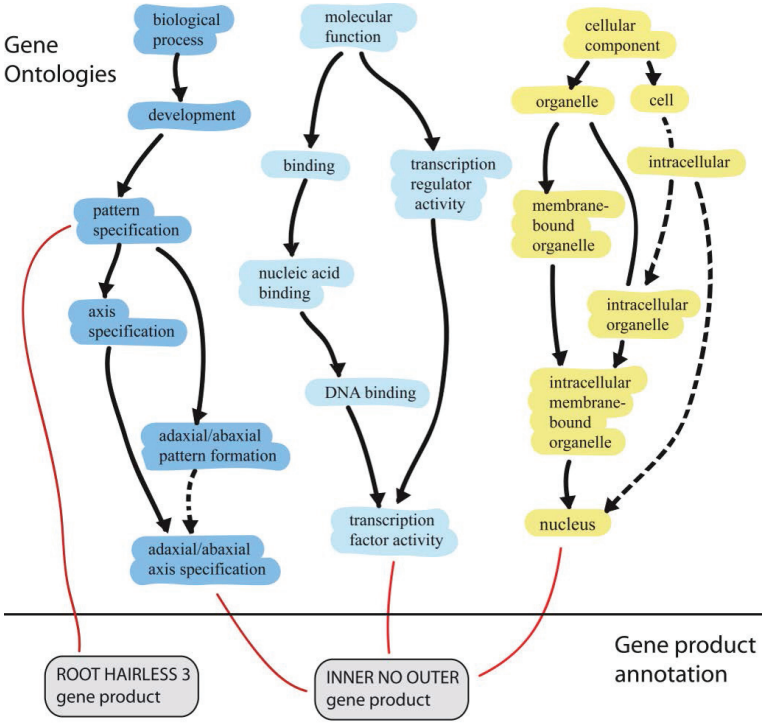


Figure 2.23: Example of Gene Ontology DAG. Image from Clark et al. [122].

the vocabulary is designed to be species neutral by including terms applicable to each type of organism.

Tool for Analysis of GO enrichments (TANGO) The TANGO[123] algorithm is one of the simplest forms of functional enrichment. The algorithm tests for the significance of a subset of genes through the use of a hypergeometric distribution. Suppose you have a set of n genes of which m are annotated with a certain function (we call this the set A). Of the whole set of genes, a subset T (cardinality m') shows a property we are interested in. T and A overlaps for k genes. The probability that the set A is not over-represented given T is

$$p(|A \cap T| \geq k) = \sum_{j \geq k} \frac{\binom{m}{k}\binom{n-k}{m'-k}}{\binom{n}{m'}}. \tag{2.22}$$

Since the test is performed on many groups, a correction is necessary in order to limit false positives. TANGO, given the set A, defines random sets of the same size of T and calculates the likelihood of over-representation. From these probabilities it estimates, therefore, an empirical distribution that is used to assess whether the cluster A is really significant, or was obtained by chance.

Gene-Set Enrichment Analysis Gene-Set Enrichment Analysis[124] (*GSEA*) is a tool that identifies groups of functionally related genes from laboratory experiments that determine their expression. Reference gene sets can be obtained from various libraries, such as Gene Ontology. The basic idea of *GSEA* is that the differences observed in a gene set will stand out more in the data than individual genes.

GSEA starts from a set D of expression data from two groups of samples, and an initial set of genes S, obtained for example by selecting from the GO all elements associated with a particular biological process. In the first place the genes in D are ranked in accordance with their difference in expression between the two groups. A Enrichment score (ES) is then calculated for S on the basis of their position, and their importance in D. Finally, a statistical significance is computed by rearranging the samples in D, and repeating the same test several times in order to estimate a distribution used to determine the probability that S is not significant. The greater the observed probability the less significant will be the result, and therefore they will be discarded. An example of *GSEA* is illustrated in figure 2.24.

Let r_j be the rank of a gene j in D according to its difference in expression, the enrichment score can be computed as:

$$ES\left(S\right) = \max_i \left| \sum_{g_j \in S, j \leq i} \frac{|r_j|}{N_r} - \sum_{g_j \notin S, j \leq i} \frac{1}{N - |S|} \right|, \tag{2.23}$$

where i is the i-th gene in S, accounting for the position, $N_r = \sum_{g_j \in S} |r_j|$, and N is the number of genes in D. The above formula also takes into account the genes that are excluded

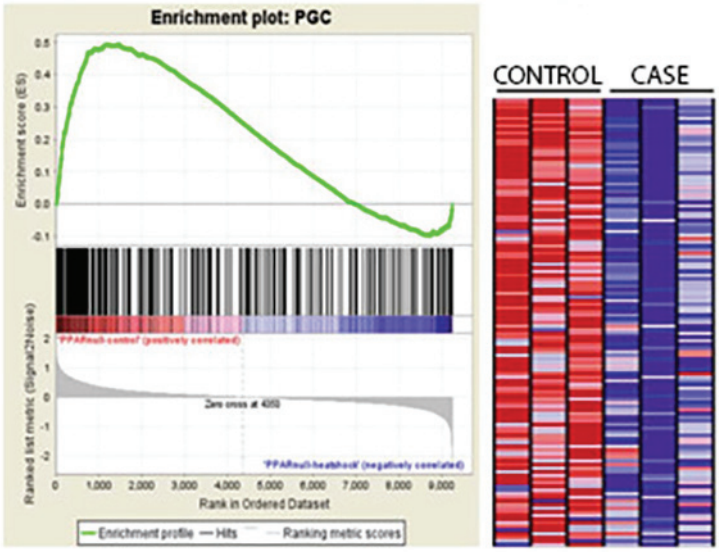


Figure 2.24: Example of *GSEA*. Starting from the left the expression profiles of genes selected by the analysis, then the screen of the *GSEA* application, with enrichment score in the top right corner and the list of all the genes sorted by rank in the bottom right one. Image from Vallanat et al. [125].

from S.

2.3.2 Next-Generation Sequencing Data Analysis

The new generation of sequencing technologies have revolutionized the analysis of genomes [126]. Compared to the classical methods, NGS platforms provide much more data at a lower cost. This represents a challenge both for their storage and analysis. Consequently, the usage of efficient algorithms and powerful computational facilities are often involved.

All current NGS technologies provide as output large text files (tens of millions of lines) containing fragments of all nucleotidic sequences in a sample. To each sequence, a quality value is assigned. The format most commonly used is **FASTQ**. Designed originally to store the results obtained by Sanger Sequencing, FASTQ is the *de facto* standard for all NGS equipment [127]. Each file is divided into blocks of four lines: the first identifies the sequence, the second contains the nuclotidic sequence, the third is a separator, and finally the fourth encodes quality of each nucleotide in the sequence through an ASCII character (greater ASCII code of a character implies

higher quality of the corresponding nucleotide).

The analysis of NGS reads in order to extract some feature of interest (i.e. expressions, variants) usually requires a preprocessing step. In such a phase, reads are first filtered (**quality filtering**), then mapped to a reference genome.

Quality Filtering During quality filtering, low quality portions of each read are removed. This is critical to increase the reliability of subsequent analysis, gaining both in terms of execution time and computational resources. A typical instrument used for this purpose is ERNE-FILTER [128]. Given a threshold Q, the algorithm works in two steps. In the first step, for each reads it calculates the first position where the quality is greater than Q. In the second phase, starting from such position, it selects the subsequence which maintains a quality higher than Q. Finally, if the result has length smaller than *min-size*, or average quality lower than *min-mean-phred-quality*, then the read is discarded. Other tools are available for quality filtering such as *AdapterRemoval* [129], *Cutadapt* [130], *Fastx-Toolkit* (`http://hannonlab.cshl.edu/fastx_toolkit`), *Sickle* [131], or *Trimmomatic* [132].

Bowtie 2 After the quality filtering process, starting from all reads, it is necessary to rebuild the genome, or transcriptome, of the sample in order to continue with other analysis. To do so, a reference genome can be used and each read could be aligned onto it, finding its most likely location. This problem is not exhaustively tractable. This implies the necessity of having clever strategies to address such issue. A common technique is **Bowtie 2** [133]. The algorithm has essentially four steps:

1. **Seed** substrings, which are short segments that are likely to have unique matches in the genome, are extracted from each read;
2. Seeds are aligned to the reference genome using and index to speed-up the process;
3. Seed placements in the genome are prioritized to find the most likely map location(s);

4. Seeds are extended into full alignments with a hardware-accelerated dynamic programming algorithm.

The results of such an algorithm are stored in **SAM** files. SAM stands for Sequence Alignment/Map format. It is a TAB-delimited text file consisting of a header section, which is optional, and an alignment section. In the header section there are general information about the alignment, such as reads sorting and grouping, and other information specific of the alignment algorithm. In the alignment section, each line represents an aligned segment. It contains an unique identifier, the position in the reference genome, a value proportional to the probability that the alignment is wrong, and other information related to its quality. Other similar tools are *BWA* [134], *MOM* [135], *SeqMap* [136] or *SOAP* [137].

Tophat RNA-seq experiments must be analyzed with robust, efficient and statistically principled algorithms. By properly analyzing the data obtained from an RNA-Seq experiment two major objectives can be achieved:

1. identification of novel transcripts from the locations of regions covered in the mapping to the reference genome;

2. estimation of the abundance of each transcript from their depth of coverage.

TopHat [138] aligns reads to the genome and discovers transcript splice sites without a reference annotation. First, by mapping RNA-Seq reads to the genome employing *Bowtie 2*, it identifies potential exons, since many RNA-Seq reads will contiguously align to the genome. Using this initial mapping information, TopHat builds a database of possible splice junctions, and then maps the reads against these junctions to confirm them. In the past people have built splice junctions based on known references, such as RefSeq. TopHat allows a user to find potentially new splice variants.

TopHat is also able to identify the type of transcript by using a set of annotations. Therefore trascripts can be mapped in different functional groups such as exons, introns, or rRNA. On

the basis of this classification an expression level for each mapped transcript can be estimated by dividing the number of reads by their length and total number of mapped reads. Such a value is called RPKM (**reads per kilobase of transcript per million mapped reads**).

Variant Calling Alignment to a reference genome allows the reconstruction of the entire genome, or transcriptome, of a sample. This enables simultaneously a variety of analysis, such as the detection of variants.

Detecting variants in the genome requires different tools according to the type of sequencing and the analysis strategy. To track **single nucleotide variations** (*SNVs*), those bases that vary with respect to a reference genome must be found. Given such bases, with high probability many of them will be due to sequencing errors. More reads, aligned in the same DNA region, present the same variation and more statistically significant it will be.

Finding **structural variations** (*SVs*) instead require a more complex procedure. First, a sequencing from both ends of the same DNA fragment (*paired-end sequencing*) should be performed. Having thus mapped each read to the reference genome, it is necessary to estimate the distribution of fragment lengths, and look for pairs that are mapped to different chromosomes or have an abnormal distance, or ordering or orientation (see figure 2.25). More discordant reads pairs can be explained by the same variant, more significant and precise will be the identification of its breaking point.

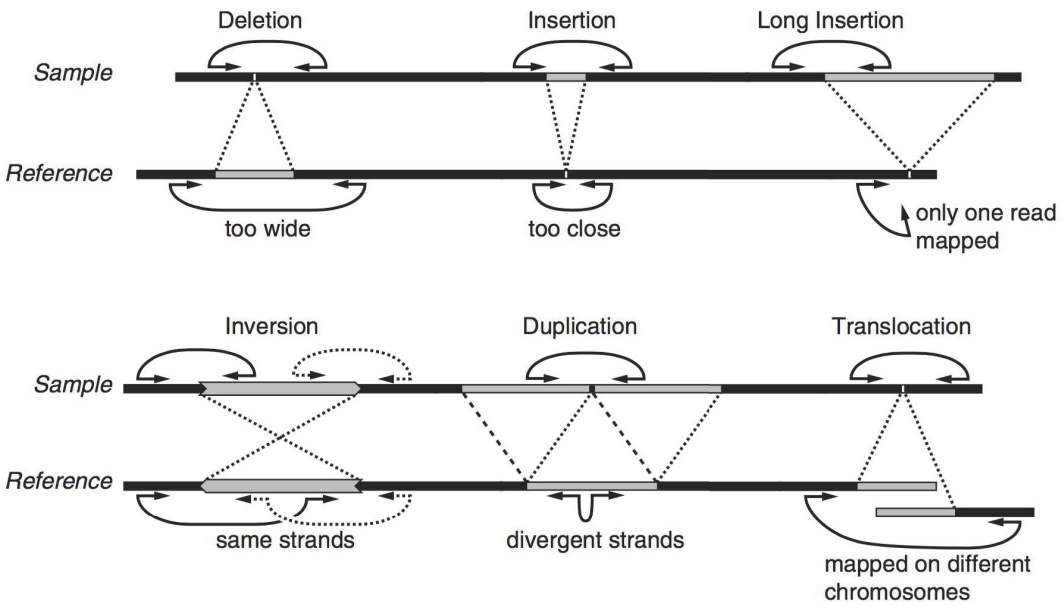


Figure 2.25: Different variant types detected by paired-end sequencing [139]. (1) Deletion: The reference contains a sequence that is not present in the sample. (2–3) Insertion and Long Insertion: The sample contains a sequence that does not exist in the reference. (4) Inversion: A part of the sample is reverse compared to the reference. (5) Duplication: A part of the reference occurs twice in the sample (tandem repeat). (6) Translocation: The sample is a combination of sequences coming from different chromosomes in the reference. Image courtesy of Gogol-Döring and Chen [140].

3

Related Works

The *in silico* development of personalized therapies is a complex process that requires the use of algorithmic techniques of different nature. A key instrument are biological pathway analysis algorithms. They may be used to classify phenotypes and on the basis of this information highly precise therapies can be determined for a single patient.

Nevertheless, the functioning of many biological elements is still unknown, and the use of laboratory experiments is impractical because of the high cost and time requirements. For this reason, techniques that analyze current biological knowledge in order to extract only the most promising hypotheses for experimentation is essential. In this sense, recommendation algorithms play a key role for their simplicity and flexibility.

3.1 Recommendation Systems to mine biological data

A recommender system is an algorithm for information filtering which aims to predict a user's preferences for some objects. Although developed mainly in e-commerce, they are currently being applied in other fields such as biology. In particular, two problems are better suited to be tackled with the tools offered by recommendation systems: the prediction of drug-target interactions (*DTI*s) and the prediction of associations between *ncRNAs* and diseases.

This section will briefly outline algorithmic methodologies currently employed for the aforementioned problems.

3.1.1 Drug-Target Interaction Prediction

Historically, some proteins have been chosen as druggable [18] and it has been shown that drugs with very different chemical structures target the same proteins and the same protein is affected by different drugs. This provides evidence that drugs are not specifically designed to diseases [19]. Recently the trend in pharmaceutical industry changed due to novel bioinformatic prediction methods. New experimental drugs have a wider variety of target proteins. In addition, the analysis of drug-target and gene-disease networks showed that only a few targets are essential, and correlated with tissue specificity to the disease [20].

Following this trend, an attractive drug discovery technique is drug repositioning [141]. The usage of known drugs for novel therapeutic scopes represents a fast and cost-effective strategy for drug discovery. Numerous studies raised a wide variety of models and computational methods to identify new therapeutic purposes for drugs already on the market and sometimes even in disuse. Such computational methods employ an high level knowledge integration in order to discover unknown mechanisms. In González-Díaz et al. [27] a compressive survey on the techniques and models is given. Such models using tools available in chemoinformatics [18, 21, 22], bioinformatics [23–26], network and system biology [18] allow the development of strategies that can speed up drug design. Following González-Díaz et al. [27], repositioning methods can

be grouped into 6 categories: blinded, target-based, knowledge-based, signature-based, pathway- or network-based, and targeted-mechanism-based.

The basic approach to repositioning is known as blinded. **Blind methods** do not include any biological information or pharmaceutical discovery. They commonly rely on serendipity and depend on random tests on specific diseases [28, 29].

Target-based repositioning includes high-throughput experiments on drug and biomarkers of interest in connection with in silico screening for the extraction of compounds from libraries based, for example, on docking [19, 20, 141] or on comparisons of the molecular structures [21, 27]. This approach compared to the blind one is more effective as different targets link directly to the mechanisms of the disease. Therefore, these methods in a short time (i.e. a few days) are used to screen all molecules for which the chemical structure is known. In Iskar et al. [18], authors designed a framework for drug repositioning based on the functional role of novel drug targets. They proceeded by detecting and annotating drug-induced transcriptional modules in cell specific contexts, which allowed also to detect novel mechanisms of action. In silico results were confirmed by an in vitro validation of several predicted genes as modulators of *cholesterol homeostasis*.

Knowledge-based drug repositioning takes into account information concerning drugs, drug-target interaction networks [22–24], drug chemical structure, target structure (including also their similarity), side-effects and affected metabolic pathways [25]. This knowledge enables the development of integrated high-performance predictive models [26]. In Yıldırım et al. [23], a bipartite graph linking US Food and Drug Administration-approved drugs to proteins by drug target binary associations is exploited. Campillos et al. [25] identified new drug target interactions (DTI) using side effect similarity. Iorio et al. [142] make use of transcriptional responses, predicted and validated drug modes of action and drug repositioning. Furthermore, Yamanishi et al. [143] presented a bipartite graph learning method to predict DTI by integrating chemical and genomic data. Cheng et al. [30] present a technique based on network-based inference (NBI)

implementing a naive version of the algorithm proposed by Zhou et al. [144]. van Laarhoven et al. [145] used a machine learning method to predict novel DTIs with high accuracy. Chen et al. [28] introduced a Network-based Random Walk with Restart on the Heterogeneous network (NRWRH) algorithm predicting new interactions between drugs and targets by means of a model dealing with an «heterogeneous» network. Mei et al. [29] proposed the Bipartite Local Model-Interaction-profile Inferring (BLM-NII) algorithm. Interactions between drugs and targets are deduced by training a classifier.

Signature-based methods use expression data to discover off-target related to known molecules for the treatment of other pathologies [146]. Some of these methods also incorporate time-course quantitative data showing that a drug can give the survival outcome in connection to the clinical conditions [147]. This allows to stratify patients. Furthermore, such methods by integrating quantitative information are able to discover additional mechanisms of action not yet known to molecules and known compounds. In Dudley et al. [148] authors predicted therapeutic drug-disease relationship so far not yet described, by combining publicly available disease microarray data of human cell lines treated with drugs or small molecules obtained from Gene Expression Omnibus (GEO) of the National Center for Biotechnology Information (NCBI). With this approach they identified about $16,000$ possible drug-disease pairs, in which $2,664$ are statistically significant and more than half suggests a therapeutic relationship. To validate the hypothesis, authors tested the *cimetidine* as a therapeutic approach for *lung adenocarcinoma* (LA). Cancer cells exposed to *cimetidine* showed a dose-dependent reduction in growth and proliferation (experiments performed on mice implanted with human cell lines of LA). Furthermore, to test the specificity of this proposal, a similar experiment was carried out in mice with cell lines of *ACHN renal cell carcinoma* (the score of such signature was not significant for *cimetidine*), and in agreement with their computational analysis no effect has been observed. In Sirota et al. [149] by integrating publicly available gene expression data, they discovered that anticonvulsant *topiramate* is a hypothetical new therapeutic agent for *inflammatory bowel diseases* (IBD). They

experimentally validated *topiramate*'s efficacy even though the exact pharmacodynamics mechanism of action is not yet known.

Pathway/network based approaches use omic data, signaling pathways and protein-protein interaction networks to build disease-specific pathways containing end-points targets of repositioned drugs [150–152]. These methods have the advantage of identifying signaling mechanisms hidden within the pathway and the signatures of its genes. The above approaches together with large-scale drug sensitivity screening led to predict combinations of drugs for therapeutically aims. In Tang et al. [153] the idea behind their inference model consists of using druggable targets resulting from taking into account drug treatment efficacies and drug-target binding affinities information. They validated the model in breast and pancreatic cancers data by using siRNA-mediated target silencing, highlighting also the drug mechanism of action in cancer cell survival pathways.

More recently, the development of multi-target drugs or drug combinations has been considered crucial to deal with complex diseases [154, 155]. Effective methods to improve the combinations prediction include, choke point analysis [156, 157], a reaction that either uniquely consumes a specific substrate or produces a specific product in a metabolic network, and comparison of metabolic networks of pathogenic and non-pathogenic strains [158]. These approaches commonly share the identification of nodes having a high ratio of incident k-shortest paths [158, 159]. On the other hand, it has been shown that the co-targeting of crucial pathway points [160, 161] is efficient against drug resistances both in anti-infective [162] and anti-cancer [163, 164] strategies. Two relevant examples are RAS [165, 166] and Survivin [167] associated diseases.

In practice, a fundamental question is if the chosen drug is effective to the treated patient. A large amount of money is spent on drugs that have no beneficial effects on patients causing dangerous side effects. It is known that this is due to the genetic variants of individuals that influence metabolism, drug absorption, and pharmacodynamics. Although this, frequently

GWAS for drugs are not replicated in either the same or different populations. Genomic and epigenomic profiling of individuals should be investigated before prescription, and a database of such profiling should be maintained to design new drugs and understand the correct use of the existing ones for the specific individual. Such profiling should exist for each individual and not as in the current era related only to publications which are sample-case specific and results are in some case difficult to replicate [168].

3.1.2 non-coding RNA-disease association prediction

As stated in early chapters, functions of non-coding RNAs (*ncRNAs*) are mostly unknown, which implies that great efforts is been employed in their study, due to their involvement in a wide variety of biological functions. Small ncRNAs, such as siRNA, miRNA and piRNA, are highly conserved in different species and have a key role in transcriptional and post-transcriptional silencing of genes. Long ncRNA instead are poorly preserved and have the task of regulating gene expression through mechanisms still largely unknown [38–40]. It has been shown that these molecules are involved in the regulation of gene expression by acting as controllers of processes such as RNA maturation or transportation, or altering chromatin structure. *ncRNAs* have great variety in structure and in gene regulation outcomes, however, several similarities can be identified in the way they act [33].

The connection between diseases and de-regulation of small *ncRNAs* has been established for years. However, recent studies show that mutations and de-regulations of *lncRNAs* are heavily involved in the onset or progression of several diseases [169]. Alterations in the structure, or in the expression levels, are the main underlying causes of diseases, from cancer to neurodegenerative disorders [169].

Pasmant et al. [35] highlight how the expression of *lncRNA ANRIL*, antisense transcript to *INK4b* gene, is correlated with the epigenetic silencing of *INK4a*, or *p16 protein*, which is involved in the regulation of cell cycle. High levels of *ANRIL* were found in prostate cancer

tissues [34]. Yap et al. [34], also, hypothesizes that this transcript is an initiating factor in tumor formation due to its silencing action on the *INK4b/ARF/INK4a* locus. Other experimental evidence link *ANRIL* de-regulation to a number of pathologies, including coronary disease, intracranial aneurysm, and type II diabetes [35].

Another example of correlation between *lncRNAs* and diseases is the *HOTAIR* transcript, which is involved in the progression of breast cancer by chromatin landscape remodeling [36]. In particular, increased expression of *HOTAIR* is an index of poor prognosis and tumor metastasis. Gupta et al. [37] show that *HOTAIR* is also responsible for invasiveness and metastasis in epithelial cancer cells and its inhibition may lead to a reduction of invasiveness in cells where *PRC2 complex* is highly activated.

Therefore, despite the enormous importance that *ncRNAs* are showing in connection with several diseases, the number of entities, which somehow have been functionally characterized and associated to pathologies, is extremely small [169]. For this purpose, developing a methodology that is able to predict ncRNA-disease interactions is crucial in order to formulate new hypotheses on the molecular mechanisms underlying complex diseases, and to identify potential new biomarkers for their diagnosis, treatment and prevention.

In this direction, Yang et al. [41] developed a method, which exploits a bipartite network and a propagation algorithm to predict new associations that can be evaluated through appropriate in vitro experiments. Yang et al. [41] based their method on the database assembled by Chen et al. [170]: a collection of approximately 1028 experimentally validated interactions among 322 *lncRNAs* and 221 diseases. The database has been further extended, through deep literature mining, to include additional interactions. The database includes also 478 experimentally validated interactions among 126 *lncRNAs* and 236 protein coding genes. For such genes a modulation in expression values is known to be carried out by such *ncRNAs*.

Another network-based approach has been devised in Liu et al. [171]. The authors built from transcriptome microarray data a co-expression network to identify *ncRNA-mRNA* interactions.

The network was then filtered through the use of differentially expressed transcripts, and for each *ncRNA-mRNA* pair a pearson correlation was computed. Only pairs with significant correlation were preserved. Then, by means of functional enrichment, *lncRNAs* that showed correlations with genes enriched in a pathology were returned. The authors were able to identify circulating *lncRNAs* whose aberrated expression is found in major depressive disorder. Their association with *mRNAs* also suggested a contribution to the molecular pathogenesis of the disease.

A similar method (*RWRlncD*) was developed by Sun et al. [172]. The proposed algorithm uses a random walk with restart in order to identify possible associations between functional *lncRNAs* and disease. The performance was verified using a leave-one-out cross-validation, obtaining an area under the **receiver operating characteristic** (*ROC*) curve of 0.822.

An additional method that exploits random walk technique has been developed by Ganegoda et al. [173]. Authors use a gaussian kernel interaction profile for estimating similarity between *lncRNAs*. A similarity between disease is also built using a text mining approach on the Online Mendelian Inheritance in Man (*OMIM*) database. This information is then combined with an input *lncRNA-disease* network, and a random walk is applied for the final predictions. The method was validated using a leave-one-out cross-validation and the results have shown the effectiveness of the approach.

A completely different strategy exploits known interactions between *lncRNAs* and *miRNAs* [174]. First a *miRNA-disease* network is built from the HMDD database [175]. So, a list of *lncRNA-miRNA* associations was obtained by taking the information contained in Starbase v2.0 [176]. Finally a hypergeometric test is implemented for each *lncRNA-disease* pair by examining whether the number of shared common *miRNAs* is statistically significant. Although the method does not rely on any validated *lncRNA-disease* association, the results show an average area under the ROC curve equal to 0.7621.

Prediction of phenotypes, such as diseases, or of responses to therapies from the large amount of high-dimensional data obtained through Next-Generation Sequencing techniques is an extremely important task in translational biology and precision medicine. However, the gap between current analysis techniques and the ability to obtain precise and accurate knowledge is broad.

High-throughput sequencing and gene profiling techniques are radically transforming medical research, allowing the full monitoring of a biological system. The use of these technologies typically generates a list of differentially expressed transcripts (i.e. genes, microRNAs, lncRNAs, etc.) whose behavior varies significantly among the phenotypes under examination.

Furthermore, compared to traditional gene expression extraction techniques, deep sequencing methods, such as RNA-Seq, provide much larger lists of differentially expressed transcripts, increasing, therefore, the complexity of their analysis. A common approach to simplify and make the analysis of such data more fruitful consisted in grouping genes into smaller sub-sets according to some relationship, leveraging on existing knowledge bases such as ontologies or pathways. The analysis of this data at the functional level is crucial since it allows a strong reduction of dimensionality, thus providing greater insights on the biology of the phenomenon under study [177].

An extensive class of techniques known as **Pathway Analysis** [178] goes in this direction. In the past, such term had been associated to the analysis of ontological terms, protein-protein interaction (PPI) networks, or to the inference of gene regulatory networks from expression data. More recently, great interest has shifted toward a class of methods called *Knowledge base-driven pathway analysis* [10]. Such methods leverage on existing databases, such as the Kyoto Encyclopedia of Gene and Genomes (KEGG) [179, 180] or Pathway Commons [181], to identify those pathways that may be affected by the expression changes in the observed phenotype. Following Khatri et al. [10], knowledge base-driven pathway analysis techniques can be grouped into three

generations of approaches: i) over-representation analysis (ORA), ii) functional class scoring (FCS), and iii) pathway topology-based (PT).

Over-representation analysis methods statistically evaluate the number of altered genes in a pathway with respect to the set of all analyzed genes. After filtering the resulting gene set of an expression assessment experiment, ORA strategies [182–188] typically divide the list of genes according to the pathway each gene belongs to. By applying an hypothesis test (i.e. hypergeometric, chi-square, or binomial) they are able to determine if the number of such genes is over- or under-represented. These methods, however, have some major limitations. Firstly, considering only the number of differentially expressed genes, while omitting their expression, implies that the magnitude of their change be unimportant for pathway activity. Furthermore, considering only statistically significant differential expression may exclude those genes whose coordinated alteration may lead to remarkable effects, although their differential expression may not be statistically significant. Finally, they consider individual genes and pathways, respectively, in a manner independent of the surrounding biological context, eluding what truly happens in reality.

Functional class scoring methods compensated some of the disadvantages of ORA approaches. Typically FCS methods compute a gene-level statistics from their expression levels, by means of a statistical approach (i.e. ANOVA, Q-statistic, signal-to-noise ratio, t-test, or Z-score). Such statistic is calculated for all genes in a pathway [189–195] and its statistical significance is estimated through an appropriate null hypothesis [190, 196–198]. FCS methods avoid some of the limitations of the ORA approaches by ranking all genes through their expression level and by considering the dependencies within a pathway. However, they do not take into account the fact that a gene can be simultaneously active in multiple pathways, utilizing the expression magnitude only to sort genes.

In order to overcome the disadvantages of FCS methods, the third class of techniques models a pathway as a graph, considering its topology when computing scores. A thorough analysis of

all PT-based approaches has been provided in [16].

The first PT-based algorithm is *ScorePAGE* [11]. As FCS methods do, it computes a gene-level statistic by applying a similarity between their expression values. Rahnenführer et al. [11] evaluated four different measures: correlation, covariance, cosine distance, and dot product. They stated that an exclusive choice among these measures is impossible, the best approach being an adaptive one, namely, selecting the measure best suited to the data in question. Such similarity is, therefore, combined to build a pathway-level score. For each pair of elements in a pathway, *ScorePAGE* averages their similarity weighting it on the basis of the distance between the nodes in the pathway. Finally, by means of a non-parametric permutation test, the authors are able to establish the significance of each pathway-level score. However, *ScorePAGE* has been designed to analyze only metabolic pathways.

In Draghici et al. [12], an analytical technique called **impact factor** (*IF*) was introduced. The impact factor is a pathway-level score that takes into account biological factors such as the magnitude of change in genes expression, the type of interactions between genes, and their location in the pathway. In Draghici et al. [12], each pathway is modeled as a graph in which nodes represent genes, while edges represent interactions between them. Authors also define a gene-level statistic (called **perturbation factor**, PF) as a linear function of the change in gene expression and the perturbation of its neighborhood. Such a statistic is then combined for each element in a pathway, and a p-value is computed by means of exponential distribution.

The analysis method presented by Draghici et al. [12], has been further improved by the **SPIA** algorithm [13] which attenuates the dominant effect exercised by the change in expression within PFs computation, while reducing the high rate of false positives when the input list of genes is small. *SPIA* uses a bootstrap procedure to evaluate the significance of the observed perturbation in a pathway. All this is combined with a p-value computed in ORA style to make a full assessment of the statistical significance of the perturbation of each pathway.

To reduce the number of false positives, and to obtain a more significant analysis, Vaske et al.

[14] presented the *PARADIGM* algorithm, which has been further improved by Sedgewick et al. [15]. *PARADIGM* is a method to infer patient-specific genetic activity by incorporating information regarding interactions between genes provided in a pathway. The method predicts the degree of alteration in the activity of a pathway by employing a probabilistic inference algorithm. The authors show that their model obtains significantly more reliable results than SPIA. However, Mitrea et al. [16] stated they could not reproduce the results reported in Vaske et al. [14], despite the full cooperation of its authors.

In Martin et al. [199], a method that quantifies the response of a network in an interpretable way was implemented. It exploits the structure of a cause-and-effect network to integrate and mine transcriptomic data, extracting signatures also to predict the phenotype of interest. The main disadvantage of this method is the assumption that the underlying network is strongly connected. Finally, in Vasilyev et al. [200], an algorithm called **sampling of spanning trees** (*SST*) has been devised to analyze networks. The algorithm is based on spanning trees and a sampling procedure that uses a random walk in the graph. *SST* provides a practical way to aggregate the values of the nodes in a network and evaluate its efficacy in discriminating phenotypes.

4

From diagnosis to prognosis

The purpose of this chapter is to illustrate the bioinformatic framework, based on synergistically working instruments, that has been built to help in every step necessary for the development of highly precise therapies, going, where possible, beyond the limits imposed by currently employed techniques. First, a method for a precise classification of samples, which goes beyond biomarkers, will be introduced. Following, two algorithms will be presented as a potential guide to the experimental activity with the aim to fully understand how drugs act, allowing the prediction of more accurate combinations. Next, algorithms for the enrichment of pathways, improving therefore the reliability of previous phases, will be illustrated. Finally, the need to guide the experimental activity reducing associated cost and time requirement led to the development of a technique for simulating the action of endogenous and exogenous elements in a biological system.

4.1 High-precision sample classification: pathway analysis

As already pointed out in previous chapters, the accurate classification of samples is a critical step in precision medicine. Indeed, it is the identification of specific molecular patterns of a patient that allows an accurate definition of therapies. In common practice this process is done through the use of biomarkers. They are measurable indicators typical of some biological state or condition. Genetic biomarkers (specific alterations in gene sequences) or transcriptomic ones (specific alterations in gene expressions) are currently used in common medical practice to identify disease states or possible responses to therapies. However, their main problem is the lack in most cases of a correlation with the origin of the disease state. This makes impossible to use them directly to understand the molecular mechanisms related to the pathology or to define novel therapeutic targets.

Pathway analysis shows all the characteristics needed to identify a novel category of functional biomarkers that not only can be used to understand the molecular processes at the origin of a pathological condition, but also employed for possible new therapies. However, at present, this type of analysis has not been used due of the unreliability of their results.

In what follows it will be shown **MITHrIL** (*miRNA enriched pathway impact analysis*) [17], a technique that extends Draghici et al. [12] and **SPIA** [13], and improves the reliability of the results. The strength of **MITHrIL** lies in the enrichment of pathways with information regarding miRNAs. Our method, starting from expression values of genes and/or miRNAs, returns a list of pathways sorted according to the degree of their de-regulation, together with the corresponding statistical significance (p-values), and a predicted degree of alteration for each endpoint.

Pathway Enrichment **MITHrIL** distinguishes itself from other pathway analysis techniques primarily for the use of *KEGG* [179, 180] pathways enriched with miRNAs and their interactions with genes. In order to achieve this, we downloaded all validated interactions between miRNA and targets from miRTarBase [201, 202] and miRecords [203]. We also obtained interactions

between transcription factors (TFs) and miRNAs from TransmiR [204]. By taking into account TFs activating miRNA genes we are able to increase the knowledge stored within each pathway. We then standardized all identifiers in their respective databases to avoid duplicates. The mapping of miRNA identifiers was performed by using miRBase release 20 [205–209] as reference database. For each target, we performed a twofold mapping procedure: firstly, each gene identifier has been converted to its *Entrez* one; then, by taking advantage of *KEGG REST API*, we mapped each *Entrez Id* to the corresponding *KEGG Id*. This standardized interactions list was, lastly, filtered to remove all duplicates. Such a procedure allowed us to build a knowledge base of $10,537$ experimentally validated interactions between 385 miRNAs and $3,080$ genes. Pathway enrichment was performed by defining a new type of node representing miRNAs in the pathway notation, along with two types of directed edges, for miRNA-target inhibition interactions and TF-miRNA interactions, respectively. The enrichment is thus performed automatically by adding to each pathway only miRNAs that interact with at least one element within it. Finally, in order to acquire information on which endpoints are contained in each pathway, we employed a depth-first search algorithm [210] to automatically mark which genes are located at the end of the chains of reactions in each pathway.

Algorithm **MITHrIL** consists in an extension of Draghici et al. [12] and Tarca et al. [13]. It requires a case/control expression data set from which statistically differentially expressed features have been extracted (genes, miRNAs, or both). For such elements, the computation of their *Log-Fold-Change* is also needed. Starting from such information, **MITHrIL** computes, for each gene in a pathway, a **Perturbation Factor** (PF), which is an estimate of how much its activity is altered considering its expression and 1-neighborhood. Positive (negative) values of PF indicate that the gene is likely activated (inhibited). By appropriately combining each PF of a pathway, the algorithm is, therefore, able to calculate an **Impact Factor** (IF) and an **Accumulator** (Acc). The IF of a pathway is a metric expressing how important are the changes detected in the pathway, the greater the value, the most significant are the changes. The Acc indicates the total

level of perturbation in the pathway and the general tendency of its genes: positive Acc values indicate a majority of activated genes (or inhibited miRNAs), while negative ones corresponds to an abundance of inhibited genes (or activated miRNAs). To the Acc is also assigned a p-value which is an estimate of the probability of getting such accumulator by chance. Finally, by applying the Benjamini and Yekutieli [211] method, we estimate the false discovery rate and p-values are adjusted on multiple hypotheses.

More precisely, let n be a node in pathway P_i. Its perturbation factor, $PF(n, P_i)$ can be defined as:

$$PF(n, P_i) = \Delta E(n) + \sum_{u \in U(n,P_i)} \beta(u, n) \cdot \frac{PF(u, P_i)}{\sum_{d \in D(u,P_i)} |\beta(u, d)|}, \tag{4.1}$$

where $\Delta E(n)$ is the *Log-Fold-Change* computed for the node n, $U(n, P_i)$ and $D(n, P_i)$ are the set of upstream and downstream nodes of n in pathway P_i respectively, and $\beta(u, n)$ is a function that indicates the strength and type of interaction between genes u and n. In particular, negative values of β indicate an inhibitory effect, while positive values an activating one. By exploiting the methodology described in Draghici et al. [12] we compute an impact factor, $IF(P_i)$, which reflects the importance of the changes observed in a pathway, as:

$$IF(P_i) = \log\left(\frac{1}{p(P_i)}\right) + \frac{\sum_{n \in P_i} |PF(n, P_i)|}{\overline{|\Delta E|} \cdot N_{de}(P_i)}, \tag{4.2}$$

where $p(P_i)$ is the probability, calculated using an hyper-geometric distribution, of obtaining a number of differentially expressed nodes at least equal to the observed one in P_i, $\overline{|\Delta E|}$ is the mean *Log-Fold-Change* in P_i; finally, $N_{de}(P_i)$ represents the number of differentially expressed nodes in the pathway.

Our methodology takes also advantage of the accumulation (or accumulator) as described by Tarca et al. [13]. Such a methodology has been revised to take into account the addition of miRNAs. In order to do so, first we need to compute two partial accumulators, $Acc_{mir}(P_i)$ and

$Acc_{gene}(P_i)$, which take into account the perturbation, respectively, of miRNAs and genes:

$$Acc_{mir}(P_i) = \sum_{m \in P_i^m} [PF(m, P_i) - \Delta E(m)], \tag{4.3}$$

$$Acc_{gene}(P_i) = \sum_{g \in P_i^g} [PF(g, P_i) - \Delta E(g)], \tag{4.4}$$

where P_i^m and P_i^g are the sets of miRNAs and genes present in P_i respectively.

In equations 4.3 and 4.4 we sum the perturbations of all miRNAs (P_i^m) and genes (P_i^g) in pathway P_i, addressing the dominant effect of the expression change in the PF computation by subtracting such values. We can now compute total perturbation accumulation, $Acc(P_i)$, which measures whether the pathway is likely activated or inhibited. The introduction of miRNAs in our model addresses the necessity to take into account the fact that an increased (decreased) expression of such elements results in an inhibition (activation) of the pathway. $Acc(P_i)$ is computed as:

$$Acc(P_i) = Acc_{gene}(P_i) - Acc_{mir}(P_i) - E[Acc(P_i)], \tag{4.5}$$

where $E[Acc(P_i)]$ is an estimate of the expected value of the distribution of all accumulators computed for pathway P_i, as explained below.

P-value estimation is then performed by combining the Z-scores, computed through an inverse Standardized Normal distribution, associated to two probabilistic terms: the first is the probability of obtaining by chance a number of differentially expressed genes in the pathway at least equal to the observed one, while the second consists in the probability of observing by chance an accumulator higher than the computed one. The first term corresponds to $p(P_i)$ introduced in equation 4.2. The second term, instead, has to be estimated through a bootstrapping procedure. In such a procedure, we assign to a random group of genes in the pathway a *Log-Fold-Change* selected randomly from the input ones, so as to compute a random accumulator.

The procedure is repeated several times and the final probability is estimated as the ratio between the number of random accumulators greater than $Acc(P_i)$ and the number of repetitions performed. In our experiments, the repetitions were set to $2,000$ in order to obtain maximum precision up to two decimal places.

At this stage we are also able to estimate expected value $E[Acc(P_i)]$ as the median value of the random accumulators.

Therefore, the final result of our algorithm consists of a list of pathways along with their impact factor, accumulator and adjusted p-values. Such list is sorted by p-value and Acc.

Performance assessment and Data Sources To perform a comprehensive test of **MITHrIL**, we exploited expression data provided by The Cancer Genome Atlas (beginning of 2014). We downloaded all patient expression profiles of genes (RNASeqV2 obtained through platforms Illumina Genome Analyzer and Illumina HiSeq) and miRNAs (miRNASeq obtained through platforms Illumina Genome Analyzer and Illumina HiSeq). The initial dataset was then filtered by removing all patients for which one of the two types of expression was unavailable. We then eliminated all tumor samples for which no healthy controls were present. By applying such a procedure, we built a dataset of $3,053$ expression profiles ($2,721$ case samples and 332 control samples) of patients affected by 10 distinct tumor pathologies (see Table 4.1 for more details). Case samples were further divided by disease stage.

To run our algorithm, we performed a differentially expressed genes analysis by using the *RNASeq* pipeline based on **Limma** [212]. The expression matrices for each disease were firstly normalized by using the Voom algorithm [213], then a linear model was trained with **Limma** and differentially expressed genes for each stage of the disease were extracted along with their *Log-Fold-Change*. In our analysis we considered as differentially expressed only those genes for which an adjusted p-value was lower than 0.01 as computed by **Limma**. For each tumor sample we also downloaded and processed *copy number variation* (*CNV*) as shown in Vaske et al. [14].

To compare **MITHrIL** with other methodologies, *PARADIGM* [14] and *SPIA* [13], we used

Table 4.1: List of cancer types extracted from The Cancer Genome Atlas (TCGA) with their codes, number of case and control samples, and Subcategories.

Code	Cancer Type	Controls	Cases	Samples Categories
BLCA	Bladder Urothelial Carcinoma	19	193	Stage I, II, III, IV
BRCA	Breast invasive carcinoma	86	642	Stage I, II, III, IV, X
COAD	Colon adenocarcinoma	8	389	Stage I, II, III, IV
KICH	Kidney Chromophobe	25	66	Stage I, II, III, IV
KIRC	Kidney renal clear cell carcinoma	71	224	Stage I, II, III, IV
LUAD	Lung adenocarcinoma	19	388	Stage I, II, III, IV
LUSC	Lung squamous cell carcinoma	37	247	Stage I, II, III, IV
PRAD	Prostate adenocarcinoma	50	191	Category 6, 7, 8, 9, 10
READ	Rectum adenocarcinoma	3	150	Stage I, II, III, IV
UCEC	Uterine Corpus Endometrial Carcinoma	14	231	Stage I, II, III, IV
	All Samples	**332**	**2721**	

the decoy pathway technique introduced in Vaske et al. [14]. For each pathway, a decoy one has been built by using the same structure and substituting each gene (or miRNA) with one randomly chosen from the set of all possible genes. After the execution of the three algorithms, the pathways were classified by each method and the fraction of real pathways versus the total number of pathways considered was computed. The higher the fraction of real pathways, the better the ability of an algorithm to extract biologically sound results. Lastly, to achieve a fair comparison with *SPIA*, we chose the same β function as Tarca et al. [13]: $\beta(u, n) = 1$ for all interactions that increase node expression level, $\beta(u, n) = -1$ for those that have the effect of decreasing node expression level, $\beta(u, g) = 0$ for irrelevant ones. However, the β function introduces a huge concealed potential in **MITHrIL**, which paves the way for possible future extensions.

Results As stated before, **MITHrIL** has been compared with PARADIGM [14] and SPIA [13] by employing the technique defined in Vaske et al. [14]. The aim is to establish whether the ranking computed with a pathway analysis algorithm is biologically significant. This is achieved by defining random pathways (called *decoy pathways*) with the same topology as the real ones but randomly selected nodes. All pathways are then evaluated by each algorithm, estimating the ability of each method to properly separate decoy pathway from real ones by means of a receiver

Table 4.2: Average areas under the curves (AUC) of 4.1 computed for all cancer dataset. Best results are in bold. In the table **A** represents **MITHrIL** , **B SPIA**, and **C PARADIGM**.

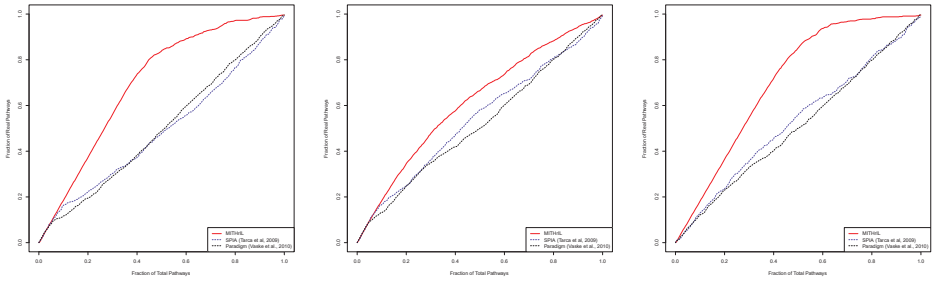
	Dataset									
	BLCA	**BRCA**	**COAD**	**KICH**	**KIRC**	**LUAD**	**LUSC**	**PRAD**	**READ**	**UCEC**
A	**0.6980**	**0.6145**	**0.7047**	**0.6208**	**0.6782**	**0.6821**	**0.5746**	**0.6714**	**0.6852**	**0.6930**
B	0.4884	0.5393	0.5275	0.5259	0.5873	0.4464	0.5273	0.4884	0.4884	0.5340
C	0.4974	0.5162	0.5062	0.5098	0.5287	0.4835	0.5081	0.4983	0.4789	0.5103

operating characteristic (ROC) curve. In principle, a method that can correctly distinguish real pathways from decoys should yield biologically significant results.

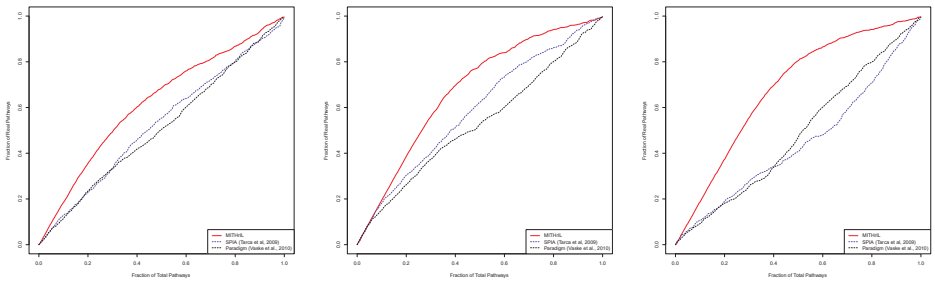
The results of the three methodologies were ranked as follows: **PARADIGM** according to the average number of significant scores, as described in Vaske et al. [14]; **SPIA** according to the adjusted p-value as obtained through their software implementation; **MITHrIL** according to the adjusted p-value and the accumulator. Figure 4.1 shows the results of the comparison which clearly highlight that **MITHrIL** gives the best performances. As further proof of the goodness of the methodology, the average area under each ROC curve (AUC) has been computed. The results were summarized in Table 4.2.

An additional validation of the methodology has been obtained by verifying the percentage of endpoints for which a correct prediction of the deregulation is obtained. Initially, **MITHrIL** has been applied with and without miRNA enrichment to estimate the perturbations for each endpoint of each sample (excluding the expression values of the endpoints in order to avoid introducing a bias in the evaluation). Subsequently, the percentage of endpoints for which the sign of perturbation coincides with that of the log-Fold Change was computed. This validation estimates the reliability of the predictions in terms of biological accuracy and the importance of the addition of miRNA knowledge to our model. The results (Figure 4.2) demonstrate that the incorporation of quantitative information on miRNAs is crucial to obtain valid predictions. This is also supported by tumoral pathologies not influenced by miRNA alterations, in which no improvement is observed when adding such elements.

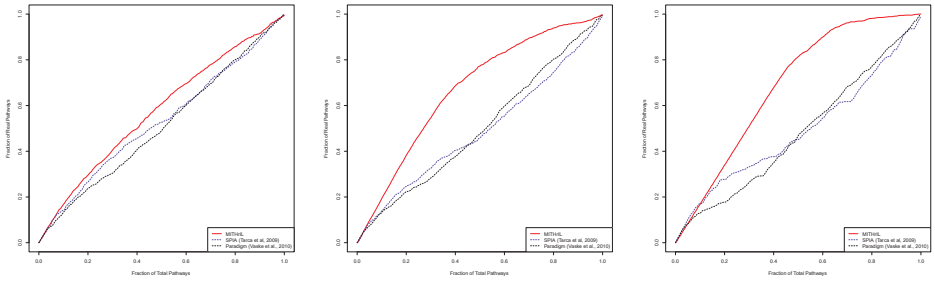
Through the enrichment with miRNA information, **MITHrIL** can greatly improve predic-



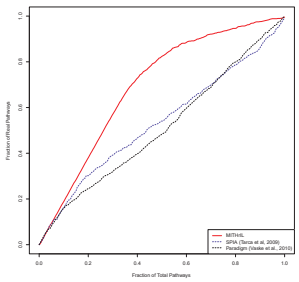
(a) BLCA **(b)** BRCA **(c)** COAD



(d) KICH **(e)** KIRC **(f)** LUAD



(g) LUSC **(h)** PRAD **(i)** READ



(j) UCEC

Figure 4.1: Comparison between **MITHrIL**, *SPIA* [13] and *PARADIGM* [14] by means of decoy pathways. Each line shows the receiver operating characteristic (ROC) curves for distinguishing real from decoy pathways using the pathway ranking.

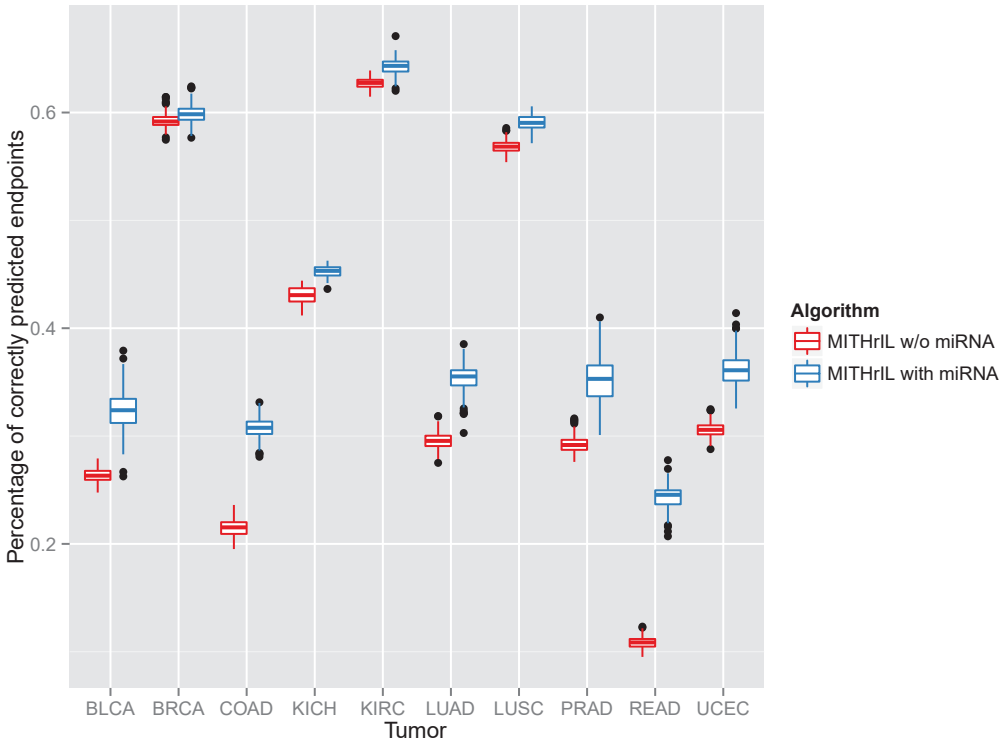


Figure 4.2: Significance of the addition of miRNA in our model by means of a comparison of the percentages of correctly predicted endpoints for each sample between our method with and without miRNAs. Each box in the figure represents the variability range of the percentage of correctly predicted endpoints for the patients of a specific tumor type. A prediction is correct when the deregulation observed in the original data correspond to the one inferred by our algorithm. Namely, the sign of an endpoint *Log-Fold-Change* corresponds to the sign of its *perturbation value*.

tions over *SPIA* and *PARADIGM*. Indeed, while *SPIA* and *PARADIGM* cannot properly distinguish between decoy pathways and real ones, **MITHrIL** is capable of obtaining much better results. Even our worst case had superior results than our two competitors. From a biological standpoint, the ability to distinguish decoy pathways from real ones addresses the fundamental necessity to be able to properly interpret the actual cellular mechanisms as possessing a biological criterion which is crucial to the life of the cell and not the result of random phenomena.

Finally, to evaluate classification performances we elected to train *PAMR* [60] algorithm and evaluate its performance by means of a 10-fold cross validation procedure. *PAMR* is an approach devised to predict cancer class from gene expression profiling, based on an enhancement of the nearest shrunken centroid classifier. The algorithm is able to identify subsets of genes that best characterize each class.

A reference classification was established by applying such a procedure to the *Log-Fold-Change* of differentially expressed genes of our cancer cases. First we computed all differentially expressed genes for each tumor type, obtaining a total of $17,326$ genes that appear to be de-regulated in at least one disease. Next, we calculated their *Log-Fold-Change* in each sample, trained a classifier and verified its performance. The results (Table 4.3) demonstrate that such a classification is quite reliable since it yields a very small error. We then ran the three algorithms on all samples of our set of selected cancer types, and trained three classification models using their scores. As before, we performed a 10-fold cross validation and evaluated errors in each class (Table 4.3). Furthermore, leveraging the ability of our algorithm to return the perturbation for each pathway endpoint, we trained an additional classifier based on such values.

The analysis clearly shows that performances can be considerably improved over reference classification, by taking into account endpoint perturbations (Table 4.3). Table 4.3 also reports the classifications based on *MITHrIL pathway accumulators*. Accumulator summarizes, with a single value, the general perturbation observed within a pathway. Hence, as a further effect this yields a strong dimensionality reduction, although a slight increase in misclassification error can

Table 4.3: Classification results of tumor samples in our dataset obtained training PAMR algorithm by means of *Log-Fold-Change*, *SPIA total accumulation*, *Paradigm scores*, *MITHrIL accumulators*, and *MITHrIL endpoint perturbations*. Each element in the table corresponds to the classification error for a specific cancer type using one algorithm. Despite the reference classification based on *Log-Fold-Change* yields a low average error (2.90%), the employment of perturbations computed for each endpoint provides a significant improvement in the classification accuracy.

		MITHrIL			
Dataset	**Log-Fold-Change**	**Perturbations**	**Accumulators**	**SPIA**	**Paradigm**
BLCA	3.11%	**1.55%**	12.95%	49.74%	82.38%
BRCA	1.86%	**1.09%**	13.08%	8.25%	73.05%
COAD	2.31%	**0.00%**	0.77%	**0.00%**	32.90%
KICH	3.03%	**0.00%**	4.54%	3.03%	31.81%
KIRC	3.12%	**1.79%**	5.80%	2.67%	35.26%
LUAD	4.89%	**1.80%**	4.38%	2.83%	64.43%
LUSC	6.07%	**1.21%**	5.26%	4.04%	71.54%
PRAD	**0.00%**	**0.00%**	2.61%	30.89%	18.94%
READ	3.33%	**0.00%**	0.66%	**0.00%**	96.66%
UCEC	1.73%	**0.00%**	4.32%	1.29%	46.32%
Total	2.90%	**0.90%**	6.40%	8.80%	57.60%

be noticed, compared to reference classification. The last two columns of Table 4.3 report the classification performances obtained by *SPIA accumulators* and *PARADIGM scores*. All of this shows that the addition of miRNA information is crucial in order to obtain more reliable results.

Leveraging on the potential provided by miRNA enrichment in pathway analysis, **MITHrIL** represents a bioinformatic resource capable of a far more precise evaluation of pathway deregulation in cancer. This can provide a decisive contribution to cancer research in terms of directing researchers more effectively, reducing costs and time requirements. Specifically, **MITHrIL** can contribute to an earlier diagnosis, an early and more accurate drug resistance assessment, as well as to more precise prognosis in terms of predicting future disease development.

4.2 Enhancing drugs knowledge: DTI Prediction

In the past, the classical drugs development approach consisted in producing chemical compounds that would act against specific families of proteins associated in some way to a pathology [23]. However, drugs work by binding to protein and modifying their biochemical and biophysical properties, consequently altering their activity. A protein operates as part of a highly

interconnected cellular network (interactome). This implies that a small change can have great effects, even unthinkable. Furthermore it was shown that drugs developed to act on specific proteins can interact with other even if during their development such an eventuality was avoided [214]. This has shown that the paradigm *one gene, one drug, one disease* is too restrictive in most cases [141]. Therefore fully understand the activities that drugs play within the cell is of fundamental importance to develop highly precise therapies, which minimize side effects. However, a comprehensive laboratory approach is unthinkable given the high cost of the experimentation process. For this purpose, drug-target interaction (DTI) prediction techniques are playing a crucial role by reducing the number of laboratory experiments selecting only the most promising candidates.

In this sense, the **DT-Hybrid** (domain-tuned hybrid) algorithm was conceived [31]. It extends *NBI*, an algorithm proposed in Zhou et al. [144] and applied by Chen et al. [28], adding application domain-specific knowledge to its model. Despite its simplicity, the technique provides a comprehensive and practical framework for in silico prediction of relationships between biological entities including drug and targets.

Algorithm **DT-Hybrid** is a recommendation algorithm that uses domain-specific knowledge to obtain accurate in silico predictions of DTIs. Let $D = \{d_1, d_2, \ldots, d_m\}$ be a set of small molecules (i.e. biological compounds, biotech drugs), and $T = \{t_1, t_2, \ldots, t_n\}$ a set of targets (i.e. genes, proteins), the network of D-T interactions can be described as a bipartite graph $G(D, T, E)$ where $E = \{e_{ij} : d_i \in D, t_j \in T\}$. A link between d_i and t_j is drawn in the graph when the small-molecule d_i is associated with the target t_j. Such a network can be summarized by an adjacency matrix $A = \{a_{ji}\}_{n \times m}$, where $a_{ji} = 1$ if d_i is connected to t_j, otherwise $a_{ji} = 0$.

In Zhou et al. [64, 144], authors proposed a recommendation method based on the bipartite network projection technique implementing the concept of resources transfer within such a network. Given the bipartite graph defined above, a two phase resource transfer is associated with

Table 4.4: List of Algorithms with the associated Γ functions.

	Algorithm	Γ function
(1)	NBI [144]	$\Gamma(i,j) = k(t_j)$
(2)	HeatS [64]	$\Gamma(i,j) = k(t_i)$
(3)	Hybrid N+H [64]	$\Gamma(i,j) = k(t_i)^{1-\lambda} k(t_j)^{\lambda}$
(4)	**DT-Hybrid**	$\Gamma(i,j) = \left(k(t_i)^{1-\lambda} k(t_j)^{\lambda}\right) / s_{ij}$

one of its projections: at the beginning a resource is transferred from nodes belonging to T to those in D, subsequently the resource is transferred back to T nodes. This process allows us to define a technique for the computation of a weight matrix ($W = \{w_{ij}\}_{n \times n}$), which represents the projection, as follows:

$$w_{ij} = \frac{1}{\Gamma(i,j)} \sum_{l=1}^{m} \frac{a_{il} a_{jl}}{k(x_l)}, \tag{4.6}$$

where Γ determines how the distribution of resources takes place in the second phase, and $k(x)$ is the degree of node x in the bipartite graph. By varying the Γ function we obtain the following algorithms (see Table 4.4 for more details):

- *NBI*, introduced in Zhou et al. [144], and used in Cheng et al. [30] for the prediction of DTIs;
- *HeatS*, introduced in Zhou et al. [64];
- *Hybrid N+H*, introduced in Zhou et al. [64], in which the functions defined in NBI and HeatS are combined in connection with a parameter called λ.

Finally, given the weight matrix W and the adjacency matrix A of the bipartite network, it is possible to compute the prediction matrix $R = \{r_{ij}\}_{n \times m}$ by the product:

$$R = W \times A. \tag{4.7}$$

For each d_i in D, its recommendation list is given by the set $R_i = \{(t_j, r_{ji}) \mid a_{ji} = 0\}$ where r_{ji} is the score of recommending t_j to d_i. This list is sorted in a descending order with respect to the

score, since the higher elements are expected to have a better interaction with the corresponding structure.

DT-Hybrid is an enhanced version of the *Hybrid N+H* algorithm in which prior domain-dependent biological knowledge is plugged into the model through (i) a similarity matrix between small molecules, and (ii) a sequence similarity matrix between targets.

Let $S = \{s_{ij}\}_{n\times n}$ be the targets similarity matrix (i.e. either BLAST bits scores [215], or *Smith-Waterman* local alignment scores [216]). This information can be taken into account by using equation 4.6 with $\Gamma(i,j)$ defined as in row 4 of table 4.4.

Including structural similarity requires more effort. Therefore, it is necessary to manipulate such information in order to obtain a variant of the S matrix, and simplify the computation of the equation 4.6. Let $S_1 = \left\{s'_{ij}\right\}_{m\times m}$ be the structure similarity matrix (i.e. *SIMCOMP* similarity score [217] in the case of small molecules). It is possible to obtain a matrix $S_2 = \left\{s''_{ij}\right\}_{n\times n}$, where each element s''_{ij} describes the similarity between two targets t_i and t_j based on the common interactions in the network, weighting each one by drugs similarity. In other words, if two targets t_i and t_j are linked by many highly similar drugs than s''_{ij} will be high. S_2 can be computed as:

$$s''_{ij} = \frac{\sum_{k=1}^{m}\sum_{l=1}^{m}\left(a_{il}a_{jk}s'_{lk}\right)}{\sum_{k=1}^{m}\sum_{l=1}^{m}\left(a_{il}a_{jk}\right)}. \tag{4.8}$$

Such a matrix can be linearly combined with the target similarity matrix S,

$$S^{(1)} = 1 + \alpha S + (1-\alpha)\, S_2, \tag{4.9}$$

where α is a tuning parameter. This additional biological knowledge yields faster computation together with higher numerical precision. The matrix defined by equation 4.9 in connection with equations 4.6 and 4.7 allows the prediction of recommendation lists, which are used to select the most promising interaction for further laboratory experimentations.

Table 4.5: Description of the dataset: number of biological structures, targets and interactions together with a measure of sparsity. The sparsity is obtained as the ratio between the number of known interactions and the number of all possible interactions.

Data set	Structures	Targets	Interactions	Sparsity
Enzymes	445	664	2926	0.0099
Ion Channels	210	204	1476	0.0344
GPCRs	223	95	635	0.0299
Nuclear Receptors	54	26	90	0.0641
Complete Drugbank	4398	3784	12446	0.0007

Performance assessment and Data Sources In order to correctly evaluate **DT-Hybrid**, the four datasets used in Cheng et al. [30] containing experimentally verified DTIs were used. The datasets were built by grouping known DTIs based on their main gene types: enzymes, ion channels, GPCRs and nuclear receptors (see Table 4.5 for the details). The following similarity measures have been used: (i) *SIMCOMP* 2D chemical similarity for small molecules [217], and (ii) *Smith-Waterman* sequence similarity of genes [216]. Similarities have been normalized according to Yamanishi et al. [143]:

$$S_{norm}(i,j) = \frac{S(i,j)}{\sqrt{S(i,i) \cdot S(j,j)}}. \tag{4.10}$$

Results are evaluated by combining the methods presented in [64] and [30]. More precisely a 10-fold cross-validation has been applied and the experiments were repeated 30 times. Notice that, the random partition used in the cross validation could cause isolation of nodes in the network on which the test is performed. Since all the tested algorithm are capable to predict new interactions only for drugs and targets for which we already have some information, we compute each partition so that for each node, at least one link remains to the other nodes in the test set. Consistent with Zhou et al. [64] the comparison between methods was performed by applying also Precision and Recall Enhancement computed as described in section 2.1.3.

Results The evaluation of results obtained by comparing **DT-Hybrid** with competing methodologies shows that adding domain-related knowledge improves the algorithms in terms of predic-

Table 4.6: Optimal values of λ and α parameters for the data sets used in the experiments (Enzymes, Ion Channels, GPCRs, Nuclear Receptors, Complete Drugbank).

Data set	λ	α
Enzymes	0.5	0.4
Ion Channels	0.5	0.3
GPCRs	0.5	0.2
Nuclear Receptors	0.5	0.4
Complete Drugbank	0.8	0.7

Table 4.7: Comparison of **DT-Hybrid**, Hybrid and NBI through the precision and recall enhancement metric computed for each dataset listed in Table 4.5. The results were obtained using the optimal values for λ and α parameters as shown in table 4.6.

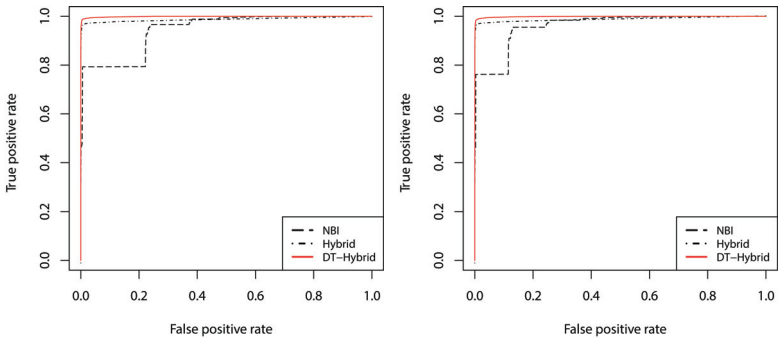
	$e_P\,(20)$			$e_R\,(20)$		
Data set	NBI	Hybrid	**DT-Hybrid**	NBI	Hybrid	**DT-Hybrid**
Enzymes	103.3	104.6	**228.3**	19.9	20.9	**32.9**
Ion Channels	22.8	25.4	**37.0**	9.1	9.7	**10.1**
GPCRs	27.9	33.7	**50.4**	7.5	**8.8**	5.0
Nuclear Receptors	28.9	31.5	**70.2**	0.3	**1.3**	**1.3**
Complete Drugbank	538.7	861.3	**1141.8**	55.0	85.7	**113.6**

tion of novel biologically significant interactions. Table 4.6 reports the optimal values found for parameters λ and α. Such values were obtained by computing the quality of results by varying each parameter and selecting those which achieved better biological significance.

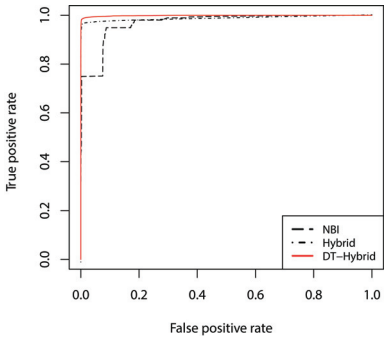
Table 4.7 illustrates the result of comparing *NBI*, *Hybrid* and **DT-Hybrid** in terms of precision and recall enhancement. **DT-Hybrid** clearly outperforms both *NBI* and *Hybrid*, in recoverying of deleted links. It is important to point out that hybrid algorithms are able to significantly improve recall (e_R) measuring the prediction ability of recovering existing interactions in a complex network.

Figure 4.3 illustrates the receiver operating characteristic (ROC) curves calculated over the complete drugbank data set. Simulations were executed 30 times and the results were averaged to obtain a performance evaluation.

Experiments show that all of the three techniques have high *True-Positive Rate* (TPR) against a low *False-Positive Rate* (FPR). However, hybrid algorithms provided better performance than *NBI*. In particular Table 4.8 clearly shows an increase of the average areas under the ROC



(a) $L = 10$ **(b)** $L = 20$



(c) $L = 30$

Figure 4.3: Comparison between **DT-Hybrid**, Hybrid and NBI by means of receiver operating characteristic curves (ROC), computed for the top-L places of the recommendation lists, which were built upon the complete drugbank data set.

Table 4.8: Comparison of **DT-Hybrid**, Hybrid and NBI through the average area under ROC curve (AUC) calculated for each dataset listed in Table 4.5. The results were obtained using the optimal values for λ and α parameters as shown in table 4.6.

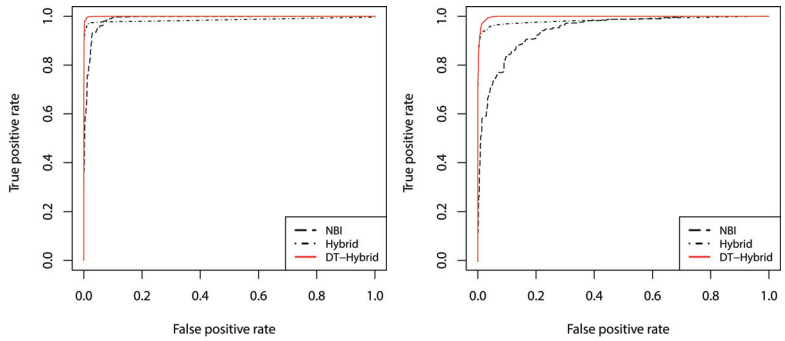
	$AUC\,(30)$		
Data set	NBI	Hybrid	**DT-Hybrid**
Enzymes	0.9789±0.0007	0.9982±0.0002	**0.9995±0.0001**
Ion Channels	0.9320±0.0046	0.9929±0.0008	**0.9973±0.0006**
GPCRs	0.9690±0.0015	0.9961±0.0007	**0.9995±0.0006**
Nuclear Receptors	0.9944±0.0007	0.9986±0.0004	**1.0000±0.0000**
Complete Drugbank	0.9619±0.0005	0.9976±0.0003	**0.9989±0.0002**

curves (AUC) in the complete data set. This indicates that hybrid algorithms improve the ability of discriminating known links from predicted ones. The increase of AUC values for **DT-Hybrid** demonstrates that adding biological information to prediction is a key choice to achieve significant results. Figure 4.4 illustrates the receiver operating characteristic (ROC) curves calculated on the Enzymes, Ion Channels, GPCRs, and Nuclear Receptors data sets using the Top-30 predictions. Finally it can be asserted that adding similarity makes prediction more reliable than an algorithm, such as *NBI*, which applies only network topology to score computation.

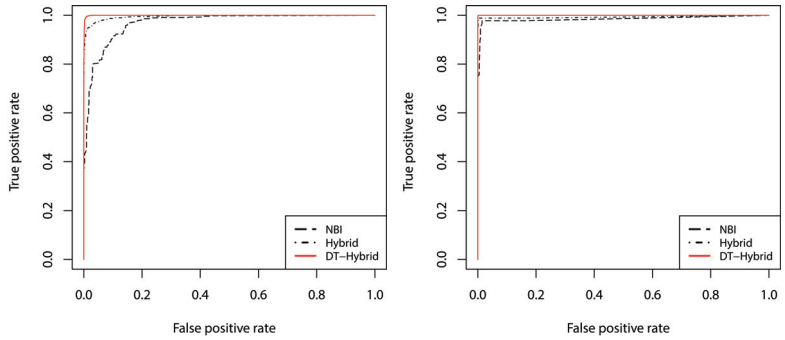
4.3 New therapies: Drug Combination Prediction

In recent years, due to the complexity of pathologies, it has become necessary the development of multi-target drugs and drug combinations. The latter, despite the complexity, is the most promising approach. Indeed, by combining multiple drugs it is possible to enhance their effect reducing or removing side effects. For example, the RAS oncogene mutation, very frequent phenomenon in many human cancers [218], has no effective therapeutic approaches due to their high toxicity, and, even if viable, a high level of resistance can be observed. To overcome such deficiencies many different approaches have been proposed, based for example on drug combinations operating on parallel pathways [165, 166].

Already researches can relay on public-domain databases of drug combination references, such as Liu et al. [219]. However there is still a lack of systematic computational approaches, and



(a) Enzymes **(b)** Ion Channels



(c) GPCRs **(d)** Nuclear Receptors

Figure 4.4: Comparison between **DT-Hybrid**, Hybrid and NBI by means of receiver operating characteristic curves (ROC), computed for the top-30 places of the recommendation lists, which were built upon the four data sets (Enzymes, Ion Channels, GPCRs, and Nuclear Receptors).

often several possible drug combinations are disclaimed by expert knowledge and verified via clinical trials. Motivated by the success of network-based approaches, a multi-purpose pathway analysis, which relies on a multi-drug, multi-target, multi-pathway approach, has been developed to provide a limited set of candidate drugs both for drug repositioning and combination, that can be directly evaluated by the experts or combined with other methods. The algorithm has been implemented as part of the **DT-Web** database [32].

Algorithm The aim of the multi-purpose pathway analysis is to discover the minimal set of drug targets that are able to affect a user-specified set of genes in a multi-pathway environment. The distances among such targets and user genes are limited to a given range in order to minimize drug side effects. The set of validated drug targets is extended with **DT-Hybrid** predictions, along with their score to give a measure of confidence on each prediction.

The implemented pipeline has been divided into two main phases. The first phase is performed off-line and kept up-to-date whenever the **DT-Web** database is synchronized with the latest version of **Drugbank** [220–222]. The second one is performed on-line and responds to the request of an user passed through a submission form.

Off-line Database Building Pipeline The calculation of a multi-pathway environment requires huge computational resources and it is a time-consuming task. Because of this, the construction of such an environment, consisting of merging all *Homo sapiens* metabolic and signaling pathways contained in *Reactome* [223] and *PID* [224], is done off-line through a proprietary Java module, and stored in a database. The steps are the following.

Step 1. All pathways are retrieved by downloading *BioPax* [225] level 3 *XML* files from the **Pathway Commons** [181] web service, using PC2 for the remote connection to the public database.

Step 2. For each pathway, first entity names are normalized if these exist (i.e. symbolic names for proteins, such as BRCA1), otherwise we consider the BioPax entity reference IDs.

Step 3. Subsequently, we collapse all nodes representing the same biological entity (protein sub-units or the same protein in different cell locations) in a single node, and we map them to the Drugbank database. Edge directions are kept as they are in the input network, except for those which connect a complex to a constituent protein that are made undirected.

Step 4. The entire set of retrieved pathways is merged into a single global network by mapping nodes and edges using their names and interaction types, respectively.

Step 5. Finally, to control the combinatorial explosion of such data, we store only directed shortest paths between proteins that lie at a distance of, at most, 9 edges. Moreover, since a path could contain edges belonging to different pathways, we decided to store for each edge the list of pathways where it appears. We, also, store the mapping to Drugbank database computed at step 3.

Real-time Prediction Pipeline The multi-pathway environment computed and indexed as described above can be easily queried to perform predictions. From a set of genes altered in some way by a disease, a list of druggable proteins can be traced back. Such proteins should not be too far away from the genes affected by the disease in order to reduce the cascading impact of drugs on other proteins, thus avoiding side effects (`Direct-Indirect Range`). Furthermore, such proteins should be as close as possible between them to ensure the a synergistic effect (`Pair Range`). Finally, the resulting set of proteins should be as small as possible to minimize the number of drugs. The steps that implement this algorithm are as follows.

Step 1. The user provides a list of genes (names in HGNC format, or Uniprot Accession Number, or Entrez Gene Id, or HGNC Id, or Ensembl Gene Id) through a web interface. He can also set the ranges (min/max) for distances between drug targets and user-provided genes `Direct-Indirect Range`, or between each pair of drug targets `Pair Range`.

Step 2. Users' data are thus filtered to remove all proteins that are not present in our database. If the filtered list is not empty, a search is performed in our multi-pathway environment. The task selects all proteins which are at distance within the `Direct-Indirect Range` specified by

the user.

Step 3. Each protein is then mapped in Drugbank, and those targeted by at least one drug are selected as a preliminary list of targets. Such a list is further filtered by removing all pairs of targets which are outside of the user-specified `Pair Range`.

Step 4. Next, by applying Chvatal [226], an approximation of the minimum list of targets needed to reach all the user-specified genes can be quickly computed.

Step 5. Finally, the list of all targets calculated in step 3, and each associated drug (experimentally validated or predicted), is returned to the user, along with the minimum set computed in step 4.

Case Study The multi-purpose pipeline, being a new software, has not yet been used for the intended studies. That is why we used such a methodology to predict the combination of Propofol and Sevoflurane whose additive action produces consciousness and movement to skin incision during general anesthesia [227]. Both drugs interact with the $GABA_A$ receptor. Propofol is a potentiator of the $\beta 2$ subunit (GABRB2) of $GABA_A$, while Sevoflurane is an agonist of the $\alpha 1$ subunit (GABRA1) of $GABA_A$ with its binding site between both subunits [228]. Probably is such a location which hinders agonist activity, thereby producing mutually substitutable actions [229].

4.4 Enhancing pathways: mining of biological data

Although in the previous sections the merits, and untapped potential, of pathway analysis algorithms have been highlighted, such methodologies are not yet fully reliable as can be seen from the ROC curves shown in section 4.1. This is mainly due to phenomena that biological pathway do not account for several reasons.

For example, long non-coding RNA are long RNA molecules not translated into proteins which to date are not fully functionally characterized but showed different correlations with the regulation of gene expression, and consequently are also involved in the onset and progression of

diseases. Another important phenomena is RNA editing, which by altering a single nucleotide in an RNA molecule can cause important alterations.

This has led to the development of several algorithmic methodologies that try to fill some gaps in current knowledge, or at least try to highlight some unclear correlations. These techniques are based on disparate, and have the ultimate goal of bringing to a further enrichment of pathways, making their analysis more fruitful and reliable. In this section some of such algorithms will be described.

4.4.1 Prediction of ncRNA-Disease Association

As stated above, long non-coding RNAs (lncRNAs) are RNA molecules not translated into proteins, whose function is mostly unknown. They compose the majority of transcribed RNA sequences in a human cell, and have been associated with gene expression regulation mechanisms, such as chromatin remodeling. This implies that the study of such molecules is of fundamental importance. For this purpose, the development of a methodology that is able to predict ncRNA-disease interactions in silico is crucial in order to formulate new hypotheses on the molecular mechanisms underlying complex diseases, and to identify potential new biomarkers for their diagnosis, treatment and prevention.

In this direction, ncPred [42], a resource propagation methodology, uses a tripartite network to guide the inference process of novel ncRNA-disease associations. The tripartite network allows the introduction of two levels of interaction: ncRNA-target and target-disease. Here, we call targets a group of biomolecules (i.e., genes, microRNAs, proteins) whose activity is modulated by a ncRNA (e.g., regulation of expression, binding to improve the efficiency of its activity, or binding to help the formation of complexes). In this way, we can exploit the greater quantity of known interactions between targets and diseases to build a wider knowledge base and obtain a greater number of high quality predictions.

Algorithm Let $O = \{o_1, o_2, \ldots, o_n\}$ be a set of lncRNAs, $T = \{t_1, t_2, \ldots, t_m\}$ a set of targets, and $D = \{d_1, d_2, \ldots, d_p\}$ a set of diseases. The ncRNA-target and target-disease interactions can be represented in a tripartite graph $G\,(O, T, D, E)$, where E is the set of interactions (edges) between nodes in O and T and nodes in T and D. Such a graph, can be represented by using a a pair of adjacency matrices $A^{OT} = \left\{a_{ij}^{OT}\right\}_{n\times m}$ and $A^{TD} = \left\{a_{rs}^{TD}\right\}_{m\times p}$ where $a_{ij}^{OD} = 1$ if o_i is connected to t_j in G, and $a_{rs}^{TD} = 1$ if t_r is connected to d_s in G.

ncPred, in the same way as Alaimo et al. [31], is based on the concept of resources transfer within a network. Due to the tripartite network, we developed a multi-level transfer approach that at each step takes into account the resource transferred in the previous one (see figure 4.5 for an example). In the first level of the transfer, the resource is moved from the nodes in T (targets) to nodes in O (ncRNAs) and vice versa. In the second level, the resource is moved from D nodes to T nodes and it is combined with the resource of the previous step. Then the resources is moved back to the D nodes. In this way, we define a methodology for the computation of a combined weight matrix $W^C = \left\{w_{ij}^c\right\}_{m\times p}$, where w_{ij}^c corresponds to the likelihood allowing us to claim that if a ncRNA interacts with a target t_i then it may be associated with the pathology d_j.

To compute such a matrix, we start by defining two partial weight matrices corresponding to the intermediate levels of transfer. These two matrices are then used to obtain the combined weight matrix and, therefore, compute the recommendations.

Let $k'\,(x)$ be the degree of node x in the ncRNA-target sub-network and $k''\,(y)$ the degree of node y in the target-disease sub-network.

The matrix $W^T = \left\{w_{ij}^T\right\}_{m\times m}$, associated with the first level of transfer, can be defined as:

$$w_{ij}^T = \frac{1}{k'\,(t_i)^{(1-\lambda_1)}\,k'\,(t_j)^{\lambda_1}} \sum_{l=1}^{n} \frac{a_{li}^{OT} a_{lj}^{OT}}{k'\,(o_l)}, \tag{4.11}$$

where w_{ij}^T corresponds to the likelihood that given a ncRNA interacting with target t_i, then it

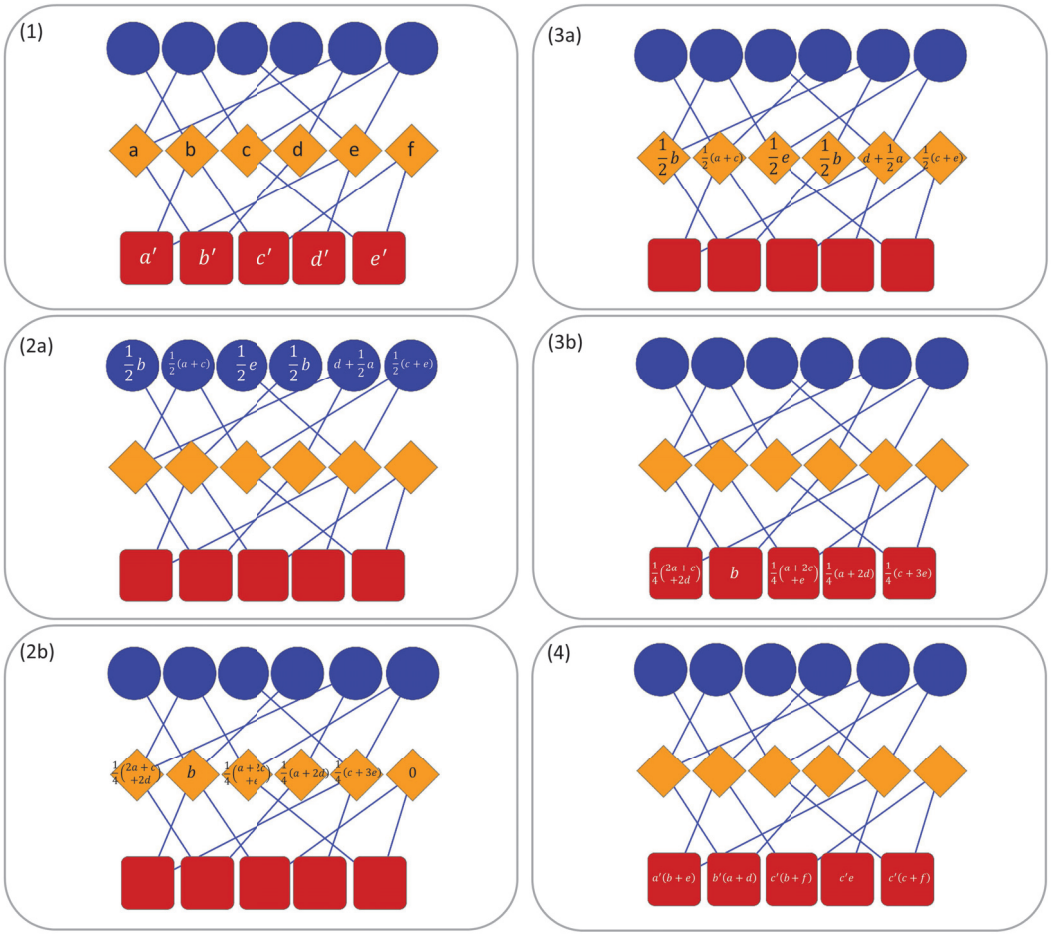


Figure 4.5: Operating principle of ncPred in a tripartite network. Here we represent ncRNAs in blue, targets in orange, and diseases in red. Without loss of generality, and in order to simplify the reading of the image we decided to put λ_1 and λ_2 to 1, so as to obtain a uniform distribution of resources in the network. In the first step, a resource is assigned to each target and disease node (**1**). Thereafter two separate transfer process are launched to compute the resource in target nodes (**2a, 2b**) and disease nodes (**3a, 3b**). Finally, resources are combined to obtain the total quantity in each disease node (**4**). In (**4**), the literals are used only for example purposes due to lack of space. They are to be replaced with the values computed in steps (**2b**) and (**3b**).

may also interact with target t_j. By using such an equation, we assign higher weights to the pairs of targets that share many ncRNAs, rather than those who share only a few.

The same applies to $W^D = \left\{w_{ij}^D\right\}_{p\times p}$, matrix associated with the second level of the transfer, where:

$$w_{ij}^D = \frac{1}{k''\left(d_i\right)^{(1-\lambda_2)}\, k''\left(d_j\right)^{\lambda_2}} \sum_{l=1}^{m} \frac{a_{li}^{TD} a_{lj}^{TD}}{k''\left(t_l\right)}. \tag{4.12}$$

In equation 4.12, w_{ij}^D indicates whether we can assert that given a target associated with the disease d_i, it may also be linked to the disease d_j. w_{ij}^D is higher for the disease pairs which are associated to many common targets with respect to those with fewer common targets.

In equations 4.11 and 4.12, the $\lambda_1 \in [0,1]$ and $\lambda_2 \in [0,1]$ parameters are used to tune the quality of the predictions. Parameter values close to zero indicate that the resource of a node is computed as the average of those in its neighborhood, while values close to one indicate that the resource is uniformly distributed among the nodes of its neighborhood. In terms of predictions, lambda values close to zero correspond to conservative predictions, while values close to one correspond to a larger number of predictions.

Therefore, the combined weight matrix $W^C = \left\{w_{ij}^c\right\}_{m\times p}$ can be obtained as:

$$w_{ij}^C = \sum_{t=1}^{m} \left[w_{it}^T \sum_{r=1}^{p} \left(a_{tr}^{TD} \cdot w_{rj}^D\right) \right]. \tag{4.13}$$

In equation 4.13, the weight of a target-disease pair is computed by taking into account both the targets with a similar neighborhood and the diseases with a similar neighborhood. In this way, a larger weight is assigned to those pairs for which more frequently there is a path which passes through them.

Given the above weights, it is now possible to compute a prediction matrix $R = \{r_{ij}\}_{n\times p}$ as:

$$R = A^{OT} \cdot W^C. \tag{4.14}$$

We call each r_{ij} prediction score for the pair (i, j). For each ncRNA o_i, its list of predictions R_i can be obtained by selecting those disease-prediction score pairs for which there is no path with o_i in the tripartite network. Such a list is sorted in descending order with respect to the value of r_{ij}, as the higher the score, the greater the belief that the ncRNA will have some connection with that particular disease.

Performance assessment and Data Sources In order to evaluate **ncPred** , two datasets containing experimentally verified interactions between ncRNAs, targets, and diseases have been used. The first data set was built by collecting from [170] 478 interactions between lncRNAs and genes. These interactions were mapped by converting each target identifier to its Entrez Id. This allowed removing about 230 duplicates or superseded interactions. From the remaining targets, we then extracted 1005 experimentally validated gene-disease associations by searching in DisGeNET [230].

The second data set was obtained by collecting about 4000 lncRNA-miRNA interactions found in [231] by applying the CLASH methodology [232]. Each association indicates that a lncRNA contains one or more binding sites for miRNAs. From such a list, we removed all targets not present in miR2Disease database [233], obtaining 1699 lncRNA-miRNA associations. Finally, using Jiang et al. [233], we recovered 1572 miRNA-disease associations. Table 4.9 provides a summary of the two datasets together with some metrics that can further elucidate their characteristics.

For the evaluation of **ncPred**, a 10-fold cross-validation procedure repeated 30 times has been applied to obtain more reliable results. Each fold is built in the the following way. Given the tripartite graph, we selected all possible pairs of ncRNA-disease interactions. Then, we randomly partitioned them into each fold. We make sure that the tripartite network generated from each fold is not disconnected. **ncPred** makes predictions only on connected networks. The four metrics of Alaimo et al. [31] have been used to assess performances: precision and recall enhancement, recovery, personalization, and Surprisal. The first two establish the ability of the

Table 4.9: Description of the datasets: number of ncRNAs, targets and diseases together with the count of interactions, average degree, density, modularity, number of connected components, and average path length.

Metrics	Chen et al.	Helwak et al.
ncRNAs	119	338
Targets	110	179
Diseases	514	134
ncRNAs-Targets Interactions	247	1699
Targets-Diseases Interactions	1005	1572
Average Degree	1.572	5.025
Density	0.002	0.008
Modularity	0.609	0.274
Number of Connected Components	24	1
Average Path Length	1.572	1.734

method to recover the interactions of the test set, therefore obtaining biologically relevant predictions. The other two measure the ability of the method to propose unexpected interactions, which may lead to novel insights onto ncRNA functions. Special care should be given to the precision and recall enhancement metrics. They measure the reliability of the prediction algorithm by comparing the standard precision and recall with a null model. Such a model is defined as a methodology that randomly assigns ncRNA-disease pairs. This implies that values greater than one are to be considered synonymous of higher quality and, therefore, reliability.

Results In Table 4.10, we illustrate the behavior of **ncPred**, comparing it with Yang et al. [41], in terms of precision and recall enhancement. The results demonstrate that **ncPred** clearly outperforms its competitor. In particular we can see that while Yang et al. [41] obtains a recall close to the null model, **ncPred** has much better results. This is crucial since the recall measures the ability of the algorithm to recover existing interactions in the network, and is therefore a sign of their reliability, namely their biological relevance.

In Figure 4.6, we report the receiver operating characteristic (ROC) curves computed on both datasets. The simulations were repeated 30 times and their results were averaged to obtain a more accurate evaluation. Both methods show a high true positive rate against low false positive rate, although **ncPred** is clearly able to achieve better results. This is also shown in Table 4.10,

Table 4.10: Comparison of **ncPred** and Yang et al. [41] through the precision and recall enhancement metric, and the average area under ROC curve (AUC) calculated for each of the two datasets listed in Table 4.9. The results were obtained using the optimal values for λ_1 and λ_2 parameters as shown in Table 4.12.

	$e_P\,(20)$		$e_R\,(20)$		$AUC\,(20)$	
Dataset	Yang et al.	**ncPred**	Yang et al.	**ncPred**	Yang et al.	**ncPred**
Chen et al.	5.5113	**12.3290**	0.7297	**1.6636**	0.6217±0.0178	**0.7566±0.0218**
Helwak et al.	1.8654	**5.8197**	1.6509	**5.6572**	0.7069±0.0084	**0.7669±0.0093**

Table 4.11: Friedman rank sum test applied to establish the statistical significance in the performance improvement of **ncPred** compared to [41].

Dataset	Friedman χ^2	p-Value
Chen et al.	1026.315	$< 2.2 \cdot 10^{-16}$
Helwak et al.	6537.915	$< 2.2 \cdot 10^{-16}$

where we can see a significant increase in the average area under the ROC curve (AUC). Such a significance is further proved by the results shown in Table 4.11. By applying the Friedman rank sum test, we determined that the performance improvement achieved by our algorithm is statistically significant (i.e. the p-value is close to zero on both datasets).

Regarding the parameters λ_1 and λ_2 , the peculiar characteristics of each dataset greatly affect the performances and, consequently, the parameters. It is, therefore, necessary to perform an *a priori* analysis in order to determine which values give the best results. In our experiments, we used such an analysis to determine the best parameters in terms of precision and recall enhancement (see Table 4.12 for details on their values). By looking at the characteristics of our data sets, the values obtained from such an analysis allowed us to suppose that the two parameters are close to zero in Helwak et al. [231] dataset because of the greater density. This implies that to maintain high quality predictions it is necessary to reduce their number to avoid the introduction of noise. On the other hand, the Chen et al. [170] dataset has a lower density. This allows us to produce a higher number of predictions before they start losing quality. Therefore, this explains the lambda values closer to one. It is important to point out that in order to determine the best parameters an analysis was performed considering only precision and recall enhancement, since they are closely related to the biological significance of the predictions.

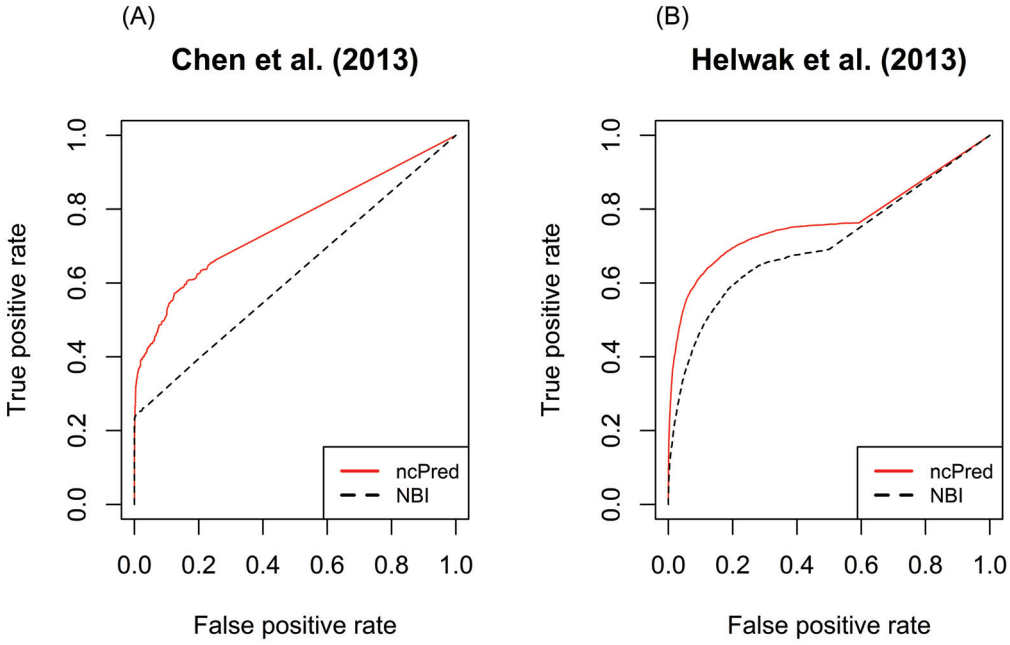


Figure 4.6: Comparison between **ncPred** and Yang et al. [41] by means of receiver operating characteristic (ROC) curves, computed for the recommendation lists built on the two datasets. Such curves measure the quality of the algorithms in terms of false positives rate against true positives rate. (A) and (B) are independent since computed on two separate datasets. The significance of the difference highlighted between **ncPred** and Yang et al. [41] was measured by applying the Friedman rank sum test as assessed in Table 4.11.

Table 4.12: Optimal values of λ_1 and λ_2 parameters for the datasets used in ncPred experiments.

Dataset	λ_1	λ_2
Chen et al.	0.5	1
Helwak et al.	0.2	0.2

4.4.2 Detecting possible A-to-I RNA editing signals

RNA editing is an important post-transcriptional change in RNA sequence. The most common alteration, carried out by a specific family of proteins called **ADARs**, is the **A-to-I editing** where an *adenosine* is converted in an *inosine*, which is treated as a *guanosine* by the translation machinery. This change albeit minimal can have profound changes both on the structure of proteins, where even a modification of a single amino acid can affect significantly its functioning, and on the expression of *mRNA*, where *miRNA* or binding sites editing can destroy or create interactions with *mRNA*, indirectly changing their expression. This phenomenon has a major impact on cellular pathway by altering or enhancing their functions. Moreover alterations of the editing mechanism have been associated with various diseases, such as neurodegenerative ones.

This shows the crucial importance of understanding such mechanisms. However due to their dynamic nature, laboratory experiment are difficult and costly. To this end, an in silico prediction of possible editing sites to restrict experiments is crucial. Moreover, it has been demonstrated that albeit the activity of **ADAR** is not random, virtually the entire genome can be edited. In this context fits **AIRlINER** [50], an algorithmic approach for the prediction of A-to-I RNA editing sites in non-repetitive regions. The method has been compared with *InosinePredict* [234], a similar technique, which analyzes the nucleotides flanking the editing site. *InosinePredict* assumes a multiplicative relationship between the coefficients necessary to compute the percentage of editing. The comparisons clearly show that **AIRlINER** improves the quality of predictions with respect to *InosinePredict* and suggest further research directions.

Algorithm, Data sources and Performance assessment Starting from the idea proposed by Pinto et al. [235], we used a logistic regression technique to determine a model from which we can compute the probability that an adenosine in a non-repetitive region of the genome is affected by the *A-to-I editing* phenomenon. **AIRlINER** determines the editing probability of an adenosine by analyzing its flanking region of 10 nucleotides. Such pattern is then combined

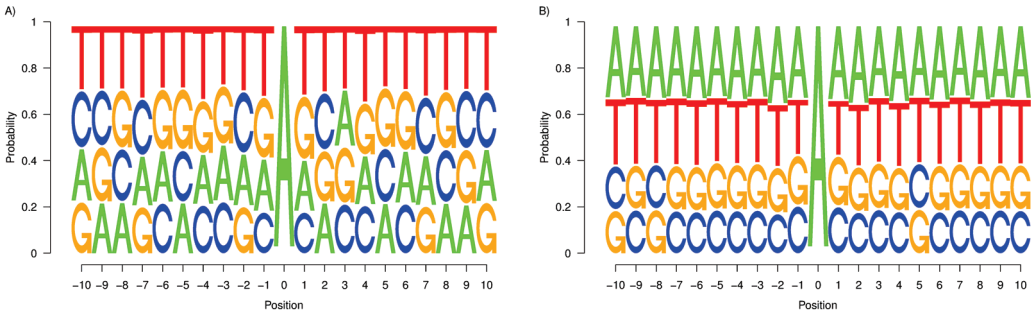


Figure 4.7: Neighborhood preferences computed for experimentally verified editing sites in non-repetitive regions (A) and random sites chosen among those for which no editing is reported (B). Neighborhood preferences are coherent with the upstream nucleotide distribution of editing site sequence contexts reported in Eggington et al. [234].

with a similar model calculated from un-edited sequences, resulting in the estimation of an unbiased editing probability.

In order to train the method, we built a dataset composed of $30,280$ sequences of 21 nucleotides centered on an adenosine, from the human genome (hg19). According to their provenance, our dataset can be divided equally into two sets: known editing sites and random sites. For the purpose of retrieving known editing sites in non-repetitive regions, only human sites which do not have any repetitive elements in their flanking regions of $2,000$ nucleotides were selected from the **RADAR database** [236]. Random sites were chosen by randomly selecting a number of sequences equal to that of the known editing sites. From such a selection we excluded known editing sites in both repetitive and non-repetitive regions.

From such a dataset, two probabilities $P(j,i)$ and $P'(j,i)$ can be estimated: the first one corresponds to the probability of finding nucleotide j in position i of a region affected by editing, while the second one represents the probability of finding nucleotide j in position i of an unedited region. Starting from these probabilities, we computed the graphs in Figure 4.7, which represent the distributions of nucleotides for the two types of regions.

Therefore, let s be a nucleotide sequence and $P(s)$ its editing probability, using the previously

defined probabilities we are able to train a logistic regression model such as:

$$\log\left(\frac{P(S)}{1-P(S)}\right) = \beta_0 + \sum_{i=1}^{21} \beta_i P(s[i], i) - \sum_{i=1}^{21} \beta_i P'(s[i], i), \tag{4.15}$$

where $s[i]$ is the i-th nucleotide in a sequence. Now we can use this model to estimate the editing probability of any sequence of 21 nucleotides centered on an adenosine, and if such probability is greater than 0.5, we can say that such a sequence may be affected by editing.

To tune and validate **AIRlINER**, a 10-fold cross validation procedure has been applied and mean error computed. To compare our method with *InosinePredict*, we used a threshold to establish the presence or absence of editing in a specific sequence. Such a threshold was set to 9.6% for *InosinePredict*, as shown in Eggington et al. [234]. For our algorithm, we choose all sites for which an editing probability > 0.5 is computed. We also took into account the fact that *InosinePredict* can produce predictions for both types of *ADAR* enzymes. We do not have this information in our dataset, so we chose to select the maximum score produced by *InosinePredict* for editing sites, and the minimum score for random sequences. Consequently, we are able to ensure a fair comparison with our method despite the absence of information on which *ADAR* affects each editing site.

Results In Tables 4.13 and 4.14, we show the confusion matrices computed using the previously described procedure. The two algorithms were applied to the dataset and the values computed for the central adenosines in each sequence were used to determine the presence or absence of editing. **AIRlINER** significantly reduces the number of false negatives compared to *InosinePredict*, thus resulting in a better editing sites prediction quality. **AIRlINER** is also able to achieve a substantial reduction of false positives, even if nothing can be stated with certainty about them, as the absence of editing in these sites can also be determined by lack of experimental tests. The best quality in predicting editing sites, however, may reflect the fact that the random sequences classified as non-edited could be with high probability considered as such.

Table 4.13: Confusion matrix computed by applying InosinePredict [234] to our dataset. Editing percentages for each site have been divided into two classes (editing/non-editing) using the thresholds defined in Eggington et al. [234]

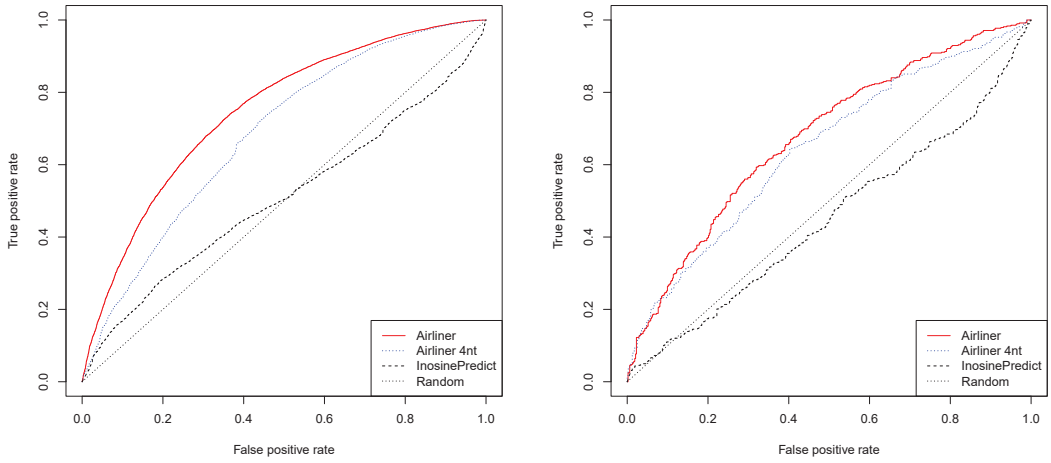
		Prediction Outcome	
		Editing site	**Random site**
Actual Value	**Editing sites**	58.48	41.52
	Random sites	60.18	39.82

Table 4.14: Confusion matrix computed by applying AIRIINER to our dataset. All editing sites for which editing probability > 0.5 were classified as editing while the remaining as non-editing.

		Prediction Outcome	
		Editing site	**Random site**
Actual Value	**Editing sites**	71.18	28.82
	Random sites	34.05	65.95

Further confirmation of the quality of our methodology is represented by the receiver operating characteristic curves (ROCs), Figure 4.8, computed from the results produced by the two algorithms. The curves demonstrate a significant improvement in performance. Such curves also show that the threshold chosen to distinguish editing sites from non-editing ones does not affect the performance difference between the two algorithms. As a confirmation of this, *InosinePredict* obtains an average area under the ROC curve (AUC) of 0.5072, while **AIRlINER** reaches 0.7466. In Figure 4.8, we also compare a variant of the method, **AIRlINER 4 nt**, with *InosinePredict*. Such a variant computes the editing probability of an adenosine by considering its flanking region of 4 nt. This comparison shows that our strategy is superior to *InosinePredict* even when the prediction is calculated from this same region around an adenosine.

Furthermore, we investigated that *ADAR* acts on each editing site in our training set by building an additional data set from editing sites experimentally identified in Bahn et al. [237]. Using human cell lines *U87MG* in which the gene expression of *ADAR1* was repressed, the authors were able to identify about $4,000$ *ADAR1-specific* editing sites. Four hundreds of such sites were identified in non-repetitive regions. From the latter, we have built a training set using the same procedure described above and trained our model. In Figure 4.8b, we show the results of this experiment by means of ROC curves. Even in this case, **AIRlINER** is significantly better than



(a) RADAR **(b)** Bahn et al. [237]

Figure 4.8: Receiver operating characteristic curve (ROC) computed for the **AIRIINER** and *InosinePredict* . We also provide a ROC curve for a variant of our algorithm (**AIRIINER 4 nt**), which takes into account only the flanking region of 4 nt around an adenosine. Such a curve is useful to compare the performance with our algorithm using the same flanking region. In 4.8a **AIRIINER** shows an average area under the ROC curve (AUC) equal to 0.7466, while *InosinePredict* gets an AUC of 0.5072, and **AIRIINER 4 nt** has an AUC of 0.7464. In 4.8b **AIRIINER** shows an AUC equal to 0.6763, while *InosinePredict* gets an AUC of 0.4498, and **AIRIINER 4 nt** has an AUC of 0.6435.

InosinePredict. As further confirmation, we also computed the AUC, which amounts to 0.6763 for **AIRlINER**, and 0.4498 for *InosinePredict*.

Finally, to verify the quality of the editing sites predicted by **AIRlINER**, we selected from the literature 52 experimentally validated sites by Sanger method and 7 sites validated as non-edited. We then applied the two methodologies and checked how many of them are correctly identified. **AIRlINER** is able to predict 42 of 52 editing sites and 5 of 7 non-editing sites while *InosinePredict* identifies 26 editing sites and 4 non-editing ones.

4.4.3 uPPeRCUT

The development of in silico predictive models of cell lines cytotoxic response to environmental toxicants and drugs exposure is an extremely important task. It allows a better understanding of the underlying mechanisms such as prediction of responses to new compounds under devel-

opment. As part of the *NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge* [56], a dataset containing estimates measuring cytotoxicity in lymphoblastoid cell lines (derived from 884 individuals following in vitro exposure to 156 chemical compounds) was provided, together with genotypic information (presence/absence of about 1.3 million of SNPs) and expression profiles obtained with RNA-seq (about 40,000 genes). These data have been used to develop an approach, which combines different techniques for dimensionality reduction, Locality-Sensitive Hashing (LSH) [238, 239] and Singular Value Decomposition (SVD) [240, 241], with a versatile and efficient technique for cytotoxicity values approximation, Conditional Inference Trees [242, 243]. The resulting pipeline produces fast and efficient predictive models when a high number of independent variables are present. An important point was the use of random projections based on *LSH* functions to build a profile of the patient's genotypic information. This allows mapping onto a smaller space producing a better approximation of the original data highlighting the latent correlations that are hard to identify in a classical analysis.

uPPeRCUT (a Pipeline for the PRediction of Cytotoxicity valUes in lymphoblasToid cell) predicts cytotoxicity values in response to the administration of compounds in cell lines for which genotypic and RNA-seq information are available. Given the high dimensionality of such input data, our approach consists of combining techniques that allow dimensionality reduction, aiming to maintain the properties of the data. The method has three main stages: (i) dimensionality reduction, (ii) construction of the regression model, and (iii) prediction of the new values.

Let S_i ($|S_i| = N_s$) be the vector that contains information about the presence or absence of SNPs for the i-th cell line. We perform a first dimensionality reduction based on **Locality-Sensitive Hashing** (*LSH*), so as to compute new vectors S'_i (with $|S'_i| \ll N_s$). S'_i is calculated by applying N_d times the random projection method of *LSH*, by approximating the cosine similarity between vectors. The method consists of defining in the original space N_d random planes: the distance between the i-th plane and a point in that space corresponds to the value to assign to the i-th dimension of the point in the reduced space. In particular, since the values

for each dimension in the original space are discrete (i.e. 0, 1, or 2), we chose to assign 0 if the distance from the random plane is negative, 1 otherwise.

Subsequently, on the vectors defined above, the *SVD* algorithm is applied, mapping them to a concepts space of N_{v1} dimensions (with $N_{v1} < N_d$), thereby obtaining a matrix where each row corresponds to the reduced genotypic profile for each cell line.

Let E_i ($|E_i| = N_g$) be the vector that contains information about the normalized RNA-seq counts of the i-th cell line. Using the *SVD* algorithm, as previously described, we can map these vectors into a concepts space of N_{v2} dimensions (with $N_{v2} < N_g$), obtaining a matrix where each row corresponds to the reduced expression profile of each cell line.

Starting from those two profiles, we can now compute a combined profile of $N_{v1} + N_{v2}$ dimensions, putting together the rows of the two profiles matrices. If, for some reason, some expression values are missing, zeroes will replace them in the combined profile vector. These profiles are now the independent variables of our problem and we use them to build a regression model based on the **Conditional Based Inference-Tree** (*C-Tree*) algorithm. The choice of this algorithm was done after a comparative analysis of the performance of different methods (including linear regression, Random Forest based regression, CART Tree based Regression, Non-linear least squares regression and Bayesian regression models).

First, for each cell line we extract the K most similar profiles, using the cosine distance as a similarity measure. Next, for each compound under investigation, a regression model is trained based on the *C-Tree* algorithm. At the end of the training procedure, all predictors are used to compute the cytotoxicity values associated with each cell line.

The parameters of our algorithm have been obtained through an exhaustive search of the space of all possible values. The optimal combination of parameters has been selected as the one that yields the minimum *RMSE* (root-mean-square error) score, according to a 10-fold cross validation procedure. Furthermore, to optimize the performances of our prediction pipeline, we selected the parameters with minimum values yielding the best performances. The combination

of optimal values found is: $N_d = 50000$, $N_{v1} = 10$, N_{v2} = 2, $K = 30$. To conclude our analysis, we have validated our method on the data published at the closure of the Leaderboard of the *NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge*, yielding a better *RMSE* score compared to the others (mean *RMSE* 0.269, minimum *RMSE* 0.18, and maximum *RMSE* 0.30).

4.4.4 A text-mining approach to infer of novel hypotheses

The inference of novel knowledge and the generation of new hypotheses from the analysis of current literature is a fundamental process in making new scientific discoveries. Especially in biomedicine, given the enormous amount of literature and knowledge bases available, this process is often complex, and researchers may focus too much on aspects already widely investigated due to poor literature mining. The automatic extraction of information in the form of semantically related terms (or tags) is becoming an aspect of great importance and extensive investigation (Kilicoglu et al., 2012; Stewart et al., 2012).

In this context, **BioTAGME**, a combination of Tagme [67, 68] and **DT-Hybrid** [31], is a technique designed to extract novel knowledge using a text-mining approach on PubMed. Such knowledge is returned as a set of tags, or annotations, whose topic is highly related to the input text one.

Algorithm Given an input set of texts annotated with terms characterizing each document, the aim of **BioTAGME** consists of computing a new set of annotation terms as much as possible related to the input set but having no synonyms among the old annotations. This is achieved by defining a semantic similarity between pairs of terms, which is properly combined over all terms extracted from each input text. All this is enriched with a statistical test supporting that the set of predictions obtained by our algorithm cannot be achieved by chance. Our method consists of the following steps: (i) the user performs a query on PubMed and the set of matched abstracts are downloaded; (ii) Tagme annotates each abstract with entities drawn from Wikipedia (here an entity is modeled as a Wikipedia page); (iii) then, we build an unweighted bipartite network

consisting of two types of nodes: abstracts and entities, with edges denoting whether an entity occurs in an abstract; (iv) we apply **DT-Hybrid** over this graph in order to recommend more entities to each abstract; (v) a correlation score is estimated for each pair of annotations (i.e. entity-entity), by means of a semantic similarity measure, which is built by annotating with TAGME each entity's Wikipedia page, and computing for each pair of entities a cosine distance between their annotations; finally, (vi) based upon such correlation measure, we compute a set of highly correlated entities along with a p-value that expresses its quality.

NOTATION AND TERMINOLOGY Given a set of n texts $T = \{t_1, t_2, \ldots, t_n\}$ and a set of m terms $S = \{s_1, s_2, \ldots, s_m\}$, we call an **annotation** of T with terms in S a function $f_T : T \rightarrow P(S)$ that associates, for each text in T, a non-empty subset of terms in S. A term which annotates a document of T is called a **tag** of T. Moreover, given a set S of terms, a **glossary** $G_S : S \rightarrow \bar{S}$ is a function associating to each term s_i a text $\bar{s}_i$ describing its semantics ($\bar{S} = \{\bar{s}_1, \bar{s}_2, \ldots, \bar{s}_m\}$).

Given a glossary G_S, a set of texts T, and an annotation f_T, we can build a matrix $M = \{m_{ij}\}_{n \times m}$, called **tags matrix** of T, where $m_{ij} = 1$ if s_j is a tag of t_i, $m_{ij} = 0$ otherwise.

Given a similarity measure Sim defined on a glossary G_S, a text t, and a term $s' \in S \setminus f_T(t)$, we can define a correlation score C as:

$$C(s', t) = \max_{s'' \in f_T(t)} Sim(s', s''). \tag{4.16}$$

This notation can be easily extended to subsets of terms $\hat{S} \subseteq S \setminus f_T(t)$ in the following way:

$$C\left(\hat{S}, t\right) = \min_{s' \in \hat{S}} C(s', t). \tag{4.17}$$

BUILDING THE TAG MATRIX Let $T = \{t_1, t_2, \ldots, t_n\}$ be the set of input texts. For each t_i, the TAGME algorithm is applied in order to obtain a list of anchors to Wikipedia pages

($L_i \ for \ i = 1, \ldots, n$). These lists are, then, combined to obtain a set of terms $S = L_1 \cup L_2 \cup \ldots L_n = \{s_1, s_2, \ldots, s_m\}$, together with an annotation f_T, where $f_T(t_i) = L_i$. Next, we build a glossary G_S describing each term by its corresponding Wikipedia page.

The similarity among terms is defined by a suitable semantic distance among the corresponding pages. For example one could define this distance in terms of the number of common anchors generated by TAGME for each Wikipedia page.

Finally, using the set of terms S, and the annotation f_T, we can build a tags matrix $M = \{m_{i,j}\}_{n \times m}$, where $m_{i,j}$ is 1 if the term s_j is a tag of t_i, 0 otherwise.

RECOMMENDATION AND CANDIDATE TERMS MATRIX The tags matrix M built at the end of the previous step can be viewed as an user-object utility matrix used in a recommendation system, where users and objects correspond to text documents and terms respectively. For this purpose any recommendation algorithm may be used to generate new candidate terms. Our approach uses the **DT-Hybrid** algorithm as a prediction method. **DT-Hybrid** produces sets $R_i = \{(s_j, r_{ji}) \mid m_{i,j} = 0\}$, where r_{ji} is a score expressing our confidence in the association of term s_j to document t_i. **DT-Hybrid** selects only the top-L predictions for each document, where L is a user-defined parameter. This information can be used to generate a **candidate terms matrix** $M^{cand} = \left\{m_{i,j}^{cand}\right\}_{n \times m}$, where $m_{i,j}^{cand}$ is 1 if and only if the term s_j has been predicted for the document t_i.

SIMILARITY AND CORRELATION COMPUTATION In order to build the final sets of tags and the associated p-values, it is crucial to choose a suitable semantic similarity measure among terms, from which the correlation values will be computed. We achieved this by designing a similarity function which combines the TAGME algorithm with the cosine similarity. Let G_S be a glossary. We apply the TAGME algorithm to build an anchor vector V_i for each text $\bar{s}_i$ associated with the term s_i in the glossary G_S. These vectors are then used to compute the cosine similarity among all pairs of terms, where each term s_i is represented by its corresponding anchor vector

V_i. Finally, The correlation values linking each text to each recommended term can be computed by equation 4.16.

Final sets of tags and p-values computation In order to refine the knowledge extracted from the input documents, we filter the above results by providing a sort of quality degree for this collection. Our approach consists of computing subsets of terms with increasing minimum correlation. Next, for each subset, we compute the probability of obtaining by chance a subset of terms whose correlation is greater than the proposed one. Such probability is our quality degree that can be assigned to each subset. The lower is this value, the more appropriate will be said set of tags. More precisely, for each text $t_i \in T$, let M_i^{cand} be the set of its candidate terms produced by the recommendation system, and suppose that the terms are sorted by correlation $C\left(M_i^{cand}[1], t_i\right) \leq C\left(M_i^{cand}[2], t_i\right) \leq \ldots \leq C\left(M_i^{cand}[L], t_i\right)$. The subsets of terms with increasing minimum score can be generated by choosing each term from M_i^{cand} as the minimum threshold and selecting all terms that have a value greater than it, that is $M_{i,j}^{cand} = \left\{x \in M_i^{cand} \mid C(x, t_i) \geq C\left(M_i^{cand}[j], t_i\right)\right\}$. Now for each $M_{i,j}^{cand}$, using equation 4.17, we can compute a correlation score, and, with that value, a p-value $p_{i,j}$, where $p_{i,j}$ is the probability of extracting $\left|M_{i,j}^{cand}\right|$ terms whose correlation is greater than the threshold from the set of all terms S. Let $\left|M_{i,j}^{cand}\right| = k_{i,j}$ and $\left|\left\{x \in S \mid C(x, t_i) \geq C\left(M_{i,j}^{cand}, t_i\right)\right\}\right| = q_{i,j}$, the p-value can be computed by using an hypergeometric distribution as:

$$p_{i,j} = \frac{\binom{q_{i,j}}{k_{i,j}}\binom{m-q_{i,j}}{k_{i,j}-k_{i,j}}}{\binom{m}{k_{i,j}}} = \frac{\binom{q_{i,j}}{k_{i,j}}}{\binom{m}{k_{i,j}}} \tag{4.18}$$

Obtained these values, we can now find the index x_i, which maximizes the correlation, and at the same time, minimizes the probability p. The final result of the algorithm is composed by the sets M_{i,x_i}^{cand} for each text t_i, along with their correlation and probability.

Table 4.15: Description of the datasets: for each one, we specified the topic which has been searched in PubMed, the number of abstracts downloaded, and the number of terms which have found by applying the first step of our methodology.

Search Term	# of abstracts	# of terms
MicroRNA	490	1364
Adenocarcinoma	494	1475
BRAF	163	459
EGFR	493	1345
Colorectal Cancer	490	1178
Thyroid Cancer	272	840
Prostate Cancer	465	1131
Malaria	469	1250

Data sources To evaluate **BioTAGME**, we used eight data sets (table 4.15) composed of abstracts extracted from PubMed. For each one we have applied our methodology and checked the quality of the results.To build these data sets we applied the following procedure: first, we selected eight terms,and (i) for each one a search was performed in PubMed, (ii) subsequently, the results of each search were downloaded, and, (iii) for each found paper, we collected the abstracts, so as to obtain all documents which composes each data set. We choose eight terms to evaluate our method with heterogeneous texts. These terms are: "MicroRNA", "adenocarcinoma", "BRAF", and "EGFR", "Colorectal Cancer", "Thyroid Cancer", "Prostate Cancer", and "Malaria".

Results Finding new and interesting latent associations between texts and topics is a very important practice, which allows the identification of possible novel hidden knowledge. The main purpose of **BioTAGME** is to provide a tool that allows the algorithmic analysis of texts and the extraction therefrom of latent interesting associations, in order to enrich our biological insight.

To verify the results produced by our method, we focused on two fronts: First, we analyzed, using the procedure described in [31], the quality of the predictions provided by **DT-Hybrid**. This has been done in order to determine the reliability of our approach with the similarity measures we defined. Next, we have applied our methodology to find new sets of terms, that,

subsequently, have been manually analyzed to determine their biological soundness. Our analysis showed that its results are highly unexpected and surprising from a biological point of view, and this allows us to gain a novel biological insight of phenomena described in the examined texts. Here, we will focus only on a single case study regarding the results of the verification stage.

DT-Hybrid evaluation Verifying **DT-Hybrid** predictions is a critical step to understand how our method will behaves, and, above all, if it has the characteristicswe are looking for: the ability to find new sets of highly correlated terms, which, at the same time, are sufficiently innovative to provide new biological insight. In order to perform such assessment, we used the methodology proposed in [31]. More precisely, for each dataset, we applied a 10-fold cross-validation which has been repeated 30 times. For each fold, a random group of interactions between texts and terms have been removed from the tags matrix and the **DT-Hybrid** algorithm was applied. Afterwards, the four metrics (Recovery, Precision and Recall Enhancement, Personalization,and Surprisal) were computed as described in [31], and the worst values were chosen as the overall results of the experiments.

In Table 4.16, we illustrate each result: the first two metrics represent the capability of the algorithm to predict new sets of terms as much as possible related to the source ones, while the other metrics represent how innovative are those terms. For this reason, the ability of **DT-Hybrid** to achieve high values (low for Recovery) for all metrics is a strong indicator of the quality and reliability of the final results. What has been obtained, in fact, shows that the algorithm is able to find many new highly related terms for each text in the data set, although the presence of completely uncorrelated terms can not be avoided. This latter aspect is, however, crucial in order to obtain novel biological insight.

"MicroRNA" Case Study As stated earlier, the ultimate goal of our methodology is to obtain new biological insight from the input documents, so as to get new latent associations from them. For this reason we have developed a method that can do this as automatically as possible.

Table 4.16: The four metrics (Recovery, Precision and Recall Enhancement, Personalization, and Surprisal) computed for our datasets using **DT-Hybrid** . As suggested in [31], the top-30 predictions where chosen to perform this analysis.

Data set	r	$e_P\,(20)$	$e_R\,(20)$	$h\,(20)$	$I\,(20)$
MicroRNA	0.1024	30.33	13.74	0.8209	5.0114
Adenocarcinoma	0.0657	89.51	48.12	0.7790	5.1985
BRAF	0.0694	34.43	18.94	0.7434	4.1291
EGFR	0.0606	66.85	46.38	0.6813	3.7854
Colorectal Cancer	0.0630	85.35	44.69	0.7389	4.9993
Thyroid Cancer	0.0964	50.42	24.83	0.8306	5.3598
Prostate Cancer	0.0783	79.91	41.12	0.7226	4.8177
Malaria	0.0653	100.04	46.61	0.7008	4.9042

However the final analysis of each result is a complex process: it requires a manual checking of all new terms, even if the use of filtering based on correlation and p-value help us to remove all uninteresting results.

In order to illustrate the quality of our approach, we show here a case study. First of all, we built a data set by extracting 500 PubMed abstracts that have some relationship with the topic "MicroRNA". On this dataset, we applied **BioTAGME** and calculated new sets of terms. For all those terms, we then checked manually if each new term had some relation to the original texts. We show here two of them as an example of how well our method behaves. In Table 4.17, we indicate for each item the number of original terms, the number of new terms, the global correlation score, and the p-value computed by our method. In table 4.18, we present the list of source and new terms. To complete the results, on table 4.18, we associate to each new term the correlation value and a citation confirming the discovered association. Our results shows that the terms computed by our method are highly related to the topics in the paper to which they refer. Papers with such information were not present in our data set, this proves, therefore, the validity of our methodology. In addition, it is important to emphasize, that for two terms we were not able to find any experimental confirmation. This result is important because it shows that our approach is able to find possible novel associations to experiment with.

Table 4.17: General information for the two papers we chose: number of source terms, number of new terms, correlation, and p-value.

#	Citation	Source terms	New terms	Correlation	p-Value
1	Kuhlmann et al. [244]	4	5	0.90	< 0.01
2	Ren et al. [245]	5	6	0.80	< 0.01

Table 4.18: List of source and new terms of the two papers in table 4.17. For each new term, we provided the correlation value and the citation to an article that experimentally confirms the association between each new term and the main topic of the source paper.

Paper	Source Terms	New Terms		
			Correlation	Citation
1	microrna u2 spliceosomal rna small nuclear rna ovarian cancer	messenger rna	0.90	-
		non-coding rna	1.00	-
		parp inhibitor	0.91	[246]
		muc1	0.98	[247]
		xiap	0.93	[248]
			Correlation	Citation
2	guangzhou senescence e2f3 hefei mir-449 microrna	ezh2	1.00	[249]
		pdgfb	0.80	[250]
		col4a1	1.00	-
		ctgf	1.00	[251]
		sox4	1.00	-
		mcl1	1.00	[252]

4.5 Towards an intuitive pathway based simulation method: new clues and preliminary results

The techniques discussed in previous sections are useful to formulate novel hypotheses and guide the experimental process with the aim of enhancing current knowledge of patho-physiological phenomena. This would make the classification and proposition of personalized therapy much more reliable than currently possible. Despite the subsequent reduction in number of laboratory experiments that these methods can archive, the time and cost associated therewith appear mostly prohibitive.

With the aim to reduce the candidates for further experimentation, the development of an in silico simulation methodology which allows the prediction of experimental outcomes on the basis of present knowledge is necessary. Recently several simulation models have been proposed and evaluated by using biological observational data [51–55], in order to understand the dynamic behavior of biological systems by estimating changes of concentration of some elements. Such methods usually represents biological phenomena described by pathways with differential equations. Their parameters are therefore estimated by some computational method, and their ability to predict data is measured. However, their results are typically incompatible with the observational data due to missing or wrong interactions. Even when accurate enough, the evaluation of such algorithms is computationally expensive, due to high number of parameters to estimate and variables to predict. In this sense **MITHrIL** offers great untapped potential, shifting the point of view from the prediction of single gene concentration, to the prediction of biological functions activity, which are represented by the state (activated or inhibited) of their pathway endpoint. From expression data, in fact, the algorithm is able to evaluate with high precision the state of pathways which are representations of biological functions. Simulating, therefore, the deregulation obtained by the introduction of exogenous or endogenous elements in a pathway it is possible to estimate their impact on biological functions and, consequently, formulate more accurate hypotheses.

SIMPATHY (*SIMulations on PATHwaY*) is a probabilistic algorithm that exploits **MITHrIL** to infer the most likely state of pathway endpoints when an alteration of some endogenous or exogenous factors happens. To do so, random expression values are computed, and **MITHrIL** is applied repeatedly in order to obtain an **Activity Score** for each endpoint, that is an assessment of how likely it is to observe a specific state compared to a null model.

Algorithm. Let e be an endpoint of a pathway. Suppose we want to estimate its alteration when a deregulation can be observed in $n_1, \ldots, n_m$ independent nodes of its pathway. In order to evaluate such alteration, a probability of activation, $P_A(e)$, and a probability of inhibition, $P_I(e)$, could be estimated.

By assigning synthetic expression values to $n_1, \ldots, n_m$ and applying **MITHrIL** , it is possible to assess whether the endpoint e is activated or inhibited by its perturbation (positive values indicate activation, negative inhibition). By repeating this procedure N times and counting the number of times it appears activated, $N_A(e)$, and inhibited, $N_I(e)$, it is possible to estimate such probabilities as:

$$P_A(e) = \frac{N_A(e)}{N}, \tag{4.19}$$

$$P_I(e) = \frac{N_I(e)}{N}. \tag{4.20}$$

Moreover, by repeating such a procedure to a set of randomly chosen nodes it is possible to estimate an a priori probability of activation, $P_A^R(e)$, and inhibition, $P_I^R(e)$. We can now estimate an activation log-likelihood, $L_A(e)$, and an inhibition one, $L_I(e)$, as:

$$L_A(e) = \log \frac{P_A(e)}{1 - P_A(e)} - \log \frac{P_A^R(e)}{1 - P_A^R(e)}, \tag{4.21}$$

$$L_I(e) = \log \frac{P_I(e)}{1 - P_I(e)} - \log \frac{P_I^R(e)}{1 - P_I^R(e)}, \tag{4.22}$$

The **activity score**, $As(e)$, summarizes both previous values. The sign indicates the type of

more likely deregulation (negative inhibition, positive activation) while the value is indicative of the probability of such an eventuality. It can be calculated as:

$$As\left(e\right)=\begin{cases}L_A\left(e\right) & \text{if } L_A\left(e\right)>L_I\left(e\right)\\ -L_I\left(e\right) & \text{if } L_I\left(e\right)>L_A\left(e\right)\\ 0 & \text{otherwise}\end{cases}. \tag{4.23}$$

The above procedure works only when deregulated nodes are independent. If such an event does not occur, a meta-node, connected to the ones that should be simulated, can be added to the pathway. This will make it possible to apply **SIMPATHY** on that node while avoiding significant changes to the model. In order to simulate exogenous factors in the pathways we can take advantage of the enrichment procedure of **MITHrIL** to introduce them in the pathway and therefore simulate their action.

Case study: simulating Epstein-Barr viral miRNAs. **SIMPATHY** is still under development. Extensive testing to assess the accuracy and compare it with the competitors have not yet been executed. However, a preliminary validation of the quality of the results was performed trying to simulate the consequences of the introduction in human of viral miRNAs due to an infection.

Since 2004, many miRNAs have been discovered in double-stranded DNA viruses [253]. This discovery opened a new field of virology, indeed, the possibility that such molecules can simultaneously regulate hundreds of genes suggested a novel and complex mechanism of host-virus interaction. Viral miRNAs can directly alter the physiology of the host, including components of the immune system. In particular, the expression of viral miRNAs has been demonstrated for many pathogenic human herpesviruses such as Epstein-Barr virus (EBV).

As a first validation of our approach, we simulated the action of viral miRNAs expressed by EBV on the following human signaling pathways:

- **B Cell Receptor Signaling Pathway**, which is activated when antigen binds to the receptor present on the surface of a B Cell leading to proliferation and antibody production;
- **mRNA Surveillance Pathway**, a mechanism of quality control which finds and degrades abnormal mRNAs;
- **p53 Signalling Pathway**, activated by a cellular stress, which leads to programmed cell death, or apoptosis.

Our choices are linked to the idea of simulating how EBV tries to act against human cells defense mechanisms. In particular we also wanted to highlight the viral oncogenic action by looking at the p53 pathway, whose inhibition causes failure to reply to an aberrated activation of oncogenes, thus progression of cancer.

The role of EBV miRNAs in the transformation of B-Cell has been characterized for some time [254, 255]. These molecules are responsible for B Cell proliferation and repression of apoptosis [254, 255], resulting in an increased viral load [256]. Figure 4.9 illustrates graphically the results of our simulation on the **B Cell Receptor Signaling** pathway. In blue we highlight endpoints probably inhibited, in red those activated. As we can see, all paths leading to activation of the immune response are inhibited. In particular, our predicted inhibition of trascription factor *AP1* is responsible for the lack of differentiation, activation of the inflammation process, and productions of immunoglobulins. Such event, although not yet demonstrated for EBV, has already been shown in another virus, the Kaposi's sarcoma-associated herpesvirus, or KSHV [257]. Our predictions also show the inhibition of factor *NFAT* and inhibitors $IKK\alpha$ and $IKK\beta$, which lead to the inhibition of apoptosis. This conclusion is also supported by the activation of *GSK3B* kinase, which leads to uncontrolled cell duplication independent of its substrate [258, 259]. This assumption is reinforced by the results shown in Figure 4.10 where an arrest of apoptosis and a strengthening of the cell cycle can be observed, resulting in increased mitotic process [255]. Finally, our predictions in Figure 4.11 illustrates the state of protein complexes

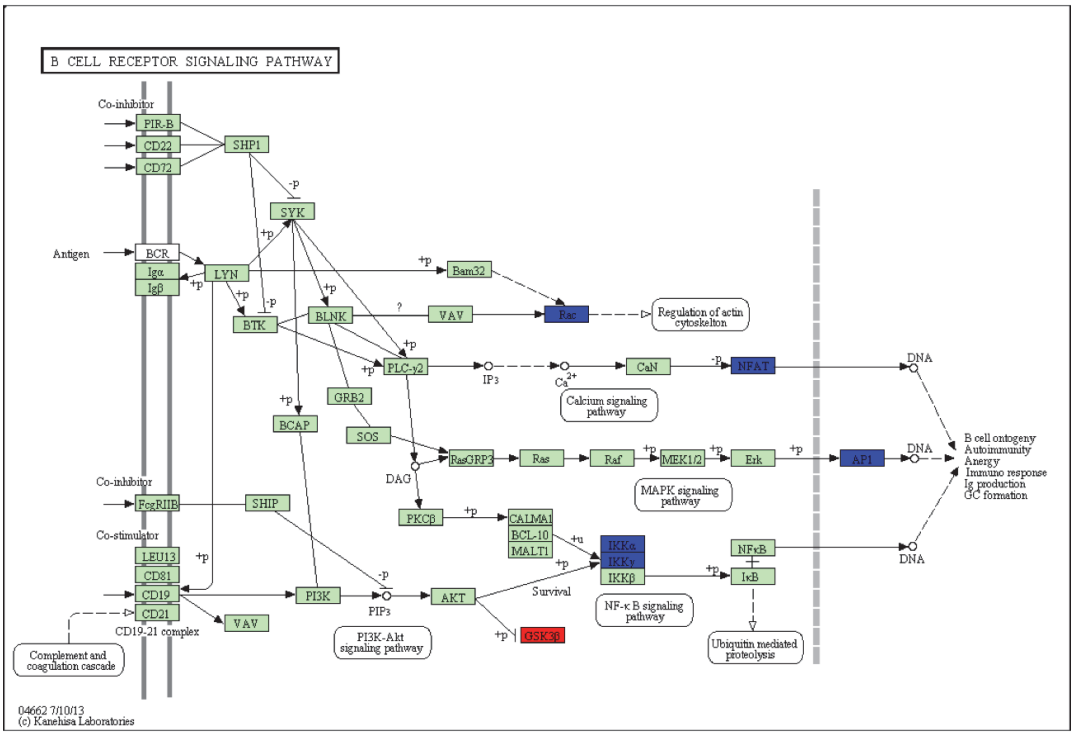


Figure 4.9: Simulating action of EBV viral miRNA on B Cell Receptor Signalling Pathway. In red we highlight positive activity score values for an endpoint, negative ones in blue. A positive (negative) score implies that after the introduction of exogenous elements we predict an activation (inhibition) of the endpoint.

that are employed to control and degrade aberrant mRNAs. An inhibition of such complexes is consistent with aim of the virus to lead to its replication. At the same time, such inhibition can cause the onset of cancer due to the formation of misfolded proteins which are not degraded by such quality control mechanisms.

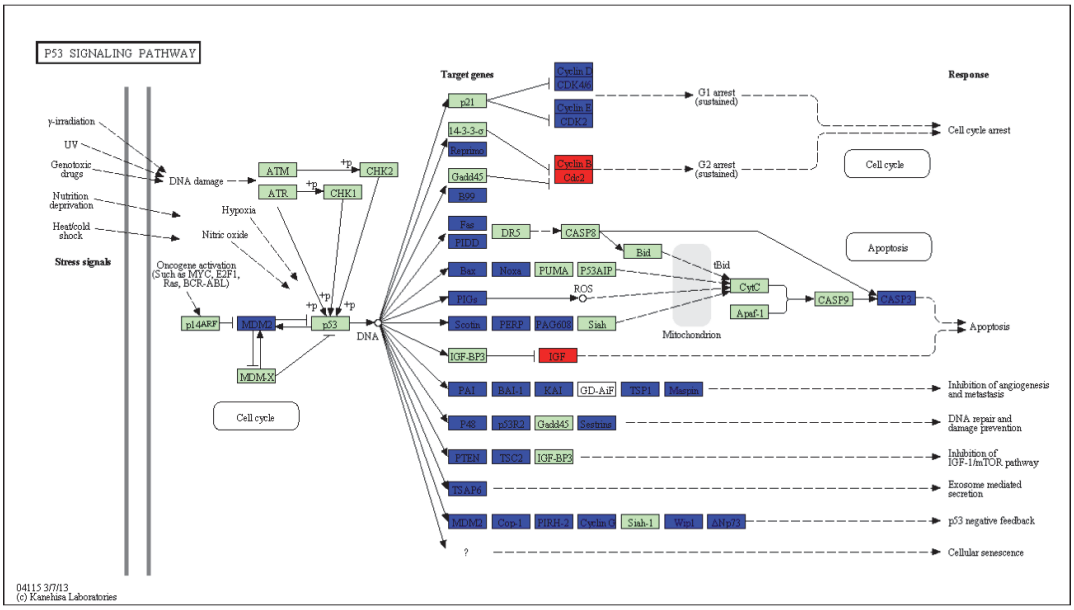


Figure 4.10: Simulating action of EBV viral miRNA on P53 Signalling Pathway. In red we highlight positive activity score values for an endpoint, negative ones in blue. A positive (negative) score implies that after the introduction of exogenous elements we predict an activation (inhibition) of the endpoint.

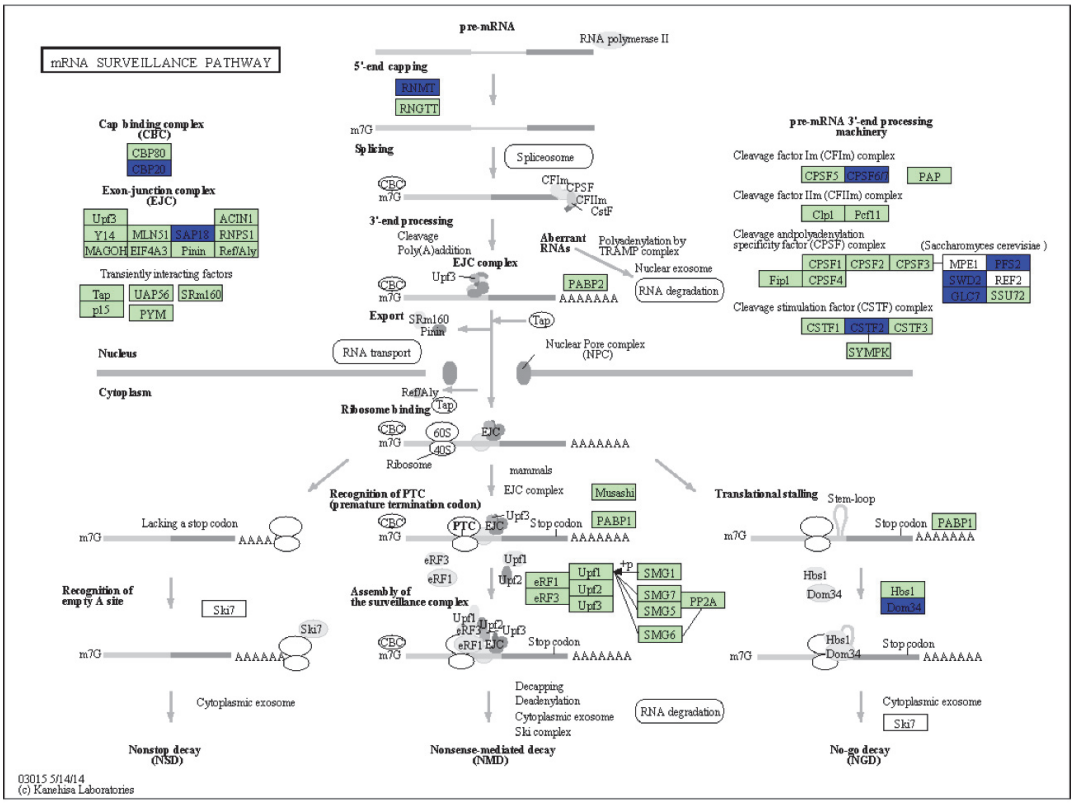


Figure 4.11: Simulating action of EBV viral miRNA on mRNA Surveillance Pathway. In red we highlight positive activity score values for an endpoint, negative ones in blue. A positive (negative) score implies that after the introduction of exogenous elements we predict an activation (inhibition) of the endpoint.

5

Conclusions

Precision medicine is the new frontier of the twenty first century. A key assumption in precision medicine is that the cause of a disease is at least partially attributable to specific genetic or epigenetic characteristics of a patient. Therefore, identifying these specificities helps building the best treatment for each individual. However, in order to develop personalized therapies, highly interdisciplinary teams are required. Such medicine depends heavily on Next-generation Sequencing techniques, and information sharing to ensure optimal results. The amount of data generated is too large to be analyzed with a classical approach, in which a physician examines such data to formulate hypotheses therefore prescribing a treatment. That is why bioinformatics has taken a prominent role in precision medicine, returning highly accurate therapies based on the unique characteristics of each patient. Naturally, this involves a significant effort to coordinate and train heterogeneous groups, providing at the same time high-precision tools.

The aim of this thesis has been the development of an integrated framework, based synergistic tools, models and algorithms, which help to fill some of the major gaps in each step of the production of highly customized therapies, overcoming, if possible, the limitations of currently employed techniques.

Classifying patients is a key step in the development of personalized therapy. It allows us to determine his unique characteristics and help diagnosis. However, the use of biomarkers is not always enough. That is why we have proposed the use of functional biomarkers determined through a pathway analysis algorithm. Despite the results demonstrate the accuracy of our technique, many factors are not yet taken into account. Future developments of our methodology should consider important genetic factors that alter the action of proteins, such as mutations and editing, and epigenetic factors such as methylation or chromatin remodeling. In this sense, some methods have already been proposed in this thesis, but an integration work is still needed to make the most out of them.

To predict possible treatments, we need to know the full functioning of drugs, and the cytotoxic response that each tissue shows, in order to predict suitable drugs combinations. In this regard, we developed several methodologies. Despite this, our cytotoxicity prediction technique still lacks precision, and our DTI prediction methodology has some important shortcomings. One major deficiency is the inability to predict any target for completely novel drugs. A variation of our tripartite recommendation technique might be used to fill this gap and provide a set of initial predictions to extend. Moreover, our drug combinations prediction methodology still does not take into account genetic or epigenetic characteristics of a patient, and is not able to exploit cytotoxicity knowledge to remove poisonous combinations. These processes are still delegated to a manual assessment, but by crossing all informations we are able to predict, we could automate such process by filtering out predictions.

Finally, our simulation technique still needs accurate tests, and improvements in the underlying pathway analysis algorithm. However, our current findings are promising, showing that

the prediction of macroscopic biological processes alterations is more reliable than predicting the concentration of single biological elements due to the many gaps in our knowledge of low level phenomena.

6
Bibliography

[1] National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease et al. *Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease*. National Academies Press (US), 2011.

[2] Xavier M Keutgen, Filippo Filicori, Michael J Crowley, Yongchun Wang, Theresa Scognamiglio, Rana Hoda, Daniel Buitrago, David Cooper, Martha A Zeiger, Rasa Zarnegar, et al. A panel of four mirnas accurately differentiates malignant from benign indeterminate thyroid lesions on fine needle aspiration. *Clinical Cancer Research*, 18(7):2032–2038, 2012.

[3] William Pao, Vincent Miller, Maureen Zakowski, Jennifer Doherty, Katerina Politi, Inderpal Sarkaria, Bhuvanesh Singh, Robert Heelan, Valerie Rusch, Lucinda Fulton, et al. Egf receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proceedings of the National Academy of Sciences of the United States of America*, 101(36):13306–13311, 2004.

[4] Edward H Romond, Edith A Perez, John Bryant, Vera J Suman, Charles E Geyer Jr, Nancy E Davidson, Elizabeth Tan-Chiu, Silvana Martino, Soonmyung Paik, Peter A Kaufman, et al. Trastuzumab plus adjuvant chemotherapy for operable her2-positive breast cancer. *New England Journal of Medicine*, 353(16):1673–1684, 2005.

[5] Gregory J Downing, Scott N Boyle, Kristin M Brinner, and Jerome A Osheroff. Information management to enable personalized medicine: stakeholder roles in building clinical decision support. *BMC medical informatics and decision making*, 9(1):44, 2009.

[6] Guy Haskin Fernald, Emidio Capriotti, Roxana Daneshjou, Konrad J Karczewski, and Russ B Altman. Bioinformatics challenges for personalized medicine. *Bioinformatics*, 27 (13):1741–1748, 2011.

[7] Richard Simon and Sameek Roychowdhury. Implementing personalized cancer genomics in clinical trials. *Nature reviews Drug discovery*, 12(5):358–369, 2013.

[8] Vincent Canuel, Bastien Rance, Paul Avillach, Patrice Degoulet, and Anita Burgun. Translational research platforms integrating clinical and omics data: a review of publicly available solutions. *Briefings in bioinformatics*, page bbu006, 2014.

[9] Andrea Sboner and Olivier Elemento. A primer on precision medicine informatics. *Briefings in bioinformatics*, page bbv032, 2015.

[10] Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*, 8(2):e1002375, 2012.

[11] Jörg Rahnenführer, Francisco S Domingues, Jochen Maydt, and Thomas Lengauer. Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

[12] Sorin Draghici, Purvesh Khatri, Adi Laurentiu Tarca, Kashyap Amin, Arina Done, Calin Voichita, Constantin Georgescu, and Roberto Romero. A systems biology approach for pathway level analysis. *Genome research*, 17(10):1537–1545, 2007.

[13] Adi Laurentiu Tarca, Sorin Draghici, Purvesh Khatri, Sonia S Hassan, Pooja Mittal, Jungsun Kim, Chong Jai Kim, Juan Pedro Kusanovic, and Roberto Romero. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82, 2009.

[14] Charles J Vaske, Stephen C Benz, J Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, 26(12):i237–i245, 2010.

[15] Andrew J Sedgewick, Stephen C Benz, Shahrooz Rabizadeh, Patrick Soon-Shiong, and Charles J Vaske. Learning subgroup-specific regulatory interactions and regulator independence with paradigm. *Bioinformatics*, 29(13):i62–i70, 2013.

[16] Cristina Mitrea, Zeinab Taghavi, Behzad Bokanizad, Samer Hanoudi, Rebecca Tagett, Michele Donato, Călin Voichiţa, and Sorin Drăghici. Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in physiology*, 4, 2013.

[17] Salvatore Alaimo, Rosalba Giugno, Mario Acunzo, Dario Veneziano, Alfredo Ferro, and Alfredo Pulvirenti. Post-transcriptional knowledge in pathway analysis increases the accuracy of phenotypes classification. *arXiv preprint arXiv:1510.08237*, 2015.

[18] Murat Iskar, Georg Zeller, Peter Blattmann, Monica Campillos, Michael Kuhn, Katarzyna H Kaminska, Heiko Runz, Anne-Claude Gavin, Rainer Pepperkok, Vera van Noort, et al. Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. *Molecular systems biology*, 9(1):662, 2013.

[19] Honglin Li, Zhenting Gao, Ling Kang, Hailei Zhang, Kun Yang, Kunqian Yu, Xiaomin Luo, Weiliang Zhu, Kaixian Chen, Jianhua Shen, et al. Tarfisdock: a web server for identifying drug targets with docking approach. *Nucleic acids research*, 34(suppl 2):W219–W224, 2006.

[20] Michael J Keiser, Bryan L Roth, Blaine N Armbruster, Paul Ernsberger, John J Irwin, and Brian K Shoichet. Relating protein pharmacology by ligand chemistry. *Nature biotechnology*, 25(2):197–206, 2007.

[21] Michael J Keiser, Vincent Setola, John J Irwin, Christian Laggner, Atheir I Abbas, Sandra J Hufeisen, Niels H Jensen, Michael B Kuijer, Roberto C Matos, Thuy B Tran, et al. Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181, 2009.

[22] Michael Kuhn, Damian Szklarczyk, Sune Pletscher-Frankild, Thomas H Blicher, Christian von Mering, Lars J Jensen, and Peer Bork. Stitch 4: integration of protein–chemical interactions with user data. *Nucleic acids research*, page gkt1207, 2013.

[23] Muhammed A Yıldırım, Kwang-Il Goh, Michael E Cusick, Albert-László Barabási, and Marc Vidal. Drug—target network. *Nature biotechnology*, 25(10):1119–1126, 2007.

[24] Sharangdhar S Phatak and Shuxing Zhang. A novel multi-modal drug repurposing approach for identification of potent ack1 inhibitors. In *Pac Symp Biocomput*, pages 29–40. World Scientific, 2013.

[25] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. Drug target identification using side-effect similarity. *Science*, 321(5886):263–266, 2008.

[26] Guangxu Jin and Stephen TC Wong. Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. *Drug discovery today*, 19(5):637–644, 2014.

[27] Humberto González-Díaz, Francisco Prado-Prado, Xerardo García-Mera, Nerea Alonso, Paula Abeijón, Olga Caamano, Matilde Yánez, Cristian R Munteanu, Alejandro Pazos, María Auxiliadora Dea-Ayuela, et al. Mind-best: Web server for drugs and target discovery; design, synthesis, and assay of mao-b inhibitors and theoretical- experimental study of g3pdh protein from trichomonas gallinae. *Journal of proteome research*, 10(4): 1698–1718, 2011.

[28] Xing Chen, Ming-Xi Liu, and Gui-Ying Yan. Drug–target interaction prediction by random walk on the heterogeneous network. *Molecular BioSystems*, 8(7):1970–1978, 2012.

[29] Jian-Ping Mei, Chee-Keong Kwoh, Peng Yang, Xiao-Li Li, and Jie Zheng. Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics*, 29(2):238–245, 2013.

[30] Feixiong Cheng, Chuang Liu, Jing Jiang, Weiqiang Lu, Weihua Li, Guixia Liu, Weixing Zhou, Jin Huang, and Yun Tang. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol*, 8(5):e1002503, 2012.

[31] Salvatore Alaimo, Alfredo Pulvirenti, Rosalba Giugno, and Alfredo Ferro. Drug–target interaction prediction through domain-tuned network-based inference. *Bioinformatics*, 29(16):2004–2008, 2013.

[32] Salvatore Alaimo, Vincenzo Bonnici, Damiano Cancemi, Alfredo Ferro, Rosalba Giugno, and Alfredo Pulvirenti. Dt-web: a web-based application for drug-target interaction and drug combination prediction through domain-tuned network-based inference. *BMC systems biology*, 9(Suppl 3):S4, 2015.

[33] Kevin C Wang and Howard Y Chang. Molecular mechanisms of long noncoding rnas. *Molecular cell*, 43(6):904–914, 2011.

[34] Kyoko L Yap, Side Li, Ana M Muñoz-Cabello, Selina Raguz, Lei Zeng, Shiraz Mujtaba, Jesús Gil, Martin J Walsh, and Ming-Ming Zhou. Molecular interplay of the noncoding rna anril and methylated histone h3 lysine 27 by polycomb cbx7 in transcriptional silencing of ink4a. *Molecular cell*, 38(5):662–674, 2010.

[35] Eric Pasmant, Audrey Sabbagh, Michel Vidaud, and Ivan Bièche. Anril, a long, noncoding rna, is an unexpected major hotspot in gwas. *The FASEB Journal*, 25(2):444–448, 2011.

[36] Christin E Burd, William R Jeck, Yan Liu, Hanna K Sanoff, Zefeng Wang, and Norman E Sharpless. Expression of linear and novel circular forms of an ink4/arf-associated noncoding rna correlates with atherosclerosis risk. *PLoS genetics*, 6(12):e1001233, 2010.

[37] Rajnish A Gupta, Nilay Shah, Kevin C Wang, Jeewon Kim, Hugo M Horlings, David J Wong, Miao-Chih Tsai, Tiffany Hung, Pedram Argani, John L Rinn, et al. Long noncoding rna hotair reprograms chromatin state to promote cancer metastasis. *Nature*, 464 (7291):1071–1076, 2010.

[38] Tim R Mercer, Marcel E Dinger, and John S Mattick. Long non-coding rnas: insights into functions. *Nature Reviews Genetics*, 10(3):155–159, 2009.

[39] Chris P Ponting, Peter L Oliver, and Wolf Reik. Evolution and functions of long noncoding rnas. *Cell*, 136(4):629–641, 2009.

[40] Jeremy E Wilusz, Hongjae Sunwoo, and David L Spector. Long noncoding rnas: functional surprises from the rna world. *Genes & development*, 23(13):1494–1504, 2009.

[41] Xiaofei Yang, Lin Gao, Xingli Guo, Xinghua Shi, Hao Wu, Fei Song, and Bingbo Wang. A network based method for analysis of lncrna-disease associations and prediction of lncrnas implicated in diseases. *PloS one*, 9(1):e87797, 2014.

[42] Salvatore Alaimo, Rosalba Giugno, and Alfredo Pulvirenti. ncpred: ncrna-disease association prediction through tripartite network-based inference. *Frontiers in bioengineering and biotechnology*, 2, 2014.

[43] Susan M Rueter, T Renee Dawson, and Ronald B Emeson. Regulation of alternative splicing by rna editing. *Nature*, 399(6731):75–80, 1999.

[44] Brenda L Bass. Rna editing by adenosine deaminases that act on rna. *Annual review of biochemistry*, 71:817, 2002.

[45] Kazuko Nishikura. Editor meets silencer: crosstalk between rna editing and rna interference. *Nature Reviews Molecular Cell Biology*, 7(12):919–931, 2006.

[46] James EC Jepson and Robert A Reenan. Rna editing in regulating gene expression in the brain. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1779(8): 459–470, 2008.

[47] Kazuko Nishikura. Functions and regulation of rna editing by adar deaminases. *Annual review of biochemistry*, 79:321, 2010.

[48] Federica Galeano, Sara Tomaselli, Franco Locatelli, and Angela Gallo. A-to-I RNA editing: The "ADAR" side of human cancer. *Seminars in Cell & Developmental Biology*, 23 (3):244–250, 2012.

[49] Sara Tomaselli, Franco Locatelli, and Angela Gallo. The RNA editing enzymes ADARs: mechanism of action and human disease. *Cell and tissue research*, 356(3):527–532, 2014.

[50] Giovanni Nigita, Salvatore Alaimo, Alfredo Ferro, Rosalba Giugno, and Alfredo Pulvirenti. Knowledge in the investigation of a-to-i rna editing signals. *Frontiers in bioengineering and biotechnology*, 3, 2015.

[51] Jiguo Cao and Hongyu Zhao. Estimating dynamic models for gene regulation networks. *Bioinformatics*, 24(14):1619–1624, 2008.

[52] Kazuyuki Nakamura, Ryo Yoshida, Masao Nagasaki, Satoru Miyano, and Tomoyuki Higuchi. Parameter estimation of in silico biological pathways with particle filtering towards a petascale computing. In *Pacific Symposium on Biocomputing*, volume 14, pages 227–238. World Scientific, 2009.

[53] Shinya Tasaki, Masao Nagasaki, Masaaki Oyama, Hiroko Hata, Kazuko Ueno, Ryo Yoshida, Tomoyuki Higuchi, Sumio Sugano, and Satoru Miyano. Modeling and estimation of dynamic egfr pathway by data assimilation approach using time series proteomic data. *Genome Informatics*, 17(2):226–238, 2006.

[54] Lars Kuepfer, Matthias Peter, Uwe Sauer, and Jörg Stelling. Ensemble modeling for analysis of cell signaling dynamics. *Nature biotechnology*, 25(9):1001–1006, 2007.

[55] Takanori Hasegawa, Masao Nagasaki, Rui Yamaguchi, Seiya Imoto, and Satoru Miyano. An efficient method of exploring simulation models by assimilating literature and biological observational data. *Biosystems*, 121:54–66, 2014.

[56] Federica Eduati, Lara M Mangravite, Tao Wang, Hao Tang, J Christopher Bare, Ruili Huang, Thea Norman, Mike Kellen, Michael P Menden, Jichen Yang, et al. Prediction of human population responses to toxic compounds by a collaborative competition. *Nature biotechnology*, 2015.

[57] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[58] Aidan Lyon. Why are Normal Distributions Normal? *The British Journal for the Philosophy of Science*, 65(3):621–649, 2014.

[59] Jagdish K Patel and Campbell B Read. *Handbook of the normal distribution*, volume 150. CRC Press, 1996.

[60] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.

[61] Xiaofei Yang, Lin Gao, Xingli Guo, Xinghua Shi, Hao Wu, Fei Song, and Bingbo Wang. A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. *PloS one*, 9(1):e87797, 2014.

[62] Joseph A Konstan, Bradley N Miller, David Maltz, Jonathan L Herlocker, Lee R Gordon, and John Riedl. GroupLens: applying collaborative filtering to Usenet news. *Communications of the ACM*, 40(3):77–87, 1997.

[63] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[64] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010.

[65] John A Swets. Information retrieval systems. *Science*, 141(3577):245–250, 1963.

[66] Myron Tribus. *Thermostatics and thermodynamics*. Center for Advanced Engineering Study, Massachusetts Institute of Technology, 1961.

[67] Paolo Ferragina and Ugo Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628. ACM, 2010.

[68] Paolo Ferragina and Ugo Scaiella. Fast and accurate annotation of short texts with wikipedia pages. *Software, IEEE*, 29(1):70–75, 2012.

[69] Wikimedia. A typical animal cell with labeled organelles., 2012. URL `https://upload.wikimedia.org/wikipedia/commons/e/e2/Eukaryote_DNA-en.svg`. This file is licensed under the Creative Commons Attribution-Share Alike 3.0.

[70] Francis H Crick. On protein synthesis. In *Symposia of the Society for Experimental Biology*, volume 12, page 138, 1958.

[71] Francis Crick et al. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[72] Sarah Leavitt, DeWitt Stetten Jr, et al. *Deciphering the genetic code: Marshall Nirenberg*. Office of NIH History, 2004.

[73] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[74] Wikimedia. Dna in eukaryote cell, 2012. URL https://upload.wikimedia.org/wikipedia/commons/e/e2/Eukaryote_DNA-en.svg. This file is licensed under the Creative Commons Attribution-Share Alike 3.0.

[75] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[76] Wikimedia. The structure of dna showing with detail the structure of the four bases, adenine, cytosine, guanine and thymine, and the location of the major and minor groove., 2011. URL https://upload.wikimedia.org/wikipedia/commons/4/4c/DNA_Structure%2BKey%2BLabelled.pn_NoBB.png. This file is licensed under the Creative Commons Attribution-Share Alike 3.0.

[77] Andreas Bolzer, Gregor Kreth, Irina Solovei, Daniela Koehler, Kaan Saracoglu, Christine Fauth, S Muller, Roland Eils, Christoph Cremer, Michael R Speicher, et al. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol*, 3(5):e157, 2005.

[78] Ignacio Tinoco and Carlos Bustamante. How rna folds. *Journal of molecular biology*, 293 (2):271–281, 1999.

[79] Paul G Higgs. Rna secondary structure: physical and computational aspects. *Quarterly reviews of biophysics*, 33(03):199–253, 2000.

[80] Wikimedia. A hairpin loop from a pre-mrna., 2009. URL https://upload.wikimedia.org/wikipedia/commons/a/a4/Pre-mRNA-1ysv-tubes.png. This file is licensed under the Creative Commons Attribution-Share Alike 3.0.

[81] Jonathan Slack. *Genes: A Very Short Introduction*. Oxford University Press, 2014.

[82] Harvey F Lodish, Arnold Berk, S Lawrence Zipursky, Paul Matsudaira, David Baltimore, James Darnell, et al. *Molecular cell biology*, volume 4. Citeseer, 2000.

[83] Wikimedia. A gene is a segment of dna that encodes function, 2015. URL https://upload.wikimedia.org/wikipedia/commons/2/2b/Chromosome_DNA_Gene.svg. This file is licensed under the Creative Commons Attribution-Share Alike 3.0.

[84] International Human Genome Sequencing Consortium et al. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[85] Robert W Holley, Jean Apgar, George A Everett, James T Madison, Mark Marquisee, Susan H Merrill, John Robert Penswick, and Ada Zamir. Structure of a ribonucleic acid. *Science*, 147(3664):1462–1465, 1965.

[86] Victor Ambros. The functions of animal micrornas. *Nature*, 431(7006):350–355, 2004.

[87] David P Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *cell*, 116 (2):281–297, 2004.

[88] David P Bartel. Micrornas: target recognition and regulatory functions. *Cell*, 136(2): 215–233, 2009.

[89] Marc Robert Fabian, Nahum Sonenberg, and Witold Filipowicz. Regulation of mrna translation and stability by micrornas. *Annual review of biochemistry*, 79:351–379, 2010.

[90] Wikimedia. Diagram of microrna, 2012. URL https://upload.wikimedia.org/wikipedia/commons/a/a7/MiRNA.svg. This file is licensed under the Creative Commons Attribution-Share Alike 3.0.

[91] Marek Mraz, Dasa Dolezalova, Karla Plevova, Katerina Stano Kozubik, Veronika Mayerova, Katerina Cerna, Katerina Musilova, Boris Tichy, Sarka Pavlova, Marek Borsky, et al. Microrna-650 expression is influenced by immunoglobulin gene rearrangement and affects the biology of chronic lymphocytic leukemia. *Blood*, 119(9):2110–2113, 2012.

[92] Jinju Han, Yoontae Lee, Kyu-Hyun Yeom, Young-Kook Kim, Hua Jin, and V Narry Kim. The drosha-dgcr8 complex in primary microrna processing. *Genes & development*, 18(24): 3016–3027, 2004.

[93] Elizabeth P Murchison and Gregory J Hannon. mirnas on the move: mirna biogenesis and the rnai machinery. *Current opinion in cell biology*, 16(3):223–229, 2004.

[94] E Lund and JE Dahlberg. Substrate selectivity of exportin 5 and dicer in the biogenesis of micrornas. In *Cold Spring Harbor symposia on quantitative biology*, volume 71, pages 59–66. Cold Spring Harbor Laboratory Press, 2006.

[95] Tariq M Rana. Illuminating the silence: understanding the structure and function of small rnas. *Nature reviews Molecular cell biology*, 8(1):23–36, 2007.

[96] Ashley J Pratt and Ian J MacRae. The rna-induced silencing complex: a versatile gene-silencing machine. *Journal of Biological Chemistry*, 284(27):17897–17901, 2009.

[97] Wikimedia. Overview of microrna processing in animals., 2009. URL https://upload.wikimedia.org/wikipedia/commons/9/95/MiRNA-biogenesis.jpg. This file is licensed under the Creative Commons Attribution-Share Alike 3.0.

[98] Michael S Paul and Brenda L Bass. Inosine exists in mrna at tissue-specific levels and is most abundant in brain mrna. *The EMBO journal*, 17(4):1120–1127, 1998.

[99] Lily Bazak, Erez Y Levanon, and Eli Eisenberg. Genome-wide analysis of alu editability. *Nucleic acids research*, 42(11):6876–6884, 2014.

[100] Glen M Borchert, Brian L Gilmore, Ryan M Spengler, Yi Xing, William Lanier, Debashish Bhattacharya, and Beverly L Davidson. Adenosine deamination in human transcripts generates novel microrna binding sites. *Human molecular genetics*, 18(24):4801–4807, 2009.

[101] Yukio Kawahara, Molly Megraw, Edward Kreider, Hisashi Iizasa, Louis Valente, Artemis G Hatzigeorgiou, and Kazuko Nishikura. Frequency and fate of microrna editing in human brain. *Nucleic acids research*, 36(16):5270–5280, 2008.

[102] Yukio Kawahara. Quantification of adenosine-to-inosine editing of micrornas using a conventional method. *Nature protocols*, 7(7):1426–1437, 2012.

[103] Sheetal A Mitra, Anirban P Mitra, and Timothy J Triche. A central role for long non-coding rna in cancer. *Genomic "dark matter": implications for understanding human disease mechanisms, diagnostics, and cures*, page 70, 2012.

[104] AAH Su and L Randau. A-to-i and c-to-u editing within transfer rnas. *Biochemistry (Moscow)*, 76(8):932–937, 2011.

[105] Shai Carmi, Itamar Borukhov, and Erez Y Levanon. Identification of widespread ultra-edited human rnas. *PLoS Genet*, 7(10):e1002317–e1002317, 2011.

[106] Helene Wahlstedt and Marie Öhman. Site-selective versus promiscuous a-to-i editing. *Wiley Interdisciplinary Reviews: RNA*, 2(6):761–771, 2011.

[107] David S Moore. *The Developing Genome: An Introduction to Behavioral Epigenetics*. Oxford University Press, 2015.

[108] Khursheed Iqbal, Seung-Gi Jin, Gerd P Pfeifer, and Piroska E Szabó. Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *Proceedings of the National Academy of Sciences*, 108(9):3642–3647, 2011.

[109] Mark Wossidlo, Toshinobu Nakamura, Konstantin Lepikhov, C Joana Marques, Valeri Zakhartchenko, Michele Boiani, Julia Arand, Toru Nakano, Wolf Reik, and Jörn Walter. 5-hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nature communications*, 2:241, 2011.

[110] Rudolf Jaenisch and Adrian Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics*, 33:245–254, 2003.

[111] National Institutes of Health. Epigenetic mechanisms, 2005. URL http://commonfund.nih.gov/epigenomics/figure. This file is licensed under the Creative Commons Attribution-Share Alike 3.0.

[112] RAY WU. Nucleotide sequence analysis of dna. *Nature*, 236(68):198–200, 1972.

[113] Ernest Jay, Robert Bambara, R Padmanabhan, and Ray Wu. Dna sequence analysis: a general, simple and rapid method for sequencing large oligodeoxyribonucleotide fragments by mapping. *Nucleic Acids Research*, 1(3):331–354, 1974.

[114] R Padmanabhan, Ernest Jay, and Ray Wu. Chemical synthesis of a primer and its use in the sequence analysis of the lysozyme gene of bacteriophage t4. *Proceedings of the National Academy of Sciences*, 71(6):2510–2514, 1974.

[115] Frederick Sanger, Steven Nicklen, and Alan R Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.

[116] Wikimedia. The sanger (chain-termination) method for dna sequencing, 2012. URL https://upload.wikimedia.org/wikipedia/commons/b/b2/Sanger-sequencing.svg. This file is licensed under the Creative Commons Attribution-Share Alike 3.0.

[117] Claudia Knief. Analysis of plant microbe interactions in the era of next generation sequencing technologies. *Frontiers in plant science*, 5, 2014.

[118] Weihua Zeng and Ali Mortazavi. Technical considerations for functional sequencing assays. *Nature immunology*, 13(9):802–807, 2012.

[119] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[120] Ilseung Cho, Shingo Yamanishi, Laura Cox, Barbara A Methé, Jiri Zavadil, Kelvin Li, Zhan Gao, Douglas Mahana, Kartik Raju, Isabel Teitler, et al. Antibiotics in early life alter the murine colonic microbiome and adiposity. *Nature*, 488(7413):621–626, 2012.

[121] Gene Ontology Consortium et al. The gene ontology project in 2008. *Nucleic acids research*, 36(suppl 1):D440–D444, 2008.

[122] Jennifer I Clark, Cath Brooksbank, and Jane Lomax. It's all GO for plant scientists. *Plant physiology*, 138(3):1268–1279, 2005.

[123] Ron Shamir, Adi Maron-Katz, Amos Tanay, Chaim Linhart, Israel Steinfeld, Roded Sharan, Yosef Shiloh, and Ran Elkon. EXPANDER–an integrative program suite for microarray data analysis. *BMC bioinformatics*, 6(1):232, 2005.

[124] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[125] Beena Vallanat, Steven P Anderson, Holly M Brown-Borg, Hongzu Ren, Sander Kersten, Sudhakar Jonnalagadda, Rajagopalan Srinivasan, and J Christopher Corton. Analysis of the heat shock response in mouse liver reveals transcriptional dependence on the nuclear receptor peroxisome proliferator-activated receptor α (pparα). *BMC genomics*, 11(1):16, 2010.

[126] Jay Shendure and Hanlee Ji. Next-generation dna sequencing. *Nature biotechnology*, 26 (10):1135–1145, 2008.

[127] Peter JA Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic acids research*, 38(6):1767–1771, 2010.

[128] Cristian Del Fabbro, Simone Scalabrin, Michele Morgante, and Federico M. Giorgi. An extensive evaluation of read trimming effects on illumina ngs data analysis. *PLoS ONE*, 8(12):e85024, 12 2013. doi: 10.1371/journal.pone.0085024. URL http://dx.doi.org/10.1371%2Fjournal.pone.0085024.

[129] Stinus Lindgreen. Adapterremoval: easy cleaning of next-generation sequencing reads. *BMC research notes*, 5(1):337, 2012.

[130] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):pp–10, 2011.

[131] NA Joshi and JN Fass. Sickle: A sliding-window, adaptive, quality-based trimming tool for fastq files (version 1.33)[software], 2011.

[132] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, page btu170, 2014.

[133] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[134] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[135] Hugh L Eaves and Yuan Gao. Mom: maximum oligonucleotide mapping. *Bioinformatics*, 25(7):969–970, 2009.

[136] Hui Jiang and Wing Hung Wong. Seqmap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics*, 24(20):2395–2396, 2008.

[137] Ruiqiang Li, Yingrui Li, Karsten Kristiansen, and Jun Wang. Soap: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, 2008.

[138] Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[139] Jan O Korbel, Alexander Eckehart Urban, Jason P Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M Kim, Dean Palejev, Nicholas J Carriero, Lei Du, et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420–426, 2007.

[140] Andreas Gogol-Döring and Wei Chen. An overview of the analysis of next generation sequencing data. In *Next Generation Microarray Bioinformatics*, pages 249–257. Springer, 2012.

[141] Andrew L Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, 4(11):682–690, 2008.

[142] Francesco Iorio, Roberta Bosotti, Emanuela Scacheri, Vincenzo Belcastro, Pratibha Mithbaokar, Rosa Ferriero, Loredana Murino, Roberto Tagliaferri, Nicola Brunetti-Pierri, Antonella Isacchi, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences*, 107(33):14621–14626, 2010.

[143] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232–i240, 2008.

[144] Tao Zhou, Jie Ren, Matúš Medo, and Yi-Cheng Zhang. Bipartite network projection and personal recommendation. *Physical Review E*, 76(4):046115, 2007.

[145] Twan van Laarhoven, Sander B Nabuurs, and Elena Marchiori. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics*, 27(21):3036–3043, 2011.

[146] Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *science*, 313(5795):1929–1935, 2006.

[147] I Amelio, M Gostev, RA Knight, AE Willis, G Melino, and AV Antonov. Drugsurv: a resource for repositioning of approved and experimental drugs in oncology based on patient survival information. *Cell death & disease*, 5(2):e1051, 2014.

[148] Joel T Dudley, Marina Sirota, Mohan Shenoy, Reetesh K Pai, Silke Roedder, Annie P Chiang, Alex A Morgan, Minnie M Sarwal, Pankaj Jay Pasricha, and Atul J Butte. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Science translational medicine*, 3(96):96ra76–96ra76, 2011.

[149] Marina Sirota, Joel T Dudley, Jeewon Kim, Annie P Chiang, Alex A Morgan, Alejandro Sweet-Cordero, Julien Sage, and Atul J Butte. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science translational medicine*, 3(96):96ra77–96ra77, 2011.

[150] Jiao Li and Zhiyong Lu. A new method for computational drug repositioning using drug pairwise similarity. *Proceedings (IEEE Int Conf Bioinformatics Biomed)*, 2012:1–4, 2012. ISSN 2156-1125. doi: 10.1109/BIBM.2012.6392722. 25264495[pmid].

[151] Jiao Li and Zhiyong Lu. Pathway-based drug repositioning using causal inference. *BMC bioinformatics*, 14(Suppl 16):S3, 2013.

[152] Yong Li and Pankaj Agarwal. A pathway-based view of human diseases and disease relationships. *PloS one*, 4(2):e4346, 2009.

[153] Jing Tang, Leena Karhinen, Tao Xu, Agnieszka Szwajda, Bhagwan Yadav, Krister Wennerberg, and Tero Aittokallio. Target inhibition networks: predicting selective combinations of druggable targets to block cancer survival pathways. *PLoS computational biology*, 9(9): e1003226, 2013.

[154] Jing Ma, Xun Zhang, Choong Yong Ung, Yu Zong Chen, and Baowen Li. Metabolic network analysis revealed distinct routes of deletion effects between essential and non-essential genes. *Molecular BioSystems*, 8(4):1179–1186, 2012.

[155] Aditya Barve, João Frederico Matias Rodrigues, and Andreas Wagner. Superessential reactions in metabolic networks. *Proceedings of the National Academy of Sciences*, 109(18): E1121–E1130, 2012.

[156] Iwei Yeh, Theodor Hanekamp, Sophia Tsoka, Peter D Karp, and Russ B Altman. Computational analysis of plasmodium falciparum metabolism: organizing genomic information to facilitate drug discovery. *Genome research*, 14(5):917–924, 2004.

[157] Shailza Singh, Balwant Kishen Malik, and Durlabh Kumar Sharma. Choke point analysis of metabolic pathways in e. histolytica: a computational approach for drug target identification. *Bioinformation*, 2(2):68, 2007.

[158] Deepak Perumal, Chu Sing Lim, and Meena K Sakharkar. A comparative study of metabolic network topology between a pathogenic and a non-pathogenic bacterium for potential drug target identification. *Summit on translational bioinformatics*, 2009:100, 2009.

[159] Segun Fatumo, Kitiporn Plaimas, Jan-Philipp Mallm, Gunnar Schramm, Ezekiel Adebiyi, Marcus Oswald, Roland Eils, and Rainer König. Estimating novel potential drug targets of plasmodium falciparum by analysing the metabolic network of knock-out strains in silico. *Infection, Genetics and Evolution*, 9(3):351–358, 2009.

[160] Grant R Zimmermann, Joseph Lehar, and Curtis T Keith. Multi-target therapeutics: when the whole is greater than the sum of the parts. *Drug discovery today*, 12(1):34–42, 2007.

[161] Rocco Savino, Sergio Paduano, Mariaimmacolata Preianò, and Rosa Terracciano. The proteomics big challenge for biomarkers and new drug-targets discovery. *International journal of molecular sciences*, 13(11):13926–13948, 2012.

[162] Karen Bush, Patrice Courvalin, Gautam Dantas, Julian Davies, Barry Eisenstein, Pentti Huovinen, George A Jacoby, Roy Kishony, Barry N Kreiswirth, Elizabeth Kutter, et al. Tackling antibiotic resistance. *Nature Reviews Microbiology*, 9(12):894–896, 2011.

[163] Hiroaki Kitano. Biological robustness. *Nature Reviews Genetics*, 5(11):826–837, 2004.

[164] Jeremy S Logue and Deborah K Morrison. Complexity in the signaling network: insights from the use of targeted inhibitors in cancer therapy. *Genes & development*, 26(7):641–650, 2012.

[165] Ruth Nussinov, Chung-Jung Tsai, and Carla Mattos. 'pathway drug cocktail': targeting ras signaling based on structural pathways. *Trends in molecular medicine*, 19(11):695–704, 2013.

[166] Genevieve Holzapfel, Greg Buhrman, and Carla Mattos. Shift in the equilibrium between on and off states of the allosteric switch in ras-gppnhp affected by small molecules and bulk solvent composition. *Biochemistry*, 51(31):6114–6126, 2012.

[167] Jan van der Greef and Robert N McBurney. Rescuing drug discovery: in vivo systems pathology and systems pharmacology. *Nature Reviews Drug Discovery*, 4(12):961–967, 2005.

[168] Benjamin Haibe-Kains, Nehme El-Hachem, Nicolai Juul Birkbak, Andrew C Jin, Andrew H Beck, Hugo JWL Aerts, and John Quackenbush. Inconsistency in large pharmacogenomic studies. *Nature*, 504(7480):389–393, 2013.

[169] Orly Wapinski and Howard Y Chang. Long noncoding rnas and human disease. *Trends in cell biology*, 21(6):354–361, 2011.

[170] Geng Chen, Ziyun Wang, Dongqing Wang, Chengxiang Qiu, Mingxi Liu, Xing Chen, Qipeng Zhang, Guiying Yan, and Qinghua Cui. Lncrnadisease: a database for long-noncoding rna-associated diseases. *Nucleic acids research*, 41(D1):D983–D986, 2013.

[171] Zhifen Liu, Xinrong Li, Ning Sun, Yong Xu, Yaqin Meng, Chunxia Yang, Yanfang Wang, and Kerang Zhang. Microarray profiling and co-expression network analysis of circulating lncrnas and mrnas associated with major depressive disorder. *PloS one*, 9(3):e93388, 2014.

[172] Jie Sun, Hongbo Shi, Zhenzhen Wang, Changjian Zhang, Lin Liu, Letian Wang, Weiwei He, Dapeng Hao, Shulin Liu, and Meng Zhou. Inferring novel lncrna–disease associations based on a random walk model of a lncrna functional similarity network. *Molecular BioSystems*, 10(8):2074–2081, 2014.

[173] Gamage Upeksha Ganegoda, Min Li, Weiping Wang, and Qilong Feng. Heterogeneous network model to infer human disease-long intergenic non-coding rna associations. *NanoBioscience, IEEE Transactions on*, 14(2):175–183, 2015.

[174] Xing Chen. Predicting lncrna-disease associations and constructing lncrna functional similarity network based on the information of mirna. *Scientific reports*, 5, 2015.

[175] Yang Li, Chengxiang Qiu, Jian Tu, Bin Geng, Jichun Yang, Tianzi Jiang, and Qinghua Cui. Hmdd v2. 0: a database for experimentally supported human microrna and disease associations. *Nucleic acids research*, 42(D1):D1070–D1074, 2014.

[176] Jun-Hao Li, Shun Liu, Hui Zhou, Liang-Hu Qu, and Jian-Hua Yang. starbase v2. 0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data. *Nucleic acids research*, page gkt1248, 2013.

[177] Galina V Glazko and Frank Emmert-Streib. Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. *Bioinformatics*, 25(18): 2348–2354, 2009.

[178] ML Green and PD Karp. The outcomes of pathway database computations depend on pathway ontology. *Nucleic Acids Research*, 34(13):3687–3697, 2006.

[179] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[180] Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: back to metabolism in kegg. *Nucleic acids research*, 42(D1):D199–D205, 2014.

[181] Ethan G Cerami, Benjamin E Gross, Emek Demir, Igor Rodchenkov, Özgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D Bader, and Chris Sander. Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, 39(suppl 1):D685–D690, 2011.

[182] Purvesh Khatri, Sorin Draghici, G Charles Ostermeier, and Stephen A Krawetz. Profiling gene expression using onto-express. *Genomics*, 79(2):266–270, 2002.

[183] Sorin Draghici, Purvesh Khatri, Rui P Martins, G Charles Ostermeier, and Stephen A Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, 2003.

[184] Gabriel F Berriz, Oliver D King, Barbara Bryant, Chris Sander, and Frederick P Roth. Characterizing gene sets with funcassociate. *Bioinformatics*, 19(18):2502–2504, 2003.

[185] Tim Beißbarth and Terence P Speed. Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004.

[186] Cristian I Castillo-Davis and Daniel L Hartl. Genemerge–post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, 19(7):891–892, 2003.

[187] David Martin, Christine Brun, Elisabeth Remy, Pierre Mouren, Denis Thieffry, and Bernard Jacq. Gotoolbox: functional analysis of gene datasets based on gene ontology. *Genome biology*, 5(12):R101, 2004.

[188] Scott W Doniger, Nathan Salomonis, Kam D Dahlquist, Karen Vranizan, Steven C Lawlor, Bruce R Conklin, et al. Mappfinder: using gene ontology and genmapp to create a global gene-expression profile from microarray data. *Genome biol*, 4(1):R7, 2003.

[189] Sek Won Kong, William T Pu, and Peter J Park. A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, 22(19):2373–2380, 2006.

[190] Lu Tian, Steven A Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S Kohane, and Peter J Park. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38): 13544–13549, 2005.

[191] Zhen Jiang and Robert Gentleman. Extensions to gene set enrichment. *Bioinformatics*, 23(3):306–313, 2007.

[192] Yan Lu, Peng-Yuan Liu, Peng Xiao, and Hong-Wen Deng. Hotelling's t2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics*, 21(14): 3105–3113, 2005.

[193] Hao Xiong. Non-linear tests for identifying differentially expressed genes or genetic networks. *Bioinformatics*, 22(8):919–923, 2006.

[194] Manuela Hummel, Reinhard Meister, and Ulrich Mansmann. Globalancova: exploration and assessment of gene group effects. *Bioinformatics*, 24(1):78–85, 2008.

[195] Lev Klebanov, Galina Glazko, Peter Salzman, Andrei Yakovlev, and Yuanhui Xiao. A multivariate extension of the gene set enrichment analysis. *Journal of bioinformatics and computational biology*, 5(05):1139–1153, 2007.

[196] Jelle J Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007.

[197] Marit Ackermann and Korbinian Strimmer. A general modular framework for gene set enrichment analysis. *BMC bioinformatics*, 10(1):47, 2009.

[198] Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes. *The annals of applied statistics*, pages 107–129, 2007.

[199] Florian Martin, Alain Sewer, Marja Talikka, Yang Xiang, Julia Hoeng, and Manuel C Peitsch. Quantification of biological network perturbations for mechanistic insight and diagnostics using two-layer causal models. *BMC bioinformatics*, 15(1):238, 2014.

[200] Dmitry M Vasilyev, Ty M Thomson, Brian P Frushour, Florian Martin, and Alain Sewer. An algorithm for score aggregation over causal biological networks based on random walk sampling. *BMC research notes*, 7(1):516, 2014.

[201] Sheng-Da Hsu, Feng-Mao Lin, Wei-Yun Wu, Chao Liang, Wei-Chih Huang, Wen-Ling Chan, Wen-Ting Tsai, Goun-Zhou Chen, Chia-Jung Lee, Chih-Min Chiu, et al. mirtarbase: a database curates experimentally validated microrna–target interactions. *Nucleic acids research*, page gkq1107, 2010.

[202] Sheng-Da Hsu, Yu-Ting Tseng, Sirjana Shrestha, Yu-Ling Lin, Anas Khaleel, Chih-Hung Chou, Chao-Fang Chu, Hsi-Yuan Huang, Ching-Min Lin, Shu-Yi Ho, et al. mirtarbase update 2014: an information resource for experimentally validated mirna-target interactions. *Nucleic acids research*, 42(D1):D78–D85, 2014.

[203] Feifei Xiao, Zhixiang Zuo, Guoshuai Cai, Shuli Kang, Xiaolian Gao, and Tongbin Li. mirecords: an integrated resource for microrna–target interactions. *Nucleic acids research*, 37(suppl 1):D105–D110, 2009.

[204] Juan Wang, Ming Lu, Chengxiang Qiu, and Qinghua Cui. Transmir: a transcription factor–microrna regulation database. *Nucleic acids research*, 38(suppl 1):D119–D122, 2010.

[205] Sam Griffiths-Jones. The microrna registry. *Nucleic acids research*, 32(suppl 1):D109–D111, 2004.

[206] Sam Griffiths-Jones, Russell J Grocock, Stijn Van Dongen, Alex Bateman, and Anton J Enright. mirbase: microrna sequences, targets and gene nomenclature. *Nucleic acids research*, 34(suppl 1):D140–D144, 2006.

[207] Sam Griffiths-Jones, Harpreet Kaur Saini, Stijn van Dongen, and Anton J Enright. mirbase: tools for microrna genomics. *Nucleic acids research*, 36(suppl 1):D154–D158, 2008.

[208] Ana Kozomara and Sam Griffiths-Jones. mirbase: integrating microrna annotation and deep-sequencing data. *Nucleic acids research*, page gkq1027, 2010.

[209] Ana Kozomara and Sam Griffiths-Jones. mirbase: annotating high confidence micrornas using deep sequencing data. *Nucleic acids research*, page gkt1181, 2013.

[210] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, Clifford Stein, et al. *Introduction to algorithms*, volume 2. MIT press Cambridge, 2001.

[211] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[212] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, page gkv007, 2015.

[213] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. Voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biol*, 15(2):R29, 2014.

[214] T.T. Ashburn and K.B. Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8):673–683, 2004.

[215] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[216] Temple F Smith and Michael S Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[217] Masahiro Hattori, Yasushi Okuno, Susumu Goto, and Minoru Kanehisa. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *Journal of the American Chemical Society*, 125(39):11853–11865, 2003.

[218] Alex A Adjei. Blocking oncogenic ras signaling for cancer therapy. *Journal of the National Cancer Institute*, 93(14):1062–1074, 2001.

[219] Yanbin Liu, Bin Hu, Chengxin Fu, and Xin Chen. Dcdb: drug combination database. *Bioinformatics*, 26(4):587–588, 2010.

[220] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl 1):D668–D672, 2006.

[221] David S Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, 36(suppl 1):D901–D906, 2008.

[222] Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, et al. Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, 39(suppl 1):D1035–D1041, 2011.

[223] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 42(D1):D472–D477, 2014.

[224] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. Pid: the pathway interaction database. *Nucleic acids research*, 37(suppl 1):D674–D679, 2009.

[225] Emek Demir, Michael P Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming Wu, Peter D'Eustachio, Carl Schaefer, Joanne Luciano, et al. The biopax community standard for pathway data sharing. *Nature biotechnology*, 28(9):935–942, 2010.

[226] Vasek Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of operations research*, 4(3):233–235, 1979.

[227] Robert S Harris, Olga Lazar, Jay W Johansen, and Peter S Sebel. Interaction of propofol and sevoflurane on loss of consciousness and movement to skin incision during general anesthesia. *Anesthesiology*, 104(6):1170–1175, 2006.

[228] Erwin Sigel. Mapping of the benzodiazepine recognition site on gaba-a receptors. *Current topics in medicinal chemistry*, 2(8):833–839, 2002.

[229] Jia Jia, Feng Zhu, Xiaohua Ma, Zhiwei W Cao, Yixue X Li, and Yu Zong Chen. Mechanisms of drug combinations: interaction and network perspectives. *Nature Reviews Drug Discovery*, 8(2):111–128, 2009.

[230] Anna Bauer-Mehren, Michael Rautschka, Ferran Sanz, and Laura I Furlong. Disgenet: a cytoscape plugin to visualize, integrate, search and analyze gene–disease networks. *Bioinformatics*, 26(22):2924–2926, 2010.

[231] Aleksandra Helwak, Grzegorz Kudla, Tatiana Dudnakova, and David Tollervey. Mapping the human mirna interactome by clash reveals frequent noncanonical binding. *Cell*, 153 (3):654–665, 2013.

[232] Grzegorz Kudla, Sander Granneman, Daniela Hahn, Jean D Beggs, and David Tollervey. Cross-linking, ligation, and sequencing of hybrids reveals rna–rna interactions in yeast. *Proceedings of the National Academy of Sciences*, 108(24):10010–10015, 2011.

[233] Qinghua Jiang, Yadong Wang, Yangyang Hao, Liran Juan, Mingxiang Teng, Xinjun Zhang, Meimei Li, Guohua Wang, and Yunlong Liu. mir2disease: a manually curated database for microrna deregulation in human disease. *Nucleic acids research*, 37(suppl 1): D98–D104, 2009.

[234] Julie M Eggington, Tom Greene, and Brenda L Bass. Predicting sites of adar editing in double-stranded rna. *Nature communications*, 2:319, 2011.

[235] Yishay Pinto, Haim Y Cohen, and Erez Y Levanon. Mammalian conserved adar targets comprise only a small fragment of the human editosome. *Genome Biol*, 15(1):R5, 2014.

[236] Gokul Ramaswami and Jin Billy Li. Radar: a rigorously annotated database of a-to-i rna editing. *Nucleic acids research*, page gkt996, 2013.

[237] Jae Hoon Bahn, Jae-Hyung Lee, Gang Li, Christopher Greer, Guangdun Peng, and Xinshu Xiao. Accurate identification of a-to-i rna editing in human by transcriptome sequencing. *Genome research*, 22(1):142–150, 2012.

[238] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 459–468. IEEE, 2006.

[239] Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388. ACM, 2002.

[240] James Baglama and Lothar Reichel. Restarted block lanczos bidiagonalization methods. *Numerical Algorithms*, 43(3):251–272, 2006.

[241] J Baglama and L Reichel. irlba: Fast partial svd by implicitly-restarted lanczos bidiagonalization. r package version 1.0. 2, 2012.

[242] Torsten Hothorn, Kurt Hornik, Mark A Van De Wiel, and Achim Zeileis. A lego system for conditional inference. *The American Statistician*, 60(3), 2006.

[243] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3): 651–674, 2006.

[244] Jan Dominik Kuhlmann, Alexander Baraniskin, Stephan A Hahn, Frank Mosel, Maren Bredemeier, Pauline Wimberger, Rainer Kimmig, and Sabine Kasimir-Bauer. Circulating u2 small nuclear rna fragments as a novel diagnostic tool for patients with epithelial ovarian cancer. *Clinical chemistry*, pages clinchem–2013, 2013.

[245] Xiao-Shuai Ren, Meng-Hui Yin, Xin Zhang, Zhe Wang, Shi-Peng Feng, Guo-Xin Wang, Ying-Jun Luo, Pei-Zhou Liang, Xiu-Qun Yang, Jian-Xing He, et al. Tumor-suppressive microrna-449a induces growth arrest and senescence by targeting e2f3 in human lung cancer cells. *Cancer letters*, 2013.

[246] Lilian T Gien and Helen J Mackay. The emerging role of parp inhibitors in the treatment of epithelial ovarian cancer. *Journal of oncology*, 2010, 2009.

[247] Ying Dong, Michael D Walsh, Margaret C Cummings, R Gordon Wright, Soo Keat Khoo, Peter G Parsons, and Michael A McGuckin. Expression of muc1 and muc2 mucins in epithelial ovarian tumours. *The Journal of pathology*, 183(3):311–317, 1997.

[248] W Chen and P Peng. Expression and clinical significance of xiap and caspase-3 protien in primary epithelia ovarian cancer]. *Xi bao yu fen zi mian yi xue za zhi= Chinese journal of cellular and molecular immunology*, 26(7):673, 2010.

[249] Masashi Takawa, Ken Masuda, Masaki Kunizaki, Yataro Daigo, Katsunori Takagi, Yukiko Iwai, Hyun-Soo Cho, Gouji Toyokawa, Yuka Yamane, Kazuhiro Maejima, et al. Validation of the histone methyltransferase ezh2 as a therapeutic target for various types of human cancer and as a prognostic marker. *Cancer science*, 102(7):1298–1305, 2011.

[250] Arne Östman and Carl-Henrik Heldin. Pdgf receptors as targets in tumor treatment. *Advances in cancer research*, 97:247–274, 2007.

[251] Cheng-Chi Chang, Jin-Yuan Shih, Yung-Ming Jeng, Jen-Liang Su, Been-Zen Lin, Szu-Ta Chen, Yat-Pang Chau, Pan-Chyr Yang, and Min-Liang Kuo. Connective tissue growth factor and its role in lung adenocarcinoma invasion and metastasis. *Journal of the National Cancer Institute*, 96(5):364–375, 2004.

[252] Lanxi Song, Domenico Coppola, Sandy Livingston, W Douglas Cress, and Eric B Haura. Mcl-1 regulates survival and sensitivity to diverse apoptotic stimuli in human non-small cell lung cancer cells. *Cancer biology & therapy*, 4(3):267–276, 2005.

[253] Sébastien Pfeffer, Mihaela Zavolan, Friedrich A Grässer, Minchen Chien, James J Russo, Jingyue Ju, Bino John, Anton J Enright, Debora Marks, Chris Sander, et al. Identification of virus-encoded micrornas. *Science*, 304(5671):734–736, 2004.

[254] Regina Feederle, Sarah D Linnstaedt, Helmut Bannert, Helge Lips, Maja Bencun, Bryan R Cullen, and Henri-Jacques Delecluse. A viral microrna cluster strongly potentiates the transforming properties of a human herpesvirus. *PLoS Pathog*, 7(2):e1001294, 2011.

[255] Eri Seto, Andreas Moosmann, S Gromminger, Nicole Walz, Adam Grundhoff, and Wolfgang Hammerschmidt. Micro rnas of epstein-barr virus promote cell cycle progression and prevent apoptosis of primary human b cells. *PLoS Pathog*, 6(8):e1001063, 2010.

[256] Angela Wahl, Sarah D Linnstaedt, Caitlin Esoda, John F Krisko, Francisco Martinez-Torres, Henri-Jacques Delecluse, Bryan R Cullen, and J Victor Garcia. A cluster of virus-encoded micrornas accelerates acute systemic epstein-barr virus infection but does not significantly enhance virus-induced oncogenesis in vivo. *Journal of virology*, 87(10):5437–5446, 2013.

[257] Feng-Chun Ye, Fu-Chun Zhou, Jian-Ping Xie, Tao Kang, Whitney Greene, Kurt Kuhne, Xiu-Fen Lei, Qui-Hua Li, and Shou-Jiang Gao. Kaposi's sarcoma-associated herpesvirus latent gene vflip inhibits viral lytic replication through nf-κb-mediated suppression of the ap-1 pathway: a novel mechanism of virus control of latency. *Journal of virology*, 82(9): 4235–4249, 2008.

[258] Monique Barel, Michelle Balbo, Muriel Le Romancer, and Raymond Frade. Activation of epstein-barr virus/c3d receptor (gp140, cr2, cd21) on human cell surface triggers pp60src

and akt-gsk3 activities upstream and downstream to pi 3-kinase, respectively. *European journal of immunology*, 33(9):2557–2566, 2003.

[259] Stefan Golz, Ulf Bruggemeier, Andreas Geerts, and Bernhard Weingartner. Diagnostics and therapeutics for diseases associated with glycogen synthase kinase 3 beta (gsk3b), February 12 2005. US Patent App. 10/590,304.