# A DATA MINING-BASED APPROACH FOR INVESTIGATING THE RELATIONSHIP BETWEEN DNA REPAIR GENES AND AGEING

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Master in Philosophy

by

Alex Alves Freitas

January 2011

# ABSTRACT

There is a clear motivation for ageing research, since ageing is the greatest risk factor for many diseases, including most types of cancer. Arguably, another strong motivation for ageing research is that, despite the large progress in this area in the last two decades, ageing is still to a large extent a poorly understood process, especially in humans.

The vast majority of biogerontology research is still based on "wet lab" experiments done with simpler organisms, due to the problems associated with performing ageing-related experiments with humans. In contrast, this thesis proposes a data mining approach, based on classification algorithms, for analysing data about human DNA repair genes and their relationship to ageing. The classification algorithms – more precisely, decision tree induction and Naive Bayes algorithms – were applied to datasets prepared specifically for this research, by adapting and integrating data from several bioinformatics resources, namely: (a) the GenAge database of ageing-related genes; (b) a web site with a comprehensive list of human DNA repair genes; (c) Uniprot, a centralized repository of richly-annotated data about proteins; (d) the HPRD (Human Protein Reference Database); and (e) the Gene Ontology – a controlled vocabulary for describing gene or protein functions. Some experiments also used a separate dataset including gene expression data.

Applying classification algorithms to such datasets aimed at producing classification models that identify which gene properties are most effective in discriminating ageing-related DNA repair genes from other types of genes – mainly non-ageing-related DNA repair genes, but in some experiments the other types of genes also included genes whose protein product interact with DNA repair genes. A related goal of this research was to analyse the automatically-built classification models from two perspectives, namely: (a) measuring the predictive accuracy (or "generalization ability") of those models from a data mining perspective; and (b) interpreting the meaning of the main gene properties relevant for classification in those models, in the light of biological knowledge about DNA repair genes and the process of ageing.

In summary, the main gene properties that were found effective in discriminating ageing-related DNA repair genes from other types of genes (mainly non-ageing-related DNA repair genes) in the datasets created in this research are as follows: ageing-related DNA repair genes' protein products tend to interact with a considerably larger number of proteins; their protein products are much more likely to interact with WRN (a protein whose defect causes the Werner's progeroid syndrome) and XRCC5 (KU80, a key protein in the initiation of DNA double-strand repair by the error-prone non-homologous end joining DNA repair pathway); they are more likely to be involved in response to chemical stimulus and, to a lesser extent, in response to endogenous stimulus or oxidative stress; and they are more likely to have high expression in T lymphocytes.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGMENTS

The author thanks his primary supervisor, Dr. Joao Pedro de Magalhaes, for his help in defining the topic of this research and for his valuable comments throughout this research project.

The author also thanks his secondary supervisor, Dr. Olga Vasieva, for her use of the Genevestigator® and Ingenuity® software tools and her help in the analysis of the results obtained by using those tools.

Thanks are also due to Dr. Fernando E.B. Otero, from the University of Kent, for his use of software to process Gene Ontology terms.

In addition, the author thanks all his colleagues in the Magalhaes' lab for creating a friendly and productive research atmosphere.

# DECLARATION

I hereby confirm that all of the experimental work in this thesis has been done by myself, with the following exceptions: the extraction of gene expression data from the Genevestigator software tool, as well as the generation of the network shown in Figure 4.3 using the Ingenuity software tool, were performed by Dr. Olga Vasieva; and the retrieval of ancestral Gene Ontology (GO) terms in the GO hierarchy, using software specific for this purpose, was performed by Dr. Fernando E.B. Otero.

# Chapter 1 – Introduction

## 1.1 WHAT IS AGEING?

### 1.1.1 Defining ageing

Ageing is a natural process, occurring in almost all species (although at very different rates in different species), and ageing signs in humans are "obvious" to a lay person. However, from a scientific point of view, ageing is still a mysterious process, whose fundamental causes are still strongly debated, and it is difficult to find a definition of ageing which is accepted as a "standard" in the literature.

Several candidate definitions are reviewed in (Arking, 2006). Two main conclusions can be drawn from that review. First, several definitions of ageing refer to changes that happen to an organism with the passage of chronological (physical) time. In principle it would be better to refer to a more relevant measure of "biological time", in terms of biomarkers of ageing involving physiological processes; but in practice it is not easy to define such biomarkers, due to our lack of knowledge of the fundamental causes of ageing.

Secondly, it seems useful to make a distinction between at least two kinds of changes affecting an organism in two different stages of its life, namely the developmental stage and the post-maturational stage. This distinction is important because organismal changes during development are not normally deleterious. In contrast, a major characteristic of the ageing process is that it produces deleterious changes (involving functional decline) in an organism, and such changes usually manifest themselves after the organism's maturation. Hence, gerontology usually focuses on the study of the ageing process after an organism has completed its development (Martin and Oshima, 2000).

Taking the above points into account, ageing can be defined as: "...the time-independent series of cumulative, progressive, intrinsic, and deleterious functional and structural changes that usually begin to manifest themselves at reproductive maturity and eventually culminate in death" (Arking, 2006).

It is worth discussing the relationship between the process of ageing and age-related diseases, which seems controversial. It has been argued that ageing and ageing-related diseases are different types of biological processes. In this spirit, (Hayflick, 2000) points out several features of the process of ageing that are not features of any specific disease, such as: ageing affects every animal that reaches adulthood, it occurs in all members of a species only after the age of reproduction, and it takes place in virtually all species. There is no specific disease with these features. On the other hand, ageing and ageing-related diseases are very related processes, and in general mutations that slow ageing also postpone age-related diseases (Kenyon, 2010). Hence, it has also been argued that the "aging is no more and no less than the *collective early stages* of the various age-related diseases" (Grey and Rae, 2007).

In any case, it is worth noting that the distinction between ageing and age-related diseases is made by several governmental agencies. For instance, the US Food and Drug Administration will not approve any drug whose purpose is "only" to slow down ageing, rather than combating a disease (Vijg and Campisi, 2008).

### 1.1.2 Ageing at the cellular and tissue levels

In order to understand ageing in complex multicellular organisms, it is often useful to study ageing at the more basic cellular and tissue levels, as discussed next.

The term "senescence" derives from the Latin word *senex*, meaning old man or old age, and the term cellular senescence was proposed to refer to the situation when a cell loses its ability to replicate in culture, based on the old assumption that the behaviour of senescent cells recapitulated organismal ageing (Campisi and Fagagna, 2007). It should be noted, however, that the loss of the ability to proliferate is not the only major change in

the behaviour of a cell when it gets senescent. Senescent cells also have altered morphology (in general an enlarged flattened shape) and altered functionality, and they are often resistant to apoptotic signals. Hence, cellular senescence is now understood as a stress response process, rather than as a process that mimics organismal ageing at the cellular level.

Functionality changes associated with cellular senescence have been extensively studied in fibroblasts – the cell type that synthesizes the stroma, the structure underlying the cells of epithelial tissues. Senescent human fibroblasts exhibit, for instance, an increased secretion of inflammatory cytokines and epithelial growth factors (Campisi, 2005).

Cellular senescence can be caused not only by the well-known shortening of telomeres (which typically occurs in humans, but not usually in mice (Lombard et al., 2005)), but also by several other factors such as DNA damage, perturbations to chromatin organization, and expression of certain oncogenes (Campisi, 2005), (Campisi and Fagagna, 2007), (Passos et al., 2009). In addition, in vitro culturing can be regarded, by itself, as a form of stress that induces premature cellular senescence (Pelicci, 2004). Cellular senescence induced by dysfunctional telomeres is often called telomere-initiated cellular senescence, whilst in general senescence induced by stress – independently of telomere dysfunction – is called stress-induced premature senescence (Zglinicki et al., 2005), (Debacq-Chainiaux et al., 2005).

There are two major ways in which cellular senescence can contribute to the ageing of an organism (Lombard et al., 2005). First, as the number of senescent progenitor or stem cells increases with age, tissue renewal is gradually impaired, leading to a loss of tissue homeostasis (Maslov and Vijg, 2009), (Pelicci, 2004). The issue of whether stem cell depletion per se is a major cause of ageing-related tissue or organ dysfunction, in the absence of high levels of genotoxic stress, is still an open question, since stem cells provide just one out of several types of homeostatic or defence mechanisms against ageing and DNA mutations or damage represent a fundamental type of change from which many harmful changes follow (Vijg, 2008), (Rando, 2006). It is also important to

note that stem cell depletion seems not to have any significant contribution to the ageing of post-mitotic tissues such as the brain and the heart – where the vast majority of cells are not replaced during adult life. Furthermore, (Maslov and Vijg, 2009) point out that in general stem cell populations are not completely lost, and they suggest that a more important role of stem cells in ageing is that mutations accumulated in stem cells are transmitted to their daughter cells that become newly differentiated cells, and this in turn contributes to a decline of tissue functionality.

The second way in which cellular senescence can contribute to organismal ageing is that the aforementioned secretory phenotype of senescent cells can modify the local tissue environment in a way that contributes to ageing-related diseases, including cancer (Campisi, 2005) – which seems ironic, considering that cellular senescence is believed to have evolved as a tumor suppressor mechanism (Campisi and Fagagna, 2007).

Turning to the tissue level, one should recall that complex multicellular organisms such as humans have a large number of different types of tissues, with very different functionality and patterns of gene expression, and therefore it is not surprising that the process of ageing manifests itself in different ways in different tissues (Arking, 2006).

Here are some examples of important differences between human tissues that may affect the ageing process. First, different tissues can be associated with very different frequencies of mutation. For instance, (Busuttil et al., 2007) reported that, in mice, the small intestine (a tissue with very high cell proliferative activity) underwent the largest increase in mutation frequency with age, among several organs studied; whilst there was virtually no increase in mutation frequency with age in the brain, which is a post-mitotic organ. In addition, different types of tissue tend to accumulate different types of mutations. For example, in mice the heart and liver exhibited many mutations involving large genomic rearrangements, whilst almost all mutations observed in the small intestine of old mice were point mutations (Maslov and Vijg, 2009).

Another example of differences between tissues – which is relevant for the DNA damage theory of ageing, which will be briefly discussed later – is that the number of endogenous abasic sites – i.e. sites in a DNA strand without a base, characterizing a type of DNA damage – has been observed to vary widely between tissues, but not within the same tissue (Nakamura and Swenberg, 1999). The greatest number of abasic sites was observed in the brain.

### 1.1.3 The motivation for ageing research

A major motivation for ageing research is that ageing is the greatest risk factor for many diseases, including virtually all types of cancer.

It is important to recall here the distinction between ageing and ageing-related diseases that is often made by policy makers in government and research-funding agencies. For instance, it has been noted that more than half of the budget of the US National Institute on Ageing is spent on Alzheimer's disease, a single ageing-related disease; even though the likelihood of dying from Alzheimer's disease is just 0.7% and a complete elimination or cure of this disease would add only about 19 days onto the average human life expectancy (Hayflick, 2000).

In contrast, if we could significantly slow down ageing in humans, we would be simultaneously postponing the onset of a large number of age-related diseases (Grey and Rae, 2007), (Kenyon, 2010), greatly increasing *both life expectancy and quality of life*. Therefore, there is a clear motivation to do research on the process of ageing, rather than just on individual age-related diseases.

One can also consider that there is a great need for ageing research because, despite the large progress in this area in the last two decades, ageing is still to a large extent a poorly understood process. To quote (Miller, 2004b):

"We now know as much about aging as scientists knew about infection after John Snow's observations on the epidemiology of cholera, which gave the first, premicroscopic hints that germs could make people sick."

## 1.2 THEORIES OF AGEING

In this section we discuss just two types of theories of ageing, namely evolutionary theories – which at present provide the most accepted explanation about why we age from an evolutionary perspective – and the DNA damage theory of ageing – which is particularly relevant for this thesis. It should be noted, however, that there are many other theories of ageing, and for a discussion of other theories the reader is referred to (Arking, 2006), (Magalhaes, 2011).

### 1.2.1 Evolutionary theories of ageing

Explaining why ageing occurs in evolutionary terms is not easy, since at first glance it seems that natural selection would not favour the evolution of ageing, given its seriously harmful (eventually lethal) effects to an organism. However, the existence of ageing can be explained, in broad terms, by evolutionary theories of ageing (Arking, 2006), (Magalhaes, 2011), as follows.

The *mutation accumulation theory* is based on Peter Medawar's insight that the force of natural selection fades with age (Medawar, 1952). That is, in the wild most animals have a very high rate of mortality – due, for instance, to predators. Hence, in general, if a mutation produces a harmful or lethal effect only in very old age, the gene allele associated with that mutation will not be eliminated from the population by natural selection, because the vast majority of animals will die from other causes before the mutation's effect is triggered. Hence, mutations that are harmful at old age tend to accumulate. The mutation accumulation theory of ageing was also influenced by Szilard, a physicist who proposed that somatic mutations are the elementary step in the ageing process based on an analogy with the known effects of radiation (Arking, 2006).

George Williams has proposed another evolutionary theory of ageing which is complementary to the mutation accumulation theory. Williams' theory, called the *antagonistic pleiotropy* theory of ageing (Arking, 2006), (Campisi, 2005), (Magalhaes, 2011), is based on the fact that some genes have pleiotropic functions – i.e., multiple,

different functions. Hence, if a gene allele has the opposite (antagonistic) effects of increasing the chances of an animal surviving until reproduction and being harmful at an older age, that gene allele will be favoured by natural selection despite its harmful effects later in life.

There is reasonable evidence for the existence of genes with antagonistic pleiotropic effects, with the caveat that most of the evidence is derived from animal models, and the evidence for genes with such effects in the wild is much weaker (Leroi et al., 2005). Evidence for the mutation accumulation theory seems less convincing and more controversial than the evidence for the antagonistic pleiotropy theory (Kirkwood, 2005), (Kirkwood and Austad, 2000).

To illustrate the difficulty of finding genes with antagonistic pleiotropy, let us consider the following case. A typical example of a gene which is often considered to have antagonistic pleiotropic effects is the well-known p53 tumor suppressor gene (Campisi and Fagagna, 2007), (Lombard et al., 2005), which contains mis-sense mutations in approximately 50% of the major forms of cancers (Ko and Prives, 1996). This gene has major roles in apoptosis and cellular senescence, and also has a role in DNA repair (Cao et al., 2006), (Ko and Prives, 1996), (Oren, 2003). Mutant mice with hyperactive p53 are, as expected, much more resistant to cancer than the wild type, but in some cases, surprisingly, they have a somewhat shorter life span (despite the lower incidence of cancer) and display multiple signs of premature ageing (Tyner et al., 2002), (Maier et al., 2004). Apparently, these results can be explained by the antagonistic pleiotrophy theory.

However, a more recent analysis of those experiments suggests that the aging phenotype of those mutant mice seems to be due to the fact they had a truncated form of the p53 protein expressed along with the wild type of the protein, and apparently it was that truncated form (with altered function) - rather than an overall increase in the normal form of the protein – that mainly contributed to the premature ageing phenotype (Mendrysa et al., 2006). In support of this analysis, Mendrysa et al. created mice with hyperactive p53 (due to decreased levels of Mdm2, a gene which is a major inhibitor of p53 activity) that

are resistant to tumor formation but do not show signs of premature ageing. Hence, it is not clear if p53 is really a gene with antagonistic pleiotropy after all.

In any case, the concept of antagonistic pleiotropy has been applied not only to genes, but to ageing-related biological processes or mechanisms. For instance, cellular senescence can be considered a tumor suppressor mechanism at relatively young ages, but senescent cells seem to contribute to an ageing phenotype in later ages – as discussed earlier. A similar comment applies to apoptosis (Campisi, 2005).

Another evolutionary theory of ageing is the disposable soma theory, proposed by Kirkwood (Kirkwood, 1996). In essence, this theory postulates that there is a trade-off between the investment of metabolic resources in the maintenance of the soma and in reproduction. Hence, organisms tend to invest more metabolic resources in the maintenance of the soma as long as they are expected to survive, but the use of metabolic resources tends to shift from soma maintenance to reproduction as the chance of the organism dying in the wild becomes higher.

Although at a high level of abstraction the disposable soma theory seems somewhat similar to the antagonistic pleiotropy theory, the disposable soma theory is more specific about the mechanisms underlying ageing (Kirkwood and Austad, 2000), (Kirkwood, 2005). In particular, the disposable soma theory emphasises the role of somatic repair mechanisms (for instance, DNA repair) in ageing, predicting that a fundamental factor underlying ageing is the accumulation of unrepaired cellular damage due to imperfect repair mechanisms. In contrast, the antagonistic pleiotropy theory is formulated in terms of a more abstract general pattern of temporal differences in gene effects.

### 1.2.2 DNA damage theory of ageing

DNA is subject to two basic types of alterations: mutations and damage (Arking, 2006). Mutations are changes in the DNA sequence, involving deletions, insertions, substitutions or re-arrangements of base pairs. Although mutations can be very harmful (leading to dysfunctional proteins), a DNA molecule with mutations but no damage still exhibits a

normal double-helix structure and consists of an uninterrupted sequence of bases. In contrast, DNA damage refers to physical or chemical alterations in the structure of a DNA molecule, which is no longer a normal double-helix. In other words, mutations change the informational content of a DNA molecule, but preserve its normal structure. Damage modifies the structure of a DNA molecule, producing an abnormal structure. Although damage to other kinds of molecules found in cells can also influence ageing, DNA damage seems a particularly important kind of damage because, unlike other cellular components which can be replaced, DNA must last the lifetime of the cell (Lombard et al., 2005).

In essence, the DNA damage theory of ageing postulates that the main cause of the functional decline associated with ageing is the accumulation of DNA damage. Note that DNA damage can have multiple effects. In particular, DNA damage can impair transcription, cause an interruption of the cell cycle until the damage is repaired or (if the damage is too serious) lead to programmed cell death (apoptosis). DNA damage can also lead to mutations when the DNA is replicated, as will be discussed later. Hence, the DNA damage theory of ageing can be interpreted in different ways, depending on how one interprets the relative contribution of each of those effects to the ageing process.

Although this theory is clearly related to the mutation accumulation theory (both involve alterations to DNA), they are also quite different, given the aforementioned differences between mutations and damage in DNA. However, the theories would tend to converge if one believed that the main effect of DNA damage causing ageing were the DNA mutations resulting from that damage.

Within the variations of the DNA damage theory of ageing, one can also distinguish between variations emphasizing either the role of nuclear DNA damage or the role of mitochondrial DNA damage in ageing. It has been argued that one reason why nuclear DNA damage might be a more important cause of ageing is that normally there are only two copies of the nuclear genome in a cell (in diploid organisms), whilst there are several

thousand copies of the mitochondrial DNA in a cell (Lombard et al., 2005). In addition, nuclear DNA (nDNA) accounts for about 99% of the cellular DNA.

On the other hand, there are also possible reasons for the greater importance of mitochondrial DNA (mtDNA) in ageing, as follows (Graziewicz et al., 2006). First, mtDNA is much more prone to damage than nDNA (since mtDNA is not protected by histone proteins and it is very close to the site of ROS – reactive oxygen species – generation in the mitochondrial membrane). In addition, overall the repair of mtDNA is less efficient than the repair of nDNA.

The relative importance of mtDNA damage and mutations for ageing is still controversial, though. In a recent review of several mouse models with increased levels of mtDNA, (Khrapko & Vijg 2008) noted that, although in general those mutant mice show multiple signs of premature ageing, the interpretation of the results requires great caution. For instance, in general the mutant mice start to have a high frequency of mutation in the developing embryo, whilst in a mice undergoing normal ageing the accumulation of mtDNA mutations would presumably be driven by ROS or other factors in a time- and tissue-dependent manner. Khrapko & Vijg conclude their review with the remark that

"Despite decades of research and recent advances in generating mouse models with increased mutational loads, the study of mitochondrial DNA mutations in aging still has not reached a stage at which clear, definitive conclusions can be drawn regarding causal relationships."

Regardless of the relative importance of nDNA damage and mtDNA damage, broadly speaking, the DNA damage theory of ageing is associated with two major predictions (Arking, 2006). The first one is that there is a positive correlation between DNA repair ability and life span. This correlation involves two aspects, namely: (a) reduced or defective DNA repair should lead to accelerated or premature ageing, ultimately leading to shorter life span; and (b) improved DNA repair should lead to slower or postponed ageing, ultimately leading to longer life span.

There is good evidence for (a), as follows. First, many human progeroid syndromes – diseases characterized by accelerated ageing, which will be reviewed later – are due to defective nDNA repair, and overall the ones with the most severe progeroid phenotype are the ones with most severe nDNA repair defects (Best, 2009), (Lombard et al., 2005). In addition, there are many DNA repair genes whose deletion leads to a premature ageing phenotype in mouse models (Hasty et al., 2003), (Magalhaes and Faragher, 2008).

In summary, as pointed out by (Arking, 2006): "Damage to nuclear DNA likely contributes to the aging process, at least in certain tissues. The data strongly suggests that the absence of DNA repair ability probably has a causal relationship to the expression of a shortened life span and accelerated senescence."

However, it is much harder to find evidence for (b), which increased DNA repair leads to increased life span – and even harder to find evidence that such increased life span is due to slower aging, since life span could be extended due only to a reduced risk of death from disease. One study showing an increase in life span with improved DNA repair in flies is presented in (Symphorien and Woodruff, 2003). In this work, *D. melanogaster* with one or two extra copies of a DNA repair gene had a significantly extended life span, although the extension was not large. Also, in this study DNA repair was not directly measured, it was simply assumed to directly vary with gene dose (Arking, 2006).

This kind of experiment artificially increasing the number of copies of a gene cannot usually be done in humans (for ethical reasons, at least), but one can study the DNA repair systems of centenarians, to determine if they are more efficient than the DNA repair systems of most old individuals. (Chevanne et al., 2007) compared the efficiency with which cells from young, old and centenarian subjects repair DNA strand breaks caused by sublethal concentrations of the oxidant hydrogen peroxide ($H_2O_2$). They observed that cells from centenarians are about as efficient in that kind of repair as the cells from young subjects, and both types of cell were considerably more efficient in that task than the cells of old subjects. They also observed that the expression level of PARP-1 – a protein with important role in DNA repair (Beneke and Burkle, 2007) – was

significantly decreased in the cells of old subjects, but not in the cells of young and centenarian subjects. In addition, centenarians have significantly higher levels of the KU70 protein (a key factor in the repair of DNA double-strand breaks – as will be discussed later). These results support the hypothesis that improved DNA repair systems may lead to longer life span.

The second major prediction of the DNA damage theory of ageing is that a systemic, age-related shift in DNA repair activity is a major factor underlying the functional decline associated with the process of ageing (Arking, 2006). There are many studies showing evidence that many kinds of DNA repair activity decrease with age, and a number of such studies will be discussed later in this thesis, in the context of specific DNA repair pathways. Here we just mention that, in a recent review of work in this area, (Gorbunova et al., 2007) concluded that: "There is sufficient evidence that all pathways of DNA repair...become less efficient with age".

There are, however, two caveats to this kind of conclusion. First, it should be noted that most studies that measure DNA repair tend to focus on just one type of DNA repair, but the decrease of one type of DNA repair may be compensated by an increase in another type (Engels et al., 2007). Also, DNA repair activity is very difficult to measure, especially in vivo (Maslov and Vijg, 2009).

## 1.3 PROGEROID SYNDROMES

There are many types of diseases where the patient shows signs of premature or accelerated ageing. Such diseases are called *progeroid syndromes* (Martin and Oshima, 2000) or *premature ageing syndromes* (Pesce and Rothe, 1996). Interestingly, most of these diseases are caused by defects in DNA repair genes (Hasty et al., 2003), (Magalhaes and Faragher, 2008).

This section is divided into two parts. First, we review some major progeroid syndromes caused by defects in DNA repair. Secondly, we discuss the relevance of such progeroid syndromes to the study of normal human ageing.

### 1.3.1 An overview of progeroid syndromes

#### *1.3.1.1 Werner syndrome (WS)*

This is usually considered the progeroid syndrome that shows more symptoms of normal ageing and ageing-related diseases (Magalhaes and Faragher, 2008). WS is caused by a variety of loss-of-function mutations in a gene coding for a protein that is a member of the RecQ DNA helicase family (WRN) (Kipling et al., 2004). The WRN protein is involved in several important biological processes, related to DNA replication, recombination, apoptosis and telomere metabolism, but its major function seems to be the re-initiation of stalled replication forks. *In vitro*, the absence of WRN leads to a mutator phenotype, characterized by an increased frequency of deletional mutations, resulting from the inability to re-initiate stalled replication forks.

WS patients usually seem to be normal during childhood, but stop growing during the teenage years (Best, 2009). WS patients usually show the following ageing symptoms (Davis and Kipling, 2006), (Martin and Oshima, 2000): premature graying of the hair and baldness, skin and muscular atrophy, hypogonadism, poor wound healing, atherosclerosis, osteoporosis, soft-tissue calcification, juvenile cataracts, a tendency toward diabetes, and an elevated cancer frequency (Arking, 2006), (Kipling et al., 2004). On the other hand, WS patients show no increased tendency for neurodegeneration or Alzheimer's Disease, and the immune system remains normal. The median age at death is 47 years, and death is usually a result of cancer or arteriosclerosis.

The cells of WS patients show significant chromosomal abnormalities and accumulation of DNA double-strand breaks (Ariyoshi et al., 2007), (Best, 2009). WS fibroblasts reach the stage of replicative senescence considerably faster than normal fibroblasts, but both types of fibroblasts have been observed to have very similar transcriptional changes and gene expression patterns after senescence (Kipling et al., 2004), (Lee et al., 2005).

*1.3.1.2 Hutchinson-Gilford progeroid syndrome (HGPS)*

This progeroid syndrome has an onset in childhood, much earlier than the onset of WS. For this reason, HGPS is sometimes called "child progeria", whilst WS is sometimes called "adult progeria". HGPS is caused by a point mutation in the gene for lamin A, a type of protein that forms a network of filaments beneath the inner nuclear membrane (among other possible locations in the nucleus) (Bridger and Kill, 2004). A-type lamins are also believed to be involved in nDNA replication and RNA polymerase II-dependent transcription, and higher-order chromatin structure (they can directly bind to DNA and to chromatin).

There are several types of HGPS, which can be divided into classical and non-classical HGPS (Hennekam, 2006). Here we focus on the classical HGPS, which is associated with a strong resemblance of symptoms among the affected patients and tends to be the more severe form of the disease.

HGPS patients show the following symptoms (Hennekam, 2006), (Arking, 2006): premature loss of hair and subcutaneous fat (starting in the first year), postnatal growth is severely disturbed, no pre-pubertal or pubertal growth spurt, osteolysis, decreased joint mobility from the second to third year, thinning of the skin, limited sexual development and severe vascular problems in the brain and elsewhere – strokes occur at the median age of nine years. The vast majority of patients have a normal cognitive development. No neurofibrillary tangles appear in the central nervous system. The median age at death is 12 years. The cause of death is usually of vascular origin, in particular myocardial infarctions.

The study of HGPS seems less relevant for the understanding of normal human ageing than the study of WS, because, although HGPS patients show symptoms that superficially resemble premature ageing, there seems to be no basic mechanism shared between HGPS symptoms and normal human ageing, as pointed out by (Arking, 2006). However, it should be noted that, although WS and HGPS patients have little overlap of clinical symptoms, at the cellular level both these progeroid syndromes are associated with

genomic instability and both lead to accelerated ageing in some tissues (Bridger and Kill, 2004).

*1.3.1.3 Trichothiodystrophy (TTD)*

TTD is caused by point mutations in the XPD gene, which encodes one of the two core transcription factor IIH (TFIIH) helicases (Hasty et al., 2003). Different mutations in this gene can give rise to TTD, XP (xeroderma pigmentosum) or CS (Cockayne syndrome). The helicase encoded by the XPD gene is involved in both DNA repair and transcription initiation (Boer et al., 2002).

TTD patients show the following symptoms (Hasty et al., 2003), (Boer et al., 2002): neurodegeneration (including cerebellar ataxia), skeletal degeneration, impaired sexual development, cachexia (i.e, a patient's unintentional loss of body mass that cannot be reversed by nutrional means), osteoporosis, cataracts, brittle hair and nails. Patients have a mean life span of just about 10 years, and show no predisposition to cancer.

*1.3.1.4 Cockayne syndrome (CS)*

CS is caused mainly by mutations in either the CSA or CSB gene. In addition, as mentioned earlier, CS can also be caused by a mutation in the XPD gene, whose mutation can also cause TTD or xeroderma pigmentosum (XP). CS patients show the following symptoms (Hasty et al., 2003), (Pesce and Rothe, 1996): neurodegeneration (including cerebellar ataxia), growth retardation (in CS type I, growth failure usually starts in the first year of life), cachexia, thin hair, retinal degeneration, hearing loss, and cataracts – which can be seen at birth in the most severe cases. Almost all CS patients are mentally retarded. Note that TTD (reviewed earlier) and CS have several symptoms in common (Boer et al., 2002). Despite chromosomal instability, patients show no predisposition to cancer. The average age at death for CS patients is estimated as 20 years in (Hasty et al., 2003) and as 12 years in (Pesce and Rothe, 1996).

*1.3.1.5 Ataxia telangiectasia (AT)*

AT is caused by a loss-of-function mutation in the ATM (ataxia-telangiectasia mutated) gene. The term ataxia refers to the shaky and unsteady limb movements, due to the brain's failure to regulate the body's posture and regulate the strength and direction of limb movements. The ATM gene's product is a protein kinase which is involved in several signal transduction pathways, which operate both under stress and in normal physiological conditions (Rotman and Shiloh, 1997). In particular, ATM is involved in cell cycle progression and checkpoint response to DNA damage.

AT patients show the following main symptoms (Pesce and Rothe, 1996), (Rotman and Shiloh, 1997), (Wong et al., 2003): progressive neurodegeneration – with cerebellar ataxia becoming apparent when the patient begins to walk, telangiectases (dilated blood vessels) – with onset typically between three and five years of age, immunodeficiency, genomic instability, strong cancer predisposition (mainly of lymphoid origin) and sensitivity to radiation, accelerated telomere loss, and growth retardation in many patients – although those who reach puberty usually have normal height and weight. The average life span of AT patients is about 20 years (Hasty et al., 2003). ATM-deficient mice exhibit most of the symptoms of the human disease (Rotman and Shiloh, 1997), (Wong et al., 2003).

*1.3.1.6 Rothmund-Thomsom (RT) syndrome*

RT is caused by a mutation in a gene coding for a RecQ-like DNA helicase (Rotman and Shiloh, 1997), (Hasty et al., 2003). RT patients typically exhibit the following symptoms (Pesce and Rothe, 1996): skin changes starting in the first year of life and leading to poikiloderma (a condition involving pigmentary and atrophic changes in the skin) – most prominent over sun-exposed skin, growth retardation, a variety of skeletal and ocular abnormalities, including osteoporosis and corneal or retinal atrophy, as well as juvenile cataracts. Malignancies have also been reported, and delayed or immature sexual development has been reported for about 28% of the patients. Most patients have normal intelligence. Surprisingly, despite the premature ageing phenotype, RT patients seem to have a normal life span.

*1.3.1.7 Xeroderma pigmentosum (XP)*

This is a disease due to a defect in one of seven genes (XPA – XPG) required for nucleotide excision repair (a form of DNA repair to be reviewed later). XP victims show dramatically accelerated aging only in areas of skin exposed to the sun and a skin cancer rate more than a thousand times greater than normal, and frequently exhibit neurodegeneration (Hasty et al., 2003), (Best, 2009). Overall, the symptoms of accelerated ageing in XP seem to be more focused than the broader symptoms associated with other progeroid syndromes such as Werner.

To summarize the previous discussion on progeroid syndromes, Table 1.1 shows, for each of those syndromes: which genetic defect is associated with it and what are the main processes affected by that defect; what is the mean life span (in years) of patients with that syndrome; and whether or not the syndrome is associated with an increased incidence of cancer.

**Table 1.1:** Summary of major human progeroid syndromes (adapted from (Hasty et al., 2003), (Arking, 2006), (Lans and Hoeijmakers, 2006))

| Syndrome | Genetic defect | Mean life span (years) | Predisposition to cancer? |
|---|---|---|---|
| Werner | RecQ-like DNA helicase and exonuclease, involved in DNA repair | 47 | yes |
| Hutchinson-Guilford | Lamin A, involved in nDNA replication, transcription, nuclear organisation | 13 | no |
| Trichothiodystrophy | TFIIH helicase, involved in DNA repair and transcription | 10 | no |
| Cockayne | CSA or CSB gene, involved in DNA repair and transcription | 12-20 | no |
| Ataxia telangiectasia | ATM protein kinase, involved in DNA damage response | 20 | yes |
| Rothmund-Thomson | RecQ-like DNA helicase | Normal? | yes |
| Xeroderma Pigmentosum | XPA – XPG genes, involved in DNA repair | Lower than normal? | yes |

### 1.3.2 On the relevance of progeroid syndromes to the study of human ageing

The relevance of the study of progeroid syndromes for the understanding of normal ageing is controversial, particularly when such progeroid syndromes are observed in mouse models of their human counterpart. (Miller, 2004a) has strongly criticized this relevance, arguing that progeroid syndromes and animal models of premature or accelerated ageing in general do not offer any significant insights into the normal human ageing process. A major point of his criticism is that patients or animal models of progeroid syndromes show just a subset of the symptoms of normal ageing, whilst normal ageing is characterized by a much broader range of symptoms. Actually, progeroid syndromes are often called *segmental progeroid syndromes*, to emphasize the fact that their pathologies are limited to one or a few organs or tissue types, so that they resemble ageing only in part (Arking, 2006).

In addition, Miller points out that it is relatively easy to make an animal have a significantly shorter life span by introducing a defect in some crucial DNA repair gene, but it is much more difficult to show that the defect is really accelerating ageing – particularly considering that we still do not have a very reliable biomarker of ageing. Furthermore, although in several cases a mutation in a single DNA repair gene leads to signs of premature or accelerated ageing in mice, not all mice mutants with defective DNA-repair genes show signs of premature or accelerated ageing.

Counter-arguments to Miller's criticism have been provided in (Hasty and Vijg, 2004b) and (Kipling et al., 2004). (Hasty and Vijg, 2004b) point out that the "segmental" nature of progeroid syndromes does not invalidate their relevance for the study of normal ageing, because every individual who undergoes normal ageing exhibits a segmental ageing phenotype, by comparison with all possible ageing phenotypes in the population. That is, segmental ageing is natural and normal. A similar argument is made by (Best, 2009). The extent to which progeroid syndromes patients have symptoms of normal ageing varies significantly among those syndromes – as discussed earlier – but at least Werner's syndrome is considered to have symptoms which have a very significant overlapping with the symptoms of normal ageing. In their review of the relevance of the study of progeroid

syndromes for the understanding of normal human ageing, (Kipling et al., 2004) conclude that there is good evidence that the premature replicative senescence of Werner's syndrome cells is a causal factor in the aspects of that syndrome that resemble premature ageing, and that this evidence supports a causal role for replicative senescence in normal human ageing.

Concerning Miller's argument that not all mice mutants with DNA-repair defects show signs of accelerated ageing, used as an argument against the relevance of DNA repair in ageing, (Hasty and Vijg, 2004b) point out that deletion of some crucial DNA repair genes leads to embryonic death or cancer at an early age, so that there is simply no time for the ageing phenotype to appear. This shows that DNA repair is crucial for survival, but this is not incompatible with the fact that DNA repair is also important for ageing. Hence, some DNA repair defects are so harmful that they lead to early death, whilst other defects allow the organism to survive long enough to show signs of premature or accelerated ageing. The authors conclude that progeroid syndromes in human and mouse mutants with defects in some DNA repair genes offer strong support to the idea that a major causal factor of ageing is the accumulation of DNA damage and mutations with time, which eventually leads to cellular senescence or apoptosis (Hasty et al., 2003).

In any case, one should recall that, unlike normal ageing, in general progeroid syndromes are subject to the force of natural selection (so that the existence of progeroid syndromes cannot be explained by the evolutionary theory of ageing), and some phenotypic features of those syndromes may result from gross abnormalities in development which are not directly relevant for gerontology (Martin and Oshima, 2000) – recall that gerontology involves mainly the study of post-maturational ageing following normal development.

## 1.4 DNA DAMAGE

This section is divided into two subsections. The first one discusses two major types of sources of DNA damage (arguably the two most studied ones), namely oxidative damage and damage induced by ultra-violet radiation. The second subsection discusses specific types of damages regardless of the cause of the damage. That subsection focuses just on describing the damage itself, not how to repair it, which will be the focus of section 1.5.

### 1.4.1 Two major sources of DNA damage

*1.4.1.1 Oxidative damage*

A common cause of DNA damage is exposure to reactive oxygen species (ROS). ROS include superoxide, hydrogen peroxide, hydroxyl radicals and singlet oxygen. Such ROS can oxidize DNA, which can produce several kinds of DNA damage, in particular oxidized bases, abasic sites and single- and double-strand breaks (Bont and Larebeke, 2004), (Jackson and Loeb, 2001).

Different types of ROS vary considerably in their degree of reactivity with cellular components. Superoxide has limited reactivity but it is converted to hydrogen peroxide by superoxide dismutase. Hydrogen peroxide is reduced to water by catalase and glutathione peroxidase; but, in the presence of some metals such as iron, hydrogen peroxide is reduced to hydroxyl radicals. These radicals' reactivity is so great that they do not diffuse more than one or two molecular diameters before reacting with a cellular component (Friedberg et al., 2006), (Bont and Larebeke, 2004). Hence, a hydroxyl radical is able to oxidize DNA only if that radical is produced immediately adjacent to DNA. On the other hand, hydrogen peroxide can be seen as a diffusible, latent form of hydroxyl radical that reacts, for instance, with iron ($Fe^{+2}$) adjacent to a DNA molecule to generate a hydroxyl radical.

Organisms have several defence mechanisms to cope with ROS (Friedberg et al., 2006). For instance, in eukaryotic cells, oxygen metabolism is concentrated on mitochondria, so

that normally the nucleus is practically anoxic, helping to protect nuclear DNA from the harmful effects of ROS. Also, in mammals the level of oxygen concentration in tissues is only 3 to 4%, which is much smaller than the environmental level of about 20%. In addition, to reduce the use of iron available to be used in the production of hydroxyl radicals as mentioned earlier, most iron in cells is safely stored in ferritin and transferrin. Furthermore, organisms produce a number of antioxidant enzymes that eliminate ROS. For instance, catalase removes hydrogen peroxide, and superoxide dismutase (SOD) eliminates superoxide. There are also several redox-activated transcription factors, indicating that mechanisms to cope with ROS are genetically regulated. Despite all these defence mechanisms against ROS-induced damage, sometimes the production of ROS is so overwhelming that those defence mechanisms are not enough, and in this case oxidative stress occurs.

ROS can produce many different kinds of damage and mutation in DNA. For instance, the cytosine base alone can undergo oxidative damage producing at least 40 different modified species (Jackson and Loeb, 2001). Some oxidatively modified bases block DNA replication, whilst others are mispaired and lead to base substitutions in the DNA. It is interesting to note that some of the progeroid syndromes caused by defective DNA repair discussed earlier – such as xeroderma pigmentosum and ataxia telangiectasia – are associated with a high amount of 8-oxo-dG (Bont and Larebeke, 2004).

In any case, in the last few years it has become clearer that ROS is not just a source of damage, but also relevant players in signalling pathways, involved in the regulation of gene expression, development, growth and apoptosis (Magalhaes and Church, 2006).

*1.4.1.2 Damage induced by ultraviolet (UV) radiation*

The UV radiation spectrum can be divided into several segments based on wavelength (Friedberg et al., 2006): UV-A (320-400nm), UV-B (295-320nm), UV-C (100 to 295nm). It should be noted that, in order to study DNA repair, most laboratories use UV-C light, but organisms in nature are not exposed to this wavelength. This suggests some caution in interpreting the results of laboratory experiments involving UV-induced damage,

although (Friedberg et al., 2006) point out that many of the DNA lesions produced by UV-C are also produced at longer wavelengths, and UV-C light is normally used because it is more efficient in producing those lesions.

In any case, in the natural world UV-A and UV-B are the major cause of DNA damage (Yaar and Gilchrest, 2007). UV-B photons are on average 1,000 times more energetic than UV-A photons, and so UV-B radiation is believed to be the main type of UV radiation responsible for direct DNA damage and photocarcinogenesis after sun exposure**.** UV-A radiation is still likely to be a relevant cause of photodamage, though, because it is at least 10-fold more abundant in terrestrial sunlight and on average it has a greater depth of penetration into the dermis, by comparison with UV-B radiation. UV-A radiation also triggers mtDNA mutations.

The main types of DNA damage caused by UV radiation are single-strand breaks and numerous photoproducts (Friedberg et al., 2006) – the most frequent type of which is cyclobutane pyrimidine dimers (Yamada et al., 2006). (These types of DNA damage will be further discussed below.) UV radiation is also a major exogenous source of ROS (Yaar and Gilchrest, 2007). UV-B light crosses the epidermis and results in the generation of ROS and DNA damage, leading to the activation of signalling pathways related to stress response and ageing (Debacq-Chainiaux et al., 2005). UV radiation also has an effect in transcriptional down-regulation of a transforming growth factor-β (TGF- β) receptor, which leads to reduced production of collagen (Quan et al., 2004). In experiments with yeast, (Kozmin et al., 2005) concluded that UVA radiation can be strongly mutagenic due to the generation of oxidative DNA damage.

### 1.4.2 An overview of major types of DNA damage

#### *1.4.2.1 Depurination and depyrimidination*

Depurination involves the loss of purine bases (adenine and guanine) from DNA. In spontaneously-occurring depurination reactions, the N-glycosyl bond to deoxyribose is broken by hydrolysis, leaving the DNA's sugar-phosphate chain intact, producing an

abasic site. In general about 10,000 purine bases are lost every day from the DNA of each mammalian cell by spontaneous depurination reactions (Friedberg et al., 2006).

Depyrimidination involves the loss of pyrimidine bases (cytosine and thymine) from DNA. Depyrimidination is much less common than depurination, however, since the N-glycosyl bond between a pyrimidine base and the deoxyribose is more stable than the corresponding bond for purine bases. The rate of loss of depyrimidination has been estimated as about only 5% of the rate of depurination (Lindahl, 1993).

*1.4.2.2 Deamination*

Deamination involves the loss of amino groups from DNA bases. Almost all DNA bases undergo deamination in spontaneous reactions, with the exception of thymine – which does not have an amino group – as summarized in Table 1.2. It should be noted that most types of deamination shown in the table produce a base that does not naturally occur in DNA (the only exception is the deamination of 5-methylcytosine), and this fact facilitates the identification and excision of the deaminated base by a DNA glycosylase enzyme, as will be discussed later.

Interestingly, the fact that DNA uses T as a base, rather than the corresponding U base in RNA, provides one possible reason why the genetic code, which is thought to have been carried in RNA bases (A, C, G, U) a long time ago, was replaced by the current code carried in DNA bases. In the current code, a deaminated C converted to a U can be easily recognized as damage and excised from DNA. However, if DNA used U, rather than T, as a natural base, a deanimated C converted into a U would not be so easily recognized as damage.

The most common type of deamination event in cells is deamination of cytosine into uracil. This event occurs at a rate of about 100-500 bases per cell per day in mammalian cells, in spontaneous deamination reactions at physiological temperatures and pH (Friedberg et al., 2006). The deamination of cytosine can be increased by several factors, including the formation of cyclobutane pyrimidine dimers – which are a form of DNA

damage by themselves. In addition, cytosine can deaminate to uracil as a result of specific biological processes, such as the somatic hypermutation phase of antibody production.

**Table 1.2:** Types of deamination in DNA bases

| Base | Deamination product | Potentially mutagenic effect? |
|---|---|---|
| Adenine | Hypoxanthine | Yes – result in base pair transitions (XH pairs as G) |
| Guanine | Xanthine | Less likely |
| Cytosine | Uracil | Yes – result in base pair transitions (U pairs as T) |
| Thymine | None | N/A |
| 5-methylcytosine | Thymine | Yes – result in base pair transitions |

DNA can also contain the base 5-methylcytosine, which base pairs with guanine and is involved in silencing gene expression at CpG sequences. The deamination of 5-methylcytosine into thymine leads to the formation of a G-T base pair. This is potentially mutagenic because, although such unnatural base pair can be corrected by the mismatch repair pathway, this process is relatively slow – by comparison with the rapid excision of a deaminated cytosine by a uracil-DNA glycosylase (which is normally abundantly present in cells). Interestingly, although only about 3% of the C bases in human DNA are methylated, GC → AT transitions at the sites of those methylated cytosines account for about one-third of the single-base mutations in inherited human diseases (Cooper and Youssoufian, 1988), (Alberts et al., 2002).

It is interesting to note that, although the deamination of cytosine to uracil is normally a form of damage to DNA that has to be repaired, in some cases the deamination of cytosine can be beneficial to the organism – at least in the short term. For instance, as part of a cell's defence against retroviruses, a cytosine deaminase catalyzes the conversion of cytosine to uracil, which decreases viral infectivity (Bishop et al., 2004).

The deamination of adenine and guanine normally occurs at a lower rate than the deamination of cytosine. More precisely, adenine is deaminated into hypoxanthine at only about 2-3% of the rate of cytosine deamination (Bont and Larebeke, 2004). The deamination of adenine into hypoxanthine is a potentially mutagenic event, because hypoxanthine can base pair with cytosine during DNA replication, which can generate A-T → G-C transitions. Hypoxanthine-DNA glycosylase is normally present at low levels in a cell, and so the excision of hypoxanthine from DNA is less efficient than the excision of uracil (deaminated cytosine).

The deamination of guanine produces xanthine, which pairs with cytosine, and so this lesion is less mutagenic than the deamination of adenine. The rate of deamination of guanine is similar to the rate of deamination of adenine (about 2-3% of the rate of cytosine deamination) (Bont and Larebeke, 2004).

*1.4.2.3 Abasic (AP) sites*

An abasic site, also called an "apurinic or apyrimidinic" (AP) site, is formed when a base is lost from the DNA by cleavage of a N-glycosyl bond, leaving the sugar-phosphate chain intact (Friedberg et al., 2006). At normal physiological conditions, it has been estimated that 50,000-200,000 AP site lesions persist at a steady-state level in mammalian cells (Nakamura and Swenberg, 1999). This number results from the balance between the continuous generation and repair of AP sites in cells.

Abasic sites are produced not only by spontaneous depurination reactions, but also by ROS (Nakamura and Swenberg, 1999), (Nakamura et al., 2000). Abasic sites are also produced in intermediate steps of the base excision repair pathway (to be discussed later). When that repair pathway is successfully completed, abasic sites are repaired, but inefficient or incomplete base excision repair might leave abasic sites in DNA. Abasic sites are potentially mutagenic, because, if they are not repaired, DNA polymerase would preferentially incorporate an adenine opposite the abasic site during DNA replication (Jackson and Loeb, 2001). In addition, 5'-cleaved AP sites might induce frameshift mutations (Bont and Larebeke, 2004).

*1.4.2.4 DNA strand breaks*

Some strand breaks are produced in intermediate steps of natural reactions and are not necessarily considered as a form of damage. An example involves DNA strand breaks occurring during lagging-strand DNA replication, which are protected by multi-protein complexes and therefore are not accessible to poly(ADP-ribose) polymerase (which modulates DNA repair) (Friedberg et al., 2006). In addition, the process of V(D)J recombination during lymphocyte development is initiated by a kind of programmed double-strand break between two recombining variable-region gene segments and their flanking sequences (Taccioli et al., 1994), (Walker et al., 2001).

However, some strand breaks are clearly a serious form of DNA damage and inhibit DNA replication, leading to the activation of DNA repair mechanisms. DNA strand breaks can be caused by oxidative damage to DNA (Sharma, 2007) or by ionizing radiation (Friedberg et al., 2006). Single-strand breaks caused by ionizing radiation and free radicals usually have complex terminal structures due to the destruction of the deoxyribose residue at the 3' or 5' end of the break, and as a result such breaks cannot usually be rejoined directly by DNA ligase. Double-strand breaks can also result from the blockage or pausing of DNA replication – which can lead to replication fork collapse and free double-stranded ends (Engels et al., 2007).

Misrepaired double-strand breaks lead to genomic rearrangements, a common and serious problem in aging organisms (Gorbunova et al., 2007). A considerably-increased frequency of DNA double strand breaks is observed in patients of some progeroid syndromes discussed earlier, such as Werner's syndrome and ataxia telangiectasia (Fishel et al., 2007). The number of single- and double-strand breaks in the neurons of rat cerebral cortex has been shown to considerably increase with age (Rao, 2007).

*1.4.2.5 Cyclobutane pyrimidine dimers (CPDs)*

CPDs are characterized by covalent linkages between adjacent pyrimidines in the same DNA strand, and they are the most frequent type of photoproduct produced when DNA is exposed to UV-B (Yaar and Gilchrest, 2007) or to UV-C radiation (Friedberg et al.,

2006). The type of CPD most frequently found in DNA consists of a thymine dimer, which is known to be mutagenic in mammalian cells (Yamada et al., 2006). The formation of CPDs can also enhance the deamination of cytosine (Friedberg et al., 2006).

## 1.5 DNA REPAIR

### 1.5.1 Base excision repair (BER)

The BER pathway corrects small alterations in a DNA strand that do not distort the overall structure of the DNA helix, such as a base altered by deamination or a missing base due to a depurination reaction. The base alterations targeted by BER may or may not block transcription and normal replication, but they frequently lead to changes in DNA sequence, being therefore potentially mutagenic (Hoeijmakers, 2001). BER is the main pathway to repair oxidative damage. The BER pathway removes lesions affecting only one DNA strand, which permits the use of the information in the complementary strand to correct the lesion in the damaged strand.

The main steps of the BER pathway are shown, in a simplified form, in Figure 1.1 (adapted from (Alberts et al., 2002)), using as example the repair of DNA with a uracil base which was accidentally produced by a deamination of a cytosine base. In the first step a uracil DNA glycosylase catalyzes the hydrolytic removal of uracil from the DNA. More precisely, the DNA glycosylase removes the uracil by cleaving its N-glycosyl bond, which leaves an abasic (or AP – apurinic or apyrimidinic) site in one strand. There are different types of DNA glycosylases, which can recognize and catalyze the removal of different altered bases, and there is partial redundancy between different glycosylases (Lindahl and Wood, 1999), which helps to make the BER pathway more robust. For instance, uracil is excised mainly by the abundant UNG glycosylase, but it can also be excised by TDG and MBD4. Next, the sugar phosphate with the missing base is cut out by the sequential action of AP endonuclease and a phosphodiesterase. Note that this leaves a gap in the damaged DNA strand. The gap of a single nucleotide is then filled by

DNA polymerase and DNA ligase. At the end of this repair process, the U that was produced by an accidental deamination has been restored to a C.

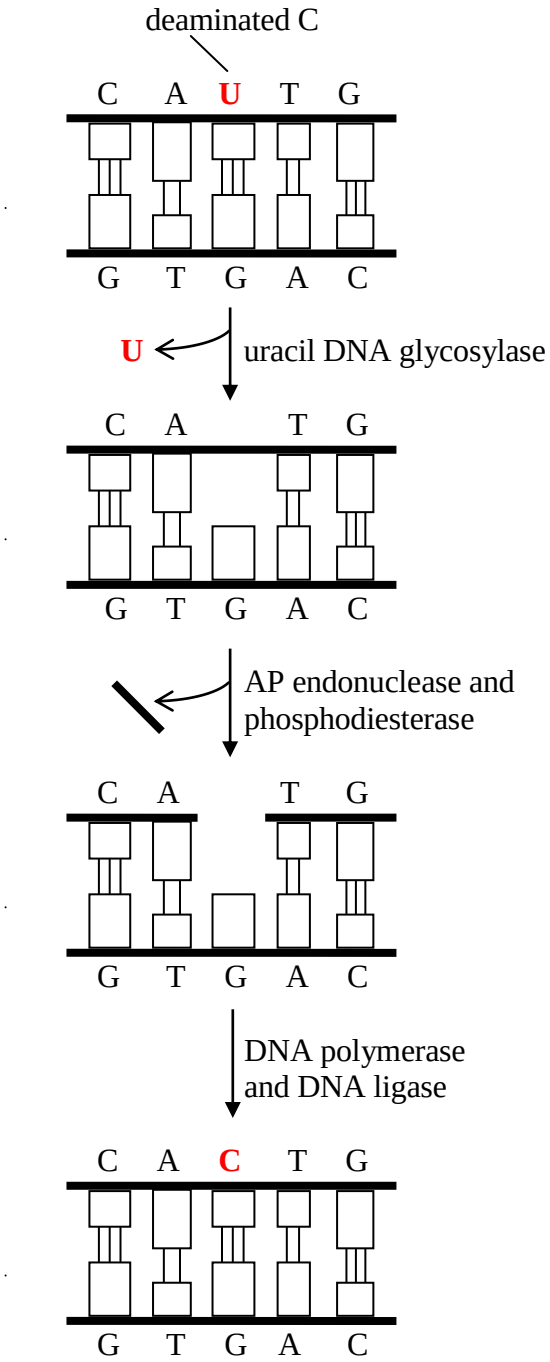


**Figure 1.1**: Main steps in the base excision repair pathway

The BER pathway can also be used to repair the result of a depurination event. In this case, since there is no need for a DNA glycosylase (the purine base was already removed by a spontaneous reaction), the BER pathway starts with the action of the AP endonuclease (Alberts et al., 2002). This is possible because the AP endonuclease recognizes any site that contains a deoxyribose sugar with a missing base, regardless of which event produced that abasic site.

The BER pathway can be categorized into two sub-pathways, namely short-patch BER, where only one nucleotide is replaced (as illustrated in Figure 1.1); or long-patch BER, where 2-13 nucleotides are replaced (Gorbunova et al., 2007). The decision between performing a short-patch or long-patch repair is modulated by PARP-1 and PARP2 (poly(ADP-ribose) polymerases) (Beneke and Burkle, 2007), and the short-patch BER sub-pathway is used by cells in approximately 80-90% of the cases (Friedberg et al., 2006).

Some differences between the short-patch and long-patch pathway are as follows (Lee et al., 2005). First, in the short-patch pathway, DNA polymerase β (polβ) is the main gap-filling enzyme; whilst in the long-patch pathway this activity seems to be performed by polβ, pol δ, polε. In addition, in the long-patch pathway, the WRN protein (defective in patients with Werner's syndrome) interacts physically and functionally with several other proteins such as PCNA and RPA, which is not the case in the short-patch pathway (Lee et al., 2005), (Rao, 2007).

The BER pathway is particularly important in the brain (Fishel et al., 2007), (Krishna et al., 2005), for at least two reasons. First, BER is the primary pathway to repair oxidative DNA damage, and this is the most likely kind of damage to occur in brain tissue, which is metabolically very active. Actually, gene expression is overall two- to three-fold higher in brain cells than in other tissues (Rao, 2007). Secondly, neurons are post-mitotic (non-

dividing) cells, and in principle other DNA repair pathways such as homologous recombination and mismatch repair (reviewed later) are not important in neurons.

There is good evidence that, overall, the level of BER activity is reduced with age. In particular, the activity of polβ – an important component of the BER pathway – has been shown to be considerably reduced with age in mice in many investigations (Rao, 2007), (Cabelof et al., 2006), (Krishna et al., 2005), (Kaneko et al., 2002), (Cabelof et al., 2002). The activity of polγ – which performs the gap-filling step of BER in mitochondrial DNA (Graziewicz et al., 2006) – has also been observed to decrease with age (Kaneko et al., 2002). There are, however, studies reporting that some BER enzymes have an increased expression with age – see, for instance, (Lu et al., 2004). This seems likely to be a response to increased levels of oxidative DNA damage with age, although the response is presumably not effective due to the aforementioned decrease in polβ activity.

### 1.5.2 Nucleotide excision repair (NER)

The NER pathway copes with lesions in a DNA strand that distort the DNA double helix. This kind of lesion interferes with base pairing and usually blocks transcription and normal replication (Hoeijmakers, 2001). NER is considered the most versatile DNA repair pathway in terms of the variety of lesions that it can recognize – it recognizes several types of bulky lesions, produced, for instance, by ultraviolet light and carcinogens. Like the lesions targeted by BER, the lesions targeted by NER affect a single DNA strand, which allows the use of the information in the complementary strand to correct the lesion in the damaged strand.

The main steps of the NER pathway are shown, in a simplified form, in Figure 1.2 (adapted from (Alberts et al., 2002), using as example the repair of a pyrimidine dimer. Once the helix-distorting pyrimidine dimer has been recognized (due to a distortion in the DNA double helix), a nuclease cleaves the phosphodiester backbone of the damaged strand at both sides of the distortion. Then, a DNA helicase removes the region of the damaged strand containing the lesion, which produces a gap in that strand. Figure 1.2

illustrates a small gap in the strand for simplicity, but the actual gap in humans is considerably larger, more than 20 nucleotides (Alberts et al., 2002). Next, the gap is filled by DNA polymerase and DNA ligase.

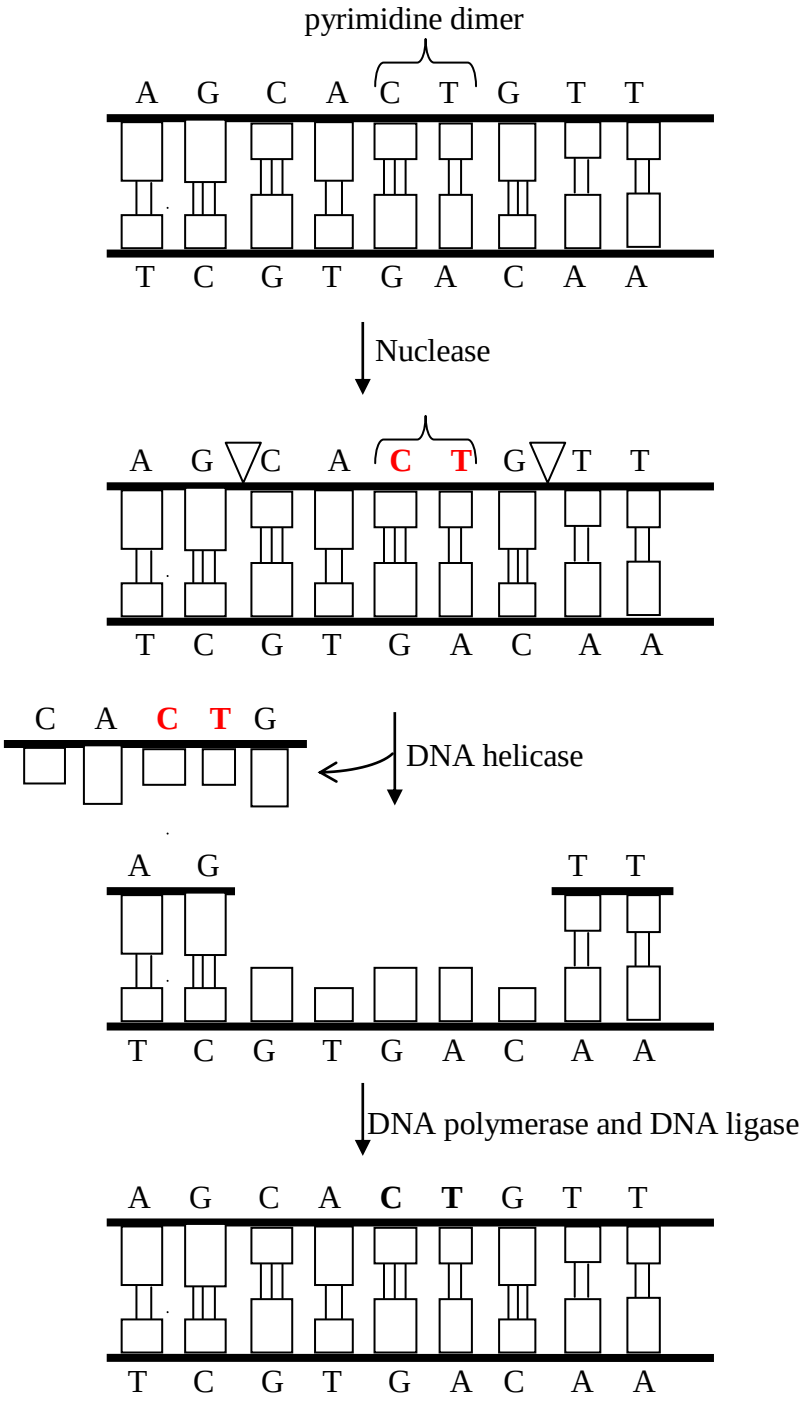


**Figure 1.2**: Main steps in the nucleotide excision repair pathway

The NER pathway is usually classified into two types, namely global genome NER (GG-NER), which occurs everywhere in the genome, and transcription-coupled NER (TC-NER), which occurs in the transcribed strand of active genes. The repair of DNA damage by TC-NER is faster than the repair by GG-NER (Moriwaki and Takahashi, 2008).

In GG-NER, the first step is the recognition of the DNA damage by the XPC-HR23B complex. In contrast, in TC-NER the repair process is believed to be triggered by a stalled RNA polymerase, and initiation of the repair requires the proteins CSB and CSA (whose mutations cause the Cockayone's progeroid syndrome) (Gorbunova et al., 2007), (Hoeijmakers, 2001). After the initial stage, GG-NER and TC-NER seem to proceed in an identical way. The presence of damage is verified by XPA, and if damage is absent the repair process is aborted. The XPB (ERCC3) and XPD (ERCC2) helicases in complex with the TFIIH transcription factor open the DNA helix double helix around the damage. RPA (Replication Protein A) stabilizes the open DNA by binding to the undamaged strand. The endonucleases XPF and XPG cleave the borders of the open segment in the damaged strand. The damaged segment is then removed, and the repair is completed by DNA polymerase and DNA ligase.

There has been many experiments investigating whether or not NER efficiency in repairing ultraviolet light-induced damage decreases with age, but the results of these experiments are sometimes conflicting (Best, 2009). For instance, NER efficiency was observed to decrease with age in (Yamada et al., 2006), (Hazane et al., 2006), (Takahashi et al., 2005), but observed not to decrease with age in (Merkle et al., 2004). It seems like these different results are due to the use of different experimental procedures and different types of damages being investigated.

However, good evidence for an association between NER and the ageing process comes from the fact that inherited defects in NER cause three major types of progeroid diseases

in humans: Xeroderma Pigmentosum (XP), Cockayne Syndrome (CS), and Trichothiodystrophy (TTD).

These diseases were reviewed earlier, but a few additional remarks are relevant here, in the context of their association with defects in the NER pathway. The severity of the symptoms in XP varies significantly across the different types of XP – associated with defects in different complementation genes – and in general the more the mutation affects the NER pathway, the more severe the symptoms are (Niedernhofer, 2008). Moreover, multiple mutations in NER genes have been observed to result in dramatically accelerated aging phenotypes (Niedernhofer et al., 2006), (Pluijm et al., 2007), (Ven et al., 2006), (Gorbunova et al., 2007).

In addition, XPD-mediated NER has been observed to have a significantly role in maintaining the functional capacity of long-term reconstituting haematopoietic stem cells (LT-HSCs) with age, by helping to preserve the proliferative capacity and to prevent apoptosis under stress (Rossi et al., 2007).

It should also be noted that XP – which is associated with a dramatic increase in skin cancers – is mainly caused by a defect in GG-NER; whilst the progeroid syndromes CS and TTD – which show no evidence of increased risk cancer – are caused mainly by defects in TC-NER (Niedernhofer et al., 2006). This is because GG-NER is responsible mainly for repairing pre-mutagenic DNA lesions, preventing carcinogenesis; whilst TC-NER is responsible mainly for repairing DNA lesions that block transcription (Ljungman and Lane, 2004). Hence, in general defects in GG-NER (but not in TC-NER) tend to greatly increase predisposition to cancer.

A particularly interesting gene for the study of the NER pathway is XPD, which encodes a helicase subunit of the transcription factor IIH complex, because different point mutations in this gene are associated with different phenotypes: cancer (XP), the TTD progeroid syndrome, or a combination of cancer and a progeroid syndrome, namely XP combined with CS (XPCS) or XP combined with TTD (XPTTD) (Ven et al., 2006).

Hence, many mouse models have been created with mutations in the XPD gene, as follows.

First, inactivation of the XPD gene led to embryonic lethality (Boer et al., 1998). Later, the same group generated mice carrying an XPD point mutation found in TTD patients, which produced mice with several symptoms of TTD, including cachexia (Boer et al., 2002). They also crossed TTD mice with $XPA^{-/-}$ mice, which greatly increased the NER defect. This produced mice with increased neonatal lethality and – among the surviving double mutants – extreme cachexia. The authors proposed that the observed premature ageing of TTD mice is due to the accumulation of DNA damage, which leads to impaired transcription, apoptosis, functional decline, and depletion of cell renewal capacity.

The multiple effects of XPD have also been exploited to create mouse models of "progeroid NER syndromes", by combining different mutant Xpd alleles with a $Xpa^{-/-}$ background (Ven et al., 2006). The authors observed that such progeroid NER mice share many similarities with long-lived dwarf and calorie-restricted mice, in particular reduced postnatal growth and small size. They concluded that this is likely due to an adaptive response to genomic instability during postnatal development, which involved dampening of the somatotropic GH/IGF-1 (growth hormone/insulin growth factor) axis, rather than being due to the proliferative defects associated with premature cell senescence – which is a common alternative explanation for this kind of progeroid phenotype.

In another related work, a mouse model was created with a mutation in the XPD gene that exhibits strong signs of progeroid TTD (Park et al., 2008), and the $Xpd^{TTD}$ mice were also observed to have reduced body and organ weight. In this work the authors used microarray gene expression analysis to investigate the impact of the $Xpd^{TTD}$ mutation in the liver (which is a major target of oxidative damage due to its central role in metabolism). They noted that the rate of apoptosis exceeded the rate of cell proliferation, resulting in homeostatic imbalance, and that this imbalance was associated with decreased energy metabolism and reduced IGF-1 signaling. Hence, similarly to (Ven et al., 2006),

(Park et al., 2008) concluded that the reduced energy metabolism is likely to reflect an adaptive response to the increased DNA damage in those mouse mutants.

### 1.5.3 Repair of double-strand breaks

Here we discuss two basic pathways for the repair of double-strand breaks, namely the homologous recombination (HR) and the non-homologous end-joining (NHEJ) pathways. For a discussion of variants of those basic pathways, the reader is referred to (Friedberg et al., 2006), (Engels et al., 2007).

#### *1.5.3.1 Homologous recombination (HR)*

This repair pathway exploits the fact that diploid cells contain two copies of the DNA double helix. In this pathway the undamaged chromosome is used as the template for the repair of the broken chromosome. This type of repair involves the two sister DNA molecules that exist in each chromosome in cells that have replicated their DNA but not divided yet – i.e., in phases S and G2 of the cell cycle (Hoeijmakers, 2001). In contrast, the error-prone NHEJ pathway (discussed below) is mainly used in phase G1 of the cell cycle, before DNA replication, when there is no sister copy of DNA to be used in homologous recombination, although NHEJ seems to occur throughout the cell cycle (Lombard et al., 2005). The type of double-strand break repair also depends on the tissue or cell type. For instance, in non-dividing cells like neurons, it seems that homologous recombination is not an option, and double-strand breaks have to be repaired by NHEJ (Vyjayanti and Rao, 2006).

The main steps in homologous recombination are as follows (Hoeijmakers, 2001), (Lombard et al., 2005). First, in response to DNA damage in the form of double-stranded DNA breaks, the ATM kinase (the product of the "ataxia-telangiectasia mutated" gene) phosphorylates the RAD50/MRE11/NSB1 complex. This complex exposes the 3' ends of both broken strands, which promotes "strand invasion". This is a process where Rad51 and a number of other proteins (including BRCA2 and several Rad51 paralogs) form a nucleoprotein complex with the DNA and direct an overhanging 3'end of a single-stranded DNA to search for, invade and pair with an undamaged homologous DNA

molecule. A DNA polymerase then carries out the repair using the undamaged DNA as a template. This process creates intermediate structures called Holliday junctions – structures in which the two double-stranded DNA complexes are intertwined. These junctions can be resolved in different ways, depending on which type of biological process is being performed (Alberts et al., 2002). In particular, in the case of DNA repair the two original DNA molecules are normally restored with the repair of the double-strand break, but in the case of meiosis a crossover event normally occurs, where some regions of DNA sequences are exchanged between the two DNA molecules. Next, a DNA ligase completes the repair process.

*1.5.3.2 Non-homologous end joining (NHEJ)*

The NHEJ pathway simply links the ends of a double-strand break together, without using any strand as a template. This pathway is more error prone than the homologous recombination pathway, and it tends to produce deletions or insertions in DNA strands. Nonetheless, in mammalian cells this seems the main pathway for the repair of double-strand breaks resulting from ionizing radiation (Walker et al., 2001), as well as being the pathway used to repair the double-strand breaks that arise during V(D)J recombination in lymphocyte formation, as mentioned earlier.

The main steps in the NHEJ repair pathway are as follows (Gorbunova et al., 2007), (Hoeijmakers, 2001), (Holcomb et al., 2008), (Seluanov et al., 2007). The repair process starts with the binding of the KU heterodimer (consisting of KU70 and KU80 subunits, with 70 and 86 kDa, respectively) to the broken DNA strand ends, which recruits DNA-dependent protein kinase catalytic subunit (DNA-$PK_{cs}$). Then, KU complexes with DNA-$PK_{cs}$ to form a holoenzyme referred to as DNA-PK. In addition, DNA-PK and Artemis form a complex that processes the broken DNA strand ends, preparing them for ligation. The broken ends are then ligated by a complex formed by XRCC4 and DNA ligase IV.

KU is one of the most abundant proteins in human cells (Chai et al., 2002), and it also forms a complex with the WRN protein (whose defect causes Werner's syndrome), and

with PARP-1 (a protein implicated in the protection of genome integrity), suggesting that these proteins act together as "caretakers" of the genome integrity (Li et al., 2004).

Good evidence for the KU complex's role in ageing has been shown in several studies with mice knockouts, as follows. In experiments carried out around the late 1990's, $Ku80^{-/-}$ mice exhibited signs of premature ageing without significantly increased cancer (Vogel et al., 1999), (Zhu et al., 1996). Surprisingly (considering that Ku70 and Ku80 form a complex), $Ku70^{-/-}$ mice exhibited instead a significant incidence of thymic lymphoma (Li et al., 1998), (Gu et al., 1997). However, more recently, (Li et al., 2007) showed that those differences were likely due to differences in genetic background or environment, rather than to differences in functions of Ku70 or Ku80. Their experiments with three types of mice cohorts, consisting of $Ku70^{-/-}$, $Ku80^{-/-}$, and $Ku70^{-/-}$ /$Ku80^{-/-}$ double-mutant mice, showed that all these cohorts exhibit a premature ageing phenotype and lower cancer levels than previously reported for $Ku70^{-/-}$ mice.

However, a different type of phenotype is obtained when combining the deletion of Ku70 or Ku80 with the deletion of the well-known tumor suppressor p53 gene. Recently, (Li et al., 2009) have shown that, surprisingly, $Ku70^{-/-}/p53^{-/-}$ mice lived significantly longer than either $Ku80^{-/-}/p53^{-/-}$ mice or $Ku70^{-/-}/Ku80^{-/-}/p53^{-/-}$ triple mutant mice, due to a much lower incidence of pro-B-cell lymphoma in the former cohort. This confirms the well-known fact that DNA repair deficiency can lead to increased cancer or premature ageing, depending on the kind of repair deficiency.

There is also evidence that NHEJ activity is considerably reduced with age in rat cortical neurons (Sharma, 2007), (Vyjayanti and Rao, 2006). Note that this decreased NHEJ activity cannot be trivially explained as a consequence of a reduced number of double-strand breaks, because actually the number of single- and double-strand breaks in the neurons of rat cerebral cortex has been shown to considerably increase with age (Rao, 2007). In addition, genetic defects in the NHEJ pathway have been shown to reduce hematopoietic stem function in an age-dependent manner under conditions of stress in mice (Rossi et al., 2007).

Turning to studies directly related to human ageing, the level of mRNA expression of KU70 was observed to decrease considerably with age in human hematopoietic stem and progenitor cells (Prall et al., 2007).

The levels of expression of proteins KU70 and MRE11 were observed to significantly decline with age (Ju et al., 2006). In addition, KU70 expression was significantly higher in the longevity community than in the control community. These facts led the authors to suggest that KU70 expression in lymphocytes may be considered a biomarker of ageing.

Moreover, NHEJ has been observed to become less efficient and more error-prone in senescent human fibroblasts (Seluanov et al., 2007). The authors suggest this decline in NHEJ activity may be due to the following combination of reasons, based on their results: KU expression is reduced in senescent cells, KU has different location distributions in young and old cells (it has both nuclear and cytoplasmic location in young cells, but only nuclear location in old cells), and the nuclear KU in senescent cells is not available for new transactions.

### 1.5.4 Mismatch repair

This pathway repairs mismatched base pairs in the two strands of a DNA molecule. Such mismatched base pairs can be occasionally produced, for instance, by errors made by DNA polymerase during DNA replication. Although DNA replication is a very accurate process, it is not perfect, and MMR reduces the number of errors made during DNA replication by a factor of 100 (Alberts et al., 2002). MMR can also repair mismatched bases produced by recombination between imperfectly matched DNA sequences or deamination of 5-methylcytosine (Gorbunova et al., 2007). MMR is a strand-specific process, where bases incorrectly added to the new strand during DNA synthesis are removed and replaced by correct bases, using information from the complementary (parent) strand, which is used as a template.

The main steps of the MMR pathway are as follows (Li, 2008), (Hoeijmakers, 2001), (Alberts et al., 2002). First, the mismatch is recognized, by detecting the distortion in the

DNA helix that results from the mismatch between non-complementary base pairs. In *E. coli* this detection is done by a MutS dimer. In mammalian cells, MutS homologs (MSH) forms heterodimers hMSH2/hMSH6 and hMSH2/hMSH3, called hMutSα and hMutSβ, respectively. This process also involves interactions between replication factors and other "Mut" proteins. (The generic name "Mut" is due to the fact that these genes were first identified in *E. coli*, where their inactivation by mutation produces hypermutable strains.)

Then, the segment of DNA containing the mismatched base pairs is removed from the newly synthesized strand; and next the excised segment in the new strand is resynthesized by using the parent strand as a template. Several proteins are involved in these last two steps, including POLδ/ε, RPA (Replication Protein A), PCNA (Proliferating Cell Nuclear Antigen), exonuclease 1 and endonuclease FEN1.

Since MMR significantly improves the accuracy of DNA synthesis, defects in crucial MMR genes tend to significantly increase the predisposition to cancer (Alberts et al., 2002). In particular, in humans a MMR defect is the major cause of hereditary non-polyposis colorectal cancer (HNPCC). The MMR pathway is also involved in stabilizing trinucleotide repeats, and failure of this process, leading to expansion of those repeats, contributes to several human diseases, including Huntington's disease (Fishel et al., 2007), (Li, 2008). Knockout mouse models have been developed for several of the "Mut" homologs, and most of such knockout mice have a mutator phenotype and are cancer-prone (Li, 2008).

## 1.6 OBJECTIVES

Broadly speaking, the general objective of this thesis is to use classification algorithms – which are a kind of predictive data mining methods (see Sections 2.4 and 2.5) – to investigate the relationship between DNA repair genes and the process of ageing. More specifically, that objective involves two related goals, as follows.

The first goal is to use classification algorithms – applying them to datasets prepared specifically for the purpose of this research – to find gene properties that effectively discriminate between ageing-related DNA repair genes and other types of genes (mainly non-ageing-related DNA repair genes). The second goal is to analyse the predictive patterns (gene properties) discovered by the classification algorithms, from two perspectives: (a) measuring the predictive accuracy of those patterns, from a data mining and statistical perspective; and (b) interpreting the meaning of the main discovered patterns in the light of biological knowledge about DNA repair genes and the process of ageing.

# Chapter 2 – Bioinformatics and Data Mining

## 2.1 BIOLOGICAL DATABASES

In this section we review the biological databases that are most related to this research, as follows. Subsections 2.1.1 and 2.1.2 give an overview of databases specialized on ageing-related genes. Out of the databases reviewed in those two subsections, the most relevant one for this research is the GenAge database, which is reviewed in subsection 2.1.1. For a recent review of ageing-related databases and bioinformatics resources, see (Wieser et al., 2011). Subsection 2.1.3 gives an overview of UniProt, a comprehensive and richly annotated repository of protein sequence data – from where we extracted data to create a number of datasets to be discussed later in this thesis. Subsection 2.1.4 gives an overview of the Human Protein Reference Database, from where we extracted protein-protein interaction data used in this research.

### 2.1.1 GenAge

The GenAge database is part of the online Human Ageing Genomic Resources (HAGR) – see http://www.genomics.senescence.info/ – which includes both the GenAge and the AnAge databases (Magalhaes et al., 2005), (Magalhaes et al., 2009). The former contains information about ageing-related genes focusing on human ageing, whilst the latter contains integrated information about the ageing phenotype of a number of species. In the remainder of this section we focus only on GenAge, which was used in this research – unlike AnAge.

Originally the selection of genes to be included in GenAge was carefully done taking into account two factors, namely (Magalhaes et al., 2005): the selection focused on genes that influence the rate of ageing, rather than genes whose expression is influenced by the ageing process; and the focus was on genes that may affect the ageing process, rather than

simply preserving health. Hence, in general genes whose defects lead to, say, embryonic or early post-natal death are not included, despite their obvious effect on reduced life span.

More recently, however, GenAge was expanded to include human genes which seem associated with human longevity (Magalhaes et al., 2009) – even though these genes are not necessarily related to the ageing process per se. For instance, a gene allele may increase longevity by reducing the incidence of some disease(s) or making the organism more resistant to some disease(s), without affecting the ageing process. In any case, information about genes associated with human longevity in GenAge is located mainly at the web site: http://www.genomics.senescence.info/genes/longevity.html, which is separated from the information about human ageing-related genes – located at the web site: http://www.genomics.senescence.info/genes/human.html. Hence, researchers can easily focus on just one of these two types of genes if desired, depending on the research goals. In this research we focus on ageing-related genes only, since we are investigating DNA repair genes related to the ageing process.

It should be noted that, although GenAge focuses on human ageing-related genes, the selection of genes to be included in the database also considered genes which are believed to be ageing-related in model organisms if there was good evidence that the orthologues of such genes in humans may be related to ageing. As a result of a careful gene selection procedure considering many types of evidence for ageing-relatedness, when a gene is included in GenAge it is annotated with the main type of evidence used to select that gene. This annotation takes one out of seven possible textual values, as follows (http://www.genomics.senescence.info/genes/allgenes.php):

(a) evidence directly linking the gene product to ageing in humans;

(b) evidence directly linking the gene product to ageing in a mammalian model organism;

(c) evidence directly linking the gene product to ageing in a non-mammalian animal model;

(d) evidence directly linking the gene product to ageing in a cellular model system;

(e) evidence linking the gene product to the regulation or control of genes previously linked to ageing;

(f) evidence linking the gene product to a pathway or mechanism linked to ageing;

(g) indirect or inconclusive evidence linking the gene product to ageing or showing the gene product to be an effector of genes related to ageing;

These different types of evidence were used in this research to create different versions of a dataset of DNA repair genes, as will be discussed in detail in the next chapter.

### 2.1.2 Other ageing-related databases

This section briefly reviews some ageing-related databases other than GenAge, which were not used in this research. For further details about these other databases the reader is referred to the references cited below.

**NetAge** – The NetAge database contains information about genes and micro RNAs (miRNA)-regulated protein-protein interaction networks related to longevity, ageing-related diseases and ageing-associated processes (Tacutu et al., 2010). Hence, this database seems particularly useful for research involving the systems biology of ageing, where the focus is on analysing ageing-related networks, considering interactions between genes and related processes such as ageing-related diseases, rather than focusing on the analysis of individual genes. NetAge is a relatively recent project; the NetAge web site – http://netage-project.org/ – was launched in June 2009.

**The CISBAN Interactome Database (CID)** – This database (Taschuk et al., 2010) integrates ageing-related interaction data from several model organisms and several public databases with protein interaction data – including BioGrid, DIP, (Database of Interacting Proteins), HPRD (Human Protein Reference Database), IntAct, KEGG pathway database, MINT (Molecular INTeraction database), and MPact (Representation of Interaction Data at MIPS), as well as in-house data. At the time of writing (August

2010), according to the web site: http://cisban-silico.cs.ncl.ac.uk/cid.html, a plugin for Cytoscape is "currently in the alpha stages of development and subject to a very limited release".

**Gene Aging Nexus (GAN)** – This is a web-based platform with ageing-related data and data analysis tools. It includes a database containing ageing-related gene expression data from microarray experiments in six species, including *H. sapiens*. This database is available at the web site: http://gan.usc.edu. For the purpose of collecting microarray datasets to be included in this database, ageing-related data has been defined as datasets including "age" as a variable, and the database also contains data about ageing-related diseases such as Alzheimer's (Pan et al., 2007).

**AGEID** – This is a database containing information about experiments with ageing-related genes. As stated in (Kaeberlein et al., 2002), the goal is "to catalog, in one location, every published experiment where life span has been measured in any organism". Users can do a search either by gene or by intervention. Examples of interventions with multiple entries include caloric restrict and human growth hormone therapy. This database is available at the web site: http://uwaging.org/genesdb/.

### 2.1.3 Uniprot

Uniprot is a centralized repository of information about proteins, which was created in 2002 with the collaboration of the European Bioinformatics Institute (EBI), Protein Information Resources (PIR) and Swiss Institute of Bioinformatics (SIB). It is available at the web site: http://www.uniprot.org. Uniprot provides rich and accurate annotation to protein sequences, and it consists of four types of protein databases (Uniprot-Consortium, 2007), (Uniprot-Consortium, 2010), namely UniProtKB (the UniProt Knowledge Base), UniParc (Uniprot Archive), UniRef (UniProt Reference Clusters) and UniMES (UniProt Metagenomic and Environmental Sequences database). Out of these, the only one used in this research was the UniProtKB database.

UniProtKB is a curated database containing an extensive annotation for protein sequences. It consists of two parts, namely UniProtKB/SwissProt and UniProtKB/TrEMBL. The former contains annotations that were carefully curated by biologists with specific expertise in each type of protein being annotated. TrEMBL contains mainly records that were automatically annotated by have not yet been fully curated by expert biologists. Hence, the SwissProt part offers more reliable data than the TrEMBL part for data mining purposes; and in this research we used only protein sequence data from UniProtKB/SwissProt.

A detailed discussion about which information was extracted from this database, and how that information was used to create datasets specific to the goals of this research, will be presented in the next chapter.

### 2.1.4 HPRD (Human Protein Reference Database)

HPRD is a database specialized on information about the human proteome (Prasad et al., 2009), including protein-protein interaction data involving human proteins, which is the part of the database that was used in this research. This database is available at the web site: http://www.hprd.org.

HPRD contains protein-protein interaction data from multiple types of experiments, including in vivo, in vitro and high-throughput methods such as the yeast two hydrid method. Each protein-protein interaction record in HPRD is annotated with the type(s) of experiments where that interaction was detected, so researchers can use this information to select only interactions whose data are more reliable (say, interactions detected via in vivo and in vitro interactions). HPRD is carefully curated by expert biologists, to ensure a high level of accuracy of the annotated information for each record. To quote from the FAQ (Frequently Asked Questions) at the HPRD web site (http://www.hprd.org/FAQ):

“We are not using any literature mining program. Our trained biologists read and interpret each and every paper related to the molecule of interest.”

Again, a detailed discussion about which information was extracted from HPRD, and how that information was used to create datasets specific to the goals of this research, will be presented in the next chapter.

## 2.2 GENE ONTOLOGY (GO)

### 2.2.1 The motivation for the gene ontology

The Gene Ontology Consortium was created in 1998 with the basic goal of producing the Gene Ontology (GO). This goal and the corresponding definition of the GO were originally stated as follows (GO-Consortium, 2000) (p. 26):

"The goal of the Consortium is to produce a structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products in any organism."

The proposed controlled vocabulary is also dynamic, in the sense that it is continuously revised and updated by the GO Consortium as our knowledge of gene, gene products and their roles in cells keeps accumulating and changing. The GO also specifies well-defined types of relationships between different genes or gene products.

A major motivation for the development of the GO was the fast growth in the amount of data about genes and proteins stored in biological databases and the fast progress of sequencing technology. These factors contributed to the recognition that there is a limited universe of genes and proteins and a surprisingly large fraction of genes and proteins is shared between distantly-related species. This recognition has been a major driving force to develop the GO as a means to achieve a "grand unification" of biology (GO-Consortium, 2000).

The GO Consortium initially consisted of researchers and developers associated with three major model organism databases, namely the Flybase, the Mouse Genome

Informatics and The Saccharomyces Genome Database, but the number of participating databases in the GO Consortium has increased fast since its creation (Lewis, 2004).

The GO has important advantages. It specifies well-defined, common, controlled vocabulary and specific types of relationships for describing gene and protein functions. Hence, it improves the interoperability of genomic databases and provides a generic framework for gene or protein functional classification. In addition, it is a *pan-organism* classification, i.e., it can potentially be applied to all organisms, contributing to the unification of biology, as mentioned earlier.

### 2.2.2 The basic structure of the gene ontology

The controlled vocabulary specified by the GO Consortium includes three categories of terms, namely biological process, molecular function and cellular component, which are implemented as three independent ontologies (GO-Consortium, 2001), (GO-Consortium, 2004). Within each of these three ontologies, each GO term has a free text definition and a stable unique identifier.

A *biological process* term refers to a biological "objective" which is accomplished by one or more ordered assemblies of molecular functions. A process involves input, processing and output, where the processing often involves a chemical or physical transformation. An example of a broad (high-level) biological process term is "cell death", whilst an example of a more specific (lower-level) biological process is "programmed cell death".

A *molecular function* term specifies a gene product's specific function from a biochemical perspective – which can depend, for instance, on binding to some ligand; without specifying how that function is carried out or where it is carried out. An example of a broad molecular function is "enzyme", whilst an example of a more specific (lower-level) biological process is "alcohol dehydrogenase (NAD) activity". Note that GO molecular function terms represent biochemical activities rather than the physical molecules that perform those activities.

A *cellular component* term refers to a location in the cell where a gene product is active. The "location" in question can be a broad place such as the "nuclear membrane" or a very specific "place" like a given complex.

The GO consists of not only terms, but also relationships among terms (GO-Consortium, 2001). In particular, the GO contains relationships of child terms to parent terms, which can be of the type "is a" or of the type "part of". The "is a" relationship type denotes that the concept represented by a child (more specific) term is an instance of the concept represented by a parent (more generic) term. For instance, the term "single strand break repair" is an instance of the term "DNA repair". The "part of" relationship type denotes that the object represented by a child term is part of the concept represented by a parent term. For instance, the Golgi membrane (a child cellular component term) is part of the endomembrane system (its parent cellular component term).

In GO, a child term can have one or more parent terms. From a computer science perspective, terms can be thought of as nodes in a graph, and relationships as directed edges connecting nodes. Since no cycles are allowed in the GO graphs, each of the three separate ontologies is represented by a DAG – Direct Acyclic Graph.

The GO also includes an important rule to guarantee the consistency of GO term annotations to gene products, the so-called *True Path Rule* (GO-Consortium, 2001). This rule states that a path from a child term to all its parents and corresponding ancestors in a DAG must always be true, considering the relationships represented by the edges in that DAG. This means that, if a gene product is annotated with a given GO term in a given database, the gene product must be assumed to be (implicitly) annotated with all parents and ancestors of that term too, even though such parents and ancestors may not be explicitly annotated to that gene product in that database. This rule should be taken into account when performing data mining based on GO term annotations, as will be discussed in the next chapter.

## 2.3 ANALYSING AGEING-RELATED GENE OR PROTEIN NETWORKS

From a mathematical viewpoint, gene or protein interaction networks are usually modelled as graphs, where each node represents a given biological entity – typically a gene or protein – and an edge connecting two nodes represents a specific type of relationship or interaction – for instance, physical binding or genetic regulation – between the genes or proteins represented by those two nodes.

In the last few years there has been a growing interest in the analysis of gene or protein interaction networks (Panchenko and Przytycka, 2008). In this section we focus on ageing-related networks, i.e., networks where the genes or proteins in the network are ageing or longevity-related genes. As mentioned in subsection 2.1.1., the literature often makes a distinction between ageing-related and longevity-related genes or proteins (Magalhaes et al., 2009), (Wang et al., 2009). The former is believed to influence the ageing process and the latter is involved in extending life span (an effect often observed in centenarians) without influencing the ageing process per se – for instance, by making the organism more resistant to some disease(s). For the purposes of the discussion of ageing-related networks in this section, however, this distinction is not very relevant, since in this section our discussion is focused on network analysis issues which are independent from the specific genes or proteins being analyzed. Hence, the discussion in this section applies to both aging-related and longevity-related genes and proteins.

### 2.3.1 Types of interactions and reference organisms in ageing-related networks

Projects studying ageing-related networks vary widely concerning the types of interactions represented in the network edges and the reference organism(s) – i.e., the organism whose genes or proteins and corresponding interactions are represented in the network.

Concerning the type of reference organism, several projects focused on relatively simple model organisms like yeast (*S. cerevisiae*) (Barea and Bonatto, 2009), (Managbanag et

al., 2008), (Promislow, 2004) or the nematode worm (*C. elegans*) (Fortney et al., 2010), (Witten and Bonchev, 2007).

However, there are also several projects that focus on human genes or proteins, as follows. (Budovsky et al., 2007) focused on longevity-associated human genes and human genes that are homologs of longevity-associated genes in model organisms; and (Budovsky et al., 2009) extended that work to consider human cancer-associated genes and their homologs in model organisms. Another study involving the human interactome is reported in (Bergman et al., 2007), which focused specifically on human genes believed to be associated with longevity due to studies with Jewish centenarians.

(Hsu et al., 2008) focused on human genes that are targets of miRNA (microRNA) regulation. In essence, miRNAs are small (~22 nucleotides) non-coding RNAs that can repress gene expression post-transcriptionally by binding to the 3' untranslated regions (3' UTRs) of their target mRNAs, in a process called RNA interference (Liang and Li, 2007), (Wolfson et al., 2008). Although Hsu et al.'s work did not focus on ageing, miRNAs are relevant to the study of ageing and age-related diseases (Wolfson et al., 2008).

Concerning the type of interaction (or relationship) represented by the edges in the network, most studies have focused on a single type of interaction. The most commonly used type of interaction has been protein-protein interaction – used for instance in (Bell et al., 2009), (Bergman et al., 2007), (Budovsky et al., 2007), (Wang et al., 2009); but there is also work on interactions representing the degree of correlation between the expressions of two genes (Fortney et al., 2010).

In (Managbanag et al., 2008) initially a network containing edges representing several interactions – including physical binding, genetic relationships and post-translational modification – was considered. However, the data analysis activity focused on a simpler network with edges representing physical binding between proteins.

Relatively few studies have focused on more than one type of interactions in the same network, as follows. (Chautard et al., 2010) focused on both protein-protein interactions

and protein-glycosaminoglycan interactions – a type of interaction particularly relevant for the type of protein that was the focus of their study, namely extracellular matrix proteins. A more sophisticated work, in the sense of considering more types of interactions, is reported in (Witten and Bonchev, 2007). In the ageing-related network constructed in this work, a node can represent a gene or a protein, and different edges can represent different types of gene-gene, gene-protein or protein-protein interactions, including physical binding, transcriptional regulatory interactions or correlation of gene co-expression. Multiple types of interactions are also used in (Wang et al., 2009), where a node can represent a gene or a disease (constituting a disease-ageing network), and the edges in the network can represent relationships such as protein-protein interactions or gene-disease associations.

Considering multiple types of interactions in the network edges seems an under-explored and important research direction. After all, although the majority of the research in this area is focusing on protein-protein interactions, other types of interactions can be at least as important – or perhaps even more important – to understand the systems biology of ageing. For instance, as a result of the analysis of their longevity network considering multiple types of interactions between genes and proteins in *C. elegans*, (Witten and Bonchev, 2007) observed that in the core longevity network the interactions are predominantly genetic regulations, rather than protein-protein interactions.

In addition, it should be noted that a network of protein-protein interactions usually is represented as a static network, but in reality such interactions are dynamic, and the actual interactions that happen *in vivo* depend on several factors such as the current levels of gene expression and the location of the proteins (Liang and Li, 2007).

Yet another limitation of current projects involving a computational analysis of protein-protein interactions in ageing-related networks is that such interactions are usually assumed to be "all-or-nothing", i.e., either two proteins interact or not; whilst in reality different pairs of proteins can interact to different degrees of strength – which can be characterized by the dissociation constants of individual interactions (Maslov, 2008). It

would be interesting to represent the strength of protein interactions as numerical weights in network edges, for instance by associating those weights with dissociation constants, as suggested – as future research – in (Chautard et al., 2010).

A summary of the types of reference organism and types of interactions (represented by edges in the network) used in the ageing-related networks discussed in this subsection is shown in Table 2.1.

**Table 2.1**: Summary of types of reference organism and types of interactions in ageing-related networks

| Work | Reference Organism | Type of Interaction |
|---|---|---|
| (Barea and Bonatto, 2009) | Yeast | protein-protein interaction |
| (Bell et al., 2009) | human and model organisms | protein-protein interaction |
| (Bergman et al., 2007) | human | protein-protein interaction |
| (Budovsky et al., 2007) | human and model organisms | protein-protein interaction |
| (Budovsky et al., 2009) | human and model organisms | protein-protein interaction |
| (Chautard et al., 2010) | human and other mammalian organisms | protein-protein interaction, protein-glycosaminoglycan interaction |
| (Fortney et al., 2010) | nematode worm | gene expression correlation |
| (Hsu et al., 2008) | human (focus on miRNA targets) | protein-protein interaction |
| (Managbanag et al., 2008) | yeast | protein-protein interaction |
| (Promislow, 2004) | yeast | protein-protein interaction |
| (Wang et al., 2009) | human | protein-protein interaction, gene-disease associations |
| (Witten and Bonchev, 2007) | nematode worm | multiple types of gene or protein interactions |

### 2.3.2 Analysing ageing-related gene or protein networks

There are several different types of data analysis that can be applied to a given ageing-related network. One of the most common types of analysis consists of identifying hubs in the target network. Hubs are essentially nodes that have a large degree – i.e., a large number of neighbouring nodes.

Several studies have reported that network nodes representing ageing-related genes tend to have a significantly higher degree than expected by chance – i.e., they tend to be hubs (Bell et al., 2009), (Budovsky et al., 2007), (Ferrarini et al., 2005), (Promislow, 2004), (Witten and Bonchev, 2007). The identification of hubs can be used for predicting new ageing-related genes, based on the observed fact that ageing-related genes are significanltly more likely to be hubs than randomly-selected genes (Witten and Bonchev, 2007). In addition, genes identified as hubs in ageing-related networks have the potential to be the targets of interventions aimed at extending life span (Bell et al., 2009). However, such potential interventions should consider the fact that many such hubs may not be considered ageing-related genes because defects in those genes may lead to embryonic or early post-natal death, since genes corresponding to hubs tend to be involved in several important biological processes (Budovsky et al., 2007).

On the other hand, genes corresponding to hubs may have pleiotropic effects, contributing to the individual's healthy during its development and adult phases but contributing to age-related diseases later in life – a possibility suggested by the antagonistic pleiotropy theory of ageing (Section 1.2). Consistently with this possibility, (Promislow, 2004) has shown that indeed hubs tend to have a significantly higher degree of pleiotropy than expected by chance. (In that work the degree of pleiotropy of a protein was simply defined as the number of different functional classes assigned to that protein in the "FunCat" functional classification scheme of the MIPS database.) This suggests that it might be possible to perform some interventions that extend life span by carefully modulating a hub gene's expression or its interaction with other genes or proteins in a selective way, starting in adult age (to avoid interfering with a hub gene's crucial function

during development), reinforcing the gene's activity in longevity extension pathways and decreasing the gene's activity in ageing-related pathways (Budovsky et al., 2007).

In any case, it should be noted that most studies do not present a precise definition of a hub, in the sense that they do not specify the minimal number of connections that a node (gene or protein) should have, in order to be considered a hub. An exception is (Chautard et al., 2010), where a hub is defined as a node with at least 20 neighbours in the network, where the value 20 seems an ad-hoc user-defined parameter. Intuitively, the setting of this parameter will have a considerable influence in the result of any network analysis based on identifying hubs; therefore this parameter should be carefully set in the context of the underlying data. For instance, in their ageing-related network analysis, (Barea and Bonatto, 2009) report that hubs had a mean node degree of 18.8. If the parameter value 20 were used in this work (as it was in (Chautard et al., 2010)), the set of hubs would be considerably smaller. Hence, in future studies it might be useful for researchers to report the results of their analysis for different values of this parameter.

In general the studies mentioned above focus on the identification of hub genes by taking into account only the degree (number of edges) of their corresponding nodes in the network. Arguably, however, another important characteristic of hubs is related to their clustering coefficient (Liang and Li, 2007). The clustering coefficient of a given gene *g* can be defined as the ratio of the number of edges in the network connecting genes that are neighbours of *g* over the maximum possible number of edges connecting genes that are neighbours of *g*. Liang & Li point out that genes or proteins with a high clustering coefficient tend to be *intra-modular hubs*, i.e. hubs in a given module or subgraph of the network (normally assumed to be a functional module), whilst genes or proteins with a low clustering coefficient tend to be *inter-modular hubs*, i.e. genes or proteins that coordinate different functional modules.

In addition to identifying hubs, another strategy to identify genes potentially relevant for ageing consists of identifying nodes with a high degree of centrality in an ageing-related gene or protein network. Intuitively, the more central the position of a node in a network,

the larger the number of paths connecting other nodes that pass through that node, and so the greater the relevance of that gene for the ageing process. Several alternative quantitative definitions of the centrality of a node (gene or protein) in a network are discussed in (Witten and Bonchev, 2007).

As an example of this kind of definition, let *dist(g)* be the average distance (in terms of number of edges) between the gene *g* and each of all the other genes in the network. Then *closeness centrality* has been defined as the reciprocal of that distance, i.e. 1/*dist(g)*. This definition of centrality has been shown in (Witten and Bonchev, 2007) to produce large values of centrality for some genes that are well-known to be ageing-related or longevity-related genes.

The measure of closeness centrality has also been used in (Wang et al., 2009). This work involved a disease-ageing network (DAN) of human protein-protein interactions, where a node can be a disease-related or an ageing-related gene. This work has reported that ageing-related genes tend to have much larger values of closeness centrality than disease-related genes.

Another form of analysing the data in an ageing-related network consists of detecting Gene Ontology (GO) terms that are overrepresented in the network or in a specific part of the network. Some examples of this type of network analysis are briefly mentioned next.

In (Chautard et al., 2010), where the ageing-related network consists of interactions between extracellular molecules (containing both protein-protein interactions and protein-glycosaminoglycan interactions), biological process GO terms overrepresented in the network include, for instance, coagulation, response to wounding, response to external stimulus, and response to stress.

In (Wang et al., 2009), where the ageing-related network consisted of interactions between ageing-related and disease-related genes, GO terms overrepresented in the disease-related genes in the network include “nucleobase, nucleoside, nucleotide and nucleic acid metabolic process” and “transcription regulator activity”.

(Fortney et al., 2010) identified 37 biological process GO terms that are significantly enriched for the genes of modular subnetworks of a network of gene expression data in *C. elegans*. Out of those 37 terms, 12 were significantly enriched for longevity genes. The set of ageing-associated GO terms identified in this analysis includes both some obvious terms, such as "determination of adult life span" and some more interesting, less expected terms such as "locomotory behaviour", which has been proposed as a biomarker of physiological ageing in *C. elegans* (Golden et al., 2008).

(Barea and Bonatto, 2009) identified several subnetworks in their yeast ageing-related network that are enriched for a number of GO terms. Of particularly relevance to this thesis, the set of GO terms enriched in those subnetworks include several terms related to DNA, such as "DNA packing", "DNA repair", "double strand break repair via NHEJ", "double strand break repair via homologous recombination" and "response to DNA damage stimulus".

**Table 2.2**: Summary of major types of network analysis in ageing-related networks

| Work | Network Analysis |
|---|---|
| (Barea and Bonatto, 2009) | Identification of GO terms enriched in subnetworks |
| (Bell et al., 2009) | Identification of hubs |
| (Budovsky et al., 2007) | Identification of hubs and their broad biological functions |
| (Budovsky et al., 2009) | Identification of hubs and their broad biological functions |
| (Chautard et al., 2010) | Identification of hubs and overrepresented GO terms |
| (Fortney et al., 2010) | Identification of GO terms enriched in subnetworks |
| (Promislow, 2004) | Identification of hubs and their degree of pleiotropy |
| (Wang et al., 2009) | Identification of central nodes and overrepresented GO terms |
| (Witten and Bonchev, 2007) | Identification of hubs and central nodes |

There are also projects that identified broad biological functions that were associated with several hubs in ageing-related networks of protein-protein interactions, without referring to the use of GO terms to describe those functions. In particular, (Budovsky et al., 2007) and (Budovsky et al., 2009) reported that several hubs in their ageing-related networks represent genes or proteins involved in signal transduction and transcription.

Table 2.2 shows a summary of the discussion of network analysis presented this section; mentioning, for each work, the major type of network analysis carried out in that work.

## 2.4 CONCEPTS AND PRINCIPLES OF DATA MINING

### 2.4.1 Basic concepts of data mining

Data mining essentially consists of concepts and methods to extract interesting knowledge (or patterns) from real-world datasets (Fayyad et al., 1996), (Witten and Frank, 2005). Data mining is a broad research area at the intersection of several fields, including machine learning (often considered a sub-area of artificial intelligence), statistics (particularly statistical pattern recognition), databases (a sub-area of computer science) and data visualisation. In this thesis we focus on data mining from a machine learning perspective (Witten and Frank, 2005), so that the concepts and methods discussed and used in this thesis are mainly derived from machine learning. Hence, in this thesis the terms data mining and machine learning are used interchangeably.

From a conceptual viewpoint, it is important to distinguish data mining *tasks* from data mining *methods*. A data mining task is a specific kind of problem, associated with a certain kind of structure of the data to be mined and a certain kind of knowledge to be extracted from that type of data. A data mining method is a computational method – specified by an algorithm – which is used to solve a particular data mining task, i.e. to extract knowledge from the type of dataset associated with the target task. Each data mining task can be solved by many different types of data mining methods, but each

specific type of data mining method is usually aimed at solving a specific data mining task.

Hence, the definition of a data mining task can be considered more fundamental than the definition of a data mining method. Therefore, before we discuss the data mining methods used in this research, we present, in the next subsection, an overview of the data mining task addressed in this research, namely classification – also called supervised learning in the machine learning literature.

### 2.4.2 The classification task of data mining

The classification task is characterized by a specific type of dataset to be mined and a specific type of knowledge to be extracted from that type of dataset (Witten and Frank, 2005), (Tan et al., 2005), (Freitas, 2002). Let us start with the structure of a dataset for classification. In this task, a dataset consists of a set of data instances (sometimes called examples). Each data instance can be thought of as a record in a file or a row in a spreadsheet. Each data instance is composed by a set of attributes (sometimes called features), and each attribute can take a value from its domain, i.e., the set of valid values for that attribute. A key characteristic of the classification task is that an instance is represented by two types of attributes, namely a set of predictor attributes (with potentially a large number of attributes) and a single special attribute, called the class attribute. The domain of the class attribute consists of the classes to be predicted by a classification algorithm. The class attribute is categorical (or nominal) – i.e., its domain consists of a set of unordered values. The other attributes, called predictor attributes (since their values will be used to predict the value of the class attribute, as discussed later), can be either categorical or continuous (real-valued) – i.e., consisting of an ordered list of numerical values.

Let us consider a hypothetical example of a classification dataset to clarify these concepts, in the application domain of the diagnosis of cancer. One might have a dataset, derived from a hospital's database, where each data instance corresponds to a patient. The predictor attributes describing that instance would be characteristics of the patient or

results of medical tests applied to the patient, each attribute with a predefined domain. For example, one attribute could be the age of the patient, whose domain would consist of integer numbers, and another attribute could be the result of a certain type of blood test, whose domain would consist of a set of possible results for that test. The class attribute could be defined as a binary variable, whose domain would be simply "yes" or "no" representing whether or not the patient has cancer. Alternatively, the class attribute could have a domain consisting of a set of values, corresponding to different types of cancer. The values of the class attribute are simply referred to as classes.

Another key characteristic of the classification task is that, for the purposes of evaluating the performance of a classification algorithm, the available dataset is divided into two parts – with no overlapping of data instances between them – namely the training set and the test set (Witten and Frank, 2005), (Freitas, 2002). The *training* set consists of data instances for which the classification algorithm has access to both the values of the predictor attributes and the values of the class attribute. The *test* set consists of data instances for which the classification algorithm has access to the values of the predictor attributes but not to the values of the class attribute. This crucial difference between the training and test sets is illustrated in Figure 2.1.

A classification algorithm analyses the relationship between the values of the predictor attributes and the classes in all instances of the training set, and builds a classification model of the data. This model represents the classification knowledge (or patterns) extracted from the data – i.e., a relationship between predictor attributes values and classes – that can be used to predict the unknown class of a data instance given the observed values of its predictor attributes.

Once a classification model has been built from the training set, that model can then be evaluated on the test set, whose class values are unknown by the algorithm – but are known by the user or data analyst. More precisely, to evaluate the predictive performance of the model, for each data instance in the test set, the model is applied, using as input to the model that instance's predictor attribute values and producing as output a predicted

class. That class is then compared with the actual class of the instance, to determine if the prediction was correct or wrong. In essence, the predictive accuracy of the model is then computed taking into account the total number of correct and wrong predictions over all instances of the test set. (There are many different measures of predictive accuracy proposed in the literature (Caruana and Niculescu-Mizil, 2004), the one used in this thesis is described in the next chapter.)

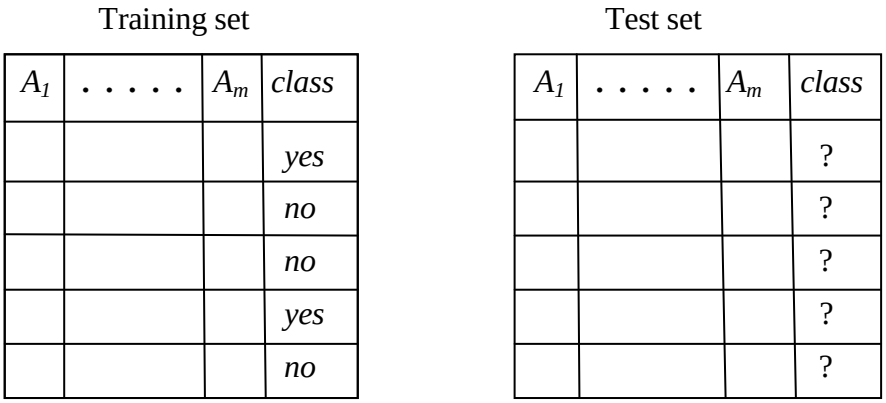


**Figure 2.1:** Basic difference between training set and test set in the classification task. $A_1$...$A_m$ denote predictor attributes, where $m$ is the number of attributes. Each row represents a data instance. The attribute values in the data instances are not shown to keep the figure simple. Note that in the training set the class of each data instance is known, whilst the class (*yes* or *no*) of each data instance in the test set is unknown – denoted by "?" – by the algorithm.

It is important to emphasize that predictive accuracy is measured on the *test* set, containing data instances which were *not* included in the training set, and so a measure of predictive accuracy is a measure of the generalization ability of the classification model built by a classification algorithm. A major goal of a classification algorithm is to discover a classification model that maximizes predictive accuracy on the test set. Classification accuracy in the training set is too easy to be maximized, since the algorithm has access to the class values of all instances in the training set, and so classification accuracy in the training set is not a valid measure of predictive accuracy (Witten and Frank, 2005).

#### *2.4.2.1 Overfitting and underfitting*

Another important aspect of the classification task, which is associated with its predictive nature, is the possibility that a classification model may overfit or underfit the data (Witten and Frank, 2005), (Tan et al., 2005). Overfitting is the phenomenon where a classification model is "too adapted" to the training set, in the sense that the model is capturing some relationships or patterns in the data that are specific to details of the training set and do not generalize well to the test set. Underfitting is the converse of overfitting, i.e., it is the phenomenon where a classification model is "under-adapted" to the training set, in the sense that the model is capturing only generic relationships or patterns in the training set, missing more specific relationships or patterns that would still generalize well to the test set.

There is a delicate trade-off between the chances of overfitting and underfitting when a classification algorithm is building a classification model. Hence, in principle every classification algorithm has to cope with this trade-off in some way, doing its best to achieve a good balance between the goals of minimizing both the chance of overfitting and the chance of underfitting.

#### *2.4.2.2 Classification versus clustering*

The classification task (supervised learning) is sometimes confused with the clustering task (*un*supervised learning) in the bioinformatics literature, even though the two tasks are fundamentally different from a data mining perspective. A relevant example of this confusion in the literature is discussed in (Kell and King, 2000), which points out the following mismatch between data mining tasks and methods – mentioning that clustering methods are often unduly applied to classification problems:

"...functional genomics is, in part, an exercise in pattern classification. Because many genes have known functional classes, the problem of predicting their functional class is a supervised learning problem. However, most pattern classification methods that have been applied to the problem have been unsupervised clustering methods. Consequently, the best classification tools have not always been used."

Hence, it is worth briefly discussing here the main differences between the classification and clustering tasks (Fayyad et al., 1996), (Tan et al., 2005).

First of all, whilst the central goal of the classification task is prediction, the clustering task does not involve prediction. Rather, clustering is a data mining task where the goal is essentially to partition the set of data instances into clusters in such way that each cluster contains data instances that are very similar to each other – i.e., have similar attribute values – and the data instances in each cluster are very different from the data instances in other clusters. The partitioning of the data instances into clusters can be regarded as a form of identifying previously-unknown patterns or structure in the data and describing those patterns or structure in a comprehensible form – since it is easier to describe a relatively small set of clusters than to describe a large number of unclustered data instances. However, the clusters produced by a clustering algorithm are not supposed to be used for prediction purposes, and there is no expectation that the clusters produced by a clustering algorithm will "generalize" well to another set of data instances, as it is the case in classification.

Furthermore, in the clustering task, data instances are not assigned to pre-defined classes, i.e., there is no special class attribute. Also, the attributes describing an instance are not referred to as "predictor attributes", since there is no class to be predicted; they are simply referred to as "attributes". In addition, since clustering does not involve prediction, there is no need to partition the dataset being mined into a training set and a test set as it is done in classification – where that dataset partitioning is used to evaluate the predictive accuracy of a classification model built from the training set. A clustering algorithm uses all the available dataset to find clusters in the data.

In conclusion, when the dataset to be mined consists of data instances that are assigned to pre-defined classes, there is a clear distinction between predictor attributes and the class attribute (values of the former are used to predict the value of the latter), and the goal is to build a classification model for prediction purposes (a model with a good generalization ability in another set of data instances unseen during training), the data mining algorithm

to be used should be a classification algorithm. It would be a serious mistake, from a data mining perspective, to use a clustering algorithm to solve a classification problem.

## 2.5 CLASSIFICATION METHODS USED IN THIS RESEARCH

### 2.5.1 Decision tree induction

From a computer science perspective, a tree is a type of graph commonly used as a data structure, consisting of a set of *n* nodes connected by *n* – 1 edges in such a way that each node (with the exception of the root node) is connected to exactly one "parent" node. A tree node is called a leaf node if it has no "child" node, and called an internal node otherwise. A decision tree is a graphical classification model represented in the form of a tree, with the following main characteristics (Quinlan, 1993): every internal node is associated with a predictor attribute; each branch (edge) coming out from an internal node is associated with one or more values of the attribute in that node; and every leaf node is associated with a class.

A very simple example of a decision tree is shown in Figure 2.2, referring to a hypothetical dataset where *Smoking Level* and *Age* are two predictor attributes; *yes* and *no* are classes representing whether or not a patient is classified as having lung cancer.

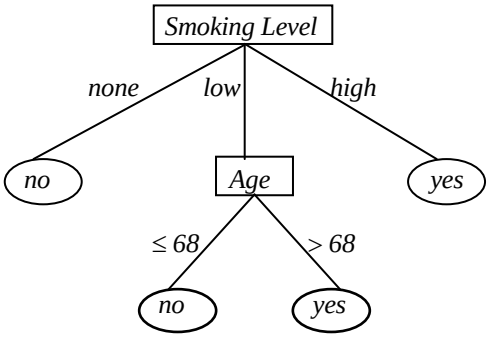


**Figure 2.2:** A very simple example of a decision tree to predict lung cancer

In order to classify a new data instance (whose class is unknown) using a decision tree, a top-down classification procedure is used. In this procedure, the data instance to be classified is passed downwards the tree (starting from the root node), following the branches whose attribute values match the instance's attribute values until the instance reaches a leaf node. Then the class label at that leaf node is assigned (as the predicted class) to that instance. Considering as an example the very simple decision tree in Figure 2.2, to classify a new data instance (representing a patient) the system would first check the value of the *Smoking Level* attribute in that instance. If the value is *none* the decision tree predicts no lung cancer for that patient. If the value is *high* the tree predicts lung cancer. If the value is *low* the system needs to check the value of another attribute of that instance, namely *Age*, before finally making a prediction – with lung cancer being predicted only when the patient's age is greater than 68 years old.

Note that a decision tree algorithm selects only the most relevant predictor attributes to be included in the decision tree, and many attributes present in the dataset being mined may not appear in the tree at all because they are considered, by the algorithm, as irrelevant for class prediction. In addition, for the attributes which are selected to be included (as labels of internal nodes) in the decision tree, in general, the closer to the root the attribute is, the more relevant for class prediction it is. In particular, the attribute at the root node will be used to classify *every* example in the test set – since, given the top-down nature of the classification, every example has to be submitted to the attribute value test at the root node. In contrast, an attribute that occurs only in a deep level of the tree will probably be used to classify just a relatively small number of test examples, namely only those examples whose "classification path" from the root to a leaf node pass through the node of that particular attribute.

It is important to note that a decision tree with $l$ leaf nodes can be easily converted into another kind of classification model consisting of a set of $l$ IF-THEN classification rules (Quinlan, 1993), where each path in the tree from the root node to a leaf node corresponds to a rule. Each rule contains, in its IF part – called the rule antecedent – a logical conjunction of conditions (i.e. conditions connected by the logical operator *AND*)

corresponding to the attribute values present in the path used to create that rule; and the rule contains, in its THEN part – the rule consequent – the predicted class for any data instance satisfying the conditions in the rule antecedent. For example, the decision tree shown in Figure 2.2 can be converted into the set of classification rules shown in Figure 2.3.

It should also be noted, however, that the conversion of a decision tree into a rule set is not the only approach to produce a set of classification rules. There are many classification algorithms that build a set of classification rules directly from the data, without producing a decision tree first. Such algorithms are called rule induction algorithms, and they are discussed in detail in (Furnkranz, 1999), (Witten and Frank, 2005).

IF (*Smoking Level = none*) THEN (*Lung Cancer = no*)
IF (*Smoking Level = low*) AND (*Age* $\leq$ *68*) THEN (*Lung Cancer = no*)
IF (*Smoking Level = low*) AND (*Age* > *68*) THEN (*Lung Cancer = yes*)
IF (*Smoking Level = high*) THEN (*Lung Cancer = yes*)
**Figure 2.3:** Rule set corresponding to the decision tree shown in Figure 2.2

So far we have discussed how a decision tree is used as a classification model. Let us now give an overview of the method used to automatically build a decision tree from a given training set – i.e. a dataset where both the attribute values and classes for all data instances are known, as discussed in Section 2.4.2. Decision tree induction algorithms typically build a tree in an iterative fashion, by adding one node at a time to the tree, starting with the root node. More precisely, a typical decision tree induction algorithm essentially works as follows (Quinlan, 1993), (Freitas, 2002), (Witten and Frank, 2005).

The algorithm starts by creating a tree with a single node, the root node, and all data instances in the training set are assigned to that node. If a certain stopping criterion is satisfied (see below) the tree building process is stopped, and that node is considered a leaf node, labelled with the most frequent class in the set of instances in the training set.

Otherwise the algorithm proceeds with the tree building process, by selecting a partitioning attribute and partitioning the set of instances at the current node according to the values of the selected attribute. In this attribute selection step, the goal is to select a partitioning attribute that best separates the classes, so that instances with different values of the attribute tend to belong to different classes. The ability of a candidate attribute in separating the classes is measured by a specific formula – many such attribute evaluation measures have been proposed in the literature, see for example (Quinlan, 1993), (Rokach and Maimon, 2005). Different attribute evaluation measures have different biases (White and Liu, 1994), so that there is no single measure which is universally "the best" for all datasets and the relative effectiveness of any given measure is dependent on the data being mined.

This attribute selection step also includes the choice of a test, over the value of the selected attribute, which produces mutually exclusive and collectively exhaustive outcomes. When the selected attribute is categorical the test usually consists of one outcome for each possible value of the attribute; whilst when the selected attribute is continuous the test usually consists of a binary test producing outcomes of the form $value \leq ths$ and $value > ths$, where *ths* is a threshold automatically chosen by the algorithm to maximize class separation.

Once a partitioning attribute and a corresponding test over its values have been chosen, the next steps of the algorithm are relatively simple. The algorithm labels the current tree node with the name of the selected attribute and creates one branch coming out from that node for each outcome of the chosen attribute value test. Those branches effectively partition the set of data instances in that node, so that each data instance is allocated to the branch whose attribute value test outcome matches the instance's attribute value. For instance, in the simple example of Figure 2.2, after selecting *Smoking Level* as the attribute for the root node the training set is partitioned into three subsets, each of them containing only instances with the *Smoking Level* value in the corresponding just-created branch. Then the decision tree induction algorithm is recursively applied to each of the

subsets produced by this partitioning procedure – i.e. to each of the newly created branches.

Concerning the stopping criterion of the algorithm, a natural criterion is that the tree-expansion process should be stopped when all the instances in the current root node belong to the same class, so that there is obviously no need to select any partitioning attribute to discriminate among classes in that node. In practice, however, this criterion tends to be too strict and conservative, and it is common to use less strict stopping criteria, such as stopping the tree expansion when the number of instances in the current tree node is smaller than a small number (a predefined threshold) – in which case there is not enough data to reliably choose a partitioning attribute, from a statistical perspective. Several stopping criteria are discussed, for example, in (Rokach and Maimon, 2005).

It should be noted that in general decision tree induction algorithms are not guaranteed to find the optimal decision tree because the number of all candidate decision trees is too large in practice. Hence, the previously described iterative procedure for building a decision tree, consisting of adding one node at a time to the current partial tree, is by far the most used procedure, since it is relatively fast and tends to build very good (though not optimal) decision trees in many cases.

Among the large number of decision tree induction algorithms proposed in the literature, the two most-well known ones, which were seminal works in the field, are C4.5 (Quinlan, 1993) and CART (Breiman et al., 1984). A comprehensive survey of decision tree algorithms can be found in (Rokach and Maimon, 2005). The decision tree induction algorithms used in this research are J4.8 (an implementation of the C4.5 algorithm) and CART – as implemented in the freely available WEKA data mining tool (Witten and Frank, 2005), version 3.6.1.

### 2.5.2 Naive Bayes

The Naive Bayes algorithm is a direct application of Bayes' theorem (Swinburne, 2002) to classification (Witten and Frank, 2005). In essence, it assigns to a data instance the class *k* that maximises the value of the following product of probabilities:

$$P(A_1|C_k) \times P(A_2|C_k) \times \ldots P(A_m|C_k) \times P(C_k), \qquad (2.1)$$

where $P(A_i|C_k)$ – $i = 1,...,m$ – is the empirical conditional probability of the value of attribute $A_i$ in the current data instance given that the instance belongs to class *k* (i.e., the number of training data instances having that value of attribute $A_i$ and having class *k* divided by the number of training data instances having class *k*), *m* is the number of predictor attributes, and $P(C_k)$ is the empirical prior probability of class *k* (i.e. the relative frequency of class *k* in the training set).

As an example of how equation (2.1) can be used to predict the class of a given data instance, consider a very simple hypothetical medical diagnosis dataset where the class attribute, called *Disease (D)*, denotes whether or not a patient has a certain disease (class labels: *yes* or *no*), and there are two predictor attributes: *Blood Test (BT)*, whose values (representing possible test outcomes for a patient) are *positive (+ve)* or *negative (–ve)*; and *Fever (F)*, whose values are *yes* or *no* (representing whether or not the patient has fever). Suppose we want to predict the class of a new patient, for whom we know that the result of the blood test was positive and the patient does not have fever. Suppose the required probabilities for this classification – computed from the training set by the Naive Bayes algorithm – are as follows:

$P(BT = +ve \mid D = yes) = 0.8$; $P(BT = +ve \mid D = no) = 0.1$

$P(F = no \mid D = yes) = 0.3$; $P(F = no \mid D = no) = 0.9$

$P(D = yes) = 0.1$; $P(D = no) = 0.9$

Given the above probabilities, Naive Bayes computes the value of equation (2.1) for each class *k* ($D = yes$ or $D = no$) as follows:

For class $D = yes$: $P(BT = +ve \mid D = yes) \times P(F = no \mid D = yes) \times P(D = yes)$

$= 0.8 \times 0.3 \times 0.1 = 0.024$

For class $D = no$: $P(BT = +ve \mid D = no) \times P(F = no \mid D = no) \times P(D = no)$

$= 0.1 \times 0.9 \times 0.9 = 0.081$

As a result, in this hypothetical example the new patient would be assigned to the class "*Disease = no*", since 0.081 > 0.024.

The "naive" part of the algorithm's name refers to the fact that the computation of the product of terms $P(A_1|C_k) \times P(A_2|C_k) \times \ldots P(A_m|C_k)$ makes the simplifying assumption that the attributes $A_1 \ldots A_m$ are independent from each other given the class. However, the use of the term "naive" can be considered a misnomer, because, although this simplifying assumption is not true in many cases, the algorithm still performs robustly well in practice, and it remains a popular data mining algorithm. It should also be noted that, intuitively, more complex variants of Bayesian classifiers, which try to detect dependences among attributes (rather than making the above simplifying assumption) would tend to lead to overfitting (subsection 2.4.2.1) in small datasets, such as the datasets mined in this research – which will be described in the next chapter.

The Naive Bayes algorithm used in this research is the version implemented in the freely available WEKA data mining tool (Witten and Frank, 2005) – version 3.6.1.

## 2.6 RELATED WORK ON PREDICTING PROTEIN FUNCTION WITH CLASSIFICATION METHODS

To the best of our knowledge, there is no published work on the use of classification algorithms to discriminate between ageing-related and non-ageing-related DNA repair genes (as investigated in this research), and there is just one recently-published work using classification algorithms to discriminate between ageing-related and non-ageing-

related genes in general, without any focus on DNA repair genes (Li et al., 2010). This latter work will be discussed at the end of this section.

However, the broader topic of using classification algorithms to predict protein functions (without any focus on ageing) has been investigated by many authors. For a comprehensive review of research in this area, the reader is referred to surveys or reviews such as (Zhao et al., 2008), (Friedberg, 2006), (Rost et al., 2003). Here we briefly review some related work on protein function prediction using classification algorithms that produce comprehensible classification models in the form of decision trees or sets of classification rules (the same type of model analysed in our computational experiments to be described in the next chapter), to show the potential that such comprehensible classification models have to represent new knowledge or patterns about protein functions extracted from protein data. The following discussion is partly based on (Freitas et al., 2010).

(Clare and King, 2001) applied the well-known C4.5 decision-tree induction algorithm (Quinlan, 1993) to data about mutant phenotype growth experiments with *S. cerivisiae*. The predictor attributes used by the algorithm had values representing information about the observed sensitivity or resistance of the mutants to different growth media, by comparison with the wild type. The classes predicted by the algorithm represented protein functional classes defined in the MIPS functional classification scheme (http://www.helmholtz-muenchen.de/en/mips/projects/funcat). The algorithm discovered many IF-THEN classification rules (extracted from the decision tree) that had just one or two conditions in its IF part and had a good predictive accuracy. These rules were simple to interpret from a biological perspective, and they identified the most relevant attributes (growth media) for predicting different functional classes of mutants.

(Clare and King, 2003) performed other experiments with the C4.5 algorithm applied to *S. cerivisiae* data, using a wider variety of types of predictor attributes to predict MIPS functional classes. The algorithm found, again, several IF-THEN classification rules which were easy to interpret from a biological perspective. That set of rules included, for

instance, a rule that had – in its IF part – conditions based on attribute values related to the predicted secondary structure (lengths and relative positions of alpha, beta and coil parts of the structure) of each protein.

(Syed and Yona, 2009) used a mixture of decision trees to predict the functional classes of enzymes – defined by their Enzyme Commission (EC) code. The dataset contained a large set of 453 predictor attributes, involving structural and functional properties of proteins. These attributes represented properties that in general were easily interpretable by biologists. The authors observed that the built decision trees identified the attributes with most predictive power, corresponding to properties that can be used to define, in a concise form, protein families.

In the aforementioned projects, the decision trees or IF-THEN classification rules extracted from the decision trees referred to predictor attribute values representing gene or protein-related information at a high level of abstraction, which facilitated the interpretation of those rules or trees by biologists. The two references mentioned next differ considerably from the aforementioned references with respect to the level of abstraction of the reported discovered rules, as follows.

(He et al., 2006) used a combination of a decision-tree induction algorithm and a support vector machine algorithm (a type of classification algorithm reviewed in detail in (Cristianini and Shawe-Taylor, 2000)) to discover rules predicting transmembrane segments. In the set of discovered rules, the rule conditions involved information about the presence of specific amino acids in specific positions of the sequence. Due to the use of this type of information, the discovered rules can be considered as pieces of knowledge expressed at a relatively "low-level" of abstraction.

This kind of rule with information at a low-level of abstraction was also discovered by (Huang et al., 2007) using a decision tree-induction algorithm, with the major difference that in this work the rules predicted changes in protein stability as a result of different types of mutations, rather than predicting transmembrane segments.

Although the classification rules reported in (He et al., 2006) and (Huang et al., 2007) represent knowledge at a lower level of abstraction than the rules reported in other projects discussed in this section, the former were still found interpretable and useful by the authors. In particular, the IF part of those rules identified which specific amino acids in specific sequence positions produce the predictions in the THEN part of the rule, and therefore specific mutations in the sequence can be made in "wetlab" experiments to validate the computational predictions. That is, wetlab experiments can be carried out in a focused way, based on the information contained in the discovered rules.

The above discussion has focused on projects using decision tree (or rule) induction algorithms to discover classification models that were interpreted in the light of biological knowledge. However, it should be noted that there are many other projects where, although this type of algorithm was used, the focus was on proposing a new variant of the algorithm and evaluating it from a computer science perspective (mainly in terms of predictive) accuracy, rather than on evaluating the discovered model in the light of biological knowledge. Hence, the latter kind of work is not discussed here, and interested readers are referred to related work on this topic such as (Schietgat et al., 2010), (C.Vens et al., 2008), (Davies et al., 2007), (Otero et al., 2010), (Pappa and Freitas, 2009).

As mentioned earlier, the author is aware of just one recently-published work on the use of classification algorithms to discriminate between ageing-related and non-ageing related genes, as reported in (Li et al., 2010). The main differences and similarities between that work and the research reported in this thesis are as follows.

First of all, in (Li et al., 2010) the research focused on data about genes of the nematode worm *C. elegans*, whilst in this thesis the research focused on data about human genes. In both projects the list of ageing-related genes was obtained from the GenAge database (discussed in Subsection 2.1.1.) and in both projects the classification problem involved two classes, which can be broadly referred to as ageing-related and non-ageing related genes – more precisely, in Li et al.'s work the classes are referred to as longevity genes and genes not yet known to be involved in longevity regulation. A major difference in the

definition of the classes (in addition to the aforementioned difference in the underlying type of organisms) is that in Li et al.'s work the definitions of the longevity-related and non-longevity-related classes were independent from the function(s) of the genes, so that in each class genes with many different types of functions were included; whilst in this thesis the ageing-related and non-ageing-related classes included only DNA repair genes, due to our focus on this type of gene in the context of the DNA damage theory of ageing (discussed in Subsection 1.2.2).

In addition, the set of predictor attributes used in (Li et al., 2010) was more diverse than the set of predictor attributes used in this thesis. More precisely, in Li et al.'s work the attributes used included information about protein sequence length, sequence conservation, expression pattern, functional annotation (including GO terms), RNAi phenotype and several types of topological features computed in a network that was built by integrating physical and genetic interactions between genes or proteins. Some of these types of features were also used in this thesis (as will be discussed in detail in Chapter 3), including GO terms, some gene expression data and information about protein-protein interactions (PPIs). However, the information about PPIs used in Li et al.'s work was considerably more extensive than the information about PPIs used in this thesis, since in the former several types of topological features were computed and both physical and genetic interactions were considered. In contrast, in this thesis genetic interactions were not considered and the only PPI attributes used were the number of interaction partners of a protein and binary attributes indicating whether or not a protein interacts with each protein in a given set of proteins (for details, see Subsection 3.1.5). Interestingly, in both projects a high value of the number of interaction partners has been observed to be a good predictor of the class of ageing-related (or longevity-related) genes.

Furthermore, the two projects used different approaches to interpret the relative importance of different attributes for the prediction of the ageing-related or longevity-related class, as follows. In Li et al.'s work the main classification algorithm used for predictions was a support vector machine (SVM) (Cristianini and Shawe-Taylor, 2000). This type of algorithm is well-known for its good predictive power in general, but it has

the disadvantage of producing a classification model that cannot be directly interpreted by users. Hence, the analysis of the importance of different attributes for classification was performed by using attribute evaluation criteria that are independent from the SVM algorithm and that measure the predictive power of each attribute separately, without taking account interactions among different attributes. In contrast, in this thesis the interpretation of the relative importance of different attributes was performed by directly analysing the constructed decision trees. This has the advantage that the relative importance of each attribute is determined by taking into account attribute interactions – for instance, an attribute might have a good predictive power by itself, but it might be redundant with respect to another attribute which measures a similar property of the underlying gene and has a greater predictive power. This analysis has been made possible by the fact that decision tree induction algorithms (unlike SVM algorithms) represent the classification model in a form that tends to be easily interpretable by users (Freitas et al., 2010). Decision tree induction algorithms were also used in (Li et al., 2010), but no interpretation of the constructed decision trees was reported in that work.

# Chapter 3 – Dataset Creation and Experimental Set Up

In order to discover gene features that discriminate between ageing-related DNA repair genes and other types of genes, we created a number of datasets for the classification task of data mining. These datasets can be categorized into three broad categories, according to the definitions of classes and predictor attributes used in each category. The creation of the classes and predictor attributes for each category of datasets is detailed in the next three sections, 3.1 through 3.3.

Next, Section 3.4 discusses the measure of predictive accuracy that was used to evaluate the effectiveness of the classification algorithms used in this research – whose computational results will be presented in the next chapter. Finally, Section 3.5 discusses how the statistical significance of the classification rules reported in the next chapter was evaluated.

## 3.1 CREATING DATASETS WITH TWO CLASSES AND MULTIPLE ATTRIBUTE TYPES

### 3.1.1 Creating two classes: ageing-related vs. non-ageing-related DNA repair

In the first category of datasets created in this research, each data instance represents a DNA repair gene that belongs to one out of two classes, namely: "ageing-related" or "non-ageing-related" DNA repair gene. In this scenario, the procedure for creating data instances belonging to each of these two classes works as follows.

First, a set of DNA repair gene names was obtained from the web site: http://sciencepark.mdanderson.org/labs/wood/DNA_Repair_Genes.html (which is hereafter called "Wood's web site" for simplicity). The list of genes in this web site is an updated supplement to a review of DNA repair genes presented in (Wood et al., 2001), (Wood et al., 2005). Next, a set of ageing-related DNA repair gene names was obtained

from the list of human genes in the GenAge database (Magalhaes, 2009), at the web site: http://genomics.senescence.info/genes/.

We then computed the intersection between the sets of gene names obtained from Wood's web site and from GenAge. Each gene name occurring in the intersection of these two sets was used to create a data instance of the "positive class", i.e., the class of ageing-related DNA repair genes. Each gene name included in the set obtained from Wood's webpage but *not* included in GenAge was used to create a data instance of the "negative" class, i.e, the class of non-ageing-related DNA repair genes. Gene names included in GenAge but *not* included in the set of gene names in Wood's web site were ignored for the purpose of creating the dataset, since our focus is on DNA repair genes.

Note that, in this scenario, the ageing-related DNA repair gene class consists of all DNA repair genes included in GenAge, regardless of the reason for inclusion of the gene in GenAge. This allowed us to obtain a reasonable number of data instances belonging to this class (33 ageing-related DNA repair genes), although at the expenses of including, in the class definition, genes whose reason for inclusion in GenAge is not very strong – for a list of all possible reasons for inclusion of a gene in GenAge, see Subsection 2.1.1. To compensate for that, we also created a different set of classes in another scenario with four classes, where one of the classes only includes ageing-related DNA repair genes that were included in GenAge for a very strong reason, as will be explained in Subsection 3.3.1.

Once the set of data instances belonging to each class has been determined, the next step is to assign, to each data instance (DNA repair gene), a number of predictor attributes describing properties or features of that gene. Several types of predictor attributes – corresponding to different types of properties of the genes in the dataset – have been created in this research, as detailed in the following subsections.

### 3.1.2 Creating the predictor attribute type of DNA repair

Wood's web site also specifies, for each DNA repair gene, the main type of DNA repair activity which the gene is associated with. Some of these types are further divided into

sub-types, resulting in about 20 types or sub-types. In principle each of those types or sub-types could be considered as an attribute value, for the purposes of defining a predictor attribute. However, this would have the drawback that some of those types or sub-types have a very small number of genes associated with them. In general, attribute values associated with very few data instances (genes, in our case) have little predictive power for classification purposes. That is, it would be difficult for a classification algorithm to discover patterns using those attribute values that had good generalization ability.

In order to mitigate this problem, we merged some specific types and sub-types into broader types, which are then associated with a larger number of genes. More precisely, we defined a predictor attribute representing the type of DNA repair in which a gene is involved, consisting of the following 12 attribute values: base excision repair, mismatch repair, nucleotide excision repair, homologous recombination, non-homologous end joining, other types of DNA repair, DNA polymerases (catalytic subunits), editing and processing nucleases, Rad6 pathway, disease, other genes with known or suspect DNA repair function, other conserved DNA damage response genes. For a definition of these types of DNA repair, the reader is referred to (Friedberg et al., 2006), (Wood et al., 2001), (Wood et al., 2005).

### 3.1.3 Creating a predictor attribute measuring the rate of evolutionary change ($K_a/K_i$ ratio)

This attribute is essentially a measure of the rate of evolutionary change of orthologous ageing-related genes in human and chimpanzees (*Pan troglodytes*), called the $K_a/K_i$ ratio. The values of this attribute are calculated as follows.

A measure of evolutionary change commonly used to study selective forces acting on two sets of genes is the $K_a/K_s$ ratio, where $K_a$ and $K_s$ essentially represent the number of substitutions per non-synonymous and synonymous (respectively) sites between corresponding genes (Magalhaes and Church, 2007), (Chimpanzee-Consortium, 2005). More precisely, for each pair $i$ of genes where each gene belongs to a different set (for example, for each pair of orthologous genes between two genomes), $K_a$ is given by $a_i/A_i$,

where $a_i$ is the number of non-synonymous nucleotide substitutions (changing the amino acid sequence) and $A_i$ is the number of non-synonymous sites between the pair of genes. Analogously, $K_s$ is given by $s_i/S_i$, where $s_i$ is the number of synonymous nucleotide substitutions (which do not change the amino acid sequence) and $S_i$ is the number of synonymous sites between the pair of genes. $K_s$ is used as a normalisation factor for $K_a$, so that $K_a/K_s$ ratio values considerably greater than 1 suggest positive selection, favouring the emergence of new phenotypes; whilst $K_a/K_s$ values considerably smaller than 1 suggest purifying selection, promoting the conservation of existing phenotypes.

(Magalhaes and Church, 2007) used a variation of the $K_a/K_s$ ratio to analyse the rate of evolutionary change of orthologous ageing-related genes in human and chimpanzees (*Pan troglydytes*). They noted that, given the high similarity between human and chimpanzee genes, there are many pairs of orthologous genes with zero synonymous substitutions, and therefore they suggested to replace the above-described $K_s$ value by another value denoted $K_i$, based on the local intergenic or intronic substitution rate. This led to the $K_a/K_i$ ratio, which is the ratio used as a predictor attribute in our dataset. The value of the $K_a/K_i$ ratio for each DNA repair protein in our dataset was the same used in (Chimpanzee-Consortium, 2005), (Magalhaes and Church, 2007). For further details about how to calculate this ratio, the reader is referred to those references.

### 3.1.4 Creating a set of predictor attributes representing GO terms

The Gene Ontology categorizes gene and protein functions into three separate "namespaces": biological process, molecular function and cellular component (GO-Consortium, 2004) – see Section 2.2. We used as predictor attributes only biological process (BP) GO terms, since overall this type of term seems to be more easily interpretable as attributes for predicting whether a DNA repair gene is ageing related or not.

It is important to note that, in most gene and protein databases with GO term annotations, only the most specific GO terms known for a gene are explicitly included in the database. Ancestors of those specific terms are not normally explicitly included in the database

record for that gene. However, the semantics of the Gene Ontology specifies a hierarchical relationship between terms, so that if a gene has a certain biological process function associated with it, this means the gene also has all its "ancestor functions" in the GO hierarchy, even though those ancestor annotations are not explicitly included in the gene database. If information about those ancestors is not included in the dataset being mined, the algorithm could easily compute wrong probabilities or related statistics from the dataset.

In order to avoid this problem, we took the hierarchical relationship among GO terms into account when creating our datasets, as follows. First, for each DNA repair gene, we obtained the list of all the most specific GO terms annotated for that gene in the Uniprot database (www.uniprot.org). Secondly, since the type of ontology of a GO term is not explicitly provided in Uniprot records, we checked the type of ontology associated with each of those terms using information from the gene ontology web site (www.geneontology.org), and selected only terms of the BP ontology. Thirdly, for each DNA repair gene, we extended its list of specific BP GO terms with the set of all GO terms that are ancestors of those specific terms according to the "is a" relationship of the GO[1]. After the above steps different genes tend to have different numbers of GO terms, but most classification algorithms assume that each data instance (gene or protein) has the same number of attributes. Hence, in the fourth step, we put the data into a format suitable for most classification algorithms, involving the same set of binary attributes for each gene. To perform this step, we first identified all GO terms that occurred in the dataset. Each of these GO terms corresponds to a binary attribute. Each gene in the dataset is then associated with a binary value ("yes" or "no") for each of those terms, indicating whether or not the gene is annotated with that GO term. These steps are summarized in Figure 3.1.

[1] We thank Dr. Fernando Otero for running software that obtained the list of all GO terms that were ancestors, in the GO hierarchy, of all the more specific GO terms that we had in our initial version of the dataset.

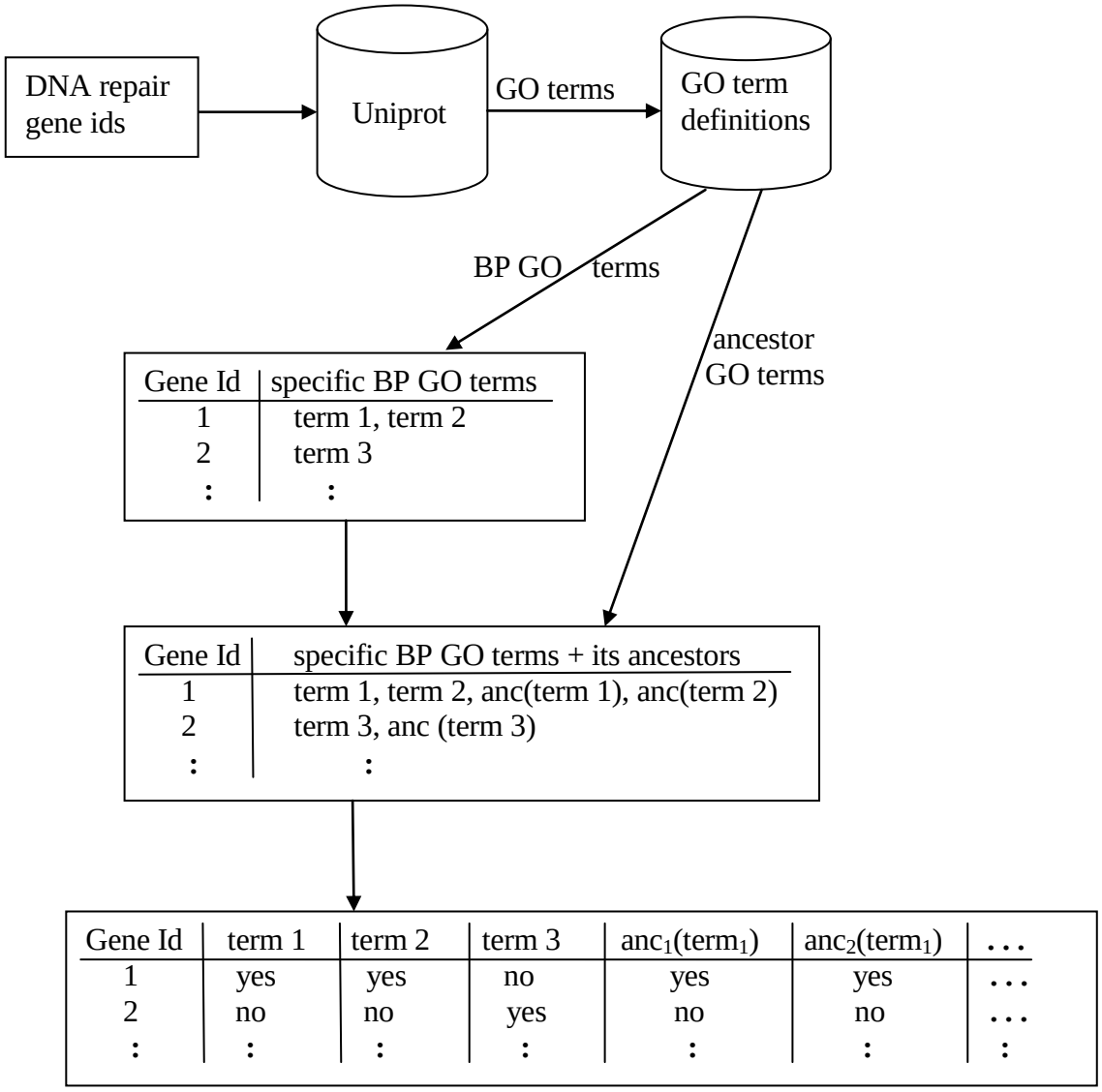


**Figure 3.1:** Summary of the procedure for creating a set of GO term-based predictor attributes. First, a list of gene ids is used to download from Uniprot the specific GO terms annotated for each gene. Next, information about GO term definitions is used to select only the biological process (BP) terms for each gene, and then to find the ancestors of those terms in the GO hierarchy. (The notation "anc($term_1$)" denotes the set of all ancestors of $term_1$, "$anc_1(term_1)$" denotes the first ancestor of term 1, and so on.) After adding those ancestor GO terms to the list of GO terms per gene, the dataset is transformed into a format having a fixed-length list of binary attributes (representing GO terms) for each gene, where each attribute value indicates whether or not the gene is annotated with the corresponding GO term.

After the above steps, a large number of predictor attributes based on GO terms is produced. Many of those GO terms are associated with just one or two genes in the

dataset, and therefore they correspond to predictor attributes with no or very low predictive power (generalization ability), statistically speaking. Hence, attributes (GO terms) whose value "yes" has a frequency of occurrence smaller than a predefined threshold – called the "GO term occurrence threshold" – are removed from the dataset. The value of this threshold is a parameter associated with the creation of the datasets.

The experiments reported in this thesis used datasets with three different values of the GO term occurrence threshold, namely 3, 7 and 11. The value 3 is an inclusive, low value, which avoids the use of rare GO terms with very little statistical support and virtually no generalisation power. Increasing the threshold value has two opposite effects, as follows. On one hand, although the classification algorithms are given a smaller set of attributes (because fewer GO terms satisfy the occurrence threshold), this smaller set has the advantage of including only GO terms with a larger frequency of occurrence in the dataset, for which the probabilities or related statistics computed by the classification algorithms are more reliable than the corresponding statistics computed for GO terms with fewer occurrences in the dataset.

On the other hand, if the threshold value is increased too much, GO terms which are relatively rare but still have some predictive power would be lost, which could lead to a decrease in classification accuracy. Hence, the experiments performed with different values of this parameter allowed us to study the aforementioned trade-off between statistical reliability and availability of relevant attributes for prediction.

### 3.1.5 Creating a set of attributes representing protein-protein interaction information

Information about protein-protein interaction (PPI) was obtained from the HPRD (Human Protein Reference Database), Release 8 – http://www.hprd.org/. Out of the many available PPI databases, HPRD was chosen because it focuses specifically on human proteins and because the curation of its data is of high quality (Prasad et al., 2009) – see Subsection 2.1.4. A set of attributes representing PPI information about the genes in our

dataset was created as follows. First, we downloaded the dataset of PPIs from HPRD and selected a subset of those interactions satisfying two conditions:

(a) at least one of the two proteins in the interacting pair is a DNA repair protein in our dataset; and

(b) the type(s) of evidence for the interaction includes *in vitro* or *in vivo* experiments – i.e. interactions for which the only type of evidence is high-throughput experiments were not selected, since this is considered a weaker type of evidence.

Once a set of PPIs has been selected in the previous step, two types of predictor attributes derived from those PPIs were used. The first type of PPI-related attribute involves a simple attribute, called NumInter (number of interaction partners). The value of that attribute for a given data instance (gene) is the number of proteins that interact with that gene's protein product.

The second type of PPI-related attribute involves a set of binary attributes, each of which refers to the presence or absence of interaction with a particular protein. That is, for each data instance (gene) in our dataset, each attribute takes on the value "yes" or "no" to indicate whether or not that gene's protein product interacts with the protein represented by that attribute. The set of PPIs selected from the HPRD database contained a large number of interacting proteins (namely 656), and it was not practical to create one predictor attribute for each of those proteins. Hence, we created a set of $N$ binary attributes referring to the $N$ most frequent proteins in the PPIs selected from HPRD, where $N$ is a parameter used in the creation of the datasets. The experiments reported in this thesis used datasets with three different values of $N$, namely 10, 20 and and 30.

### 3.1.6 Removing duplicate data instances

After a dataset has been created by generating the above-described set of predictor attributes, it usually happens that a few data instances – representing DNA repair genes which are similar to each other – contain the same values for all attributes. This happens because the majority of the attributes are binary attributes representing relatively sparse

data. That is, for most binary attributes representing the presence or absence of a GO term annotation or interaction with a given protein, the number of data instances with the value "yes" is much smaller than the number of data instances with the value "no". From a data mining perspective, duplicated instances should be removed (i.e. only one of all equal instances should be kept) from the dataset, to avoid the possibility that an instance in the test set is equal to an instance in the training set when performing the cross validation procedure (Subsection 2.4.2) – which would be a serious violation of the fundamental principle of measuring predictive accuracy on a test set containing data instances that were not observed in the training set, during the construction of the classification model.

Hence, as a final step in the creation of the datasets, we detect and remove duplicate instances in the data. The actual number of duplicates depends on the specific composition of the created dataset – which depends on parameters determining the number of GO terms used as attributes and the number of PPI-related attributes, as explained above. Hence, different datasets used in our experiments have slightly different numbers of data instances (just a few duplicated instances needed to be removed in each dataset). More precisely, after all the above-described dataset creation steps, the datasets created in this work have between 135 and 140 data instances, out of which 33 represent ageing-related DNA repair genes and the others represent non-ageing-related DNA repair genes. The precise number of data instances in each dataset is mentioned in the next subsection.

### 3.1.7 Dataset specifications

After the creation of the classes and multiple types of predictor attributes described in the previous subsections, five different types of datasets were created. These five dataset types use the same class definitions – i.e, in all datasets each data instance belongs to one out of two classes: "ageing-related" or "non-ageing-related" DNA repair gene.

Furthermore, all datasets contain the attribute "type of DNA repair", the attribute "$K_a/K_i$ ratio" (a measure of evolutionary change) and a number of GO term-based predictor attributes (each taking the value "yes" or "no" to indicate whether or not the gene is

annotated with the corresponding GO term). The actual number of GO term-based attributes is determined by the parameter GO term occurrence threshold – see Subsection 3.1.4. We did experiments with three values of this threshold, namely 3, 7 and 11, which led to datasets with 301, 157 and 101 terms, respectively.

In addition to this common set of types of attributes, the five types of datasets vary according to their use of attributes related to protein-protein interaction information, as follows:

**Dataset D1** contains no attribute related to protein-protein interaction information.

**Dataset D2** contains only one attribute related to protein-protein interaction information, namely the attribute NumInter. The value of this attribute for each data instance (gene) is the number of proteins that interact with that gene's protein product.

**Dataset D3** contains the NumInter attribute plus 10 attributes referring to binary protein interactions (BPI). The value of each BPI attribute for each data instance is "yes" or "no", indicating whether or not that gene's protein product interacts with the protein represented by that attribute.

**Dataset D4** contains the NumInter attribute plus 20 attributes referring to BPIs.

**Dataset D5** contains the NumInter attribute plus 30 attributes referring to BPIs.

In datasets D3, D4 and D5, the BPI attributes refer to the 10, 20 and 30 proteins with the greatest number of protein-protein interactions. For a review of the procedure for the creation of the attributes related to protein-protein interaction information, please see Subsection 3.1.5.

Table 3.1 shows the main characteristics of the datasets D1-D5. Each of these datasets is produced in three versions, each with a different value of the GO term occurrence threshold (second column). The column "Class distribution" mentions the number of data

instances belonging to the "ageing-related" and "non-ageing-related" classes, respectively, followed by the total number of data instances between brackets.

**Table 3.1:** Main characteristics of datasets with two classes and multiple attribute types

| Dataset | GO term occur. thres. | Class distribution | Predictor attributes |
|---|---|---|---|
| D1 | 3 | 33/104 (137) | DNA repair type, Ka/Ki, 301 GO terms |
| | 7 | 33/103 (136) | DNA repair type, Ka/Ki, 157 GO terms |
| | 11 | 33/102 (135) | DNA repair type, Ka/Ki, 101 GO terms |
| D2 | 3 | 33/106 (139) | DNA repair type, Ka/Ki, 301 GO terms, NumInter |
| | 7 | 33/105 (138) | DNA repair type, Ka/Ki, 157 GO terms, NumInter |
| | 11 | 33/105 (138) | DNA repair type, Ka/Ki, 101 GO terms, NumInter |
| D3 | 3 | 33/106 (139) | DNA repair type, Ka/Ki, 301 GO terms, NumInter, 10 binary protein interactions |
| | 7 | 33/105 (138) | DNA repair type, Ka/Ki, 157 GO terms, NumInter, 10 binary protein interactions |
| | 11 | 33/105 (138) | DNA repair type, Ka/Ki, 101 GO terms, NumInter, 10 binary protein interactions |
| D4 | 3 | 33/106 (139) | DNA repair type, Ka/Ki, 301 GO terms, NumInter, 20 binary protein interactions |
| | 7 | 33/105 (138) | DNA repair type, Ka/Ki, 157 GO terms, NumInter, 20 binary protein interactions |
| | 11 | 33/105 (138) | DNA repair type, Ka/Ki, 101 GO terms, NumInter, 20 binary protein interactions |
| D5 | 3 | 33/106 (139) | DNA repair type, Ka/Ki, 301 GO terms, NumInter, 30 binary protein interactions |
| | 7 | 33/105 (138) | DNA repair type, Ka/Ki, 157 GO terms, NumInter, 30 binary protein interactions |
| | 11 | 33/107 (140) | DNA repair type, Ka/Ki, 101 GO terms, NumInter, 30 binary protein interactions |

## 3.2 CREATING A DATASET WITH TWO CLASSES AND GENE EXPRESSION ATTRIBUTES

The two classes for this dataset were created in the same way as the two classes for the category of datasets described in the previous section – for details, please see Subsection 3.1.1. Hence, the set of genes in each class of this dataset is similar to the set of genes in the corresponding class in the category of datasets described in Section 3.1. However, the predictor attributes used in this dataset are very different, and they consist of only one type of predictor attribute, namely attributes representing gene expression values. Another characteristic that distinguishes this dataset from the others used in this research is that the gene expression values used to create this dataset were obtained from Genevestigator®, a commercial transcriptome meta-analysis tool produced by NEBION (http://www.nebion.com/nebion/doc/products.jsp); whilst all the other datasets were created from data in freely available biological databases on the web.

Genevestigator is a system for investigating gene expression and gene regulation (Hruz et al., 2008). In this work gene expression data was obtained by using the system's Anatomy tool, which reports how strongly a gene is expressed in different anatomical categories, including tissues and cell cultures. The information reported by this tool does not contain arrays from samples classified as cancer.

To create attributes representing gene expression levels reported by the anatomy tool, the lists of ageing-related and non-ageing-related DNA repair genes mentioned in Subsection 3.1.1 was used to search for their expression profiles pre-selected by the annotation adult human tissue in the Genevestigator's anatomy array collection. The data is plotted against a tree of anatomical categories. In some cases the tree was obviously simplified, extended, or adapted to avoid redundancies and trees with too many levels. For each anatomy category, Genevestigator displays the average expression value calculated from all arrays in the focused array selection that are annotated as belonging to this category or a child category. This means that the values from child nodes are included into parent nodes, which is reflected in the number of arrays that increases as one goes up the tree. For data

mining purposes, the data underlying Genevestigator's graphical displays were exported to a file in tabular form in tab-delimited format[2].

Note that Genevestigator contains expression data from multiple types of microarrays, for example, different generations of Affymetrix GeneChips. On these arrays, individual genes are sometimes represented by different sets of probes, which are not mixed. To get an advantage of all the existing data, we have used all probes corresponding to one gene in our analysis, by computing the arithmetic average of all gene expression values (one for each probe) for each gene. Hence, we created a dataset where each instance corresponds to a DNA repair gene and each column (attribute) corresponds to an anatomical category – i.e., each attribute value is the average expression level of a given gene for all probes in the corresponding anatomical category. After all data preparation steps, the created dataset has 109 predictor attributes and 148 data instances, out of which 33 belong to the "ageing-related DNA repair gene" class and 115 belong to the "non-ageing-related DNA repair gene" class. This dataset is hereafter denoted as **Dataset D6**.

Since each of the attributes is a continuous (real-valued) number, the issue of duplicate data instances that occurred in other datasets was not an issue in this gene expression dataset, and therefore there was no need to remove duplicate instances as described earlier for the other datasets.

For details related to data normalization and quality control in Genevestigator, see the Genevestigator manual: https://www.genevestigator.com/userdocs/manual/index.html, and for an overview of Genevestigator, see (Hruz et al., 2008).

[2] We thank Dr. Olga Vasieva for obtaining gene expression data from the Genevestigator tool and export it to a text file.

## 3.3 CREATING DATASETS WITH FOUR CLASSES AND MULTIPLE ATTRIBUTE TYPES

### 3.3.1 Creating the four classes to be predicted

The first class, denoted by *Age-HM-DNA* for short, includes ageing-related DNA repair genes whose reason for inclusion in GenAge was evidence directly linking the gene product to ageing in humans (H) or in a mammalian (M) model organism.

The second class, denoted by *Int-Age-HM-DNA* for short, includes genes whose protein product (denoted $p_1$) interacts with the protein product of another gene (denoted $p_2$) and the two interacting proteins satisfy the following two conditions: $p_1$ is the product of a gene that is not in GenAge – i.e., it is not considered an ageing-related gene – *and* $p_2$ is the product of a gene of the *Age-HM-DNA* class – as defined in the previous paragraph.

The third class, denoted *Non-Age-DNA* for short, includes DNA repair genes that are not in GenAge – i.e., they are not considered ageing-related genes.

The fourth class, denoted *Int-Non-Age-DNA* for short, includes genes whose protein product (denoted $p_1$) interacts with the protein product of another gene (denoted $p_2$) and the two interacting proteins satisfy the following two conditions: $p_1$ is the product of a gene that is not in GenAge – i.e., it is not considered an ageing-related gene – *and* $p_2$ is the product of a gene of the *Non-Age-DNA* class – i.e., a DNA repair gene that is not considered an ageing-related gene (because it is not in GenAge).

It should be noted that the definition of the first class, *Age-HM-DNA*, is significantly less inclusive than the definition of the class "ageing-related DNA repair gene" used in the datasets with two classes (see Subsection 3.1.1), since the latter included all DNA repair genes in the GenAge database, regardless of their reason for inclusion in that database. As a result, the *Age-HM-DNA* class consists of only nine data instances.

The motivation for defining this new *Age-HM-DNA* class in such a less inclusive way was that the genes in the second class (*Int-Age-HM-DNA*) are related to ageing only in an

indirect way – via protein-protein interaction with ageing-related genes in the *Age-HM-DNA* class – and so it seemed sensible to consider only interactions with genes having a very strong reason for inclusion in the GenAge database (evidence directly linking the gene product to ageing in humans or mammals).

In order to create the classes *Int-Age-HM-DNA* and *Int-Non-Age-DNA*, we used a set of protein-protein interactions (PPIs) which was produced using the same procedure that was used to create the predictor attributes related to PPI information, based on retrieving PPIs from the HPRD database whose evidence includes *in vivo* or *in vitro* experiments – as described in Subsection 3.1.5.

### 3.3.2 Creating the predictor attributes

The datasets with four classes included the following types of attributes used in the category of datasets described in Section 3.1: type of DNA repair, $K_a/K_i$ ratio, a set of GO term-based attributes, and, in some versions of the datasets, the NumInter attribute. It should be noted, however, that the attribute type binary protein interaction, whose definition was presented in Subsection 3.1.5, was not included in the 4-class datasets described here. The reason for this exclusion is that the definitions of two out of the four classes are already directly specified in terms of protein-protein interactions, and so it was decided that it would be better not to use that same kind of information in the definition of the predictor attributes. The exception is that, as mentioned earlier, the attribute NumInter was included in some versions of the dataset, since this attribute refers just to the total number of proteins interacting with the current gene's protein product, rather than refering to specific protein-protein interactions.

After creating the predictor attributes, duplicate data instances were removed from the dataset, as discussed in Subsection 3.1.6.

### 3.3.3 Dataset specifications

After the creation of the classes and multiple types of predictor attributes described in the previous subsections, two different types of datasets were created. These two dataset

types use the same definitions for the four classes, but somewhat different combinations of attribute types, as follows.

Both dataset types contain the attribute "$K_a/K_i$ ratio" (a measure of evolutionary change) and a number of GO term-based predictor attributes (each taking the value "yes" or "no" to indicate whether or not the gene is annotated with the corresponding GO term). The actual number of GO term-based attributes is determined by the parameter GO term occurrence threshold – see Subsection 3.1.4. We did experiments with three values of this threshold, namely 3, 7 and 11, which led to datasets with 560, 272 and 200 terms, respectively.

In addition to this common set of types of attributes, the two types of datasets vary according to the presence or absence of a single attribute, as follows:

**Dataset D7** contains no attribute related to protein-protein interaction information.

**Dataset D8** contains only one attribute related to protein-protein interaction information, namely the attribute NumInter. The value of this attribute for each data instance (gene) is the number of proteins that interact with that gene's protein product.

**Table 3.2:** Main characteristics of datasets with two classes and multiple attribute types

| Dataset | GO term occur. thres. | Class distribution | Predictor attributes |
|---|---|---|---|
| D7 | 3 | 9/133/96/90 (328) | Ka/Ki, 560 GO terms |
| | 7 | 9/133/96/90 (328) | Ka/Ki, 272 GO terms |
| | 11 | 9/132/96/88 (325) | Ka/Ki, 200 GO terms |
| D8 | 3 | 9/133/97/91 (330) | Ka/Ki, 560 GO terms, NumInter |
| | 7 | 9/133/97/91 (330) | Ka/Ki, 272 GO terms, NumInter |
| | 11 | 9/133/97/90 (329) | Ka/Ki, 200 GO terms, NumInter |

Table 3.2 shows the main characteristics of the datasets D7 and D8. Each of these datasets is produced in three versions, each with a different value of the GO term occurrence threshold (second column). The column "Class distribution" mentions the number of data instances belonging to the classes *Age-HM-DNA*, *Int-Age-HM-DNA*, *Non-Age-DNA*, *Int-Non-Age-DNA*, respectively, followed by the total number of data instances between brackets.

## 3.4 MEASURING PREDICTIVE ACCURACY

As discussed in Subsection 2.4.2, the performance of a classification model is essentially measured by its predictive accuracy in data that was not used to build the model, as follows. First, the classification model is built from a subset of the data called the training set, where the algorithm knows the values of both predictor attributes and classes for the data instances. After the model is built, its predictive accuracy is then measured in a separate subset of the data, called the test set, where the algorithm knows only the values of the predictor attributes (and not classes) for data instances. Therefore, this measure of predictive accuracy measures the generalization ability of the classification model.

In practice, measuring the predictive accuracy on a single test set is not the ideal, from a statistical point of view, because different test sets tend to result in somewhat different measures of predictive accuracy, given the non-determinism associated with the creation of the test set. (Any actual test set is just a sample out of a potentially infinite number of data instances that could be produced for a given set of predictor attributes and classes.) Hence, to get a more statistically reliable measure of predictive accuracy, it is usual to compute such a measure averaged over a number of test sets, with different sets of data instances – but of course the same definitions of predictor attributes and classes. A common approach to implement this idea – which is also the approach used in this research – consists of performing a 10-fold cross-validation procedure, which essentially works as follows (Witten and Frank, 2005).

First, the dataset is divided into 10 folds of approximately equal size. Next, the classification algorithm is run 10 times, each time with a different fold used as the test set and all the other nine folds merged into the training set. Then, any chosen measure of predictive accuracy (see below) is computed as the average value of that measure in the test set over the 10 experiments. Hence, each data instance is used exactly once in the test set and nine times in the training set.

The cross-validation procedure is very generic and it can be used with any given measure of predictive accuracy – for a review of several of such measures, see for example (Caruana and Niculescu-Mizil, 2004). In this work predictive accuracy is measured in terms of the Area Under the ROC curve (AUC) – using 10-fold cross-validation.

In essence, a ROC curve for a given classification model is plotted in a graph having the model's True Positive Ratio (TPR) in the Y axis and the model's False Positive Ratio (FPR) in the X axis. For each point in the curve, these ratios are defined as follows. Let one of the classes be the "positive" class, and consider all other classes as a single "negative" class. The TPR is the proportion of data instances that are correctly predicted to have the positive class out of all instances that really have the positive class – or the number of "true positive" instances divided by the total number of "positive" instances, in data mining terminology (Witten and Frank, 2005). The FPR is the proportion of data instances that are wrongly predicted to have the positive class out of all instances that have the negative class – or the number of "false positive" instances divided by the total number of "negative" instances.

A single point in the curve represents a single value of the TPR and FPR ratios, associated with a given value of a parameter affecting the class predictions made by the classification model. That parameter value controls the trade-off between two types of classification errors: predicting the positive class for an instance of the negative class (a "false positive"), or predicting the negative class for instance of the positive class (a "false negative"). In general the misclassification costs associated with these two types of errors are not equal and they depend on the application domain and the dataset being mined.

Hence, the motivation to compute a ROC curve is essentially to consider a range of values of a parameter controlling the trade-off between the numbers of false positives and false negatives, which gives us a more robust evaluation of predictive performance than a single value of the TPR and FPR ratios (Karwath and King, 2002), (Bradley, 1997).

The TPR and FPR ratios are computed by considering each of the classes in turn as the positive class, and then the results are averaged to produce the TPR and FPR coordinates for a point in the ROC curve. An example of a ROC curve is illustrated in Figure 3.2. In this figure, the diagonal dashed line represents the expected predictive performance of a random classification model – i.e., a model that makes class predictions in the test set at random. The ideal predictive performance is represented by the point at the extreme upper-left part of the graph in Figure 3.2, where the classification model would have TPR = 100% and FPR = 0%. Once a ROC curve has been computed, the predictive accuracy of a classifier is then measured as the Area Under the ROC Curve (AUC). In this research the AUC values were computed by using the WEKA data mining tool (Witten and Frank, 2005), the same tool used to run the classification algorithms used in this research.

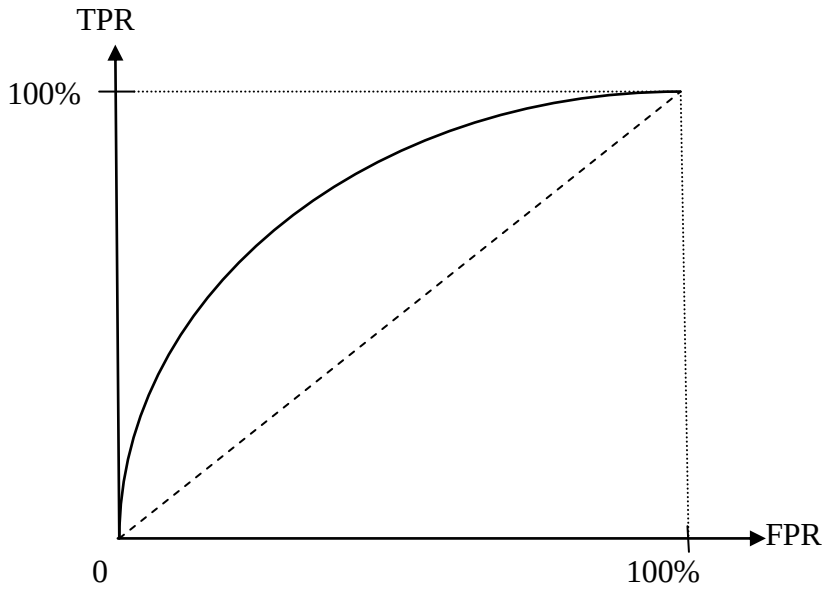


**Figure 3.2:** The ROC curve for measuring a classification model's predictive accuracy. TPR = True Positive Ratio. FPR = False Positive Ratio. (See text for definition of these ratios.)

In order to interpret the AUC values in the tables of results reported in Chapter 4, the main points are as follows. The AUC value varies in the range from 0 to 100%; and the larger the AUC value, the better the predictive accuracy of the classification model. In particular, a perfect predictive model would have an AUC value of 100%, whilst a model that makes predictions entirely at random would be expected to have an AUC value of 50%. For more details about ROC analysis and the AUC measure please see (Bradley, 1997), (Karwath and King, 2002).

## 3.5 STATISTICAL SIGNIFICANCE

Recall that a classification rule has the general form "IF (condition(s)) THEN (predicted class)". A number of such rules extracted from decision trees produced in our computational experiments will be reported in Chapter 4. The statistical significance of each of those rules was measured by using a hypothesis test based on the binomial distribution, as follows.

In the context of statistics in general, consider an experiment consisting of *n* independent random trials, where the result of each trial is either *success* or *failure*, with the same probability *s* of success in each trial. The distribution of the number of trials with the result *success*, out of the *n* independent trials, is binomial (DeGroot and Schervish, 2002). The values of *s* and *n* are parameters of the binomial distribution.

In the context of this research, in order to specify the null hypothesis for a test of statistical significance based on the binomial distribution, the classification made by a rule for a given data instance can be regarded as the result of a random trial with two outcomes: *success* (the predicted class equals the true class of that instance) or *failure* (the predicted class is different from the true class). In addition, the instances being classified by a rule are assumed to be independent from each other, and the number of trials *n* corresponds to the number of data instances being classified by the rule – i.e., the number

of data instances satisfying all conditions in the IF part of he rule. Furthermore, *under the null hypothesis assumption that the IF part of the rule is irrelevant to predict the class of an instance (i.e. assuming the rule has no predictive power),* for each data instance, the probability *s* of observing a *success* is given by the relative frequency of the class predicted by the rule in the dataset – i.e., the ratio of the number of instances of that class in the dataset divided by the total number of instances (of any class) in the dataset.

Hence, to set up a test of hypothesis for the statistical significance of the predictive power of a classification rule, we take into account the observed number of data instances correctly classified by the rule, denoted *k*, where $k \leq n$. Let *X* be a random variable with binomial distribution with probability of success *s* and number of trials *n*, so that *X* can take integer values in the range from 0 to *n*. Under the above null hypothesis that the rule has no predictive power, the probability of observing exactly *k* successes, according to the binomial distribution, would be (DeGroot and Schervish, 2002):

$$p(X = k) = C_{n,k}\, s^k\, (1 - s)^{n-k}, \quad (3.1)$$

where $C_{n,k}$ is the number of combinations of *k* elements out of *n* elements, given by $n!/(k!(n-k)!)$. For the test of hypothesis, however, we need to compute the probability that the observed number of successes will be a value $\geq k$, which is given by:

$$p(X \geq k) = 1 - \mathrm{p}(X < k), \quad (3.2)$$

where the term $\mathrm{p}(X < k)$ is given by the cumulative binomial distribution, adding up the values $p(X = x)$ for all *x* values between 0 and $k - 1$, i.e.: $\mathrm{p}(X < k) = \Sigma_{x=0,k-1}\, p(X = x)$.

Hence, the null hypothesis that the rule has no predictive power can be rejected – so that the rule is considered statistically significant – when the value of the probability given by equation (3.2) (the so-called *p value*) is very small, say less than 1%.

# Chapter 4 – Computational Results and Discussion

This chapter is divided into three sections, each of them presenting the computational results and discussion for one of the categories of datasets whose creation was explained in detail in Sections 3.1, 3.2 and 3.3. The computational experiments involved running three classification algorithms – namely the J4.8 and CART decision tree induction algorithms and the Naive Bayes algorithm (see Section 2.5) – in each of the datasets created in this research. To run these algorithms we used the freely available data mining tool WEKA (Witten and Frank, 2005), version 3.6.1, and in all experiments these algorithms were ran with their default parameter values in that tool.

Note that the experiments produced a large number of classification models (outputs of the previously mentioned algorithms), and it is not practical to discuss all those models in detail. In this chapter we are selective and focus on discussing the most relevant predictive patterns extracted from those classification models. Some results reported in Sections 4.1 and 4.2 have been published in (Freitas et al., 2011).

## 4.1 RESULTS AND DISCUSSION FOR DATASETS WITH TWO CLASSES AND MULTIPLE ATTRIBUTE TYPES

All the results reported in this section refer to the previously described datasets with two classes of DNA repair genes and multiple types of predictor attributes. A detailed explanation about the creation of classes and attributes for these datasets is provided in Subsections 3.1.1 through 3.1.6, whilst the specification of the major characteristics of these datasets – such as the number and types of attributes in each dataset – is presented in Subsection 3.1.7.

In summary, these datasets have the following characteristics. Each data instance belongs to one out of two classes: “ageing-related” or “non-ageing-related” DNA repair gene. In

addition, the datasets contain multiple types of predictor attributes. All datasets contain the attribute "type of DNA repair", the attribute "$K_a/K_i$ ratio" (a measure of evolutionary change) and a number of GO term-based predictor attributes (each taking the value "yes" or "no" to indicate whether or not the gene is annotated with the corresponding GO term). The actual number of GO term-based attributes is determined by the parameter GO term occurrence threshold. We did experiments with three values of this threshold, namely 3, 7 and 11, which led to datasets with 301, 157 and 101 terms, respectively. In addition to this common set of types of attributes, the created datasets vary according to their use of attributes related to protein-protein interaction (PPI) information. This led to five types of datasets, with different numbers of attributes related to PPI information, as will be shown in the tables of results reported in this chapter. (For details, please see Subsection 3.1.7.)

This section can be conceptually divided into two parts. In the first part, the results for each classification algorithm are discussed in a separate subsection (4.1.1 through 4.1.3), and the discussion focuses on the predictive accuracy obtained by each of the algorithms. In the second part, in Subsection 4.1.4, we discuss the interpretation of the results in the light of biological knowledge, focusing on specific predictive patterns extracted from the decision trees built in the experiments – rather than on the predictive performance of each algorithm as a whole.

### 4.1.1 Results for the J4.8 decision tree induction algorithm

Table 4.1 shows the predictive accuracy – measured by the AUC value (see Section 3.4) – obtained by the J4.8 decision tree induction algorithm. Several relevant remarks can be made about this table. First, overall, using binary protein interaction (BPI) attributes considerably increases predictive accuracy. For each of the three values of the GO term occurrence threshold, the AUC value obtained using BPI attributes is considerably greater than the AUC value obtained without that type of attribute. This tendency is particularly clear in the column for the threshold value of 3, where the AUC value for dataset D1 (with no BPI attribute) was 63% and the AUC values for datasets D3-D5 varied from 72.3% to 80%.

**Table 4.1:** Area Under ROC curve (AUC, in %) for J4.8 algorithm, for datasets with two classes and multiple attribute types

| Dataset | PPI-related attributes | GO term occurrence threshold (number of GO terms) | | |
|---|---|---|---|---|
| | | 3 (301 terms) | 7 (157 terms) | 11 (101 terms) |
| D1 | none | 63.0 | 68.0 | 65.3 |
| D2 | NumInter | 66.1 | 63.3 | 59.6 |
| D3 | NumInter + 10 BPI attr's | 72.3 | 74.2 | 75.4 |
| D4 | NumInter + 20 BPI attr's | **80.0** | 73.5 | 74.6 |
| D5 | NumInter + 30 BPI attr's | 79.2 | 67.7 | 77.5 |

Concerning the effect of different values of the GO term occurrence threshold in the predictive accuracy of J4.8, increasing the value of that threshold to 7 or 11 had mixed effects. In particular, those increased threshold values led to higher AUC values in datasets D1 and D3, but lower values in datasets D2, D4 and D5, by comparison with the AUC values associated with the original threshold value of 3. Overall, taking into account all datasets, the best results are achieved with the GO term occurrence threshold set to 3, and the best two results in the entire table are achieved for datasets D4 and D5 with the GO term occurrence threshold value of 3, corresponding to AUC values of 80.0% (boldfaced in Table 4.1) and 79.2%, respectively.

To illustrate the kind of decision tree produced by J4.8 in these experiments, the decision tree associated with the highest AUC value in these experiments is shown in Figure 4.1. This tree is drawn using indentation (denoted by the "|" symbol) to indicate further levels of tree depth, to keep the figure simple – this is the tree-drawing form directly produced by the data mining tool used in the experiments, WEKA (Witten and Frank, 2005). The root node is WRN_interaction, and the tree has seven leaf nodes, each of them associated with a predicted class: "ageing" (i.e. ageing-related DNA repair gene) or "non-ageing" (i.e., non-ageing-related DNA repair gene). To facilitate the identification of the leaf nodes, its predicted classes are boldfaced in the figure. For each leaf node, right after the predicted class there are one or two numbers between brackets. The first one is the

number of data instances (DNA repair genes) classified by that node – i.e., data instances that satisfy all the attribute-value conditions in the path from the root until that leaf node – regardless of the predicted class being correct or not. The second one is the number of data instances wrongly classified by that leaf node, the so-called "false positives" for that leaf node, in data mining terminology. Hence, the number of DNA repair genes correctly classified by a leaf node is given by the first number minus the second one. If there is no second number for a given leaf node, this means there is no false positive associated with that leaf node, and all DNA repair genes associated with that leaf node are correctly classified.

WRN_interaction = no
| GO:0006295 (nucleotide-excision repair, DNA incision, 3'-to lesion) = no
| | CHEK1_interaction = no
| | | GO:0006351 (transcription, DNA-dependent) = no
| | | | GO:0009719 (response to endogenous stimulus) = no
| | | | | GO:0006283 (transcription-coupled NER) = no: **non-ageing** (112.0/7.0)
| | | | | GO:0006283 (transcription-coupled NER) = yes: **ageing** (4.0/1.0)
| | | | GO:0009719 (response to endogenous stimulus) = yes: **ageing** (2.0)
| | | GO:0006351 (transcription, DNA-dependent) = yes: **ageing** (2.0)
| | CHEK1_interaction = yes: **ageing** (2.0)
| GO:0006295 (nucleotide-excision repair, DNA incision, 3'-to lesion) = yes: **ageing** (3.0)
WRN_interaction = yes: **ageing** (14.0)

**Figure 4.1:** Decision tree built by J4.8 for dataset D4 and GO term occurrence threshold 3 in Table 4.1

As can be observed in Figure 4.1, most leaf nodes are predicting the ageing-related class for small numbers of DNA repair genes – varying from two to four genes. These specific predictions help to achieve a good predictive accuracy for the decision tree as a whole, but individually they may not be very statistically reliable, due to the small number of genes involved. However, the value "yes" for the attribute WRN_interaction (at the bottom of the tree) leads to a very reliable prediction of the ageing-related class, involving

14 DNA repair genes and with no false positive. The statistical significance and biological interpretation of this and other patterns extracted from the decision trees built by both J4.8 and CART will be discussed later, in Subsection 4.1.4.

### 4.1.2 Results for the CART decision tree induction algorithm

Table 4.2 shows the predictive accuracy (measured by the AUC value) obtained by the CART decision tree induction algorithm. The AUC value was quite low – around 51-55% – for dataset D1, which does not include any attribute related to protein-protein interaction (PPI) information. It should be recalled that an AUC value of 50% is expected from random predictions, so the predictions made by the decision trees built by CART for dataset D1 are just slightly better than random predictions. Similarly to the results observed for J4.8 in the previous subsection, CART's AUC values are considerably better in the other datasets, D2 through D5, which include predictor attributes with PPI-related information. In general, the highest AUC values in that table are observed for datasets D4 and D5, using not only the NumInter attribute, but also 20 and 30 binary protein interaction (BPI) attributes, respectively. An exception to this trend was that, in the column for GO term occurrence threshold = 11, the AUC value for dataset D5 (64.4%) was considerably lower than for datasets D4 (72.7) and D3 (67.0%), but that former value is still greater than the AUC value for databases D2 and D1 in the same column.

**Table 4.2:** Area Under ROC curve (AUC, in %) for CART algorithm, for datasets with two classes and multiple attribute types

| Dataset | PPI-related attributes | GO term occurrence threshold (number of GO terms) | | |
|---|---|---|---|---|
| | | 3 (301 terms) | 7 (157 terms) | 11 (101 terms) |
| D1 | none | 51.4 | 53.4 | 55.0 |
| D2 | NumInter | 68.9 | 65.4 | 62.5 |
| D3 | NumInter + 10 BPI attr's | 67.3 | 69.1 | 67.0 |
| D4 | NumInter + 20 BPI attr's | 68.5 | 72.0 | **72.7** |
| D5 | NumInter + 30 BPI attr's | 70.1 | 72.0 | 64.4 |

Concerning the effect of different values of the GO term occurrence threshold in the predictive accuracy of CART, increasing the value of that threshold to 7 or 11 had again mixed effects – like observed in the results for J4.8 reported in the previous subsection. In particular, those increased threshold values led to higher AUC values (by comparision with the threshold value of 3) in datasets D1 and D4, but led to lower or approximately the same AUC values in the other datasets.

The overall best result, across all 15 entries in the table, was an AUC value of 72.7% (boldfaced in the table), obtained for dataset D4 and GO term occurrence threshold = 11; but this value is, in general, similar to the AUC values obtained for datasets D4 and D5 with all GO term occurrence thresholds. Actually, a manual inspection of the trees built by CART for datasets D4 and D5 across all threshold values revealed that in all cases the decision tree built was the same, namely the decision tree shown in Figure 4.2. (The different AUC values observed across these different cases is due to the non-determinism of the cross-validation procedure used to measure the AUC values.) This very simple decision tree uses just one predictor attribute, WRN_interaction, and classifies data instances (DNA repair genes) as follows. If a DNA repair gene's protein product interacts with the Werner's protein, that gene is classified as ageing-related, otherwise it is classified as non-ageing-related. A discussion about the relevance of this pattern will be presented later, in Subsection 4.1.4.

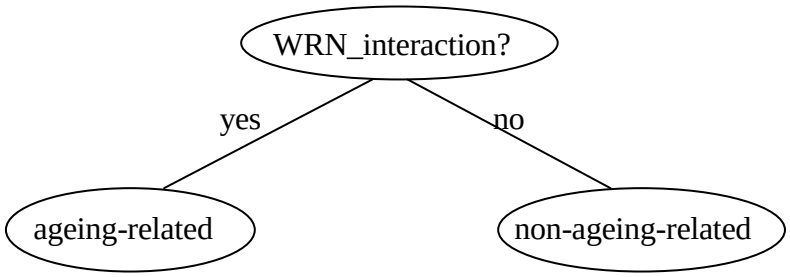


**Figure 4.2:** Decision tree consistently built by CART for datasets D4 and D5 and the three values of the GO term occurrence threshold in Table 4.2

This trend in the CART results, i.e. building the same decision tree for all six versions of datasets D4 and D5, was not observed in the results for J4.8 reported in the previous

subsection. J4.8 results showed a reasonable diversity of decision trees across those six dataset versions.

Although a detailed analysis of the relative effectiveness of different types of classification algorithms is a data mining research issue which is not the main goal of this thesis – we are rather mainly interested in the biological interpretation of the patterns discovered by those algorithms, as will be discussed later – it is interesting to make some brief comments about the relative predictive performance of J4.8 and CART, as follows.

In the experiments reported in this subsection and the previous subsection, J4.8 obtained, overall, better predictive accuracy results than CART. Indeed, comparing the AUC values in Table 4.1 (for J4.8) with the corresponding values in Table 4.2 (for CART), one can see that J4.8 consistently obtains considerably higher values than CART, for all GO term occurrence thresholds, in three datasets, namely D1, D3, D4. There is only one dataset where CART obtains AUC values somewhat better than J4.8, namely dataset D2. Dataset D5 presents more mixed results, but overall J4.8 obtains better AUC values for this dataset.

A possible explanation for this relatively superiority of J4.8 over CART involves the fact that J4.8 built, in general, larger decision trees than CART, i.e., the trees built by J4.8 tended to have more nodes and therefore to use more predictor attributes than the trees built by CART. More precisely, over the 15 combinations of datasets and GO term occurrence threshold values used in our experiments, the decision trees built by J4.8 had on average 9.4 leaf nodes, whilst the decision trees built by CART had on average just 3.0 nodes.

Hence, one can conclude that, at least in the datasets used in our experiments, CART tends to use a considerably more "conservative" approach in selecting attributes to add to a decision tree than J4.8. That is, CART seems to require a stronger evidence for the predictive power of a candidate attribute in order to include that attribute in the decision tree. In terms of the previously mentioned trade-off between the chances of overfitting

and underfitting, mentioned in Subsection 2.4.2.1, it seems that CART is therefore more prone to underfitting than J4.8; whilst J4.8 is more prone to overfitting than CART. In the experiments reported here, the "less conservative" approach of J4.8 led, overall, to better predictive accuracy results – although in some cases CART obtained better results, confirming the well-known fact that it is worth using more than one classification algorithm instead of simply relying on a single algorithm.

### 4.1.3 Results for the Naive Bayes algorithm

The results for Naive Bayes are reported in Table 4.3. In general, for all three values of the GO term occurrence threshold, Naive Bayes' AUC values increased monotonically, from the first row (D1) to the last row (D5), with an increase in the number of protein-protein interaction (PPI)-related attributes – i.e., NumInter and binary protein interaction (BPI) attributes.

In the case of Naive Bayes, increasing the value of the GO term occurrence threshold to 7 or 11 led to somewhat lower AUC values in four datasets (D1-D4), by comparison with the AUC values associated with the original threshold value of 3. However the highest AUC value in Table 4.3 was achieved with that threshold set to 11, for dataset D5 (AUC = 82.6%).

**Table 4.3:** Area Under ROC curve (AUC, in %) for Naive Bayes, for datasets with two classes and multiple attribute types

| Dataset | PPI-related attributes | GO term occurrence threshold (number of GO terms) | | |
|---|---|---|---|---|
| | | 3 (301 terms) | 7 (157 terms) | 11 (101 terms) |
| D1 | none | 75.9 | 74.9 | 71.9 |
| D2 | NumInter | 76.0 | 75.3 | 74.0 |
| D3 | NumInter + 10 BPI attr's | 78.3 | 77.1 | 76.6 |
| D4 | NumInter + 20 BPI attr's | 80.5 | 80.1 | 79.4 |
| D5 | NumInter + 30 BPI attr's | 80.7 | 80.2 | **82.6** |

In summary, varying the value of the GO term occurrence threshold had, in general, little effect on the predictive accuracy of Naive Bayes, which was more affected by the types of predictor attributes used in the dataset.

### 4.1.4 Discussion on predictive patterns extracted from the decision trees

The previous subsections focused on evaluating the predictive accuracy of the classification algorithms used in the experiments. In terms of interpretation of the results, the previous subsections only showed the most accurate decision tree built by each of the decision tree induction algorithms, J4.8 and CART, and discussed its main broad characteristics.

By contrast, in this section we discuss in more detail predictive patterns extracted from the decision trees built by J4.8 and CART, and interpret them in the light of biological knowledge. More precisely, we discuss two kinds of patterns. First, we discuss the distribution of attributes chosen as root nodes in decision trees, in Subsection 4.1.4.1. Then, after discussing some important issues on the selection and interpretation of classification rules extracted from decision trees in Subsection 4.1.4.2, we discuss several such classification rules in Subsection 4.1.4.3.

The following subsections focus on extracting predictive patterns from decision trees, rather than from the output of the Naive Bayes algorithm, because patterns extracted from decision trees – particularly classification rules – were considered easier to be identified and interpreted, as follows. First, a decision tree contains only the relevant attributes for classification purposes – as determined by the decision tree algorithm – whilst the output of Naive Bayes contains probabilities for all predictor attributes in the dataset being mined (on the order of hundreds of attributes, in this research). Secondly, most classification rules extracted from decision trees identify combinations – more precisely, logical conjunctions – of different attribute values in their IF part, taking into account interactions between different attributes. Such combinations can give an insight about predictive patterns in the data that would not be available by analysing the attributes individually, like Naive Bayes does.

*4.1.4.1 Discussion on attributes chosen as root nodes in the decision trees*

Recall that, out of all attributes which were selected to be included (as labels of internal nodes) in the decision tree, in general the most important attribute is the one selected to label the root node of the tree, since the value of this attribute will be used (potentially together with other attributes) to classify every data instance in the test set – see Subsection 2.5.1 for details.

Hence, it is worthy to analyse the distribution of the root attributes selected by the decision tree induction algorithms across the datasets used in the previously-reported experiments. From a biological perspective, the detail about whether a given root node attribute was selected by the J4.8 or CART algorithm is not very relevant. Hence, we analyse the distribution of the root attributes in all trees built by any of these algorithms. This gives us a total of 30 decision trees – 15 dataset versions times two decision trees per dataset version (one tree built by J4.8, the other built by CART) – to be analysed, which has the advantage of providing statistically more reliable information, by comparison with the analysis of just 15 decision trees produced by a single algorithm.

In this spirit, Table 4.4 shows how many times each attribute was selected to be at the root node of the decision tree, for all attributes which were selected as root node in at least one decision tree. It should be noted that the actual frequency of selection of an attribute for the root node across all decision trees depends not only on the predictive power of the attribute (as evaluated by the decision tree induction algoritm), but also on the number of datasets in which the attribute was included. Although some attributes were included in all datasets, recall that many GO term-based attributes and attributes related to protein-protein interaction information were included only in some datasets. Hence, to make the comparison between different attributes mentioned in the table fairer, the table shows – in its second column – both the number of times the attribute was selected as a root node and the number of datasets in which the attribute was included (the number after "out of"), which was the maximum number of times the attribute could have been selected.

The attributes are listed in the table in decreasing order of selection frequency. Note that the summation of the frequencies across all rows equals 29, rather than 30 – the total number of decision trees analysed. This is because one of those decision trees was an "empty" tree, with no attribute selected for the root node – i.e, one of the algorithms was not able to find any attribute with a predictive power good enough to be included in the decision tree. That empty tree was produced by the CART algorithm in the dataset D1 – with no protein-protein interaction-related attribute – with GO term occurrence threshold = 7. (That empty tree consists of a single leaf node that assigns, to all data instances to be classified, the most frequent class in the dataset – i.e., class "not-ageing-related DNA repair gene".)

**Table 4.4:** Frequency of selection of an attribute as a root node in decision trees, for datasets with two classes and multiple attribute types

| Attribute | Frequency |
|---|---|
| WRN_interaction | 12 (out of 12) |
| NumInter | 8 (out of 24) |
| GO:0009719 (response to endogenous stimulus) | 3 (out of 10) |
| XRCC5_interaction | 2 (out of 18) |
| GO:0042221 (response to chemical stimulus) | 2 (out of 30) |
| GO:0045935 (positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process) | 2 (out of 30) |

Note: The GO terms GO:0009719 (response to endogenous stimulus) and GO:0042221 (response to chemical stimulus) are sibling terms in the GO's hierarchy, since both are a child of the more general GO term GO:0050896 (response to stimulus), via the "is-a" relationship. On the other hand, the GO term GO:0045935 (positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process) is not related to any of the above GO terms via the "is-a" relationship.

Interestingly, protein-protein interaction (PPI)-related attributes were chosen to be decision tree root nodes much more often than GO term-based attributes, even though there are much fewer PPI-related attributes than GO term-based attributes in the datasets.

More precisely, the number of PPI-related attributes varies from 0 to 31 across the datasets, whilst the number of GO term-based attributes varies from 101 to 301 across the datasets. Yet, as can be seen in the table, only three GO term-based attributes were chosen to be root nodes of a decision tree. Two out of those three GO terms refer to response to stimulus, and in both cases the definition of the kind of stimulus in question seems somewhat broad.

By far the most relevant attribute for building decision trees was the WRN_interaction attribute, which takes the value "yes" or "no" depending on whether or not the current DNA repair gene's protein product interacts with the WRN protein. This attribute was selected as the root node of the decision tree in 100% (12 out of 12) of the experiments where it was actually contained in the dataset used as input by J4.8 or CART. The WRN protein is a DNA helicase and an exonuclease, and humans with a defect in this gene suffer from the so-called Werner's progeroid syndrome (Subsection 1.3.1.1). Although there is some controversy about to what extent progeroid syndromes are good models of the real ageing process, Werner's syndrome is considered the progeroid syndrome that most presents characteristics of real ageing (Magalhaes and Faragher, 2008).

The second most relevant attribute in Table 4.4 was NumInter, which was selected as a root node in 33% (8 / 24) of the cases. The GO term "response to endogenous stimulus" was also found to be very relevant, being selected as a root node in 30% (3 / 10) of the cases. The other three attributes mentioned in Table 4.4 were, by comparison, found to be less relevant for the purpose of acting as a root node in a decision tree; they were selected to have this role in about 10% or less of the cases.

Note that Table 4.4 only identifies the attributes which were considered most relevant – in the sense of having the greatest predictive power – by the decision tree algorithms in general, but class predictions are made for specific values of those attribute, and often in combination with other values of other attributes. Hence, it is also important to interpret specific classification rules, extracted from the decision trees, which use the predictor attributes listed in Table 4.4. We discuss several classification rules in Subsection 4.1.4.3.

Before moving to that subsection, however, we discuss in the next subsection some relevant issues about the interpretation of rules extracted from the decision trees.

*4.1.4.2 Issues on selecting and interpreting rules extracted from decision trees*

Recall that each path in a decision tree from the root node to a leaf node corresponds to an IF-THEN classification rule, as follows. The IF part of the rule consists of the attribute-value conditions along that path; and the THEN part predicts, for each data instance satisfying the conditions in the IF part, the class associated with the leaf node.

Not all classification rules extracted from a decision tree are interesting, though; and it is not practical to try to interpret each rule that can be extracted from each of the decision trees built in the experiments reported in the previous subsection. Hence, it is necessary to be selective and focus on the potentially more interesting classification rules.

There are several reasons why a rule may not be "interesting" from a biological perspective. Some leaf nodes correspond to rules with many false positives – because the algorithm was unable to find attributes with enough predictive power to discriminate well among classes in that tree node. In addition, some leaf nodes correspond to rules which are very specific, covering two few (say, just two) data instances – and therefore do not seem very reliable. Furthermore, even if a given rule covers a reasonable number of data instances and has no or very few false positives, it might not be interesting from a biological perspective due to the difficulty of interpreting the relationship between the predictor attributes in its IF part and the class predicted in the THEN part of the rule.

In this research we used predictor attributes which have good interpretability, in general, but there are two caveats concerning this issue. First, we used a large number of predictor attributes representing terms of the "biological process" namespace of the Gene Ontology (GO). These attributes represent information about biological processes at various levels of abstraction, varying from very generic to very specific biological processes – recall that GO terms are arranged into a hierarchical structure (Section 2.2). Hence, in some cases the IF part of a rule contains attributes referring to GO terms which represent information

that might be too generic or too specific, failing to represent interesting knowledge about which properties of DNA repair genes makes them ageing-related or not.

Another caveat involves the fact that, among the several types of predictor attributes used in this research, there are two types of attributes, namely GO terms and binary protein interactions, which are binary attributes which can take on the value "yes" or "no". In the case of attributes representing GO terms, these values indicate whether or not a data instance (DNA repair gene) has the GO term associated with the attribute. In the case of attributes representing binary protein interaction information, the "yes" or "no" values indicate whether or not a DNA repair gene's protein product interacts with the protein associated with the attribute. Hence, for both types of attribute, the values "yes" tends to be considerably more informative and easier to be interpreted than the value "no". After all, the value "yes" is a definite statement about a property of a gene, whilst the value "no" represents a less definite statement.

For instance, many of the GO term-based attributes in our datasets have a relatively low number of "yes" values – which is the reason for the specification of the parameter "GO term occurrence threshold" (Subsection 3.1.4). For such "sparse" attributes, the value "no" is not very informative – since in the case of such attributes the value "no" is shared by many DNA repair genes without implying that they have a similar well-defined property. In addition, the value "no" seems somewhat less reliable from a data mining perspective, because when the value "no" is assigned to a given data instance, that only means that the instance is not known, at present, to have the GO term's biological process function or binary protein interaction associated with that attribute; but new biological experiments might reveal the presence of such function or interaction in the future.

Taking into account the above caveats, we have manually selected a set of classification rules, extracted from the decision trees built by J4.8 or CART, to be discussed here, interpreting them in the light of biological knowledge. As general criteria for this selection, we gave priority to rules that cover at least four data instances, have relatively few false positives (ideally no false positives in the case of specific rules covering around

four or five data instances, or a small proportion of false positives for rules covering many more data instances), and rules whose IF part's attributes refer to biological information that is reasonably interpretable – trying to avoid the selection of rules whose attributes represent information which is either too generic or too specific.

It should be noted that the selection of classification rules to be reported in this section according to the above criteria involves, of course, subjective decisions; unlike the completely automated procedure used to build the decision trees from the datasets created in this research. At the current state of the art, data mining algorithms simply are not advanced enough to "understand" the meaning of the information they are processing, and therefore such subjective interpretation of the results of those algorithms is needed, constituting an important component of this research.

There is, however, a more objective aspect in the evaluation of the classification rules selected to be reported here, which concerns the evaluation of their statistical significance. Hence, for each classification rule reported here, we mention whether or not its predictive pattern is statistically significant, according to the statistical hypothesis test based on the binomial distributioin discussed in Section 3.5.

It should be noted that, in the discussion of the classification rules reported here, the focus is on the interpretation of the rules from a biological perspective, and in this context the specific decision tree induction algorithm (J4.8 or CART) that was used to discover a rule is not very relevant. Hence, instead of discussing the results one algorithm at a time, as presented in the previous subsections, we discuss the rules together in the next subsection. For the sake of completeness, however, we mention, for each rule, the identification of the decision tree where the rule was extracted from. This identification consists of the dataset from where the rule was discovered and the algorithm used to discover it. In addition, we focus mainly on rules predicting the class "ageing-related DNA repair gene", since this is the main class of interest in this research, but we also discuss a couple of rules predicting the class "non-ageing-related DNA repair gene".

*4.1.4.3 Discussion on selected rules extracted from decision trees*

The following very simple IF-THEN classification rule – with a single condition in its part – was extracted from the three decision trees built by J4.8 for datasets D1, D2 and D3 – in all cases with the parameter GO term occurrence threshold set to 3.

Rule 1:
IF GO:0009719 (response to endogenous stimulus) = yes
THEN class is ageing-related DNA repair gene

There are four DNA repair genes satisfying the IF part of this rule, and all of them belong to the predicted class. This rule is highly statistically significant – $p = 0.003$ under the null hypothesis that the rule makes has no predictive power (i.e., the rule predictions are not better than random predictions). See Section 3.5 for details of the statistical test of significance.

The following rule was extracted from a decision tree built by J4.8 in the version of dataset D2 (with NumInter attribute but no binary protein interaction attributes) with parameter GO term occurrence threshold set to 11.

Rule 2:
IF NumInter > 15
AND GO:0050896 (response to stimulus) = yes
AND GO:0048518 (positive regulation of biological process) = yes
THEN class is ageing-related DNA repair gene

There are 10 DNA repair genes satisfying the IF part of this rule, and all of them belong to the predicted class. This rule is highly statistically significant – $p = 6.1 \times 10^{-7}$, with the aforementioned test of statistical significance.

This rule uses as predictor attributes not only GO terms but also the NumInter attribute, which was the attribute with the second largest frequency of selection to be a root node (most relevant attribute) in a decision tree, as mentioned earlier. This rule shows that the

values of this attribute that are most associated with ageing are relatively high values (> 15 in this particular rule); and indeed a manual inspection of all decision trees as a whole confirmed that, broadly speaking, high values of this attribute tend to be associated with the ageing-related class.

This result is consistent with other investigations showing that ageing-related proteins in general (without focusing on DNA repair genes) tend to have a higher number of interaction partners than non-ageing-related proteins (Budovsky et al., 2007), (Ferrarini et al., 2005), (Promislow, 2004).

The following rule, referring to a somewhat more specific kind of stimulus, namely chemical stimulus, was extracted from two decision trees built by J4.8 in two versions of the D1 dataset (with no attribute related to protein-protein interaction information) with the parameter GO term occurrence threshold set to 7 and 11.

Rule 3:
IF GO:0042221 (response to chemical stimulus) = yes
AND GO:0050789 (regulation of biological process) = yes
THEN class is ageing-related DNA repair gene

There are 11 DNA repair genes satisfying the IF part of this rule, and 10 of them belong to the predicted class. This rule is highly statistically significant – $p = 6.5 \times 10^{-6}$, with the aforementioned test of statistical significance.

Taken as a whole, the above three rules indicate that involvement in a biological process related to response to stimulus in general is a good predictor of ageing-relatedness for DNA repair genes. It is worth mentioning that the GO term "response to external stimulus" was one of the GO terms overrepresented in an ageing-related interaction network of a very different type of gene or protein, namely extracellular proteins (Chautard et al., 2010). This suggests that the relevance of response to stimulus for predicting ageing-relatedness is not limited to DNA repair genes.

Let us now turn to classification rules using only predictor attributes related to protein-protein interaction information, and not GO terms. One of these rules, referring to the WNR_interaction attribute – which was considered by far the most relevant attribute by the decision tree induction algorithms, as mentioned earlier – is as follows:

Rule 4:
IF WRN_interaction = yes
THEN class is ageing-related DNA repair gene

This rule was consistently found in all the 12 decision trees built by J4.8 and CART for all versions of the D4 and D5 datasets (for all GO term occurrence threshold values) – which are all datasets where the attribute WRN_interaction is included. There are 14 DNA repair genes satisfying the IF part of this rule, and all of them belong to the predicted class. This rule is highly statistically significant – $p = 2 \times 10^{-9}$, with the aforementioned test of statistical significance. Hence, this pattern can to some extent be interpreted as supporting the argument that WRN syndrome is a useful model for ageing research (Hasty and Vijg, 2004a), (Kipling et al., 2004).

However, it is well-known that the WRN protein is a hub (a node with a large number of neighbours) in ageing-related protein interaction networks (Magalhaes and Toussaint, 2004), (Beneke and Burkle, 2007), and so it is not surprising that interaction with the WRN protein is associated with ageing. What is somewhat surprising is that the rule has no false positives – i.e., there is no DNA repair gene in the dataset that interacts with WRN and belongs to the non-ageing-related class. It should also be noted that there is a certain bias in this result, since the WRN protein and its interaction partners tend to be more studied than other types of proteins.

A more interesting predictive pattern is represented by the following rule, which refers to interaction with the protein XRCC5 (X-ray repair complementing defective repair in Chinese hamster cells 5, also called KU 80).

Rule 5:
IF XRCC5_interaction = yes
THEN class is ageing-related DNA repair gene

This rule was extracted from two decision trees built by J4.8 for two versions of dataset D3 (with NumInter and 10 binary protein interaction attributes) with parameter GO term occurrence threshold set to 7 and 11. The IF part of this rule is a strong predictor of the ageing-related class, despite using only one attribute to make that prediction. More precisely, there are 11 DNA repair genes satisfying the IF part of this rule, and 10 of them belong to the predicted class. This rule is highly statistically significant – $p = 5.3 \times 10^{-6}$, with the aforementioned test of statistical significance.

XRCC5 is another DNA helicase, and is involved in double-strand-break repair. As discussed in Subsection 1.5.3, KU is a heterodimer composed of KU70 and KU80 subunits. When a double-strand break occurs, KU binds to DNA ends and recruits DNA-dependent protein kinase catalytic subunit, which is believed to phosphorylate and activate downstream targets in the non-homologous end joining (NHEJ) DNA repair pathway (Seluanov et al., 2007). Ku80 $^{-/-}$ mice, which are defective in double-strand DNA break repair via the NHEJ pathway, exhibit multiple symptoms of ageing (Hasty et al., 2003), (Ven et al., 2006).

So far we focused on rules predicting the "ageing-related DNA repair gene" class, which is the main class of interest in these experiments. However, it is also worth discussing a couple of rules predicting the "non-ageing-related DNA repair gene" class.

Analysing the decision trees built by J4.8 and CART, in general those decision trees include a leaf node predicting the "non-ageing-related" class for a large number (usually on the order of 90-100) of data instances. The vast majority of the data instances classified by those leaf nodes do have the predicted class, but usually a few instances in each node have the other class, "ageing-related", constituting therefore false positives for the rules extracted from those leaf nodes. This result can be partly explained by the fact that the

"non-ageing-related" class has many more data instances in the dataset than the ageing-related class; hence, it seems normal that it is difficult for J4.8 and CART to find rules with no false positive for such a large number of data instances. In any case, since the number of false positives in such rules is still much smaller than the number of true positives (data instances correctly predicted by the rule), the rules are still highly statistically significant, as shown in the next couple of rules, chosen as interesting examples of such rules.

The following rule was extracted from the decision tree built by J48 for dataset D1 (with no attribute related to protein-protein interaction information), with the GO term occurence threshold set to 11.

Rule 6
IF GO:0042221 (response to chemical stimulus) = no
AND GO:0009314 (response to radiation) = no
AND GO:0006302 (double-strand break repair) = no
THEN class is non-ageing-related DNA repair gene

There are 94 DNA repair genes satisfying the IF part of this rule, out of which 85 have the class predicted by the rule. This rule is highly statistically significant – $p = 0.000215$, with the aforementioned test of statistical significance.

In this rule the *absence* of the GO term "response to chemical stimulus" is associated with the class "non-ageing-related". This is consistent with the fact that, in a previous rule (Rule 3), the *presence* of that GO term was associated with the other class, "ageing-related". In the case of the GO term "response to radiation", its presence is not as strongly associated with the "ageing-related" class as the presence of the GO term "response to chemical stimulus", but the absence of the former term is still a good predictor of the "non-ageing-related class" by itself. Actually, an analysis of the relative frequency of each class for each of the attribute values in the rule shows that the condition "response to

radiation = no" is actually the strongest predictor of the "non-ageing-related" class among the three conditions in the rule.

It is also interesting to note that the absence of the GO term annotation "double-strand break repair" is associated with the class "non-ageing-related". This type of DNA repair seems indeed quite associated with ageing, as discussed in Subsection 1.5.3, and therefore it makes sense that the absence of such GO term annotation helps to predict the "non-ageing-related" class.

The following rule was extracted from the decision tree built by J48 for dataset D2 (with attribute NumInter but no binary protein interaction attribute), with the GO term occurrence threshold set to 7.

Rule 7
IF NumInter $\leq$ 15
AND GO:0006979 (response to oxidative stress) = no
AND GO:0009411 (response to UV) = no
AND GO:0006284 (base-excision repair) = no
THEN class is non-ageing-related DNA repair gene

There are 94 DNA repair genes satisfying the IF part of this rule, out of which 88 have the class predicted by the rule. This rule is highly statistically significant – $p = 6.9 \times 10^{-6}$, with the aforementioned test of statistical significance.

This rule indicates that relatively low values of the NumInter attribute are associated with the "non-ageing-related" class, which is consistent with rules indicating that relatively high values of that attribute are associated with the other class, "ageing-related" – see for example Rule 2. An analysis of the relative frequency of each class for each of the attribute values in the rule shows that none of the three conditions referring to the absence of GO terms is a strong predictor of the "non-ageing-related" class by itself – it is the synergistic combination of the absence of the three GO terms that gives the rule its good predictive power.

It should be noted that in the above examples of rules predicting the "non-ageing-related" class, the rules' IF parts consist mainly of conditions referring to the absence of GO term annotations, and the interpretation of such *no* values for the GO term-based attributes involves some caveats, as discussed earlier. This pattern of using mainly a combination of *no* values of GO term-based attributes to classify a large number of data instances in the "non-ageing-related" class also can be partly explained by the fact that such class has many more data instances than the "ageing-related" class. Rule conditions referring to the *yes* value of GO term-based attributes tend to be very selective (i.e., tend to be satisfied by relatively few data instances), and a conjunction of such conditions would tend to create a very specific rule, covering few data instances. By contrast, since rule conditions referring to the *no* value of GO term-based attributes tend to be less selective, they seem more appropriate to be used in rules covering a larger number of data instances from a statistical point of view, although at the expenses of biological interpretability.

## 4.2 RESULTS AND DISCUSSION FOR DATASETS WITH TWO CLASSES AND GENE EXPRESSION ATTRIBUTES

This section reports the results for a dataset containing the same two classes of DNA repair genes used in the experiments reported in the previous section, but containing a single type of predictor attribute, namely gene expression values – unlike the datasets with multiple types of attributes used in the experiments reported in the previous section. Since the creation of this dataset with gene expression values does not involve any parameter (unlike previous datasets), a single version of the dataset was created. Therefore, it is more practical to analyse the predictive accuracy results of all classification algorithms applied to this dataset in a single subsection, 4.2.1, rather than having one separate subsection for each algorithm as in the previous section. In addition, Subsection 4.2.2 discusses a classification rule extracted from a decision tree built for this dataset; and Subsection 4.2.3 discusses the results in a broader biological context,

integrating the classification rule discussed in Subsection 4.2.2 with some results presented in Section 4.1.

### 4.2.1 Predictive accuracies for J4.8, CART and Naive Bayes algorithms

The AUC values obtained by the J4.8, CART and Naive Bayes algorithms in this dataset were 51.1%, 45.8% and 52.1%, respectively. These values are much lower than the AUC values for the datasets with multiple types of attributes but not gene expression attributes (reported in Subsections 4.1.1 through 4.1.3).

Actually, the J4.8 and Naive Bayes algorithms obtained AUC values which are just slightly higher than the AUC value expected from random predictions, which is 50%; and the AUC value obtained by CART was even somewhat lower than 50%. Hence, the entire classification models built from this data are not reliable. However, part of the decision tree built by J4.8 corresponds to a reliable classification rule, as discussed in the next subsection.

### 4.2.2 Interpreting a rule extracted from the decision tree built by J4.8

Although the predictive accuracy of the entire decision tree built by J4.8 was not good, as reported earlier, the decision tree built by J4.8 contains a path corresponding to a classification rule which is a good predictor of ageing-related DNA repair genes. This rule – which is a modular component of the classification model that can be interpreted independent from the logical conditions in the rest of the decision tree – is as follows:

Rule 8:
IF (the gene's expression in T-lymphocyte > 6,265.926)
AND (the gene's expression in tongue_squamous_cell ≤ 11,127.391)
THEN class is aging-related DNA repair gene

The actual numerical gene expression values in the IF part of the rule are not easily interpretable, but we can interpret the conditions in the IF part of the rule by computing how many data instances (DNA repair genes) satisfy each of those conditions

individually. This will allow us to understand "how high" and "how low" are the threshold values in those conditions, in the context of the dataset used in the experiments.

The first condition, referring to the value of a gene's expression in T-lymphocyte, is a very selective condition; it is satisfied by only eight genes – out of 148 genes in the dataset. Therefore, the condition's threshold of 6,265.926 can be interpreted as a high value of gene expression in T-lymphocytes – since the condition selects relatively few genes with values greater than that threshold. In other words, that condition can be broadly interpreted as "IF (the gene's expression in T-lymphocyte is high)".

By contrast, the condition referring to the value of a gene's expression in tongue_squamous_cell has very low selectivity; it is satisfied by 143 out of the 148 genes in the dataset. Although this condition does not have a good predictive power by itself, this condition is useful to improve the predictive power of the above rule, as follows. If we consider a simpler rule with just one condition: "IF (the gene's expression in T-lymphocyte > 6,265.926)", this condition is satisfied by eight DNA repair genes, six of which are ageing-related (APEX1, ERCC5, POLB, RPA1, XRCC5, XRCC6) and two of which are non-ageing-related (DUT and TDG). By adding the condition "(the gene's expression in tongue_squamous_cell ≤ 11,127.391)" to the IF part of the rule, the set of genes satisfying the rule is reduced to five (APEX1, ERCC5, RPA1, XRCC5, XRCC6). Hence, adding the latter condition has the desirable effect of removing, from the rule coverage, the two non-ageing-related genes (DUT and TDG), as well as having the undesirable effect of removing one ageing-related gene (POLB). (Unfortunately the algorithm was not able to find any rule that does not cover DUT and TDG but still covers all the other aforementioned six ageing-related genes.)

Hence, the rule with the two above conditions – in its form discovered by J4.8 – covers five "true positives" (genes that are predicted by the rule to be ageing-related and actually belong to that class) and no "false positives" (genes that are predicted by the rule to be ageing-related and that do not belong to that class). This rule is highly statistically significant – $p = 0.00055$, with the aforementioned test of statistical significance.

In summary, the above rule can be broadly interpreted as: "IF a DNA repair gene has a high expression (by comparison with other DNA repair genes) in T-lymphocytes and has a gene expression value in tongue squamous cells that is not very high, then the gene is predicted to be ageing-related". The main condition in this rule – i.e. the condition which is better at predicting the ageing-related class – is a high expression in T-lymphocytes; the other condition was added to the rule by J4.8 only to make the rule more consistent with respect to the underlying dataset.

### 4.2.3 Integrating results for gene expression and other types of predictor attributes

In this subsection the results for the dataset with gene expression data only – focusing on the classification rule analysed in the previous subsection – are integrated with the results for the datasets with multiple types of predictor attributes discussed in Section 4.1. Recall that, although Sections 4.1 and 4.2 refer to datasets with very different and non-overlapping types of predictor attributes, the definitions of the classes to be predicted by the classification algorithms were basically the same for the experiments reported in both these sections.

The Ingenuity® software (Gerling et al., 2005), (Mayburd et al., 2006) – see also www.ingenuity.com – was used to define crosstalk between genes or proteins involved in the patterns that were discovered by the classification algorithms. The network automatically generated by Ingenuity is mainly composed by known protein-protein interactions (PPIs). In addition, known (and stored in Ingenuity database) links to physiological processes and diseases were imposed on the reconstructed network, shown in Figure 4.3[3]. In the figure, pink links connect proteins to the process of double-stranded DNA repair, green links connect proteins to the process of telomere maintenance, dark blue to T cell development, light blue to V(D)J recombination, and yellow to apoptosis.

As can be observed in Figure 4.3, XRCC5 (KU80) and XRCC6 (KU70) – two of the genes whose expression was observed to be high in T lymphocytes – have particularly

[3] We thank Dr. Olga Vasieva for using the Ingenuity software to generate the network shown in Figure 4.3 and for her help in the discussion of the results presented in this subsection.

important roles in this network. More precisely, XRCC6 is directly connected to all five aforementioned processes, whilst XRCC5 is directly connected to four of those processes (the only exception being V(D)J recombination).

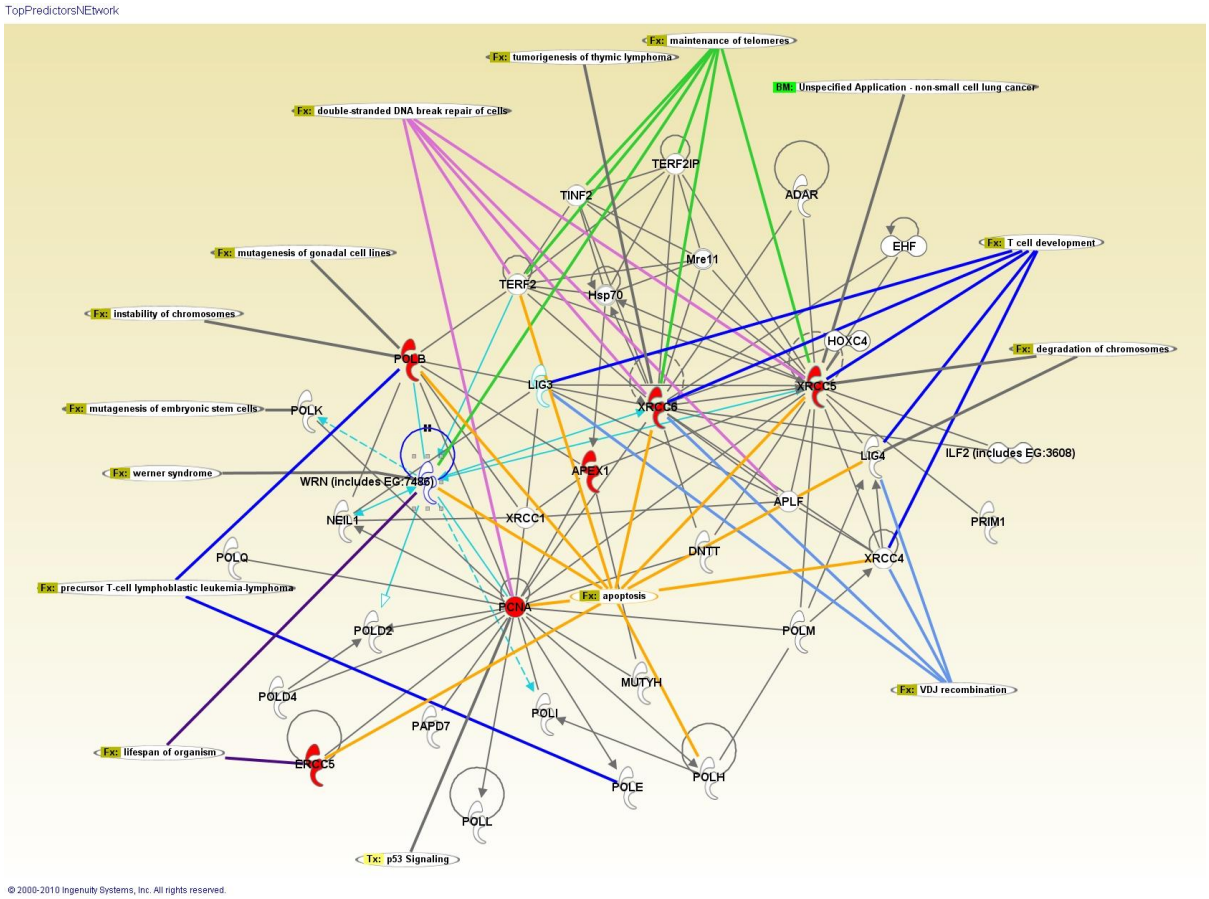


**Figure 4.3:** Network of genes or proteins and biological processes produced by using the Ingenuity tool to integrate results for gene expression and other types of attributes

As discussed in Subsection 1.5.3, KU80 and KU70 form the KU heterodimer, which has a crucial function in the initiation step of non-homologous end joining DNA repair pathway – binding to the broken DNA strand ends. KU is one of the most abundant proteins in human cells (Chai et al., 2002). Good evidence for the role of KU on ageing is provided by a number of investigations with mice knockouts (Vogel et al., 1999), (Zhu et al., 1996), (Li et al., 2007). In addition, the levels of expression of the protein KU70 were observed to significantly decline with age in humans (Ju et al., 2006), and KU70

expression was significantly higher in the group of subjects with higher longevity than in the control group used in that study.

As discussed in Subsection 4.2.2, the main pattern discovered in the gene expression dataset was that DNA repair genes having a high expression in T lymphocytes tend to be ageing-related genes. Among the genes satisfying this pattern are XRCC5 (KU80) and XRCC6 (KU70). In addition, as discussed in Subsection 4.1.4.3, interaction with XRCC5 and WRN are strong predictors of ageing-relatedness for DNA repair genes.

Integrating these patterns, it is interesting to note that WRN, XRCC5, XRCC6, and T lymphocytes are all related to Non-Homologous End Joining (NHEJ), an important pathway for the repair of double-strand DNA breaks (Burmaa et al., 2006). This process is required for proper telomere maintenance, and NHEJ is also required for joining double-strand breaks induced during V(D)J recombination (Rooney et al., 2004), (Rassool, 2003), the process that generates diversity in B-cell and T-cell receptors in the vertebrate immune system (Mombaerts et al., 1992).

Relations of DNA double-strand break frequency and telomere maintenance to ageing are established facts (Ju et al., 2006), (Ariyoshi et al., 2007). However, the link between aging and autoimmunity is more striking, although it is also discussed in the literature. Increased autoimmunity is observed in Down syndrome (Karlsson et al., 1998), (Prasher, 1999), which is also characterized by accelerated ageing (Arking, 2006) – including premature T-cell ageing (Vaziri et al., 1993), (Rabinowe et al., 1989). Increased autoimmunity was also shown to be associated with the normal ageing process (Prelog, 2006), (Hosaka et al., 1996).

Human T lymphocytes represent a well-characterized example of a cell type which retains the ability to up-regulate telomerase as part of their response to a proliferative stimulus. Interestingly, (James et al., 2000) have shown that *WRN* gene mutations do not significantly reduce the lifespan of T-cells, although such mutations are known to significantly reduce the lifespan of human fibroblasts (which cannot up-regulate

telomerase). The authors note this result is consistent with the hypothesis that *WRN* gene mutations reduce the lifespan of fibroblasts due to those mutations' deleterious effects in telomere maintenance; whilst in T-cells such deleterious effects are compensated by the up-regulation of telomerase.

## 4.3 RESULTS AND DISCUSSION FOR DATASETS WITH FOUR CLASSES AND MULTIPLE ATTRIBUTE TYPES

Unlike the experiments in the previous two sections, which involved only two classes, the experiments reported in this section are characterized by the use of datasets where there are four classes of data instances (genes). A detailed explanation about the creation of classes and attributes for these datasets is provided in Subsections 3.3.1 and 3.3.2, respectively; whilst the specification of the major characteristics of these datasets – such as the number and types of attributes in each dataset – is presented in Subsection 3.3.3.

In summary, these datasets have the following characteristics. Each data instance belongs to one out of four classes: "ageing-related DNA repair genes with evidence for ageing-relatedness in humans or mammals" (*Age-HM-DNA*), "non-ageing-related genes that interact with genes in the class *Age-HM-DNA*" (*Int-Age-HM-DNA*), "non-ageing-related DNA repair genes" (*Non-Age-DNA*), or "non-ageing-related genes that interact with genes in the class *Non-Age-DNA*" (*Int-Non-Age-DNA*). In addition, the datasets contain multiple types of predictor attributes, as follows. All datasets contain the attribute "$K_a/K_i$ ratio" (a measure of evolutionary change) and a number of GO term-based predictor attributes (each taking the value "yes" or "no" to indicate whether or not the gene is annotated with the corresponding GO term). The actual number of GO term-based attributes is determined by the parameter GO term occurrence threshold. We did experiments with three values of this threshold, namely 3, 7 and 11, which led to datasets with 560, 272 and 200 terms, respectively.

Furthermore, two types of datasets were created. Dataset type D7 has just the aforementioned attribute types, and no attribute related to protein-protein interaction information. By contrast, dataset type D8 has the aforementioned attribute types plus the attribute NumInter – the value of this attribute for each data instance (gene) is the number of proteins that interact with that gene's protein product.

The results and discussions presented in this section can be conceptually divided into two parts. In the first part, the results for each classification algorithm are discussed in a separate subsection (4.3.1 through 4.3.3), and the discussion focuses on the predictive accuracy obtained by each of the algorithms. In the second part, in subsection 4.3.4, we discuss the interpretation of the results in the light of biological knowledge, focusing on specific predictive patterns extracted from the decision trees built in the experiments – rather than on the predictive performance of each algorithm as a whole.

### 4.3.1 Results for the J4.8 decision tree induction algorithm

The predictive accuracies obtained by the J4.8 algorithm, measured by the AUC value, are reported in Table 4.5. As shown in the table, the inclusion of the predictor attribute NumInter in the dataset being mined (dataset D8) led to higher AUC values than the ones obtained in the dataset without that attribute (D7). The increase in the AUC values was negligible (0.1%) in the experiments with GO term occurrence threshold = 11, but it was more noticeable in the experiments with GO term occurrence threshold = 3 and 7 – where the increase in AUC value associated with the use of the attribute NumInter was 4.3% and 5.3%, respectively. This positive effect of the attribute NumInter is particularly interesting considering that this attribute is just one among hundreds of other attributes included in each dataset – more precisely, the number of GO term-based attributes in a dataset varies from 200 to 560 depending on the value of the GO term occurrence threshold, as shown in Table 4.5.

The highest value of AUC observed in Table 4.5, 76.3%, was obtained for all three versions of dataset D8, with the three different values of the GO term occurrence threshold. Despite the same AUC value obtained via the cross-validation procedure, the

decision trees built for these three versions of the dataset D8 are actually different – although they share some important parts, such as having the same predictor attribute at the root node.

**Table 4.5:** Area Under ROC curve (AUC, in %) for J4.8 algorithm, for datasets with four classes

| Dataset | PPI-related attributes | GO term occurrence threshold (number of GO terms) | | |
|---|---|---|---|---|
| | | 3 (560 terms) | 7 (272 terms) | 11 (200 terms) |
| D7 | none | 72.0 | 71.0 | 76.2 |
| D8 | NumInter | **76.3** | **76.3** | **76.3** |

As an example of a decision tree built in these experiments, Figure 4.4 shows the smallest of the decision trees built for a version of dataset D8, which is a tree having 16 leaf nodes. Like the tree in Figure 4.1, this tree is drawn using indentation (denoted by the "|" symbol) to indicate further levels of tree depth, the classes predicted by the leaf nodes are boldfaced, and the numbers in brackets after a predicted class denote, respectively, the number of data instances classified by that leaf node and (if there is a second number) the number of data instances wrongly classified by that leaf node, the so-called "false positives" for that leaf node. The absence of a second number means there are no false positives associated with that leaf node.

In this decision tree, the attribute selected for the root node was NumInter, which divided the tree into two subtrees: the one associated with data instances satisfying the condition "NumInter > 2" – the part of the tree at the bottom of Figure 4.4, and the subtree associated with the other data instances, satisfying the condition "NumInter ≤ 2". The former subtree is quite small, it contains only three leaf nodes, two of which are predicting the class *Age-HM-DNA* with no false positive; whilst the other leaf node predicts the *Non-Age-DNA* class with less precision (having eight false positives) – but covering many more data instances (78).

NumInter ≤ 2

| GO:0006308 (DNA catabolic process) = no

| | GO:0006281 (DNA repair) = no

| | | GO:0022403 (cell cycle phase) = no

| | | | GO:0033043 (regulation of organelle organization) = no

| | | | | GO:0022415 (viral reproductive process) = no

| | | | | | GO:0022414 (reproductive process) = no

| | | | | | | GO:0044093 (positive regulation of molecular function) = no

| | | | | | | | GO:0070647 (protein modif. by small protein conjugation or removal) = no

| | | | | | | | | GO:0009056 (catabolic process) = no: **Int-Age-HM-DNA** (158.0/50.0)

| | | | | | | | | GO:0009056 (catabolic process) = yes: **Int-Non-Age-DNA** (16.0/6.0)

| | | | | | | | GO:0070647(prot. modif. by small conj./rem.) = yes: **Int-Age-HM-DNA** (8.0)

| | | | | | | GO:0044093 (posit. regul. of molec. funct.) = yes: **Int-Non-Age-DNA** (12.0/2.0)

| | | | | | GO:0022414 (reproductive process) = yes: **Int-Non-Age-DNA** (7.0/1.0)

| | | | | GO:0022415 (viral reproductive process) = yes: **Int-Age-HM-DNA** (4.0)

| | | | GO:0033043 (regulation of organelle organization) = yes: **Int-Non-Age-DNA** (2.0)

| | | GO:0022403 (cell cycle phase) = yes

| | | | GO:0008152 (metabolic process) = no: **Non-Age-DNA** (5.0/1.0)

| | | | GO:0008152 (metabolic process) = yes: **Int-Non-Age-DNA** (4.0)

| | GO:0006281 (DNA repair) = yes

| | | GO:0042770 (DNA damage response, signal transduction) = no

| | | | GO:0050794 (regulation of cellular process) = no: **Non-Age-DNA** (15.0/1.0)

| | | | GO:0050794 (regulation of cellular process) = yes: **Int-Non-Age-DNA** (7.0)

| | | GO:0042770 (DNA damage response, signal transduction) = yes: **Int-Age-HM-DNA** (3.0)

| GO:0006308 (DNA catabolic process) = yes: **Non-Age-DNA** (5.0)

NumInter > 2

| GO:0042221 (response to chemical stimulus) = no

| | GO:0007050 (cell cycle arrest) = no: **Non-Age-DNA** (78.0/8.0)

| | GO:0007050 (cell cycle arrest) = yes: **Age-HM-DNA** (2.0)

| GO:0042221 (response to chemical stimulus) = yes: **Age-HM-DNA** (4.0)

**Figure 4.4:** Decision tree built by J4.8 for dataset D8 with GO term occurrence threshold = 7 in Table 4.5

The subtree below the condition "NumInter ≤ 2" is much larger, though, reflecting a greater difficulty in predicting classes of data instances satisfying that condition. Although many leaf nodes in that subtree make precise predictions without any false positives, there are also parts of that subtree where class predictions are much less reliable. In particular, in the first leaf node from the top of the tree, predicting class *Int-Age-HM-DNA*, that class is predicted for 158 data instances, but 58 of them have a different class, characterizing a wrong prediction.

A detailed analysis of several classification rules extracted from the decision trees produced in the experiments reported here will be presented in Subsection 4.3.4.

### 4.3.2 Results for the CART decision tree induction algorithm

The predictive accuracies obtained by the CART algorithm, measured by the AUC value, are reported in Table 4.6.

**Table 4.6:** Area Under ROC Curve (AUC, in %) for CART algorithm, in datasets with four classes

| Dataset | PPI-related attributes | GO term occurrence threshold (number of GO terms) | | |
|---|---|---|---|---|
| | | 3 (560 terms) | 7 (272 terms) | 11 (200 terms) |
| D7 | none | 72.5 | 73.6 | 72.9 |
| D8 | NumInter | 77.7 | 76.9 | **78.9** |

As can be seen in this table, for every value of the GO term occurrence threshold, the AUC measure is higher for the dataset where the NumInter attribute is included (dataset D8) than for the dataset where that attribute is missing (D7). This was the same overall trend observed in the results of the J4.8 algorithm, discussed in the previous section. However, this trend is somewhat stronger in Table 4.6, where the increase in AUC associated with the use of the attribute NumInter varies from 3.3% (for GO term occurrence threshold = 7) to 6% (for GO term occurrence threshold = 11).

The highest AUC value in Table 4.6 was 78.9%, and the decision tree built by CART associated with this AUC value is shown in Figure 4.5. This tree has 11 leaf nodes, whose predicted classes are boldfaced in the figure.

NumInter < 2.5
| GO:0006259 (DNA metabolic process) = yes
| | GO:0050794 (regulation of cellular process) = yes
| | | GO:0006281 (DNA repair) = yes
| | | | GO:0042770 (DNA damage response, signal transduc.) = yes: **Int-Age-HM-DNA** (3.0/0.0)
| | | | GO:0042770 (DNA damage response, signal transduct.) = no: **Int-Non-Age-DNA** (7.0/0.0)
| | | GO:0006281 (DNA repair) = no: **Int-Age-HM-DNA** (10.0/3.0)
| | GO:0050794 (regulation of cellular process) = no: **Non-Age-DNA** (20.0/6.0)
| GO:0006259 (DNA metabolic process) = no
| | GO:0044093 (positive regulation of molecular function) = yes: **Int-Non-Age-DNA** (11.0/2.0)
| | GO:0044093 (positive regulation of molecular function) = no
| | | GO:0022403 (cell cycle phase) = yes
| | | | GO:0008152 (metabolic process) = yes: **Int-Non-Age-DNA** (3.0/0.0)
| | | | GO:0008152 (metabolic process) = no: **Non-Age-DNA** (4.0/1.0)
| | | GO:0022403 (cell cycle phase) = no: **Int-Age-HM-DNA** (115.0/60.0)
NumInter ≥ 2.5
| NumInter < 31.5
| | GO:0042221 (response to chemical stimulus) = yes: **Age-HM-DNA** (3.0/0.0)
| | GO:0042221 (response to chemical stimulus) = no: **Non-Age-DNA** (70.0/7.0)
| NumInter ≥ 31.5: **Age-HM-DNA** (4.0/0.0)

**Figure 4.5:** Decision tree built by CART for dataset D8 and GO term occurrence threshold = 11 in Table 4.6

Like the decision trees in Figures 4.1 and 4.4 built by J4.8, the tree in Figure 4.5 built by CART is drawn using indentation (denoted by the "|" symbol) to indicate further levels of tree depth, and the classes predicted by the leaf nodes are boldfaced. There is, however, a difference in the meaning of the numbers in brackets after a predicted class – as they are produced by the WEKA data mining tool used in the experiments. In the trees built by

CART, such as the tree in Figure 4.5, the first number in brackets after a predicted class denotes the number of data instances *correctly* classified by that leaf node, rather than denoting the total number of data instances classified (correctly or not) by the leaf node as in the case of the trees built by J4.8. The second number in brackets – if it is is present – has the same meaning in both types of trees, denoting the number of data instances wrongly classified by that leaf node; and the absence of the second number means it is zero.

It is interesting to compare the decision tree in Figure 4.5 (most accurate tree built by CART) with the decision tree in Figure 4.4 (most accurate tree built by J4.8). Although the two algorithms use different procedures for building a decision tree, there are some interesting broad similarities between those two decision trees, as follows. In both trees, the predictor attribute selected for the root node was NumInter, which divided the tree into two subtrees: the one associated with data instances having a small value of NumInter – the part of the tree at the top of the corresponding figures – and the subtree associated with the other data instances, having larger values of NumInter – at the bottom of the corresponding figures. Also, in both trees the condition based on the NumInter value used to create the subtrees has the same logical meaning in practice – since the NumInter attribute can take only integer values, the condition "NumInter < 2.5" used by CART is logically equivalent to the condition "NumInter ≤ 2" used by J4.8. Furthermore, in both trees (in Figures 4.4 and 4.5), the subtrees associated with the condition "NumInter > 2" or "NumInter ≥ 2.5" have three leaf nodes, two of which predict the class *Age-HM-DNA* with no false positive for a small number of data instances (2-4), and the other leaf node predicts the class *Non-Age-DNA* with a relatively small number of false positives (7-8) for a much larger number of data instances (77-78). Moreover, both subtrees use the GO term "response to chemical stimulus". The main difference between those two bottom subtrees is that CART uses a further condition based on NumInter, whilst J4.8 uses the GO term "cell cycle arrest". Despite this difference, it is clear that both subtrees are broadly similar in structure, selected predictor attributes and class predictions.

The top subtrees in Figures 4.4 and 4.5 – i.e, the subtrees rooted at the conditions "NumInter ≤ 2" or "NumInter < 2.5", respectively – show a larger difference; but they still present some similarities, like the selection of the attributes based on the GO terms "DNA catabolic process" and "DNA repair" at shallow levels of the tree – where attributes have more relevance than at deep levels, as discussed earlier.

### 4.3.3 Results for the Naive Bayes algorithm

The predictive accuracies obtained by the Naive Bayes algorithm, measured by the AUC value, are reported in Table 4.7. As shown in the table, the inclusion of the predictor attribute NumInter in the dataset being mined (dataset D8) led to slighthly (around 2.5%) higher AUC values than the ones obtained in the dataset without that attribute (D7) when the GO term occurrence threshold was set to 3 or 7. However, there was a negligible reduction (0.4%) in the AUC values obtained for datasets D7 and D8 when the GO term occurrence threshold was set to 11.

**Table 4.7:** Area Under ROC curve (AUC, in %) for Naive Bayes, for datasets with four classes

| Dataset | PPI-related attributes | GO term occurrence threshold (number of GO terms) | | |
|---|---|---|---|---|
| | | 3 (560 terms) | 7 (272 terms) | 11 (200 terms) |
| D7 | none | 76.2 | 76.1 | 76.5 |
| D8 | NumInter | 78.8 | 78.4 | 76.1 |

These results regarding the influence of the NumInter attribute in the predictive accuracy of Naive Bayes are in contrast with the results shown in Tables 4.5 and 4.6, where the NumInter attribute was found to have a greater influence in the predictive accuracy of the J4.8 and CART decision tree induction algorithms, respectively. A possible explanation for this difference in the results between these two types of classification algorithms is that Naive Bayes uses all predictor attributes to classify new data instances, whilst decision tree induction algorithms select just a relatively small subset of attributes to include in the decision tree that will be used to classify new data instances. Hence, intuitively one would expect the influence of a single attribute to be smaller in the case of Naive Bayes, since the influence of that attribute will be more "diluted" among the

hundreds of predictor attributes used by the algorithm for classification. Conversely, one would expect the influence of a single attribute to be larger in the case of decision trees, since the influence of that attribute will be less "diluted" among a relatively small subset of attributes used for classification. This seems to be the case considering in particular that, as will be discussed in the next subsection, the NumInter attribute was selected to be a root node in a decision tree for every dataset where that attribute was present, and root node attributes have a disproportionally high influence in the classification of new instances by a decision tree, as discussed in Section 2.5.1.

### 4.3.4 Discussion on predictive patterns extracted from the decision trees

In this subsection we discuss two kinds of predictive patterns extracted from the decision trees built by J4.8 or CART, namely the distribution of attributes selected as root nodes of the decision trees (discussed in Subsection 4.3.4.1) and selected classification rules extracted from the decision trees (discussed in Subsection 4.3.4.2).

#### *4.3.4.1 Discussion on attributes chosen as root nodes in the decision trees*

Table 4.8 shows how many times each attribute was selected to be at the root node of the decision tree, for all attributes which were selected as root node in at least one decision tree in the experiments with datasets with four classes. The structure of this table is the same as the structure of Table 4.4 (which referred to datasets with two classes).

**Table 4.8:** Frequency of selection as a root node attribute in decision trees for datasets with four classes

| Attribute | Frequency |
|---|---|
| NumInter | 6 (out of 6) |
| GO:0006259 (DNA metabolic process) | 5 (out of 12) |
| GO:0007568 (aging) | 1 (out of 4) |

As shown in the table, the NumInter attribute was selected as the root node for every dataset where that attribute was available, indicating the great relevance of this attribute for classification purposes. The values of this attribute are also easily interpretable, as will be seen later in the discussion of some rules using this attribute. The attribute based on the

GO term “DNA metabolic process” was chosen as the root attribute for almost all (5 out of 6) datasets where the NumInter attribute was not available, which also indicates the high relevance of this attribute for classification purposes.

The only other attribute chosen as root node – and in just one dataset, out of the four datasets where it was available – was the attribute based on the GO term “aging”. This GO term is annotated for six genes included in our datasets, and so it was available only in the dataset versions when the parameter GO term occurrence threshold was set to 3, and not 7 or 11. An example of a rule referring to this GO term will be discussed in the next subsection.

*4.3.4.2 Discussion on selected rules extracted from decision trees*

Let us start with the discussion of rules predicting the class “ageing-related DNA repair gene, with evidence in humans or mammals” (*Age-HM-DNA*). One can extract from the decision trees several simple rules – with just one or two conditions each – predicting this class.

First, the following rule was consistently discovered by J4.8 in all three versions of the dataset D8, for the three different values of the GO term occurrence threshold (3, 7, 11).

Rule 9:
IF NumInter > 2
AND GO:0042221 (response to chemical stimulus) = yes
THEN class = *Age-HM-DNA*

There are four DNA repair genes satisfying the IF part of this rule, and all of them belong to the predicted class. This rule is highly statistically significant – $p = 5.6 \times 10^{-7}$, with the aforementioned test of statistical significance.

In addition, the following very simple rule was consistently discovered by CART in all three versions of the dataset D8, for the three different values of the GO term occurrence threshold.

Rule 10:
IF NumInter ≥ 31.5
THEN class = *Age-HM-DNA*

There are four DNA repair genes satisfying the IF part of this rule, and all of them belong to the predicted class. This rule is highly statistically significant – $p = 5.6 \times 10^{-7}$, with the aforementioned test of statistical significance.

In Rule 9 the main condition (the stronger predictor of the class) is "response to chemical stimulus = yes". This condition was also used as a condition in Rule 3, predicting the "ageing-related DNA repair" class in the experiments with two classes (Subsection 4.1.4.3); which reinforces the relevance of this condition.

The other condition in Rule 9 – i.e, "NumInter > 2" – is a very broad condition, given the conservatively low value of the threshold 2. It is true that NumInter can be a good predictor of the class *Age-HM-DNA* by itself, but this requires much higher values of this attribute. This is shown by Rule 10, where the fact that a DNA repair gene's protein product interacts with at least 32 proteins is sufficient - without the help of other attributes – to predict the class *Age-HM-DNA* without any false positives.

The following rule is found in two decision trees built by J4.8 for two versions of the dataset D7 (with no attributed related to protein-protein interaction) where the GO term occurrence threshold was set to 7 and 11.

Rule 11:
IF GO:0006259 (DNA metabolic process) = yes
AND GO:0040008 (regulation of growth) = no
AND GO:0048584 (positive regulation of response to stimulus) = yes
AND GO:0006325 (chromatin organization) = no
THEN class = *Age-HM-DNA*

There are four DNA repair genes satisfying the IF part of this rule, and all of them belong to the predicted class. This rule is highly statistically significant – $p = 5.9 \times 10^{-7}$, with the aforementioned test of statistical significance.

This rule is relatively complex, having four conditions in its IF part, all of them referring to GO terms. As discussed in Subsection 4.1.4.2, for binary attribute such as the attributes referring to the presence or absence of GO terms, conditions with the value "yes" tend the be more informative than conditions with the value "no". Hence, out of the four conditions in this rule, the most informative ones seem to be the two conditions having the value "yes" for their corresponding attributes, namely the conditions "DNA metabolic process = yes" and "positive regulation of response to stimulus = yes".

Each of the three previous rules has four true positives and no false positives – i.e, there are four data instances satisfying the IF part of the rule, and all of those instances represent genes belonging to the class "*Age-HM-DNA*". At first glance, one criticism that could be made to those three rules is that they cover such a small number of data instances (genes) that they might not have good generalization ability. However, there are two factors that contribute to the relevance of these rules.

First, one should recall that there are only nine instances of the *Age-HM-DNA* class in each dataset, and therefore, in the context of such a rare class, each of these rules has a relatively good coverage, covering approximately 44% (4 / 9) of the data instances of that class. Secondly, despite their low coverage, the rules are highly statistically significant, having $p$ values on the order of $10^{-7}$. The rules have such very small p values – despite covering few data instances – due to the fact that the probability of observing the class predicted by the rules ("ageing-related DNA repair genes with evidence in humans or mammals") in the dataset as a whole, without using the conditions in the IF parts of the rules to make more focused predictions, is very small – approximately 3%.

Let us now show an example of a rule predicting a different class, namely the class "interact with ageing-related DNA repair gene having evidence in humans or mammals"

(*Int-Age-HM-DNA*). The following rule has been extracted from the decision tree built by J4.8 in dataset D7 (with no attribute related to protein-protein interaction) with the parameter GO term occurrence threshold set to 3.

Rule 12:
IF GO:0007568 (aging) = no
AND GO:0006283 (transcription-coupled nucleotide-excision repair) = no
AND GO:0006259 (DNA metabolic process) = yes
AND GO:0040008 (regulation of growth) = yes
THEN class = *Int-Age-HM-DNA*

There are four DNA repair genes satisfying the IF part of this rule, and all of them belong to the predicted class. This rule is not as highly statistically significant as the other rules discussed earlier – $p = 0.027$, using the aforementioned test statistical of statistical significance. This $p$ value is not enough to reject the null hypothesis that the rule has no predictive power using a significance level of 1%, but it is small enough to give good reason to believe that the rule is unlikely to satisfy the null hypothesis of having no predictive power.

This rule makes a prediction based on the absence of two GO term annotations and the presence of two other GO term annotations for the gene being classified. As discussed earlier, in general attribute values representing the presence of GO terms tend to be more informative and easier to interpret. Indeed, the conditions "DNA metabolic process = yes" and "regulation of growth = yes" have a straightforward interpretation – although the former, in particular, is a very broad term.

In the case of this particular rule, however, even the attribute values representing the absence of GO terms can be interpreted and seem relevant for understanding the predictive pattern represented by the rule, as follows. The first condition referring to an absent GO term annotation is "aging = no". This makes sense, since the rule is predicting

the class *Int-Age-HM-DNA* and one of the criteria for a gene to belong to this class is that it should not be an ageing-related gene.

At this point one could, perhaps, argue that the use of the GO term "aging" in this rule represents a trivial part of the rule, given the class definition. However, it should be noted that, before doing the experiments, it was not clear if some value of the GO term "ageing" would have enough predictive power to be included in some rule. Actually, the GO term "ageing" was not used at all in any of the decision trees built in the experiments with two classes reported in Section 4.1, even though that GO term was included in many of the datasets used in those experiments. In addition, the class definition that we used to create the datasets used in the experiments referred to the class "ageing-related" as defined by the presence of the gene in the GenAge database; which is not the same criterion used by the Uniprot curators to decide if a gene is annotated with the GO term "aging". Hence, although the condition "aging = no" in this rule is expected and makes sense, it should *not* be considered a "trivial" condition.

The second condition referring to an absent GO term annotation in the above rule is "transcription-coupled nucleotide-excision repair = no". As discussed in Subsection 1.5.2, the two major types of nucleotide excision repair (NER), namely transcription-coupled and global genome NER, are usually associated with different dysfunctions. Defects in transcription-coupled NER genes – which are responsible for repairing DNA lesions that block transcription (Ljungman and Lane, 2004) – are often associated with ageing. In particular, defects in transcription-coupled NER genes cause the progeroid syndromes Trichothiodystrophy and Cockayne Syndrome (Niedernhofer et al., 2006). By contrast, defects in global-genome NER genes are mainly associated with cancer, since those genes are mainly responsible for preventing carcinogenesis by repairing pre-mutagenic DNA lesions. Hence, considering again that the rule's predicted class does not include any ageing-related gene, it makes sense to have in the rule a condition referring to the absence of the GO term "transcription-coupled nucleotide-excision repair".

It is also worth discussing here a few rules predicting the class "non-ageing-related DNA repair" (*Non-Age-DNA*), to contrast such rules with rules predicting the class *Age-HM-DNA*.

The following rule was extracted from the two decision trees built by CART in dataset D8 with the GO term occurrence threshold set to 3 and 7.

Rule 13:
IF $2.5 \leq \text{NumInter} \leq 31.5$
AND GO:0006979 (response to oxidative stress) = no
THEN class = *Non-Age-DNA*

There are 77 data instances satisfying the IF part of this rule, and 70 of them belong to the predicted class. This rule is highly statistically significant – $p = 2.9 \times 10^{-15}$, using the aforementioned test statistical of statistical significance.

The following rule was extracted from the decision tree built by CART in dataset D8 with the GO term occurrence threshold set to 11.

Rule 14:
IF $2.5 \leq \text{NumInter} \leq 31.5$
AND GO:0006979 (response to chemical stimulus) = no
THEN class = *Non-Age-DNA*

Again, there are 77 data instances satisfying the IF part of this rule, and 70 of them belong to the predicted class. This rule is highly statistically significant – $p \approx 0$, using the aforementioned test statistical of statistical significance.

Both Rule 13 and Rule 14 have, in their IF part, a condition referring to a moderate value of the NumInter attribute and another condition referring to the absence of response to either oxidative stress or chemical stimulus. Note that NumInter values between 3 and 31 are considered "moderate", rather than large, in the context of our datasets, where there

are several genes – in general belonging to the class *Age-HM-DNA* – with higher values for this attribute. The condition "response to chemical stimulus = no" also occurred in Rule 6; whilst the condition "response to oxidative stress = no" also occurred in Rule 7. Both those rules (reported in Subsection 4.1.4.3, for datasets with two classes) predict the class "non-ageing-related DNA repair gene" – which is essentially the same as the class *Non-Age-DNA* predicted by Rule 14. This reinforces the robustness of these predictive patterns.

Finally, the following rule was extracted from the two decision trees built by J4.8 in the dataset D8 with the GO term occurrence threshold set to 7 and 11.

Rule 15:
IF NumInter $\leq 2$
AND GO:0006308 (DNA catabolic process) = no
AND GO:0006281 (DNA repair) = yes
AND GO:0042770 (DNA damage response, signal transduction) = no
AND GO:0050794 (regulation of cellular process) = no
THEN class = *Non-Age-DNA*

There are 15 data instances satisfying the IF part of this rule, out of which 14 have the class predicted by the rule. This rule is highly statistically significant – $p = 1.6 \times 10^{-7}$, using the aforementioned test of statistical significance.

This is another rule that confirms that in general low values of the NumInter attribute are associated with non-ageing-related DNA repair genes – whilst large values of that attribute are associated with ageing-related DNA repair genes. In addition, this rule characterizes non-ageing related DNA repair genes not only by the trivial condition "DNA repair = yes", but also by three non-trivial conditions referring to the absence of the GO terms "DNA catabolic process", "DNA damage response, signal transduction" and "regulation of cellular process". An analysis of the relative frequency of each class

for each of the attribute values in these three non-trivial conditions of the rule shows that the condition "regulation of cellular process = no" is the strongest predictor of the "*non-ageing-related*" class among those conditions.

# Chapter 5 – Conclusions

## 5.1 CONTRIBUTIONS

This thesis proposed the use of classification algorithms – a type of predictive data mining algorithm – to discover gene properties that are effective in discriminating ageing-related DNA repair genes from non-ageing-related DNA repair genes and genes that interact (in the sense of protein-protein interaction) with DNA repair genes. We focused on DNA repair genes due to their strong association with the process of ageing, as discussed in Subsection 1.2.2. This research presents the following contributions to the relatively new field of bioinformatics-based ageing research.

First, to the best of our knowledge, this is the first work to propose the use of classification algorithms to investigate the relationship between DNA repair genes and the process of ageing; and, more generally, one of the first projects to propose the use of classification algorithms to discriminate between ageing-related and non-ageing-related genes. We described in detail the methodology for the creation of the datasets for the target classification task, as well as the corresponding experimental set up, in Chapter 3. The methodology described in that chapter could be useful for other researchers interested in applying classification algorithms to the analysis of ageing-related data.

Secondly, we created a number of datasets for the classification task of discriminating ageing-related DNA repair genes from other types of genes. In total 22 dataset versions were created, consisting in general of multiple types of predictor attributes. These datasets are available from the author on request.

Thirdly, many predictive patterns discovered by the classification algorithms – in particular, a number of highly statistically significant classification rules – were discussed, in Chapter 4, in the light of biological knowledge. A summary of this type of

result is also presented in Section 5.2. Some results reported in Sections 4.1 and 4.2 have been published in (Freitas et al., 2011).

Fourthly, we presented a review of related bioinformatics work on analysing ageing-related gene or protein networks, in Section 2.3. To the best of our knowledge there is no published review paper focusing on this topic yet.

## 5.2 SUMMARY OF THE MAIN DISCOVERED PREDICTIVE PATTERNS

The experiments involved many datasets with different combinations of types of predictor attributes. Hence, in this section we focus on identifying the types of predictor attributes and specific combinations of attribute values that seem biologically meaningful and had the greatest predictive power in the classification task of discriminating between ageing-related DNA repair genes and other types of genes.

Broadly speaking, the attributes that had the greatest predictive power – as evaluated by their role in the decision trees and corresponding classification rules produced in the experiments – were protein-protein interaction (PPI)-related attributes (Subsection 3.1.5), followed by GO term-based attributes (Subsection 3.1.4). The other types of attributes investigated in this research, namely the type of DNA repair function according to Wood's taxonomy (Subsection 3.1.2), a measure of the rate of evolutionary change based on the $K_a/K_i$ ratio (Subsection 3.1.3), and gene expression attributes (Section 3.2) turned out to have little predictive power in general. In the case of the type of DNA repair function and the $K_a/K_i$ ratio, each of these is a single attribute, and so it does not seem fair to compare their effectiveness with the sets of PPI-related and GO term-based attributes, which involved at least tens of attributes of each of these two types. However, the set of gene expression attributes consisted of 109 attributes, a number on the same order of magnitude than the number of GO term-based attributes in some datasets, and considerably greater than the number of PPI-related attributes. Hence, the performance of the gene expression attributes was disappointing, which may be a result of the well-

known fact that this type of attribute is usually associated with high levels of noise, which tends to be a significant problem for data mining methods (Aris et al., 2004).

Overall, considering all experiments where GO term-based attributes and protein-protein interaction (PPI)-related attributes were used – i.e., considering all experiments reported in Sections 4.1 and 4.3 – PPI-related attributes were selected to be decision tree root nodes (associated with the attributes having the greatest predictive power) much more often than GO term-based attributes, even though there were much fewer PPI-related attributes than GO term-based attributes in each of the datasets. Hence, let us first discuss the predictive power of PPI-related attributes.

In the experiments with four classes, reported in Section 4.3, no predictor attributes representing binary protein interactions were used. In this scenario, the only PPI-related attribute used, NumInter, was selected to be a decision tree root node in 100% (6 out 6) of the datasets where that attribute was available. In the experiments with two classes, in Section 4.1, the NumInter attribute was selected less often as root node due to the strong "competition" with other PPI-related attributes; but the NumInter attribute still had enough predictive power to be selected in 33% (8 out of 24) of the datasets where that attribute was available.

In the decision trees produced in the experiments, in general, larger values of NumInter were observed to be associated mainly with ageing-related DNA repair genes, whilst smaller values of NumInter were observed to be associated mainly with non-ageing-related genes. This result is consistent with other investigations – which did *not* focus on DNA repair genes – showing that ageing-related proteins tend to have a higher number of interaction partners than non-ageing-related proteins (Ferrarini et al., 2005), (Budovsky et al., 2007), (Promislow, 2004).

It is important to note that our research and these other investigations obtained this result by using two very different kinds of methods, and they produced different types of information, as follows. In (Budovsky et al., 2007), (Ferrarini et al., 2005), (Promislow,

2004), this result was obtained by simply counting the number of neighbours of each protein in a protein-protein interaction network. This produces information in the form of the number of interaction partners for each protein, which is effectively the same information captured in the NumInter attribute used in our datasets.

However, in our research this information is part of the *input* (as a predictor attribute) to a sophisticated classification algorithm; whilst in the aforementioned other investigations that information is the *output* of their system – i.e., it is simply reported to the user, for her or his interpretation. By using the number of interaction partners for each protein as one of the predictor attributes for a classification algorithm, our approach goes somewhat further than those other investigations, as follows.

First, the classification algorithms – in particular the decision tree building algorithms used in this research – are able to evaluate the predictive power of the NumInter attribute among a number of other attributes, and automatically decide to what extent the NumInter should be used for classification purposes. In other words, our data mining approach does not only tell us that larger values of NumInter are associated with ageing-related DNA repair genes, but also tell us that NumInter is in general a better predictor of ageing-relatedness than other types of attributes such as GO term-based attributes. This comparison of the predictive power of different types of attributs is not presented in those other investigations.

In addition, in our approach the information incorporated in the NumInter attribute can be combined with other types of information incorporated in other attributes for classification purposes – several examples of classification rules using both the NumInter attribute and some GO term-based attributes were discussed in Chapter 4. Furthermore, the classification rules extracted from the decision trees refer to specific cut-off (threshold) values of NumInter that were automatically found to be useful for discriminating between the "ageing-related DNA repair" class and other classes. Such cut-off values are dynamically determined by the algorithms for each dataset, since the optimal choice of cut-off value depends on the class definitions, the data instances and the

full set of predictor attributes in the dataset being mined. By contrast, in (Budovsky et al., 2007), (Ferrarini et al., 2005), (Promislow, 2004) there is no such automatically determined cut-off value of NumInter for predicting that a gene is ageing-related.

Out of all binary protein interaction attributes, from a statistical perspective, by far the most relevant one for discriminating ageing-related DNA repair genes from other classes of genes was the attribute WRN_interaction, which takes the value "yes" or "no" for a given gene depending on whether or not that gene's protein product interacts with the WRN protein – whose defect leads to the Werner's progeroid syndrome in humans (Subsection 1.2.1.1). This attribute was selected as the root node of the decision tree in 100% (12 out of 12) of the datasets where this attribute was available. The "yes" value of this attribute is a very strong predictor of the ageing-related DNA repair class. From a biological perspective, as discussed in Subsection 4.1.4.3, the predictive power of this attribute can to some extent be interpreted as supporting the argument that Werner's syndrome is a useful model for ageing research (Hasty and Vijg, 2004a, Davis and Kipling, 2006).

However, there are two caveats about the biological interpretation of the predictive power of this attribute. First, this predictive power is not surprising, since it is well-known that the WRN protein has many interaction partners, as a result of other investigations about protein-protein interaction networks – see for example (Magalhaes and Toussaint, 2004), (Beneke and Burkle, 2007). Secondly, there is a certain bias in the creation of the datasets, since the WRN protein and its interaction partners tend to be more studied than other types of proteins.

A more interesting, more surprising pattern discovered by the decision tree induction algorithms is that interaction with the XRCC5 (KU80) protein is also a strong predictor of the ageing-related DNA repair class – as discussed in Subsection 4.1.4.3. The XRCC5 protein plays an in important role in non-homologous end joining, an error-prone DNA repair pathway whose association with ageing was discussed in Subsection 1.5.3.

Let us now turn to the discussion of the predictive power of GO term-based attributes. The results reported in Sections 4.1 and 4.3 indicate that individual GO terms have, in general, less predictive power than individual PPI-related attributes. However, there are many GO term-based attributes whose predictive power was good enough to allow them to be selected for inclusion – sometimeas as root nodes – in the decision trees produced in the experiments.

In the experiments with two classes reported in Section 4.1, the only GO term-based attributes that were selected to be root nodes of decision trees were "response to endogenous stimulus", selected in 30% (3 out of 10) of the datasets where that attribute was used; "response to chemical stimulus", selected in 6.7% (2 out of 30) of the datasets; and "positive regulation of nucleobase, nucleoside and nucleic acid metabolic process", also selected in 6.7% of the datasets.

In general the value "yes" for each of these three attributes turned out to be a good predictor of the class "ageing-related DNA repair". This was observed, for instance, in some highly statistically significant classification rules whose IF part included the condition "response to endogenous stimulus = yes" or "response to chemical stimulus = yes", as reported in Subsection 4.1.4.3.

Importantly, in some rules, the presence of a single stimulus-related GO term in the rule's IF part was not enough to create a very precise rule, but the algorithms combined one of those conditions with another condition – based, for example, on a number of interaction partners (represented by attribute NumInter) greater than a certain automatically-adjusted threshold or based on the presence of a GO term related to (positive) regulation of biological process – in order to create a very precise rule.

In the experiments with four classes reported in Section 4.3, the GO term "DNA metabolic process" was selected as a decision tree root node in 42% (5 out of 12) of the datasets where that attribute was used. If we consider only the datasets where no PPI-related attribute was used, then the selection frequency of that GO term as a root node

increases to 83% (5 out of 6) of the datasets. In the experiments with two classes, where all datasets versions contain only DNA repair genes, this GO term was not selected as root node because it is too generic, being annotated to 130 DNA repair genes. However, in the experiments with four classes, in all datasets, the majority of the data instances represent genes other than DNA repair genes, and in this context this GO term acquires a strong predictive power, since it is very useful to discriminate between DNA repair genes and other types of genes.

In the experiments with four classes, the only other GO term-based attribute selected as a decision tree root node was "aging", which was selected as a root node in 25% (1 out of 4) of the datasets where that attribute was used. Interestingly, this GO term was never selected to be included in any of the decision trees produced in the experiments with two classes reported in Section 4.1 – i.e. its predictive power is restricted to the dataset with four classes. As mentioned in Subsection 4.3.4.2, the criterion used to consider a gene as ageing-related in this work is basically its inclusion in the GenAge database, which is presumably different from the criteria used by curators to annotate the GO term "aging" to a given gene in the Uniprot database – the source of GO term annotations for the creation of GO term-based attributes in this research. Actually, in our datasets only six data instances have the values "yes" for the attribute based on the GO term "aging", even though there are 33 DNA repair genes included in the GenAge database, nine of which with a strong reason for their inclusion in the database – namely, evidence directly linking the gene to ageing in humans or a mammalian model organism. Therefore, it seems that the Uniprot curators responsible for the annotation of the GO term "aging" tend to be considerably more conservative in considering a gene as an ageing-related one.

In terms of the classification rules extracted from decision trees, out of the previously mentioned GO terms whose value "yes" turned out to be a good predictor of the class "ageing-related DNA repair" in the experiments with two classes, the GO term "response to chemical stimulus" deserves to be highlighted because it was also strongly predictive enough to appear in some highly statistically significant rules produced in the experiments with four classes – involving a stricter definition of the class "ageing-related DNA repair

gene" that consists only of genes with direct evidence linking them to ageing in humans or mammals.

By contrast, the GO term "response to endogenous stimulus", which was a good predictor of the class "ageing-related DNA repair" in the experiments with two classes, did not turn out to be a good predictor in the experiments with four classes. This is because the majority of the DNA repair genes in GenAge annotated with this GO term were included in that database for reasons other than evidence in humans or mammals, and therefore those genes do not count as positive examples of the "ageing-related" class in the experiments with four classes.

So far we discussed the most predictive GO terms in the experiments with two and four classes separately because the class "ageing-related DNA repair" was defined in substantially different ways in those two experiments. However, the class "non-ageing-related DNA repair" was defined in the same way in the experiments with both two and four classes, so we now discuss the GO terms that were relevant predictors of this class.

First of all, it should be recalled that most classification rules predicting this class contained, in their IF part, conditions referring to the absence of GO terms (value "no" for the attribute), whose interpretation is subject to some caveats, as discussed in Subsection 4.1.4.2. In any case, the conditions "response to chemical stimulus = no" and "response to oxidative stress = no" occurred in rules extracted from the experiments with both two and four classes. It is also interesting to note that these patterns are consistent with the fact that the opposite value ("yes") of these attributes are in general good predictors of the opposite class, "ageing-related DNA repair". (The condition "response to oxidative stress = yes" is not among the previously reported most predictive conditions of the class "ageing-related DNA repair" simply because there are other GO terms that are stronger predictors of that class; but that condition is nonetheless a reasonably good predictor of that class.) Again, in general the aforementioned conditions often had to be combined with other conditions – usually relatively low values of the attribute NumInter and other GO terms – in order to create a rule with strong predictive power.

Finally, although gene expression attributes in general had poor predictive accuracy, the condition "IF gene expression in T lymphocytes is high" turned out to be a good predictor of the class "ageing-related DNA repair". This pattern was integrated with other patterns – referring to other types of attributes – in a discussion involving DNA double-strand break repair by the non-homologous end joining (NHEJ) pathway, this pathway's role in joining double-strand breaks during V(D)J recombination in lymphocytes (the process that generates diversity in T cell receptors in the vertebrate immune system) and the increased autoimmunity that is associated with ageing. For details of this discussion, see Subsection 4.2.3.

In summary, a number of highly statistically significant classification rules manually selected from the automatically-built decision trees, representing patterns with a strong predictive power, were discussed in the light of biological knowledge. Considering the results as a whole, these rules provide evidence that many DNA repair genes can be discriminated into two classes, ageing-related versus non-ageing related ones, based on several relevant gene properties. Broadly speaking, the main gene properties that were found effective in such a class discrimination task were as follows: ageing-related DNA repair genes' protein products tend to interact with a considerably larger number of proteins; their protein products are much more likely to interact with WRN and XRCC5; they are more likely to be involved in response to chemical stimulus and, to a lesser extent, in response to endogenous stimulus or oxidative stress; and they are more likely to have high expression in T lymphocytes.

## 5.3 FUTURE RESEARCH DIRECTIONS

One natural direction for future research would be to extend the set of predictor attributes in the target classification task to include other types of attributes, such as protein domains and other types of protein properties available in the Uniprot database. Protein domains, in particular, seem a potentially relevant type of predictor attribute because each

domain is usually associated with a well-defined biological function and domains can be regarded as "higher-level" building blocks" – rather than the "low-level" building blocks of amino acids – from which proteins are created. Actually, it has even been argued that protein domains, rather than genes, are the fundamental units of evolution (Nagl, 2003). However, it is possible that the information represented by protein domains would be to a significant extent redundant with respect to the GO term-based attributes, which could limit the predictive power of protein domains as additional attributes.

In addition, the use of gene expression attributes was under-explored in this research, by comparison with the use of PPI-related and GO term-based attributes – since only one dataset was produced based on the former, while many different dataset versions were produced based on the latter two types of atributes. Hence, in future research it would be interesting to create more datasets based on gene expression values for investigating the relationship between DNA repair genes and ageing. The caveat is that, in the experiments with the dataset using gene expression data as predictor attributes, all three classification algorithms used in this research obtained a low predictive accuracy (around the accuracy expected from random predictions), when measuring the accuracy of a classification model as a whole. Hence, in future experiments the type of gene expression data to be used should be chosen carefully to try to obtain better predictive accuracies.

In this research the target classification task has focused mainly on discriminating ageing-related DNA repair genes from non-ageing-related DNA repair genes – although other classes of genes interacting with DNA repair genes (in the sense of protein-protein interaction) were also included in the experiments with four classes. In future research the same data mining-based approach proposed in this thesis could be applied to a much larger set of genes, of different functional types.

For instance, one could consider all genes included in the GenAge database (or a subset of those genes with stronger evidence for their association with ageing), regardless of their functional type, as the "positive" class. This introduces the problem of defining a "negative" class, for training the classification algorithms. Perhaps the negative class

could be the set of genes that interact with genes in the GenAge database but have not been shown to be associated with ageing – i.e., are not in GenAge. This definition of the target classification problem would have the advantage that many more data instances (genes) would be available for training the classification algorithms, which, on one hand, would be an improvement of the dataset from a statistical perspective. On the other hand, this problem definition would have the potential disadvantage that the set of genes in the "positive" class would be quite diverse, and it is not clear if good predictive accuracies and many biologically interesting classification rules referring to genes of multiple functional types would be obtained.

Another variation in the definition of the target classification task would be to focus on discriminating ageing-related versus non-ageing related genes in the context of a set of genes of a single functional type other than DNA repair. For instance, one could use classification algorithms to classify genes related to oxidative stress into ageing-related and non-ageing-related ones.

In addition, the predictive patterns discoved by the classification algorithms could potentially be used to make predictions of the "ageing-related" class that could, in the future, be investigated in "web lab" experiments. That is, if some genes are currently considered as "non-ageing-related", but a very reliable pattern discovered by a classification algorithm predicts that those genes belong to the "ageing-related" class, this prediction could be used as a hint that it is worth doing a "wet lab" experiment to try to detect those genes' association with ageing. In order to better exploit the potential of this research direction, it would probably be useful to apply classification algorithms to a larger set of genes, rather than just DNA repair genes as in this research, since there are relatively few DNA repair genes and many of them have already been extensively studied in "wet lab" experiments. This research direction seems more promising for mining a larger set of genes including genes that have been less investigated by "wet lab" methods, where the computational predictions would seem to be more likely to lead to new discoveries of ageing-related genes.

The previously mentioned future research directions are mainly related to biological issues. From a computer science perspective, other types of classification algorithms could be applied. In this research we focused on decision tree induction algorithms as the main kind of classification algorithm producing predictive patterns that can be interpreted in the light of biological knowledge. The Naive Bayes algorithm used in this research uses all available predictor attributes for classification and assumes that the predictor attributes are independent from each other given the class of a data instance; and therefore it does not select a subset of most relevant attributes nor discover combinations of predictor attribute values with strong predictive power. However, there are more sophisticated types of Bayesian classification algorithms which automatically select a subset of most relevant attributes for classification and detect dependences between different predictor attributes, and such algorithms could be used in future research.

The caveat is that such algorithms in general require considerably larger datasets – in terms of the number of data instances (genes) – to obtain a good predictive accuracy, by comparison with the Naive Bayes algorithm used in this work. This is because the former typically have to compute a considerably larger number of probabilities from the data, and hence they are more prone to overfitting (Subsection 2.4.2.1) in small datasets.

# References


ALBERTS, B., JOHNSON, A., LEWIS, J., RAFF, M., ROBERTS, K. & WALTER, P. (2002) *Molecular Biology of the Cell. 4th Ed*, New York, NY, USA, Garland.

ARIS, V. M., CODY, M., CHENG, J., DERMODY, J. J., SOTEROPOULOS, P., RECCE, M. & TOLIAS, P. P. (2004) Noise filtering and nonparametric analysis of microarray data underscores discriminating markers of oral, prostate, lung, ovarian and breast cancer. *BMC Bioinformatics,* 5**,** 9 pages.

ARIYOSHI, K., SUZUKI, K., GOTO, M., WATANABE, M. & KODAMA, S. (2007) Increased chromosome instability and accumulation of DNA double-strand breaks in Werner Syndrome Cells. *Journal of Radiation Research,* 48**,** 219-231.

ARKING, R. (2006) *The Biology of Aging: Observations and Principles. 3rd Ed,* Oxford, UK, Oxford University Press.

BAREA, F. & BONATTO, D. (2009) Aging defined by a chronologic–replicative protein network in Saccharomyces cerevisiae: An interactome analysis. *Mechanisms of Ageing and Development,* 130**,** 444-460.

BELL, R., HUBBARD, A., CHETTIER, R., CHEN, D., MILLER, J. P., KAPAHI, P., TARNOPOLSKY, M., SAHASRABUNHDE, S., MELOV, S. & HUGHES, R. E. (2009) A Human Protein Interaction Network Shows Conservation of Aging Processes between Human and Invertebrate Species. *PLoS (Public Library of Science) Genetics,* 5**,** 12 pages.

BENEKE, S. & BURKLE, A. (2007) Poly(ADP-ribosyl)ation in mammalian ageing. *Nucleic Acids Research,* 35**,** 7456-7465.

BERGMAN, A., ATZMON, G., YE, K., MACCARTHY, T. & BARZILAI, N. (2007) Buffering mechanisms in aging: a systems approach toward uncovering the genetic component of aging. *PLoS (Public Library of Science) Computational Biology,* 3**,** 1648-1656.

BEST, B. P. (2009) Nuclear DNA damage as a direct cause of ageing. *Rejuvenation Research,* 12**,** 199-208.

BISHOP, K. N., HOLMES, R. K., SHEEHY, A. M., DAVIDSON, N. O., CHO, S. J. & MALIM, M. H. (2004) Cytidine deamination of retroviral DNA by diverse APOBEC proteins. *Current Biology,* 14**,** 1392-1396.

BOER, J. D., ANDRESSOO, J. O., WIT, J. D., HUIJMANS, J., BEEMS, R. B., STEEG, H. V., WEEDA, G., HORST, G. T. J. V. D., LEEUWEN, W. V., THEMMEN, A. P. N., MERADJI, M. & HOEIJMAKERS, J. H. J. (2002) Premature aging in mice deficient in DNA repair and transcription. *Science,* 296**,** 1276-1279.

BOER, J. D., DONKER, I., WIT, J. D., HOEIJMAKERS, J. H. J. & WEEDA, G. (1998) Disruption of the mouse xeroderma pigmentosum group D DNA repair/basal transcription gene results in preimplantation lethality. *Cancer Research,* 58**,** 89-94.

BONT, R. D. & LAREBEKE, N. V. (2004) Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis,* 19**,** 169-185.

BRADLEY, A. P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition,* 30**,** 1145-1159.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. & STONE, C. J. (1984) *Classification and Regression Trees,* Pacific Grove, CA, USA, Wadsworth.

BRIDGER, J. M. & KILL, I. R. (2004) Aging of Hutchinson-Gilford progeria syndrome fibroblasts is characterised by hyperproliferation and increased apoptosis. *Experimental Gerontology,* 39**,** 717-724.

BUDOVSKY, A., ABRAMOVICH, A., COHEN, R., CHALIFA-CASPI, V. & FRAIFELD, V. (2007) Longevity network: construction and implications. *Mechanisms of Ageing and Development,* 128**,** 117–124.

BUDOVSKY, A., TACUTU, R., YANAI, H., ABRAMOVICH, A., WOLFSON, M. & FRAIFELD, V. (2009) Common gene signature of cancer and longevity. *Mechanisms of Ageing and Development,* 130**,** 33-39.

BURMAA, S., CHENA, B. P. C. & CHEN, D. J. (2006) Role of non-homologous end joining (NHEJ) in maintaining genomic integrity. *DNA Repair,* 5**,** 1042-1048.

BUSUTTIL, R. A., GARCIA, A. M., REDDICK, R. L., DOLLE, M. E. T., CALDER, R. B., NELSON, J. F. & VIJG, J. (2007) Intra-organ variation in age-related mutation accumulation in the mouse. *PLoS (Public Library of Science) One,* 2**,** 9 pages.

C.VENS, STRUYF, J., SCHIETGAT, L., DZEROSKI, S. & BLOCKEEL, H. (2008) Decision trees for hierarchical multi-label classification. *Machine Learning,* 73**,** 185-214.

CABELOF, D. C., RAFFOUL, J. J., GE, Y., REMMEN, H. V. & L.H. MATHERLY, A. R. H. (2006) Age-related loss of the DNA repair response following exposure to oxidative stress. *Journal of Gerontology: Biological Sciences,* 61A**,** 427-434.

CABELOF, D. C., RAFFOUL, J. J., YANAMADALA, S., GANIR, C., GUO, Z. M. & HEYDARI, A. R. (2002) Attenuation of DNA polymerase beta-dependent base excision repair and increased DMS-induced mutagenicity in aged mice. *Mutation Research,* 500**,** 135-145.

CAMPISI, J. (2005) Aging, tumor suppression and cancer: high wire-act! *Mechanisms of Ageing and Development,* 126**,** 51-58.

CAMPISI, J. & FAGAGNA, F. D. D. (2007) Cellular senescence: when bad things happen to good cells. *Nature Reviews,* 8**,** 729-740.

CAO, L., KIM, S., XIAO, C., WANG, R. H., COUMOUL, X., WANG, X., LI, W. M., XU, X. L., SOTO, J. A. D., TAKAI, H., MAI, S., ELLEDGE, S. J., MOTOYAMA, N. & DENG, C. X. (2006) ATM-Chk2-p53 activation prevents tumorigenesis at an expense of organ homeostasis upon Brca1 deficiency. *The EMBO Journal,* 25**,** 2167-2177.

CARUANA, R. & NICULESCU-MIZIL, A. (2004) Data mining in metric space: an empirical analysis of supervised learning performance criteria. *Proceedings of the ACM 2004 International Conference on Knowledge Discovery and Data Mining (KDD-04).* New York, NY, USA, ACM Press.

CHAI, W., FORD, L. P., LENERTZ, L., WRIGHT, W. E. & SHAY, J. W. (2002) Human Ku70/80 associates physically with telomerase through interaction with hTERT. *The Journal of Biological Chemistry,* 277**,** 47242-47247.

CHAUTARD, E., THIERRY-MIEG, N. & RICARD-BLUM, S. (2010) Interacting networks as a tool to investigate the mechanisms of aging. *Biogerontology,* 11**,** 463-473.

CHEVANNE, M., CALIA, C., ZAMPIERI, M., CECCHINELLI, B., CALDINI, R., MONTI, D., BUCCI, L., FRANCESCHI, C. & CAIAFA, P. (2007) Oxidative DNA damage repair and parp 1 and parp 2 expression in Epstein-Barr virus-immortalized B lymphocyte cells from young subjects, old subjects and centenarians. *Rejuvenation Research,* 10**,** 191-203.

CHIMPANZEE-CONSORTIUM (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature,* 437**,** 69-87.

CLARE, A. & KING, R. D. (2001) Knowledge discovery in multi-label phenotype data. *Proc. 2001 European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2001), Lecture Notes in Artificial Intelligence 2168.* Berlin, Germany, Springer.

CLARE, A. & KING, R. D. (2003) Predicting gene function in Saccharomyces cerevisiae. *Bioinformatics,* 19**,** ii42-ii49.

COOPER, D. N. & YOUSSOUFIAN, H. (1988) The CpG dinucleotide and human genetic disease. *Human Genetics,* 78**,** 151-155.

CRISTIANINI, N. & SHAWE-TAYLOR, J. (2000) *An Introduction to Support Vector Machines and Other Kernel-Based Methods,* Cambridge, UK, Cambridge University Press.

DAVIES, M. N., SECKER, A., FREITAS, A. A., MENDAO, M., TIMMIS, J. & FLOWER., D. R. (2007) On the hierarchical classification of G Protein-Coupled Receptors. *Bioinformatics,* 23**,** 3113-3118.

DAVIS, T. & KIPLING, D. (2006) Werner Syndrome as an example of inflamm-aging: possible therapeutic opportunities for a progeroid syndrome? *Rejuvenation Research,* 9**,** 402-407.

DEBACQ-CHAINIAUX, F., BORLON, C., PASCAL, T., ROYER, V., ELIAERS, F., NINANE, N., CARRARD, G., FRIGUET, B., LONGUEVILLE, F. D. & BOFFE, S. (2005) Repeated exposure of human skin fibroblasts to UVB at subcytotoxic level triggers premature senescence through the TGF-beta1 signaling pathway. *Journal of Cell Science,* 118**,** 743-758.

DEGROOT, M. & SCHERVISH, M. J. (2002) *Probability and Statistics (3rd Ed.),* Reading, MA, USA, Addison Wesley.

ENGELS, W. R., JOHNSON-SCHLITZ, D., FLORES, C., WHITE, L. & PRESTON, C. R. (2007) A third link connecting aging with double strand break repair. *Cell Cycle,* 6**,** 131-135.

FAYYAD, U. M., PIATETSKY-SHAPIRO, G. & SMYTH, P. (1996) From data mining to knowledge discovery: an overview. IN FAYYAD, U. M. & ET AL. (Eds.) *Advances in Knowledge Discovery and Data Mining.* Palo Alto, CA, USA, AAAI/MIT Press.

FERRARINI, L., BERTELLI, L., FEALA, J., MCCULLOCH, A. D. & PATERNOSTRO, G. (2005) A more efficient search strategy for aging genes based on connectivity. *Bioinformatics,* 21**,** 338-348.

FISHEL, M. L., VASKO, M. R. & KELLEY, M. R. (2007) DNA repair in neurons: so if they don't divide what's to repair? *Mutation Research,* 614**,** 24-36.

FORTNEY, K., KOTLYAR, M. & JURISICA, I. (2010) Inferring the functions of longevity genes with modular subnetwork biomarkers of Caenorhabditis elegans aging. *Genome Biology,* 11**,** 14 pages.

FREITAS, A. A. (2002) *Data Mining and Knowledge Discovery with Evolutionary Algorithms,* Berlin, Germany, Springer.

FREITAS, A. A., VASIEVA, O. & MAGALHAES, J. P. D. (2011) A data mining approach for classifying DNA repair genes into ageing-related or non-ageing-related. *BMC Genomics,*12.

FREITAS, A. A., WIESER, D. C. & APWEILER, R. (2010) On the importance of comprehensible classification models for protein function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* 7**,** 172-182.

FRIEDBERG, E. C., WALKER, G. C., SIEDE, W., WOOD, R. D., SCHULTZ, R. A. & ELLENBERGER, T. (2006) *DNA Repair and Mutagenesis. 2nd Ed,* ASM Press.

FRIEDBERG, I. (2006) Automated protein function prediction – the genomic challenge. *Briefings in Bioinformatics,* 7**,** 225-242.

FURNKRANZ, J. (1999) Separate-and-conquer rule learning. *Artificial Intelligence Review,* 13**,** 3-54.

GERLING, I. C., SINGH, S., LENCHIK, N. I., MARSHALL, D. R. & WU, J. (2005) New data analysis and mining approaches identify unique proteome and transcriptome markers of susceptibility to autoimmune diabetes. *Molecular and Cellular Proteomics,* 5**,** 293-305.

GO-CONSORTIUM (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics,* 25**,** 25-29.

GO-CONSORTIUM (2001) Creating the gene ontology resource: design and implementation. *Genome Research,* 11**,** 1435-1433.

GO-CONSORTIUM (2004) The Gene Ontology (DB) and informatics resources. *Nucleic Acids Research,* 32**,** D258-D261.

GOLDEN, T. R., HUBBARD, A., DANDO, C., HERREN, M. A. & MELOV, S. (2008) Age-related behaviors have distinct transcriptional profiles in C.elegans. *Aging Cell,* 7**,** 850–865.

GORBUNOVA, V., SELUANOV, A., MAO, Z. & HINE, C. (2007) Changes in DNA repair during aging. *Nucleic Acids Research,* 35**,** 7466-7474.

GRAZIEWICZ, M. A., LONGLEY, M. J. & COPELAND, W. C. (2006) DNA polymerase in mitochondrial DNA replication and repair. *Chemistry Review,* 106**,** 383-405.

GREY, A. D. & RAE, M. (2007) *Ending Aging: the rejuvenation breakthroughs that could reverse human aging in our lifetime,* New York, NY, USA, St. Martin's Press.

GU, Y., SEIDL, K. J., RATHBUN, G. A., ZHU, C., MANIS, J. P., STOEP, N. V. D., DAVIDSON, L., CHENG, H. L., SEKIGUCHI, J. M., FRANK, K., STANHOPE-BAKER, P., SCHLISSEL, M. S., ROTH, D. B. & ALT, F. W. (1997) Growth retardation and leaky SCID phenotype of Ku70-deficient mice. *Immunity,* 7**,** 653-665.

HASTY, P., CAMPISI, J., HOEIJMAKERS, J., STEEG, H. V. & VIJG, J. (2003) Aging and genome maintenance: lessons from the mouse? *Science,* 299**,** 1355-1359.
HASTY, P. & VIJG, J. (2004a) Accelerating aging by mouse reverse genetics: a rational approach to understanding longevity. *Aging Cell,* 3**,** 55-65.
HASTY, P. & VIJG, J. (2004b) Rebuttal to Miller: 'Accelerated Aging': a primrose path to insight? *Aging Cell,* 3**,** 67-69.
HAYFLICK, L. (2000) The future of ageing. *Nature,* 408**,** 267-269.
HAZANE, F., SAUVAIGO, S., DOUKI, T., FAVIER, A. & BEANI, J. C. (2006) Age-dependent DNA repair and cell cycle distribution of human skin fibroblasts in response to UVA irradiation. *Journal of Photochemistry and Photobiology B: Biology,* 82**,** 214-223.
HE, J., HU, H. J., HARRISON, R., TAI, P. C. & PAN, Y. (2006) Transmembrane segments prediction and understanding using support vector machine and decision tree. *Expert Systems with Applications,* 30**,** 64-72.
HENNEKAM, R. C. M. (2006) Hutchinson-Gilford progeria syndrome: review of the phenotype. *American Journal of Medical Genetics Part A,* 140A**,** 2603-2624.
HOEIJMAKERS, J. H. (2001) Genome maintenance mechanisms for preventing cancer. *Nature,* 411**,** 366-374.
HOLCOMB, V. B., RODIER, F., CHOI, Y. J., BUSUTTIL, R. A., VOGEL, H., VIJG, J., CAMPISI, J. & HASTY, P. (2008) Ku80 deletion suppresses spontaneous tumors and induces a p53-mediated DNA damage response. *Cancer Research,* 68**,** 9497-9502.
HOSAKA, N., NOSET, M., KYOGOKUT, M., NAGATA, N., MIYASHIMA, S., GOOD, R. A. & IKEHARA, S. (1996) Thymus transplantation, a critical factor for correction of autoimmune disease in aging MRL/+ mice. *Proceedings of the National Academy of Sciences (PNAS), USA, Immunology,* 93**,** 8558-8562.
HRUZ, T., LAULE, O., SZABO, G., WESSENDORP, F., BLEULER, S., OERTLE, L., WIDMAYER, P., GRUISSEM, W. & ZIMMERMANN, P. (2008) Genevestigator V3: a reference expression database for the meta-analysis of transcriptomes. *Advances in Bioinformatics***,** 5 pages.
HSU, C. W., JUAN, H. F. & HUANG, H. C. (2008) Characterization of microRNA-regulated protein-protein interaction network. *Proteomics,* 8**,** 1975-1979.
HUANG, L. T., GROMIHA, M. M. & HO, S. Y. (2007) iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics,* 23**,** 1292-1293.
JACKSON, A. L. & LOEB, L. A. (2001) The contribution of endogenous sources of DNA damage to the multiple mutations in cancer. *Mutation Research,* 477**,** 7-21.
JAMES, S. E., FARAGHER, R. G. A., BURKEC, J. F., SHALLD, S. & MAYNEA, L. V. (2000) Werner's syndrome T lymphocytes display a normal in vitro life-span. *Mechanisms of Ageing and Development,* 121**,** 139-149.
JU, Y. J., LEE, K. H., PARK, J. E., YI, Y. S., YUN, M. Y., HAM, Y. H., KIM, T. J., CHOI, H. M., HAN, G. J., LEE, J. H., LEE, J., HAN, J. S., LEE, K. M. & PARK, G. H. (2006) Decreased expression of DNA repair proteins Ku70 and Mre11 is associated with aging and may contribute to the cellular senescence. *Experimental and Molecular Medicine,* 38**,** 686-693.

KAEBERLEIN, M., JEGALIAN, B. & MCVEY, M. (2002) AGEID: a database of aging genes and interventions. *Mechanisms of Ageing and Development,* 123**,** 1115-1119.

KANEKO, T., TAHARA, S., TANNO, M. & TAGUCHI, T. (2002) Age-related changes in the induction of DNA polymerases in rat liver by gamma-ray irradiation. *Mechanisms of Ageing and Development,* 123**,** 1521-1528.

KARLSSON, B., GUSTAFSSON, J., HEDOV, G., IVARSSON, S. A. & ANNERZN, G. (1998) Thyroid disfunction in Down's syndrome: relation to age and thyroid autoimmunity. *Archives of Diseases in Childhood,* 79**,** 242-245.

KARWATH, A. & KING, R. D. (2002) Homology induction: the use of machine learning to improve sequence similarity searches. *BMC Bioinformatics,* 3**,** 13 pages.

KELL, D. B. & KING, R. D. (2000) On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning. *Trends in Biotechnology,* 18**,** 93-98.

KENYON, C. (2010) The genetics of ageing. *Nature,* 464**,** 504-512.

KIPLING, D., DAVIS, T., OSTLER, E. L. & FARAGHER, R. G. A. (2004) What can progeroid syndromes tell us about human aging? *Science,* 305**,** 1426-1431.

KIRKWOOD, T. B. L. (1996) Human senescence. *BioEssays,* 18**,** 1009-1016.

KIRKWOOD, T. B. L. (2005) Understanding the odd science of aging. *Cell,* 120**,** 437-447.

KIRKWOOD, T. B. L. & AUSTAD, S. N. (2000) Why do we age? *Nature,* 408**,** 233-238.

KO, L. J. & PRIVES, C. (1996) p53: puzzle and paradigm. *Genes & Development,* 10**,** 1054-1072.

KOZMIN, S., SLEZAK, G., REYNAUD-ANGELIN, A., ELIE, C., RYCKE, Y. D., BOITEUX, S. & SAGE, E. (2005) Uva radiation is highly mutagenic in cells that are unable to repair, dihydro-oxoguanine in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences (PNAS), USA,* 102**,** 13538-13543.

KRISHNA, T. H., MAHIPAL, S., SUDHAKAR, A., SUGIMOTO, H., KALLURI, R. & RAO, K. S. (2005) Reduced DNA gap repair in aging neuronal extracts and its restoration by DNA polymerase ß and DNA-ligase. *Journal of Neurochemistry,* 92**,** 818-823.

LANS, H. & HOEIJMAKERS, J. H. (2006) Ageing nucleus gets out of shape. *Nature,* 440**,** 32-34.

LEE, J. W., HARRIGAN, J., OPRESKO, P. L. & BOHR, V. A. (2005) Pathways and functions of the Werner syndrome protein. *Mechanisms of Ageing and Development,* 126**,** 79-86.

LEROI, A. M., BARTKE, A., BENEDICTIS, G. D., FRANCESCHI, C., GARTNER, A., GONOS, E., FEDER, M. E., KIVISILD, T., LEE, S., KARTAL-OZER, N., SCHUMACHER, M., SIKORA, E., SLAGBOOM, E., TATAR, M., YASHIN, A. I., VIJG, J. & ZWAAN, B. (2005) What evidence is there for the existence of individual genes with antagonistic pleiotropic effects? *Mechanisms of Ageing and Development,* 126**,** 421-429.

LEWIS, S. E. (2004) Gene Ontology: looking backwards and forwards. *Genome Biology,* 6**,** 4 pages.

LI, B., NAVARRO, S., KASAHARA, N. & COMAI, L. (2004) Identification and biochemical characterization of a Werner's Syndrome protein complex with Ku70/80 and Poly(ADP-ribose) Polymerase-1. *The Journal of Biological Chemistry,* 279**,** 13659-13667.

LI, G. C., OUYANG, H., LI, X., NAGASAWA, H., LITTLE, J. B., CHEN, D. J., LING, C. C., FUKS, Z. & CORDON-CARDO, C. (1998) Ku70: a candidate tumor suppressor gene for murine T cell lymphoma. *Molecular Cell,* 2**,** 1-8.

LI, G. M. (2008) Mechanisms and functions of DNA mismatch repair. *Cell Research,* 18**,** 85-98.

LI, H., CHOI, Y. J., HANES, M. A., MARPLE, T., VOGEL, H. & HASTY, P. (2009) Deleting Ku70 is milder than deleting Ku80 in p53-mutant mice and cells. *Oncogene,* 28**,** 1875-1878.

LI, H., VOGEL, H., HOLCOMB, V. B., GU, Y. & HASTY, P. (2007) Deletion of Ku70, Ku80, or both causes early aging without substantially increased cancer. *Molecular and Cellular Biology,* 27**,** 8205-8214.

LI, Y. H., DONG, M. Q. & GUO, Z. (2010) Systematic analysis and prediction of longevity genes in Caenorhabditis elegans. *Mechanisms of Ageing and Development,* 131**,** 700-709.

LIANG, H. & LI, W. H. (2007) MicroRNA regulation of human protein-protein interaction network. *RNA,* 13**,** 1402-1408.

LINDAHL, T. (1993) Instability and decay of the primary structure of DNA. *Nature,* 362**,** 709-715.

LINDAHL, T. & WOOD, R. D. (1999) Quality control by DNA repair. *Science,* 286**,** 1897-1905.

LJUNGMAN, M. & LANE, D. P. (2004) Transcription - guarding the genome by sensing DNA damage. *Nature Reviews - Cancer,* 4**,** 727-737.

LOMBARD, D. B., CHUA, K. F., MOSTOSLAVSKY, R., FRANCO, S., GOSTISSA, M. & ALT, F. W. (2005) DNA repair, genome stability and aging. *Cell,* 120**,** 497-512.

LU, T., PAN, Y., KAO, S. Y., LI, C., KOHANE, I., CHAN, J. & YANKNER, B. A. (2004) Gene regulation and DNA damage in the ageing bring. *Nature,* 429**,** 883-891.

MAGALHAES, J. P. D. (2009) Ageing research in the post-genome era: new technologies for an old problem. IN FOYER, C., FARAGHER, R. & THORNALLEY, P. (Eds.) *Redox Metabolism and Longevity Relationships in Animals and Plants.* London, UK, Taylor and Francis.

MAGALHAES, J. P. D. (2011) The Biology of Ageing: a primer. IN STUART-HAMILTON, I. (Ed.) *An Introduction to Gerontology.* Cambridge, UK, Cambridge University Press, In Press.

MAGALHAES, J. P. D., BUDOVSKY, A., LEHMANN, G., COSTA, J., LI, Y., FRAIFELD, V. & CHURCH, G. M. (2009) The Human Ageing Genomic Resources: online databases and tools for biogerontologists. *Aging Cell,* 8**,** 65-72.

MAGALHAES, J. P. D. & CHURCH, G. M. (2006) Cells discover fire: employing reactive oxygen species in development and consequences for aging. *Experimental Gerontology,* 41**,** 1-10.

MAGALHAES, J. P. D. & CHURCH, G. M. (2007) Analyses of human-chimpanzee orthologous gene pairs to explore evolutionary hypotheses of aging. *Mechanisms of Ageing and Development,* 128**,** 355-364.
MAGALHAES, J. P. D., COSTA, J. & TOUSSAINT, O. (2005) HAGR: the Human Ageing Genomic Resources. *Nucleic Acids Research,* 33**,** D537-D543.
MAGALHAES, J. P. D. & FARAGHER, R. G. A. (2008) Cell divisions and mammalian aging: integrative biology insights from genes that regulate longevity. *BioEssays,* 30**,** 567-578.
MAGALHAES, J. P. D. & TOUSSAINT, O. (2004) GenAge: a genomic and proteomic network map of human ageing. *FEBS Letters,* 571**,** 243-247.
MAIER, B., GLUBA, W., BERNIER, B., TURNER, T., MOHAMMAD, K., GUISE, T., SUTHERLAND, A., THORNER, M. & SCRABLE, H. (2004) Modulation of mammalian life span by the short isoform of p53. *Genes & Development,* 18**,** 306-319.
MANAGBANAG, J. R., WITTEN, T. M., BONCHEV, D., FOX, L. A., TSUCHIYA, M., KENNEDY, B. K. & KAEBERLEIN, M. (2008) Shortest-path network analysis is a useful approach towards identifying genetic determinants of longevity. *PLoS (Public Library of Science) One,* 3**,** 9 pages.
MARTIN, G. M. & OSHIMA, H. (2000) Lessons from human progeroid syndromes. *Nature,* 408**,** 263-266.
MASLOV, A. Y. & VIJG, J. (2009) Genome instability, cancer and aging. *Biochimica et Biophysica Acta,* 1790**,** 963-969.
MASLOV, S. (2008) Topological and dynamical properties of protein interaction networks. IN PANCHENKO, A. & PRZYTYCKA, T. (Eds.) *Protein-protein interactions and networks.* Berlin, Germany, Springer.
MAYBURD, A. L., MARTINEZ, A., SACKETT, D., LIU, H., SHIH, J., TAULER, J., AVIS, I. & MULSHINE, J. L. (2006) Ingenuity network-assisted transcription profiling: identification of a new pharmacologic mechanism for MK886. *Clinical Cancer Research,* 12**,** 1820-1827.
MEDAWAR, P. B. (1952) An unsolved problem of biology. London, UK, H.K. Lewis.
MENDRYSA, S. M., O'LEARY, K. A., MCELWEE, M. K., MICHALOWSKI, J., EISENMAN, R. N., POWELL, D. A. & PERRY, M. E. (2006) Tumor suppression and normal aging in mice with constitutively high p53 activity. *Genes & Development,* 20**,** 16-21.
MERKLE, T. J., O'BRIEN, K., BROOKS, P. J., TARONE, R. E. & ROBBINS, J. H. (2004) DNA repair in human fibroblasts, as reflected by host-cell reactivation of a transfected UV-irradiated luciferase gene, is not related to donor age. *Mutation Research,* 554**,** 9-17.
MILLER, R. A. (2004a) Accelerated Aging: a primrose path to insight? *Aging Cell,* 3**,** 47-51.
MILLER, R. A. (2004b) Rebuttal to Hasty and Vijg: Accelerating aging by mouse reverse genetics: a rational approach to understanding longevity. *Aging Cell,* 3**,** 54-55.
MOMBAERTS, P., IACOMINI, J., JOHNSON, R. S., HERRUPA, K., TONEGAWA, S. & PAPAIOANNOU, V. E. (1992) RAG-1-deficient mice have no mature B and T lymphocytes. *Cell,* 68**,** 869-877.

MORIWAKI, S. & TAKAHASHI, Y. (2008) Photoaging and DNA repair. *Journal of Dermatological Science,* 50**,** 169-176.

NAGL, S. B. (2003) Molecular evolution. IN C.A. ORENGO, D. T. J. J. M. T. (Ed.) *Bioinformatics: genes, proteins, computers.* New York, NY, USA, BIOS Scientific Publisher.

NAKAMURA, J., DAVID, K. L. & SWENBERG, J. A. (2000) 5'-nicked apurinic/apyrimidinic sites are resistant to ?-elimination by ?-polymerase and are persistent in human cultured cells after oxidative stress. *The Journal of Biological Chemistry,* 275**,** 5323-5328.

NAKAMURA, J. & SWENBERG, J. A. (1999) Endogenous apurinic/apyrimidinic sites in genomic DNA of mammalian tissues. *Cancer Research,* 59**,** 2522-2526.

NIEDERNHOFER, L. (2008) Tissue-specific accelerate aging in nucleotide excision repair deficiency. *Mechanisms of Ageing and Development,* 129**,** 408-415.

NIEDERNHOFER, L. J., GARINIS, G. A., RAAMS, A., LALAI, A. S., ROBINSON, A. R., APPELDOORN, E., ODIJK, H., OOSTENDORP, R., AHMAD, A., LEEUWEN, W. V., THEIL, A. F., VERMEULEN, W., HORST, G. T. J. V. D., MEINECKE, P., KLEIJER, W. J., VIJG, J., JASPERS, N. G. J. & HOEIJMAKERS, J. H. (2006) A new progeroid syndrome reveals that genotoxic stress suppresses the somatotroph axis. *Nature,* 444**,** 1038-1043.

OREN, M. (2003) Decision making by p53: life, death and cancer. *Cell Death and Differentiation,* 2003**,** 431-432.

OTERO, F. E. B., FREITAS, A. A. & JOHNSON, C. G. (2010) A hierarchical multi-label classification ant colony algorithm for protein function prediction. *Memetic Computing,* 2**,** 165-181.

PAN, F., CHIU, C. H., PULAPURA, S., MEHAN, M. R., NUNEZ-IGLESIAS, J., ZHANG, K., IAMATH, K., WATERMAN, M. S., FINCH, C. E. & ZHOU, X. J. (2007) Gene Aging Nexus: a web database and data mining platform for microarray data on aging. *Nucleic Acids Research,* 35**,** D756-759.

PANCHENKO, A. & PRZYTYCKA, T. (Eds.) (2008) *Protein-protein interactions and networks: identification, computer analysis and prediction,* Berlin, Germany, Springer.

PAPPA, G. L. & FREITAS, A. A. (2009) Automatically evolving rule induction algorithms tailored to the prediction of postsynaptic activity in proteins. *Intelligent Data Analysis,* 13**,** 243-259.

PARK, J. Y., CHO, M. O., LEONARD, S., CALDER, B., MIAN, I. S., KIM, W. H., WINJHOVEN, S., STEEG, H. V., MITCHELL, J., HORST, G. T. J. V. D., HOEIJMAKERS, J., COHEN, P., VIJG, J. & SUH, Y. (2008) Homeostatic imbalance between apoptosis and cell renewal in the liver of premature aging XpdTTD mice. *PLoS (Public Library of Science) One,* 3**,** 1-10.

PASSOS, J. F., SEMILLION, C., HALLINAN, J., WIPAT, A. & ZGLINICKI, T. V. (2009) Cellular senescence: unravelling complexity. *Age,* 31**,** 353-363.

PELICCI, P. G. (2004) Do tumor-suppressive mechanisms contribute to organism aging by inducing stem cell senescence? *The Journal of Clinical Investigation,* 113**,** 4-7.

PESCE, K. & ROTHE, M. J. (1996) The premature ageing syndromes. *Clinics in Dermatology,* 14**,** 161-170.

PLUIJM, I. V. D., GARINIS, G. A., BRANDT, R. M. C., GORGELS, T. G. M. F., WIJNHOVEN, S. W., DIDERICH, K. E. M., WIT, J. D., MITCHELL, J. R., OOSTROM, C. V., BEEMS, R., NIEDERNHOFER, L. J., VELASCO, S., FRIEDBERG, E. C., TANAKA, K., HARRY VAN, S., HOEIJMAKERS, J. H. J. & HORST, G. T. J. V. D. (2007) Impaired Genome Maintenance Suppresses the Growth Hormone–Insulin-Like Growth Factor 1 Axis in Mice with Cockayne Syndrome. *PLoS (Public Library of Science) Biology,* 5**,** 23-38.

PRALL, W. C., CZIBERE, A., JAGER, M., SPENTZOS, D., LIBERMANN, T. A., GATTERMANN, N., HAAS, R. & AIVADO, M. (2007) Age-related transcription levels of KU70, MGST1 and BIK in CD34+ hematopoietic stem and progenitor cells. *Mechanisms of Ageing and Development,* 128**,** 503-510.

PRASAD, T. S. K., GOEL, R., KANDASAMY, K. & ET AL. (2009) Human Protein Reference Database - 2009 update. *Nucleic Acids Research,* 37**,** D767–D772.

PRASHER, V. P. (1999) Down Syndrome and thyroid disorders: a review. *Down Syndrome Research and Practice,* 6**,** 25-42.

PRELOG, M. (2006) Aging of the immune system: a risk factor for autoimmunity? *Autoimmune Reviews,* 5**,** 136-139.

PROMISLOW, D. E. L. (2004) Protein networks, pleiotropy and the evolution of senescence. *Proceedings of the Royal Society of London B,* 271**,** 1225-1234.

QUAN, T., HE, T., KANG, S., VOORHEES, J. V. & FISHER, G. J. (2004) Solar ultraviolet irradiation reduces collagen in photoaged human skin by blocking transforming growth factor-ß type II receptor/smad signalling. *American Journal of Pathology,* 165**,** 741-751.

QUINLAN, J. R. (1993) *C4.5: Program for Machine Learning,* Palo Alto, CA, USA, Morgan Kaufmann.

RABINOWE, S. L., RUBIN, I. L., GEORGE, K. L., ADRI, M. N. & EISENBARTH, G. S. (1989) Trisomy 21 (Down's syndrome): autoimmunity, aging and monoclonal antibody-defined T-cell abnormalities. *Journal of Autoimmunity,* 2**,** 25-30.

RANDO, T. A. (2006) Stem cells, ageing and the quest for immortality. *Nature,* 441**,** 1080-1086.

RAO, K. S. (2007) DNA repair in aging rat neurons. *Neuroscience,* 145**,** 1330-1340.

RASSOOL, F. V. (2003) DNA double strand breaks (DSB) and non-homologous end joining (NHEJ) pathways in human leukemia. *Cancer Letters,* 193**,** 1-9.

ROKACH, L. & MAIMON, O. (2005) Top-down induction of decision tree classifiers - a survey. *IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews,* 35**,** 476-487.

ROONEY, S., CHAUDHURI, J. & ALT, F. W. (2004) The role of the non-homologous end-joining pathway in lymphocyte development. *Immunology Review,* 200**,** 115-131.

ROSSI, D. J., BRYDER, D., SEITA, J., NUSSENZWEIG, A., HOEIJMAKERS, J. & WEISSMAN, I. L. (2007) Deficiencies in DNA damage repair limit the function of haematopoietic stem cells with age. *Nature,* 447**,** 725-730.

ROST, B., LIU, J., NAIR, R., WRZESZCZYNSKI, K. O. & OFRAN, Y. (2003) Automatic prediction of protein function. *Cellular and Molecular Life Sciences,* 60**,** 2637-2650.

ROTMAN, G. & SHILOH, Y. (1997) Ataxia-telangiectasia: is ATM a sensor of oxidative damage and stress? *BioEssays,* 19**,** 911-917.

SCHIETGAT, L., VENS, C., STRUYF, J., BLOCKEEL, H., KOCEV, D. & DZEROSKI, S. (2010) Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC Bioinformatics,* 11**,** 14 pages.

SELUANOV, A., DANEK, J., HAUSE, N., GORBUNOVA, V., CHANGES IN THE, L. & DISTRIBUTION OF KU PROTEINS DURING CELLULAR, S. (2007) Changes in the level and distribution of Ku proteins during cellular senescence. *DNA Repair,* 6**,** 1740-1748.

SHARMA, S. (2007) Age-related nonhomologous end joining activity in rat neurons. *Brain Research Bulletin,* 73**,** 48-54.

SWINBURNE, R. (Ed.) (2002) *Bayes's Theorem,* Oxford, UK, Oxford University Press.

SYED, U. & YONA, G. (2009) Enzyme function prediction with interpretable models. IN SAMUDRALA, R., MCDERMOTT, J. & BUMGARNER, R. (Eds.) *Computational Systems Biology.* Berlin, Germany, Humana Press.

SYMPHORIEN, S. & WOODRUFF, R. C. (2003) Effect of DNA repair on aging of transgenic Drosophila melanogaster: I.mei-41 locus. *Journal of Gerontology. A. Biological Science Medical Science,* 58**,** B782-B787.

TACCIOLI, G. E., GOTTLIEB, T. M., BLUNT, T., PRIESTLEY, A., DEMENGEOT, J., MIZUTA, R., LEHMANN, A. R., ALT, F. W., JACKSON, S. P. & JEGGO, P. A. (1994) Ku80: product of the XRCC5 gene and its role in DNA repair and V(D)J recombination. *Science,* 265**,** 1442-1445.

TACUTU, R., BUDOVSKY, A. & FRAIFELD, V. (2010) The NetAge database: a compendium of networks for longevity, age-related diseases and associated processes. *Biogerontology,* 11**,** 513-522.

TAKAHASHI, Y., MORIWAKI, S.-I., SUGIYAMA, Y., ENDO, Y., YAMAZAKI, K., MORI, T., TAKIGAWA, M. & INOUE, S. (2005) Decreased Gene Expression Responsible for Post-Ultraviolet DNA Repair Synthesis in Aging: A Possible Mechanism of Age-Related Reduction in DNA Repair Capacity. *The Journal of Investigative Dermatology,* 124**,** 435-442.

TAN, P. N., STEINBACH, M. & KUMAR, V. (2005) *Introduction to Data Mining,* Boston, MA, USA, Pearson.

TASCHUK, M. L., SIMILLION, C., HALLINAN, J. & WIPAT, A. (2010) CID: CISBAN Interactome Database. *Poster paper at the 6th Int. Symposium on Integrative Bioinformatics (Cambridge, UK).*

TYNER, S. D., VENKATACHALAM, S., CHOI, J., JONES, S., GHEBRANIOUS, N., IGELMANN, H., LU, H., SORON, G., COOPER, B., BRAYTON, C., PARK, S. H., THOMPSON, T., KARSENTY, G., BRADLEY, A. & DONEHOWE, L. A. (2002) p53 mutant mice that display early ageing-associated phenotypes. *Nature,* 415**,** 45-53.

UNIPROT-CONSORTIUM (2007) The Universal Protein Resource (Uniprot). *Nucleic Acids Research,* 35**,** D193–D197.

UNIPROT-CONSORTIUM (2010) The Universal Protein Resource (Uniprot) in 2010. *Nucleic Acids Research,* 38**,** D142–D148.

VAZIRI, H., SCHACHTER, F., UCHIDA, I., WEI, L., ZHU, X., EFFROS, R., COHEN, D. & HARLEY, C. B. (1993) Loss of telomeric DNA during aging of normal and trisomy 21 human lymphocytes. *American Journal of Human Genetics,* 52**,** 661-667.

VEN, M. V. D., ANDRESSOO, J. O., HOLCOMB, V. B., LINDERN, M. V., JONG, W. M. C., ZEEUW, C. I. D., SUH, Y., HASTY, P., HOEIJMAKERS, J. H. J., HORST, G. T. J. V. D., MITCHELL, J. R. & GENETICS, P. L. (2006) Adaptive response in segmental progeria resembles long-lived dwarfism and calorie restriction in mice. *PLoS (Public Library of Science) Genetics,* 2**,** 2013-2025.

VIJG, J. (2008) The role of DNA damage and repair in aging: new approaches to an old problem. *Mechanisms of Ageing and Development,* 129**,** 498-502.

VIJG, J. & CAMPISI, J. (2008) Puzzles, promises and a cure for ageing. *Nature,* 454**,** 1065-1071.

VOGEL, H., LIM, D. S., KARSENTY, G., FINEGOLD, M. & HASTY, P. (1999) Deletion of Ku86 causes early onset of senescence in mice. *Proceedings of the National Academy of Sciences (PNAS), USA,* 96**,** 10770-10775.

VYJAYANTI, V. N. & RAO, K. S. (2006) DNA double strand break repair in brain: reduced NHEJ activity in aging rat neurons. *Neuroscience Letters,* 393**,** 18-22.

WALKER, J. R., CORPINA, R. A. & GOLDBERG, J. (2001) Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. *Nature,* 412**,** 607-613.

WANG, J., ZHANG, S., WANG, Y., CHEN, L. & ZHANG, X. S. (2009) Disease-aging network reveals significant roles of aging genes in connecting genetic diseases. *PLoS (Public Library of Science) Computational Biology,* 5**,** 12 pages.

WHITE, A. P. & LIU, W. Z. (1994) Bias in information-based measures in decision tree induction. *Machine Learning,* 15**,** 321-329.

WIESER, D., PAPATHEODOROU, I., ZIEHM, M. & THORNTON, J. M. (2011) Computational Biology for Ageing. *Philosophical Transactions of the Royal Society B - Biological Sciences,* 366**,** 51-63.

WITTEN, I. H. & FRANK, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques (2nd Ed.),* Palo Alto, CA, USA, Morgan Kaufmann.

WITTEN, T. M. & BONCHEV, D. (2007) Predicting aging/longevity-related genes in the nematode Caenorhabditis elegans. *Chemistry & Biodiversity,* 4**,** 2639-2655.

WOLFSON, M., TACUTU, R., BUDOVSKY, A., AIZENBERG, N. & FRAIFELD, V. E. (2008) MicroRNAs: relevance to ageing and age-related diseases. *Open Longevity Science,* 2**,** 66-75.

WONG, K. K., MASER, R. S., BACHOO, R. M., MENON, J., CARRASCO, D. R., GU, Y., ALT, F. W. & PINHO, R. A. D. (2003) Telomere dysfunction and Atm deficiency compromises organ homeostasis and accelerates ageing. *Nature,* 421**,** 643-647.

WOOD, R. D., MITCHELL, M. & LINDAHL, T. (2005) Human DNA repair genes. *Mutation Research,* 577**,** 275-283.

WOOD, R. D., MITCHELL, M., SGOUROS, J. & LINDAHL, T. (2001) Human DNA repair genes. *Science,* 291**,** 1284-1289.

YAAR, M. & GILCHREST, B. A. (2007) Photoageing: mechanism, prevention and therapy. *British Journal of Dermatology,* 157**,** 874-887.

YAMADA, M., UDONO, M. U., HORI, M., HIROSE, R., SATO, S., MORI, T. & NIKAIDO, O. (2006) Aged human skin removes UVB-induced pyrimidine dimers from the epidermis more slowly than younger adult skin in vivo. *Archives of Dermatology Research,* 297**,** 294-302.

ZGLINICKI, T. V., SARETZKI, G., LADHOFF, J., FAGAGNA, F. D. A. D. & JACKSON, S. P. (2005) Human cell senescence as a DNA damage response. *Mechanisms of Ageing and Development,* 126**,** 111–117.

ZHAO, X. M., CHEN, L. & AIHARA, K. (2008) Protein function prediction with high-throughput data. *Amino Acids,* 35**,** 517–530.

ZHU, C., BOGUE, M. A., LIM, D. S., HASTY, P. & ROTH, D. B. (1996) Ku86-deficient mice exhibit severe combined immunodeficiency and defective processing of V(D)J recombination intermediates. *Cell,* 86**,** 379-389.