A Thesis submitted for the degree of Master of Science

A Comparative Analysis of Thermodynamic Stability of Electron transport Chain Complex I Proteins and the Maximum Life Span of Host Organisms

University of Tampere

Aruj Joshi

20/05/2014

BioMediTech
University of Tampere

Acknowledgement

The thesis work was carried on in the Bio-informatics research group, BioMediTech, University of Tampere, Finland.

I would like to extend my sincerest gratitude to Professor Csaba Ortutay for his help and constant guidance and for his quick and patient review of this thesis. I am grateful to researcher, Jouni Valiaho for his expert technical guidance to me whenever I needed it. I would like to thank Alberto Sanz, head of Mitochondrial Gerontology and Age-related Diseases group who proposed this thesis idea to Csaba who then gave me the opportunity to work on it. I would also like to thank head of Bio-informatics group, Professor Matti Nykter for his help with the thesis and for reviewing it.

I am thankful to university lecturer Martti Tolvanen who has been a great mentor throughout the entire degree program. I am grateful to all the teachers here in Tampere who have guided and motivated me in the entire duration of my Master's degree. I am really happy I got the opportunity to study in Finland, a country that highly values education and got the chance to experience a society of such helpful, honest and diligent people.

I am grateful to my parents for their emotional and financial support. They both have been an immense source of inspiration to me. My mother has always been a rock solid anchor who has always supported me and my father has instilled in me a scientific curiosity ever since I was a child. I am also thankful to my caring husband Ankur, for standing by my side through thick and thin.

I would like to thank all the amazing new friends I made in Finland who have been like an extended family and who have always helped and cheered me up whenever I needed it. Last but not the least I would like to thank all my friends and family members who have helped me in some way or the other through this whole process.

Master's Thesis

Place:	University of Tampere, Finland BioMediTech
Author:	Aruj Joshi
Title:	A Comparative Analysis of Thermodynamic Stability of Electron transport Chain Complex I Proteins and the Maximum Life Span of Host Organisms
Pages:	60 pages + Appendix 12 pages
Supervisors:	Professor Csaba Ortutay and Professor Matti Nykter
Reviewers:	Professor Csaba Ortutay and Professor Matti Nykter
Date	May 2014

Abstract

Background and Aims: "Mitochondrial theory of ageing" is an established theory which states that there is a correlation between Maximum Life Span (MLSP) of an organism and its mitochondrial free radical production. Here we are trying to extend this theory to determine if there is a correlation between stability of Complex I proteins of the electron transport chain (ETC) and MLSP. Through this we want to establish if there are Complex I proteins which might be involved in ageing of organisms. We selected Complex I because it is the first complex of the ETC and ETC is located in the mitochondria (free radical damage is the maximum in mitochondria as per the free radical theory of ageing).

Methods: Downloaded orthologs for all the 45 proteins of Complex I and used I-mutant program to find the stability of proteins after mutation. We applied all possible mutations at each site of the protein and then applied various methods to establish the change in protein stability which tells us whether the protein becomes more stable or less stable after the mutations. We then used the stability value and MLSP to find if there is a correlation (Spearman's ρ) between protein stability and MLSP and for the significance of the result, we find p-value corresponding to each ρ.

Results: Out of the 45 proteins, we got 19 proteins which had a significant correlation between MLSP and protein stability. Moreover some of these proteins were found to be involved in catalytic activity of the complex and some of the proteins with a significant MLSP and protein stability correlation were found to be encoded by the mitochondria.

Conclusion: Since there is a significant correlation between protein stability and MLSP for 19 out of 45 proteins, we can say that these proteins of Complex I might have some role in the process of ageing of organisms.

Contents

List of Figures

List of Tables

List of Abbreviations

MLSP	Maximum Life Span
ROS	Reactive Oxygen Species
NADH	Nicotinamide Adenine Dinucleotide - Hydrogen
ATP	Adenosine triphosphate
DNA	Deoxyribonucleic acid
mtDNA	Mitochondrial DNA
FMN	Flavin MonoNucleotide
ADP	Adenosine diphosphate
ETC	Electron transport chain
8-oxodG	8-oxo-7,8-dihydro-2'-deoxyguanosine
MELAS	Mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes
LHON	Leber Hereditary Optic Neuropathy
LDYT	Leber Hereditary Optic Neuropathy with Dystonia
LS	Leigh Syndrome
MT-C1D	Mitochondrial Complex I Disease
AD-MT	Alzheimer Disease Mitochondrial
LLS	Leigh like Syndromes
MLD	Macrocephaly with progressive Leukodystrophy
FILA	Fatal Infantile Lactic Acidosis
SVM	Support Vector Machine
HTML	Hyper Text Markup Language
XML	Extensible Markup Language
PBS	Portable Batch System
OMIM	Online Mendelian Inheritance in Man

1. Introduction

The stability of a protein depends on the change in free energy between a folded and unfolded protein which is also called the change in Gibbs free energy. A negative value of Gibbs free energy ΔG indicates a stable protein. When a protein is formed, it is in an unfolded state. From that state, it changes to a folded state (also called the native state of a protein) (Baldwin, 2007). The three dimensional structure of a protein which is also called tertiary structure is of prime importance because it determines the functionality of the protein (Petsko and Ringe, 2004). Tertiary structure is the way the whole protein folds to make a functional protein. It is the free energy change from the initial unfolded protein to this folded state protein which determines the stability of the protein (ΔG) (Creighton, 1990).

In many cases if a single amino acid in the protein is substituted by another residue, it makes the protein unstable (Capriotti et al., 2005; Cheng et al., 2006). In that case, the ΔG of the mutated protein is more than the ΔG of the wild type protein. Sometimes a substitution can even decrease the ΔG value in which case the mutated protein becomes more stable than the wild type protein. The result of reduced stability can be an unstable protein or a protein which is unsuitable because of loss of activity. There are many diseases which are caused by a mutation (which results in an amino acid change in protein). One such disease is sickle cell anaemia.

In the thesis, we studied electron transport chain Complex I proteins. The electron transport chain is in the mitochondria and its function is ATP production (Berg JM, 2002a). However as by products the electron transport chain also produces Reactive Oxygen Species (ROS) (Finkel and Holbrook, 2000). These are harmful for proteins, DNA and lipids. The cells have mechanisms to counter the effect of ROS however they are still able to cause some oxidative damage (Berg JM, 2002a). The electron transport chain is composed of four complexes (not including the ATP synthase which is sometimes called the fifth complex) (Berg JM, 2002a). Complex I is the first complex in the respiratory chain and is shaped like an L (Carroll et al., 2006; Yamaguchi et al., 1998; Grigorieff, 1999). It is composed of 45 proteins out of which 7 are encoded by mitochondrial DNA and 38 are encoded by nuclear DNA (Carroll et al., 2006).

There have been several studies which study the effect of oxidative damage to mitochondrial DNA and Maximum Life Span (MLSP) of species or the relationship between ROS produced by mitochondrial DNA and MLSP of organisms (Barja, 1998; Ku et al., 1993; Ku and Sohal, 1993). The mitochondrial DNA is present in the mitochondria and thus it is closely affected by ROS produced by the electron transport chain. By these studies, it was discovered that there is a negative correlation between damage to mitochondrial DNA and MLSP of various species (the same effect was not discovered with damage to

nuclear DNA).

In our study we will try to find out if there is a relationship between stability of electron transport chain Complex I proteins and MLSP of species. No such study has been made till now.

2. Review of Literature

The genes we have for our analysis are mitochondrial genes. More specifically they are genes which form the sub-components of NADH: ubiquinone oxidoreductase complex of the respiratory chain (Complex I) enzyme (also called NADH dehydrogenase (ubiquinone) in humans. In the eukaryotes, it is located in the inner mitochondrial membrane. This complex is composed of mitochondrial genes and nuclear genes.

2.1. Mitochondria and its Structure

Mitochondria have their own DNA. Each mitochondria contains multiple copies of mitochondrial genome. A mammalian mitochondrial genome consists of approximately 16.5 kilobases of circular genome (Hebert et al., 2010). Though mitochondria has its own DNA, a number of proteins used in the mitochondria are produced by the nuclear genome and are then transported to the mitochondria. NADH: ubiquinone oxidoreductase complex (complex I) which we will see in more detail is one such complex which is composed by proteins made from mitochondrial genome as well as nuclear genome.

Mitochondria are known as the powerhouse of the cell because they generate energy for the cell to use. They are rod shaped organelles and convert oxygen and nutrients to ATP. This process is called aerobic respiration. It produces 15 times more energy than anaerobic respiration. Mitochondria has two membranes called the inner membrane and outer membrane. The space between these two membranes is called the inter-membrane space. The inner membrane has folds which are called cristae and the space within the inner membrane is called matrix (Cooper, 2000). To understand the working of NADH: ubiquinone oxidoreductase complex and the position of our genes, we should look at the structure of a mitochondria first. It has a structure as given by the figure: 2.1.

2.2. Complex I

As we have seen above, Complex I is a complex which is composed of proteins made from nuclear genome as well as mitochondrial genome. There are basically four enzyme complexes which make up the respiratory chain. Out of these four, three are involved in pumping protons across the inter-membrane space. They are: NADH dehydrogenase or complex I, cytochrome reductase or complex III and cytochrome oxidase or complex IV (Weiss et al., 1992). Of these three enzymes, complex I is the first enzyme of the mitochondrial electron transport chain.

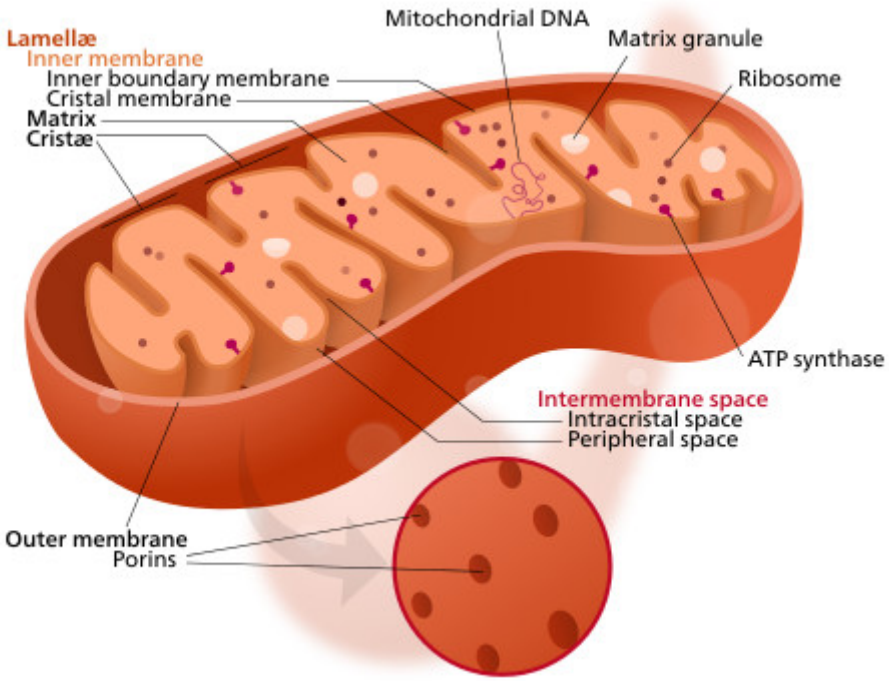


Figure 2.1.: The structure of a mitochondria showing the matrix, inner membrane, outer membrane and inter-membrane space. Source: (Kelvinsong, 2013).

2.2.1. Composition of Complex I

The enzyme has a mass of approximately 1 MDa (Hirst et al., 2003; Carroll et al., 2006; Murray et al., 2003). Analysis of bovine enzyme shows that it is composed of 45 subunits (Carroll et al., 2006). Another article listed the total components of Complex I as 46 (Lenaz et al., 2006). However in (Carroll et al., 2006) it has been proved to be a complex of 45 different proteins, a non-covalently bonded FMN and eight iron-sulphur complexes. Yet another recent article indicates that NDUFA4 which was initially thought to be a sub-unit of mammalian Complex I is in fact a sub-unit of Complex IV (Balsa et al., 2012). Thus making the total number of proteins which are a constituent of mammalian Complex I to be 44 instead of the previously known 45. In our study however we have included the NDUFA4 protein. The list of all genes which make up Complex I and their genomic location is given in table A.3 in appendix. These are the genes whose orthologs we have analysed for protein stability.

2.2.2. Structure of Complex I

Complex I has an L shape where one arm of L is in the inner mitochondrial membrane and the other perpendicular arm which is the hydrophilic arm (also called the peripheral arm) is in the mitochondrial matrix (Carroll et al., 2006; Yamaguchi et al., 1998; Grigorieff, 1999). The hydrophilic arm contains the NADH substrate binding site and flavin mononucleotide (FMN) (Brandt, 2006; Grigorieff, 1999). The hydrophilic arm also contains iron-sulphur clusters (8 of the 9 Fe-S clusters). 7 iron-sulphur clusters help in the transport of electrons along the complex (Brandt, 2006). The L shape of Complex I

is given by the figure: 2.2

2.2.3. "Core" of Complex I

Further analysis of Complex I indicates that out of the 44 subunits, 14 are called the central subunits also called the "core" of the complex (Brandt, 2006; Carroll et al., 2006). Of these 14 sub-units seven subunits are highly hydrophobic whereas the other seven are not(Brandt, 2006). The hydrophobic prediction is made by certain algorithms which predict the number of transmembrane helices in the various segments of the protein. The seven hydrophobic subunits are encoded by the seven mitochondrial genes (ND1-ND6 and ND4L) (Lenaz et al., 2006; Brandt, 2006). The other seven sub-units are encoded by nuclear genes (in fact except for the 7 proteins encoded by the mitochondrial genes, rest all the 37 proteins which form the other 37 subunits of the complex are encoded by nuclear genes). The genes encoding the other 7 proteins are: NDUFS1, NDUFV1, NDUFS2, NDUFS3, NDUFV2, NDUFS8, NDUFS7 (Brandt, 2006). The "Core" of complex I are also genes with catalytic activity. The genes with catalytic activity are mentioned in paper (Benit et al., 2004) and these names were also found from entrez (Maglott et al., 2005). The genes involved in catalytic activity are thus: NDUFS1, NDUFS2, NDUFS3, NDUFS8, NDUFS7, NDUFV1, NDUFV2, ND1, ND2, ND3, ND4, ND4L, ND5 and ND6.

2.2.4. Sub-complexes in Complex I

Another manner in which Complex I can be roughly segregated is by its 4 sub-complexes into which it readily separates in the presence of chaotropes (which are agents that in a water solution can disrupt hydrogen bonds) (Brandt, 2006; Carroll et al., 2006). The 1λ corresponds to the peripheral arm (Carroll et al., 2006) Sub-complex 1α is basically a combination of 1λ and 8 more sub-units (Hirst et al., 2003). The rest of the complex arm contains 1β and 1γ. The sub-complexes for Complex I are shown in the figure: 2.2.

2.3. Electron transport chain (also called the respiratory chain)

Electron transport chain is a series of protein complexes which help in moving electrons along the chain and in process, pump protons H^+ across a membrane. The electron transport chain in mitochondria consists of four complexes (Alberts B, 1994). They are NADH dehydrogenase (Complex I), succinate dehydrogenase (Complex II), cytochrome reductase (Complex III) and cytochrome c oxidase (Complex IV) (Berg JM, 2002b).

2.3.1. Overview of the electron (respiratory) transport chain

In the respiratory chain NADH and H^+ supply 2 electrons at Complex I. These electrons then pass through the complex and in turn pump 4 protons from matrix, across the in-

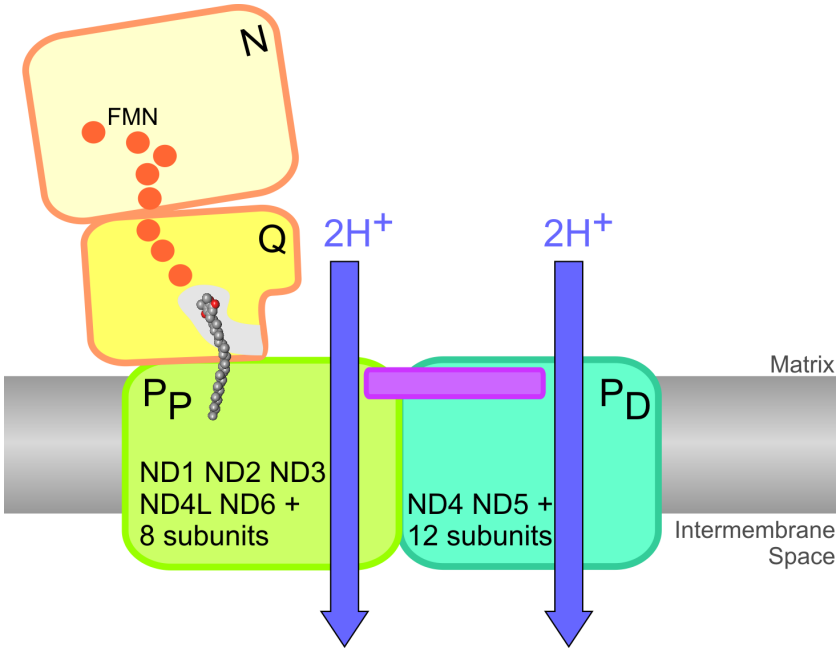


Figure 2.2.: The L shape of Complex I and the location of the peripheral arm and the hydrophobic arm. The spheres in N and Q are iron-sulphur clusters. The image also shows the sub-complexes of Complex I. N and Q (minus the FMN unit) which form the peripheral arm form 1λ. 1α is 1λ and 8 more sub-units from P_P. The remaining P_P and P_D is 1γ and 1β respectively. Source: (Dröse et al., 2011)

ner mitochondrial membrane to the inter-membrane space. Likewise complex III and IV pump protons across the membrane (4 protons and 2 protons respectively) (Berg JM, 2002b). In complex IV oxygen molecule in turn gets reduced to produce water molecule. Each NADH thus supplies 2 electrons to the chain which pumps 10 protons across the membrane and in the last complex (complex IV) one half oxygen molecule produces a water molecule. (Complex II does not pump any protons across the membrane). In complex II succinate is oxidised to fumarate and in this process two electrons are channelled into the chain from Complex II. However in this process no protons are pumped across the membrane like it happens in complex I, III and IV. Water production at Complex IV (Alberts B, 1994) happens as follows:

$$\frac{1}{2}O_2 + 2\,e^- + 2\,H^+ \longrightarrow 2\,H_2O \tag{2.1}$$

The electron transport chain is shown by the figure: 2.3

2.3.2. ATP production

Now this proton gradient that is developed helps in the production of ATP from ADP and phosphate(P_i). For this process ATP synthase enzyme which is also sometimes called Complex V is used. Due to the proton gradient, protons which are in excess in the inter-membrane space, cross the inner mitochondrial membrane through the ATP synthase enzyme. This crossing over causes rotation in the F_o subunit and the γ subunit of ATP

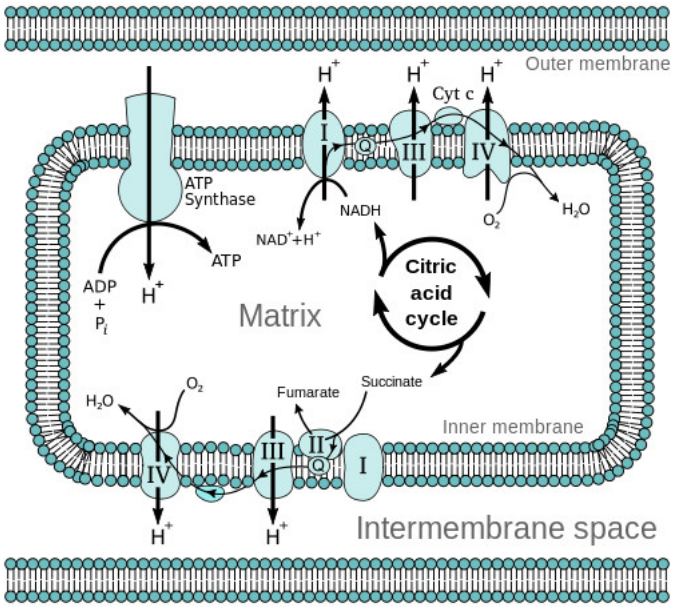


Figure 2.3.: Electron Transport Chain or Respiratory chain. Source: (Fvasconcellos, 2007).

synthase. This movement produces energy for the active sites in the β subunit which produces ATP (Schultz and Chan, 2001; Alberts B, 1994). This reaction is represented as follows:

$$\mathrm{ADP} + \mathrm{P_i} + \mathrm{H^+} + 4\,\mathrm{H^+_C} \longrightarrow \mathrm{ATP} + \mathrm{H_2O} + 4\,\mathrm{H^+_M} \tag{2.2}$$

The M and C subscript are for mitochondrial matrix and cytosol respectively.

2.3.3. Role of Complex I in the Electron transport chain

The role of Complex I in the electron transport chain is that it oxidises NADH and in that process, 2 electrons are donated to the complex and pass through the complex (being helped by the iron-sulphur clusters in their transfer to ubiquinone). Ubiquinone is a freely moving molecule which gets reduced to ubiquinol. In this complete process Complex I pumps 4 protons across its membrane and the ubiquinol travels via the inner mitochondrial membrane to complex III (bypassing complex II) (Schultz and Chan, 2001).

The equation for the reaction catalysed by Complex I is as follows (Schultz and Chan, 2001; Berg JM, 2002b):

$$\mathrm{NADH} + \mathrm{H^+} + \mathrm{CoQ} + 4\,\mathrm{H^+_M} \longrightarrow \mathrm{NAD^+} + \mathrm{CoQH_2} + 4\,\mathrm{H^+_C} \tag{2.3}$$

The M and C subscript are for mitochondrial matrix and cytosol respectively.

2.4. Oxidative Phosphorylation

As we saw above, in the electron transport chain, NADH which is produced in the citric acid cycle is oxidised to NAD^+, succinate is oxidised to fumarate and oxygen is reduced to water and in turn energy in the form of ATP is produced. This whole metabolic pathway is an example of oxidative phosphorylation occurring in the Electron transport chain (ETC) in the mitochondrial membrane.

2.4.1. Reactive oxygen species

Reactive oxygen species (ROS) are reactive molecules which have oxygen. We can also call them free radicals with oxygen. Free radicals are chemicals with unpaired electrons in their outer orbit (Turrens, 2003). Free radicals and ROS components are short-lived and highly reactive (because they are unstable and try to achieve stability by reacting with other molecules).There can be endogenous and exogenous factors which produce ROS. Exogenous factors would be smoking, pollution, UV rays and the like. Endogenous oxygen radical generation occurs in vivo and is caused by redox reactions (Beckman and Ames, 1998). Among other sites, the electron transport chain is a major site for ROS production. We are mainly talking about ROS regarding mitochondria because major intracellular ROS production stems from the mitochondria (Finkel and Holbrook, 2000). Superoxide anions are produced in Complex I, Complex II and Complex III of the ETC (Turrens, 2003). They are said to "leak electrons" to oxygen. The list of all ROS from the electron transport chain are (Turrens, 2003; Finkel and Holbrook, 2000):

- Oxygen (O_2)
- Superoxide anion ($O_2^{-\bullet}$)
- Hydroxyl ion (OH^-)
- Hydroxyl radical ($OH^\bullet$)
- Peroxide (O_2^{2-})
- Hydrogen Peroxide (H_2O_2)

The equation for superoxide and peroxide production is as follows (Berg JM, 2002b)

$$O_2 \xrightarrow{e^-} \underset{\text{superoxide}}{O_2^-} \xrightarrow{e^-} \underset{\text{peroxide}}{O_2^{2-}} \tag{2.4}$$

A point worth mentioning here is that the cells by themselves have a good anti-oxidant defence mechanism (Turrens, 2003; Finkel and Holbrook, 2000; Berg JM, 2002b) and despite that there is oxidative damage. Oxidative stress is the difference between ROS production and antioxidant defence (Finkel and Holbrook, 2000). The stress causes mutation in proteins, lipids and DNA. One of the enzymes produced by the cells to counter the effect of ageing is superoxide dismutase. It calalyzes the conversion of two superoxide

radicals into an oxygen and hydrogen peroxide(Berg JM, 2002b). The reaction is given below:

$$2\,O_2^{-\bullet} + 2\,H^+ \rightleftharpoons O_2 + H_2O_2 \tag{2.5}$$

2.5. Free radical theory of Ageing

The free radical theory of ageing was first proposed in the year 1956 (Beckman and Ames, 1998). The theory states that ROS play an important role in ageing of organism. Several experiments support the theory that free radicals play an important role in ageing even if they might not determine the life span of the organism (Beckman and Ames, 1998). Free radical theory of ageing gave way to the "rate of living hypothesis".

2.6. Rate of living hypothesis

According to the rate of living hypothesis, there is a negative correlation between metabolic rate and maximum life span(MLSP) (Beckman and Ames, 1998). The metabolic rate corresponds to the consumption of oxygen during a lifetime. However certain discrepancies were found in this theory. As per (Beckman and Ames, 1998), oxygen consumption of pigeon, canary, rats and guinea pigs is: 465, 1222, 28 and 48 litre O_2/g respectively and their MLSP is 37, 24, 4 and 8 years respectively. It was proved that MLSP is not correlated with the rate of consumption of oxygen. Also generation of oxidants is not correlated with the metabolic rate (which is oxygen consumption during a lifetime). Several experiments have confirmed that MLSP is correlated with the capacity of mitochondrial oxidant generation (Barja, 1998; Ku et al., 1993; Ku and Sohal, 1993).

In (Barja, 1998), they have compared adult male rats (with MLSP=4years) and adult pigeons (with MLSP 35 years). The O_2 consumption of the whole animal at rest, as well as the O_2 consumption of brain, liver and lung mitochondria were higher in the pigeon than the rat. However the mitochondrial free radical production of rat was considerably higher. This clearly indicates that MLSP is not related to O_2 consumption but related to the rate of oxygen radical production.

The free radical theory of ageing has evolved. Since it was first proposed, there have been some variations in the theory. One such idea proposed is the "mitochondrial theory of ageing". It deals with the role of mitochondrial DNA in ageing. This theory becomes clearer when we look at the "rate of living" hypothesis. As we saw above that there is an indication of a negative correlation between mitochondrial free radical production and MLSP. We will therefore concentrate on oxygen radical production in mitochondria. Also mutations occur at a higher rate in mitochondrial DNA (Linnane et al., 1989). As we have seen this above too, this is because mitochondria generates energy and in that process generates ROS which causes mutation (Shokolenko et al., 2009).

2.7. Further exploring the mitochondrial free radical theory of ageing

We will now look further into relation between:

- Free radical or reactive oxygen species production and lifespan of species.
- Oxidative damage to mitochondrial DNA and maximum lifespan of species.

In the initial rate of living hypothesis we had seen that it was proposed that there was a correlation between oxygen consumption (which is the metabolic rate) and MLSP (Beckman and Ames, 1998). However this theory was revised and it was proved that there was a correlation between mitochondrial oxidant production and MLSP (Barja, 1998; Ku et al., 1993; Ku and Sohal, 1993; Sohal et al., 1990). Slowly ageing species were found to have a lower rate of oxygen radical production in mitochondria and rapidly ageing species had a higher rate of oxygen radical production.

Another great article (Barja and Herrero, 2000) has explored if oxygen radical production is linked to oxidative damage as well. This gives a new dimension to the "mitochondrial theory of ageing". The article tried to establish a relationship between oxidative damage and MLSP instead of just oxygen radical production and MLSP. Oxidative damage was measured in the form of 8-oxo-7,8-dihydro-2'-deoxyguanosine (8-oxodG) which is an oxidative damage marker. They discovered that there is a strong correlation between oxidative damage in mitochondrial DNA and MLSP. According to their experiment, oxidative damage in heart for mitochondrial DNA was inversely correlated with MLSP ($r=-0.92$; $P<0.001$). MLSP was also inversely correlated with oxidative damage of mtDNA of brain ($r=-0.88$; $P<0.016$). However they did not find a significant correlation between oxidative damage in nDNA and MLSP (because 0.05 was selected as the point of minimum statistical significance and the p-value for correlation between MLSP and oxidative damage in nDNA in heart was $P<0.06$ and in brain it was $P<0.27$).

2.8. Thermodynamic stability of proteins

Protein stability depends on free energy change also called Gibbs free energy change. Change in Gibbs free energy function is represented as equation 2.6 (Creighton, 1990):

$$\Delta G = \Delta H - T\Delta S \tag{2.6}$$

It gives the free energy change between folded and unfolded states. Where ΔH is the enthalpy change between folded and unfolded states and ΔS is the change in entropy between folded and unfolded state. The factors responsible for enthalpy are

- Bonding energies: like disulphide bonds
- Non-bonding energies: Electrostatic interactions, Hydrogen bonds and Van der Waals forces.

We will review these terms in a bit more detail below. Likewise the factor responsible for entropy is: hydrophobic effect.

Gibbs free energy function for proteins can also be represented as 2.7 (Creighton, 1990):

$$\Delta G = G_{folded} - G_{unfolded} \tag{2.7}$$

Protein stability depends on free energy change between the folded and unfolded state of a protein. The folded state is also called the native state. The more negative the value of ΔG, the more stable the protein in its folded state. What the I-mutant program does is (Capriotti et al., 2005)

$$\Delta\Delta G = \Delta G^{Wt} - \Delta G^{Mut} \tag{2.8}$$

Where ΔG^{Wt} means change in free energy of the original protein (wild type) and ΔG^{Mut} is change in Gibbs free energy of the protein with a single site mutation. If the mutated protein is more stable, it will have a positive $\Delta\Delta G$ (ddG) and if the mutated protein is unstable, it will have a negative $\Delta\Delta G$ (ddG) value. The three dimensional structure of a protein is important because it determines the functionality of the protein (Petsko and Ringe, 2004). A protein which is not folded properly will not function correctly. The sequence of a protein is called its primary structure. Next, due to H-bonds between N-H and C=O of the backbone residues of the protein, there are secondary structure formations like alpha helices and beta sheets. The tertiary structure of a protein is a description of the way the whole chain (including the secondary structures) folds itself into its final 3 dimensional shape. The tertiary structure of a protein is held together by interactions between the side chains that is the "R" groups. When two or more polypeptide chains come together to form a structure, it is called a quaternary structure.

We have seen above that secondary structure is formed by H-bonds. The formation of tertiary structure is called folding of a protein and a stable tertiary structure is important for proper functioning of a protein (Petsko and Ringe, 2004). So we should look at factors responsible for formation of a stable 3-D structure. Some of the main factors which help in protein folding and on which the stability of a protein depends are (Petsko and Ringe, 2004; Berg JM, 2002a):

- Electrostatic interaction: Non-covalent interaction between atoms which are oppositely charged. For example it can form between carboxylate anion (COO^-) and protonated amino group (NH_3^+) (Nelson and Cox, 2004)

- Hydrogen bond: It is a form of non-covalent interaction between a hydrogen which is attached to an electronegative molecule (which pulls the electron cloud towards itself thus leaving a partial positive charge on hydrogen δ^+) and any atom which has a partial negative charge δ^-. In secondary structure (that is alpha helix and beta sheets) the hydrogen bonds are formed between the peptide backbone atoms and in tertiary structure, the hydrogen bonds form between the side chains. There are lots of side chains which have a hydrogen atom attached to either an oxygen or a nitrogen. They can easily form hydrogen bonds.

- Salt bridge: Salt bridges are interactions between oppositely charged amino acid side chains (Nelson and Cox, 2004). An example is that salt bridges can form between residues Glu^- and Arg^+. Both the residues are polar and charged though oppositely charged.

- Van der Waals force: Many amino acids have quite large hydrocarbon groups in their side chains (eg. Leucine, isoleucine and phenylalanine) Temporary fluctuating dipoles in one of the groups can easily induce an opposite dipole in another group on a nearby folded chain and their force would be strong enough to hold the folded structure together.

- Sulphur bridges: It is a covalent bond which is formed when two cysteine residues react to form a S-S link. Like sulphur bridges between two Cysteine.

Though the non-covalent interactions listed above are weak, yet in a protein these interactions are very large in number. That is why they play a very significant role in protein folding.

The amino acids can be classified as follows:

- Non-polar, hydrophobic amino acid: Contains: Gycine (Gly) G, Alanine (Ala) A, Proline (Pro) P, Valine (Val) V, Leucine (Leu) L, Isoleucine (Ile) I, Methionine (Met) M, Phenylalanine (Phe) F.

- Polar, non-charged amino acid: Contains: Serine (Ser)S, Threonine(Thr)T, Tyrosine(Tyr)Y, Cysteine(Cys)C, Aspargine(Asn)N, Glutamine(Gln)Q

- Polar, charged amino acid: Contains: Aspartic Acid (Asp)D, Glutamic Acid(Glu)E, Tryptophan(Trp)W, Histidine(His)H, Lysine(Lys)K, Arginine(Arg)R.

Thus we can see from the above classification that the kind of interaction between residues would depend on the kind of amino acid and if there is a mutation in amino acid, there will be a change in its chemical properties. It is the chemical properties that account for various "weak forces" that were discussed above. Thus they can be disrupted and in return can cause a change in the tertiary structure of the protein (which in return will affect the function of the protein or can make the protein unstable).

2.9. Diseases associated with Complex I genes in humans

2.9.1. Mitochondrial encephalomyopathy with lactic acidosis and stroke-like episodes syndrome (MELAS)

MELAS is a neuro-degenerative disease wherein the cells lose their ability to produce enough energy in the form of ATP (Santa, 2010). Not only is there energy deficiency, to make up for the deficiency, the cells switch to an alternative metabolic pathway which causes a build up of lactic acid in the body thus causing lactic acidosis (it is presence of low pH and build up of lactate in blood and body tissues) (Sproule and Kaufmann, 2008). There is also a presence of what is known as ragged-red fibres. This is basically a build-up of mitochondria in cells to compensate for the dysfunction of the respiratory chain (which produces ATP)(Santa, 2010).

Some of the symptoms of the disease are as follows (Hirano et al., 1992; Santa, 2010; Sproule and Kaufmann, 2008):

- There are stroke like episodes that too before the age of 40 years.
- Encephalopathy manifested in the form of seizures or loss of cognitive ability or both
- Exercise intolerance
- Lactic acidosis and presence of ragged-red fibres.

There are some other features too which are not present in all the cases and they are: recurrent headaches and vomiting.

2.9.2. Leber Hereditary Optic Neuropathy (LHON)

The disorder starts with visual impairment in one eye with the second eye generally following shortly (Yu-Wai-Man et al., 2009). This disorder can affect individuals ranging from 10 to 70 years of age and it affects males and females in the 3:1 ratio (Riordan-Eva and Harding, 1995). The peak age of onset of the disorder is however 15-30 years of age (Yu-Wai-Man et al., 2009). The progression of the disease can be in various phases (Yu-Wai-Man et al., 2009):

- Initially a loss in color vision is observed
- Blurring or clouding of vision in one eye. Which is generally followed by the other eye. There have however been cases where the other eye doesn't get affected for long periods of time. The visual impairment is painless and the vision is 6/60.
- After the above phase, the retinal nerve fibre degenerates.

In a few patients there even has been recovery after disease symptoms have appeared. Though the genes mutated in LHON disease are mitochondrial complex I genes (complex I is involved in the respiratory chain) there is no significant affect to the oxidative phosphorylation process (Yu-Wai-Man et al., 2009). However as they have stated in (Yu-Wai-Man et al., 2009), the studies could be tissue specific and need to be studied further.

Although the primary characteristic of the disease is loss in vision, there can be other features as well: cardiac arrhythmias and neurological abnormalities (Yu-Wai-Man et al., 2009).

2.9.3. Leber Hereditary Optic Neuropathy with Dystonia (LDYT)

This disease too is caused by mutations in the mitochondrial genes of complex I of the respiratory chain. The disease shows strict maternal inheritance just like LHON (De Vries et al., 1996). The disease is characterized by vision loss like LHON as well as with dystonia (De Vries et al., 1996; Jun et al., 1994). Dystonia is a neurological movement disorder which is characterized by involuntary muscle contractions. In the case study in paper (De Vries et al., 1996) the patients did not have any cadiological abnormalities however there were neurological abnormalities.

2.9.4. Leigh Syndrome (LS)

The disorder is due to deficiency in mitochondrial ATP production (Dahl, 1998). The symptoms can be motor or intellectual retardation, signs of brainstem dysfunction, and other neurological disorders like ataxia, dystonia and optic atrophy (Rahman et al., 1996). In most cases the disease is characterized by an early onset of the disease (within one year of being born) and can cause death within two years of onset of the disease though in some cases there is later onset and slower progression of the disease (Dahl, 1998; Rahman et al., 1996).

Leigh syndrome is found to be due to complex I , complex II, complex III and complex IV deficiency (Morris et al., 1996). However here we will focus on the disorder with respect to complex I deficiency. In complex I, all genes which are responsible for the catalytic part of Complex I except ND4L are involved in Leigh syndrome (Benit et al., 2004). The severity of the disease varies considerably depending on the mutations that are present (Dahl, 1998).

2.9.5. Mitochondrial Complex I deficiency (MT-C1D)

Mitochondrial Complex I deficiency is classified as an energy metabolism disorder. The clinical features of the disease vary widely. It can cause the following diseases: (Loeffen et al., 2000; Fassone and Rahman, 2012)

- Leigh and Leigh like syndromes (LS and LLS): Leigh syndrome presents in late infancy with neurological problems. Clinical signs include respiratory abnormalities, nystagmus, ataxia, dystonia and hypotonia (Fassone and Rahman, 2012).
- Cardiomyopathy: Refers to disease of the heart muscle. It is sometimes not recognised because there are already severe neurological disorders present in the patient (Fassone and Rahman, 2012).
- Macrocephaly with progressive Leukodystrophy (MLD): Leukodystrophies are white matter disorders which present themselves as progressive neurological disabilities (Kohlschütter and Eichler, 2011)
- Encephalomyopathy: encephalomyopathy refers to degenerative disease of the brain. This classification is for those barin degenrative diseases which have not been classified in some disorder already.

- Fatal Infantile Lactic Acidosis (FILA): FILA presents itself in neonatal or early infancy. It is a rapidly progressing disease which results in death in infancy (Fassone and Rahman, 2012).
- Mitochondrial encephalomyopathy, lactic acidosis and stroke-like episodes (MELAS): The symptoms include, seizures, migraines, vomiting, exercise intolerance, proximal limb weakness and short stature. Symptoms appear early on in childhood (Fassone and Rahman, 2012).

The nuclear encoded genes involved in this disorder are: NDUFV1, NDUFV2, NDUFS1, NDUFS2, NDUFS3, NDUFS4, NDUFS6, NDUFS7, NDUFS8, NDUFA2, NDUFA11, NDUFA10, NDUFB3, NDUFB9. The mitochondrial genes associated are (6 mitochondrial-encoded components of Complex I out of 7): ND1, ND2, ND3, ND4, ND5 and ND6.

2.10. Statistical Review

2.10.1. Spearman's Rho ρ.

It is a statistical test for measuring the relationship between two quantities (variables). This test is used when we want to measure the correlation between two variables and they do not necessarily have a straight line relationship. However the relationship should be monotonic. Monotonic means that the relationship is either entirely non-increasing or nondecreasing and is not for example like a 'U' turn (Weisstein, 2011). An example showing a monotonically decreasing function is figure 2.4 and an example of a non-monotonic function is figure 2.5. Spearman's rank correlation coefficient is a non-parametric method.

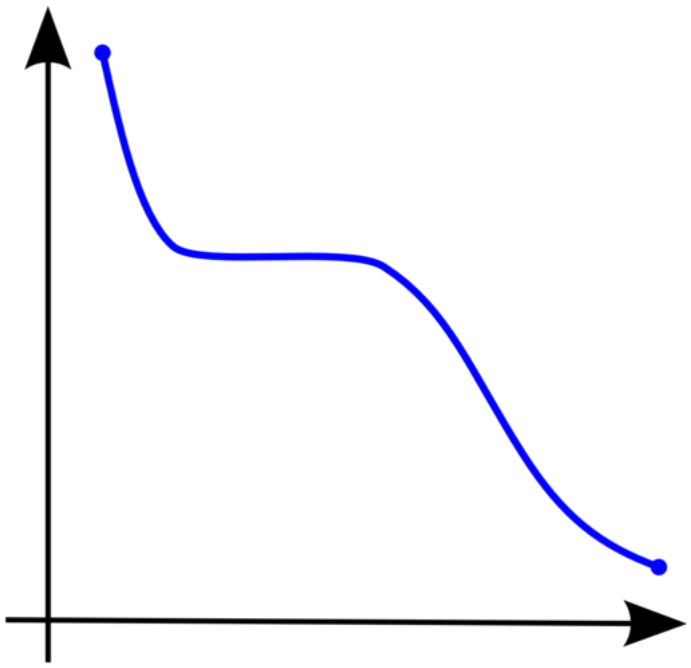


Figure 2.4.: Monotonically decreasing function. Source: (Oleg, 2007a).

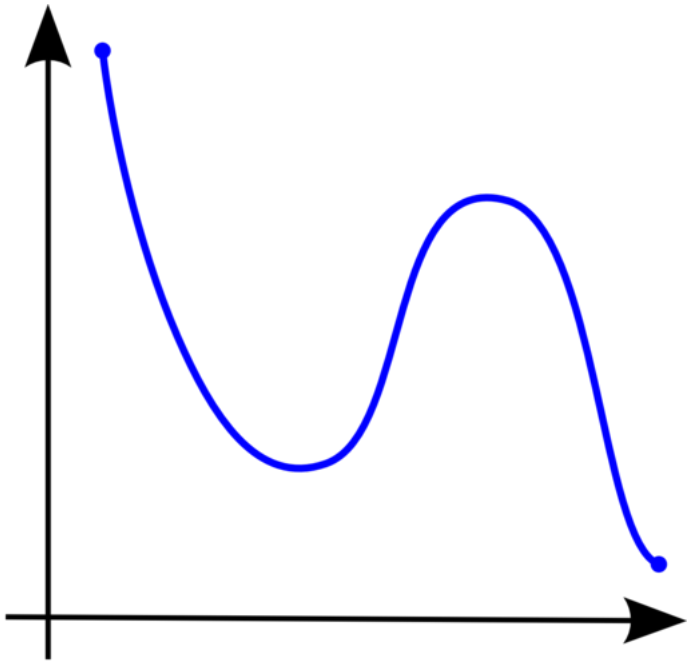


Figure 2.5.: Non-Monotonic function. Source: (Oleg, 2007b).

2.10.2. Parametric and non-parametric methods

When we cannot make an assumption about the shape of the distribution of measurements, we have to use methods which do not depend much on the shape of the distribution. These methods are non-parametric methods (Mosteller et al., 1983). They are called non-parametric because their behaviour does not rely a lot on the shape of the distribution (Mosteller et al., 1983). In non-parametric tests, very few assumptions are made about the distribution of the data and in particular, the data is not assumed to follow a normal distribution (Clarke and Cooke, 1998).

Some of the non-parametric methods are based on ranks or counts of objects or different sign assignment. Example of a method based on ranks is Spearman's ρ and based on counts of objects is median determination. And an example of method based on sign assignment can be that given a distribution, we assign a positive sign to all the numbers which are above a set threshold and a negative sign to all the numbers below a threshold. Then in such a case if we change the value of a number which was above the threshold to some value 10 times higher, the distribution will change, the mean and variance will all change however there will be no change in the sign assignment to the distribution. The ranks of the distribution as well as the median will not change. This example was mentioned in the explanation of non-parametric methods in (Mosteller et al., 1983).

Non-parametric methods are said to be resistant and robust (Mosteller et al., 1983). Example of resistance can be seen by the example discussed above (the median, sign and ranks did not change hence the term resistant for non-parametric methods). Non-parametric methods are robust meaning that where a parametric method can be used (that is based on the shape of the population we can use a parametric test) there, if we use a non-parametric method, we will lose very little information about the data. Parametric

methods on the other hand are based on the normal distribution which is also called a parametric distribution (Clarke and Cooke, 1998).

2.10.3. Spearman's correlation coefficient in a bit more detail

Spearman's rank correlation coefficient. Denoted by Rho ρ. Shows the statistical dependence between two variables.

First the data is ranked. Both the variables are ranked separately (Hollander et al., 2013). Suppose the the variables are height and age, then we rank them separately. In case there is a tie in value then we take an average of the ranks. Suppose two people have height 150 cm and the ranking till then is 4 then they both get a rank of 5.5 instead of one having a rank of 5 and other 6. Once we have calculated the rank in this manner, we use the following equation to calculate ρ, where x_i and y_i are ranks from our data. The equation for ρ is:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{2.9}$$

Another method to calculate rho is given by equation 2.10 (Maritz, 1995). This equation is used when there are no duplicate values in rank. Here $d_i = x_i$-y_i and n is the number of pairs or sample size.

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)} \tag{2.10}$$

Determining ρ by itself is not sufficient. We need to determine the significance of the ρ. For determining significance, we can use the Student's t distribution given by equation 2.11.

$$t = r \sqrt{\frac{n-2}{1-r^2}} \tag{2.11}$$

Where n-2 is the degree of freedom and n is the sample size.

2.11. I-mutant program and SVM

We used I-mutant program to find the relative stability of proteins after mutation. To understand the I-mutant program, we first need to understand Support Vector Machines. Support Vector Machine (SVM) is an algorithm which can be used for assigning continuous values or just assigning discrete labels. For example we use an SVM for an amino acid secondary structure prediction given the primary sequence or predicting the stability of a protein given the primary sequence and the single point mutation that is made to the original protein. In simpler terms, we can say that an SVM algorithm produces a model (based on training dataset) to predict class labels for test dataset (Hsu et al., 2003). For example in email classification task, there will be several features in each training dataset sample (which determine whether the email should be labelled as spam or non-spam). The spam and non-spam value is the label. There will be one label per input sample and the sample will have various features like occurrence of certain words or capitalization of words and the like. There will be certain datasets called training dataset which will be

used to train the algorithm or model (labels are available for it) and test dataset which is used to verify the model. Test dataset is a dataset which is similar to the training dataset however has no overlapping values. We also have labels for this dataset and we use the labels to verify the value predicted by our model. Once the prediction accuracy is good on the training dataset, the model can be used on an actual dataset. Training is a process by which a model assigns certain parameters like weight or bias so that the model can be used for actual dataset.

To explain the working of SVM in detail here, we will focus on SVM as a binary classifier. SVMs are of two types (Steeb, 2011):

1. Linear decision boundaries.

2. Non-linear decision boundaries.

2.11.1. Linear decision boundaries

Suppose our dataset is separable into two classes. Then in linear decision boundary, the data will be easily separable into the two classes by two margins which are parallel to the hyperplane (Steeb, 2011).

Hyperplane: Suppose we have an n-dimensional space then in that n dimensional space, a flat of (n-1) dimensions is called a hyperplane (Stanley, 2004). A hyperplane is a subspace of dimension one less than the dimension of the whole space (Heden et al., 2013) Suppose we have a two dimensional data which is clearly separable into two classes (that is it is linearly separable). Then in that case the hyperplane would be a straight line. In practical applications, our data is not two dimensional. Likewise for a three dimensional data, the hyperplane would be a two dimensional plane.

For a space, there are several hyperplanes (Heden et al., 2013). However the main aim of SVM is to find the optimal weight and bias parameters such that hyperplane corresponding to them separates the positive and negative training data with maximum margin (Steeb, 2011).

2.11.2. Non-linear decision boundaries

Here the data is not linearly separable that is no hyperplane to separate the data. In such a case, one procedure for classification could be to use a higher order polynomial feature vector (Ng, 2012; Burges, 1998). Suppose our input vector has two features x_1 and x_2 with a binary classification. The data type can be such that it is not separated by a straight line but instead by a circle or ellipse or some other irregular shape. Thus in this case, we will not be able to separate data simply by using x_1 and x_2 but will have to use a combination of terms like x_1x_2, x_1^2,x_2^2 and the like. This is simply when there are two features in our input vector (Ng, 2012). As the dimensionality of the vector increases, the complexity of the polynomial terms will also increase. Besides manually selecting the features would be cumbersome.

Thus a better method would be to use kernels (Steeb, 2011). The data (training data) is mapped to a higher dimensional space by a function ϕ such that it is transformed into a linear decision boundary problem and then a maximum margin hyperplane separating the data can be found just like for linear decision boundary problem (Steeb, 2011; Hsu et al., 2003). The kernel function is as: $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$. We don't have to worry about ϕ since we will be using kernels directly without focussing on ϕ.

Some of the common kernels are as follows (Hsu et al., 2003)

1. Linear: $K(x_i, x_j) = x_i^T x_j$
2. Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$
3. Radial basis function(RBF): $K(x_i, x_j) = exp(-\gamma ||x_i - x_j||^2), \gamma > 0$
4. Sigmoid: $K(x_i, x_j) = tanh(\gamma x_i^T x_j + r)$

γ , r and d are all kernel parameters. (x_i, y_i) is the training set where $i = 1, ..., l, x_i \in R^n$ (meaning there are l training samples and each sample has n features) and $y \in \{1, -1\}^l$.

2.11.3. I-mutant

I-mutant is a tool for predicting the protein stability upon single site mutation. It is a machine learning algorithm and uses support vector machines. The program can be used in two ways:

- Where we know the tertiary structure of our protein
- Where we only know the primary sequence of the protein

For I-mutant program, the accuracy of stability prediction in case of just primary sequence availability is 77% (Capriotti et al., 2005) The program predicts the stability by giving the relative stability change (ddG) value of the protein when a mutation is applied to it at any amino acid site. This process is composed of two parts: predicting sign of ddG value and predicting the actual ddG value. In the first method, it tells whether the mutation has caused the protein to become stable or unstable. In it, the magnitude of the relative stability change is not taken into account. In fact in many applications, the sign of relative stability is more important than the actual magnitude of (ddG) (Cheng et al., 2006).

The dataset to train the support vector machine(SVM) and to test it is obtained from ProTherm database (Gromiha et al., 2000). Further they have extracted the tertiary structure information about the proteins from Protein Data Bank (Berman et al., 2000). They selected data from ProTherm in such a manner that only proteins with single point mutations were selected and those values were selected where the protein stability change (ddG) was experimentally determined with known experimental conditions.

The program works for both cases: where only sequence is available and where sequence and structure both are available. We get to choose the position and new residue as well as the temperature and pH values. If we are not dealing with changes in temperature and pH we leave those two parameters as such (the default values are pH 7 and temperature 25 °C. In our case we made no change to the default pH and temperature values).

For the encoding scheme, they have an input vector with 42 values. The first two values of the vector correspond to temperature and pH at which the stability of the mutated protein is measured. Next 20 values are used to represent the amino acids. The amino acid which will be mutated is assigned a value of -1, the new amino acid is assigned a value of 1 and the rest of the positions corresponding to the remaining 18 amino acids are assigned a 0. The next 20 input values serve different purposes when the sequence structure is known and when just the sequence is known. Since our usage of the I-mutant program deals with just sequences, we will look at that case. It is used to represent a total of 19 amino acids including the mutated residue. Thus we can say that we consider a window of size 19 for this program. The input vector is generally called a feature vector. Each input vector would also have a corresponding (ddG) whose value would be known. Several of such feature vectors and corresponding (ddG) value (or sign when we just have labels) will compose a training data and test data. The kernel used is radial basis function(RBF) which is: $K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2), \gamma > 0$ (Capriotti et al., 2005). The function $f(x)$ for SVM is defined as 2.12 (Cheng et al., 2006; Burges, 1998).

$$f(x) = \sum_{x_i \in S^+} \alpha_i K(s_i, x_i) - \sum_{x_i \in S^-} \alpha_i K(s_i, x_i) + b \tag{2.12}$$

$f(x)$ is the predicted value of (ddG) and simply the sign of $f(x)$ will be enough to classify the data as: stable after mutation or unstable after mutation (Cheng et al., 2006). S^+ denotes data points where (ddG) > 0 and S^- denotes (ddG) < 0. α_i is the non-negative weight assigned to the training data, b is bias and s_i is the support vector.

Thus to understand s_i, we need to understand what a support vector is. While training, there is a sub-set of the training data that is very close to the hyperplane separating the dataset into the two labels (in case of a binary classification task). Classifying such data points is a bit tricky since they lie close to the separating hyperplane. Thus they could be classified on either side. However their correct classification is essential for the SVM to make a correct prediction. This training data subset plays a determining role while modelling the SVM (that is in the learning of the weight and bias vector). This sub-set of training data is called the support vector (Alpaydin, 2004).

After training, the weight vector can be written down in terms of this training data sub-set. As we know, there are many separating hyperplanes for the same training data. However the main aim of a Support Vector Machine is to find the optimal weight and bias parameters such that hyperplane corresponding to them separates the positive and negative training data with maximum margin (Steeb, 2011).

What is provided to us users is a trained SVM which can be used for the prediction of protein stability upon single site mutation. The user can provide the site of mutation and the new amino acid (as well as change in temperature and pH) and the program

will predict the relative stability change. The program can be accessed from its website (`http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi`) (Capriotti et al., 2005)). In our case we had to make stability prediction for a large dataset so we used the program in our department cluster. We used I-mutant 3.0 and we just used the sequence only option because we did not have tertiary structure available for our sequences. For the tests that were made by the designers of I-mutant, the accuracy for sequence only data (we need the sequence only case because our dataset is sequence only) is 77%. In the I-mutant program, a positive (ddG) value indicates that the protein is stable after the mutation and a negative (ddG) value indicates that the protein is unstable after the mutation.

3. Objectives

The objectives of our study are as follows:

1. Check if there is a significant correlation between the stability of complex I proteins and MLSP of the species which have that protein. That is find if proteins in Complex I emulate the MLSP and oxidative damage correlation.

2. Check if there is a difference between complex I proteins: the "core" proteins of complex I which are involved in catalysis and those with non-catalytic properties. In short we want to check if the proteins which belong to "core" of complex I are in any way more or less stable than the rest of the proteins or vice versa.

3. Check if there is a difference between proteins encoded by the mitochondria and nuclear encoded proteins when it comes to protein stability vs MLSP relationship.

4. We will apply three different methods to check for the stability of proteins. So we will check how many proteins behave similarly across the three methods.

5. Make pair comparisons between methods and find similarity or differences in the behaviour of proteins across the different methods. For example we will compare the positive correlation result of one method with the positive result of another method. Likewise for negative correlation values. We will also check if for a protein there is a positive correlation between stability and MLSP by one method and a negative correlation by another method.

4. Tools and Methods

4.1. Tools

4.1.1. Perl, BioPerl, Ensembl and Perl API

Perl: The scripting language used in the initial phase of the project wherein we had to download all the orthologs of the human genes for Complex I was done using Perl. We used Perl 5 (`http://www.perl.org`) because that was compatible with the Ensembl API version we used. We had the requirement to use Perl because we were accessing Ensembl database via its Perl API.

BioPerl: BioPerl (bioperl.org) is an open source project which provides bio-informatics modules for use with Perl. The tool-kit is written completely in Perl (Stajich et al., 2002). We did not use BioPerl directly but we used Ensembl Perl API which makes use of BioPerl. For our use, BioPerl version 1.2.3 was used. Like Perl, it can be used in Unix, MacOS or Windows.

Ensembl and Perl API: There are three methods to access the Ensembl database: Usings APIs:Perl or REST, through MySQL and Via BioMart. For our requirement the best option was to use Perl API. It has a middle level of learning required and ease of data extraction. We wanted to extract the orthologs of all the genes in our list. To use Perl API, we need to install the API and also install BioPerl version 1.2.3. While using Perl we do not have to know much about the database schema of Ensembl but we have sufficient power to extract any information that is present on the website. We indirectly queried the Ensembl core database. Since we used the API we did not need to know the database schema. There is also a short Ensembl API tutorial provided which makes using the API slightly easier (`http://www.ensembl.org/info/docs/api/core/core_tutorial.html`).

4.1.2. Shell scripting

We used biocluster.uta.fi for mutating our protein using the I-mutant program. The task of submitting the jobs to the I-mutant program was done using Unix shell scripting. We also used shell scripting to compile the needed information from the files generated by the I-mutant program.

4.1.3. Python, Biopython, Beautiful Soup

Python: We used Python (`https://www.python.org/`) an open source scripting language for simple operations like removing ortholog sequences which are incomplete (which had Xs) and making a fasta file from the protein sequences that were extracted from Ensembl. Apart from that we had to extract the list of species names from the ortholog sequences list and then find its corresponding MLSP from (`http://genomics.senescence.info/species/`). To extract MLSP from the website, we used BeautifulSoup which needs Python and Numpy installed. I have used Python 2.7 because Beautiful soup and Biopython modules work well with it.

Biopython: It is a freely available open source tool for common Bio-informatics tasks (`http://biopython.org/`). Biopython has modules for interacting with various tools like BLAST, ClustalW and EMBOSS. It can access various online databases like Entrez and ExPASy and read and write different sequence file formats and perform several operations on them (Cock et al., 2009). Biopython can be used on major operating systems like Unix, Mac OS and Windows. We have used BioPython on Ubuntu (a Linux based operating system). Version used is biopython-1.63. To use Biopython, we need Python and NumPy installed. We used the module to download information about all of our 45 genes from Entrez.

Beautiful Soup: It is a Python library used to extract information from HTML and XML files (Richardson, 2007). We used it to extract MLSP value for our ortholog species. We used Beautiful Soup 4. Instead of manually searching for each specie name on the web to extract its MLSP, we used Beautiful Soup 4 to extract that information along with extracting the common name of the organism.

4.1.4. R: Statistical analysis tool

R is a free software programming language used for statistical analysis (`http://www.r-project.org/`). It can be used on all platforms like Unix, Mac OS and Windows (Hornik, 2014). I used R version 3.0.2 on a Linux system.

4.1.5. I-mutant

The I-mutant program was installed in biocluster50 (bc50) which was reachable from Biocluster.uta.fi. We used shell scripting to perform mutation at each site for all the orthologs of all the proteins. Each site was mutated to 19 possible amino acids. We had a separate file for each site mutation which consisted of ddG value of the protein for each of the 19 possible mutations for that site. Once we had all the mutation files, we used shell scripting to extract information from those files so that we had a single file for a single ortholog. This file contained information compiled from all the site files of that ortholog. This site had the wild type amino acid value, the sum of positive and negative mutations for each site and the number of positive and negative mutations for each site. I sshed this compiled information to my Linux system on which R version 3.0.2

was installed and it was used for data analysis.

4.2. Methods

4.2.1. Overall procedure

The overall procedure applied for the study is as follows:

1. Download the protein sequences of all the orthologs for all the genes in our gene list. (Our gene list has genes which form a part of Complex I of the electron transport chain).

2. Remove those protein sequences which are incomplete (these are sequences which have Xs in their sequence) These make up less than 5% of the orthologs.

3. Submit the protein sequences to I-mutant program. Apply all possible mutations at each site for each ortholog.

4. There is one file per site containing all 19 possible mutations for the protein (the file contains information of whether the mutation was destabilizing or stabilizing). There are however as many files per ortholog as there are sites in it. So we combine information from all sites for that ortholog and write it to a single file. The file name contains species name as well as the Ensembl Id for that transcript of the ortholog.

5. Once we have one file per ortholog, we extract the information we need: The wild type Amino acid, the number of stabilizing mutations (mutations which give a positive ddG value), the number of destabilizing mutations (mutations which give a negative ddG value), the sum of stabilizing ddG values and the sum of destabilizing ddG values.

6. Make a list of all the species for which we need to find the Maximum Life Span (MLSP).

7. Find the Maximum Life Span(MLSP).

8. Find general information about all the 45 genes from Entrez (from this information, we will extract information about the diseases associated with these genes and which gene is involved in which disease).

9. Use the MLSP and the files with extracted I-mutant information in R to find if a relationship exists between the stability of the proteins and the MLSP of the species for the various electron transport chain Complex I proteins that we have.

4.2.2. Get all the orthologs for each human gene in our gene list

This program was written in Perl using Ensembl Perl API. It was used to extract all the homologs for each gene in our gene list. We extracted information for all homologs but didn't write the paralogs for the human gene (since we only need orthologs in our analysis.)

Algorithm 1 Get all the orthologs for all the genes

Require: Open the file with Ensembl gene Id names
while stable Id in file with human Ensembl gene Ids **do**
Import Registry module (This is the module that helps connect to the Ensembl database)
Then make a connection to the Ensembl Registry
Obtain the Gene adaptor using Registry (with species name as 'Human' and database type as 'Core') (Gene adaptors are object adaptors which can be used to extract information from the Ensembl database)
Use Gene adapter to fetch the Gene by its stable Id
For the Gene extract the canonical transcript
Get all the homologs for the human transcript
Create one fasta file for one gene Id (this file will have all the homologs of the human protein sequence)
Create description for fasta sequence for human protein by concatenating Ensembl gene Id and species name (here homo sapiens) with symbol "_" as a separator in between the Ensembl Id and species name
Write the description line
Write the translation of the canonical transcript extracted
for Each homolog in homologs **do**
Obtain species name from the homolog information
Obtain new Gene adaptor using Registry (with species name as obtained by the homolog and database type as 'Core')
Use the Gene adapter to fetch the Gene by its stable Id
For the Gene extract the canonical transcript
If the translation for the transcript is defined, we define the description line for the species (to add to our homologs fasta file)
Write the description line
Write the translation (of the canonical transcript extracted)for the species to the homolog file too
end for
Close the gene homologs file
end while
Close the file with Ensembl gene Id names

4.2.3. Remove incomplete protein sequences

Incomplete sequences are those which have 'Xs' in the sequence. Those are sites for which the amino acid is not known. There are, on average, less than 5% of sequences which

fall in this category. Those sequences were removed from our fasta files. The script for removal of incomplete sequences was written in Python.

Algorithm 2 Remove incomplete protein sequences

```
  function REMOVEX(name)
      Open file with name in read mode
      Open another file in write mode (for files without X character in sequences)
     for line in file do
        if line is odd numbered then
      description = line
        else if sequence line has no 'X' then
      Write the description line corresponding to this sequence to file
      Write the sequence line to file
        end if
     end for
      Close both files
  end function
Require: Open file with all fasta file names
  for name in file do
      REMOVEX(name)
  end for
      Close file
```

4.2.4. Submission of protein sequences to the I-mutant program

We use shell script for the purpose of submitting our jobs to Biocluster for execution by I-mutant. We make use of a cluster here. When we can divide a task into multiple smaller tasks which can be executed in parallel, then we make use of clusters. Clusters are basically large number of computers which are also called nodes. There are different softwares that manage jobs between the systems. The jobs are submitted in a queue. The name of the cluster in our lab was Biocluster which had 50 nodes and out of these 50 nodes, 23 were assigned for mutating sequences by I-mutant program and finding the stability of the mutated protein. To submit jobs in a cluster, we login to any one node and submit our jobs. I submitted my jobs from bc50 (which was biocluster 50) which was accessible from biocluster.uta.fi.

First we use a shell script which creates different directories for different Ensembl Ids and invokes the script that makes setting changes to the cluster and submits jobs to it. The algorithm for it is given by Algorithm 3. Each directory will contain files produced by the I-mutant program.

Algorithm 3 Submission of orthologs to I-mutant program for mutation

Require: Open the Ensembl Ids file
for stable ensembl Id in the file of Ensembl Ids **do**
Make a directory with ensembl Id as name
Copy the fasta file with that ensembl Id name into that directory
Change directory to the newly created directory
Extract all descriptions from all the fasta sequences from the fasta file
These descriptions minus the ">" symbol form the names of the protein files whose sequences will be submitted to the I-mutant program
Execute the shell script which makes settings in the cluster and submits jobs to the cluster. (How changes to cluster settings is made and what are the various jobs is discussed below)
end for

For making settings to bc50, we had used PBS which is a job resource manager. It is used on clusters and provides queuing facility and job execution services. There are several job queues. We used the queue named "batch". Each queue has a "maximum walltime". Walltime basically means that if any job submitted to it takes more than the "walltime", the job is terminated. In our case, a walltime of 24 hours was allotted (however each job in our case just took a few seconds). We make these setting by a PBS script which is basically just a normal unix script with some comments in the beginning which are called PBS directives. These are comments that start with #PBS. We can use the following commands to make some settings:

- #PBS -l walltime=HH:MM:SS
- #PBS -l nodes=N:ppn=M
- #PBS -q queuename

In #PBS -l nodes=N:ppn=M, nodes=N specifies the number of nodes and ppn=M specifies the number of processors per node. In our case we had not specified any processors but just the number of nodes (which was 23).

We used qsub to submit our jobs and after submitting the jobs, we could monitor the jobs at any time using mon and qstat. In our case the mutation of proteins (each site gets mutated by the remaining 19 amino acids for all the orthologs for each Ensembl human protein of complex I). Each mutation that is applied to a site calls the I-mutant program and is treated as a job. Each job is assigned a process Id and is submitted to one of the nodes of the cluster.

We submitted the bash file with an & in the end so that the processing of the script can continue even after we have logged out from the server. This is because not all the jobs get submitted to the queue in one go but it is an ongoing process where mutations are created and they are submitted to the I-mutant program and the result obtained and stored in a file such that there is one file per site of a protein. That one site file contains information about the stability of the protein for each of the 19 mutations that can be applied at that site.

4.2.5. Extract information from files produced by I-mutant

For each ortholog, there are n number of files (created by the I-mutant program) where n is the number of sites in that protein sequence. So first we run a bash script that concatenates the information of all the n files into one file. Each such concatenated file is then processed by a perl script which parses the I-mutant file and extracts information from it. What the perl script does is that it finds the wild type amino acid for each site of the protein sequence (from the .out file) passed to it, along with the sum of positive (ddG) values, sum of negative ddG values, number of positive ddG mutations and number of negative ddG mutations per site. These files along with the MLSP file is used for data analysis in R. The algorithm for the Perl script is as follows:

Algorithm 4 Extract information from concatenated I-mutant files

Require: open the I-mutant concatenated file
for each line in file **do**
if a match with " Position" pattern is found (this is because position is found only in the beginning of each site) **then**
Read the next line and split it so as to obtain AA position, wild type AA at that site, new AA after mutation.
Initialize npos (number of mutations with positive ddg value at that site) to 0
initialize nneg (number of mutations with negative ddg value at that site) to 0
Initilize sumneg and sumpos to 0 as well
sumneg is sum of negative ddG values for a site
sumpos is sum of postive ddG values for the site
end if
if a match with " DDG Value Prediction:" pattern is found **then**
if the ddG value is greater than 1 **then**
Increment the npos by 1
Add the value of positive ddG mutation to sumpos
else
Increment the nneg (number of negative ddG values) by 1
Add the value of negative ddG mutation to sumneg
end if
end if
if a match with an http pattern is found **then**
Make a single record with position of AA, the wild type AA, sumpos, npos, sumneg, nneg.
This is information for one site
Write the record to a file which has information about all sites of the protein (for that particular species)
end if
end for

We now have a single file with information extracted from the I-mutatnt file for a particular protein, for a particular species.

4.2.6. Collect names of species for which we have protein sequences

We need to collect the names of all the species for which we have protein sequences available because we need to extract the MLSP of those species (which we will need in our further analysis with R). The script for extraction is written in Python. The name of the species is mentioned in the description row of the fasta sequences. The description line starts with the symbol '>'. The Ensembl Id and the species name form the description line and the two are separated by an "_" symbol. We use this information to extract the species name. The algorithm for it is given below.

Algorithm 5 Extract species name

```
function extractSpecies(line, speciesList)
    Find position of "_" symbol call it position
    Species name is a slice of line from position + 1
    if the species name is not in speciesList then
    Add the species name to the speciesList
    end if
end function
Require: Open file with list of genes in read mode
    Open a new file to which we will write names of species
    speciesList = [] (empty list)
for gene name in file do
    Open the fasta file with the gene name (this has all the orthologs of the human gene by that name)
    for line in the fasta file do
        if first element of line is '>' then
        extractSpecies(line, speciesList)
        end if
    end for
    Close the fasta file that was opened
end for
    Write the species name to file
    Close file with species name
    Close file with gene names
```

4.2.7. Extract MLSP

Maximum Life Span of species has been extracted from (Tacutu et al., 2013).The script is written in Python and makes use of Beautiful Soup and requests. Both are Python libraries. Requests is an HTTP library and Beautiful Soup is a library for extracting data out of HTML and XML files. Here requests is used to obtain web page content and then Beautiful Soup is used to parse the webpage to extract required data from it. Apart from this we take

Algorithm 6 Extract MLSP

```
Require: Import BeautifulSoup and requests
  function EXTRACTWEBPAGE(base url, species)
      In species name, replace " " with a "+"
      New url is base url + new species name
      Request a session
      Requests.session is used to obtain the content of the webpage whose url we have
      We create a BeautifulSoup object on the webpage content
      Find all the "td" tags in the webpage using Beautiful Soup object
     for row in rows with "td" tag do
        if row has "em" tag too then
      val = row, where text is True
      add val to an empty info list
      val contains value of MLSP and common name of species
        end if
     end for
      return info
  end function
Require: Define the base url from which information has to be extracted
      Open file with species name
      Open maximum lifespan file to which info extracted from webpage will be written
  for species in the species name file do
      Speciesinfo = EXTRACTWEBPAGE(base url, species)
      Convert Speciesinfo extracted from function to string
      Write to maximum life span file
      Wait for 6 seconds and then continue
  end for
```

4.2.8. Fetch Entrez Information

To be able to extract all possible information about our genes from NCBI, we used a python script (which also used Biopython) to extract information about the human genes in our gene list. The information mainly consisted of the diseases the genes are involved in. This was important to get a better idea of our genes and to see which out of the 45 genes had been researched more and had actual diseases associated with it. We could get the general idea of the diseases associated with the genes just by looking at a few genes but for an exhaustive list we had to look at each gene individually. Instead of the cumbersome task of looking up each gene in NCBI (`http://www.ncbi.nlm.nih.gov/`) we used Python and Biopython to extract the required information.

Algorithm 7 Fetch Entrez Information

Require: Import Entrez and SeqIO from Bio (part of Biopython)

Open the gene Ids file in read mode

Open the file to which we will save the gene information (in write mode)

for gene in gene Id file **do**

In the protein database of Entrez, search for "Gene" and "Organism" where "Gene" is the gene name from file and "Organism" as human.

Once we have obtained the handle by the above method, we read the record and from there get the gene Id used in Entrez for that gene for humans.

We use that gene Id to fetch the record from Entrez using the protein database.

We get our required information from the comments section of annotations of the sequence record.

We write it to a file

end for

Close both the files.

4.2.9. Analysis of data by R

We use R to find Spearman's correlation co-efficient and p-value as well as to get various scatter plots depicting the relationship between the MLSP and stability of proteins. The data is analyzed in three ways. We have a file for each ortholog which consists of the wild type amino acid at each site along with how many stabilizing mutations occur at that site, how many destabilizing mutations occur at that site, the sum of the stabilizing mutations and the sum of de-stabilizing mutations. From this information, we have made three different types of analysis.

1. In the first case, we find the mean value of ddG for each site and then the mean value for the whole protein. We then find the Spearman's correlation coefficient between the mean ddG value and the MLSP of the species (along with the p-value). We also make scatter plots for all the 45 genes.
2. In the second case we find the mean ddG value for each site and then pick the minimum value of ddG among all those sites for the whole protein (we have picked the most destabilizing mutation for the whole protein) and then find the Spearman's correlation coefficient and scatter plots as we had done before.
3. In the third case, we use a ratio of number of stabilizing mutations for the protein and length of the protein. This is because the paper (Cheng et al., 2006) mentions that in many cases, the sign of ddG matters more than its actual magnitude. Even if we consider this particular case, by just taking into account the ratio of stabilizing mutations and length of protein, we get the general idea of how stable a protein is after mutations.

5. Results

5.1. The sub-cellular location of proteins

We used Biopython to collect information about the complex I genes (on our list) from Entrez (Maglott et al., 2005). We then parsed that information to obtain sub-celluar location of all the genes for which the information was available. For our list of 45 genes, sub-cellular location of 34 genes was available. The information obtained is given in table A.7 in Appendix. We also parsed information from the file to find out the list of diseases in which Complex I proteins are involved. The diseases from the list have been discussed in review of literature chapter. From the information along with the names of diseases, I collected the genes responsible for the various diseases which have been discussed in subsection 5.2. I cross-referenced the gene list of each disease from OMIM (omi, 2014). Explanation of the various sub-cellular locations are as follows:

- Peripheral membrane proteins: They are attached to the exterior of the lipid bilayer. They can be removed without harming the lipid bilayer. It is in stark contrast to integral proteins which are embedded within the lipid bilayer. (Integral proteins can be transmembrane in which case they extend through the membrane. In this case part of the chain is in the bilayer and part of it is outside. The hydrophobic parts of the protein are inside the membrane and hydrophilic parts are outside).

- Single-pass membrane protein: It is a transmembrane protein where the polypeptide chain goes through the membrane only once. They thus have their C terminus and N terminus at opposite ends

- Multi-pass membrane protein: It is a transmembrane protein where the polypeptide chain goes through the membrane more than once.

5.2. The diseases associated with complex I genes

The diseases Complex I genes are involved in are as follows: (This information was parsed from protein information extracted for all the genes from entrez using biopython) (have not listed the generic symptoms that are caused by some genes but complete diseases caused by them)

- Mitochondrial Complex I deficiency (MT-C1D)

- Leber Hereditary Optic Neuropathy (LHON)

- Leber hereditary optic neuropathy with dystonia (LDYT)
- Mitochondrial encephalomyopathy with lactic acidosis and stroke-like episodes syndrome (MELAS)
- Leigh syndrome (LS)
- Alzheimer disease mitochondrial (AD-MT)

5.2.1. Mitochondrial Complex I deficiency (MT-C1D)

As per the gene information downloaded from entrez we get to know that the OMIM Id of the disease is: 252010 and that the following genes are involved in Complex I deficiency: NDUFV1, NDUFS1, NDUFS2, NDUFS4, NDUFS7, NDUFA1, NDUFA11, MT-ND6, MT-ND3 and MT-ND5. As per OMIM databse, the Complex I genes involved in Mitochondrial Complex I deficiency are as follows: NDUFS2, NDUFB3, NDUFS1, NDUFS6, NDUFS4, NDUFB9, NDUFS3, NDUFV1, NDUFV2, NDUFA11 and NDUFA1.

5.2.2. Leber's Hereditary Optic Neuropathy (LHON)

The genes involved in this disease are as follows: ND1, ND2, ND6, ND5, ND4 and ND4L. This information is extracted using biopython program which uses entrez (Maglott et al., 2005) and cross-referenced manually with (omi, 2014). As we see above all the genes are mitochondrial genes. The OMIM Id of the disease is: 535000.

5.2.3. Leber Hereditary Optic Neuropathy with Dystonia (LDYT)

Genes involved in LDYT are as follows: ND1, ND6, ND4 and ND3. The OMIM Id of the disease is: 500001.

5.2.4. Mitochondrial encephalomyopathy with lactic acidosis and stroke-like episodes syndrome (MELAS)

As per OMIM the Complex I genes responsible are: ND1, ND5, ND6. As per information extracted from entrez the genes involved are: ND1, ND4, ND5, ND6. From entrez, we get to know that the OMIM Id of the disease is: 540000.

5.2.5. Leigh Syndrome (LS)

Leigh syndrome is caused by deficiency in Mitochondrial complex I, II, III, IV and V. Here however we have just discussed it in terms of Complex I. The Complex I genes involved in Leigh syndrome (as per OMIM) are as follows:

- NDUFA2 : Leigh syndrome due to mitochondrial complex I deficiency
- NDUFS3 : Leigh syndrome due to mitochondrial complex I deficiency
- NDUFS8 : Leigh syndrome due to mitochondrial complex I deficiency
- NDUFA9 : Leigh syndrome due to mitochondrial complex I deficiency
- NDUFA12 : Leigh syndrome due to mitochondrial complex I deficiency
- NDUFS7 : Leigh syndrome due to mitochondrial complex I deficiency
- NDUFA10
- NDUFS4

As per information extracted from entrez the genes involved are as follows: NDUFV1, NDUFS7, NDUFS8, NDUFA12, ND3 and ND5. The OMIM Id of the disease is: 256000.

5.2.6. Alzheimer Disease Mitochondrial (AD-MT)

As per information extracted from entrez, the genes involved in this disease are as follows: ND1 and ND2. The OMIM Id of the disease is: 502500.

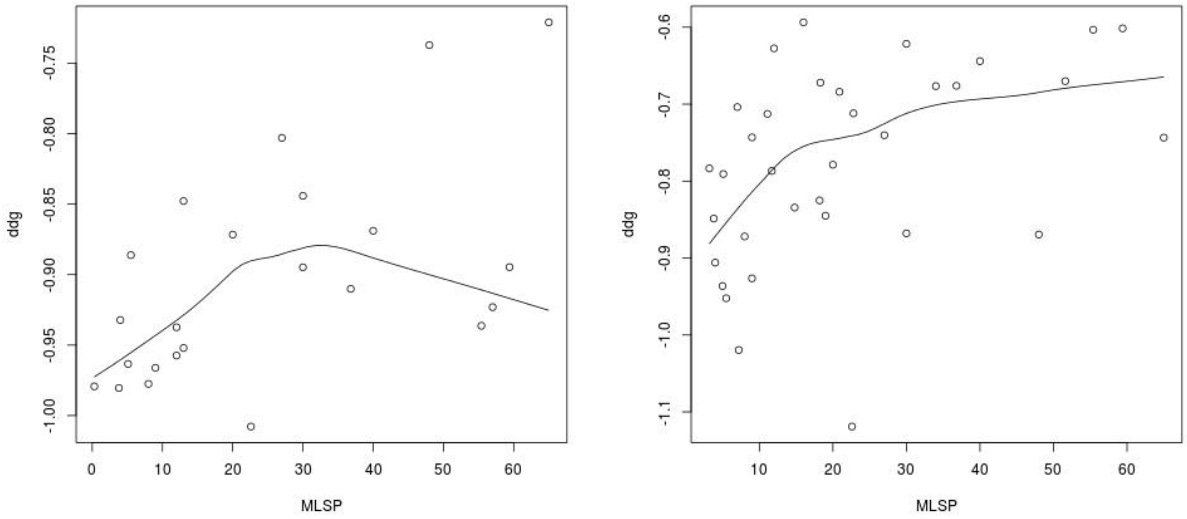
5.3. Results of statistical analysis

The MLSP of all the species involved is given in appendix A.1 and A.2. Anywhere in this thesis, where MLSP term is used, it means MLSP of the species in years. We have also removed Homo sapiens from the graphs because MLSP for humans is 122.5 years which is much higher than the second highest MLSP (almost double) thus it was making reading and analysing of graphs difficult.

5.3.1. Analysis for mean of mean of ddG

In mean of mean of ddG, we take mean ddG value at each site and then we take the mean ddG value of all sites and that is considered the stability of the protein. Greater the value, more stable the protein. To find the correlation, we use Spearman's correlation coefficient which is given by ρ. And to find the significance of each of the ρ, we have also listed the p-value associated with it. Here, there are 11 proteins with p-value less than 0.05 (which we set as the threshold). They are given in table 5.1 with their ρ and p-value.

Out of these 11 proteins, 5 proteins have a positive value of ρ. That means for those proteins, the relative stability value ddG increases with increase in MLSP. These proteins are: NDUFS8, ND2, NDUFB4, NDUFB6 and NDUFB10. The plots for ND2 and NDUFB6 are given by figure 5.1 and discussed below. The rest of the plots with positive correlation are given by B.1 in appendix. Proteins with negative ρ are: NDUFS1, NDUFS3, NDUFS5, NDUFA6, NDUFB5 and NDUFC2. That means for these proteins, the relative stability value ddG decreases with increase in MLSP. The plots for NDUFA6 and NDUFC2 are given below by figure 5.2 and discussed below. The rest of the plots are given by figure B.4 in appendix. The complete table for ρ and p-value for this method is given by table A.4 in appendix. Significant p-value values (11 proteins) are marked in blue color.



(a) ND2: ρ=0.62, p-val=0.01348 (b) NDUFB6: ρ=0.52, p-val=0.00172

Figure 5.1.: Relationship of mean mean ddG and MLSP with positive correlation

genes	ρ	p-val mean of mean
NDUFS1	-0.44	0.01348
NDUFS3	-0.35	0.04199
NDUFS8	0.45	0.00716
NDUFS5	-0.37	0.02835
NDUFA6	-0.58	0.00017
ND2	0.62	0.00177
NDUFB4	0.34	0.04872
NDUFB5	-0.46	0.00569
NDUFB6	0.52	0.00172
NDUFB10	0.45	0.00543
NDUFC2	-0.55	0.00328

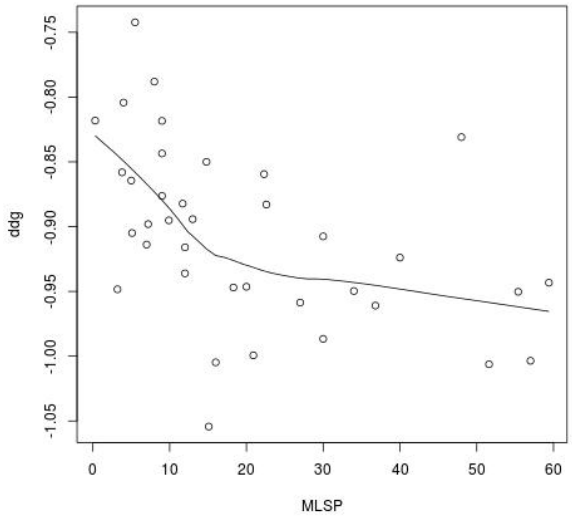
Table 5.1.: Rho and p-value for mean of mean ddG

In figure, 5.1a we see that there is a strong positive correlation between mean mean ddG and MLSP for protein ND2. Meaning that the greater the stability of ND2 for an organism, the higher the MLSP for it. In the graph, the four lowest ddG values are: -1.0078, -0.9804, -0.9793 and -0.9775 (in increasing order) and they correspond to species *Ornithorhynchus anatinus* (Common name: Duck-billed platypus, MLSP: 22.6), *Rattus norvegicus* (Common name: Norway rat, MLSP: 3.8), *Drosophila melanogaster* (Common name: Fruit fly, MLSP: 0.3) and *Gasterosteus aculeatus* (Common name: Alaskan stickleback, MLSP: 8) respectively. In the same figure, the four highest ddG values are: -0.7211, -0.7373, -0.803 and -0.8441 (in decreasing order) and they correspond to species *Loxodonta africana* (Common name: African elephant, MLSP: 65), *Latimeria chalumnae* (Common name: Coelacanth, MLSP: 48), *Sus scrofa* (Common name: Wild boar, MLSP: 27) and *Felis catus* (Common name: Domestic cat, MLSP: 30) respectively.

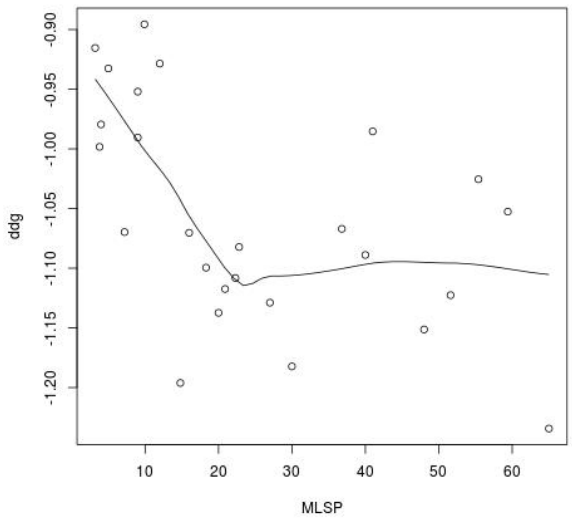
In the values above, we can see that the Duck-billed platypus does not conform to the general ddG and MLSP trend of the graph. It has a high MLSP despite low protein stability.

In 5.1b we see that there is a strong positive correlation between mean mean ddG and MLSP for protein NDUFB6. Meaning that the greater the stability of NDUFB6 for an organism, the higher the MLSP for it. In the graph, the four lowest ddG values are: -1.1190, -1.0196, -0.9523 and -0.9363 (in increasing order) and they correspond to species *Ornithorhynchus anatinus* (Common name: Duck-billed platypus, MLSP: 22.6), *Anolis carolinensis* (Common name: Green anole, MLSP: 7.2), *Danio rerio* (Common name: Zebrafish, MLSP: 5.5) and *Oryzias latipes* (Common name: Japanese medaka, MLSP: 5) respectively. In the same figure, the four highest ddG values are: -0.5936, -0.6015, -0.6032 and -0.6215 (in decreasing order) and they correspond to species *Tarsius syrichta* (Common name: Philippine tarsier, MLSP: 16), *Pan troglodytes* (Common name: Chimpanzee, MLSP: 59.4), *Gorilla gorilla* (Common name: Gorilla, MLSP: 55.4) and *Felis catus* (Common name: Domestic cat, MLSP: 30) respectively.

In the values above, we can see that the lowest ddG value, the Duck-billed platypus with MLSP: 22.6 does not conform to the general trend. Like ND2, for NDUFB6 too the Duck-billed platypus has an exceptionally high MLSP despite lowest protein stability among all species. The highest ddG value, Philippine tarsier with MLSP: 16 does not conform to the general trend too. Despite the highest protein stability, it has a considerably low MLSP.



(a) NDUFA6: ρ=-0.58, p-val=0.00017



(b) NDUFC2: ρ=-0.55, p-val=0.00328

Figure 5.2.: Relationship of mean mean ddG and MLSP with negative correlation

In 5.2a we see that there is a strong negative correlation between mean mean ddG and MLSP for protein NDUFA6. Meaning that the higher the stability of NDUFA6 for an organism, the lower the MLSP for it. In the graph, the four lowest ddG values are: -1.0543, -1.0062, -1.0048 and -1.0035 (in increasing order) and they correspond to species *Macropus eugenii* (Common name: Tammar wallaby, MLSP: 15.1), *Tursiops truncatus* (Common name: Bottlenosed dolphin, MLSP: 51.6), *Tarsius syrichta* (Common name: Philippine tarsier, MLSP: 16) and *Equus caballus* (Common name: Horse, MLSP: 57) respectively. In the same figure, the four highest ddG values are: -0.7423, -0.7880, -0.8042 and -0.8181 (in decreasing order) and they correspond to species *Danio rerio* (Common name: Zebrafish, MLSP: 5.5), *Gasterosteus aculeatus* (Common name: Alaskan stickle-

back, MLSP: 8), *Mus musculus* (Common name: House mouse, MLSP: 4) and *Drosophila melanogaster* (Common name: Fruit fly, MLSP: 0.3) respectively.

In the values above, Tammar wallaby, MLSP: 15.1 and Philippine tarsier, MLSP: 16 do not follow the trend of the graph; that lower protein stability protein value corresponds to higher MLSP. These two species have quite low MLSPs in comparison to Bottlenosed dolphin, MLSP: 51.6 and Horse, MLSP: 57 which have similar protein stability values as Tammar wallaby and Philippine tarsier.

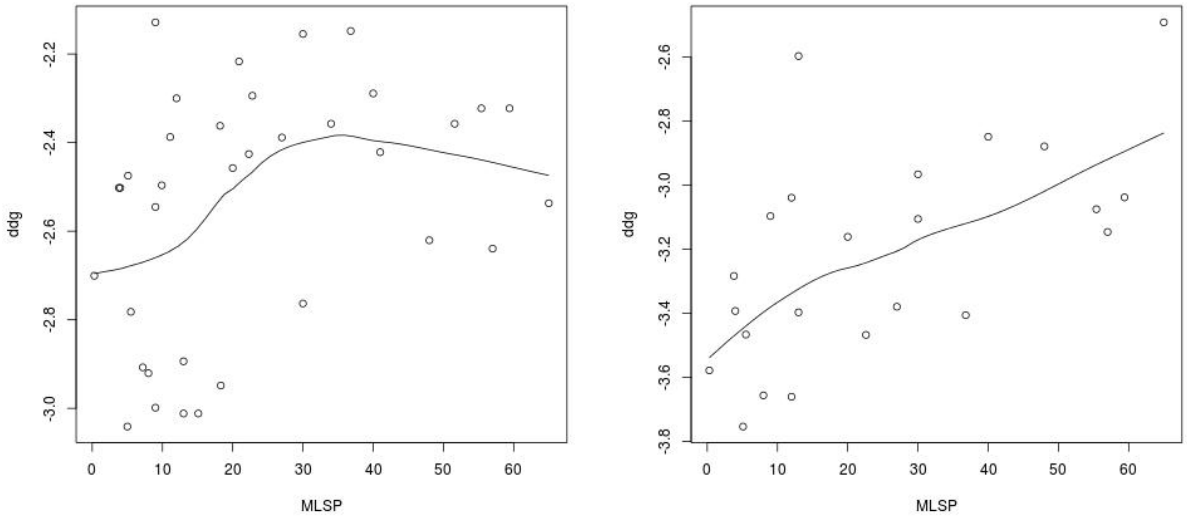
In 5.2b we see that there is a strong negative correlation between mean mean ddG and MLSP for protein NDUFC2. Meaning that the higher the stability of NDUFC2 for an organism, the lower the MLSP for it. In the graph, the four lowest ddG values are: -1.2343, -1.1961, -1.1822 and 1.1513 (in increasing order) and they correspond to species *Loxodonta africana* (Common name: African elephant, MLSP: 65), *Procavia capensis* (Common name: Rock hyrax, MLSP: 14.8), *Felis catus* (Common name: Domestic cat, MLSP: 30) and *Latimeria chalumnae* (Common name: Coelacanth, MLSP: 48) respectively. In the same figure, the four highest ddG values are: -0.8956, -0.9154, -0.9285 and -0.9326 (in decreasing order) and they correspond to species *Dipodomys ordii* (Common name: Ord's kangaroo rat, MLSP: 9.9), *Sorex araneus* (Common name: Eurasian shrew, MLSP: 3.2), *Cavia porcellus* (Common name: Guinea pig, MLSP: 12) and *Oryzias latipes* (Common name: Japanese medaka, MLSP: 5) respectively.

In the above values, Rock hyrax, MLSP: 14.8 does not follow the trend of NDUFC2 protein graph. It has a considerably lower MLSP for low stability of NDUFC2 protein unlike other species for the same protein example: African elephant, MLSP: 65 and Coelacanth, MLSP: 48 (which also have low stability values).

5.3.2. Analysis for minimum of mean ddG

In minimum of mean ddG value, we take the mean ddG value of all possible mutations at each site. However we select the minimum value from all the site means as the ddG value to represent the stability of protein. The minimum ddG value means we are selecting the most destabilizing mean ddG value for each protein. After that Spearman's correlation is found for minimum ddG value and MLSP of all the organisms in which the protein is found (and for which we have the MLSP). In total 10 proteins with p-value less than 0.05 were found they are given in the table 5.2. Out of the 10 proteins, 4 have a positive correlation and rest 6 have a negative correlation.

Those proteins with a positive correlation are: NDUFA7, ND2, NDUFB5 and NDUFB6. The ddG value vs MLSP plots for NDUFA7 and ND2 are given in the figure 5.3 and discussed in detail below. The plots for the rest of the proteins with positive correlation are given in figure: B.2 in appendix. The proteins with negative correlation are: NDUFS3, NDUFA13, NDUFC1, ND4L, NDUFA4 and NDUFB11. The plots for NDUFB11 and NDUFA13 are given in figure 5.4 and discussed below. The plots for the rest of the proteins are given by figure B.5 in appendix. The complete table for ρ and p-value for this method is given by table A.5 in appendix. Significant p-value values (10 proteins) are marked in blue color.



(a) NDUFA7: ρ=0.42, p-val=0.01117 (b) ND2: ρ=0.62, p-val=0.0015

Figure 5.3.: Relationship of min mean ddG and MLSP with positive correlation

genes	ρ	p-val min of mean
NDUFS3	-0.44	0.00755
NDUFA7	0.42	0.01117
NDUFA13	-0.51	0.00445
NDUFC1	-0.45	0.02428
ND2	0.62	0.0015
ND4L	-0.41	0.04619
NDUFA4	-0.46	0.00527
NDUFB5	0.36	0.03155
NDUFB6	0.39	0.02421
NDUFB11	-0.62	0.00076

Table 5.2.: Rho and p-value for min of mean ddG

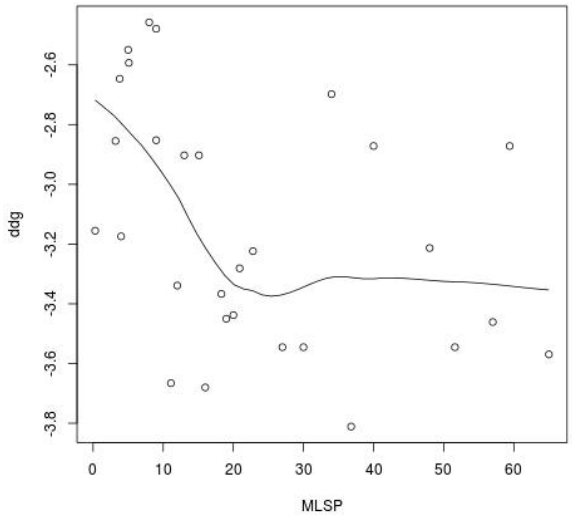
In figure, 5.3b we see that there is a strong positive correlation between min mean ddG and MLSP for protein ND2. Meaning that the greater the stability of ND2 for an organism, the higher the MLSP for it. In the graph, the four lowest ddG values are: -3.7537, -3.6605, -3.6563 and -3.5784 (in increasing order) and they correspond to species *Monodelphis domestica* (Common name: Shorttailed opossum, MLSP: 5.1), *Cavia porcellus* (Common name: Guinea pig, MLSP: 12), *Gasterosteus aculeatus* (Common name: Alaskan stickleback, MLSP: 8) and *Drosophila melanogaster* (Common name: Fruit fly, MLSP: 0.3) respectively. In the same figure, the four highest ddG values are: -2.4911, -2.5963, -2.8484 and -2.8784 (in decreasing order) and they correspond to species *Loxodonta africana* (Common name: African elephant, MLSP: 65), *Meleagris gallopavo* (Common name: Wild turkey, MLSP: 13), *Macaca mulatta* (Common name: Rhesus monkey, MLSP: 40) and *Latimeria chalumnae* (Common name: Coelacanth, MLSP: 48) respectively.

In the values above, we can see that the Guinea pig, MLSP: 12 does not conform to the

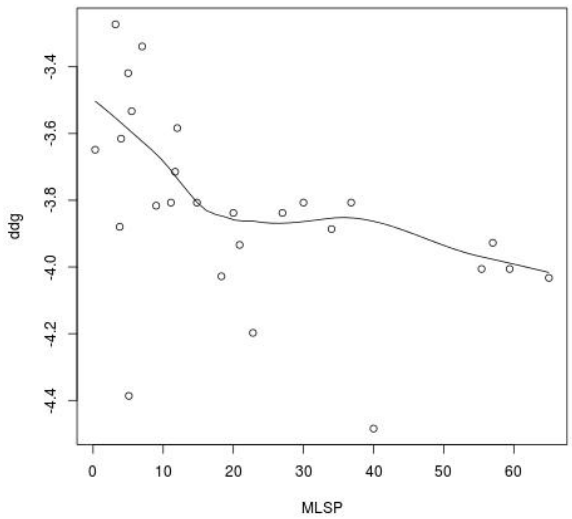
general ddG and MLSP trend of the graph for ND2. It has a considerable high MLSP corresponding to its low ND2 protein stability. Other species with similar stability value are: Shorttailed opossum, MLSP: 5.1 and Fruit fly, MLSP: 0.3 which have quite a low MLSP. Another organism in the above values which doesn't conform to the graph is Wild turkey, MLSP: 13. It has a low MLSP even though it has quite a high ND2 protein stability.

In 5.3a we see that there is a positive correlation between min mean ddG and MLSP for protein NDUFA7. Meaning that the greater the stability of NDUFA7 for an organism, the higher the MLSP for it. In the graph, the four lowest ddG values are: -3.0411, -3.0116, -3.0116 and -2.9984 (in increasing order) and they correspond to species *Oryzias latipes* (Common name: Japanese medaka, MLSP: 5), *Macropus eugenii* (Common name: Tammar wallaby, MLSP: 15), *Sarcophilus harrisii* (Common name: Tasmanian devil, MLSP: 13) and *Oreochromis niloticus* (Common name: Nile tilapia, MLSP: 9) respectively. In the same figure, the four highest ddG values are: -2.1284, -2.1479, -2.1547 and -2.2168 (in decreasing order) and they correspond to species *Oryctolagus cuniculus* (Common name: Old World rabbit, MLSP: 9), *Ailuropoda melanoleuca* (Common name: Giant panda, MLSP: 36.8), *Felis catus* (Common name: Domestic cat, MLSP: 30) and *Pteropus vampyrus* (Common name: Large flying fox, MLSP: 20.9) respectively.

In the values above, we can see that the highest ddG value, Old World rabbit, MLSP: 9 does not conform to the general trend. It has a considerably lower MLSP in comparison to other species which have similar protein stability for NDUFA7.



(a) NDUFA13: ρ=-0.51, p-val=0.00445



(b) NDUFB11: ρ=-0.62, p-val=0.00076

Figure 5.4.: Relationship of min mean ddG and MLSP with negative correlation

In 5.4a we see that there is a strong negative correlation between min mean ddG and MLSP for protein NDUFA13. Meaning that the higher the stability of NDUFA13 for an organism, the lower the MLSP for it. In the graph, the four lowest ddG values are: -3.8111, -3.6800, -3.6658 and -3.5695 (in increasing order) and they correspond to species *Ailuropoda melanoleuca* (Common name: Giant panda, MLSP: 36.8), *Tarsius syrichta* (Common name: Philippine tarsier, MLSP: 16), *Tupaia belangeri* (Common name: Northern tree shrew, MLSP: 11.1) and *Loxodonta africana* (Common name: African elephant, MLSP: 65) respectively. In the same figure, the four highest ddG values are: -2.4574, -2.4789, -2.5495 and -2.5932 (in decreasing order) and they correspond to species *Gasterosteus aculeatus* (Common name: Alaskan stickleback, MLSP:

8), *Oreochromis niloticus* (Common name: Nile tilapia, MLSP: 9), *Oryzias latipes* (Common name: Japanese medaka, MLSP: 5) and *Monodelphis domestica* (Common name: Short-tailed opossum, MLSP: 5.1) respectively.

In the values above, Philippine tarsier, MLSP: 16 and Northern tree shrew, MLSP: 11.1 do not follow the trend of the graph; that lower protein stability corresponds to higher MLSP. These two species have quite low MLSPs in comparison to African elephant, MLSP: 65 and Giant panda, MLSP: 36.8 which have high MLSP corresponding to low protein stability of NDUFA13.

In 5.4b we see that there is a strong negative correlation between min mean ddG and MLSP for protein NDUFB11. Meaning that the higher the stability of NDUFB11 for an organism, the lower the MLSP for it. In the graph, the four lowest ddG values are: -4.4842, -4.3863, -4.1974 and -4.0326 (in increasing order) and they correspond to species *Macaca mulatta* (Common name: Rhesus monkey, MLSP: 40), *Monodelphis domestica* (Common name: Shorttailed opossum, MLSP: 5.1), *Callithrix jacchus* (Common name: White-tufted-ear marmoset, MLSP: 22.8) and *Loxodonta africana* (Common name: African elephant, MLSP: 65) respectively. In the same figure, the four highest ddG values are: -3.2726, -3.3389, -3.4189 and 3.5332 (in decreasing order) and they correspond to species *Sorex araneus* (Common name: Eurasian shrew, MLSP: 3.2), *Ochotona princeps* (Common name: North American pika, MLSP: 7), *Oryzias latipes* (Common name: Japanese medaka, MLSP: 5) and *Danio rerio* (Common name:Zebrafish, MLSP: 5.5) respectively.

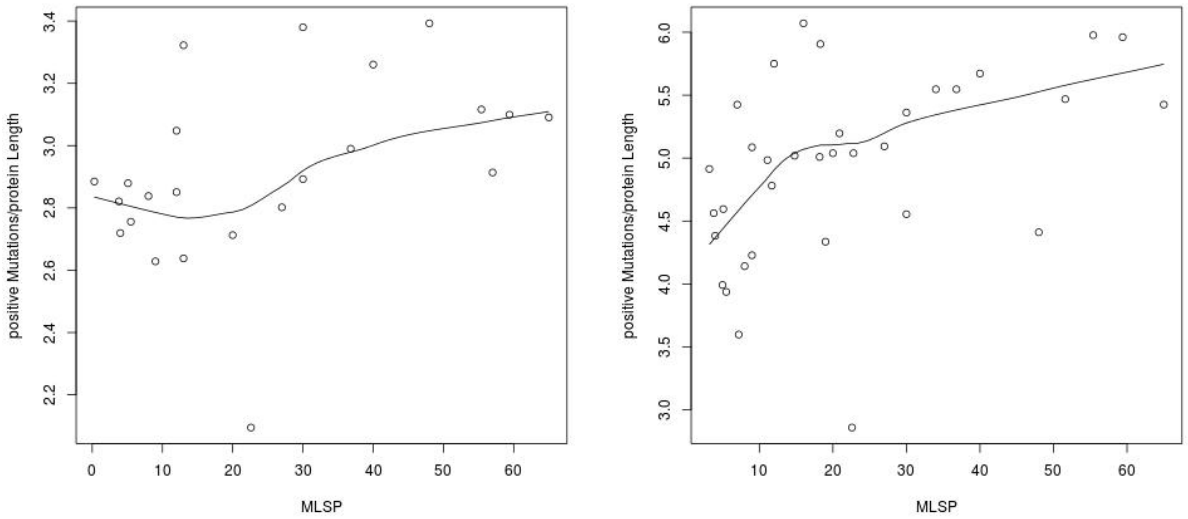
In the values above, Shorttailed opossum, MLSP: 5.1 does not follow the trend of the graph; that lower protein stability corresponds to higher MLSP. This organism has quite a low MLSP in comparison to Rhesus monkey, MLSP: 40, African elephant, MLSP: 65 and even White-tufted-ear marmoset, MLSP: 22.8 which have high MLSP corresponding to low protein stability (for NDUFB11).

5.3.3. Analysis for ratio of positive mutations and length of protein

In analysis of ratio of positive mutations, we calculate the ratio of the sum of all positive mutations in a protein and the length of the protein. Here we have discarded proteins which are less than two-thirds the average length of all the orthologs. This ratio is then plotted against the MLSP of all the species where the ortholog is found. The Spearman's correlation coefficient and p-value are then calculated. There were 10 proteins with p-value of less than 0.05 that is 10 proteins are statistically significant. They are given in table 5.3 along with their ρ and p-value.

The proteins with positive correlation coefficient are: NDUFS8, NDUFB6, NDUFB10 and ND5. The plots for NDUFB6 and ND5 are given in figure 5.5 and discussed below. The plots for the rest of the proteins with positive correlation are given in figure B.3 in appendix. The proteins with negative correlation are: NDUFV2, NDUFS3, NDUFS5, NDUFC2, NDUFA6 and NDUFB5. The plots for NDUFA6 and NDUFB5 are given in figure 5.6 and discussed below. The plots for the rest of the proteins with negative

correlation are given in figure B.6 in appendix.The complete table for ρ and p-value for this method is given by table A.6 in appendix. Significant p-value values (10 proteins) are marked in blue color.



(a) ND5: ρ= 0.55, p-val= 0.00606 (b) NDUFB6: ρ=0.55, p-val=0.00081

Figure 5.5.: Relationship of ratio and MLSP with positive correlation

genes	ρ	p-val of ratio
NDUFV2	-0.39	0.0126
NDUFS3	-0.43	0.00921
NDUFS8	0.47	0.00496
NDUFS5	-0.37	0.02665
NDUFA6	-0.59	0.00013
NDUFB5	-0.5	0.00229
NDUFB6	0.55	0.00081
NDUFB10	0.38	0.02112
NDUFC2	-0.52	0.0059
ND5	0.55	0.00606

Table 5.3.: Rho and p-value for ratio of number of positive mutations and length of protein ddG

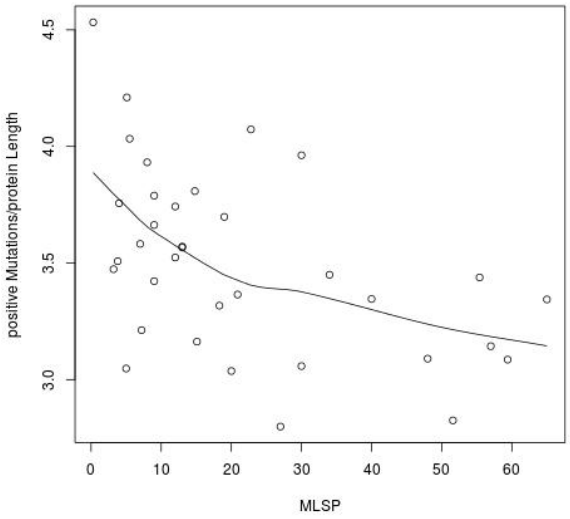
In figure, 5.5a we see that there is a strong positive correlation between ratio and MLSP for protein ND5. Meaning that the greater the stability of ND5 for an organism, the higher the MLSP for it. In the graph, the four lowest ratio values are: 2.0949, 2.6285, 2.6380 and 2.7129 (in increasing order) and they correspond to species *Ornithorhynchus anatinus* (Common name: Duck-billed platypus, MLSP: 22.6), *Oryctolagus cuniculus* (Common name: Old World rabbit, MLSP: 9), *Sarcophilus harrisii* (Common name: Tasmanian devil, MLSP: 13) and *Bos taurus* (Common name: Domestic cattle, MLSP: 20) respectively. In the same figure, the four highest ratio values are: 3.3928, 3.3802, 3.3223 and 3.2604 (in decreasing order) and they correspond to species *Latimeria chalumnae*

(Common name: Coelacanth, MLSP: 48), *Gallus gallus* (Common name: Red junglefowl (chicken), MLSP: 30), *Meleagris gallopavo* (Common name: Wild turkey, MLSP: 13) and *Macaca mulatta* (Common name: Rhesus monkey, MLSP: 40) respectively.

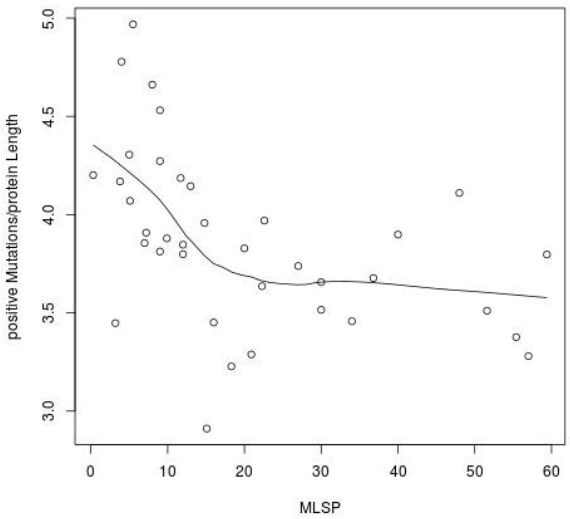
In the values above, we can see that the Duck-billed platypus, MLSP: 22.6 does not conform to the general ratio and MLSP trend of the graph. It has a high MLSP despite low protein stability. Likewise in the higher protein stability values, Wild turkey, MLSP: 13 does not conform to the ratio values of protein ND5 graph since it has a low MLSP despite a good protein stability. Other species which have comparable protein stability (to Wild turkey) are: Coelacanth, MLSP: 48, Rhesus monkey, MLSP: 40 and Red junglefowl (chicken), MLSP: 30. They all have considerably higher MLSPs.

In 5.5b we see that there is a strong positive correlation between ratio and MLSP for protein NDUFB6. Meaning that the greater the stability of NDUFB6 for an organism, the higher the MLSP for it. In the graph, the four lowest ratio values are: 2.8595, 3.5984, 3.9375 and 3.9921 (in increasing order) and they correspond to species *Ornithorhynchus anatinus* (Common name: Duck-billed platypus, MLSP: 22.6), *Anolis carolinensis* (Common name: Green anole, MLSP: 7.2), *Danio rerio* (Common name: Zebrafish, MLSP: 5.5) and *Oryzias latipes* (Common name: Japanese medaka, MLSP: 5) respectively. In the same figure, the four highest ratio values are: 6.0709, 5.9766, 5.9609 and 5.9062 (in decreasing order) and they correspond to species *Tarsius syrichta* (Common name: Philippine tarsier, MLSP: 16), *Gorilla gorilla* (Common name: Gorilla, MLSP: 55.4), *Pan troglodytes* (Common name: Chimpanzee, MLSP: 59.4) and *Otolemur garnettii* (Common name: Small-eared galago, MLSP: 18.3) respectively.

In the values above, we can see that the lowest ratio value, the Duck-billed platypus with MLSP: 22.6 does not conform to the general trend. Like ND5, for NDUFB6 too the Duck-billed platypus has an exceptionally high MLSP despite lowest protein stability value. The highest ddG value, Philippine tarsier, MLSP: 16 and Small-eared galago, MLSP: 18.3 do not conform to the general trend for protein NDUFB6. These two species have quite low MLSPs in comparison to Gorilla, MLSP: 55.4 and Chimpanzee, MLSP: 59.4 which have comparable protein stability for NDUFB6.



(a) NDUFB5: ρ= -0.5, p-val= 0.00229



(b) NDUFA6: ρ= -0.59, p-val= 0.00013

Figure 5.6.: Relationship of ratio and MLSP with negative correlation

In 5.6a we see that there is a negative correlation between ratio and MLSP for protein NDUFB5. Meaning that the higher the stability of NDUFB5 for an organism, the lower the MLSP for it. In the graph, the four lowest ratio values are: 2.7989, 2.8254, 3.0370 and 3.0479 (in increasing order) and they correspond to species *Sus scrofa* (Common name: Wild boar, MLSP: 27), *Tursiops truncatus* (Common name: Bottlenosed dolphin, MLSP: 51.6), *Bos taurus* (Common name: Domestic cattle, MLSP: 20) and *Oryzias latipes* (Common name: Japanese medaka, MLSP: 5) respectively. In the same figure, the four highest ratio values are: 4.5323, 4.2103, 4.073 and 4.0332 (in decreasing order) and they correspond to species *Drosophila melanogaster* (Common name: Fruit fly, MLSP: 0.3), *Monodelphis domestica* (Common name: Shorttailed opossum, MLSP: 5.1),

Callithrix jacchus (Common name: White-tufted-ear marmoset, MLSP: 22.8) and *Danio rerio* (Common name: Zebrafish, MLSP: 5.5) respectively.

In the values above, Japanese medaka, MLSP: 5 shows considerable lower MLSP than its counterparts with similar protein stability value. It does not follow the trend of the graph; that lower protein stability corresponds to higher MLSP. In the four highest ratio values, White-tufted-ear marmoset, MLSP: 22.8 shows considerably higher MLSP for similar protein stability values as its counterparts: Fruit fly, MLSP: 0.3, Shorttailed opossum, MLSP: 5.1 and Zebrafish, MLSP: 5.5 which have low MLSP corresponding to high protein stability of NDUFB5.

In 5.6b we see that there is a strong negative correlation between ratio and MLSP for protein NDUFA6. Meaning that the higher the stability of NDUFA6 for an organism, the lower the MLSP for it. In the graph, the four lowest ratio values are: 2.9111, 3.2272, 3.2797 and 3.2886 (in increasing order) and they correspond to species *Macropus eugenii* (Common name: Tammar wallaby, MLSP: 15.1), *Otolemur garnettii* (Common name: Small-eared galago, MLSP: 18.3), *Equus caballus* (Common name: Horse, MLSP: 57) and *Pteropus vampyrus* (Common name: Large flying fox, MLSP: 20.9) respectively. In the same figure, the four highest ratio values are: 4.9688, 4.7786, 4.6615 and 4.5313 (in decreasing order) and they correspond to species *Danio rerio* (Common name: Zebrafish, MLSP: 5.5), *Mus musculus* (Common name: House mouse, MLSP: 4), *Gasterosteus aculeatus* (Common name: Alaskan stickleback, MLSP: 8) and *Oreochromis niloticus* (Common name: Nile tilapia, MLSP: 9) respectively.

In the values above, Tammar wallaby, MLSP: 15.1 and Small-eared galago, MLSP: 18.3 do not follow the trend of the graph; that lower protein stability corresponds to higher MLSP and higher protein stability corresponds to lower MLSP for NDUFA6.

6. Discussions

There were three methods applied: Mean mean ddG value of protein: where we take the mean ddG value of each site after all the possible 19 mutations and then for the whole protein, we take the mean value of all the ddG values per site.

Min mean ddG value of the protein: wherein we take the mean value of ddG per site (after all the possible 19 mutations) but then we take the min ddG value from all the site mean values to represent the ddG of the protein. Min value means we take the most de-stabilizing mean ddg value for each protein.

Ratio method: where we take the ratio of the total positive mutations for a protein and length of the protein. Sometimes the sign of the mutation is more important than the actual ddG value (Cheng et al., 2006),that is why this method of comparing the stability should also work because here we are counting the total positive mutations occurring in a protein (per site again there will be 19 mutations). So that the length of the protein does not affect our analysis, we take ratio of positive mutations and length.

Now we compare the stability of proteins and MLSP and the different stability results as well.

6.1. Compare stability of proteins and MLSP

Mean mean ddG value of protein vs MLSP: For the 45 proteins, we got significant ρ value for 11 proteins. Out of those, 5 had positive ρ value and 6 proteins had negative ρ value. Which means that for 5 proteins, the stability of protein increases with increase in MLSP. For 6 proteins, the stability decreases with increase in species MLSP. Whereas for 34 proteins, there is no significant relationship between stability and MLSP (as the p-value for those is > 0.05)

Min mean ddG value of protein vs MLSP: For the 45 proteins, we got significant ρ value for 10 proteins only. It means that for the remaining 35 proteins, there is no significant relationship between stability of proteins and MLSP of the species for which there are orthologs for that protein. For the 10 proteins with significant ρ value, for 4 proteins, there is a positive ρ value and for 6 proteins, there is a negative ρ value. It means that the stability of the protein increases with MLSP for 4 proteins and decreases for 6 proteins.

Ratio method of protein vs MLSP: For the 45 proteins, we got significant ρ value for 10 proteins. Out of these 10 proteins, there are 4 proteins with a positive ρ value and 6

proteins with a negative ρ value. That means for 4 proteins there is a positive correlation between protein stability and MLSP and for 6 proteins a negative correlation between protein stability and MLSP. Whereas for the remaining 35 proteins there is no significant correlation between protein stability and MLSP.

Relationship between stability and MLSP: The compiled result for stability calculation with the three different methods is given by table 6.1. First we will see all those proteins which have either a positive ρ value or a negative ρ value by all the three methods. They are: NDUFB6 (which has a positive ρ value by all three methods) and NDUFS3 (which has a negative ρ value by all the three methods). This is a bit low number than what I had expected. I was hoping that probably we will get same proteins by the three methods.

Proteins with significant stability and MLSP ρ			
	Mean mean	Min mean	Ratio
Positive ρ	ND2	ND2	ND5
	NDUFB4	NDUFB5	NDUFB6
	NDUFB6	NDUFB6	NDUFB10
	NDUFS8	NDUFA7	NDUFS8
	NDUFB10		
Negative ρ	NDUFS1	NDUFA13	NDUFV2
	NDUFS3	NDUFS3	NDUFS3
	NDUFS5	NDUFC1	NDUFS5
	NDUFA6	ND4L	NDUFA6
	NDUFB5	NDUFA4	NDUFB5
	NDUFC2	NDUFB11	NDUFC2

Table 6.1.: Proteins with a significant relationship between stability and MLSP

Since we do not have many proteins which are present in all the three methods, we check proteins which are present in atleast two methods. For positive ρ value we have two proteins which are common by mean mean method and min mean method: ND2 and NDUFB6, two proteins by mean mean and ratio methods: NDUFS8 and NDUFB10, one protein by min mean method and ratio method: NDUFB6. For negative ρ value we have just one protein by mean mean and min mean method: NDUFS3, one protein by min mean and ratio method: NDUFS3 and five proteins by mean mean method and ratio method: NDUFS3, NDUFS5, NDUFA6, NDUFB5 and NDUFC2. This is also shown by table 6.2.

Proteins with significant stability and MLSP ρ			
	Mean mean and Min mean	Mean mean and Ratio	Min mean and Ratio
Positive ρ	ND2 NDUFB6	NDUFS8 NDUFB10	NDUFB6
Negative ρ	NDUFS3	NDUFS3 NDUFS5 NDUFA6 NDUFB5 NDUFC2	NDUFS3

Table 6.2.: Proteins common in two methods

We also have one protein which had a positive ρ value by one method and a negative ρ value by another method. It is NDUFB5 which has a negative ρ value by mean mean method and ratio method and a positive ρ value by min mean method. For 45 proteins, there are 19 proteins which have a significant ρ value by any of the three methods (8 proteins with significant positive correlation, 12 with significant negative correlation and 1 protein which has a positive and negative correlation both). Of these 19 proteins, there is just one protein which has a positive ρ value by one method and a negative ρ value by another method. Proteins with a significant positive correlation: ND2, NDUFB4, NDUFB5, NDUFB6, NDUFB10, NDUFS8, NDUFA7 and ND5 and proteins with a significant negative correlation: NDUFB5, NDUFS1, NDUFS3, NDUFS5, NDUFA6, NDUFC2, NDUFA13, NDUFC1, ND4L, NDUFA4, NDUFB11 and NDUFV2. All these 19 proteins are important in the process of ageing as per our analysis, since we got a significant stability and MLSP correlation for all 19 of them.

6.2. Relation between stability of "core" proteins and MLSP

The core proteins are: NDUFS1, NDUFS2, NDUFS3, NDUFS8, NDUFS7, NDUFV1, NDUFV2, ND1, ND2, ND3, ND4, ND4L, ND5 and ND6. The core proteins are the ones which are catalytically active in the complex. We will see how many out of these proteins show a significant relationship between stability and MLSP and what sort of relation there is. First we will look at the above proteins with different stability methods and then a generalized view:

Mean mean ddG value of "Core" proteins vs MLSP: For NDUFS1, NDUFS3, NDUFS8 and ND2 there is a significant correlation between MLSP and protein stability. Thus for 14 "core" proteins, there are 4 proteins with significant relation. Out of these 4 proteins, NDUFS1 (-0.44; 0.01348) and NDUFS3 (-0.35; 0.04199) have a negative correlation and NDUFS8 (0.45; 0.00716) and ND2 (0.62; 0.00177) have a positive correlation.

Min mean ddG value of "Core" proteins vs MLSP: For NDUFS3, ND2 and ND4L there is a significant correlation between MLSP and protein stability. Thus for 14 "core"

proteins, there are 3 proteins with significant relation. Out of these 3 proteins, stability of NDUFS3 (-0.44; 0.00755) and ND4L (-0.41; 0.04619) has a negative correlation with MLSP and stability of ND2 (0.62; 0.0015) has a positive correlation with MLSP.

Ratio of positive mutations and protein length for "core" proteins vs MLSP: For NDUFS3, NDUFV2, NDUFS8 and ND5 there is a significant correlation between MLSP and protein stability. Thus for 14 "core" proteins, there are 4 proteins with a significant relation between MLSP and stability. Out of these 4 proteins, stability of NDUFS3 (-0.43; 0.00921) and NDUFV2 (-0.39; 0.0126) has a negative correlation with MLSP and stability of NDUFS8 (0.47; 0.00496) and ND5 (0.55; 0.00606), has a positive correlation with MLSP.

Consolidated analysis: All the proteins which have a significant relation between stability and MLSP by even one method are as follows: NDUFS1, NDUFS3, NDUFS8, ND2, ND4L, NDUFV2 and ND5. NDUFS3 has a negative ρ value by all the three methods. ND2 has a positive ρ value by mean mean method and min mean method. NDUFS8 has a positive ρ value by mean mean method and by ratio method. For "core" proteins there are no proteins with positive ρ value by all three methods. Also there are no proteins except NDUFS3 which have a negative ρ value by more than one method. For "core" proteins, there are also no proteins which have a positive ρ value by one method and a negative ρ value by another method (that is no contradictory results by the different methods for "core" proteins). Out of the 14 core proteins, 7 are involved in the process of ageing as per our analysis. In fact ND2 and ND5 have a high significant positive correlation for stability and MLSP. Further, out of the 19 candidates which are indicated to be involved in the process of ageing by our analysis, 7 are "core" proteins that is they have a catalytic function which is quite a high number.

6.3. Stability difference between proteins encoded by the mitochondria and nuclear encoded proteins

There are total 7 mitochondrial proteins and 38 nuclear proteins.

Mean mean method: In mean mean method, there are 11 proteins with a significant relationship between MLSP and protein stability. Out of these 11, 1 protein is encoded by mtDNA (ND2). So that is 1 protein out of 7 and for nuclear encoded proteins, 10 out of 38. ND2 has positive ρ value.

Min mean method: In min mean method, there are 10 proteins with a significant relationship between MLSP and protein stability. Out of these 10, 2 proteins are encoded by mitochondria (ND2 and ND4L). So that is 2 proteins out of 7 and for nuclear encoded proteins, 8 out of 38. ND2 has positive ρ value and ND4L has negative ρ value.

Ratio method: In ratio method, there are 10 proteins with a significant relationship between MLSP and protein stability. Out of these 10, 1 protein is encoded by mitochondria (ND5). So that is 1 protein out of 7 and for nuclear encoded proteins, 9 out of 38.

ND2 has positive ρ value.

If we take all proteins for which we got a significant ρ value, among mitochondrially encoded proteins, we have: ND2, ND4L and ND5 (which is 3 out of 7). Out of these 3, 2 proteins have a positive MLSP and stability correlation. Remaining 16 are nuclear encoded proteins out of 38. Out of those 16, 6 proteins have a significant positive correlation of protein stability and MLSP where as 11 have a significant negative correlation of protein stability and MLSP. Out of the 19 statistically significant proteins of our analysis, only 3 are encoded by the mitochondria (mitochondria encodes a total of 7 proteins of complex I). Thus Mitochondrial proteins of complex I which are involved in the process of ageing by our analysis are: ND2, ND4L and ND5.

6.4. Future Research

The main aim of this study was to find proteins from Complex I which might have a role in the process of ageing. We have found those proteins however there is much scope for further research.

We have found 8 proteins which have a significant positive correlation between stability and MLSP. They are: ND2, ND5, NDUFB4, NDUFB5, NDUFB6, NDUFB10, NDUFS8 and NDUFA7. There are 12 proteins which show a significant negative correlation between MLSP and stability. They are: NDUFS1, NDUFS3, NDUFS5, NDUFA6, NDUFB5, NDUFC2, NDUFA13, NDUFC1, ND4L, NDUFA4, NDUFB11 and NDUFV2. Of the 8 proteins that demonstrate a significant positive correlation, ND2, ND5 and NDUFS8 have a catalytic function (in humans). Likewise for the 12 proteins with a significant negative correlation, NDUFS1, NDUFS3, ND4L and NDUFV2 have a catalytic function (in humans). They form part of what is known as the catalytic "core". The rest of the 5 out of 8 and 8 out of 12 proteins however have no catalytic function. Thus we can say that some catalytic proteins also demonstrate MLSP and stability correlation. However not all the proteins that demonstrate a correlation of stability with MLSP belong to the catalytic category. Further detailed study of functional information of all the statistically significant proteins could be quite significant in age related studies regarding Complex I. Thus a detailed study of the functions of these 19 proteins can be made. A comprehensive study could also me made regarding the involvement of these 19 significant proteins in various diseases we have discussed here.

Further analysing our statistically significant proteins based on their sub-cellular location:

- ND2: Mitochondrial membrane; Multi-pass
- NDUFB4: Mitochondrial inner membrane; Single pass
- NDUFB5: Mitochondrial inner membrane; Single pass
- NDUFB6: Mitochondrial inner membrane; Single pass
- NDUFB10: Peripheral
- ND5: Mitochondrial inner membrane; Multi-pass
- NDUFS1: Inner membrane (not mentioned whether single-pass or multi-pass)
- NDUFS5: Inner membrane; Single pass

- NDUFC2: Mitochondrial inner membrane: Single pass
- ND4L: Mitochondrial membrane; Multi-pass
- NDUFA4: Peripheral

Sub-cellular location of NDUFS8, NDUFA7, NDUFS3, NDUFA6, NDUFA13, NDUFC1, NDUFB11 and NDUFV2 is not available from information downloaded from entrez. As we can see from the above listing, most proteins which have a statistically significant relationship between stability and MLSP are in mitochondrial inner membrane. However sub-cellular location information of 8 out of 19 proteins is not readily available. A better analysis can be made if we explore more about the sub-cellular location of the 19 proteins.

The results of our study could have been more conclusive regarding targeting certain species for further age related studies if we had same organisms across all the 45 proteins. In our present analysis we have different species (for which orthologs were available) for each protein. If we had orthologs for all the species across all the 45 proteins, we could have made a good comparative analysis of anomalous behaviour of species. For example in the results chapter where we have discussed about some of the graphs, we observed that some species behave anomalously: like the duck-billed platypus which showed a high MLSP despite having the lowest stability among all organisms for ND2, ND5 and NDUFB6 proteins (though all these 3 proteins had a positive MLSP and stability correlation). This kind of analysis would have been easier if we had sequences for all the species. It would also have given us some species with an anomalous behaviour as subjects for further study.

7. Conclusion

The main aim of our study was to find proteins in Complex I which emulate the MLSP and oxidative damage correlation. We have found eight proteins which have a significant positive correlation between stability and MLSP. Of these eight proteins, NDUFB6 demonstrates a significant positive correlation by all the three methods that we had applied. ND2, NDUFB10 and NDUFS8 demonstrate a significant positive correlation by two methods. Thus these eight proteins and specially the four proteins (which demonstrate a significant correlation by more than one method) can be subjects of further analysis in studies related to the process of ageing because they show a significant positive correlation between stability and MLSP much like the mitochondrial free radical theory of ageing.

Apart from these proteins, there are twelve proteins which show a significant negative correlation between MLSP and stability. Out of these proteins NDUFS3 demonstrates a significant negative correlation by all the three methods applied and proteins NDUFS5, NDUFA6, NDUFB5 and NDUFC2 demonstrate a significant negative correlation by two methods. Even these twelve proteins and specially the five proteins (which demonstrate a significant correlation by more than one method) can be good candidates for further analysis as they show a significant correlation between MLSP and stability. Further, out of our nineteen statistically significant proteins, seven are involved in catalytic activity and three of the nineteen proteins are encoded by mtDNA.

The nineteen proteins, out of the forty five Complex I proteins, which demonstrate a significant metabolic activity and MLSP correlation are: ND2, ND5, NDUFB4, NDUFB5, NDUFB6, NDUFB10, NDUFS8, NDUFA7, NDUFS1, NDUFS3, NDUFS5, NDUFA6, NDUFC2, NDUFA13, NDUFC1, ND4L, NDUFA4, NDUFB11 and NDUFV2. Out of these proteins, NDUFS1, NDUFS3, NDUFS8, ND2, ND4L, NDUFV2 and ND5 proteins have a catalytic function and ND2, ND5 and ND4L are encoded by the mitochondria.

Bibliography

Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), 2014. URL http://omim.org/.

Lewis J et al. Alberts B, Johnson A. The respiratory chain and atp synthase. In *Molecular Biology of the Cell. 3rd edition.* Garland Science, New York, 1994. URL http://www.ncbi.nlm.nih.gov/books/NBK28380/.

Ethem Alpaydin. *Introduction to machine learning.* MIT press, 2004.

Robert L Baldwin. Energetics of protein folding. *Journal of molecular biology*, 371(2): 283–301, 2007.

Eduardo Balsa, Ricardo Marco, Ester Perales-Clemente, Radek Szklarczyk, Enrique Calvo, Manuel O. Landazuri, and Jose Antonio Enriquez. {NDUFA4} is a subunit of complex {IV} of the mammalian electron transport chain. *Cell Metabolism*, 16(3): 378 – 386, 2012. ISSN 1550-4131. doi: http://dx.doi.org/10.1016/j.cmet.2012.07.015. URL http://www.sciencedirect.com/science/article/pii/S1550413112002938.

Gustavo Barja. Mitochondrial free radical production and aging in mammals and birdsa. *Annals of the New York Academy of Sciences*, 854(1):224–238, 1998.

Gustavo Barja and Asuncion Herrero. Oxidative damage to mitochondrial dna is inversely related to maximum life span in the heart and brain of mammals. *The FASEB Journal*, 14(2):312–318, 2000.

Kenneth B. Beckman and Bruce N. Ames. The free radical theory of aging matures. *Physiological Reviews*, 78(2):547–581, 1998. URL http://physrev.physiology.org/content/78/2/547.

P Benit, A Slama, F Cartault, I Giurgea, D Chretien, S Lebon, C Marsac, A Munnich, A Rötig, and P Rustin. Mutant ndufs3 subunit of mitochondrial complex i causes leigh syndrome. *Journal of medical genetics*, 41(1):14–17, 2004.

Stryer L Berg JM, Tymoczko JL. *Biochemistry. 5th edition.* W H Freeman, New York, 2002a.

Stryer L Berg JM, Tymoczko JL. Section 18.3, the respiratory chain consists of four complexes: Three proton pumps and a physical link to the citric acid cycle. In *Biochemistry. 5th edition.* 2002b.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235. URL http://nar.oxfordjournals.org/content/28/1/235.abstract.

Ulrich Brandt. Energy converting nadh: Quinone oxidoreductase (complex i). *Annual Review of Biochemistry*, 75(1):69–92, 2006. doi: 10.1146/annurev.biochem.75.103004.142539. URL `http://www.annualreviews.org/doi/abs/10.1146/annurev.biochem.75.103004.142539`. PMID: 16756485.

Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

Emidio Capriotti, Piero Fariselli, and Rita Casadio. I-mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research*, 33(suppl 2):W306–W310, 2005. doi: 10.1093/nar/gki375.

Joe Carroll, Ian M. Fearnley, J. Mark Skehel, Richard J. Shannon, Judy Hirst, and John E. Walker. Bovine complex i is a complex of 45 different subunits. *Journal of Biological Chemistry*, 281(43):32724–32727, 2006. doi: 10.1074/jbc.M607135200. URL `http://www.jbc.org/content/281/43/32724.abstract`.

Jianlin Cheng, Arlo Randall, and Pierre Baldi. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Structure, Function, and Bioinformatics*, 62(4):1125–1132, 2006. ISSN 1097-0134. doi: 10.1002/prot.20810. URL `http://dx.doi.org/10.1002/prot.20810`.

Geoffrey Mallin Clarke and Dennis Cooke. *A basic course in statistics*. Arnold New York, 1998.

Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

Geoffrey M. Cooper. Mitochondria. In *The Cell: A Molecular Approach. 2nd edition.* 2000.

Thomas E Creighton. Protein folding. *Biochemical journal*, 270(1):1, 1990.

HH Dahl. Getting to the nucleus of mitochondrial disorders: identification of respiratory chain-enzyme genes causing leigh syndrome. *American journal of human genetics*, 63 (6):1594, 1998.

DD De Vries, LN Went, GW Bruyn, HR Scholte, RM Hofstra, PA Bolhuis, and BA Van Oost. Genetic and biochemical impairment of mitochondrial complex i activity in a family with leber hereditary optic neuropathy and hereditary spastic dystonia. *American journal of human genetics*, 58(4):703–711, 1996.

Stefan Dröse, Stephanie Krack, Lucie Sokolova, Klaus Zwicker, Hans-Dieter Barth, Nina Morgner, Heinrich Heide, Mirco Steger, Esther Nübel, Volker Zickermann, et al. Functional dissection of the proton pumping modules of mitochondrial complex i. *PLoS biology*, 9(8):e1001128, 2011.

EMBL-EBI. Mitochondrial respiratory chain complex, 2014. URL `http://www.genenames.org/genefamilies/mitocomplex`.

Elisa Fassone and Shamima Rahman. Complex i deficiency: clinical features, biochemistry and molecular genetics. *Journal of medical genetics*, 49(9):578–590, 2012.

Toren Finkel and Nikki J Holbrook. Oxidants, oxidative stress and the biology of ageing. *Nature*, 408(6809):239–247, 2000.

Fvasconcellos. Mitochondrial electron transport chain, 2007. URL `\url{http://en.wikipedia.org/wiki/File:Mitochondrial_electron_transport_chain%E2%80%94Etc4.svg}`.

Nikolaus Grigorieff. Structure of the respiratory nadh:ubiquinone oxidoreductase (complex i). *Current Opinion in Structural Biology*, 9(4):476 – 484, 1999. ISSN 0959-440X. doi: http://dx.doi.org/10.1016/S0959-440X(99)80067-0. URL `http://www.sciencedirect.com/science/article/pii/S0959440X99800670`.

M. Michael Gromiha, Jianghong An, Hidetoshi Kono, Motohisa Oobatake, Hatsuho Uedaira, P. Prabakaran, and Akinori Sarai. Protherm, version 2.0: thermodynamic database for proteins and mutants. *Nucleic Acids Research*, 28(1):283–285, 2000. doi: 10.1093/nar/28.1.283. URL `http://nar.oxfordjournals.org/content/28/1/283.abstract`.

Sadie L Hebert, Ian R Lanza, and K Sreekumaran Nair. Mitochondrial dna alterations and reduced mitochondrial function in aging. *Mechanisms of ageing and development*, 131(7):451–462, 2010.

Olof Heden, J Lehmann, E Năstase, and P Sissokho. The supertail of a subspace partition. *Designs, codes and cryptography*, 69(3):305–316, 2013.

Michio Hirano, Enzo Ricci, M. Richard Koenigsberger, Richard Defendini, Steven G. Pavlakis, Darryl C. DeVivo, Salvatore DiMauro, and Lewis P. Rowland. Melas: An original case and clinical criteria for diagnosis. *Neuromuscular Disorders*, 2(2):125 – 135, 1992. ISSN 0960-8966. doi: http://dx.doi.org/10.1016/0960-8966(92)90045-8. URL `http://www.sciencedirect.com/science/article/pii/0960896692900458`.

Judy Hirst, Joe Carroll, Ian M. Fearnley, Richard J. Shannon, and John E. Walker. The nuclear encoded subunits of complex i from bovine heart mitochondria. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1604(3):135 – 150, 2003. ISSN 0005-2728. doi: http://dx.doi.org/10.1016/S0005-2728(03)00059-8. URL `http://www.sciencedirect.com/science/article/pii/S0005272803000598`.

Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric statistical methods*, volume 751. John Wiley & Sons, 2013.

Kurt Hornik. R FAQ, 2014. URL `http://CRAN.R-project.org/doc/FAQ/R-FAQ.html`.

Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification, 2003.

Albert S Jun, Michael D Brown, and Douglas C Wallace. A mitochondrial dna mutation at nucleotide pair 14459 of the nadh dehydrogenase subunit 6 gene associated with maternally inherited leber hereditary optic neuropathy and dystonia. *Proceedings of the National Academy of Sciences*, 91(13):6206–6210, 1994.

Kelvinsong. Mitochondrion structure, 2013. URL \url{http://commons.wikimedia.org/wiki/File:Mitochondrion_structure.svg}.

Alfried Kohlschütter and Florian Eichler. Childhood leukodystrophies: a clinical perspective. 2011.

Hung-Hai Ku and R.S. Sohal. Comparison of mitochondrial pro-oxidant generation and anti-oxidant defenses between rat and pigeon: possible basis of variation in longevity and metabolic potential. *Mechanisms of Ageing and Development*, 72(1):67 – 76, 1993. ISSN 0047-6374. doi: http://dx.doi.org/10.1016/0047-6374(93)90132-B. URL http://www.sciencedirect.com/science/article/pii/004763749390132B.

Hung-Hai Ku, Ulf T Brunk, and Rajindar S Sohal. Relationship between mitochondrial superoxide and hydrogen peroxide production and longevity of mammalian species. *Free Radical Biology and Medicine*, 15(6):621–627, 1993.

Giorgio Lenaz, Romana Fato, Maria Luisa Genova, Christian Bergamini, Cristina Bianchi, and Annalisa Biondi. Mitochondrial complex i: Structural and functional aspects. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1757(9âĂŞ10):1406 – 1420, 2006. ISSN 0005-2728. doi: http://dx.doi.org/10.1016/j.bbabio.2006.05.007. URL http://www.sciencedirect.com/science/article/pii/S0005272806001319. Mitochondria: from Molecular Insight to Physiology and Pathology.

AnthonyW. Linnane, Takayuki Ozawa, Sangkot Marzuki, and Masashi Tanaka. Mitochondrial dna mutations as an important contributor to ageing and degenerative diseases. *The Lancet*, 333(8639):642 – 645, 1989. ISSN 0140-6736. doi: http://dx.doi.org/10.1016/S0140-6736(89)92145-4. URL http://www.sciencedirect.com/science/article/pii/S0140673689921454. Originally published as Volume 1, Issue 8639.

JLCM Loeffen, JAM Smeitink, JMF Trijbels, AJM Janssen, RH Triepels, RCA Sengers, and LP Van den Heuvel. Isolated complex i deficiency in children: clinical, biochemical and genetic aspects. *Human mutation*, 15(2):123–134, 2000.

Donna Maglott, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 33(suppl 1):D54–D58, 2005.

Johannes S Maritz. *Distribution-free statistical methods*, volume 17. CRC Press, 1995.

AAM Morris, JV Leonard, GK Brown, SK Bidouki, LA Bindoff, DM Turnbull, CE Woodward, AE Harding, BD Lake, BN Harding, et al. Deficiency of respiratory chain complex i is a common cause of leigh disease. *Annals of neurology*, 40(1):25–30, 1996.

Frederick Mosteller, Stephen E Fienberg, and Robert EK Rourke. *Beginning statistics with data analysis*. Allison Wesley, 1983.

James Murray, Bing Zhang, Steven W Taylor, Devin Oglesbee, Eoin Fahy, Michael F Marusich, Soumitra S Ghosh, and Roderick A Capaldi. The subunit composition of the human nadh dehydrogenase obtained by rapid one-step immunopurification. *Journal of Biological Chemistry*, 278(16):13619–13622, 2003.

David L Nelson and Michael M Cox. *Lehninger Principles of Biochemistry, Fourth Edition*. Freeman, fourth edition edition, 2004.

Andrew Ng. Stanford machine learning. Coursera Lecture Notes, 2012.

Alexandrov Oleg. Monotonicity example2, 2007a. URL \url{http://commons.wikimedia.org/wiki/File:Monotonicity_example2.png}.

Alexandrov Oleg. Monotonicity example3, 2007b. URL \url{http://commons.wikimedia.org/wiki/File:Monotonicity_example3.png}.

Gregory A Petsko and Dagmar Ringe. *Protein structure and function*. New Science Press, 2004.

S Rahman, RB Blok, H-HM Dahl, DM Danks, DM Kirby, CW Chow, J Christodoulou, and DR Thorburn. Leigh syndrome: clinical features and biochemical and dna abnormalities. *Annals of neurology*, 39(3):343–351, 1996.

Leonard Richardson. Beautiful soup documentation, 2007.

Paul Riordan-Eva and Anita E Harding. Leber's hereditary optic neuropathy: the clinical relevance of different mitochondrial dna mutations. *Journal of medical genetics*, 32(2): 81, 1995.

Kristin M Santa. Treatment options for mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke-like episodes (melas) syndrome. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 30(11):1179–1196, 2010.

Brian E. Schultz and Sunney I. Chan. Structures and proton-pumping strategies of mitochondrial respiratory enzymes. *Annual Review of Biophysics and Biomolecular Structure*, 30(1):23–65, 2001. doi: 10.1146/annurev.biophys.30.1.23. URL http://www.annualreviews.org/doi/abs/10.1146/annurev.biophys.30.1.23. PMID: 11340051.

Inna Shokolenko, Natalia Venediktova, Alexandra Bochkareva, Glenn L. Wilson, and Mikhail F. Alexeyev. Oxidative stress induces degradation of mitochondrial dna. *Nucleic Acids Research*, 37(8):2539–2548, 2009. doi: 10.1093/nar/gkp100. URL http://nar.oxfordjournals.org/content/37/8/2539.abstract.

R.S. Sohal, I. Svensson, and U.T. Brunk. Hydrogen peroxide production by liver mitochondria in different species. *Mechanisms of Ageing and Development*, 53(3):209 – 215, 1990. ISSN 0047-6374.

Douglas M. Sproule and Petra Kaufmann. Mitochondrial encephalopathy, lactic acidosis, and strokelike episodes. *Annals of the New York Academy of Sciences*, 1142(1):133–158, 2008. ISSN 1749-6632. doi: 10.1196/annals.1444.011. URL http://dx.doi.org/10.1196/annals.1444.011.

Jason E Stajich, David Block, Kris Boulez, Steven E Brenner, Stephen A Chervitz, Chris Dagdigian, Georg Fuellen, James GR Gilbert, Ian Korf, Hilmar Lapp, et al. The bioperl toolkit: Perl modules for the life sciences. *Genome research*, 12(10):1611–1618, 2002.

Richard P Stanley. An introduction to hyperplane arrangements. In *Lecture notes, IAS/Park City Mathematics Institute*. Citeseer, 2004.

Willi-Hans Steeb. *The nonlinear workbook: chaos, fractals, cellular automata, neural networks, genetic algorithms, gene expression programming, support vector machine, wavelets, hidden Markov models, fuzzy logic with C++, Java and SymbolicC++ programs*. World Scientific, 2011.

Robi Tacutu, Thomas Craig, Arie Budovsky, Daniel Wuttke, Gilad Lehmann, Dmitri Taranukha, Joana Costa, Vadim E Fraifeld, and João Pedro de Magalhães. Human ageing genomic resources: integrated databases and tools for the biology and genetics of ageing. *Nucleic acids research*, 41(D1):D1027–D1033, 2013.

Julio F Turrens. Mitochondrial formation of reactive oxygen species. *The Journal of Physiology*, 552(2):335–344, 2003. doi: 10.1113/jphysiol.2003.049478. URL `http://jp.physoc.org/content/552/2/335.abstract`.

Hanns Weiss, Thorsten Friedrich, GÃűtz Hofhaus, and Dagmar Preis. The respiratory-chain nadh dehydrogenase (complex i) of mitochondria. In P. Christen and E. Hofmann, editors, *EJB Reviews 1991*, volume 1991 of *EJB Reviews 1991*, pages 55–68. Springer Berlin Heidelberg, 1992. ISBN 978-3-540-55012-9. doi: 10.1007/978-3-642-77200-9_5. URL `http://dx.doi.org/10.1007/978-3-642-77200-9_5`.

Eric W. Weisstein. Monotonic function." from mathworld–a wolfram web resource., 2011. URL `\url{http://mathworld.wolfram.com/MonotonicFunction.html}`.

Mutsuo Yamaguchi, Grigory I Belogrudov, and Youssef Hatefi. Mitochondrial nadh-ubiquinone oxidoreductase (complex i): Effect of substrates on the fragmentation of subunits by trypsin. *Journal of Biological Chemistry*, 273(14):8094–8098, 1998. doi: 10.1074/jbc.273.14.8094. URL `http://www.jbc.org/content/273/14/8094.abstract`.

Patrick Yu-Wai-Man, Philip G Griffiths, Gavin Hudson, and Patrick F Chinnery. Inherited mitochondrial optic neuropathies. *Journal of medical genetics*, 46(3):145–158, 2009.

A. Appendix

Species	Scientific Name	MLSP in years
Roundworm	*Caenorhabditis elegans*	0.16
Fruit fly	*Drosophila melanogaster*	0.3
Eurasian shrew	*Sorex araneus*	3.2
Norway rat	*Rattus norvegicus*	3.8
House mouse	*Mus musculus*	4
Japanese medaka	*Oryzias latipes*	5
Shorttailed opossum	*Monodelphis domestica*	5.1
Zebra danio or zebrafish	*Danio rerio*	5.5
North American pika	*Ochotona princeps*	7
Green anole	*Anolis carolinensis*	7.2
Alaskan stickleback	*Gasterosteus aculeatus*	8
Sea lamprey	*Petromyzon marinus*	9
Nile tilapia	*Oreochromis niloticus*	9
Old World rabbit	*Oryctolagus cuniculus*	9
Ord's kangaroo rat	*Dipodomys ordii*	9.9
Northern tree shrew	*Tupaia belangeri*	11.1
Western European hedgehog	*Erinaceus europaeus*	11.7
Guinea pig	*Cavia porcellus*	12
Zebra finch	*Taeniopygia guttata*	12
Tasmanian devil	*Sarcophilus harrisii*	13
Wild turkey	*Meleagris gallopavo*	13

Table A.1.: Maximum Lifespan of species

Species	Scientific Name	MLSP in years
Tammar wallaby	*Macropus eugenii*	15.1
Philippine tarsier	*Tarsius syrichta*	16
Gray mouse lemur	*Microcebus murinus*	18.2
Small-eared galago	*Otolemur garnettii*	18.3
Lesser hedgehog tenrec	*Echinops telfairi*	19
Domestic cattle	*Bos taurus*	20
Large flying fox	*Pteropus vampyrus*	20.9
Common long-nosed armadillo	*Dasypus novemcinctus*	22.3
Duck-billed platypus	*Ornithorhynchus anatinus*	22.6
White-tufted-ear marmoset	*Callithrix jacchus*	22.8
Wild boar	*Sus scrofa*	27
Domestic cat	*Felis catus*	30
Red junglefowl (chicken)	*Gallus gallus*	30
Little brown bat	*Myotis lucifugus*	34
Giant panda	*Ailuropoda melanoleuca*	36.8
Rhesus monkey	*Macaca mulatta*	40
Hoffmann's two-toed sloth	*Choloepus hoffmanni*	41
Coelacanth	*Latimeria chalumnae*	48
Bottlenosed dolphin	*Tursiops truncatus*	51.6
Gorilla	*Gorilla gorilla*	55.4
Horse	*Equus caballus*	57
Chimpanzee	*Pan troglodytes*	59.4
African elephant	*Loxodonta africana*	65
Human	*Homo sapiens*	122.5

Table A.2.: Maximum Lifespan of species

Approved Symbol	Location
MT-ND1	mitochondria
MT-ND2	mitochondria
MT-ND3	mitochondria
MT-ND4	mitochondria
MT-ND4L	mitochondria
MT-ND5	mitochondria
MT-ND6	mitochondria
NDUFA1	Xq24
NDUFA2	5q31.2
NDUFA3	19q13.42
NDUFA4	7p21.3
NDUFA5	7q31.33
NDUFA6	22q13.2
NDUFA7	19p13.2
NDUFA8	9q33.2
NDUFA9	12p13.3
NDUFA10	2q37.3
NDUFA11	19p13.3
NDUFA12	12q22
NDUFA13	19p13.11
NDUFAB1	16p12.3
NDUFB1	14q31.3
NDUFB2	7q34
NDUFB3	2q33.1
NDUFB4	3q13.33
NDUFB5	3q27.1
NDUFB6	9p13.2
NDUFB7	19p13.12
NDUFB8	10q24.31
NDUFB9	8q24.13
NDUFB10	16p13.3
NDUFB11	Xp11.3
NDUFC1	4q31.1
NDUFC2	11q14.1
NDUFS1	2q33-q34
NDUFS2	1q23.3
NDUFS3	11p11.11
NDUFS4	5q11.1
NDUFS5	1p34.2-p33
NDUFS6	5p15.33
NDUFS7	19p13
NDUFS8	11q13.2
NDUFV1	11q13
NDUFV2	18p11.22
NDUFV3	21q22.3

Table A.3.: Complex I genes and their genomic locations. Source: (EMBL-EBI, 2014)

Number	Gene Name	ρ	p-val
1	NDUFV1	-0.19	0.26969
2	NDUFV2	-0.23	0.15807
3	NDUFV3	0.01	0.97506
4	NDUFS1	-0.44	0.01348
5	NDUFS2	0.06	0.77122
6	NDUFS3	-0.35	0.04199
7	NDUFS4	0.09	0.58194
8	NDUFS6	0.37	0.05524
9	NDUFS7	-0.32	0.12507
10	NDUFS8	0.45	0.00716
11	NDUFA2	-0.23	0.16541
12	NDUFA5	-0.13	0.50665
13	NDUFA7	0.33	0.05195
14	NDUFA12	0	0.98973
15	NDUFA13	0.02	0.90309
16	NDUFS5	-0.37	0.02835
17	NDUFA1	-0.29	0.08147
18	NDUFA3	0.1	0.61198
19	NDUFA6	-0.58	0.00017
20	NDUFA8	0.16	0.36692
21	NDUFA9	-0.07	0.68399
22	NDUFA10	0.11	0.54629
23	NDUFA11	-0.11	0.60857
24	ND6	0.07	0.74342
25	NDUFC1	0.01	0.95746
26	ND1	0.28	0.22448
27	ND2	0.62	0.00177
28	ND3	0.21	0.3417
29	ND4L	-0.28	0.18767
30	NDUFA4	-0.16	0.36545
31	NDUFAB1	-0.09	0.60363
32	NDUFB4	0.34	0.04872
33	NDUFB1	-0.1	0.66821
34	NDUFB2	0.19	0.26325
35	NDUFB3	0.32	0.05459
36	NDUFB5	-0.46	0.00569
37	NDUFB6	0.52	0.00172
38	NDUFB7	-0.01	0.95707
39	NDUFB8	-0.03	0.86997
40	NDUFB9	0.08	0.6615
41	NDUFB10	0.45	0.00543
42	NDUFB11	-0.24	0.24419
43	NDUFC2	-0.55	0.00328
44	ND4	0.42	0.05948
45	ND5	0.39	0.06244

Table A.4.: MLSP vs Mean of mean ddG value per site mutation

Number	Gene Name	ρ	p-val
1	NDUFV1	-0.3	0.0738
2	NDUFV2	0.09	0.58147
3	NDUFV3	-0.16	0.43431
4	NDUFS1	-0.21	0.24973
5	NDUFS2	-0.08	0.66985
6	NDUFS3	-0.44	0.00755
7	NDUFS4	0.27	0.11658
8	NDUFS6	0.13	0.51406
9	NDUFS7	-0.14	0.52387
10	NDUFS8	-0.28	0.10901
11	NDUFA2	0.04	0.79358
12	NDUFA5	0.29	0.13151
13	NDUFA7	0.42	0.01117
14	NDUFA12	-0.21	0.23795
15	NDUFA13	-0.51	0.00445
16	NDUFS5	-0.19	0.26824
17	NDUFA1	-0.05	0.79211
18	NDUFA3	-0.18	0.34194
19	NDUFA6	-0.13	0.45419
20	NDUFA8	0.27	0.12367
21	NDUFA9	-0.14	0.45271
22	NDUFA10	-0.21	0.25896
23	NDUFA11	0.01	0.95041
24	ND6	-0.21	0.3383
25	NDUFC1	-0.45	0.02428
26	ND1	0.15	0.5204
27	ND2	0.62	0.0015
28	ND3	0.1	0.63625
29	ND4L	-0.41	0.04619
30	NDUFA4	-0.46	0.00527
31	NDUFAB1	-0.21	0.24908
32	NDUFB4	0.26	0.13046
33	NDUFB1	-0.18	0.4119
34	NDUFB2	0.12	0.49161
35	NDUFB3	-0.06	0.74779
36	NDUFB5	0.36	0.03155
37	NDUFB6	0.39	0.02421
38	NDUFB7	0.28	0.1343
39	NDUFB8	0.19	0.26348
40	NDUFB9	-0.01	0.965
41	NDUFB10	0.28	0.09034
42	NDUFB11	-0.62	0.00076
43	NDUFC2	-0.15	0.47399
44	ND4	-0.05	0.81829
45	ND5	0.03	0.90433

Table A.5.: MLSP vs Min of mean ddG value per site of mutation

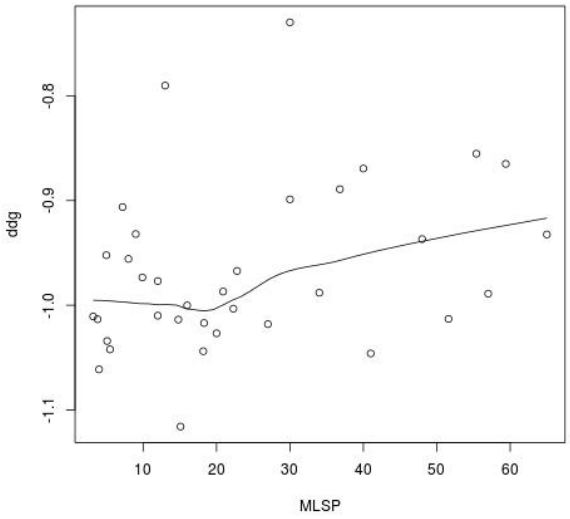
Number	Gene Name	ρ	p-val
1	NDUFV1	-0.15	0.39768
2	NDUFV2	-0.39	0.0126
3	NDUFV3	-0.23	0.26097
4	NDUFS1	-0.24	0.18717
5	NDUFS2	0.01	0.94325
6	NDUFS3	-0.43	0.00921
7	NDUFS4	0.11	0.52193
8	NDUFS6	0.07	0.72998
9	NDUFS7	-0.23	0.27846
10	NDUFS8	0.47	0.00496
11	NDUFA2	-0.24	0.15162
12	NDUFA5	-0.21	0.26497
13	NDUFA7	0.17	0.31104
14	NDUFA12	-0.05	0.79676
15	NDUFA13	0.16	0.39943
16	NDUFS5	-0.37	0.02665
17	NDUFA1	-0.25	0.13921
18	NDUFA3	0.19	0.33017
19	NDUFA6	-0.59	0.00013
20	NDUFA8	-0.02	0.92719
21	NDUFA9	0	0.99444
22	NDUFA10	0.18	0.33304
23	NDUFA11	-0.17	0.39403
24	ND6	0.24	0.27642
25	NDUFC1	-0.15	0.45984
26	ND1	0.31	0.16592
27	ND2	0.32	0.13356
28	ND3	0.21	0.34441
29	ND4L	-0.24	0.25744
30	NDUFA4	-0.06	0.73037
31	NDUFAB1	0.05	0.76461
32	NDUFB4	0.29	0.08964
33	NDUFB1	0.08	0.71669
34	NDUFB2	0.14	0.40369
35	NDUFB3	0.16	0.33729
36	NDUFB5	-0.5	0.00229
37	NDUFB6	0.55	0.00081
38	NDUFB7	-0.14	0.46803
39	NDUFB8	0	0.99389
40	NDUFB9	0.18	0.32841
41	NDUFB10	0.38	0.02112
42	NDUFB11	0.04	0.84633
43	NDUFC2	-0.52	0.0059
44	ND4	0.31	0.17153
45	ND5	0.55	0.00606

Table A.6.: MLSP vs Ratio of all positive mutations and protein length

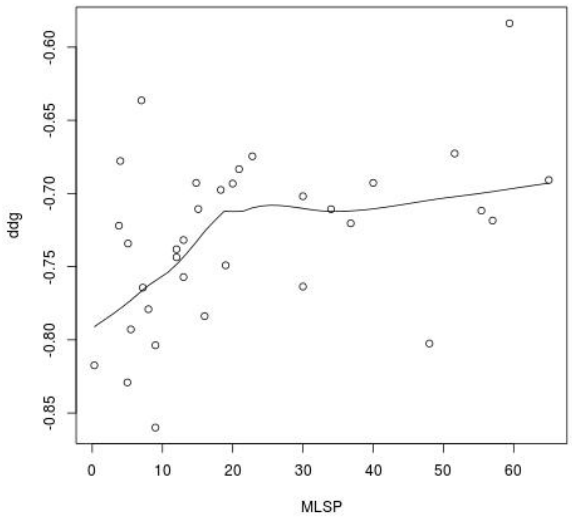
Genes	Sub-cellular location
NDUFV1	Mitochondrion inner membrane; Peripheral
NDUFV3	Mitochondrion inner membrane; Peripheral
NDUFS1	Mitochondrion inner membrane
NDUFS2	Mitochondrion inner membrane; Peripheral
NDUFS4	Mitochondrion inner membrane; Peripheral
NDUFS7	Mitochondrion
NDUFS8	Mitochondrion (Probable)
NDUFA2	Mitochondrion inner membrane; Peripheral
NDUFA12	Mitochondrion inner membrane; Peripheral
NDUFS5	Mitochondrion. Mitochondrion inner membrane;
NDUFA1	Mitochondrion inner membrane; Single-pass
NDUFA3	Mitochondrion inner membrane; Single-pass
NDUFA8	Mitochondrion. Mitochondrion intermembrane
NDUFA9	Mitochondrion matrix.
NDUFA10	Mitochondrion matrix
NDUFA11	Mitochondrion inner membrane; Multi-pass
MT-ND6	Mitochondrion membrane; Multi-pass membrane
MT-ND1	Mitochondrion membrane; Multi-pass membrane
MT-ND2	Mitochondrion membrane; Multi-pass membrane
MT-ND3	Mitochondrion membrane; Multi-pass membrane
MT-ND4L	Mitochondrion membrane; Multi-pass membrane
NDUFA4	Mitochondrion inner membrane; Peripheral
NDUFB4	Mitochondrion inner membrane; Single-pass
NDUFB2	Mitochondrion inner membrane; Peripheral
NDUFB3	Mitochondrion inner membrane; Single-pass
NDUFB5	Mitochondrion inner membrane; Single-pass
NDUFB6	Mitochondrion inner membrane; Single-pass
NDUFB9	Mitochondrion inner membrane; Peripheral
NDUFB10	Mitochondrion inner membrane; Peripheral
NDUFB11	Mitochondrion inner membrane; Single-pass
NDUFC2	Mitochondrion inner membrane; Single-pass
MT-ND4	Mitochondrion membrane; Multi-pass membrane
MT-ND5	Mitochondrion inner membrane; Multi-pass

Table A.7.: The sub-cellular location of the proteins

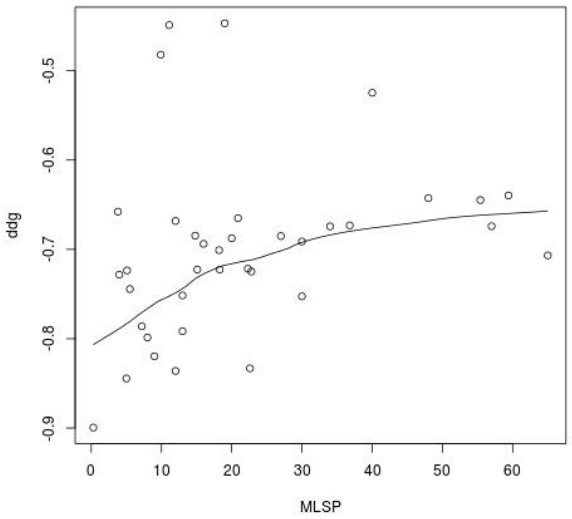
B. Appendix



(a) NDUFB4: ρ=0.34, p-val=0.04872

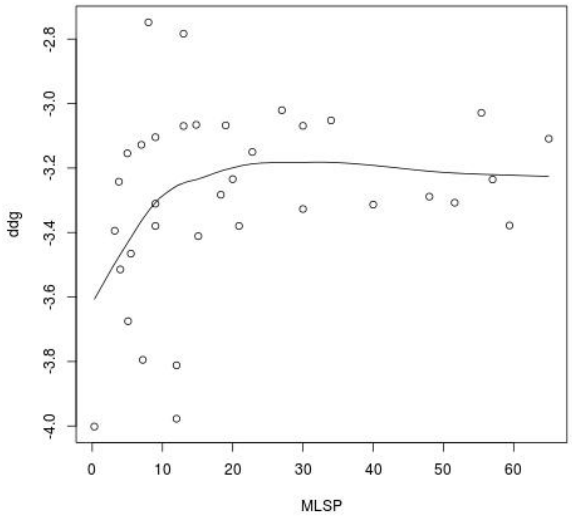


(b) NDUFS8: ρ=0.45, p-val=0.00716

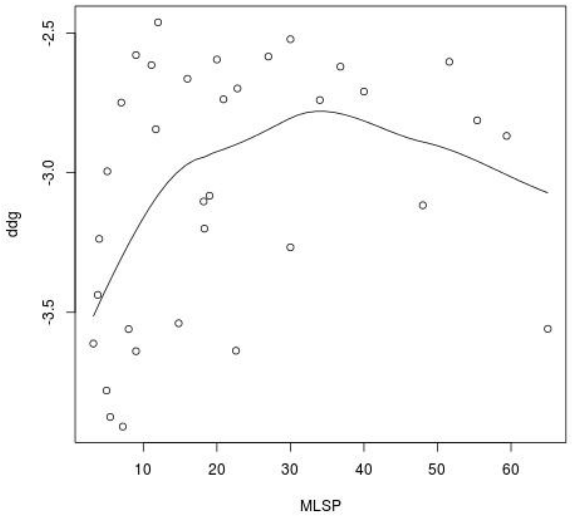


(c) NDUFB10: ρ=0.45, p-val=0.00543

Figure B.1.: Relationship of mean mean ddG and MLSP with positive correlation

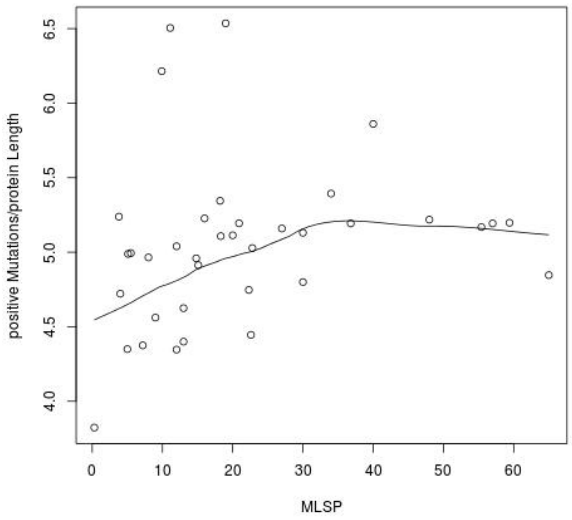


(a) NDUFB5: ρ=0.36, p-val=0.03155

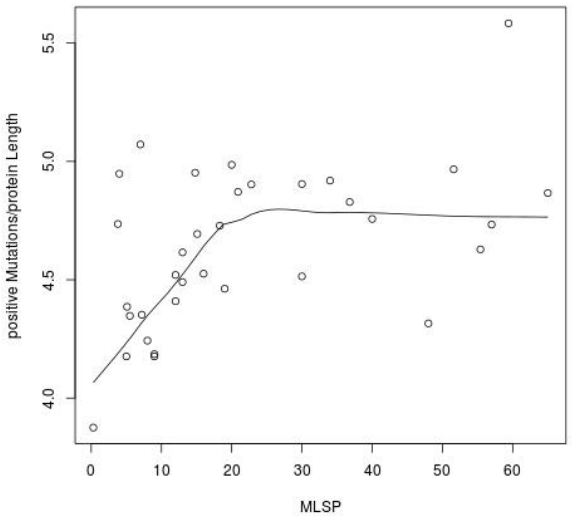


(b) NDUFB6: ρ=0.39, p-val=0.02421

Figure B.2.: Relationship of min mean ddG and MLSP with positive correlation

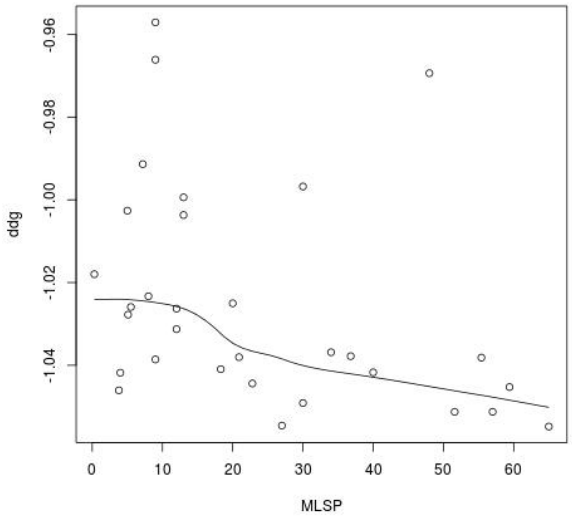


(a) NDUFB10: ρ=0.38, p-val= 0.02112

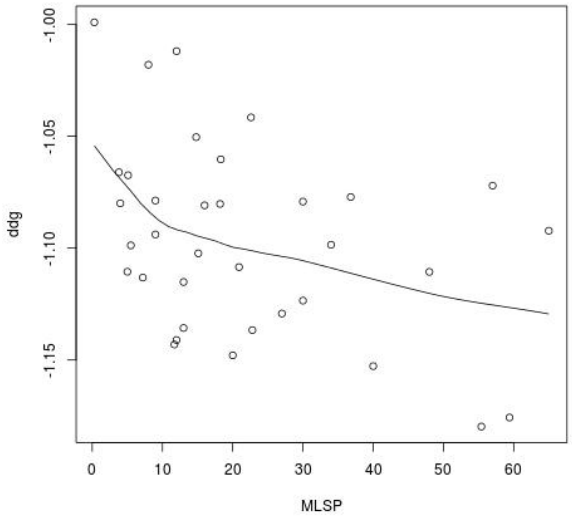


(b) NDUFS8: ρ= 0.47, p-val=0.00496

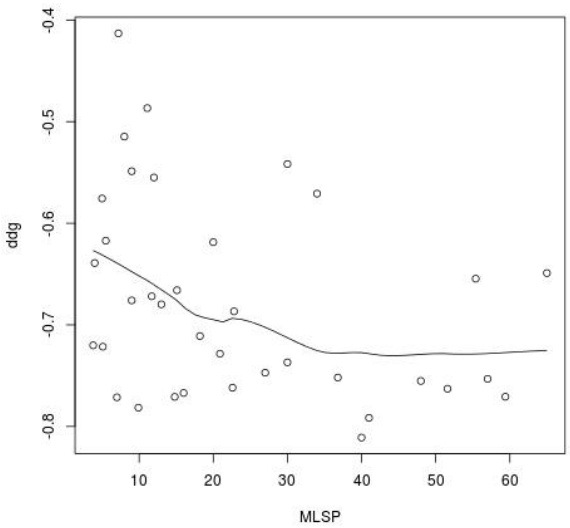
Figure B.3.: Relationship of ratio and MLSP with positive correlation



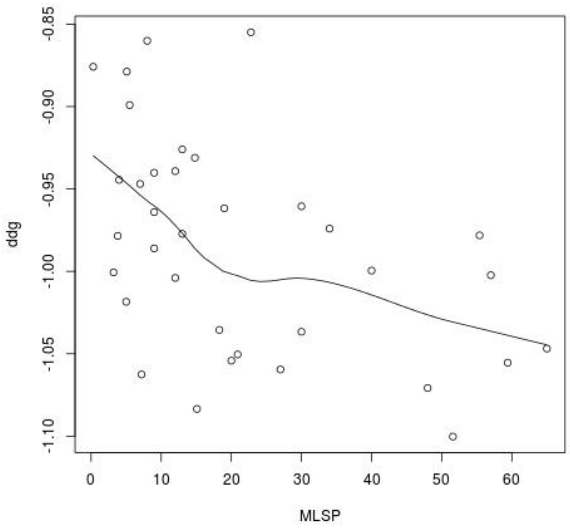
(a) NDUFS1: ρ=-0.44, p-val=0.01348



(b) NDUFS3: ρ=-0.35, p-val=0.04199

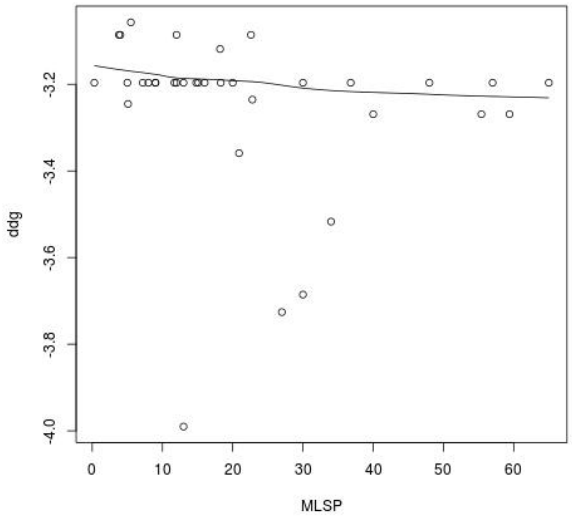


(c) NDUFS5: ρ=-0.37, p-val=0.02835

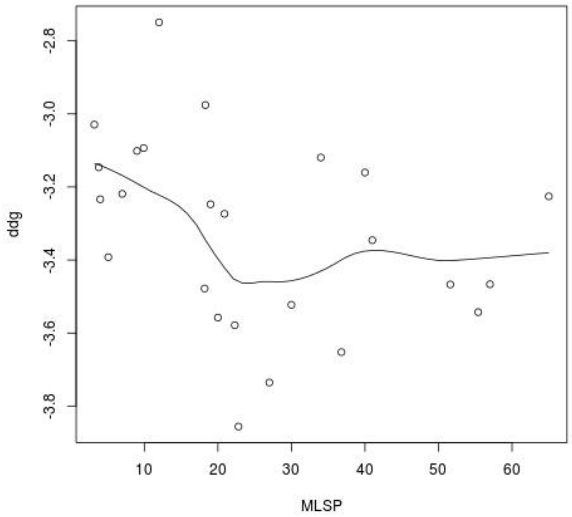


(d) NDUFB5: ρ=-0.46, p-val=0.00569

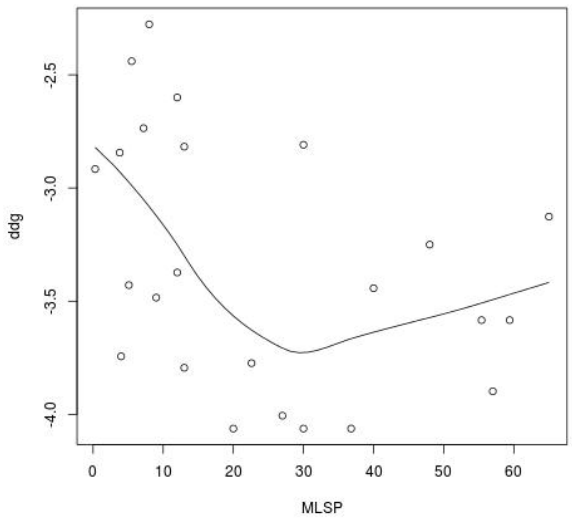
Figure B.4.: Relationship of mean mean ddG and MLSP with negative correlation



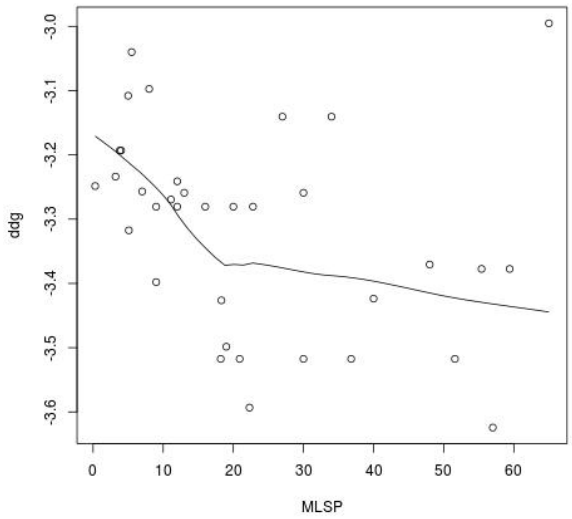
(a) NDUFS3: ρ=-0.44, p-val=0.00755



(b) NDUFC1: ρ=-0.45, p-val=0.02428

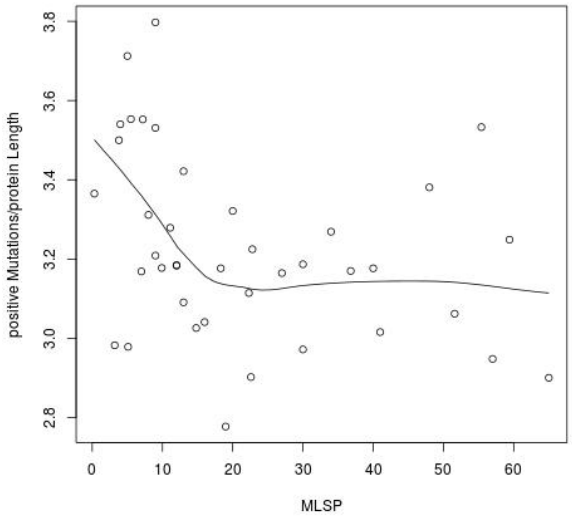


(c) ND4L: ρ=-0.41, p-val=0.04619

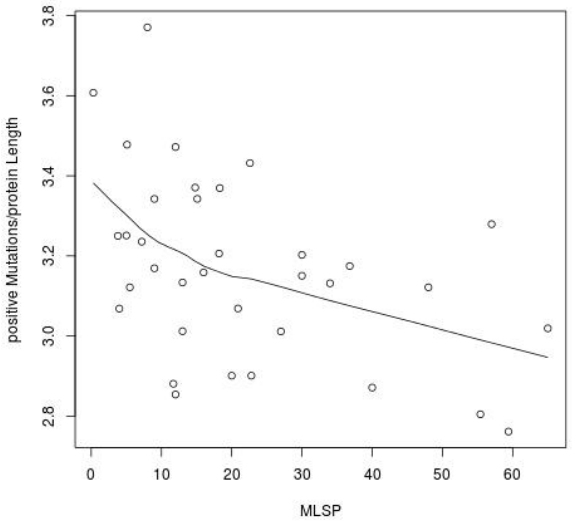


(d) NDUFA4: ρ=-0.46, p-val=0.00527

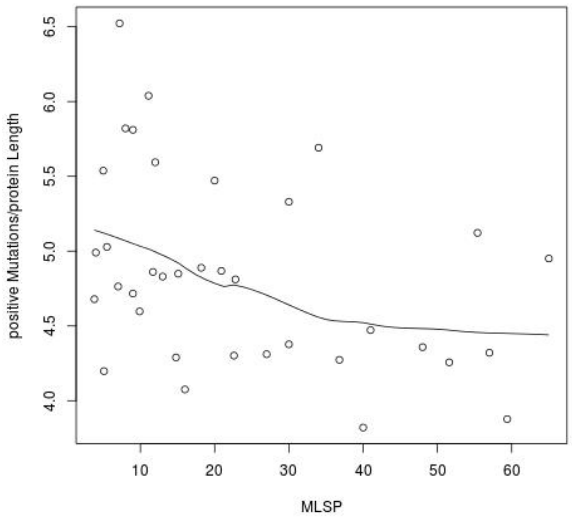
Figure B.5.: Relationship of min mean ddG and MLSP with negative correlation



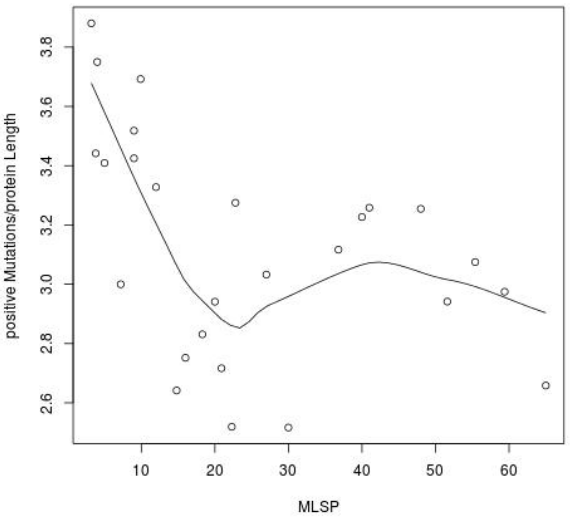
(a) NDUFV2: ρ= -0.39, p-val= 0.0126



(b) NDUFS3: ρ= -0.43, p-val= 0.00921



(c) NDUFS5: ρ= -0.37, p-val= 0.02665



(d) NDUFC2: ρ= -0.52, p-val= 0.0059

Figure B.6.: Relationship of ratio and MLSP with negative correlation