A Search for the Physical Basis of the Genetic Code and Modeling Cancer Cell Response to Chemotherapy Using the Ising Model

by

Sahar Arbabimoghadam

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Physics

University of Alberta

Abstract

Genetic code and its origin are one of the most challenging problems in biochemistry and cell biology. Studying the genetic code evolution and the logic behind it is an interesting but a very complicated problem. The logic of the genetic code from an energetic and probabilistic perspective, the occurrence frequency of protein mutations, and statistics of cytotoxicity effects on surviving cancer cell have been the main investigated topics in this thesis. The aim of this research is to implement the methods rooted in statistics, thermodynamics, and the physics of phase transitions in order to better face the challenges that experimental observations from genetics, molecular, and cell biology bring to the field of computational biophysics.

In this thesis, the first aim has been to find an underlying correlation between the Gibbs free energy and the naturally occurring frequency of codons and amino acids across extant life forms analyzed statistically. Using GAMESS software, the amino acid thermochemistry was estimated. For these calculations, we used the Hartee-Fock method with the PM3 basis sets. These energies were compared to the codon energies obtaining involving three energetic terms; nearest neighbor, stacking and nucleotide Gibbs free energy. The correlation between codon and amino acid energies could shed light on the rules behind the codon assignments in the genetic code. Unfortunately, only weak correlations were found in our study. Moreover, our investigation showed that, in human, amino acids that have a higher redundancy occur more commonly in nature, with examples including arginine and leucine. However, the higher abundance amino acids were not energetically cheaper to make in nature. In addition, among the dataset we studied such as; animal and fungal mitochondrial proteins, human body tissues and various species according to the phylogenetic tree of life (from bacteria to *homo sapiens*), the amino acid occurrence frequency was highly conserved. Also, we attempted to address the entropy reduction paradox in the transcription and translation process by accounting for the involvement of macromolecules ATP and GTP in these process and affecting the overall thermodynamic energy balance. We also investigated the hypothesis whether the amino acids have a higher affinity for their codons or anticodons according to the binding energy values obtained using computational docking simulations. However, the obtained docking scores showed no correlation between the codons or anticodons and the corresponding amino acids, and we have found some paradoxical examples that disprove the proposed hypothesis.

The next goal was to study p53 proteins mutations across a large set of various cancer types. The p53 protein has been selected due to its significant role in the cell cycle, cancer initiation, and progression. We showed that the highly represented mutants are R-H(79%), R-W(71%), R-Q(73%), G-S(55%), and R-S(48%) and at least one of these amino acid mutations occurs in 84% of the cases. Moreover, the Shannon entropy of p53 mutations has been computed in an effort to shed light on the epidemiological findings in terms of five-year-survival rate for cancer patients. However, the entropic approach to the analysis of the role of these important somatic mutations in cancer did not emerge as a prognostic factor in the analysis of cancer epidemiology data.

Finally, using the physical concepts of bistability and phase transitions, we were able to model the cancer cell response to a number of cytotoxic agents used in cancer chemotherapy. We applied the well-known model in the physics of critical phenomena, namely the Ising model and represented the two spin states (spin ‘up’ and ‘down’) in the context of cancer cell biology as a ‘dead’ and ‘alive’ state of cancerous cells, respectively. We explored both an interacting and non-interacting case of cancer cells in a culture with the latter corresponding to the well-studied “bystander effect”. The proposed model has been tested on 13 different cytotoxic compounds applied to various cancer cell lines in culture. The results were in strong agreement with our model showing high consistency among the tested chemotherapy agents. Also the results confirmed the prediction that the EC50 value corresponds to the peak of the susceptibility function, which is an important characteristic of systems at a critical point. The model has been tested successfully on experimental data from both a two-dimensional well-plate cell culture and a three-dimensional spheroid model.

Preface

This thesis is based on materials that has been already submitted for revisions or published in different journals. The research has been done under the supervision of my supervisor, Prof. Tuszynski, at University of Alberta, physics Department.

Section 1.1 of Chapter 1, and Chapter 2 are partially based on a paper accepted for publications in BioSystems journal as: Arbabi Moghadam S., Klobukowski M., and Tuszynski, J. A., “A Search for the Physical Basis of the Genetic Code”. I was responsible for collecting the data from databases, doing all the calculations and data analysis of the paper. Drs. M. Klobukowski and J. A. Tuszynski, supervised the project. Dr. Klobukowski helped with the energy calculations using semi-empirical quantum methodology. I was also responsible for writing the draft and all the author participated in revising and editing the manuscript.

Chapter 3 is partially based on a paper which is under review in BioSystems journal as: Arbabi Moghadam S., Preto J., Klobukowski M., and Tuszynski J. A., “Testing amino acid-codon affinity hypothesis using molecular docking”. I drafted the manuscript and was responsible for designing and structural preparation, performing molecular docking simulation and analyzing the results. Dr. J. Preto was involved in the development of the methodology and SMD simulation as well as manuscript editing and composition. Drs. Klobukowski M., and Tuszynski J. A. were supervisory authors and they participated in the revisions and editing of the manuscript.

Chapter 4 is partially based on a paper which is already submitted in Theoretical Biology and Medical Modeling journal as: Arbabi Moghadam S., Omar S. I., and Tuszynski J. A., “Probability Distributions of p53 Mutations and their Corresponding Shannon Entropies in Different Cancer

Cell Types”. I drafted the manuscript and was responsible for collecting the data from IARC database, performing the calculations and analysis. All the Author were responsible for the editing, revisions and composition of the manuscript.

Chapter 5 has been partially published as a research article as: Arbabi Moghadam S., Rezania V., and Tuszynski J. A., “Cell death and survival due to cytotoxic exposure modeled as a two-state Ising system”, in Royal Society of Open Science (*R. Soc. Open Sci)*, **7**: 191578*,* doi: https://doi.org/10.1098/rsos.191578. In this article, I performed all the simulation, calculations and analysis of the results. I drafted the manuscript, and was involved in developing the methodology. Drs. Rezania V., and Tuszynski J. A. conceived and supervised the project. Dr. Rezania V. was involved in the methodology development. All of the authors participated in editing the manuscript in the submission process. Electronic supplementary material of this work has been published online at at https://doi.org/10.6084/m9.figshare.c.4832877 and the experimental data is available at Dryad Digital Repository at https://dx.doi.org/10.5061/dryad.4qrfj6q6d. Note that Chapter 5 has been modified with the rest of the thesis for consistency and the published supplementary material has been transferred to the content partially.

To my parents for their true love and support

Acknowledgments

I first would like to express my profound gratitude to my honored supervisor, Jack Tuszynski, who has supported me during the life changing-experience undertaking my Ph.D. Jack is a true example of a wonderful human being, incredibly supportive supervisor and a friendly mentor. I will forever be thankful to Jack for the opportunity to work under his supervision. I would also like to acknowledge the financial support provided by Jack from his NSERC funding sources.

Also, I would like to appreciate my supervisory committee members, Drs. Mariusz Klobukowski, Charles Doran and Al Meldrum for their excellent advice and guidance. I am especially thankful to Dr. Klobukowski for his time, guidance, and valuable advice in our long discussions.

Also, I would like to thank all my lovely, friendly, and supportive colleges in the Tuszynski lab, Philip Winter, Jordane Preto, Sara Ibrahim Omar, Mahshad Moshari, Francesco Gentile, Holly Freedman and Aarat Kalra. We have supported each other in happy moments, during periods of sadness and stressful situations. I am thankful of Hamed, my fiancé, who has been consistently supported me and stayed by my side during the bright and dark days of my Ph.D. journey.

Beyond thanks to my family, my mother, Rezvaneh, my father, Reza, my sister, Sara and her husband, Shadmehr for supporting me with positive vibes from Iran. Special thanks go to Sara and Shadmehr for upgrading me to an auntie while I was working hard to finish my dissertation and I would also thank little Radin for being born to change the world to a better place to live.

Table of Contents

List of Tables

List of Figures

`

Chapter 1- Introduction

1.1 Origin of the genetic code [1]

The origin of the genetic code has always been an important and extremely complicated question in biochemistry, biology, and biophysics. The genetic code, which links codons to amino acids, is nearly universal, and its structure is well known. The codons are made of triplets of nucleotides (adenine, guanine, uracil/thymidine, and cytosine) [1−4]. In 1961, Crick *et al.* introduced the concept of codons, which are a triplet of nucleotides [5]. Based on this report, within the same year, Nirenburg *et al.* implemented more experiments that paved the way for deciphering the genetic code by decoding the codon UUU, which codes for the corresponding amino acid phenylalanine [6]. By decoding the structure of the genetic code of *Escherichia coli*, the non-random mapping of 64 codons to 20 amino acid and stop codons was recognized [5−13]. With relatively minor exceptions, nearly all forms of living organisms contribute the same genetic pathways (summarized in chapter 1 Table 2.1) [5,7].

The assignment of the 20 amino acids to particular codons and the degeneracy of the genetic code (meaning multiple codons correspond to the same amino acid) are fundamentally interesting questions since there are about 10^{84} potential algorithms for a genetic code with three nucleotide codons. Therefore, it is important to consider the reasons behind the production of standard amino acid assignments, such as evolutionary forces, historical accidents, or chemical constraints. This information can help to explain remarkable properties of the genetic code, for example the relation

[1] This section is partially based on an introduction of a paper accepted for publication as: Arbabi Moghadam S., Klobukowski M., and Tuszynski, J.A., “A Search for the Physical Basis of the Genetic Code” in BioSystems journal.

between hydrophobic amino acids and the codons that have uracil in the second position, and the inverse relation between the amino acid degeneracy and the molecular weight [14−17]. Excellent reviews on the topic would be found in these references [15,18−24].

The genetic code is evolvable, and Crick proposed the "frozen accident theory" in order to explain its origin [7]. However, as of today, it has been observed that the standard code is not universal, and without changing the basic fundamentals, it is subjected to some important modifications [4,7,8]. Also, a point mutation in a nucleotide in tRNA, RNA editing, or base modification is included in the mechanism of reassignment of codons [4,25−30]. These changes have been addressed in three main theories, the 'ambiguous intermediate theory' [31,32], 'codon capture theory' [33,34], and 'genome streamlining theory' [35,36]. More detailed studies on this topic are provided by these references [4,25−30,37,38].

The main theories on the origin, nature, and evolution of the genetic code can be listed as follows: the RNA world hypothesis, stereochemical theory, adaptive theory, and coevolutionary theory. The RNA world hypothesis postulates that a simple molecule of RNA is at the origin of forming the life on the Earth before the DNA and protein's evolution [39,40]. In the primitive cells, the RNA molecule carries the genetic information and can derive the chemical reactions and it also has the ability to self-replicate. However, in the next phases of the evolutionary time, the DNA molecules contained in living cells within their membranes acquired the function of carrying the genetic information necessary to provide instructions how to make proteins within the machinery of the cell [39,40]. The stereochemical theory asserts that physio-chemical affinity between the amino acids and codons (or anticodons) contributed to the assignment of codons to particular amino acids; thus the code is not an evolutionary accident but was determined by physico-

`

chemical factors [13,41–44]. Adaptive theory states that selective forces caused the structure of the genetic code. The robustness of the genetic code stems from these forces and the mutations happen to minimize the physio-chemical alteration of the created amino acid [20,22,23,28,45–47]. The coevolutionary theory, which is the most popular theory for the origin of the genetic code structure, states that the assignment of codons occurred alongside the evolution of the biosynthesis pathways for the amino acids, and furthermore an early simpler form of the genetic code included only those amino acids which can be formed by prebiotic processes. The basic idea of the coevolutionary theory is similar to the Crick's idea of code expansion; however this theory gained more acceptance by introducing the concept of precursor-product pairs of amino acids. It is worth mentioning that, notwithstanding the history behind these three mentioned theories and the experimental and theoretical evidence obtained so far, none of these hypotheses can be definitively proved about the genetic code. There are still some crucial questions regarding the origin of the genetic code and adaptation theory which are of fundamental importance and remain to be answered. These issues are as follows: the mechanism and basis behind the codon(s) that code for their corresponding amino acids, the genetic code's evolution and the inclusion order of amino acids into codons, as well as the theory behind the existence of the 20 standard amino acids in the genetic code while there are numerous other α-amino acids that are not in the code, but are involved in the metabolism of living organisms such as humans. These are the fundamental questions that scientists have been struggling with since the initial stages of introducing the genetic code, which was more than 50 years ago, till today, and it will probably remain unclear for another 50 years [4–8,28]. In this thesis, in Chapter 2, we aim to study these problems with a probabilistic and energetic view of amino acids and codons and their probability distribution across the species in animal and fungal mitochondrial proteins, and human body tissues.

Also, it is known that all the information about the building blocks of a living organism is stored in a macromolecule called DNA. Messenger RNA, mRNA, is produced by DNA in the transcription process. Transfer ribonucleic acid, tRNA, helps the protein synthesis to make peptide chains in the translation process in the ribosome. It is interesting to study whether the amino acids have a tendency to bind to their codon or anticodon in the ribosome machinery or there are some other mediators that affect the process. In Chapter 3, we use computational methods, specifically, Steered Molecular Dynamics, SMD and molecular docking simulations, to obtain the binding affinity of codons to their corresponding amino acids. The goal is to find if there is any correlation between the binding affinity of an amino acid and cognate codons or anticodons. The results can be used to shed light on one or more of the above-mentioned theories.

1.2 p53 mutations

Tumor protein p53 is a transcription factor that regulates the expression of multiple genes. Mutations in p53 are observed in more than 50 percent of human cancers. Tumor protein p53 is encoded by the TP53 gene in humans. Somatic mutations of TP53 are reported as one of the most frequent changes in human cancer [48,49]. The *TP53* gene codes for a protein that has an important role in the cell, which inhibits the growth and development of the cell by controlling abnormal proliferation and cell division. This protein, called p53, is also known as “Guardian of the Genome” due to its significant responsibility for suppressing tumors [50]. There are two types of mutations in the *TP53*, somatic and germline mutations, which are discussed in Chapter 4 in detail. The region of the gene sequence where the mutation happens is an important consideration for the design of putative activators of mutated p53. p53 protein has four domains, the transactivation domain (TA), the proline-rich domain (PR), the DNA binding domain (DBD), and the tetramerization domain (TED). The DNA-binding domain is where most of the p53 missense mutations occur and there are six hotspot positions that are the most frequently mutated in cancer [51,52]. These mutations occur at codons R175, G245, R248, R249, R273 and R282 in the p53 sequence [51,53,54].

Mutant p53's can be classified as either contact mutants or structural mutants. Mutations that occur in the DNA-binding domain at residues R248 and R273 are examples of contact mutants because these residues interact directly with DNA. Mutations that happen in the other hotspot positions do not have direct interactions with the DNA and are classified as structural mutants, which indirectly affect the p53 binding ability [51,53–55]. The spectrum of the p53 mutations varies among cancers types, such as breast, liver, colon, lung, and other common cancers and further analysis is needed to understand the functionality of the p53 and the etiology of these tumors [48]. Studies show that G:C to T:A mutations happen more frequently in certain cancers such as brain, breast, lymphoid malignancies, liver and lung, whereas T:A mutations occur more frequently in esophageal carcinomas compared with solid tumors [48]. In this thesis, the aim is to study the occurrence frequency of p53 somatic mutations in various cancer types. Furthermore, we analyze the distributions of the mutations of p53 in the studied cancers by defining a dissimilarity factor to one of the most mutated cancers. In addition, the entropy of these mutations and the relationship to the five-years survival rates will be considered for the cancer types.

1.3 Cancer treatments and Ising model

The past decades of cancer research have shed light on tumor development and cancer treatments. Chemotherapy and radiation therapy are the most common modalities of cancer treatments, which in both cases ideally target the damaged cells with less harm inflicted on the normal cells. In

`

chemotherapy treatment, cytotoxic drugs or compounds are used to kill or damage cancer cells, via activity against molecular targets, such as a DNA or an over-expressed protein. In radiotherapy, only the tumor is targeted, while chemotherapy is generally used systemically as a cancer treatment [56−59]. However, research advances in genomics have provided new methods for targeted therapies in which the cytotoxic compounds can be combined with antibodies or nanoparticles to attack a specific target tissue. Choosing the optimal pharmacological agent, drug dose, and interval for the administration of the drug is complicated and depends on various factors. Furthermore, to prevent tumor cells to become resistant to the cytotoxic drugs, which stems from the various features of the cancerous cells, selecting the cytotoxic agent and the corresponding dosage is crucial and needs special attention. Moreover, by diffusion, the transfected cells resulting from treatment can be transferred into the nearby untransfected cells, and respectively can damage the cells neighboring the tumor site. This process is known as the bystander effect [60,61]. The bystander effect refers to the indirect death or lethal damage to cells that are not directly affected by therapeutic interventions. This effect is seen in both radiation therapy and chemotherapy [61−75].

The aim here is to find a model that can explain this dynamic process, both in chemotherapy and radiation therapy treatments, and develop a statistical analysis of chemotherapeutic agent effects on cancer cells or tumors. In this study, we propose using concepts developed in the physics of phase transitions such as and bistability and order parameters. Bistability could be an appropriate representation of the cells under the stochastic transition from alive (proliferating cells) to dead/senescent (non-proliferating cells). In a system, which is undergoing a phase transition, multistability, or to be more specific bistability, is used as a common characteristic. Phase transitions are used to analyze the behavior of a system that has more than one distinct equilibrium

`

(or stable) state. Two other parameters, called the control parameter and order parameter, are critical properties of systems undergoing a phase transition. The control parameter is an external factor (a knob), and by changing the value of the control parameter, the critical system responding to this change can switch between its distinct states at the transition point. The order parameter has a defined value in each of the distinct states of the system and describes the response of the system to changes in external conditions. The Ising model is a two-state spin-1/2 model and is considered as a powerful, yet very simple, mathematical model of phase transitions. The effect of cytotoxicity on cancer cells can be elegantly described by the Ising model by assuming that the spin up and down states can be interpreted as alive and dead states in cancer cells, respectively [59,61,76–82]. The aim in Chapter 5 is to implement the Ising model to study the response of cancer cells exposed to chemotherapeutic agents. In Chapter 5, we study the cytotoxicity effect of different compounds in chemotherapy and model this behavior with the use of the Ising model of phase transitions. Using the biological data of collaborators in the Netherlands Translational Research Center B.V. (Oncolines), we apply the Ising model methodology on various cytotoxic drugs and cancer cell lines in a dose-response manner [83,84]. We are also interested in applying another well-known and widely-used model of physical systems at criticality, namely the Landau theory of phase transition, as well as the susceptibility function to describe the dose-response curves in order to reveal the deeper meaning of the commonly used EC50 (extinction coefficient such that 50% of the cells become unviable at the corresponding concentration) values that determine the sensitivity of the system to external perturbations. This is discussed in detail in Chapter 5.

`

Chapter 2 - A Search for the Physical Basis of the Genetic Code [2]

2.1 Abstract

DNA contains the genetic code, which provides complete information about the synthesis of proteins in every living cell. Each gene encodes for a corresponding protein but most of the DNA sequence is non-coding. In addition to this non-coding part of the DNA, there is another redundancy, namely a multiplicity of DNA triplets (codons) corresponding to code for a given amino acid. In this work, we investigate possible physical reasons for the coding redundancy, by exploring free energy considerations and abundance probabilities as potential insights.

2.2 Introduction

DNA carries the genetic code, which provides information regarding the proper development and well-functioning of any living organism and many viruses [5,7,85–88]. While DNA includes genes, that is, regions coding for their corresponding proteins (or protein domains), most of the DNA sequence in most eukaryotes is non-coding. This non-coding part of DNA is not involved in protein synthesis but is used in other mechanisms (e.g., transcription of functional non-coding RNA has an as yet unknown function or no function at all as so-called "junk DNA", and regulation of gene expression). In any organism, gene expression is carried out in three steps: (i) transcription from a specific gene to messenger RNA (mRNA) (ii) removing the mRNA introns in the splicing process, and (iii) translation from mRNA leading to the synthesis of new proteins *via* ribosomes,

[2] This chapter is partially based on a paper accepted for publication as: Arbabi Moghadam S., Klobukowski M., and Tuszynski, J.A., “A Search for the Physical Basis of the Genetic Code” in BioSystems journal.

`

macromolecular complexes responsible for assembling amino acids into the correct sequence [88–90]. The main focus of this study is not on modeling or exploring molecular mechanisms accounting for transcription and translation processes but on investigating typical features of gene expression, in particular codon/amino-acid relationships, from energetic and probabilistic perspectives. Special attention is also given to codon degeneracy and its implication for protein synthesis and expression.

Amino acids are small organic molecules, which are the building blocks of proteins. Proteins play a key role in most of the biological processes, with each protein being a sequence of hundreds of amino acids folded into a specific functional 3D structure. There are twenty naturally-occurring amino acids [88,89]. Each amino acid has its own unique physico-chemical properties such as hydrophobicity, charge, dipole moment and size. Beside the twenty standard amino acids, some non-standard residues exist, which do not participate in peptide synthesis (we are only interested in standard amino acids here). Regarding their chemical structure, amino acids share common elements with each other including a carboxyl group and an amine group. These two groups are attached to a carbon atom called an α-carbon. Each of these amino acids has a specific side chain, called an R-group that is linked to the α-carbon (except proline which has a secondary amine group and is an imino acid; however, it is commonly classified within the amino acid group). R-groups have various features such as shape, size, and charge that allow amino acids to be grouped according to the chemical properties of the side chains [4,13,85,88–91].

Genetic experiments have proven that an amino acid is encoded by a group of three base pairs of the DNA called a codon. Codons are transcribed from three base pairs of DNA to three nucleotides in messenger RNA, which eventually become translated into the production of specific amino

`

acids *via* the ribosome [12,85,86]. Since there are four distinct nitrogenous nucleotide bases: Adenine (A), Guanine (G), Cytosine (C), and Thymine (T), this results in $4^3 = 64$ permutations, i.e., 64 types of codons that can be identified in a DNA sequence [12,85,86]. Sixty-one of these combinations represent amino acids while the three remaining ones are stop codons implying the end of the transcription process [88,89]. From statistical arguments, one might expect that every amino acid is coded by approximately the same number of codons. However, this is not the case: some amino acids are coded by only one codon, others are coded by up to six codons. Table 2.1 provides a list of all the amino acids and the corresponding codons.

Table 2.1 Amino acid table and their corresponding codons [7].

	Second Letter								
First Letter	T		C		A		G		Third Letter
T	TTT	Phe F	TCT	Ser S	TAT	Tyr Y	TGT	Cys C	TACG
	TTC		TCC		TAC		TGC		
	TTA	Leu L	TCA		TAA	Stop	TGA	Stop	
	TTG		TCG		TAG	Stop	TGG	Trp W	
C	CTT	Leu L	CCT	Pro P	CAT	His H	CGT	Arg R	TACG
	CTC		CCC		CAC		CGC		
	CTA		CCA		CAA	Gln Q	CGA		
	CTG		CCG		CAG		CGG		
A	ATT	Ile I	ACT	Thr T	AAT	Asn N	AGT	Ser S	TACG
	ATC		ACC		AAC		AGC		
	ATA		ACA		AAA	Lys K	AGA	Arg R	
	ATG	Met M	ACG		AAG		AGG		
G	GTT	Val V	GCT	Ala A	GAT	Asp D	GGT	Gly G	TACG
	GTC		GCC		GAC		GGC		
	GTA		GCA		GAA	Glu E	GGA		
	GTG		GCG		GAG		GGG		

`

Since 1960, molecular biologists have attempted to find out whether the amino acid composition was a reflection of the genetic code sequence or a result of natural selection of amino acids [92,93]. Based on the assumption that allocation of amino acids to their corresponding codons is accidental, random and highly improbable, amino acids which are related to each other would be expected to have unrelated codons [28,86]. The genetic code is also expected to be frozen, meaning that any change in the genetic code will cause a lot of simultaneous changes in the gene/protein sequences and consequently will disrupt thousands of genes [7,85–87]. Therefore, from these observations, one can legitimately wonder why only four nucleotides exist in the genetic code or why 20 amino acids can be encoded by the standard codon table [94–97]. The redundancy of the genetic code implies that most of the amino acids are encoded by multiple synonymous codons but there is no ambiguity in the assignment of an amino acid to a codon [85,86]. For example, although codons GCT and GCC both specify alanine, which implies redundancy, neither specifies another amino acid, which means that there is no ambiguity. Amino acids can be classified in many different ways based on their characteristics, e.g., aliphatic, aromatic, acidic, basic, polar, hydrophobic, etc. However, since many amino acids belong to multiple families, assigning each amino acid to an invariant group is difficult, see Figure 2.1 [90,98].

`

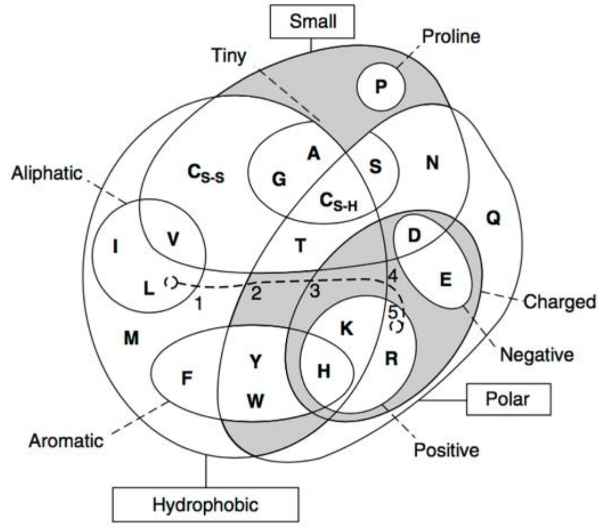


Figure 2.1 Venn diagram illustrating the properties of amino acids [99].

2.3 Methodology

Our main objective here is to obtain and analyze amino acids' thermodynamic properties, especially the Gibbs free energy with quantum chemistry methods using GAMESS and Gaussian software and find out if there is a correlation with codon Gibbs free energy [100–102]. On the other hand, the other approach to this study would be to seek a correlation between the free energy and the occurrence frequency of amino acids. Shen *et al.* reported the probability distribution (frequency) of all amino acids contained in the Swiss-Prot database, $P_{\mathrm{AA-SP}}$ [103]. The probability distribution of amino acids, $P_{\mathrm{AA-SP}}$, was obtained by averaging over the total number of amino acids found within the Swiss-Prot database [103−105]. In addition, occurrence frequency of the amino acids was studied in different species. Considering the evolutionary tree of life, some species are selected judiciously to provide a spectrum from the simplest species to the most complex ones. Using the NCBI databank the occurrence frequency of the amino acids and codons for *E. coli*, Halobacteriales salinarum, Haloferax volcanii, Physcomitrella patens, Arabidopsis

`

thaliana, Paramecium, Porphyria, Cerevisiae, sponge, spider, fruit fly, octopus, starfish, salmon, frog, crocodile, snake, pigeon, chicken, elephant, dog, rabbit, chimpanzee and human has been obtained, giving an insight into the uniformity of the amino acid abundance frequency across diverse species [106,107].

2.4 Results and discussion

2.4.1 Calculations of Codon and Amino acids' Gibbs free energies

An important quantity characterizing every amino acid (and every nucleotide) is the Gibbs free energy, G_{AA}, required to build their chemical structures. To estimate G_{AA}, a semiempirical quantum chemistry method was applied *via* the GAMESS program [101,102]. Semi-empirical methods are based on an approximate scheme that uses experimental data in order to determine many of the integrals involved in the Hartree-Fock method [108–110]. Among available basis sets for the semi-empirical Hamiltonians, the parameterized Method 3, PM3, was used [111–114]. Note that the PM3 Hamiltonian is parameterized in such a way that the thermodynamic properties associated with a large number of molecules can be reproduced. Among all the thermochemistry quantities, such as internal energy, E, enthalpy, H, Gibbs free energy, G, and the entropy, S, we are mainly interested in the Gibbs free energy values, which are shown in Table 2.2 for the 20 standard amino acids. In addition, the last four rows show the thermochemistry of the nitrogen bases with sugar and phosphate, so called deoxyadenosine monophosphate, dAMP, deoxyguanosine monophosphate, dGMP, deoxycytidine monophosphate, dCMP and deoxythymidine monophosphate, dTMP. Using GAMESS, the energy and the correction to the final state will be obtained. However, the final energy will be represented as $G_{AA} = G_0 + G$.

`

Table 2.2 Thermochemistry of amino acids and four nucleotides at PM3 level calculated by GAMESS [101,102].

Amino acid	E_{corr} kJ/mol	H_{corr} kJ/mol	S_{corr} J/mol/K	G_{corr} kJ/mol	G_0 kJ/mol	$G_{AA} = G_0 + G$ kJ/mol
Alanine	302.6	305.1	343.2	202.8	-115994.1	-115791.3
Arginine	611.9	614.4	494.1	467.0	-206753.1	-206286.1
Asparagine	383.7	386.2	406.3	265.1	-172350.3	-172085.2
Aspartic acid	349.2	351.7	392.4	234.7	-183455.8	-183221.1
Cysteine	294.8	297.3	374.9	185.5	-133740.5	-133555.0
Glutamic acid	426.1	428.6	434.3	299.1	-197749.3	-197450.2
Glutamine	459.0	461.5	431.8	332.7	-186635.7	-186303.0
Glycine	225.8	228.3	312.6	135.1	-101692.7	-101557.6
Histidine	444.0	446.5	430.8	318.1	-183828.1	-183510.0
Isoleucine	533.6	536.1	417.5	411.6	-158829.0	-158417.4
Leucine	533.7	536.2	412.4	413.2	-158840.8	-158427.6
Lysine	580.4	582.9	460.7	445.5	-175716.4	-175270.9
Methionine	460.5	463.0	429.8	334.8	-162310.6	-161975.8
Phenylalanine	524.7	527.2	445.6	394.3	-189722.0	-189327.7
Proline	391.1	393.6	356.3	287.4	-141579.8	-141292.4
Serine	317.2	319.7	368.2	209.9	-144045.2	-143835.3
Threonine	394.7	397.1	385.1	282.3	-158333.7	-158051.4
Tryptophan	607.9	610.4	493.7	463.2	-229313.3	-228850.1
Tyrosine	542.0	544.4	456.1	408.4	-217818.8	-217410.4
Valine	456.8	459.2	397.7	340.7	-144553.4	-144212.7
Nucleotide						
dAMP	748.0	750.5	667.1	551.6	-391607.7	-391056.1
dCMP	711.9	714.4	658.2	518.2	-377384.2	-376866.0
dGMP	765.9	768.3	695.5	561.0	-419655.6	-419094.6
dTMP	758.2	760.7	680.9	557.7	-402808.7	-402251.0

`

In Table 2.3, the base pair stacking energy G_{S}, and the nearest neighbor energy, G_{NN}, have been reported for each nucleic acid [90,92,115–118]. The experimental stacking energy, G_{S}, has been obtained assuming that equilibrium was set between stacked and unstacked conformations at the nick site of a DNA duplex [116,119]. Moreover, the nearest neighbor energy G_{NN}, is counted based on two nearest neighbor base pair doublets. There are ten possible doublets in the DNA double strings $(5' - 3')$. These are: $AG = CT$, $AA = TT, GG = CC, AC = GT$, $GA = TC, TG = CA, CG$, $GC, AT,$ and TA. Using the nucleotide energies reported in Table 2.2 and 2.3, it is possible to work out the Gibbs free energy for all the 64 codons, G_{codon}, using the following relation

$$G_{\mathrm{codon}} = G_{NA} + G_S + G_{NN} \tag{2.1}$$

where G_{NA} stands for the nucleic acid energy, G_{S} represents the energy of the melting stability of base pair stacks and G_{NN} is the energy of nearest neighbor nucleic acid pairs in a DNA strand [92,115–118]. As an example, for the arginine codon CGA, the complementary nucleobases would correspond to GCT, meaning that the nucleic acid energy G_{NA} and the stacking energy G_S would be

$$G_{NA} = (2G_C + 2G_G + G_A + G_T) \tag{2.2}$$

$$G_S = (G_{CG} + G_{GC} + G_{AT}) \tag{2.3}$$

and the nearest neighbor energy reads as

$$G_{NN} = \left(G_{CG/GC} + G_{GC/AT}\right) \tag{2.4}$$

where these energies will be discussed in section 2.4.2.

`

Table 2.3 Gibbs free energy for base pair stacking and nearest neighbor in double-stranded DNA base-pairs [90,92,115–118].

Stacking Gibbs free energy G_S (kJ/mol)		**Nearest neighbor Gibbs free energy G_{NN} (kJ/mol**	
TA	-0.5	AA/TT	-4.2
TG-CA	-3.3	AT/TA	-3.7
CG	-6.0	TA/AT	-2.4
AG-CT	-5.4	CA/GT	-6.1
AA-TT	-4.4	GT/CA	-6.0
AT	-5.3	CT/GA	-5.4
GA-TC	-6.9	GA/CT	-5.4
CC-GG	-8.2	CG/GC	-9.1
AC-GT	-8.5	GC/CG	-9.4
GC	-11.3	GG/CC	-7.7
TA	-0.5		
TG-CA	-3.3		

2.4.2 Energy Correlations

Using the Gibbs free energies computed in the previous section, we now attempt to find out whether any correlation exists between the energy for each amino acid and those for the corresponding codons. In addition, in the next step, we aim to find out a possible correlation between amino acid energies and empirical probability of their occurrence, as might be expected for example from the Arrhenius relation for chemical reactions [120–122]. In Figure 2.2.a, the Gibbs free energy of amino acids, G_{AA}, is plotted versus the Gibbs free energy of the corresponding codons, G_{codon}, and the lowest codon energy of each amino acid is labeled. Some of the codon energies are overlaid on each due to being close in value. For example, glycine which

`

has four codons, has the following energies: $G_1 = -2411382.5$, $G_2 = -2411380.3$, $G_3 = -2411378.9$ and $G_4 = -2411378.6$ kJ/mol. Despite no visible correlation being found between the two sets of values, several observations can be made. As a general tendency, the higher the free energy of the amino acids, the lower the free energy of the corresponding codon. This result becomes more evident when plotting $\mathrm{G_{AA}}$ as a function of G_{codon} averaged over the possible codons coding for the same amino acid in Figure 2.2.b. It is also clear that charged amino acids (arginine, lysine, glutamic acid, aspartic acid) are likely to have a lower energy. Also, results representing the data for each codon yield a large amount of scatter, even though we have combined the values for all codons representing a given amino acid into their average energies as plotted in Figure 2.2.b. As shown, the average codon energies are correlated only weakly with the average of amino acid energies within the standard deviation. The standard deviations are shown by the blue horizontal lines. In Figure 2.2.b, there appear to be two outliers, namely arginine and tryptophan. Arginine has the highest redundancy by being encoded by six codons. Tryptophan has the lowest amino acid energy. It is an uncommon amino acid and it is special, being encoded by only one triplet of base pairs. Also, tryptophan has two rings in the chemical structure and has a large number of atoms associated with it, which causes it to have the lowest amino acid energy (but the highest in magnitude). Moreover, glycine, which has the smallest number of atoms has the highest amino acid energy (but the lowest in magnitude).

`

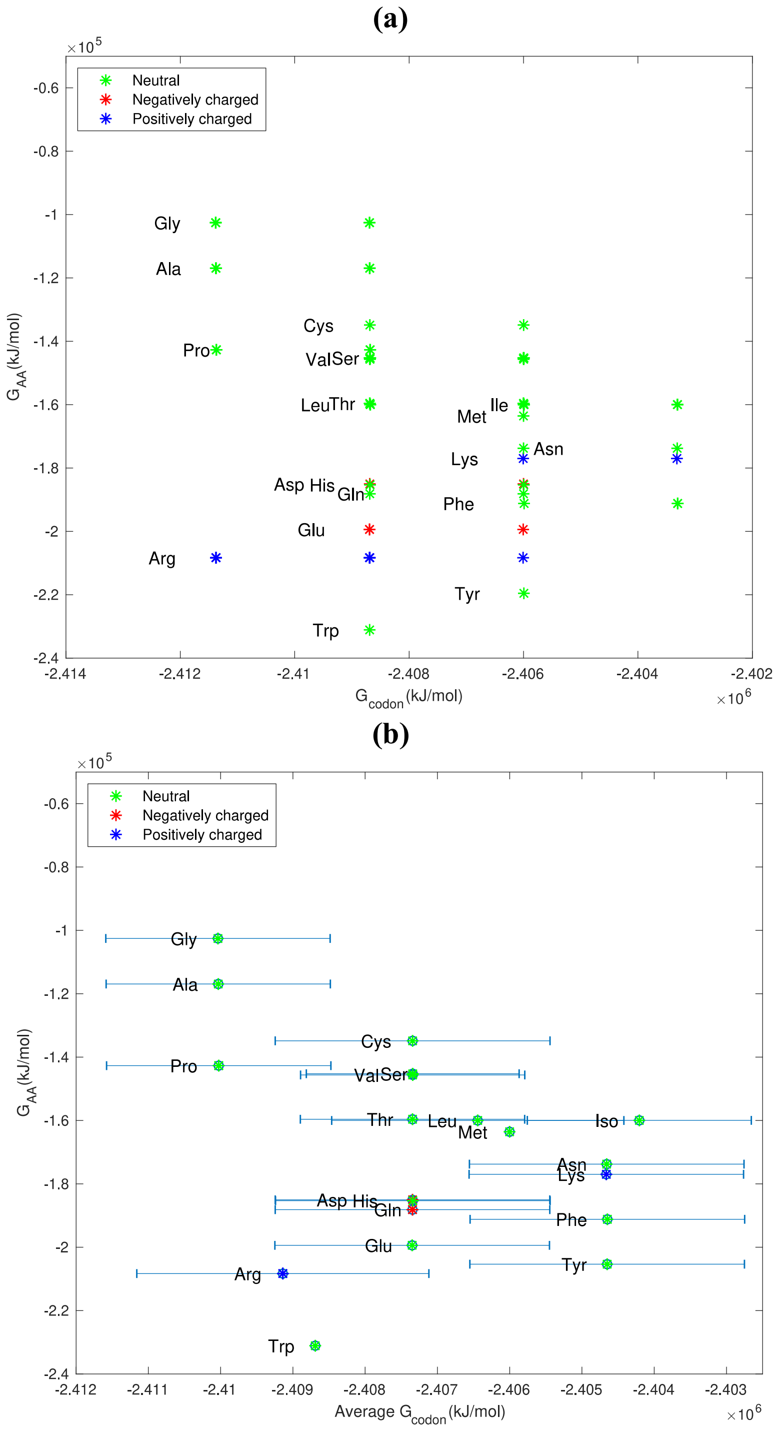


Figure 2.2 (a) Gibbs energy of amino acids (G_{AA})versus Gibbs energy of codons, (G_{codon}) **(b)** Gibbs energy of amino acids (G_{AA})versus average Gibbs energy of corresponding codons G_{codon} for each amino acid. In both plots, colors green, red and blue refer to the neutral, negative and positively charged amino acids. Standard deviations in panel b are shown by blue

`

horizontal lines. Amino acid Gibbs free energy are calculated using GAMESS software and the codons Gibbs energy are calculated using Eq. (2.3).

To compare and check the accuracy of the obtained energies using GAMESS , another software package has been chosen for additional testing. Gaussian software [100], which is one of the most popular computational chemistry tools, has been used to calculate the amino acid energies. Using the Gaussian program we computed amino acid energies using the density functional theory (DFT) employing the B3LYP functional [123–127] and 6-311++G(df, pd) basis set [128–130]. In this approach all electrons were included explicitly, and the effects of electron correlation taken into account *via* the density functional theory. This approach can be compared with the results of GAMESS software that was mentioned earlier. However, using GAMESS, the semi-empirical method, which is based on the Hartree-Fock formalism with some approximations, some of the integrals will be disregarded and some approximations will be applied [131−135]. The first approximation is that only the valence electrons are treated explicitly and the core electrons are removed. Secondly, many difficult integrals are neglected and thirdly, the effects of electron correlation are ignored in solving the Schrödinger equation. The errors introduced by these approximations are expected to be alleviated by the use of empirical parameters derived by fitting of computed results to the corresponding experimental data. The two different approaches that we used, the all-electron DFT(B3LYP/6-311++G(df,pd)) used in Gaussian and the semi-empirical method HF/PM3 used in GAMESS, use different definitions of the zero of the total energy [111–114]. In the all-electron approach the zero of the total energy is evaluated with respect to the energy of the same system with all particles (electrons and nuclei) at infinite distances. On the other hand, in the semi-empirical calculations the zero of the total energy is evaluated as the heat of formation. Therefore, it is expected that there are some differences in the geometric structures

`

which leads to the differences in the energies. Table 2.4 shows the energies obtained in the all-electron approach and in the semi-empirical method with both GAMESS and Gaussian software.

Table 2.4 Gibbs energies calculated from GAMESS and Gaussian [100–102].

Amino acid	**GAMESS G_{AA} (hartree)**	**Gaussian G_{AA} (hartree)**
Alanine	-44.4	-323.8
Arginine	-79.2	-606.6
Asparagine	-66.0	-492.5
Aspartic acid	-70.3	-512.4
Cysteine	-51.1	-722.0
Glutamic acid	-75.8	-551.7
Glutamine	-71.5	-531.8
Glycine	-39.0	-284.5
Histidine	-70.4	-548.8
Isoleucine	-60.8	-441.7
Leucine	-60.8	-441.7
Lysine	-67.3	-497.5
Methionine	-62.2	-800.6
Phenylalanine	-72.7	-554.8
Proline	-54.2	-401.2
Serine	-55.2	-399.0
Threonine	-60.6	-438.3
Tryptophan	-87.8	-686.4
Tyrosine	-83.4	-630.1
Valine	-55.3	-402.4
Nucleotide		
dAMP	-150.1	-1456.3
dCMP	-144.6	-1383.9
dGMP	-160.8	-1531.5
dTMP	-154.4	-1443.1

`

Moreover, Figure 2.3 illustrates the comparison between the Gibbs free energy of amino acid and four nucleotides using Gaussian and GAMESS resulting in the linear regression equation given by $y = 1.1x - 0.1$ and the r-squared value is found to be 0.92 (the values are normalized to the maximum value of each methods). However, cysteine and methionine, both of which contain sulphur in their chemical structure, are the two outliers found in this fitting procedure.

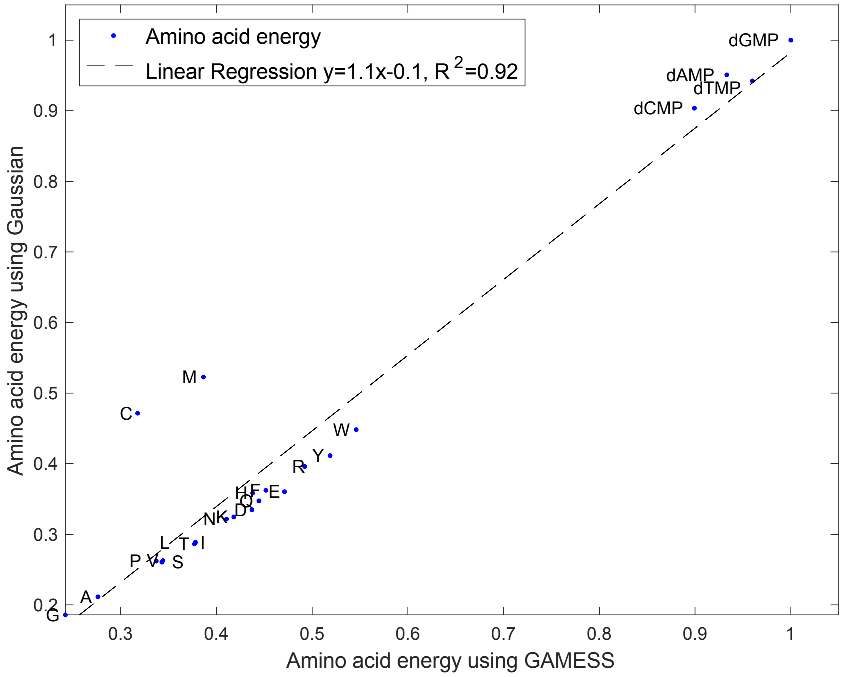


Figure 2.3 Comparison between the amino acid and nucleotide's Gibbs free energy using Gaussian and GAMESS (normalized to maximum energy for each method), the dashed line shows the linear regression $y = 1.1x - 0.1$ with $R^2 = 0.92$.

2.4.3 Probability and energy correlation

In this section, we investigate the correlation between the energy and the occurrence frequency of amino acids [136]. Based on this information and the calculated codon energies we have generated a plot in Figure 2.4, which shows $P_{\mathrm{AA-SP}}$ versus G_{codon}. The probability distribution of amino acids in human, $P_{\text{AA}-\text{SP}}$, was obtained by averaging over the total number of amino acids found within the Swiss-Prot database [104,105,136]. Also, in Figure 2.4.a, amino acids are plotted using three colors based on their charge, such that blue star stands for positive amino acids, red for negative amino acids, and the rest are represented in green. From this figure, it is clear that leucine with six codons is the most probable amino acid while tryptophan with only a single codon is the least probable one. Second and third most probable amino acids are alanine and glycine, respectively. The second and third least probable amino acids are cysteine and histidine, whose energy is approximately in the middle of the energy range. A weak correlation can be found among the probability and the energy of codons. However, it is expected that amino acids with lower formation energy are more probable. In other words, the cheaper energy cost amino acids by nature are more probable. Figure 2.4.b, shows the plot of $P_{\text{AA}-\text{SP}}$ with respect to the Boltzmann factor pre-multiplied by the redundancy factor, i.e. $g(x)Exp\left(\frac{-\langle G_{AA}\rangle}{kT}\right)$, where $g(x)$ represents the amino acid redundancy of codons and $\langle G_{AA}\rangle$ represents the average energy of codons in each amino acid (the energy is normalized to the maximum value). By increasing the magnitude of the energy, the probability increases.

`

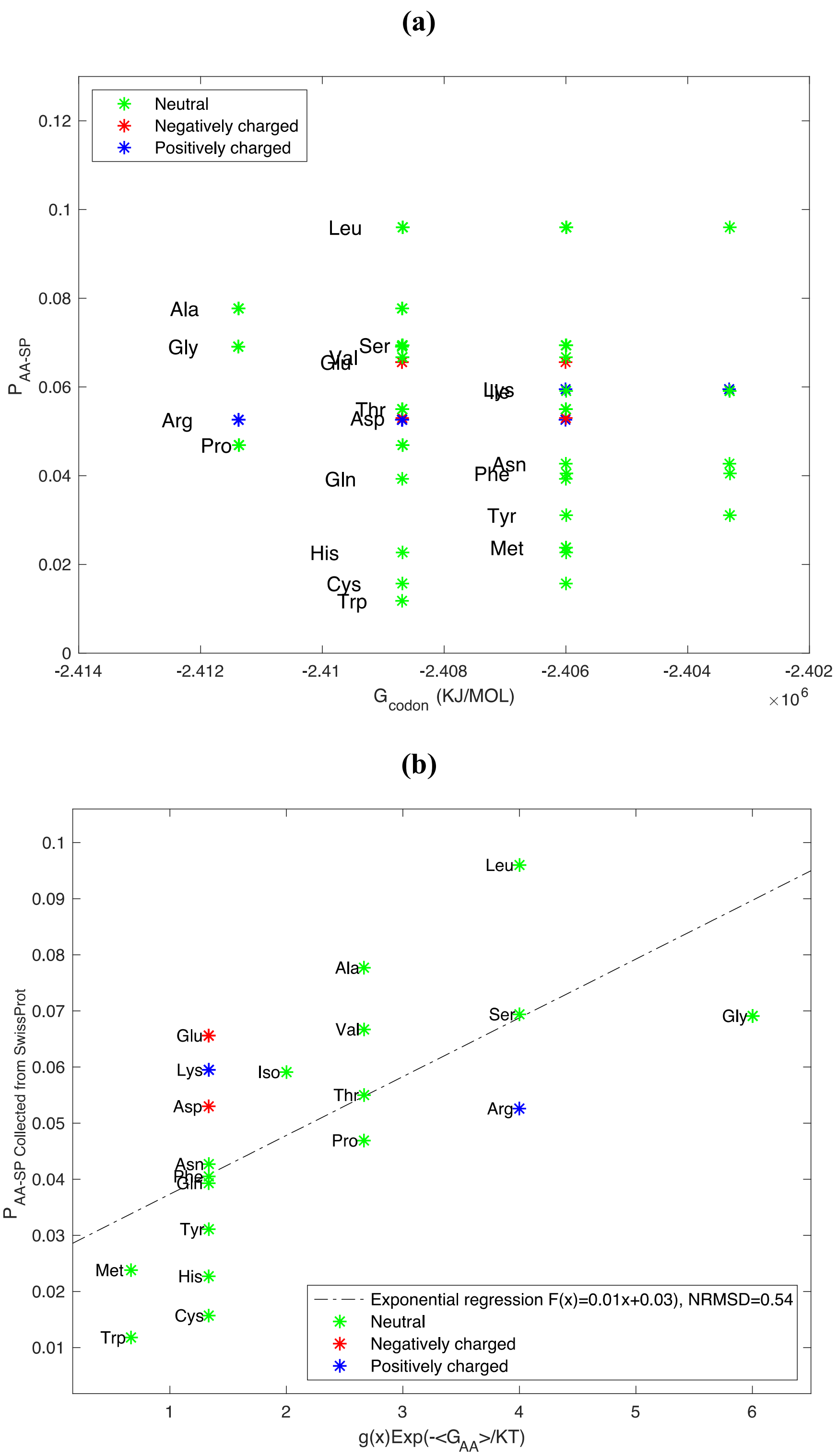


Figure 2.4 (a) The probability distribution of amino acids in human, P_{AA-SP} versus Gibbs energy of codons, G_{codon}, **(b)** P_{AA-SP} versus the Boltzmann factor taking to account for the

`

corresponding the free energy of amino acids, $g(x)Exp\left(\frac{-\langle G_{AA}\rangle}{k_BT}\right)$, $g(x)$ stands for the amino acid degeneracy and $\langle G_{AA}\rangle$ represents the averaging over the Gibbs energy of codons that code for the corresponding amino acid. Electrostatic charge of the amino acids is shown by green (neutral), red (negative) and blue (positive) star.

2.4.4 Investigating amino acid frequency across different species

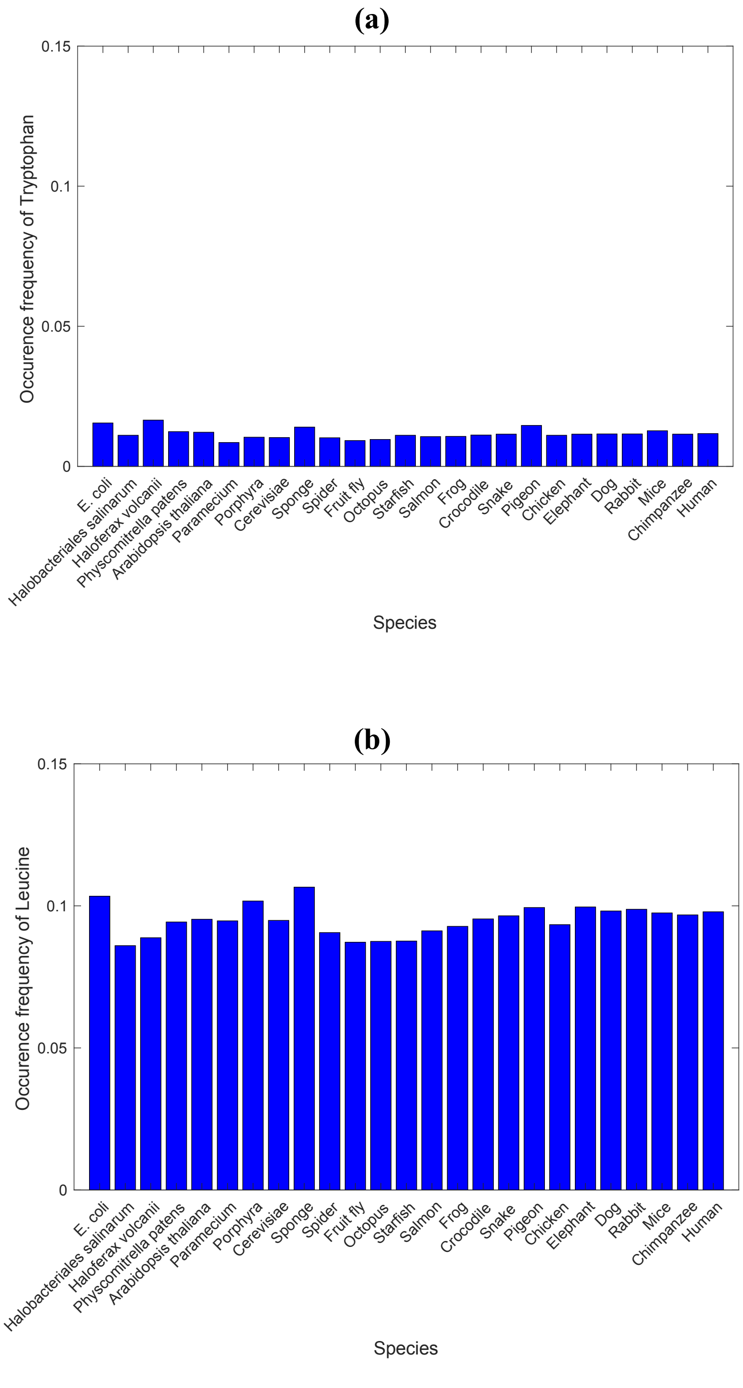
To obtain an insight into the consistency of the probability distribution for the occurrence of amino acids, we need to compare their abundances in different species. This involves considering the evolutionary tree of life [90,98]. Our analysis will aim to show an inferred evolutionary relationship within different biological species or other entities based on genetic characteristics and physical similarities/differences [90]. Some species are selected and analyzed somewhat judiciously to represent a spectrum from the simplest species to the most complex ones. The corresponding sequences were extracted from NCBI, a databank of biomedical and genomic information (http://www.ncbi.nlih.gov) [106,107]. The species we selected for analysis include: *E. coli* (Bacteria), Halobacteriales salinarum (Archaea), Haloferax volcanii (Archaea), Physcomitrella patens (Moss), Arabidopsis thaliana (Plants), Paramecium (Algae), Porphyra (Red algae), Cerevisiae (Fungus), sponge, spider (Protostomes), fruit fly (Drosophila melanogaster, Insects), octopus (Protostomes), starfish (Acanthaster planci, Echinoderms), salmon (Fish), frog (Western clawed frog, Amphibians), crocodile (Reptile), snake (Python bivittatus, Reptile), pigeon (Birds), chicken (Gallus gallus, Birds), elephant (Loxodonta africana, Mammals), dog (Canis lupus familiaris, Mammals), rabbit (Mammals), chimpanzee (Pan troglodytes, Mammals), human (Mammals). All the sequences for different species are extracted from the UniProt data bank [104] and the occurrence frequency of each species amino acids has

`

been calculated such that the sum of the probabilities for each species equals one. In Figure 2.5.a and b, the occurrence frequency of leucine and tryptophan are shown and the rest of the amino acid plots are shown in Appendix A.1. Despite some fluctuations, especially for Sponge, it seems the amino acid distribution is consistent among the species. This is in itself an interesting finding indicating that amino acid abundance is a stable attractor in this multi-dimensional space, which has not been affected by billions of years of the evolution of life on this planet.

To summarize, the average probabilities over all species considering the standard deviation versus $P_{\mathrm{AA-SP}}$ are plotted in Figure 2.5.c. . In addition, to check for the consistency of the obtained probabilities over different species, the probability of amino acid collected from Swiss-Prot, $P_{\mathrm{AA-SP}}$, is displayed versus the average probability over the species. Clearly, the probabilities are consistent and the outliers are within the standard deviation. By taking another approach, occurrence frequency of all the amino acids has been plotted for each species in Figure 2.5.d. It can be seen that tryptophan is the least probable amino acid among all of the studied species (Archaea (Halobacteriales salinarum and Haloferax volcanii) and sponge) and leucine is the most probable amino acid in all species (except for Archaea (Halobacteriales salinarum and Haloferax volcanii), spider, sponge and octopus). All of the other plots have been shown in Appendix A.2.

`


(a)
Occurence frequency of Tryptophan
0.15
0.1
0.05
0
Species
E. coli
Halobacteriales salinarum
Haloferax volcanii
Physcomitrella patens
Arabidopsis thaliana
Paramecium
Porphyra
Cerevisiae
Sponge
Spider
Fruit fly
Octopus
Starfish
Salmon
Frog
Crocodile
Snake
Pigeon
Chicken
Elephant
Dog
Rabbit
Mice
Chimpanzee
Human
(b)
Occurence frequency of Leucine
0.15
0.1
0.05
0
Species
E. coli
Halobacteriales salinarum
Haloferax volcanii
Physcomitrella patens
Arabidopsis thaliana
Paramecium
Porphyra
Cerevisiae
Sponge
Spider
Fruit fly
Octopus
Starfish
Salmon
Frog
Crocodile
Snake
Pigeon
Chicken
Elephant
Dog
Rabbit
Mice
Chimpanzee
Human

`

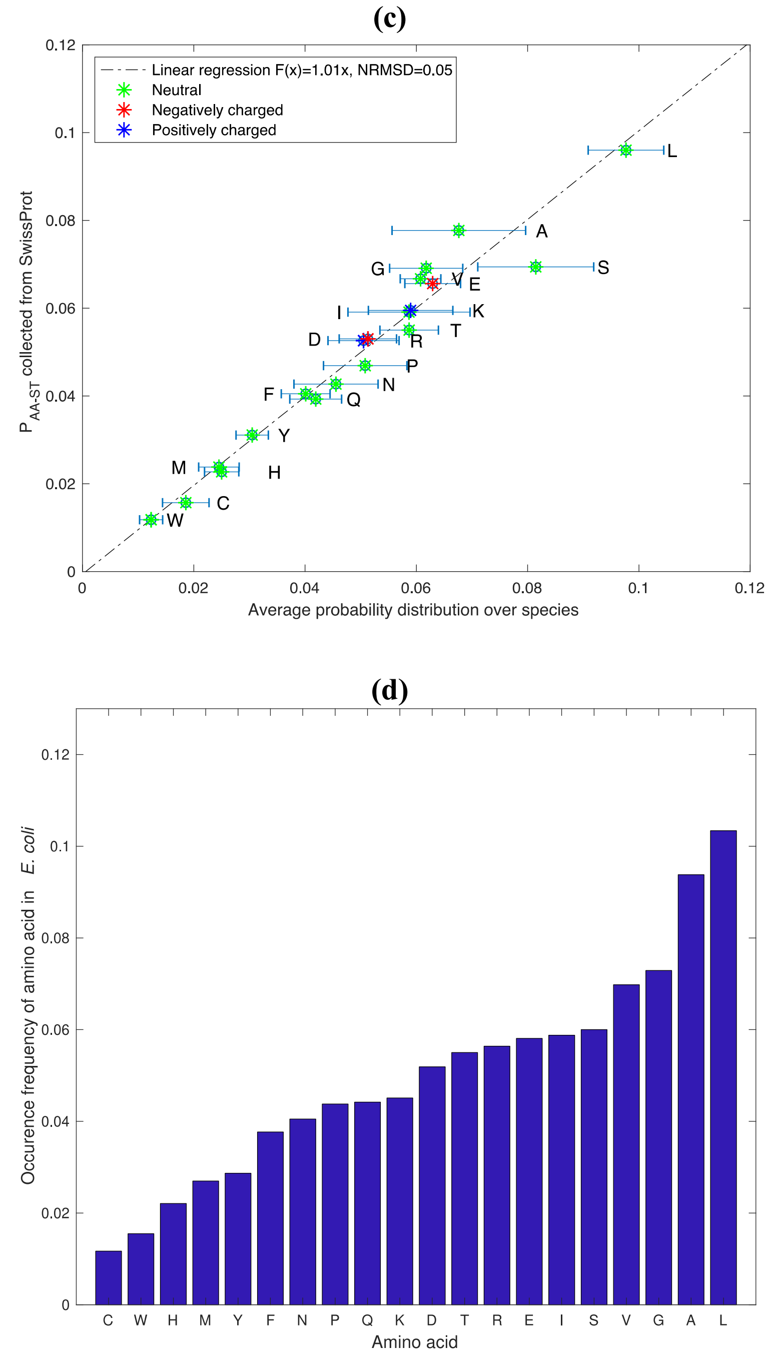


Figure 2.5 (a) The occurrence frequency of tryptophan for different species, the average value and the standard deviations are (0.011 ± 0.001), **(b)** The occurrence frequency of leucine for different species, the average value and the standard deviations are (0.095 ± 0.005), **(c)** Probability of amino acids obtained from Swiss-Prot database for human versus the average probability over all species considering obtained from NCBI database, the standard deviation

`

showed by blue bars, and the amino acids are color coded based on their electrostatic charge, **(d)** amino acid occurrence frequency in *E. coli* by an increasing order.

2.4.5 Probability distribution in terms of amino acid degeneracy

In this section, we inquire whether there might be a correlation between the redundancy and the probability of occurrence of each amino acid. Figure 2.6.a, shows the amino acid probability distribution, $P_{\mathrm{AA-SP}}$ [104], versus the number of corresponding codons or amino acid degeneracy. Clearly, an increase in amino acid degeneracy correlates with an increase in the amino acid's probability of occurrence. As was discussed in section 2.4.3, tryptophan with one codon has the lowest probability and leucine with six codons has the highest probability. It is somewhat analogous to the lottery rules, the more tickets one buys, the higher the chance to win. Although this interpretation might not be precise for methionine with one codon and arginine with six codons, it seems to be precise for serine. Figure 2.6.b, also demonstrates the probability of amino acids in human in an increasing order. This illustrates the idea that the more codons for a single amino acid, the higher the probability of occurrence. In addition, some essential amino acids are also distinguished by pink circles in both plots.

`

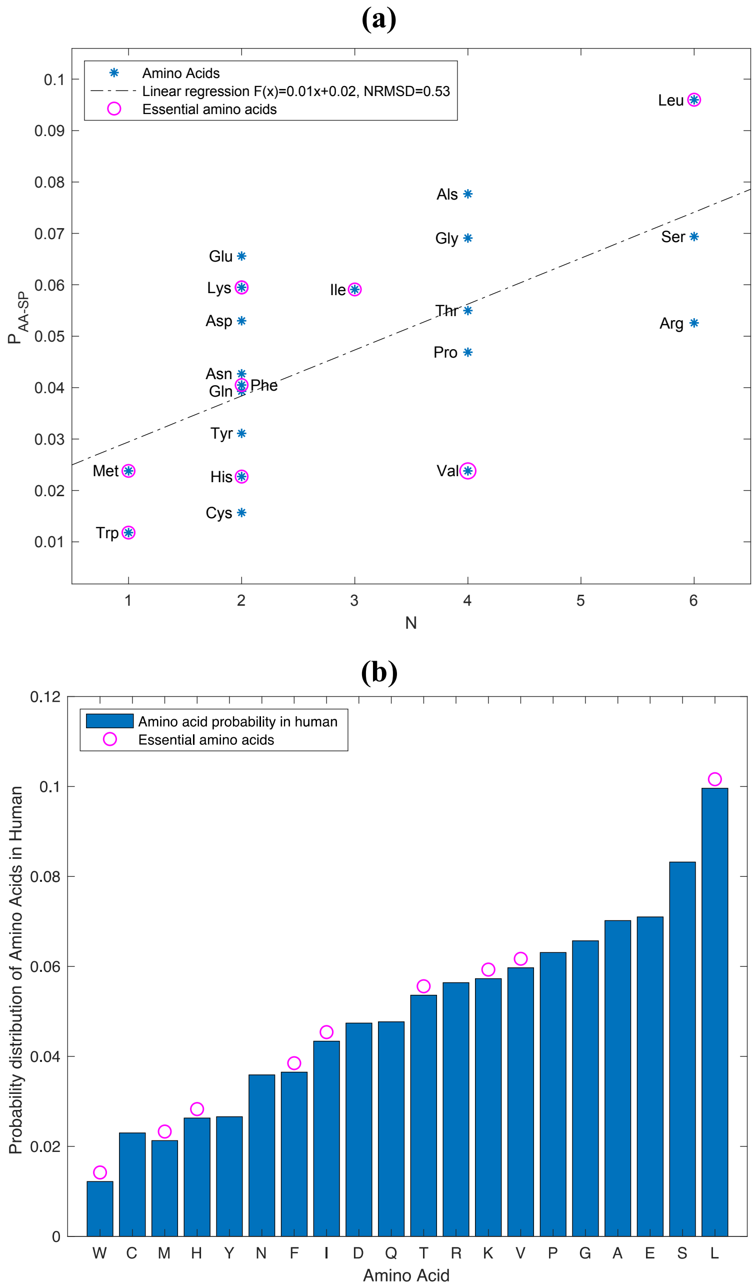


Figure 2.6 (a) Amino acid probability of occurrence $P_{\mathrm{AA-SP}}$, obtained from Swiss-Prot database for human versus the amino acid redundancy number, N, (the number of coons code for each amino acid). The linear regression with the equation of $y = 0.01x + 0.02$ with $NRMSD = 0.53$ is illustrated with a dashed line. **(b)** Amino acid probability distribution of human sorted with an increasing order of redundancy. In both plots the essential amino acids are circled in pink.

`

Next, we delve into the question of whether the frequencies of amino acids are a product of natural selection or a random permutation of the genetic code. There are two main hypotheses in this regard, called the Darwinian and non-Darwinian models, respectively [4,19,23,91,96,137–139]. In the Darwinian model (natural selection), the number of amino acids does not influence the probability of occurrence and the most optimum codon will be replaced. On the other hand, in the non-Darwinian model, permutations of amino acid would be directly dependent to the genetic code and it happens based on the random mutations [19,140]. To address the question of amino acid evolution, we calculated the expected probability of amino acids and codons using the frequencies of the DNA bases in the sequences for different species. In our dataset containing sequences for different species, the probability of adenine is 25.85%, cytosine is 25.7%, guanine is 25.92% and thymine 22.53% (only one strand has been considered). Knowing the probability of occurrence for A, C, T and G, the codon's expected frequency can be calculated multiplying each nucleotide frequency and then summing up, for example, for histidine, CAT and CAC, the random expectation value for frequency would be $[(0.25 \times 0.25 \times 0.22) + (0.25 \times 0.25 \times 0.25)] = 0.032$. In Figure 2.7, the observed probability of amino acids, obtained by averaging of the amino acid probability over the species, is plotted in terms of the expected codon probability. It may be concluded that the compared characteristics are correlating well and the outliers are within the standard deviation. In other words, the average composition among amino acids shows that the amino acids are a reflection of the genetic code in a passive manner. This means that the number of triplets coding for an amino acid will determine the frequency of the amino acid instead of having an optimal amino acid. This offers another support for the idea that the more codons exist for a single amino acid, the higher its probability of occurrence [90].

`

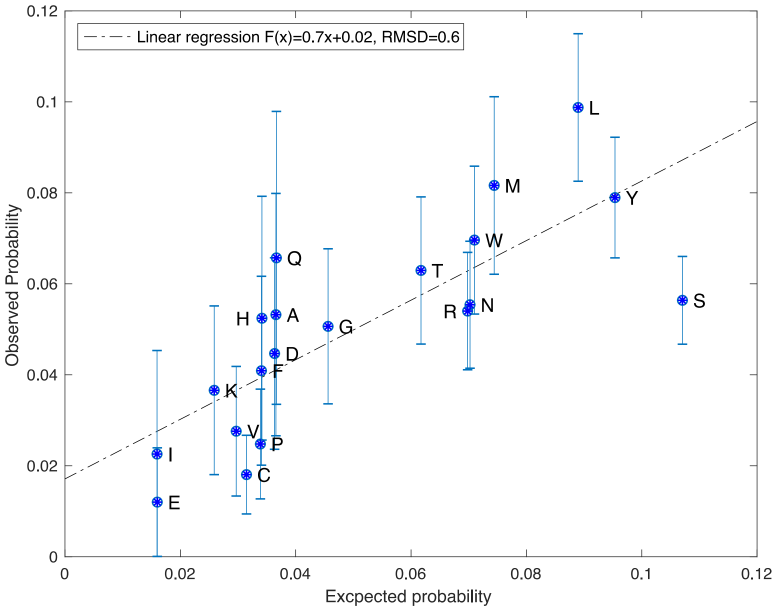


Figure 2.7 Amino acid observed probability over the species versus the expected probability of amino acids. The linear regression of $F(x) = 0.7x + 0.02$ with $RMSD = 0.6$ is demonstrated in the dashed line. Blue lines represent the standard deviation on the observed occurrence frequency for each amino acid.

2.4.6 Entropy of amino acid in different species

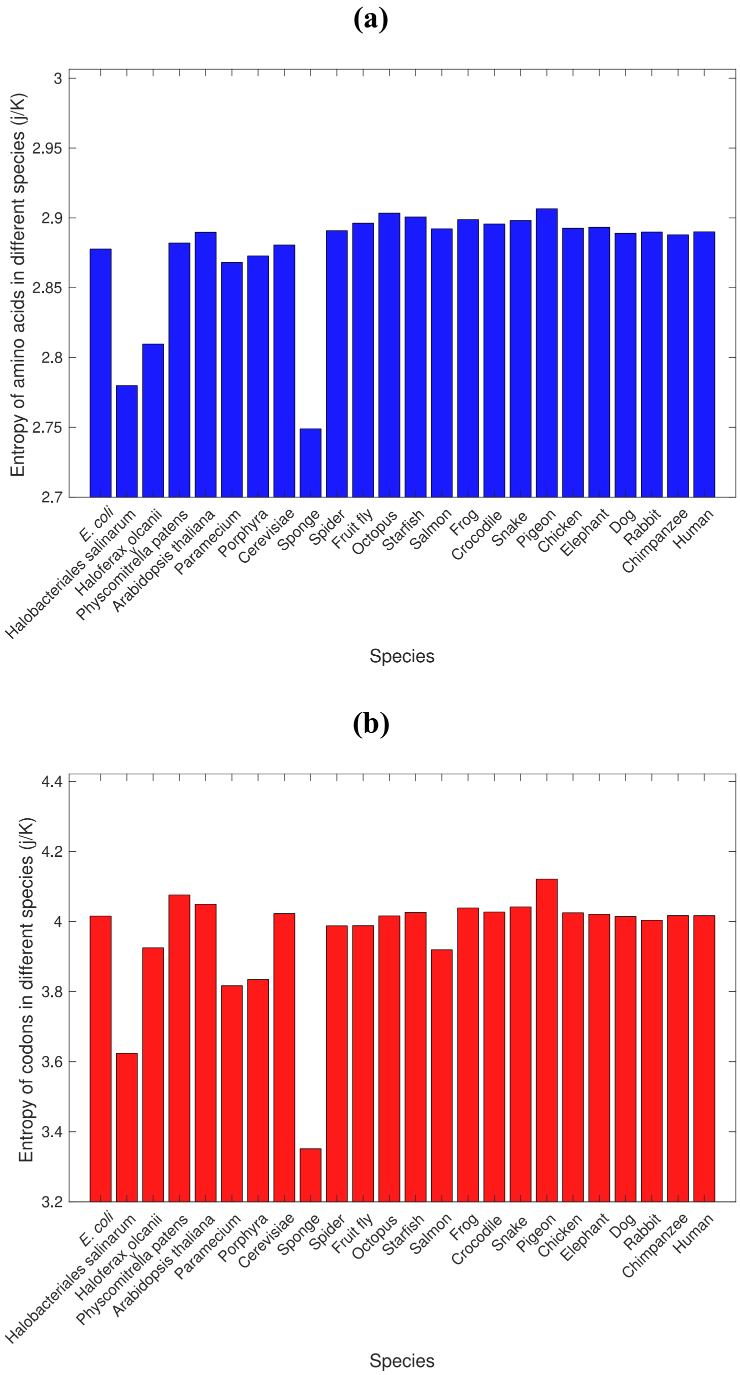
It is also interesting to consider the amino acid and codon entropy in different species. Using the described database for different species obtained from NCBI database [106,107] (section 2.4.4), the entropy values for amino acids and codons, respectively, are calculated based on the Boltzmann formula for a probability distribution in the following form

$$S_{\alpha\_AA} = -\kappa_B \sum_{i=1}^{20} P_i lnP_i \qquad (2.5)$$

where $i = 1,..,20$ represents the entropy for twenty amino acids for all 18 studied species according to the phylogenetic tree of life ($\kappa_B$considered to be 1). Figures 2.8.a and 2.8.b show the entropy of amino acids and codons in different species according to an evolutionary order. We expected to see an evolutionary trend, for example, an increasing trend in the entropy of E.coli, a simple bacterium to the human, the most complex organism. Comparing the entropy of human with E.coli, we can see there is a slight increase in the entropy but there can be found some species in between whose entropy values are even higher than that for the human. It should be stated here that there is no unambiguous quantitative measure of a given organism's rank on the evolutionary scale. However, we have selected those species that span a wide range of evolutionary advancement so their relative position on an approximate evolutionary scale can be located using general arguments such that bacteria are less advanced than multi-cellular organisms and humans are more advanced than rodents, for example.

It seems that evolution is a multiscale process and amino acid formation is its lowest scale. Our interpretation is that formation of multicellular organism is at a higher scale and it creates greater entropy by making differentiated cells. At a yet higher scale organs are formed with their own entropy generation. Hence, evolution at the level of amino acid formation did not necessarily progress in a linear fashion since it moves to multi-cellular structure organization and then on to an organ formation level. Despite an initial increase in the entropy, the process of evolution involves formation of more complex structures out of amino acids and other biomolecules later on. Therefore, we see that amino acid entropy does not have a monotonically increasing trend from E. coli to human. However, the difference in these entropies are very small in magnitude.

`



(a)

(b)

Entropy of amino acids in different species (j/K)
3
2.95
2.9
2.85
2.8
2.75
2.7
E. coli
Halobacteriales salinarum
Haloferax olcanii
Physcomitrella patens
Arabidopsis thaliana
Paramecium
Porphyra
Cerevisiae
Sponge
Spider
Fruit fly
Octopus
Starfish
Salmon
Frog
Crocodile
Snake
Pigeon
Chicken
Elephant
Dog
Rabbit
Chimpanzee
Human
Species
Entropy of codons in different species (j/K)
4.4
4.2
4
3.8
3.6
3.4
3.2
E. coli
Halobacteriales salinarum
Haloferax olcanii
Physcomitrella patens
Arabidopsis thaliana
Paramecium
Porphyra
Cerevisiae
Sponge
Spider
Fruit fly
Octopus
Starfish
Salmon
Frog
Crocodile
Snake
Pigeon
Chicken
Elephant
Dog
Rabbit
Chimpanzee
Human
Species

`

(c)

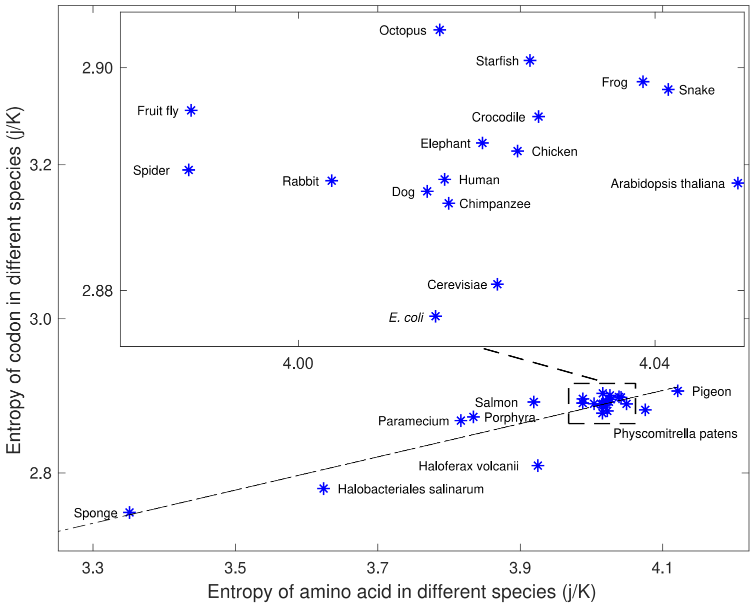


Figure 2.8 Amino acid entropy **(a)** and codon entropy **(b)** calculated for different species using Eq. (2.5) where the Boltzmann factor is assumed to be 1. **(c)** Entropy of codons in terms of entropy of amino acids in different species. The species in the range from 3.98 to 4.05 are shown as an inset plot on the top. Each species is indicated by a blue star. The dashed line shows the linear regression with the equation of $F(x) = 0.21x + 2.02$ and an acceptable $NRMSD = 0.11$. This shows that by increasing the amino acid entropy the codon entropy is also increased.

Figure 2.8.c represents the entropy of amino acids as a function of codons for all the studied species. Although there are some outliers, there can be seen an increasing trend in the entropy level. The data points within the interval $x = [3.98 - 4.05]$ are not distinguishable due to the dense distribution of data points. Hence, they are plotted as an inset figure determined by a box. In Ref. [141], the authors studied the entropy of a living cell, especially the obtained information on assembling and protecting a living state of a cell. A living cell counts as an open system that is far from an equilibrium state and each cell exchanges material and heat with the environmet

`

cyclically [142]. This process involves absorption of high energy molecules by cells through their membranes in order to produce metabolic energy for maintaining the temperature in the cell and synthesizing the various cellular components [141,143]. Similar to the protein-synthesis process, in which genetic information contained in DNA is transcribed into RNA and translated into amino acids, the difference between the entropy of a protein and the corresponding entropy of the gene that codes for it is negative, it is expected that the entropy of a codon and the corresponding amino acid in different species follows the same pattern. In the former case (gene to protein), that entropy is reduced due to the relation between the entropy and the molar heat of reaction,

$$Q = T\Delta S \tag{2.6}$$

heat would be $Q < 0$ which would seemingly contradict the second law of thermodynamics. However, this process is a controlled process that consumes energy, much the same as is the case in the temperature and entropy decrease in a refrigerator that is balanced by work performed by the engine. In a living system, this energy comes from ATP and GTP molecules. To be specific, in protein synthesis, in order to form the aminoacyl-tRNA ester linkage, one ATP molecule is required for the transfer of RNA (tRNA) (another ATP molecule is needed to drive the reaction forward), one GTP molecule for tRNA binding to the ribosome, and another GTP molecule as input for translocation. Therefore, three high energy molecules can be estimated as used in the protein synthesis for each amino acid. Note that here we are only interested in a simple estimate of the energy needed for protein synthesis [141,143] in which the additional costs of energy in the DNA transcription are excluded. We will compare the energy of ATP molecules needed for protein synthesis with the energy obtained using equation (2.6) for each amino acid. Using equation (2.5) and (2.6) the heat of reaction can be evaluated at 37 °C for *E. coli* and human to give $Q_{Ecoli} = -2.704$ kJ/mol and $Q_{Human} = -2.91$ kJ/mol, respectively. According to Ref. [141], the energy required for protein synthesis can be estimated as the energy of creating four high energy phosphate bonds per every amino acid. The energy of ATP hydrolysis into ADP can be estimated at around 30 kJ/mol, and the energy of GDP hydrolysis is also comparable to the ADP hydrolysis energy; however, it is highly substrate-dependent [143–151]. Hence, to add an amino acid to a peptide sequence at least ~ 90 kJ/mol of work is required by the cell in protein synthesis. Comparing this amount of energy with Q_{Ecoli} and Q_{human}, it is clear that this value is higher by almost two orders of magnitude and hence more than sufficient to balance the associated entropy reduction. Furthermore, the protein folding process causes an additional entropy reduction of the system which could account for additional bridging of the energy balance [152–154]. Another aspect worth considering is the change in the translational entropy of water molecules interacting with the protein's surfaces and the surfaces of DNA molecules. These water molecules attracted to hydrophilic surfaces experience a loss of some degrees of freedom, particularly translational but also some rotational degrees. Hence, by including the above mentioned effects and modifying the entropy and the calculated energy of the cellular machinery, the entropy reduction paradox can be resolved by the first law of thermodynamics, i.e. the energy conservation law [153,155].

2.4.7 Amino acids probability of occurrence in body tissues

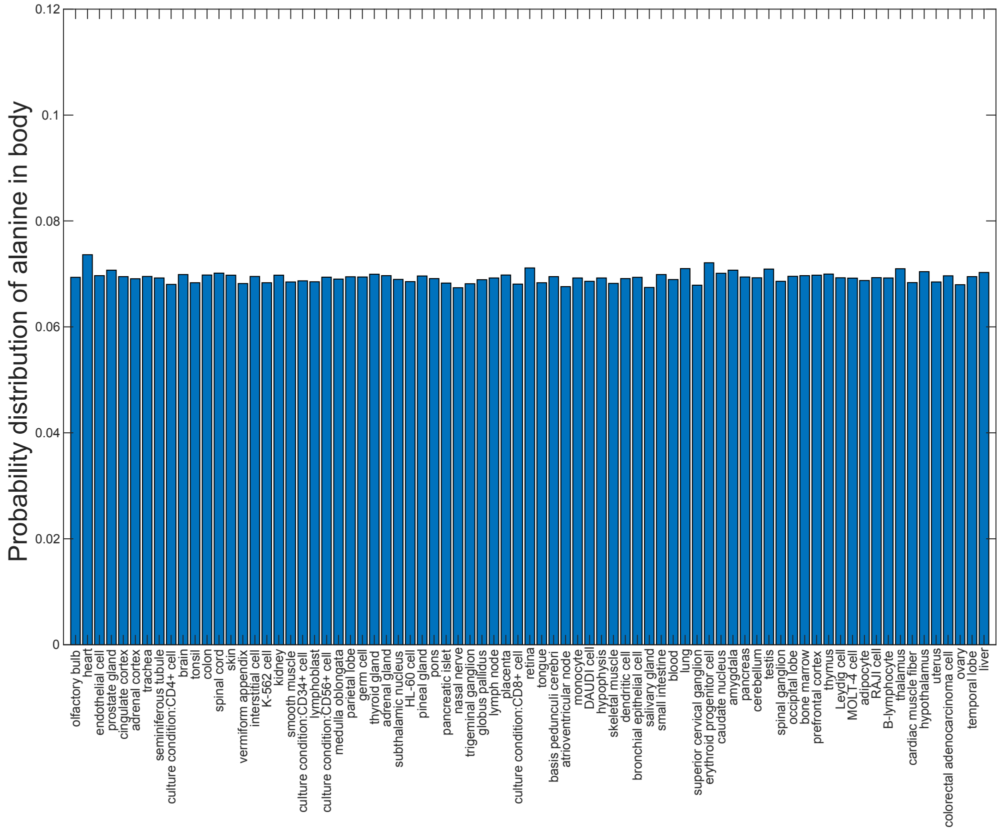
It is also interesting to consider the probability distribution of amino acids across different body tissues and also check the consistency of amino acid distributions in different organisms. Using the BioGPS databank, sequences of various body tissues, such as heart, brain, kidney, and etc., were extracted and the corresponding probabilities of amino acids have been calculated [156,156–

`

163]. Interestingly, the abundance of amino acids in different human body tissues is compatible with the human probability obtained in section 2.4.4. In Figure 2.9.a, the probability distribution of alanine is plotted for the studied cell types. It is clear that the occurrence probability of alanine is consistent among the human body tissues. The detailed list of the cell types as well as the twenty amino acid probability distribution plots can be found in Table A.1 in Appendix A.3. In addition, the average probabilities over body tissue in different cell types have been plotted and compared with the amino acid probability in the human in Figure 2.9.b. The results are fairly consistent except for arginine, which shows a slight difference between the body tissues and human average (variations less than 1% are considered to be acceptable). To conclude, amino acid probability distributions across the human body tissues are consistent with the overall probability of amino acids in human. Figure 2.9.c shows the calculated Shannon entropy for the studied body tissues and it demonstrates a high consistency of entropy among different body tissues [164]. In addition, in Figure 2.9.d the Shannon entropy of amino acid is illustrated for the studied body tissues.

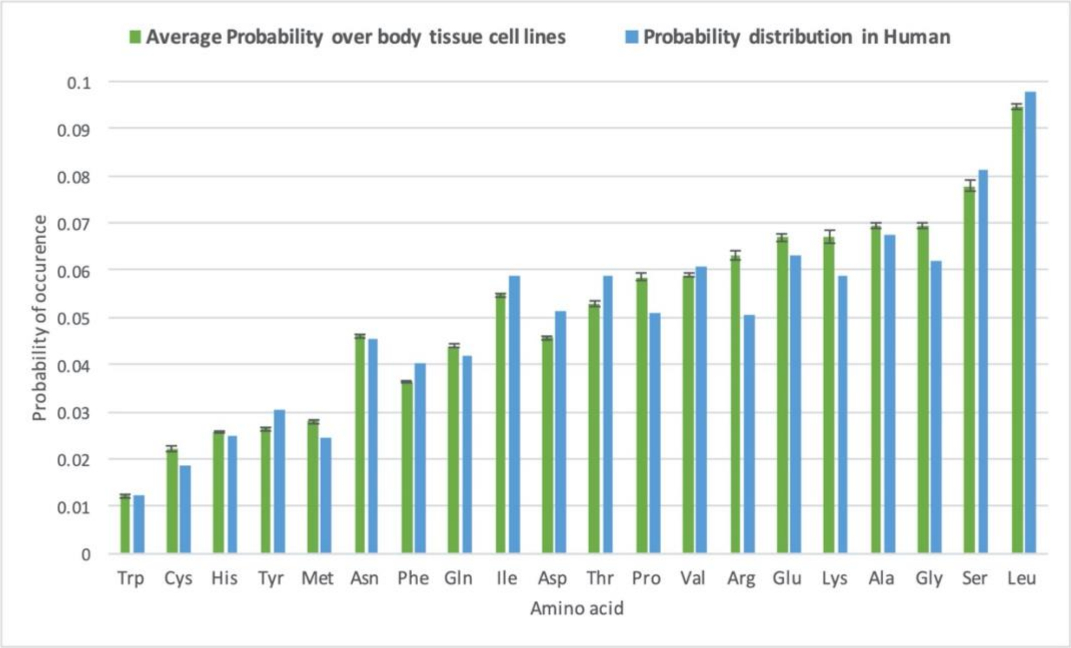
`

(a)



Probability distribution of alanine in body
0
0.02
0.04
0.06
0.08
0.1
0.12
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver

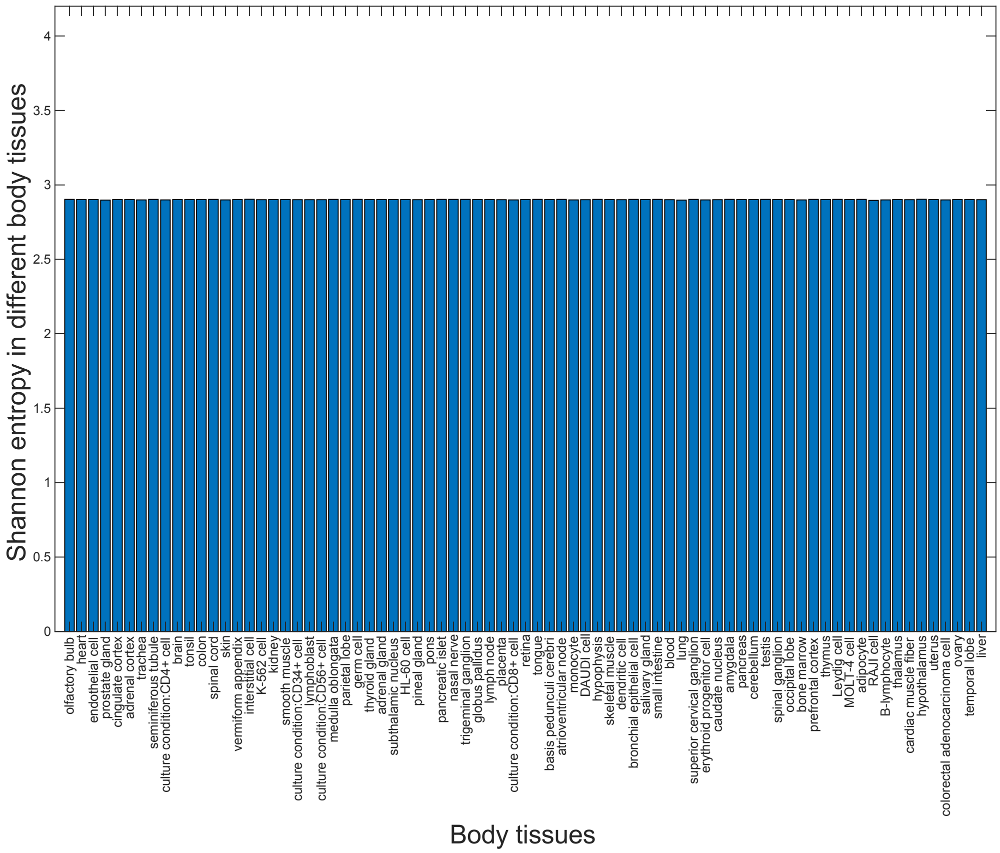
(b)



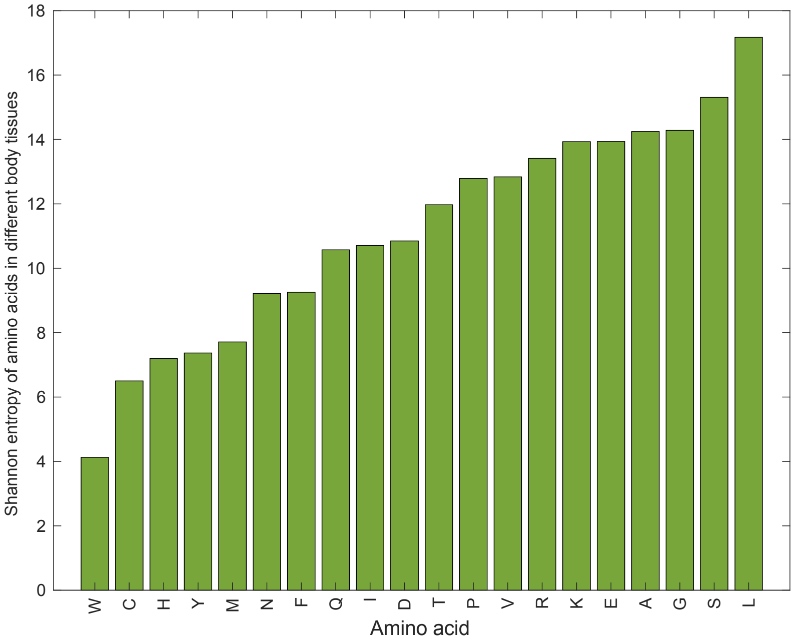
Average Probability over body tissue cell lines
Probability distribution in Human
Probability of occurence
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
Trp
Cys
His
Tyr
Met
Asn
Phe
Gln
Ile
Asp
Thr
Pro
Val
Arg
Glu
Lys
Ala
Gly
Ser
Leu
Amino acid

`

(c)


Shannon entropy in different body tissues
4
3.5
3
2.5
2
1.5
1
0.5
0
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissues

(d)


Shannon entropy of amino acids in different body tissues
18
16
14
12
10
8
6
4
2
0
W
C
H
Y
M
N
F
Q
I
D
T
P
V
R
K
E
A
G
S
L
Amino acid

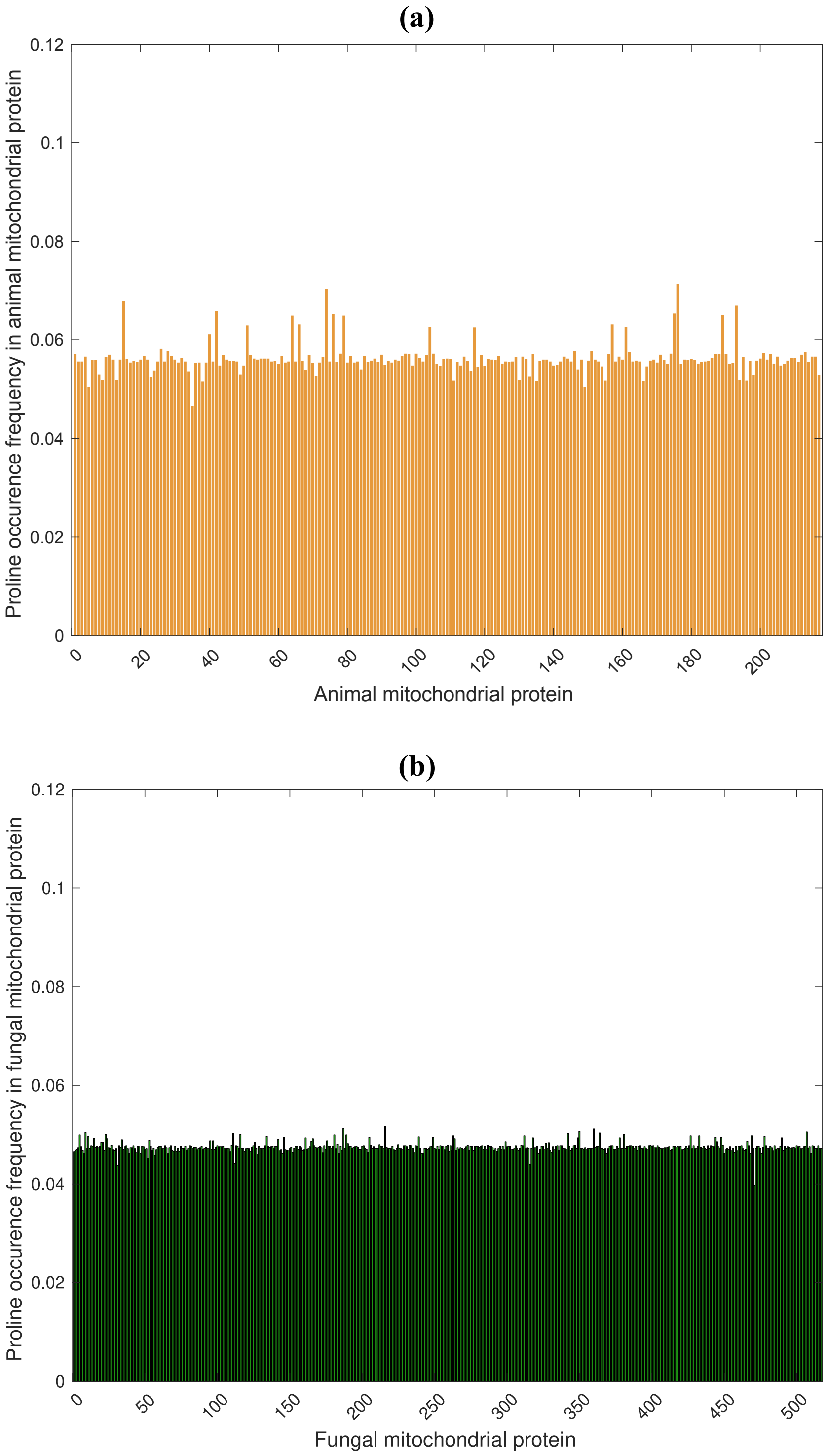
`

Figure 2.9 (a) Probability distribution alanine in different body tissues (listed in Table A.1 in the Appendix) the average value and the standard deviation for alanine is 0.07 ± 0.001, **(b)** comparing the average probability distribution of all 20 amino acids over the body tissues including the error bar (in green bars) with the amino acid occurrence frequency in human obtained from Swiss-Prot database $P_{\mathrm{AA-SP}}$ (blue bars), **(c)** the Shannon entropy of all amino acids for different studied body tissues is shown to be highly conserved across the body tissues, and, **(d)** summation over the Shannon entropy for all body tissues based on different amino acid concentrations shown in an increasing order. Leucine has the highest entropy 17.17 ± 0.001 and tryptophan has the lowest entropy 4.12 ± 0.002 and the highest and the lowest standard deviations are respectively for histidine ± 0.001 and lysine ± 0.0046.

2.4.8 Amino acids probability of occurrence in mitochondrial protein

Since mitochondrial DNA codes for different proteins, it is interesting to investigate if there are any variations in the probability distributions resulting from these two types of DNA sequences. Hence, we have next analyzed the occurrence frequency of mitochondrial proteins for all twenty amino acids. Based on the compiled data (received courtesy of Dr. Eric. A. Schon, Columbia University; personal communication) about gene products found in fungi and animal mitochondria, an attempt was made to obtain all the genes available in either animals (213 proteins) or fungi (516 proteins), from NCBI database [106, 107] (see Table A.2 for the name of the genes). As an example, Figures 2.10.a, and b represent the occurrence frequency of animal and fungi mitochondrial proteins for proline. Within a good approximation, the frequency of the proline is constant over all the selected proteins. In Figure 2.10.c, the average overall occurrence frequency of different proteins is illustrated for each amino acid for animal (orange) and fungi (dark green) cases. The mitochondrial proteins for animal and fungi cases are reasonably consistent. In Figures A.4 and A.5 in the Appendix, the plots of mitochondrial protein for animal and fungal is provided for all the amino acids.

`


(a)
Proline occurence frequency in animal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
20
40
60
80
100
120
140
160
180
200
Animal mitochondrial protein
(b)
Proline occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein

`

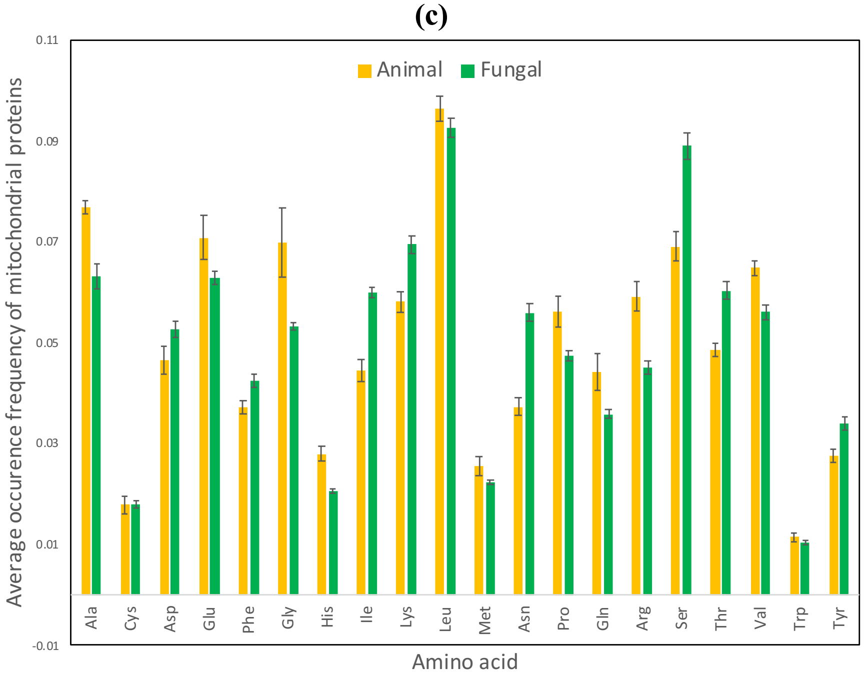


Figure 2.10 (a) Occurrence frequency of 213 animal mitochondrial proteins (listed in Table A.2) for proline (standard deviation ±0.0009), **(b)** occurrence frequency of 516 fungal mitochondrial proteins for proline (listed in Table A.2) (standard deviation ±0.003), **(c)** average amino acid probability distribution for animal and fungal mitochondrial proteins is shown in orange and green respectively, the average standard deviation is ±0.001 [103].

2.4.9 Probability of codons in different species

As mentioned earlier, 64 triplets of DNA base pairs (or RNA bases) code for twenty amino acids (with three codons representing stop instructions) and each amino acid is denoted by one or more codons. Now our aim is to obtain the codon frequency for the species studied in section 2.4.6. The corresponding sequences were extracted from NCBI, a databank of biomedical and genomic information [106,107]. Due to the huge and scattered types of data, the averages for each codon over the species were calculated and plotted with the standard deviation for each codon. In Appendix A.5 Figure A.6, the plots of all 64 codons in each species are given. The consistency of the occurrence frequency of codons is illustrated in Figure 2.11. Comparing this with the result found in section 2.4.4, where leucine has the highest probability of occurrence, it can be observed that only one codon of leucine has the highest probability. This codon also has the lowest energy among all the leucine codons. Although this conclusion is valid for leucine, histidine, cysteine and asparagine, it is not possible to generalize it for all the amino acids (See Figure A.6).

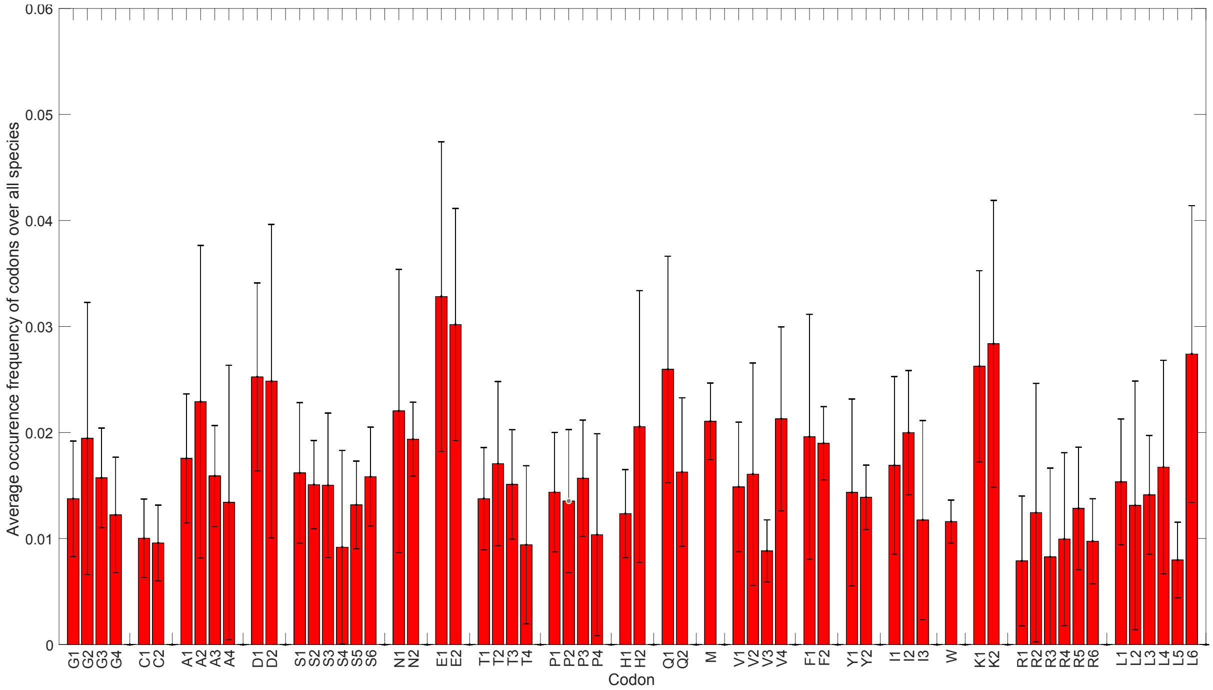


Figure 2.11 Occurrence frequency of codons in different species from *E. coli* to human among synonymous codons (normalized to the highest frequency codon values). The frequencies were obtained using the NCBI database. Each codon is represented by the one-letter representation code followed by a number (See Table A.3 for the conversion to the three letter code and orders). The corresponding codons are grouped together along the x-axis. For example, for glycine these are G1, G2, G3 and G4. Bars represents the standard deviation from the average value of probability for each codon and it ranges between ± 0.002 to ± 0.014.

`

2.5 Conclusion

In this work, we have analyzed the energy and probability of occurrence of amino acids and the corresponding codons in order to find a possible correlation between the amino acid and codon energy values and the corresponding frequency of amino acids found in human and a diverse set of species according to the evolutionary tree of life. Our results show that there seems to be no correlation between amino acid energy and the corresponding frequency of occurrence in human. However, we found that amino acids with higher degeneracy are more probable. In addition, occurrence frequencies of amino acids and codons were also studied across different human organs and body tissues as well as specifically for mitochondrial proteins in fungi and animals. The results show that amino acid distribution in different species and different organs in human are quite consistent and the fluctuations around the mean are within the standard deviation. On the other hand, there are some essential amino acids that the human body cannot produce by itself and they need to be obtained from nutrients to survive. This trend could not be explained well by the energetic or probabilistic consideration approach, however it indicates a stable attractor in the protein composition space, which can be counted as a manifestation of biochemical stability of biological systems in terms of their amino acid composition. We also examined the paradoxical aspect of entropy reduction across the species in terms of amino acid probability distributions compared to the corresponding nucleic acid distribution. By analyzing the process of transcription and translation, it can be concluded that an explanation of this paradox is given by the energetic input involving ATP and GTP molecules that are required in these processes.

`

Chapter 3 - Testing Amino Acid-Codon Affinity Hypothesis Using Molecular Docking [3]

3.1 Abstract

Genetic code refers to a set of rules that assign trinucleotides called codons to amino acids in the process of protein synthesis. Investigating the genetic code's logic and its evolutionary origin has always been both intriguing and challenging. While the correspondence rules between codons and amino acids in the genetic code are well-known, it is still unclear whether those assignments can be explained based on energetic or/and entropic arguments. As an attempt at deciphering basic thermodynamic rules governing DNA translation, we used molecular docking to investigate the ability of amino acids to bind to their corresponding anticodon compared to other codons. Based on docking scores, which are expected to correlate with binding affinity, no correlation with genetic correspondence rules was observed suggesting a more subtle process, other than direct binding, to explain codon-amino-acid specificity.

3.2 Introduction

DNA is a molecule that stores the genetic information of a living organism from generation to generation. It has a double helix structure made of two strands coiled around each other. Based on the central dogma of molecular biology, DNA is involved in protein synthesis, a two-step process, which includes transcription and translation [7,10,90,165,166]. In the transcription

[3] This chapter is partially based on a paper which is under review as: Arbabi Moghadam S., Preto J., Klobukowski M., and Tuszynski J.A., "Testing amino acid-codon affinity hypothesis using molecular docking" in BioSystems journal.

`

process, the DNA molecule is transcribed into a messenger ribonucleic acid (mRNA). The copied mRNA carries all the genetic information for the synthesis of a target protein and leaves the nucleus of the cell. With the help of the transfer ribonucleic acid (tRNA) in the cytoplasm, the mRNA directs protein synthesis in a process called translation [7,10,90,165,166]. The tRNA is responsible for bringing amino acids together in the translation process in order to make a peptide chain. Translation occurs in the ribosome, a macromolecular complex made of RNA and polypeptides, in which the small ribosomal subunits bind to mRNA and initiator tRNA, which adheres to the triplets of nucleotides (codons) and eventually induce the assembling of amino acids into polypeptide chains. Although this process are well understood, a dynamic or even thermodynamic description, specifically and quantitatively explaining the codon-amino-acid correspondences, is still lacking. There is indeed a strong motivation to study those correspondences from the point of view of energetics and structure complementarity, for example to investigate whether direct binding of those two actors together may explain such specificity. The issue of whether specific amino acids are coded by certain codons in the genetic code based on their chemical interactions was raised in earlier publications [28,167]. The stereochemical hypothesis postulates that assigning a codon to an amino acid involves a stereochemical basis [28]. There are several experimental studies on whether the amino acids bind to their cognate anticodons or not for some specific amino acids such as tryptophan, isoleucine, histidine, tyrosine, arginine and phenylalanine but not for all [10,28,168−175]. However, the question of codon-amino-acid specificity has not yet been tackled from a structural perspective.

In this study, the goal is to look at this issue using a computational structure-based approach. Since the amount of molecular biological data is increasing dramatically, computational tools to estimate and analyze molecular interactions are critical to probe protein or DNA-related

`

mechanisms. Among all the computational methods for drug design such as molecular dynamics and homology modeling, molecular docking counts as an essential tool enabling to estimate the binding affinity quickly as well as to determine the binding mode of a ligand to a protein on an atomic scale [176]. For this study, we use a combination of steered molecular dynamics (SMD) [9,177−182] and molecular docking to probe the possibility of direct interactions between amino acids and their corresponding anti-codons which could eventually be used as a molecular basis for genetic correspondence rules. However, as a result of our computations we found that there is no obvious trend that can be seen from the results of amino-acid-codon docking. Although further improvements to our model can be made, e.g. regarding receptor flexibility, the inability of docking scores to select the correct codon-amino-acid pairs consistent with the genetic code, gives some indication that direct binding between the two entities may never occur during the translation process and that other mediators in the ribosome machinery may be involved that explain genetic code assignment.

3.3 Material and methods

RNA-structures preparation. A single-stranded RNA helix was created using MOE software containing all the 64 codons of the genetic code in sequence, thus resulting in a 192-nucleotide-long RNA structure, which was further protonated at neutral pH. Due to the length of the RNA strand, the structure was split into 8 fragments with 8 codons, corresponding to 24 nucleotides each.

Steered Molecular Dynamics. Each RNA fragment was used as a starting structure for our SMD simulations. SMD was run using Amber 14 subsequent to minimization (See Appendix B.1 for more details). For minimization, 2500 steps of steepest decent followed by 5000 steps of conjugate gradient were performed. For each SMD run, the pulling force was applied to the backbone oxygen of each residue at both ends of the RNA fragment. Minimization and SMD simulations were run using Amber's multisander utility allowing to run multiple independent simulations in parallel. Simulations were performed in implicit solvent using a generalized Born model and the FF14SB force field. The temperature was set to 298K and a time step of 0.002ps was used for the SMD simulation. Regarding pulling parameters, a spring constant of 6 $\mathrm{kcal/mol/A^2}$ and a pulling speed of 100 $\text{Å}/ns$ were chosen. Figure 3.1 illustrates RNA-strand pulling with the SMD method [9,177−186], showing the structure before (helix strand) and after the pulling process (unfolded structure).

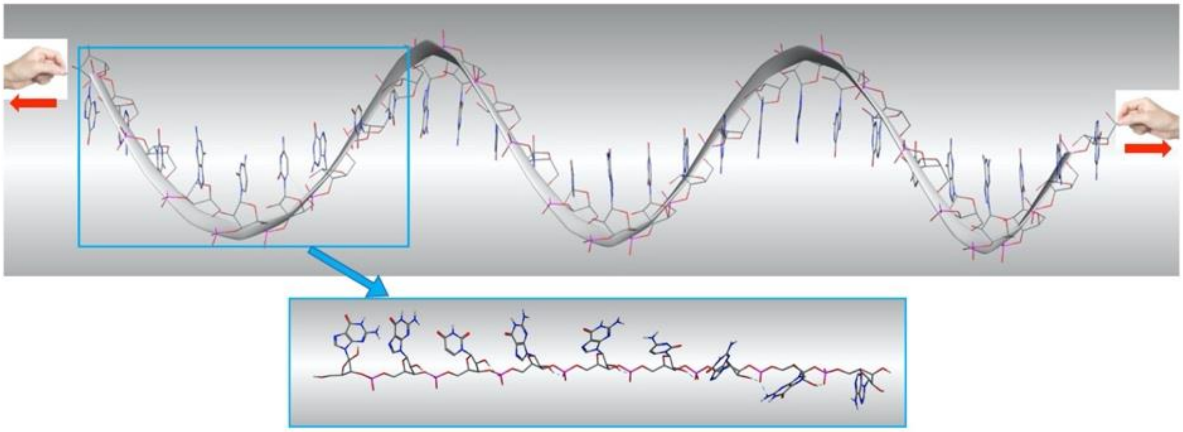


Figure 3.1 Pulling the RNA strand using SMD simulation. Top figure shows a folded RNA structure containing codons and the bottom plot represents unfolded structure for the first 9 nucleotides after performing SMD simulations.

Amino-acids preparation. Initial structures of the 20 standard amino acids were generated with MOE's Protein Builder utility. The structures were protonated at neutral pH and energy-

`

minimized using MOE. The minimized amino acids were then used as ligands in our docking simulations.

Docking. For each RNA fragment, the final structure sampled at the end of SMD simulations was used in our docking simulations. Visual inspection confirmed that each structure was completely unfolded. Since we are interested in estimating the ability of each amino acid to bind each anticodons/codon docking of each amino acid was performed separately to each codon excluding inter-codon regions. Therefore, the total number of codon-amino-acid docking simulations was $20 * 64 = 1280$. In the present study, docking was carried out *via* the 3dRPC method. 3dRPC is a new computational method that is particularly convenient to predict three-dimensional RNA-protein interactions [187–190]. 3dRPC's scoring function is a combination of two methods including a built-in force-field-based score and a FFT-based score computed from the RPDOCK algorithm [187,189,190] (See Appendix B.2 for more details).

3.4 Results and discussion

In this work, molecular docking was performed to test the hypothesis whether amino acids have a tendency to bind to their corresponding anticodons/codon from the genetic code. Consistent with the native double helical shape of DNA, RNA molecules are usually modeled as helical structures although in the ribosome machinery, the straightened structure of the mRNA is involved in the translation process. In order to better understand this situation within a cell environment, RNA helices made of the 64 existing codon types (8 RNA fragments made with 8 codons each) were first unfolded using SMD that were subsequently used as targets in docking experiments [177]. A summary of docking results is reported in Table 3.1 showing, for each amino acid, the top-scored codon, i.e. the codon that led to the best score from docking simulations (2nd

`

column). We also reported the (best-ranked) codon, which is genetically-assigned to each amino acid as well as its rank (3rd column). We expected to see that amino acids had higher binding affinities to their respective anticodons/codon in comparison to other codons. However, the docking results disprove the hypothesis, and there was no obvious trend showing that amino acids were binding to their anticodon except leucine, which has the strongest binding score to UUG, one of its codons. Several possible reasons can be discussed here. As an example, as mentioned elsewhere [10,28], glycine has a preference to bind to its anticodon but does not bind to a hairpin-bearing phenylalanine or tryptophan anticodon. In the same study, alanine was also found to prefer binding to its anticodon but not to a hairpin-bearing serine or phenylalanine anticodon [10,28]. However, our docking results did not correlate with these experimental results [10,28]. For example, there were no obvious trends in binding of serine to any of its anticodon/codons more than to the other codons. It had a higher binding score and stronger binding affinity to codons such as I2 (AUC) C2 (UGC) and L4 (CUC). Based on our results, none of the amino acids tends to bind to their anticodon except for phenylalanine, which has the highest binding score to its anticodon, which is glutamic acid (GAA). Table 3.1 also shows the number of mutual categories that each amino acid shares with the amino acid genetically-assigned to the top-scored codon (4th column). Since amino acids can be categorized in different groups based on their chemical features [90], it is tempting to investigate whether each amino acid shares common features with the amino acid genetically-assigned to the top-scored codon. Notably, amino acid groups overlap meaning that an amino acid can be assigned to more than one group. These groups include polar, aliphatic, hydrophobic, hydroxylic, aromatic and small/tiny as shown in Figure B.1 in the Appendix B. For instance, based on these characteristics, alanine is involved in three categories and the third codon of proline, P3, which has the highest binding score to alanine, is in one of the categories (See Table B.1 for amino acid/codon notation and order in the Appendix B). About 30% of amino acids do not match a category with the top-scored codon and only 15% of the amino acids assigned to the same group with the top-scored codon. Moreover, 70% of the amino acids belong to at least one mutual group with their highest ranked codon.

Table 3.1 Amino-acid-codon docking results using the 3dRPC scoring method. The table lists the top-scored codons, the best-ranked codon to an amino acid and its rank and number of mutual group highest ranked codon shared with the amino acid.

Amino acid	Codon with highest score	Codon assigned to amino acid from genetic code (rank)	Number of mutual categories/total number of categories
Alanine	P3	A2 (8)	1/3
Arginine	P3	R2 (3)	0/3
Asparagine	R2	N2 (21)	1/3
Aspartic acid	R4	D2 (8)	2/4
Cysteine	K1	C1 (12)	1/4
Glutamic acid	R5	E1 (30)	2/3
Glutamine	P3	Q1 (30)	0/2
Glycine	P3	G3 (6)	1/2
Histidine	L2	H2 (31)	1/4
Isoleucine	V1	I1 (5)	2/2
Leucine	L2	L2 (1)	3/3
Lysine	P1	K1 (43)	0/3
Methionine	P4	M (37)	0/1
Phenylalanine	E1	F2 (42)	0/2
Proline	C1	P1 (5)	1/1
Serine	I2	S3 (9)	0/2
Threonine	H2	S3 (24)	2/4
Tryptophan	K1	W (27)	1/3
Tyrosine	I2	Y1 (27)	1/3
Valine	H2	V3 (29)	1/3

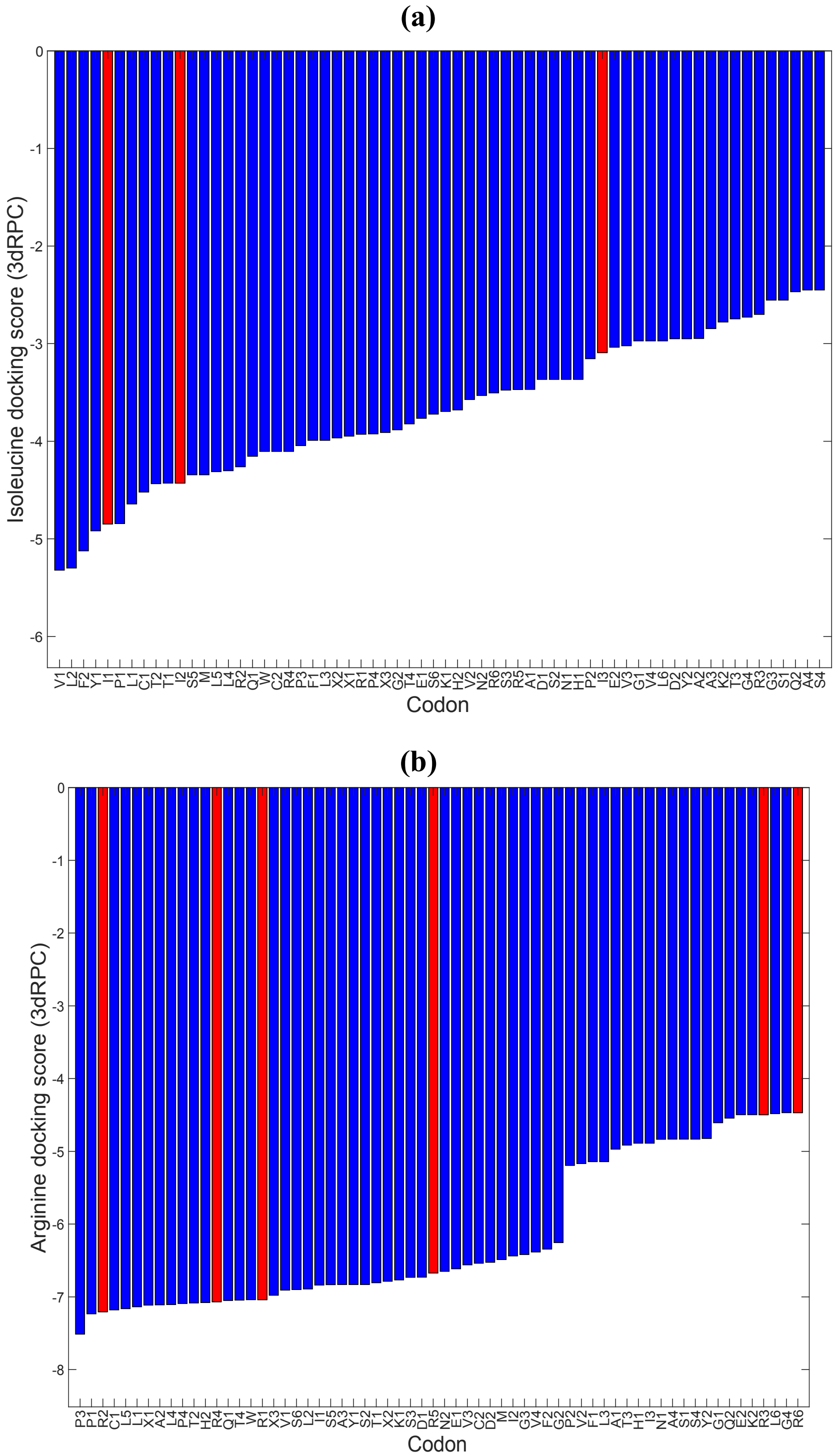
`

For further analysis, we studied the highest and the lowest binding score of each amino acid to find any correlation between the amino acids and the corresponding codons falling into the same category, based on the charge, size, and other characteristics. For instance, we investigated whether an amino acid, like alanine belonging to the hydrophobic, small and tiny groups, has a better/worse binding score than a codon in the same/different group. However, our results show that some of the amino acids bind strongly to a codon in the same group, although some other counterexamples were found. Figure 3.2 shows that isoleucine has a stronger binding affinity to the V1 codon (GUU) where they are involved in two mutual groups; they both belong to the aliphatic and hydrophobic categories. On the other hand, isoleucine has the lowest binding affinity to the S4 codon (UCG) and they are not involved in any group together. In another example, arginine has a stronger binding affinity to P3 codon (CCA), which is not similar to arginine in terms of its chemical characteristics, and has the weakest binding affinity to its codon, namely R6 (AGG). These two amino acids are examples that demonstrate that no obvious trend could be extracted from these docking results. In addition, the plot of the docking score for each amino acid to all 64 codons is shown in Figure B.2 in the Appendix B, and the codon for each amino acid is represented by different color (red).

Authors in refs. [10,28] claimed that amino acids bind selectively to their cognate anticodons in the hairpin experiment. Therefore, their results showed that a complex of four nucleotides (C4N) RNA's hairpin, with the help of aspartic acid-valine dipeptide, bound to its cognate amino acids of aminoacyl-adenylates [10]. In the present study, single amino acids (rather than dipeptide molecules) were docked onto each codon as we expect this situation to better account for the translation process in the ribosome machinery. On the other hand, codons have several single

`

bonds in their chemical structures. These single bonds create a huge conformational space for a given molecule, with various allowed dihedral angles.



`

Figure 3.2 Amino acid-codon docking scores for isoleucine **(a)** and arginine **(b)** using 3dRPC scoring method. The red bars illustrate the cognate codon that codes for the corresponding amino acids. Isoleucine anticodons are N1, D1 and Y1 while arginine anticodons are T4, A2, S4, P4, S1 and P1. This figure shows that there is no correlation found between the amino acid and the corresponding codon or anticodon.

Figure 3.3 shows different docked poses of alanine to GGA along with their docking scores. The 3dRPC docking score for the yellow, green and orange is -5.001, -4.58 and -4.46, respectively. In addition, we used another software called HDOCK for glycine docking to the 64 codons to confirm the results of the 3dRPC method. Figure B.3 shows the glycine docking scores for all 64 codons using 3dRPC and HDOCK. Needless to say, the results for each method can only be qualitatively compared, and the obtained values of the two methods are not expected to be identical. However, we were able to compare the trends. It was observed that all four codons of glycine are scattered, and no obvious trend could be observed to prove the hypothesis that an amino acid preferentially binds to its cognate anticodon/codon(s).

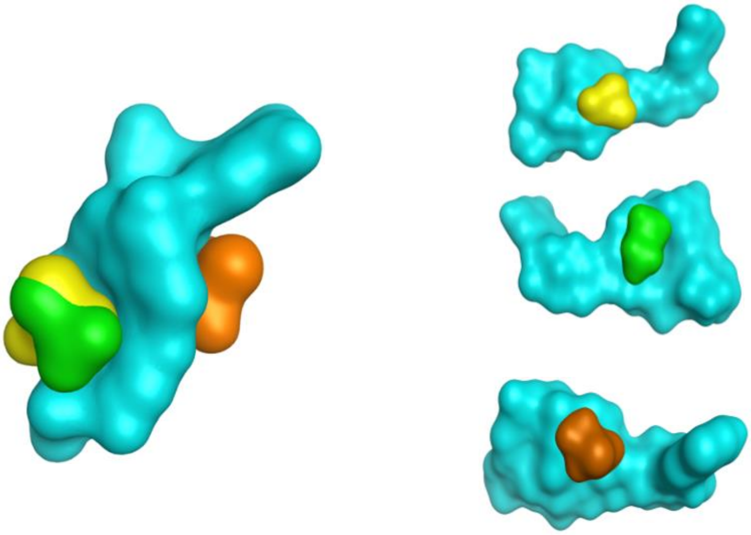


Figure 3.3 Right: Top three best docked poses of GGA to alanine with respect to the docking score. Alanine is displayed using yellow, green and orange in different poses with various

`

docking scores, -5.001, -4.58 and -4.46, respectively. The GGA codon is represented by the cyan molecular surface. Left: all three docking poses shown in one complex structure.

3.5 Conclusions

In the present work, we have reported our attempts to find a relation between the binding affinity of a codon to its corresponding amino acid in the translation process. To tackle this problem, we explored whether amino acids preferentially bind to their genetically-assigned anticodon/codons. Previous publications [10,28] claimed that glycine (alanine) preferentially bind to their anticodons and do not bind to a hairpin-bearing phenylalanine or tryptophan (serine or phenylalanine) anticodon. To investigate these findings computationally, we used molecular docking algorithms to determine amino acids' affinity for RNA codons. Using SMD, RNA structures were stretched out from a helix form to a linear strand. A total of 1280 molecular docking simulations were performed, one for every codon-amino acid pair. The docking results for each amino acid turned out to be randomly distributed between codons, and no obvious trend could be extracted. This lack of correlation might be related to the fact that we did not consider the structure for docking as a hairpin experiment, which was a simplified version of what occurs in nature within the ribosomal machinery. Another reason could be that in the hairpin experiment, a peptide chain was used rather than a single amino acid, while we considered the RNA or ligand as triplets, which may result in significant changes in the obtained results. Finally, the receptor (i.e., RNA) flexibility was not considered in our docking protocol but may improve correlation with the genetic code rules. We expect that the fairly small size of a single codon makes the conformational space of the receptor relatively fast to explore so that more accurate results can be provided in future publications.

`

Chapter 4 - Probability Distributions of p53 mutations and their Corresponding Shannon Entropies in Different Cancer Cell Types [4]

4.1 Abstract

Due to the vital role of the p53 protein mutations in about 50 percent of the human cancers, in this work, we investigate the probability distributions of different mutations in the p53 across various human cancer cells. Using the p53 database (IARC TP53), we employed statistical analysis to determine the frequency of occurrence of amino acid mutations across various cancer types. We show that amino acid hotspot mutations of p53 are highly frequent in cancers regardless of their codon location in the sequence, and at least one of the hotspot mutations has the highest probability in various cancers. We also calculated the associated Shannon entropy values for all the possible mutations in a number of cancer types and compared them to the five-year survival rate for various cancer types. We have found no evidence of correlation between mutation entropy and 5-year survival probability values.

4.2 Introduction

A permanent change in the nucleic acid sequence of a gene is known as a gene mutation. A mutation stems from an error in the DNA replication process, meiosis, mitosis, or for any other DNA damage reason. The smallest mutation happens when a single base pair (in a codon) is replaced by another base pair. In synonymous mutations, replacing a base pair does not change

[4] This chapter is partially based on a paper submitted as: Arbabi Moghadam S., Omar S. I., and Tuszynski J.A., “Probability Distributions of p53 Mutations and their Corresponding Shannon Entropies in Different Cancer Cell Types” in Theoretical Biology and Medical Modeling journal.

`

the codon that codes for the amino acid in the corresponding protein peptide sequence [191–193]. In contrast, in nonsynonymous mutations, the amino acid will be changed. Gene mutations can be attributed to two different origins, namely somatic (acquired) or hereditary (also called germline) mutations [191–193]. A somatic mutation occurs locally in a tissue or an organ, often due to some environmental factors such as UV radiation. Parents have a significant role in the former category since these hereditary mutations or germline mutations are existed in every cell of the body [52,194]. Hence, considering the importance of p53 in the pathogenesis of human cancer gives us a great motivation to study the probability of amino acids represented by p53 mutations in different cancer types in order to find any correlation between them.

TP53 codes for the tumor suppressor protein called p53 [195,196]. In human DNA, *TP53* is located on the 17th chromosome (17P13.1). *TP53* codes for over 15 various isoforms of its product protein denoted p53 [195–197]. The p53 protein is made of 393 amino acids (aa), which are divided into five main domains:

1-N-terminal transactivation domain (aa 1-43 and 44-60), which is involved in the activation of different transcription factors, binds to transcription factors and plays the role of a mediator in some interactions [198–201].

2-Pro-rich domain (aa 61-100), which is important for p53 stability and also has a function in transcription activation and induction of transcription-independent apoptosis [201–203].

3-DNA binding domain (DBD) (aa 101-300), which primarily binds to DNA. It is also responsible for binding with the p53 corepressor [201,203].

4-Tetramerization domain (aa 301-323), which plays a role in the regulation of the oligomeric state of p53 [201,204–206].

5-Basic C-terminal domain (aa 360-393), which is important for the regulation of the sequence [204,207].

The DBD of p53 is made of an immunoglobulin like β-sandwich of two antiparallel β-sheets, providing a scaffold for a flexible DNA-binding surface. This DNA-binding surface is created by two large loops stabilized by a zinc atom and a loop–sheet–helix motif [208−212]. Zinc binding is critical for correct protein folding and requires a reduction of thiol groups on cysteines [197,208−213]. In its role as a tumor suppressor protein, p53 binds to the DNA regulating the cell cycle [197,201,213]. The p53 protein controls the following cellular processes: (a) cell proliferation, (b) cell death, (c) nutrient deprivation, (d) nucleotide depletion, (e) hypoxia and oxidative stress and (f) hyperproliferative signals [201]. These and other cellular functions are performed by p53 primarily by triggering apoptosis, DNA repair, regulation of energy metabolism and anti-oxidant defense [197]. Stimuli that activate p53 include DNA damage, nutrient deprivation, nucleotide depletion, hypoxia, oxidative stress, and hyperproliferative signals [197,201,213]. The activated protein plays its role by virtue of being a transcription factor as it binds to the promoter region of different genes to activate their expression in order to induce the above-listed functions as well as cell cycle arrest when required [197,201,213].

Numerous studies show that virtually all cancer types exhibit p53 protein mutations, and several studies used computational methods, such as molecular docking, to find pharmacological compounds that are predicted to restore the function of the p53 mutant to its wild-type state [50,214−230]. It has been hypothesized that on their own, these mutations can lead to tumor initiation, and progression [50,214−225]. Due to the importance of preventing the formation of cancer in multicellular organisms and the significant role of p53 protein in conserving the cell's stability, p53 has been described as "the guardian of the genome" [213,231−240]. A vast majority

`

of the p53 mutations, approximately 95%, take place in the DNA binding domain. Interestingly, about 40% of these amino acid mutations happen in only six specific positions, known as hotspot mutations, in which the frequency of the hotspot mutations is much higher than in other mutations. These hotspot mutations involve the following specific residue changes R175H, G245S, R248W, R249S, R273H, R282Q [52]. The most common type of mutation in cancer is mainly missense, nonsense and deletion but the pattern of mutation is different in different ethnic groups, which also depends on the geographical location [241]. Most mutations in the DBD region are missense; in contrast, outside this region, missense mutations represent only about 40%, the majority of mutations being nonsense or frameshift [49]. *TP53* mutations occur in nearly all types of cancer, such as: ovarian, esophageal, colorectal, head and neck, laryngeal and lung cancers, sarcomas, breast, brain, testicular cancer, cervical cancers malignant melanoma, and leukemia. Mutations have been found to be more abundant in advanced stages of the disease. Interestingly, it was also found that in elephants, cancer prevalence was significantly lower than expected based on extrapolation from other species, including humans, which stems partly from the number of copies of the p53 protein in elephants and humans, namely twenty copies in elephants and one in humans [238,239]. The p53 gene counts as the highly frequent mutated gene in human cancers, and more than half of the human tumors include deletions or mutations of the p53 gene bases. For instance, individuals having a single p53 gene's functional copy develop Li-Fraumeni syndrome (LFS), which leads to their predisposition to developing cancer. These rare conditions create multiple autonomous tumors in different tissues. This demonstrates the importance of studying p53 mutations and their consequences for cell division.

Using different experimental biological techniques, such as gene knockout in mice, has revealed vital information regarding the mechanisms of initiation and progression of cancer in molecular level [242–244]. When the p53 protein binds to the promoter region of the p21 gene, it activates its transcription and hence its expression. The p21 protein interacts with a cyclin-dependent kinase2 (CDK2), which is a protein normally involved in cell division [197,213,245–248]. The formation of the p21-CDK2 complex inhibits the function of the latter protein and hence progression of the cell-cycle is inhibited [197,213,245–248]. Mutations in p53 can, therefore, inhibit its transcriptional activity and hence alter its control over the cell cycle. Thus, cell division would progress without control and consequently, a tumor can form. A recent study by Baugh et al., discussed the causes behind the hotspot mutations in p53 [52,249], which were listed as: 1) the mutations in the gene alter the structure of the expressed protein, 2) in a specific DNA sequence, such as a methylated cytosine residue in a CpG dinucleotide, changing it to thymidine, causes hotspot mutations to occur at these residues, 3) environmental mutagens create specific changes in the p53 gene and 4) the altered protein causes cancer due to an allele-specific gain of function [52]. In the present work, we investigate the probability distribution of p53 mutations among the various amino acids and across a number of cancer types. Below, we discuss the methodology employed to this end.

4.3 Methodology

For this study, we extracted information regarding the probability distributions of the available p53 mutations from the IARC TP53 database (http://p53.iarc.fr/). This database has organized and gathered all the published information on the *TP53* gene variations from peer-reviewed literature on human cancers since 1989 [213]. The IARC dataset provides valuable information on *TP53* gene variations and mutations associated with each human cancer sample. This information includes *TP53* germline mutations, somatic mutations, synonymous or nonsynonymous

`

mutations, functional classifications (based on the transcriptional activity), exon numbers, and several other details. Among these categories, *TP53* somatic mutations were mainly considered in this research. Somatic mutations refer to the mutations in sporadic (as opposed to genetic) cancers reported in primary tissues, cell lines, and fluids in the body.

We are interested in finding how frequent is a specific conversion of an amino acid into another amino acid in the given gene sequence. For instance, we need to know the frequency of mutating arginine to the other 19 amino acids and compare it with other amino acids, so that this would result in a matrix of 19×19 possibilities or 20×20 including non-mutated cases. In the gene sequence, different types of mutations occur, and they are recorded in the p53 database as well. The mutation types are missense, silent, nonsense, frameshift, splice, insertions or in-frame deletions, intronic, and upstream mutations in the 5' or 3' UTR (untranslated region). In missense mutations, which are in the nonsynonymous substitution category in the genetic code, a single nucleotide is altered and the produced codon codes for a different amino acid. This type of point mutation changes the protein sequence encoded. Silent mutations are those types of point mutations in which the changed nucleotide still codes for the same amino acid, and the encoded protein remains the same. The other mutation type involves nonsense mutations that arise when a point mutation of a nucleotide is an introduction of a stop codon. In this case, this mutation in the DNA sequence leads to a premature termination of a protein [213,250]. Splice mutations refer to the mutations that delete, insert or change the number of nucleotides in the specific site at which splicing occurs during the processing of precursor messenger RNA into mature messenger RNA and are located in the two first and last intron nucleotides, which remain conserved and hence, nominated for change in splicing. Also, intronic mutations happen in introns that are located outside of the splicing site. In human cancers, approximately 90% of the mutations are missense

`

mutations, and the produced protein by these mutations is not sufficiently able to bind to the DNA sequence to regulate the transcriptional pathway of p53 [52]. Among the 189 different mutations in the trinucleotides, eight of them are referred to the codons that contain about 28% of all p53 mutations. Therefore, in our calculations, we are interested in finding the frequency of missense and silent mutations in p53 protein in different types of cancer. Using the IARC database, all wild-type to mutants of p53 in 75 different cancer types have been found. Table 4.1 shows all the human cancer types studied in the present work.

Table 4.1 Cancer types studied with respect to p53 mutations [213].

Topography	Database Total	Topography	Database Total	Topography	Database Total
Adrenal gland	65	Liver	1196	Prostate	373
Anus	5	Lung	3047	Pyriform sinus	12
Biliary tract	73	Lymph nodes	762	Rectosigm. Junct.	40
Bladder	1516	Meninges	2	Rectum	691
BONES (limbs)	53	MOUTH (floor)	94	Renal pelvis	58
BONES (other)	231	MOUTH (other)	689	Salivary gland	22
Brain	1840	Nasal cavity	190	Sinuses	219
Breast	2874	Nasopharynx	62	Skin	1052
Cervix uteri	117	Nerves	79	Small intestine	13
Colon	1144	Oropharynx	259	Soft tissues	406
Colorectum, nos	1758	Other digestive org.	3	Spinal cord	5
Corpus uteri	217	Other endocrine gl.	7	Stomach	978
Endocrine glands, nos	1	Other female gen. org.	25	Testis	29
Esophagus	1873	Other head & neck	6	Thymus	21
Eye and adnexa	29	Other male gen. org.	2	Thyroid	121
Female genital org., nos	3	Other respir. Syst.	22	TONGUE (base)	13

`

Gallbladder	110	Other sites	4	TONGUE (other)	208
Gum	81	Other urinary org.	28	Tonsil	18
Head & neck, nos	665	Ovary	2303	Unknown site	25
Heart/med/pleura	13	Palate	28	Up. Urinary tract, nos	172
Hematop. System	925	Pancreas	490	Ureter	26
Hypopharynx	183	Parotid gland	29	Urinary tract, nos	5
Kidney	147	Penis	14	Uterus	73
Larynx	437	Peritoneum	46	Vagina	3
Lip	30	Placenta	2	Vulva	108

The probability distribution of each of the mutations to other amino acids has been obtained considering the somatic mutations among all the cancer diseases listed in Table 4.1. Each mutation of the p53 protein is associated with a number of the human samples in the IARC database. The probability for each mutation from the wild-type sequence is obtained using the formula

$$P_{ij} = \frac{n_{ij}}{N} \tag{4.1}$$

where $1 < i < 20$ refers to each amino acid for all somatic mutations available in the database, n_{ij} is the frequency of missense or silent mutations involving ij amino acid pairs, N is the total number of mutations reported in the database and $\sum_{ij} P_{ij} = 1$. Figure 4.1 shows glycine to alanine mutations of p53 protein in different human cancer types extracted from the IARC database. The total number of mutations for this amino acid is 38, and they are distributed unevenly between the 17 cancer types. For instance, liver cancer has a total number of 1198 mutations, among which 8 stem from glycine to alanine mutations, which gives us the probability distribution for this specific amino acid. Similarly, using equation (4.1), the probability distribution for other amino acid mutations were extracted.

`

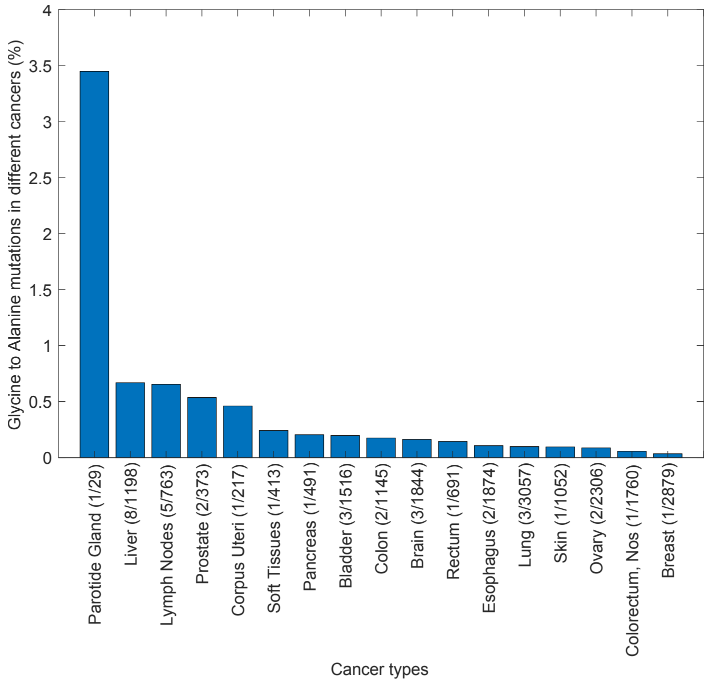


Figure 4.1 Somatic mutations of glycine to alanine in different cancers, IARC TP53 database, R20, July 2019 [213].

4.4 Results and Discussion

Having used the IARC TP53 database, the amino acid mutations of p53 protein in different human cancer have been analyzed. Similar to the example of glycine to alanine shown in Figure 4.1, all of the amino acid mutations can be presented as the elements of a matrix whose size is 20×20. In the p53 protein, some of the amino acids do not mutate as reported in the IARC database. From all 400 possible permutations, 189 cases were mutated and the rest (400-189 = 211) did not involve any mutations. The terminology "mutated" means that there is at least one mutation between two amino acid regardless of the number of repetitions in cases in (IARC TP53 database, R20, July 2019) [213]. For instance, there is some information about glycine to alanine mutations in the database and this number is 38 and it is repeated in 17 different cancer types.

`

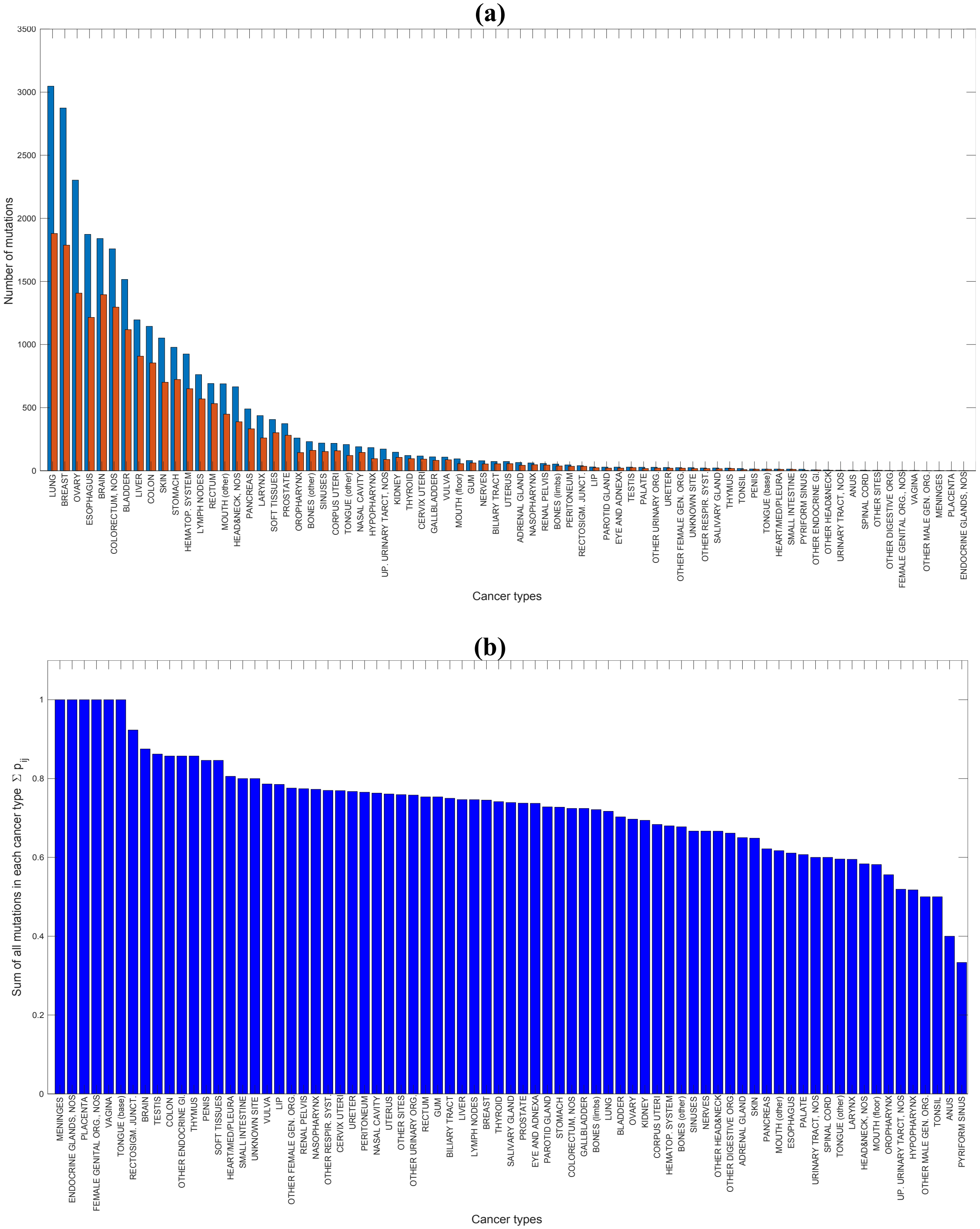
Moreover, mutations for each category were extracted based on wild-type to mutant changes in amino acids, including missense and silent mutations. Figure 4.2.a shows the number of all mutations found in different cancer types for each amino acid. In this figure, for each type of cancer, blue bars show the total number of mutations and the red bars show missense and silent mutations. Lung cancer has a total number of 3047 different reported mutations (shown by blue bars in Figure 4.2.a), which is the highest number of mutations compared to the other cancers. Among this number, 1880 are missense and silent mutations (shown by red bars in Figure 4.2.a). Other cancers, such as bladder and breast cancer are the second and third highest mutated cancers, respectively. Summation over the elements of the 20×20 mutation matrix has been calculated using equation (4.2) as:

$$P_{M,S} = \sum_{i,j=1}^{20} \left(p_{ij}\right)^{\alpha} \tag{4.2}$$

where $P_{M,S}$ refers to the sum over all missense and silent mutations, p_{ij} stands for the occurrence frequency of each mutation (i to j), and α refers to the cancer type ($1 < \alpha < 75$). The results of equation (4.2) are presented in Figure 4.2.b. Since we only focus on the missense and silent mutations, the summation over all the probabilities is not equal to 1. The rest of the contributions are for the other types of mutations (in order to compare Figure 4.2.a and 4.2.b for each cancer type, the same order is chosen for the cancer type, in the x-axis, see Appendix C Figure C.1). In addition, Figure 4.3.a and b show the summation over occurrence frequency, p_{ij}, of all cancer types in one graph. The red bars demonstrate the hot spot mutations of p53 protein, which are R175H, G245S, R248W, R249S, R273H and R282Q. Mutation of arginine to histidine, R-H,

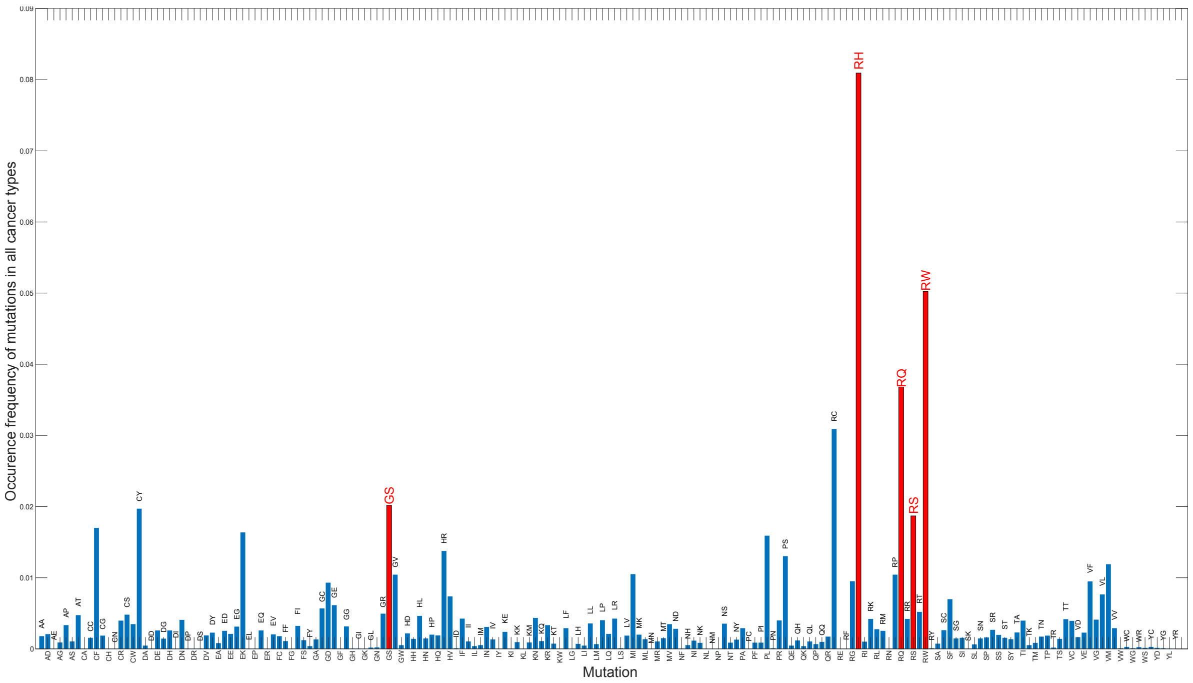
`

arginine-to-tryptophan R-W, arginine-to-glutamine R-Q, arginine-to-cysteine R-C, glycine-to-serine G-S and arginine-to-serine R-S, are the top-six highly mutated amino acid pairs.



`

Figure 4.2 (a) Total number of amino acid mutations for each cancer type in decreasing order, blue bar shows all types of mutations recorded for each cancer type in the IARC database, the red bar shows the missense and silent mutations for each cancer type. **(b)** Sum over all mutation occurrence frequencies in each cancer type as given in equation (4.2) (The summation reaches one for each cancer if other types of mutations are taken into account (not only missense and nonsense).



`

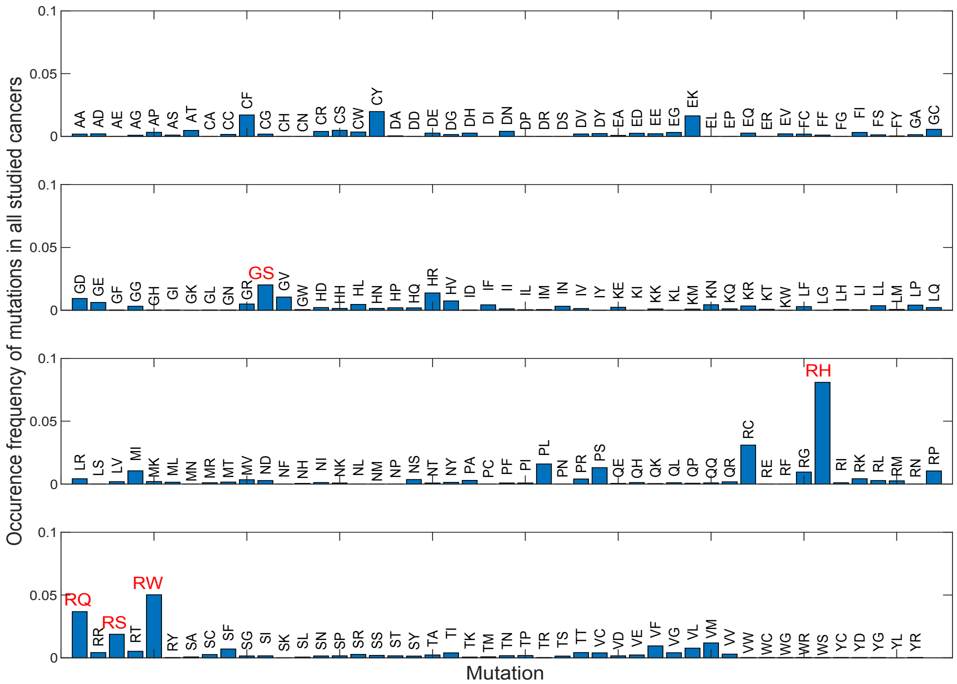


Figure 4.3 (a) Summation over all the occurrence frequencies of p53 mutations in 75 different cancers types, red bars show the occurrence frequency of p53 hotspot mutations, **(b)** to represent these large data points better, the same data are shown in 4 subplots in the same order as plot **(a)** and the hotspot mutations are labeled in red.

Among all 75 studied cancer types, 79% have at least one arginine-to-histidine mutation, 73% have at least one arginine to glutamine, 71% have arginine to tryptophan. For the next two hotspot mutations, this number drops to 55% for glycine to serine and 48% for arginine to serine. Moreover, in ~84% of the cancer types at least one of the hotspot mutations has a higher frequency compared to other mutants. Figure 4.4 shows the mutation frequency of p53 in two of the highly mutated cancer types, which are lung (a) and breast (b) cancers. In Figures C.2 in the Appendix C, a histogram of all the mutations in the different types of cancer has been plotted separately. Similarly, in most of them, the highest frequency mutations belong to one or more hotspot mutations of p53 protein.

`

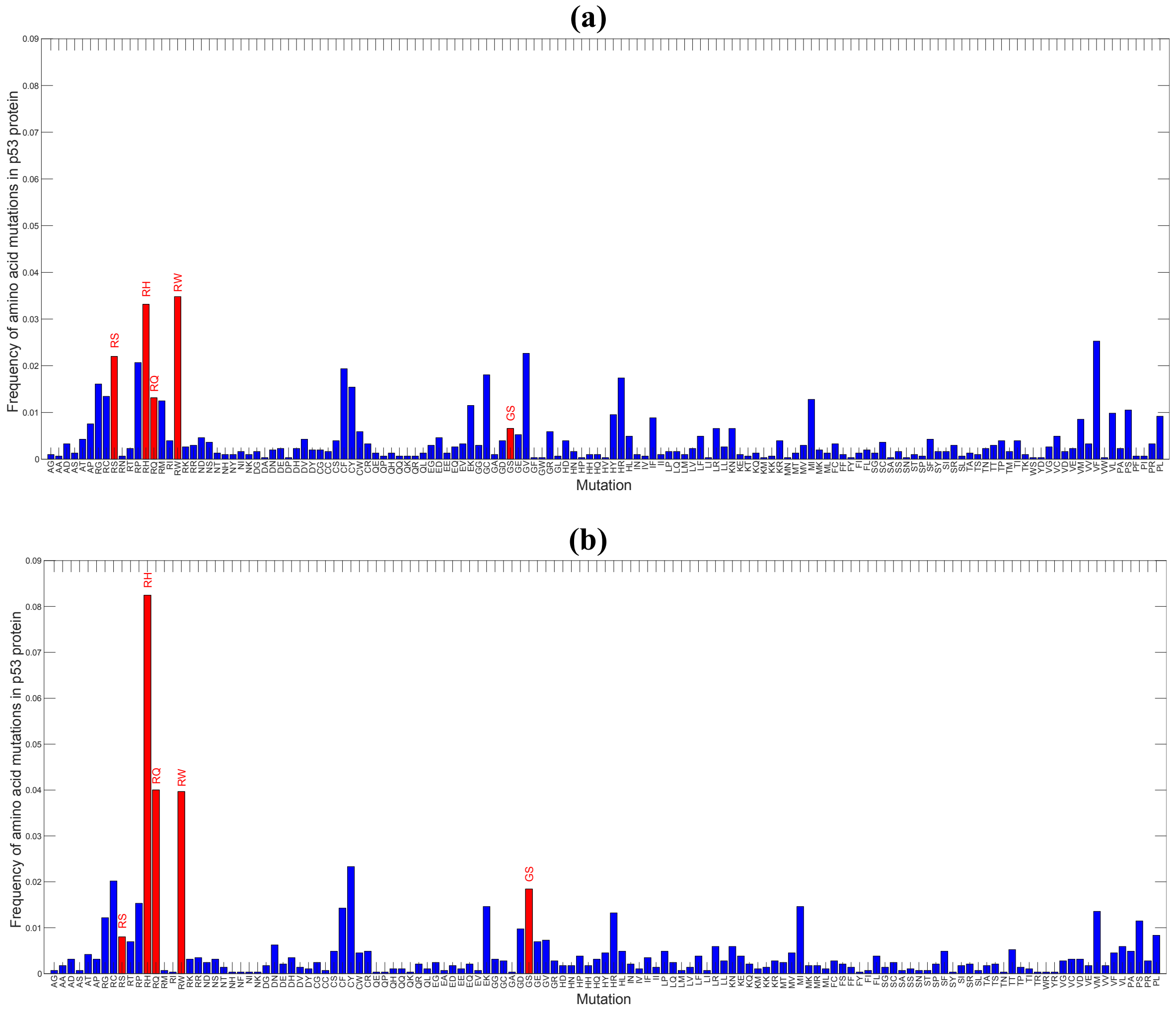
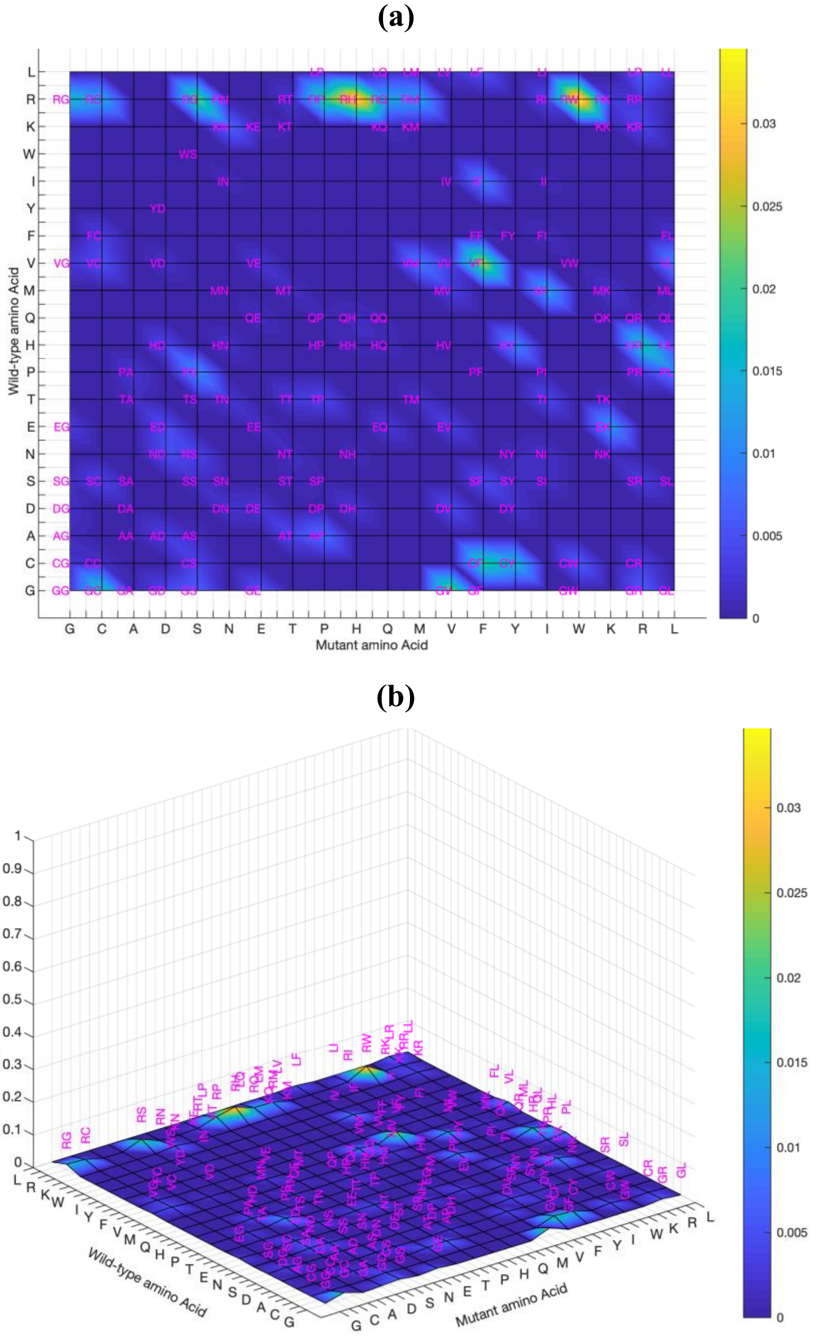


Figure 4.4 Frequency of amino acid mutations in the p53 protein in **(a)** lung and **(b)** breast cancer using the IARC database (for only missense and nonsense mutations). The red bars in both plots represent the hotspot mutations of the p53 protein. They are more frequent in these cancer types as well.

In order to produce clear and easy to understand graphs, two-dimensional (2-D), and three-dimensional (3-D) heat map representations of the amino acid mutations' frequency have been plotted. Figure 4.5.a demonstrates a 20×20 matrix with the occurrence frequency of the corresponding mutations of the p53 protein in 2-D and Figure 4.5.b is a 3-D representation, in which zero means there is no amino acid mutation in that cancer type. The results are color-coded

`

starting from blue, which means there were no mutations, to yellow as the frequency of that p53 mutation increases. As mentioned earlier, lung cancer has been reported to have the highest number of mutations among all cancers. Arginine to histidine and arginine to tryptophan are the highest frequency mutations in lung cancer.


(a)

(b)

`

Figure 4.5 The 2-D **(a)** and 3-D **(b)** plot of p53 mutation frequency in lung cancer obtained from the IARC database. The color bar changes from blue to yellow, which represents the mutation frequency from 0 to 0.04. Zero means there is no mutation from Wild-Type to that specific mutant reported in the database, and the higher the frequency of mutations, the more yellow it is represented in both **(a)** and **(b)** plots. For each mutation, the Wild-Type to mutant is represented by the first letter representation of the amino acids shown in pink.

Next, we investigate the dissimilarity factor relative to a reference number. First, we consider lung cancer, which has the highest number of reported mutations. The dissimilarity factor $\Delta^{\alpha\beta}$ is defined as

$$\Delta^{\alpha\beta} = \sqrt{\frac{\sum_{i,j=1}^{20}\left(p_{ij}^{\alpha} - p_{ij}^{\beta}\right)^2}{N}} = \sqrt{\frac{\sum_{i,j=1}^{20}\left(\delta_{ij}^{\alpha\beta}\right)^2}{N}} \tag{4.3}$$

where N is the normalization factor, α corresponds to all the other cancer types relative to cancer type β, and p_{ij} is the occurrence probability of a mutation of i to j. Equation (4.3), $\Delta^{\alpha\beta}$, varies from zero to one ($0 < \Delta^{\alpha\beta} < 1$), whereby 1 indicates that the two compared cancer type mutations are dissimilar (the higher the value, the lower the similarity). Moreover, the similarity factor can be obtained from $\Delta^{\alpha\beta\prime} = 1 - \Delta^{\alpha\beta}$. In Figure 4.6.a and 4.6.b, dissimilarity and similarity coefficients are plotted for all the 75 cancer types. It should be noted that both plots in Figure 4.6 are complementary to each other. One conclusion that can be readily drawn is that most of the cancers have similar mutations to those found in lung cancer. Furthermore, some of them, such as cancer of the endocrine glands, placenta and meninges, which have a small number of mutations, are less likely to have similar amino acid mutations to those in lung cancer and hence the similarity factor is low.

`

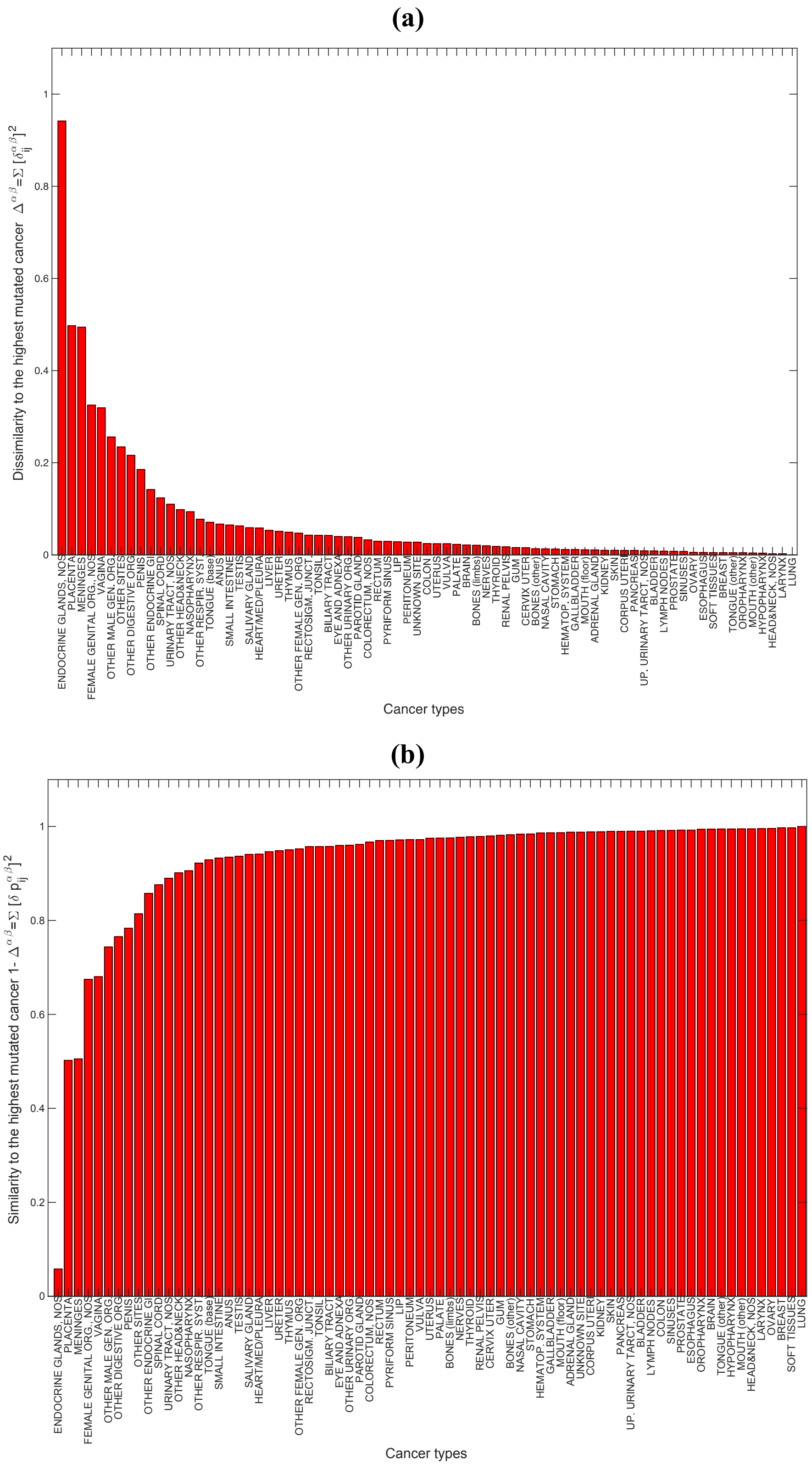


Figure 4.6 Mutation dissimilarity factors **(a)** and mutation similarity factors **(b)** obtained from Eq. (4.3) for all cancers with respect to the lung cancer (these plots are complementary to each

other). Lung cancer has been assumed to be the reference due to its highest number of mutation. The more similar to the reference cancer type, the closer the value to zero is (i.e., similarity for lung cancer to itself is zero and it means they are identical in terms of mutation types).

As a general calculation, by taking each cancer as a reference, dissimilarity and similarity factors have been obtained for all other cancer types and plotted in Figure 4.7 in the Appendix C, which shows a 75 × 75 symmetric matrix for dissimilarity factors. As explained in Equation (4.3), $\delta_{ij}^{\alpha\beta}$ represents the value of each matrix element. The diagonal elements are zero since they represent the dissimilarity of a mutation frequency of each cancer to itself, $\delta_{ij}^{\beta\beta} = 0$ and the off-diagonal elements show the dissimilarity factor between two cancer types.

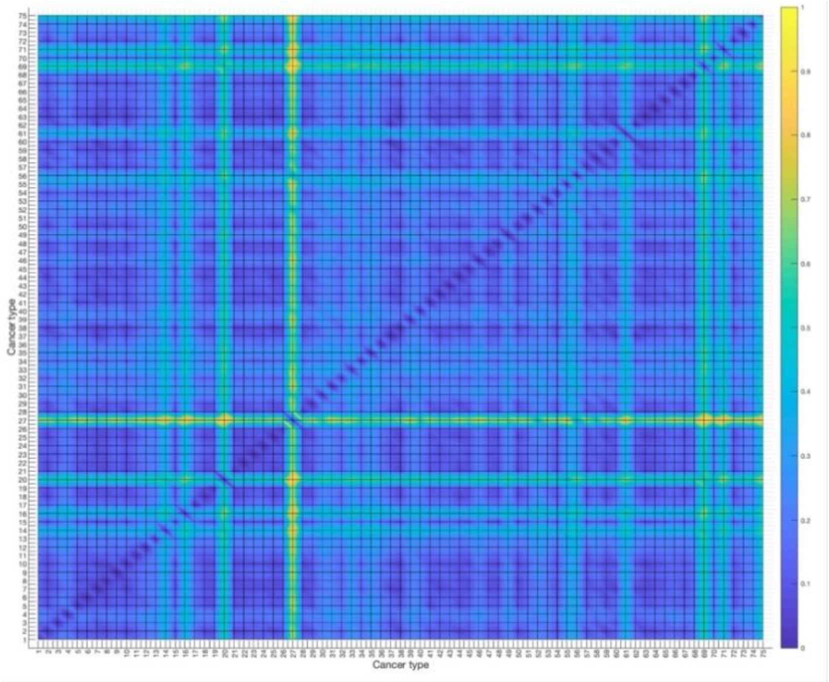


Figure 4.7 The dissimilarity factors between different cancer types. Using Eq. (4.3), a 75 × 75 matrix of dissimilarities has been obtained with respect to each cancer type, $i = j$ columns are zero since they show the dissimilarity factors of each cancer to itself, and $i \neq j$ shows the

`

dissimilarity of each cancer to the rest of the 74 cancer types. Color bar changes from blue to yellow to show the dissimilarity of each cancer to the reference cancer type.

The next stage to this approach is to consider the entropy values corresponding to the probability distribution of mutations in different cancers. Entropy is an important concept in studying cancer from the view point of information theory and statistical thermodynamics [251–259]. Based on the second principle of thermodynamics, in an isolated system, entropy always increases. Also, at the macroscopic level, entropy is a statistical measure of disorder [260]. Entropy can be computed as a system-specific entity that allows us to predict the gap between the present and estimate the final stage of a biological system based on the statistics of macroscopic characteristics of the system. The dynamics of the carcinogenesis process, which, among other processes that are dysregulated, is associated with the misplacement of internal cellular information leading to pathological transformations. It can be quantified by accumulation of genomic mutations, which can be studied by using concepts from information theory [260]. Previous studies showed that Shannon entropy is a useful concept for creating a theoretical model of carcinogenesis and prognostic models for patient survival. In this study, we apply the Shannon entropy relation to obtain the entropy of p53 mutations in different cancers [259–262]. The Shannon entropy of a system, which is characterized by a probability distribution p_{ij}^{α}, can be computed using the relation

$$S^{\alpha} = -\kappa_B \sum_{i,j}^{20} p_{ij}^{\alpha} \; ln(p_{ij}^{\alpha}) \tag{4.4}$$

where p_{ij}^{α} is the occurrence probability of an amino acid mutation and κ_B is the Boltzmann constant, $\kappa_B = 0.0083144621 \, \frac{kJ}{molK}$. With the help of equation (4.4) the entropy values for the

`

studied mutations for various cancers are obtained and plotted in Figure 4.8. As can be clearly seen, lung cancer has the highest entropy. Recall that lung cancer also has the highest number of p53 mutations. However, this trend is not seen in the rest of the cancer types. For example, bladder and ovarian cancers are the second and third ranked cancers, respectively, in terms of the number of mutations, although here they rank fourth and eleventh among the cancers ordered by their mutation entropy values.

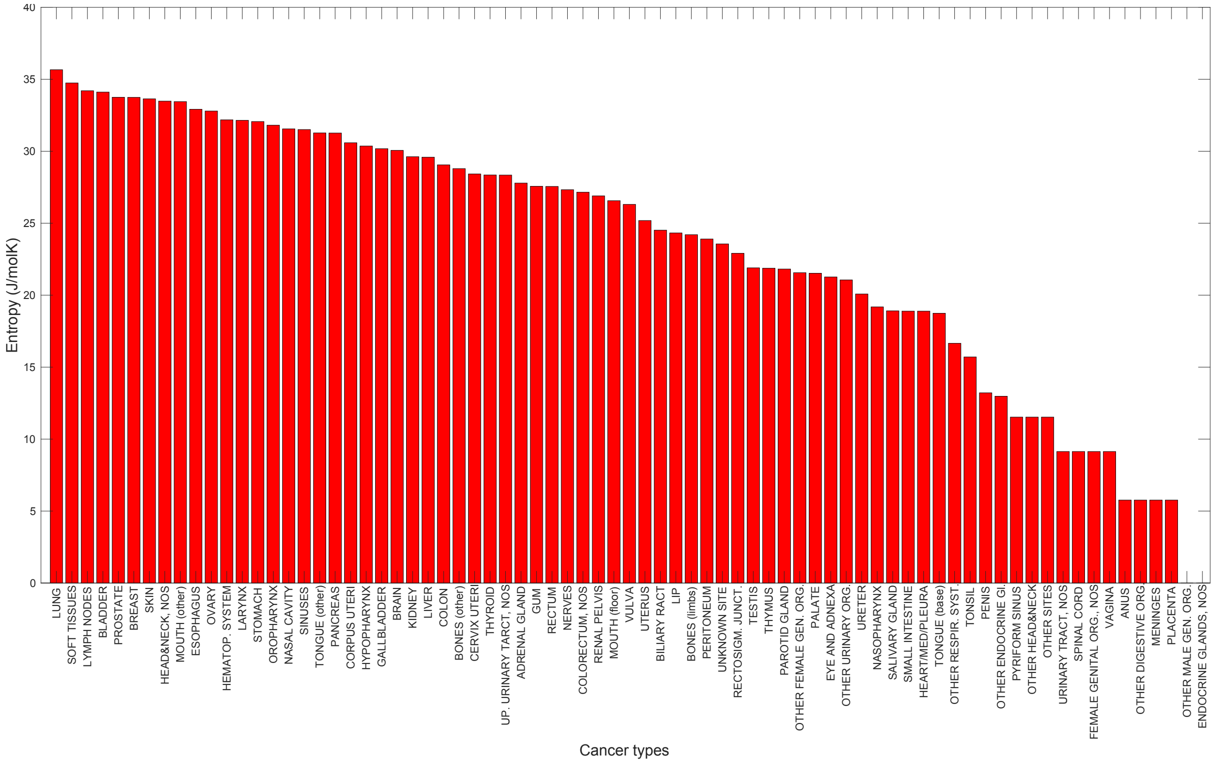


Figure 4.8 The Shannon entropy for different cancer types calculated from Eq. (4.4) (using the occurrence frequency obtained from the IARC database).

Also, it is interesting to find out whether there is a correlation between the entropy of p53 mutations for a given type of cancer and the corresponding 5-years survival rate. Therefore, the 5-year survival rates for the available cancer type can be compared using the available databases.

`

Using the statistical information provided by the Surveillance, Epidemiology, and End Results (SEER) Program, which is an authoritative source for cancer statistics located in the United States, the statistical data on 5-years survival of cancer patients were collected. There is a dedicated websites at https://seer.cancer.gov/. SEER collects and curates information on cancer cases from all around the world. The 5-years survival rates were collected based on the patient's information for the period 2009-2015. These values are obtained comparing survival rates in people who are diagnosed with cancer with those who are healthy without diagnosed cancer, having the same age, sex, and race [263,264]. Figure 4.9 shows a plot of entropy as a function of the survival rate for all the available cancer types in the SEER database. The data are scattered and only a weak correlation can be found [263,264].

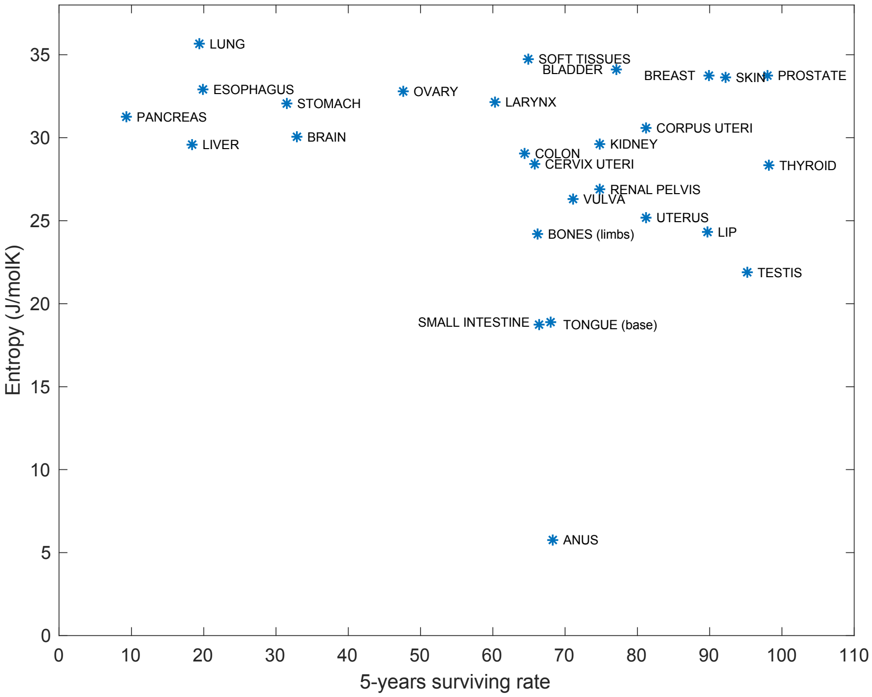


Figure 4.9 Mutation entropy of the p53 protein as a function of 5-years survival rate (using SEER database). Each of the cancer types is shown with a blue star symbols and a weak correlations can be seen.

`

As discussed earlier, if it is equally likely for any amino acid to be mutated to any other amino acid, then it is a reasonable expectation to have the same probability distribution for any amino acid mutations in every cancer type. However, our results showed that some of the amino acids are more probable than others in general. Moreover, in recent decades studies specifically focused on the p53 protein showed that there are some hotspot mutations in well-defined locations of the p53 sequence. Our results show that most of the p53 hotspot mutations have a higher occurrence frequency as well. These observations are in contradiction to the random permutation theory and indicate that the mutation location in the genetic sequence is important for a mutation to happen and it is not a random event. Therefore, there should be a correlation between the occurrence probability of a mutation and the location in the p53 protein sequence. Despite the findings in this study, mutations in p53 could still be random. However, p53 might activate cell death in abnormal cells if the mutations do not affect the wild-type activity of p53 [195,265]. If this is the case, then only cells with p53 mutations that alter the protein's tumor suppressor activity would be reported and hence bias the results showing that some mutations are more likely than others.

4.5 Conclusions

In this study, the frequency of p53 mutations of amino acid has been studied in a large number of cancer types. In terms of the number of somatic mutations, lung cancer has the highest number of such mutations. After lung cancer, breast, ovarian, esophageal, brain, and colorectal cancers have the next highest numbers of mutations. We showed that in ~84% of somatic mutations, at least one of the hotspot mutations has the highest frequency. The top-five highly mutated amino acids are; arginine-to-histidine, arginine-to-tryptophan, arginine-to-glutamine glycine-to-serine, arginine-to-serine. Moreover, the Shannon entropy of the mutations was also computed and analyzed as a possible characteristic of the associated malignancy. Lung cancer has the highest entropy value of all cancer types and also the highest number of p53 mutations. However, our results indicate there is no correlation between the entropy of p53 mutations and the number of mutations for all cancer types in general. We also examined the hypothesis that entropy may be correlated with the five-year survival rate for the available cancers types as listed in the SEER database. Except for the lung cancer, which is the most highly mutated cancer, no obvious trend could be found between the p53 mutation entropy and either the five-year patient survival rate or the occurrence frequency of mutations across all cancer types.

`

Chapter 5 – Cell Death and Survival Due to Cytotoxic Exposure Modeled as a Two-State Ising System [5]

5.1 Abstract

Cancer chemotherapy agents are assessed for their therapeutic utility primarily by their ability to cause apoptosis of cancer cells and their potency is given by an EC50 value. Chemotherapy uses both target-specific and systemic-action drugs and drug combinations to treat cancer. It is important to judiciously choose a drug type, its dosage, and schedule for optimized drug selection and administration. Consequently, the precise mathematical formulation of cancer cells response to chemotherapy may assist in the selection process. In this work, we propose a mathematical description of the cancer cell response to chemotherapeutic agent exposure based on a time-tested physical model of two-state multiple-component systems near criticality. We describe the Ising model methodology and apply it to a diverse panel of cytotoxic drugs administered against numerous cancer cell lines in a dose-response manner. The analyzed dataset was generated by the Netherlands Translational Research Center B.V.(Oncolines). This approach allows for an accurate and consistent analysis of cytotoxic agents' effects on cancer cell lines and reveals the presence or absence of the bystander effect through the interaction constant. By calculating the susceptibility function, we see the value of EC50 coinciding with the peak of this measure of the system's sensitivity to external perturbations.

[5] This chapter has been partially published as a research article as: Arbabi Moghadam S., Rezania V., and Tuszynski J. A., "Cell death and survival due to cytotoxic exposure modeled as a two-state Ising system", Royal Society of Open Science (*R. Soc. Open Sci)*, **7**: 191578*,* doi: https://doi.org/10.1098/rsos.191578].

`

5.2 Introduction

Chemotherapy is a standard cancer therapy modality based on the concept of cytotoxicity of drugs or drug combinations inflicting lethal damage to cancer cells but being less damaging to normal cells. Cytotoxicity refers to killing or damaging a viable cell by chemical compounds or pathogens. Toxic agents usually damage molecular targets such as metabolic sites, signaling proteins or DNA, which are essential to the cell's reproductive ability and survival. The result is dose-dependent cell death or inhibition of its proliferative potential. Unlike radiotherapy, which specifically targets cancer cells in a tumor, chemotherapy is typically applied systemically and is not site-specific unless combined with target-specific antibodies or *via* special drug delivery strategies. Hence, it may also affect metastasized cells distant from the primary tumor site and also the entire body of the patient with concomitant detrimental side effects [56−59]. Besides the collateral damage to healthy cells, another major shortcoming of chemotherapy is the emergence of drug resistance in the population of tumor cells. This is commonly due to the heterogeneity of tumor cells, some of which are sensitive to a given drug and others resist it. Due to the survival of the fittest, the resistant subpopulation proliferates when exposed to a cytotoxic agent while sensitive subpopulation is eradicated making the tumor more malignant over time. Consequently, we should judiciously choose a pharmacological agent, its dose and scheduling, which can be a very complex, multi-factorial problem, considering both tumor destruction and the collateral damage to the healthy tissues. Furthermore, chemotherapy-transfected cells are likely to host toxic anabolites resulting from the therapy, which can directly transfer into neighboring untransfected cells through diffusion resulting in a secondary wave of damaged cells. This refers to a so-called bystander effect extensively reported in the literature [61−71]. In general, it describes the population of dead and/or damaged cells that are not directly targeted by either chemotherapy or irradiation.

Cytotoxicity mechanisms are commonly analyzed using the Hill model representing an empirical sigmoidal fit describing the binding equilibria in ligand-receptor interactions [266] based on a simple reaction scheme: $R + nL \overset{K_d}{\leftrightarrow} RL_n$, where R is the receptor, n (called the Hill coefficient) is the number of ligands L, and K_d is the dissociation constant. Despite its simplicity, the Hill equation is not always physically plausible and the Hill coefficient n can only be accurately estimated for extremely positive cooperative interactions among multiple ligand binding sites. Even for a reaction with a high degree of positive cooperativity, e.g. binding four oxygen molecules to hemoglobin, the Hill coefficient ranges from 1.7 to 3.2 rather than 4 [266]. Therefore, other physically plausible reaction schemes such as a “two-state’’ (activated and inactivated) receptor model have been proposed to account for complex cases with various ligand-receptor cooperativities [266]. The main limitation of such models is a large number of adjustable parameters required to fit experimental data, e.g. seven parameters are needed for the hemoglobin-oxygen two-state receptor model. In addition, the observed bystander effects cannot be addressed by the above models. For these reasons, we propose a more accurate and better-motivated modeling approach to cytotoxicity, following on its successful applications in physics and other fields. To improve a statistical analysis of the effects of chemotherapeutic agents on tumor cells we adopt the concepts introduced for phase transitions and multistability. This is appropriate for cells under cytotoxic attack since cytotoxicity (and irradiation) is a dynamical process, which triggers a stochastic transition from a proliferating cell (live) to a non-proliferating cell (dead or senescent). As a result, cytotoxicity can be viewed as a transition between two different biological states present in replicas of manifestly identical systems (cancer cell cultures). Phase transitions have been successfully analyzed in many-body systems in physics, chemistry, and even social sciences and economics. Phase transitions ranging from the solid ↔ liquid ↔ gas transitions at the macroscopic level to the superconductor-metal transition at the microscopic level, have been exquisitely understood employing statistical model systems such as the Ising or the Landau model (see *SI Text* for details). Bistability is a common motif in systems undergoing phase transitions, which can exist in two distinct states and switch between them at the transition point in response to a change in the so-called control parameters [60]. The system's response is reflected in its order parameter, i.e. a macroscopic property that is zero in the disordered phase and nonzero in the ordered phase. The so-called generalized susceptibility function is the first order derivative of the order parameter with respect to the control parameter and it describes the system's sensitivity to perturbations. Generalized susceptibility of an infinitely large system diverges at the critical point (the tipping point) as the system switches from one stable phase to the other. Biological systems such as cancer cells positioned at a threshold of viability can indeed be viewed as dynamical systems at criticality, which exhibit clear bistability characteristics between being alive and dead. This perspective leads to a meaningful connection between biology and physics providing a physical model with a better mathematical insight into cancer cells' behavior [60,267–269,40].

The Ising model, first introduced to statistical physics by Wilhelm Lenz in 1920, is one of the simplest examples of dynamical systems undergoing a phase transition [270]. An Ising system must be at least two-dimensional in space, in the absence of an external magnetic field, for a spontaneous phase transition to occur, i.e., no phase transition takes place in one-dimensional Ising systems [40,60,267–269,271–274]. The Ising model is a standard mathematical model of a phase transition in a lattice of spins ½ (with only two states: +1/2 and -1/2) where each spin is allowed to interact with its neighbors [76,77]. Beneath a characteristic temperature called the

`

Curie temperature, the Ising system exhibits a ferromagnetic phase (spins are aligned along the same axis) while above this temperature a paramagnetic phase is stable where spins are disordered and no net magnetization of the sample exists. The transition from a non-magnetized state to a magnetized state depends on both temperature and the applied magnetic field's strength. Generally, the Ising model corresponds to any N-dimensional lattice whose each site is occupied by a spin with two possible states, pointing either 'up' or 'down' [59]. The mathematical notation used for the spin variable is $s_i = \pm 1/2$ where +1/2 refers to spin up and -1/2 refers to spin down. In this study, the Ising model has been applied to both interacting and non-interacting generalized spin systems [59,76–82]. We arbitrarily assign a spin-up state to a live cell while a spin-down state to a dead cell. This bystander effect [61,72–75,275] may account for the cell-cell interactions in a manner similar to spin-spin interactions between neighboring spins, hence we propose a model whereby a damaged cancer cell might affect the survival status of the neighboring cells. Due to the influence of neighboring cells on individual cell fate, the drug concentration reduces the chance of survival of other cancer cells in the neighborhood [61]. These types of interactions are distance-dependent, so the farther apart the neighboring cells are, the weaker their intercellular interactions [61,72–75,275].

Our aim is to implement the spin-1/2 Ising model of phase transitions as an elegant and powerful mathematical approach to study a cancer cells exposed to cytotoxic chemotherapeutic agents (See *SI Text* for details) [60,72–82,270–273,275–280]. Recently, a similar approach has been applied to ionizing radiation response of tumors [61]. In analogy to the Ising model for ferromagnetic materials with long-range interactions, these authors proposed to study tumor response to a uniform ionizing radiation field. In particular, using the mean-field approach individual cells are averaged out and characteristic features such as cell survival curves, tumor control probabilities,

`

fractionation and bystander effects emerge naturally showing that the bystander effects cannot be ignored at low-dose radiotherapy [61]. Below, we elaborate on the use of the Ising model to a dose-response cancer cell survival dataset provided by the Netherlands Translational Research Center B.V. (Oncolines). These experimental data correspond to inhibition profiles obtained in a uniform manner for numerous cell lines exposed to a number of chemotherapy agents [83,84].

5.3 Methods

Similarly to the Ising model with two spin states, up and down, the effect of chemotherapeutic drugs on cancer cells is associated with two possible outcomes: either survival or death states of cancer cells, respectively. For simplicity, we assign two states as $s_i = \{0,1\}$, where $s_i = 0$ refers to the live state of the ith cell and $s_i = 1$ represents the dead state of the ith cell. Following the radiation-induced bystander effect [61,62], we invoke the bystander effect for chemotherapy-exposed cancer cells in a culture by introducing a classical Ising Hamiltonian applicable for interacting spin systems as:

$$\mathcal{H} = -\sum_{ij}^{N} J_{ij} s_i s_j + \sum_{i=1}^{N} h_i (1 - 2s_i) \tag{5.1}$$

where ij indices are summed over all nearest-neighbour cell pairs at site i and j, J_{ij} denotes the strength of the interaction between neighbouring cells i and j (namely the strength of the bystander effect), and h_i represents the potency of the external agent at location of s_i. $J_{ij} = 0$ means there is no interaction between cells. Similarly to spin systems, it is expected that the interaction strength is always positive for all cells i and j, so $J_{ij} \geq 0$, and the summation $\sum_{ij}^{N} J_{ij}$ over all the

`

neighbouring cells is finite. The partition function and the probability of finding the system in state $\{s_1, s_2, \dots, s_N\}$ with $s_i = \{0,1\}$ is obtained in a standard manner as [61]

$$Z = \sum_{s_1} \sum_{s_2} \dots \sum_{s_N} e^{-\mathcal{H}(s_1,s_2,\dots,s_N)/k_B T} \tag{5.2}$$

$$P(s_1, s_2, \dots, s_N) = \frac{e^{-\mathcal{H}(s_1,s_2,\dots,s_N)/k_B T}}{Z} \tag{5.3}$$

where k_B is the Boltzmann constant and T is the absolute temperature in Kelvin. The corresponding average value of the order parameter (magnetization for spins or survival rate for cells) $M_i = \langle s_i \rangle$, is found from Eq. (5.1)-(5.3) as

$$2M_i - 1 = k_B T \frac{\partial}{\partial h_i} (\ln(Z)) \tag{5.4}$$

This order parameter in the case of cytotoxicity corresponds to the average number of surviving cells at a given concentration of the toxic agent acting on them. In general cell survival is a function of temperature, which is typically kept constant in cell-based assays. Future experiments with temperature as a variable parameter, could further test the model. However, in cancer cell system, the most commonly used control parameter affecting cell viability is the toxic agent's concentration. In what follows, non-interacting and interacting cell situations are discussed separately [61].

<u>The case of non-interacting cells</u>: In this case, the interaction strength, J_{ij} vanishes, which corresponds to the absence of the bystander effect. The partition function in the non-interacting case Z_{NI} with a constant field h, which represents the control parameter, is

`

$$Z_{NI} = (e^{-h/k_BT} + e^{h/k_BT})^N \tag{5.5}$$

The order parameter, i.e. the survival rate for cancer cells, M, is calculated as

$$N(2M-1) = k_BT\frac{\partial \ln(Z_{NI})}{\partial h} = \tanh\left(\frac{h}{k_BT}\right) \tag{5.6}$$

From the partition function, we then calculate the probability of the system occupying each of the two states. It is especially interesting to calculate the probability of the state in which all cancerous cells are dead and there would be no correlations between cells, i.e. $s_i = 1$, which is obtained using Eq. (5.5) as

$$P_{NI}(1,1,\ldots,1) = \left(\frac{e^{h/k_BT}}{e^{-h/k_BT} + e^{h/k_BT}}\right)^N = R(C)^N \tag{5.7}$$

where $R(C)$ is the death rate as a function of drug concentration, and $S(C) = 1 - R(C)$ is the survival probability of a cell. Hence, magnetization and external field in the Ising model correspond to death rate, R, and drug concentration, C, respectively, and Eq. (5.6) is written as

$$2R - 1 = \tanh\left(\frac{h}{k_BT}\right) \tag{5.8}$$

As shown below, Equation (5.8) gives very good agreement with most of the cytotoxicity assays. When h/k_BT approaches its maximum value, all cells are dead and the death rate is at a maximum. Conversely, when h/k_BT approaches its minimum value, all cells are in the survival state and the death rate is zero. Knowing the functional dependence of the survival rate on drug concentration, $S = S(C)$, allows one to determine the values of the model parameters for each chemotherapeutic agent and for each cancer cell line. For example, considering diffusion of drug molecules throughout the tumor with a heterogeneous microenvironment, one would expect the survival rate to follow the generalized Michaelis-Menten dynamics as

$$S(C) = \frac{1}{1 + \left(\frac{C}{C_M}\right)^{\alpha}} \tag{5.9}$$

where C_M is a Michaelis constant representing the drug concentration associated with reaching the half-maximal inhibition effect [281]. The cytotoxic drug concentration, which gives a half-maximal effective cytotoxic concentration, is known as EC50. Another metric related to a drug's potency is called IC50 (inhibition coefficient 50) and refers to the inhibition of a process or reaction to the level of 50% of its maximum value. The EC50 and IC50 values are both measures of drug's potency and are related and sometimes used interchangeably despite a subtle difference between them [282,283]. The parameter α shows the slope of the dose-response curve and depends on the system's heterogeneity, drug-binding efficacy and diffusion of drug molecules. It can be estimated by fitting the solution to experimental data points. In cytotoxicity assays, C_M is usually denoted as EC50. Equating Eqs. (5.7) and (5.9), we find that the control parameter, h/k_BT, is associated with the logarithm of the anti-cancer drug concentration according to

$$h = \frac{\alpha k_B T}{2} \ln\left(\frac{C}{C_M}\right). \tag{5.10}$$

<u>The case of interacting cells</u>: Here, the coupling constant J_{ij} is non-zero, which makes the exact solution impossible to calculate analytically at least in the 3D case while it is very complicated in 2D and hence we resort to approximations. A convenient approach to solving this problem is to apply the mean-field approximation where the quadratic terms in spin fluctuations are neglected. The Hamiltonian in this case becomes

`

$$\mathcal{H} = \mathcal{H}_0 \ + \mathcal{H}_1 \tag{5.11}$$

where

$$\mathcal{H}_0 = \sum_{i<j=1}^{N} J_{ij}\langle s_i\rangle\langle s_j\rangle + \sum_{i=1}^{N} h_i \tag{5.12}$$

$$\mathcal{H}_1 = -\sum_{i=1}^{N}\left(2h_i + J_i^{\mathrm{eff}}\right)s_i \tag{5.13}$$

and $J_i^{\mathrm{eff}} = \sum_{j\neq i}^{N} J_{ij}\langle s_j\rangle$ is the effective interaction coefficient between cells and h_i is the direct effect of cytotoxicity on cells. From Eq. (5.12), it follows that $\mathcal{H}_0$ is an average value for the Hamiltonian with no effect on any particular cell directly while the first term in Eq. (5.13) signifies the effect of the control parameter on the ith cell s_i and the second term can be interpreted as the average bystander effect from all other cells on the ith cell s_i [61,62]. Hence, the partition function Z given by Eq. (5.2) becomes

$$Z = e^{-\mathcal{H}_0/k_B T} \prod_{i=1}^{N}\left(1 + e^{\left(2h_i + J_i^{\mathrm{eff}}\right)/k_B T}\right) \tag{5.14}$$

Note that Eq. (5.11) is a general result valid for any drug concentration distribution used in numerical computations [61]. To derive some analytical results, however, we need to assume again a uniform drug distribution *via* $h = h_i$, which leads to the uniform magnetization, $M = M_i$, representing an average survival response of the cell culture, and the effective constant interaction strength, *J*. As a result, the partition function and the order parameter are found after several steps of calculations as;

`

$$Z = \left[2e^{-\lambda M(M-1)} \cosh\left(\frac{h}{k_B T} + \lambda M\right)\right]^N \tag{5.15}$$

and

$$2M - 1 = \tanh\left(\frac{h}{k_B T} + \lambda M\right) \tag{5.16}$$

where $\lambda = J/2k_B T$ [61]. Using the fact that the order parameter, $M,$ and control parameter h, are equivalent in the case of cancer cells to the death rate, R and the logarithm of concentration, $\log(C)$, respectively, we rewrite equation (5.16) as

$$2R - 1 = \tanh\left[\frac{h}{k_B T} + \lambda R\right] \tag{5.17}$$

which reduces to Eq. (5.8) when $\lambda = 0$. Inserting Eq. (5.10) in Eq. (5.16), we find that

$$R = \frac{1}{2}\left(1 + \tanh\left[1.15\alpha \log\left(\frac{C}{C_M}\right) + \lambda R\right]\right) \tag{5.18}$$

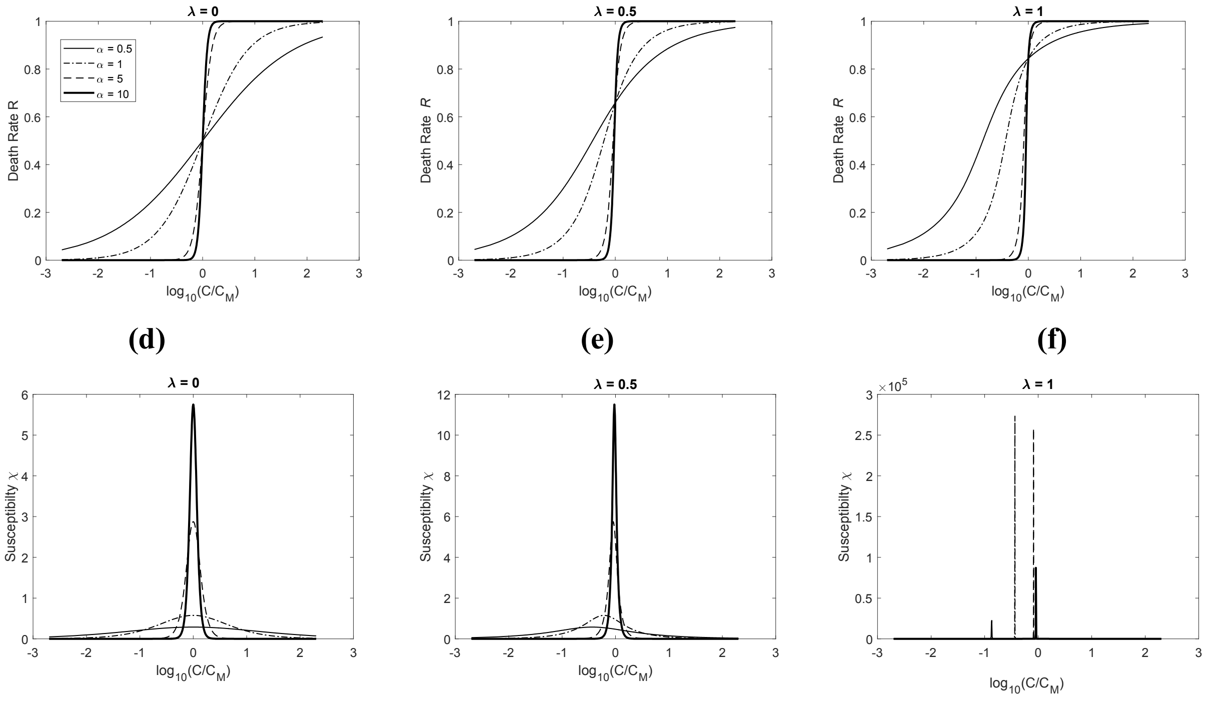
where $1.15 = \ln(10)/2$ [281]. As shown in the next section, equation (5.18) provides excellent agreement with the cytotoxicity experiment data. Equation (5.18) has a critical value for λ, at which a discontinuity in the death rate emerges and can be solved numerically for given values of α and λ. Upper panels in Figure 5.1 illustrate the death rate, $R,$ as a function of $\log(\,C/C_M)$ for four values of α and λ, respectively. Figure 5.1.a shows the result for uncorrelated cells, which presents similar behavior to that found in the Landau theory of phase transition (See section D.1 and Figure D.1 in Appendix D) [73,74,275,277]. As shown in Figures 5.1.a-c, increasing the interaction strengths between cells for $\lambda = 0, 0.5,$ and 1, triggers their transit from live to dead states as a result of only a subtle change in the drug concentration. As stated above the generalized susceptibility function, $\chi = \partial M/\partial h$, describes the sensitivity of the order parameter to a change in the control parameter [284,285], which here corresponds to the survival rate's change due to the change in the drug concentration. Using the Ising model results, equation (5.18), the cancer cell susceptibility function is calculated as the first derivative of the death rate, R, with respect to the log(C/C_M):

$$\chi = \frac{1}{2}\left[\frac{1.15\alpha(1-\tanh^2\delta)}{1-\frac{\lambda}{2}(1-\tanh^2\delta)}\right] \tag{5.19}$$

where δ=1.15$\alpha\log(C/C_M) + \lambda R$. In the case of no interaction, λ=0, using equation (5.19) we show that the susceptibility function maximum occurs at "zero-field" or $C = C_M$ with $\chi_{\max} =$ 0.575α. The susceptibility of the Ising model was extensively studied by Fisher [284,285]. Lower panels in Figure 5.1 depict susceptibility calculated using equation (5.19) as a function of log(C/C_M) for various values of α and λ, respectively. In the case λ=0, Figure 5.1.d, the susceptibility maximum occurs at zero-field, while by increasing λ it peaks at concentrations below C_M. Interestingly, by increasing α in the non-zero λ case, the maximum shifts towards zero-field. This demonstrates that there is a competition between α and λ, where larger values of α provide solutions similar to $\lambda = 0$. This is observed in the parameter values found using the experimental data.

(a) **(b)** **(c)**

`



(d)

(e)

(f)

Figure 5.1 Solution plots for the death rate (**a**, **b** and **c**) and susceptibility (**d**, **e**, and **f**) of equations (5.18) and (5.19), for four values of α= 0.5, 1, 5 and 10; **(a)** and **(d)**: λ= 0, **(b)** and **(e)**: λ= 0.5, **(c)** and **(f)**: λ=1.

5.4 Results and discussion

In this study, we applied the Ising model methodology to better understand and more accurately describe cancer cell response to chemotherapy agents. Inhibition profiles of 13 diverse anti-cancer compounds were analyzed from proliferation assays performed on 66 cancer cell lines provided by Oncolines, Inc., the Netherlands [83,84] (See section D.2 in Appendix D for experimental methodology). The anti-cancer compounds tested were: Afatinib, Bortezomib, Busulfan, Cisplatin, Doxorubicin, Idelalisib, Irinotecan, Methotrexate, Paclitaxel, Palboliclib, Tazemetostat, Trametinib and Vincristine [286–290]. For all the 13 compounds and 66 cell lines (See Table D.1

`

in Appendix D), the death rate, R, is plotted in terms of the logarithm of anti-cancer drug concentration, $\log(C)$, and fitted to the following function

$$R = a\left[b + \tanh\left(c(log(C) + d)\right)\right] \tag{5.20}$$

where a , b, c and d represents the best-fit parameters. As an example, Figure 5.2.a shows the results for Bortezomib acting on a melanoma cell line (A375). Based on the cell response-Ising model, the EC50 value marks the drug concentration at which the phase transition from a live system to a dead one occurs. In Figure 5.2.a, dashed lines represent the best-fitting function and a red star and a purple circle represent the experimental and predicted EC50 values, respectively. Taylor expansion of the fitted function around EC50 is shown using a green solid line in Figure 5.2.a. Note that the solid line in cyan shows the susceptibility in equation (5.19) and its highest value coincides with the EC50 concentration, which provides a rationale for the thus far arbitrary use of EC50 as a significant parameter for cytotoxicity estimates. These findings also demonstrate good agreement between our model and the Landau mean-field theory of phase transition (see *SI Text*). Figure 5.2.b illustrates a similar behavior for Paclitaxel applied to other cell lines, namely: 769-P (Kidney), A-172 (Blood), A-375 (Skin), A-427 (Lung), BxPC-3 (Pancreas), BT-549 (Breast) and Colo-205 (Colon). All findings in this section demonstrate good agreement between the proposed model and the experimental data.

`

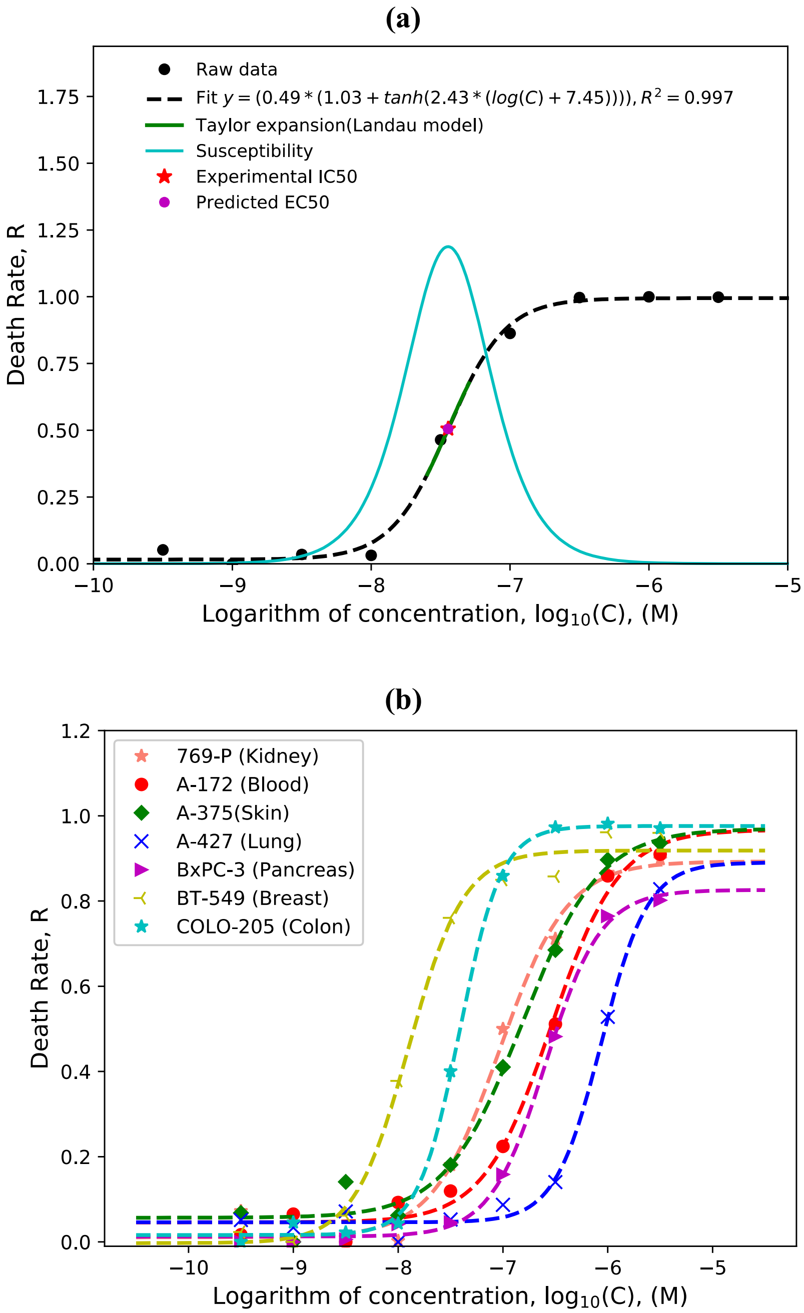


Figure 5.2 (a) Profile of the death rate in terms of the logarithm of the drug concentration for the melanoma A375 cell line upon its exposure to the Bortezomib drug. Dashed line represents the fit to Eq. (5.18), the solid line in cyan is the corresponding susceptibility obtained from Eq. (5.19), the red star represents the experimental EC50 value, the purple circle shows the predicted EC50 value from our model and the solid green line shows the Taylor expansion around the coefficient *d* (which shows the correlation to the Landau theory of phase transition).

`

(b) Profile of the death rate in terms of the $\log(C)$ of Paclitaxel drug for different cell lines in different colors and the data points are demonstrated in various symbols.

One of this model's advantages is its ability to predict a precise value of EC50 by finding the parameter d of the fitting function. For those cytotoxic drugs, which follow the model, the difference between the predicted and experimental values of EC50is very minor, on the order of 10^{-3}. Comparing the fitted curve with equation (5.18) derived from the cell response-Ising model in Table 5.1, one finds excellent agreement between the predicted and observed results for non-interacting cell lines ($\lambda = 0$ in equation (5.18)). Table 5.1 shows the average value of the fitting parameter for all the 66 cell lines tested followed by the correlation coefficient for each cytotoxic drug between the experiment and the cell response-Ising model. The best-fit parameter values among the cell lines are fairly consistent for a, b, R^2 and χ among the cytotoxic drugs, although some small fluctuations can be seen for parameter c and d values. To study the possibility of the bystander effect, we use equation (5.18), derived using the Ising model, with an assumption that $\alpha = 1$. The obtained values are fairly consistent among all the drugs (See Figure D.3 in the appendix D). It emerges that drugs with higher correlations to our model have a high susceptibility values as well. The correlation coefficient includes the cell lines with $R^2 > 0.5$ and the susceptibility $\chi < 10$. These two conditions drop ~16% of the cases studied, which do not follow the model. One possible reason for this is that the experimental EC50 reported is not appropriate, possibly meaning that the drug was not cytotoxic for that particular cell line or the drug dosage was not sufficiently high.

Table 5.1 The best-fit parameters to equation (5.20), a, b, c and d, susceptibility, χ, and correlations with the theoretical model (± stands for standard deviation)

`

Predicted parameters based on the Ising model

Cytotoxic drugs	**A**	**b**	**c**	**d**
Equation (5.20)	0.5	1	1.15α	$-\log_{10}$ (EC50)

Estimated fitting parameters of equation (5.20) when $\lambda = 0$

Cytotoxic drugs	**A**	**b**	**c**	**d**	R^2	χ	**Total correlation (%)**
Bortezomib	0.47 ± 0.02	1.08 ± 0.06	5.98 ± 6.49	7.68 ± 0.36	0.99 ± 0.01	2.57 ± 2.7	100.0
Methotrexate	0.45 ± 0.06	1.09 ± 0.08	8.16 ± 7.54	7.66 ± 0.35	0.99 ± 0.01	2.99 ± 2.6	100.0
Paclitaxel	0.43 ± 0.07	1.07 ± 0.07	2.37 ± 1.01	7.52 ± 0.55	0.99 ± 0.01	1.01 ± 0.5	100.0
Vincristine	0.45 ± 0.04	1.09 ± 0.07	4.37 ± 4.14	8.00 ± 0.61	0.99 ± 0.02	1.80 ± 1.4	100.0
Doxorubicin	0.48 ± 0.06	1.07 ± 0.09	2.16 ± 2.60	6.96 ± 0.49	0.99 ± 0.03	0.99 ± 1.1	97.0
Irinotecan	0.43 ± 0.11	1.14 ± 0.12	3.37 ± 4.72	5.51 ± 0.54	0.97 ± 0.04	1.21 ± 1.5	95.5
Cisplatin	0.54 ± 0.26	1.10 ± 0.06	2.37 ± 3.45	5.16 ± 0.54	0.98 ± 0.02	1.11 ± 1.4	93.9
Afatinib	0.51 ± 0.09	1.14 ± 0.16	2.38 ± 1.25	5.47 ± 0.37	0.98 ± 0.02	1.15 ± 0.6	87.7
Trametinib	0.31 ± 0.20	1.15 ± 0.36	3.92 ± 6.63	7.69 ± 0.90	0.92 ± 0.11	0.60 ± 0.7	72.7
Idelalisib	0.67 ± 0.75	1.16 ± 0.20	3.95 ± 6.09	4.43 ± 0.88	0.91 ± 0.10	1.00 ± 1.1	71.2
Tazemetostat	0.20 ± 0.17	1.51 ± 0.45	6.50 ± 10.57	5.18 ± 0.71	0.80 ± 0.16	1.32 ± 1.6	62.1
Palbociclib	0.63 ± 0.27	1.10 ± 0.10	2.95 ± 4.84	5.14 ± 0.48	0.98 ± 0.02	1.50 ± 2.2	59.1
Busulfan	0.52 ± 1.65	1.44 ± 0.45	10.12 ±9.13	4.49 ± 3.50	0.70 ± 0.17	1.76 ± 2.3	53.0

In addition, Table 5.2 lists the fitting results of our model to all the cytotoxic drugs for interacting cells, $\lambda \neq 0$, and we fitted the data with the equation $R = 0.5(1 + \tanh(1.15(\log(C) - \log(\mathrm{EC50})) + \lambda R)$. Although for some drugs such as Trametinib, Idelalislib, Tazemetostat and

Busulfan, only a few of the cell lines have been fitted very well to the model, the results does not show the existence of the bystander effect in most cell lines. It is important to note that no obvious trend has been found between the interacting and non-interacting cell lines for each drug. In other words, the cell lines that were not following the model in the non-interacting case, are not well fitted to the model in the interacting case either (See section D.3, Table D.2 and Figure D.2 in the appendix D).

Table 5.2 Results of the application of the Ising model for interacting cells yielding the equation R= $0.5(1 + \tanh(1.15(\log(C) - \log(\mathrm{EC50})) + \lambda R)$ (± stands for the standard deviation).

Cytotoxic drugs	$\boldsymbol{\lambda}$	$\boldsymbol{R^2}$	**Total correlation (%)**
Bortezomib	0.41±0.21	0.96±0.03	98.5
Methotrexate	1.01±1.53	0.95±0.04	75.8
Afatinib	0.71±0.71	0.87±0.20	66.7
Vincristine	1.18±0.93	0.95±0.08	63.6
Doxorubicin	1.00±0.95	0.97±0.03	57.6
Paclitaxel	0.98±0.79	0.93±0.09	56.1
Cisplatin	0.89±0.72	0.96±0.03	53
Palboliclib	0.86±0.59	0.83±0.12	51.5
Irinotecan	1.22±1.06	0.93±0.10	40.9
Trametinib	0.72±0.34	0.80±0.16	10.6
Idelalilsib	0.31±0.24	0.90±0.08	9.1
Busulfan	0.21±0.12	0.89±0.01	4.5
Tazemetostat	0.17±0.04	0.81±0.04	4.5

`

In the original dataset, there are some cell lines for which the corresponding IC50 (or EC50) value was not reported. This is due to it exceeding the maximum tested concentration $(< 31600\ nM)$. At this concentration, the compounds still show less than 50% inhibition of cell proliferation. Figure 5.3 shows the correlation of each cytotoxic drug's profile with the cell response-Ising model prediction and the percentage of cells with missing IC50 values. We found that Busulfan, Palbociclib, Tazemetostat, Idelalisib and Trametinib exhibit a relatively low correlation with the model while these cytotoxic drugs, except Palbociclib, have a high number of missing IC50 values. On the other hand, cytotoxic drugs with perfect correlations have only one cell line without an IC50 value measured. Therefore, there appears to be an inverse relationship between the correlation coefficient and the missing IC50 (or EC50) values. Those cell lines, which are missing the IC50 values appear not to follow the Ising model. One conclusion that can be reached is that the Ising model as applied to cytotoxicity is sensitive to the IC50 values. For those cell lines for which the measurement has been made either below or above the IC50 values only, no phase transition has occurred so the model will not work (recall that the maximum value of the experimentally measured IC50 was 31.6 μM). Figure 5.3 shows correlation of the drugs with the cell response-Ising model in blue and the drugs with missing IC50 in red. Interestingly, eight of the drugs tested, namely Bortezomib, Methotrexate, Paclitaxel, Vincristine, Doxorubicin, Irinotecan, Cisplatin and Afatinib, exhibit excellent consistency with the cell response-Ising model (exceeding an 87% correlation coefficient) while for the rest of the drugs the correlation coefficient is between 52% and 73%.

Since a diverse set of cancer cell lines was used among the 66 cell lines tested, this could have contributed to a low correlation coefficient for the drugs, which target a specific cancer type. For example, Busulfan is mostly used for treating bone marrow transplantation, especially in chronic

`

myelogenous leukemia (CML), such as represented by the following cell lines: SR, MOLT-4, K-562, SK-N-AS, SK-N-FI. The Ising model was fitted very well for bone marrow and chronic myelogenous leukemia (CML) cell lines in this case. This trend can also be seen for Trametinib, Idelalisib, Tazemetostat and Palbociclib. Correlated and uncorrelated cell lines have also been represented using green and red colors, respectively.

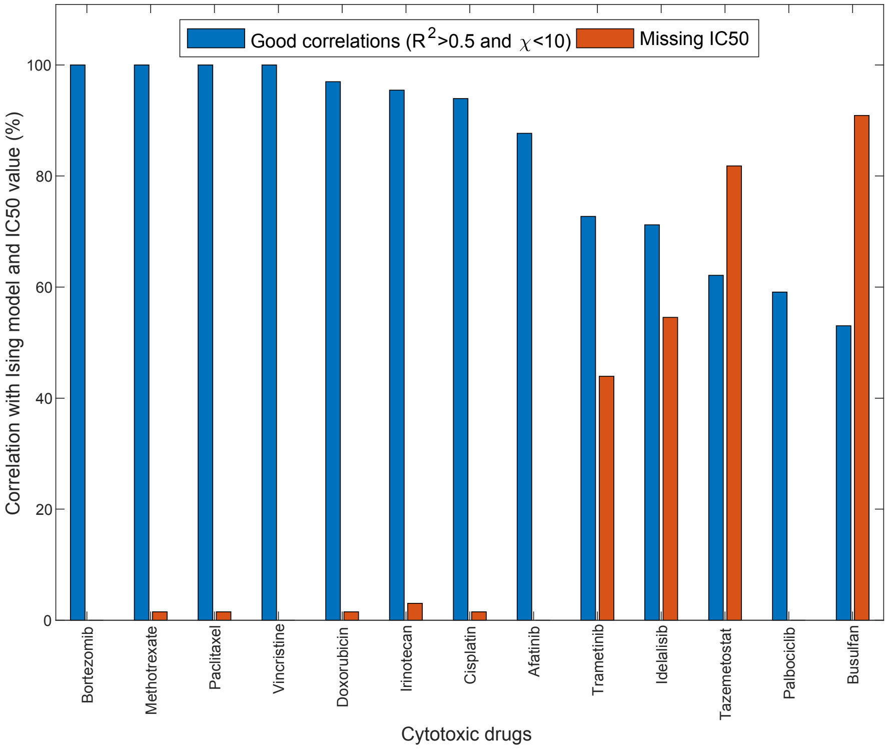
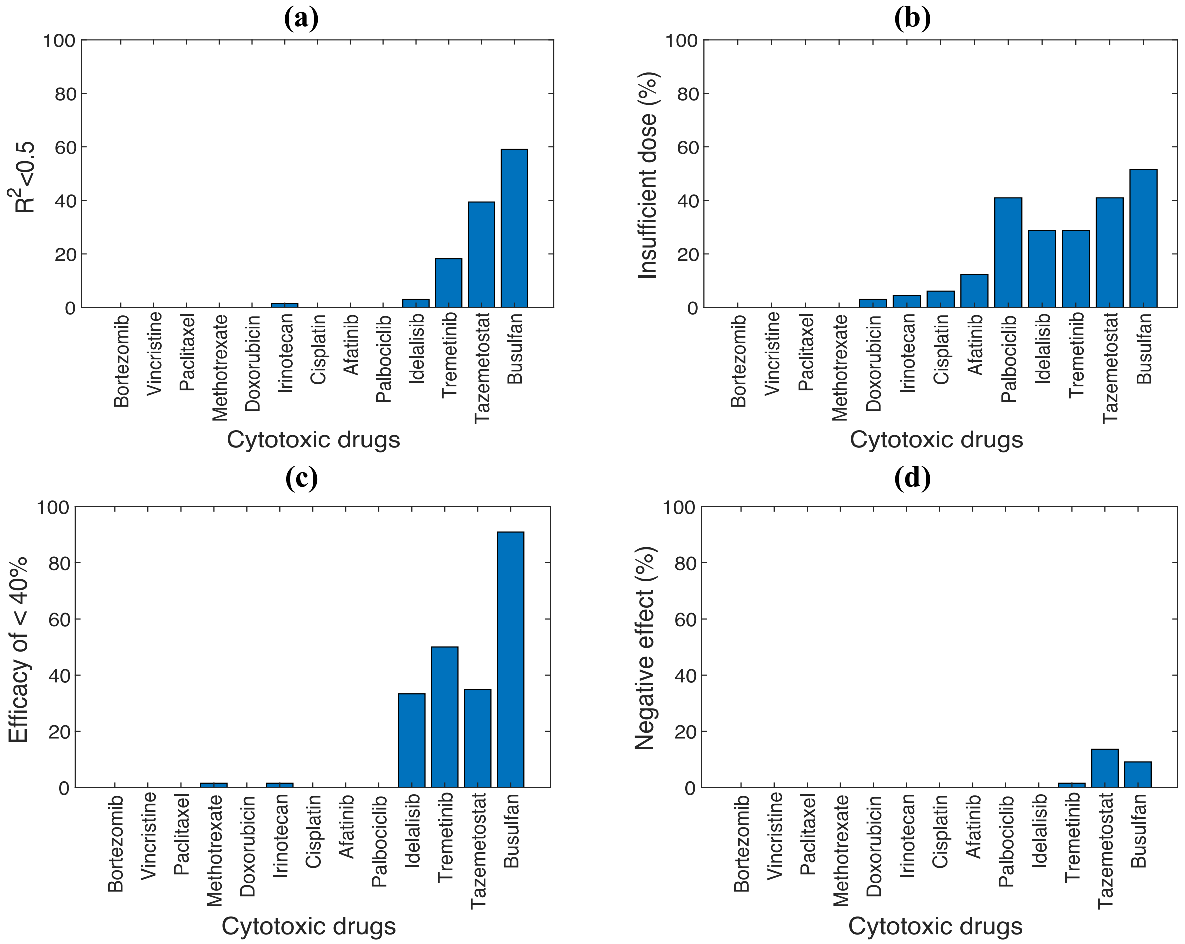


Figure 5.3 Good correlations with the cell response-Ising model and the corresponding EC50 values, blue bars represent the good correlations (%) and the red bars shows the missing EC50 reported for each cytotoxic drug (%)

Figure 5.4 shows the effect of cytotoxic drugs on the cell lines in different categories such as: cell lines with $R^2 < 0.5$, exhibiting insufficient cytotoxic dose, showing a bizarre reversal effect (i.e. increasing the dose decreases the cells death rate), showing efficacy of less than 40%. Figure 5.4.a shows the number of the cells exposed to different drugs, which kill less than 40% of the cells. It

`

demonstrates the cells that either were not responding well to the cytotoxic drugs or the drug was not cytotoxic enough. Comparing Figure 5.4.a and Figure 5.3, one can see that the drugs with a missing IC50 value have a lower death rate, such as Busulfan, Tazemetostat, Trametinib and Idelalisib. Figure 5.4.b illustrates the cell lines for which the dosage used was insufficient to cause the death of these cell lines. Also, Figure 5.4.c shows the efficacy of the drug on these cancer cell lines. Over 90% of the cell lines have a lower efficacy for Busulfan. The efficacy value is 35% and 33% for Tazemetostat and Idelalisib, respectively and 1.6% for Irinotecan and Methotrexate. Finally, Figure 5.4.d shows that some of the cell lines show a toxicity reversal effect when exposed to Tazemetostat, Busulfan and Trametinib such that increasing the cytotoxic dosage decreases the death rate, although the number of such cases is negligible.



`

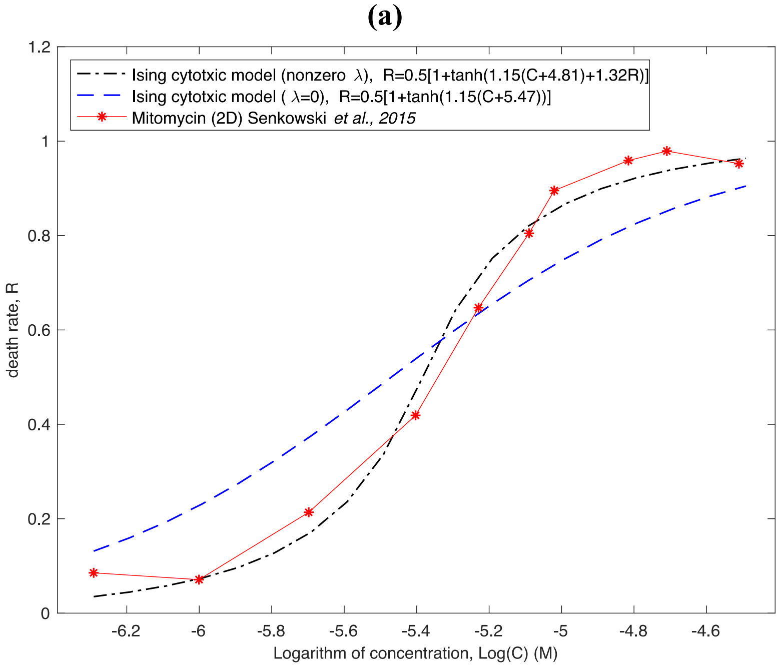
Figure 5.4 Weak correlation cell lines in different categories of **(a)** having $R^2 < 0.5$, **(b)** exhibiting insufficient cytotoxic dose, **(c)** showing reversal effect and **(d)** efficacy of less than 40% in the death rate.

As part of our study, the interacting cases with a non-zero λ, have also been considered. The same datasets have been fitted to equation (5.18) and the implicit equation for $\log(C)$ was solved. The interaction coefficients were assumed to be positive, $\lambda = J/2k_BT \geq 0$ in order to conform to the known biological effect ($\lambda_C = 2$) (See Table D.2 and Figure D.3 in the appendix D).

Until now all the cases analyzed in this work involved typical well-plate cytotoxicity assays, in which cell culture grows within a plane and is then exposed to toxic chemotherapy agents. Spatial dimensionality of physical systems undergoing phase transitions plays a crucial role in the response of the system to control parameter changes. In order to explore whether this can also be seen in biological systems such as cancer cells, we have found some experimental data in the literature that provide examples of dimensionality dependence. Figure 5.5 shows a comparison between the cell response-Ising model and the experimental data from reference [291] in which the drug Nitazoxanide was studied as a colorectal cancer therapy candidate. In Figure 5.2 of reference [291], HCT116 and HCT116 GPF cells were exposed to two cytotoxic drugs, Nitazoxanide and Mitomycin in monolayer two-dimensional and multicellular tumor spheroid (3D) cell cultures for 72 hours. Here, dose-response curves for the two drugs have been compared to the cell response-Ising model in Figure 5.5. The black dashed lines and blue dashed lines represent the best fit values of the cell response-Ising model in the cases with a non-zero and zero λ and the red line shows the HCT116 cell line exposed to Mitomycin a) 2D and b) 3D and Nitazoxanide c) 2D and d) 3D. It can be seen that in the case with no cell-cell interactions, $\lambda = 0$,

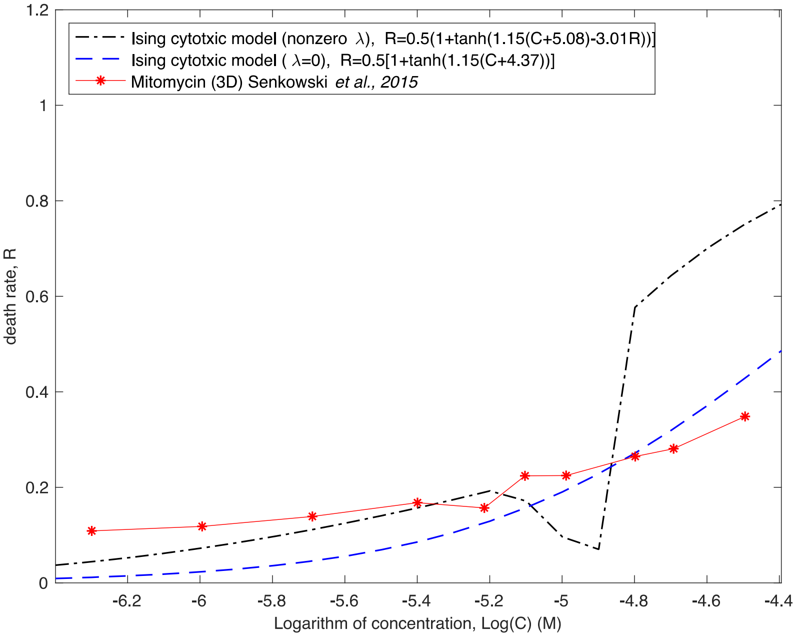
`

the two cytotoxic drugs have a better correlation than in the interaction case. In the presence of the interaction term, $\lambda \neq 0$, the values of λ in the curves a to d, are $1.32, -3.01, 0.74$ and 0.05, respectively. These values show that the model doesn't work well for the Mitomycine 3D experiment, while for Nitazoxanide 3D, the λ value shows that the cell-cell interactions are very weak (almost zero). However, in the case of the 2D experiments for both drugs, we see a better correlation in the cell-cell interaction cases, see Figures 5.5.a and c. We can, therefore, tentatively conclude that spatial dimensionality does indeed affect the response of cell cultures to cytotoxic agents but a more in depth analysis of larger datasets is required to develop an appropriate mathematical model that captures these complex systems' behavior better than the present Ising model.

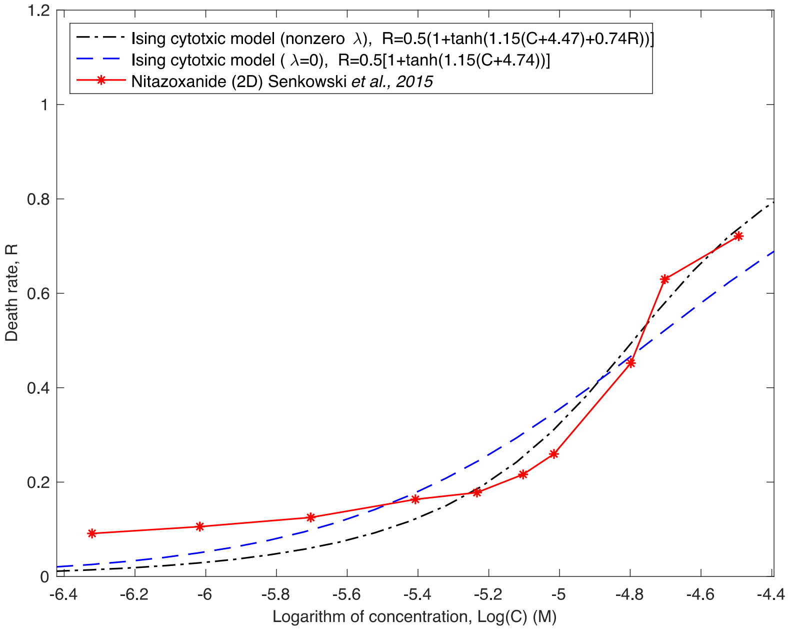


`

(b)


Ising cytotxic model (nonzero λ), R=0.5(1+tanh(1.15(C+5.08)-3.01R))]
Ising cytotxic model (λ=0), R=0.5[1+tanh(1.15(C+4.37))]
Mitomycin (3D) Senkowski *et al., 2015*
death rate, R
Logarithm of concentration, Log(C) (M)
0
0.2
0.4
0.6
0.8
1
1.2
-6.2
-6
-5.8
-5.6
-5.4
-5.2
-5
-4.8
-4.6
-4.4

(c)


Ising cytotxic model (nonzero λ), R=0.5(1+tanh(1.15(C+4.47)+0.74R))]
Ising cytotxic model (λ=0), R=0.5[1+tanh(1.15(C+4.74))]
Nitazoxanide (2D) Senkowski *et al., 2015*
Death rate, R
Logarithm of concentration, Log(C) (M)
0
0.2
0.4
0.6
0.8
1
1.2
-6.4
-6.2
-6
-5.8
-5.6
-5.4
-5.2
-5
-4.8
-4.6
-4.4

`

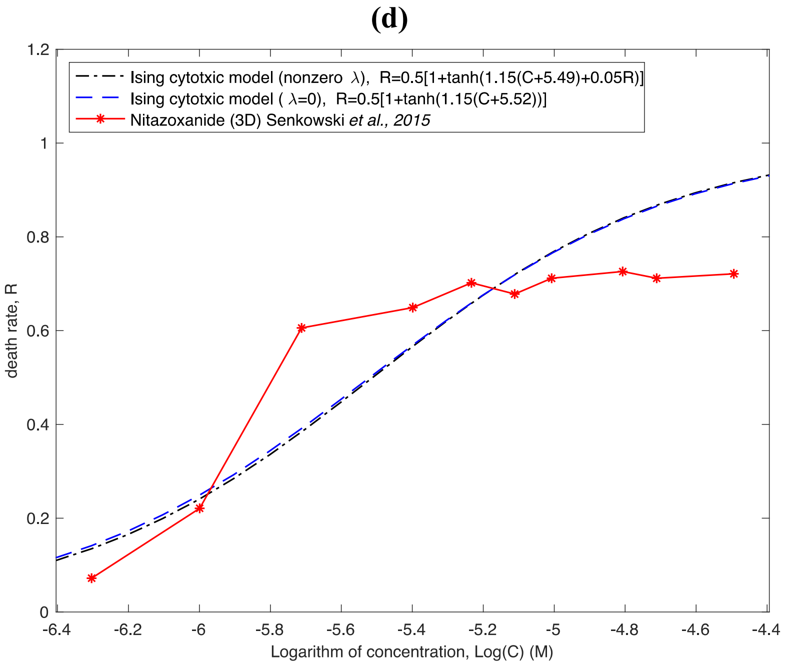


Figure 5. 5 Death rate profile for the colorectal cancer cell line, HCT116, exposed to Nitazoxanide and Mitomycin cytotoxic drugs (solid red line) Ref. [291]. Blue and black dashed lines represent the best fit curves to the cell response-Ising model for interacting and non-interacting cases, respectively.

5.5 Conclusions

In this study, the physical concepts developed for the theory of phase transitions occurring in bistable systems were for the first time applied to describe the effects of various chemotherapeutic agents on cancer cell lines. Specifically, we adopted the Ising model of a spin system and applied it to the survival plots of cancer cells at different concentrations of the various chemotherapeutic agents these cells were exposed to. This model was originally proposed for a spin system in a uniform external field with a constant interaction parameter and a variable temperature. In the case of cancer cells, the external field is analogous to the logarithm of the drug concentration while the interaction parameter describes the cell-cell interactions and hence accounts for the bystander effect. Unlike in physical systems, temperature is kept constant for cancer cells in the reported assays. It should be noted that this model has been successfully applied to both

`

interacting and non-interacting cells depending on the underlying biological situation. We have tested the model on a consistently produced data set of 66 cancer cell lines exposed to 13 different cancer chemotherapy drugs. The results show good agreement between the cell response-Ising model and the biological data. Using the bistabiliy concept in the Ising model, EC50 (or IC50) values can be very accurately determined with an error on the order of $1\,nM$ by one of the parameters of the fitting function in the non-interacting case. The cell-cell interaction was also applied to the experimental data, although in our case, most of the cell lines tend to be non-interacting. Nonetheless, the presence of interactions can be determined using our fitting procedures and it offers a clear biological insight that bare experimental data do not reveal. We have additionally introduced the thermodynamic concept of the susceptibility function and found its peak to closely coincide with the value of EC50. Further studies should be performed considering a non-constant interaction term as well as non-uniform fields in the Ising model applied to cytotoxicity assays with both 2D and 3D geometries and various cell concentrations. In addition, spatial dimensionality of the cancer cell culture was shown to affect the response to cytotoxic agents, which requires a future study to gain insight into how 2D culture may not be an appropriate proxy for tissue-based studies. This approach is expected to introduce a high level of consistency in cytotoxic data analysis and hence better confidence in the preclinical data assessment for cancer chemotherapy and related applications.

`

Chapter 6 – Conclusions and Future Work

The research reported in this thesis concerns fundamental questions in biological systems and performed a search to find underlying physical bases for the empirical observations in biology. The issues investigated in this connection include the logic of the genetic code, the probability distribution of protein mutations, and the statistics of surviving cancer cells under toxic stress. While spanning a range of seemingly disparate topics, all these questions concern the applicability of physical approaches to basic biological units of living systems such as genes, proteins, and cells. It has been our goal to shed light on empirical facts from genetics, molecular, and cell biology applying the lens of statistics, thermodynamics, and the physics of phase transitions.

6.1 Amino acid and codon energy and probability in the genetic code

6.1.1 Summary

In Chapter 2, we investigated the occurrence frequency of amino acids and codons from the point of view of energy and occurrence frequency in order to obtain an underlying correlation from physical approaches instead of purely experimental observations between amino acid and codons. We were motivated to find correspondence correlation between probability of amino acid over the species in the evolutionary tree of life. The energy estimates of the 20 natural amino acids were evaluated using GAMESS software based on a semi-empirical method employing and the Hartree-Fock method and PM3 basis sets. The energy values of codons and those of amino acids have been obtained and contrasted in search of a pronounced correlation between the two, which would give a logical explanation for the assignment of codons to amino acids found in nature. However, the results generated using our methods were scattered, and no correlation could be

`

found. In addition to the amino acid energy obtained from GAMESS, another software called Gaussian was also used to obtain the amino acid energy. These results showed that except for the two outliers, cysteine and methionine, the results of the two methods were well correlated. This discrepancy could be due to the parametrization differences between the two software dealing with Sulphur containing atoms. Moreover, we showed that higher degeneracy amino acids are more probable, shedding some light on the question of amino acid abundance in nature. However, from the energetic point of view, our results did not confirm the hypothesis put forward that the higher frequency stems from the cheap energetic cost in natural systems, including humans. Moreover, our results interestingly show that the amino acid probability distribution is highly conserved across the species according to the evolutionary tree of life that included in our analysis various species from bacteria to human body tissues, as well as animal and fungal mitochondrial proteins. Finally, in our research focused on the fundamental issues in cell biology, we analyzed the paradox of the apparent entropy reduction in the process of transcription and translation, which starts from DNA and takes it to RNA ending in protein synthesis. We showed that the entropy reduction paradox occurring across the biological species could be explained by the involvement ATP and GTP macromolecules.

6.1.2 Future work

There is a lot of room for future work and unanswered questions regarding the topics broached in this thesis. This is because questions remain unanswered about the origin of the genetic code as an algorithm providing rules for codon-to-amino acid assignments and its advantages, the evolution of the genetic code, and eventually, the reasons behind having exactly 20 amino acids with their specific physico-chemical characteristics.

`

6.2 Amino acid-codon docking

6.2.1 Summary

We followed one of these particular questions with an in-depth analysis In Chapter 3, whose aim was to test the hypothesis if amino acids have an increased propensity to bind to their cognate codon or anticodon due to the differences in the associated binding free energies. We tested this hypothesis using computational structure-based methods, in particular as RNA-protein docking simulations. The 3D structure of the RNA strand used contained all 64 codons created with the help of MOE software. After protonation and structural minimization, we used steered molecular dynamics (SMD) simulations of the RNA structure and applied the docking protocol using the 3dRPC method to all the amino acid-codon pairs ignoring the codon-codon interactions. Our results show that there is no obvious trend that can be seen to confirm the hypothesis that the amino acids are more likely to bind to their codon or anticodon. In fact, two paradoxical examples of isoleucine and arginine directly contradict this hypothesis.

6.2.2 Future work

Further improvements in the calculations performed can be made in the future. Namely, the interaction between codons can be taken into account before implementing the docking simulation. In addition, instead of a single amino acid, a peptide chain can be used to find the binding affinity of the docking complex. This might restrict the flexibility of the receptor in the RNA, making this a more realistic representation of the situation.

`

6.3 Probability of amino acid mutations in the p53 protein

6.3.1 Summary

Having attempted to understand if any potential free energy-based considerations can shed light on the rules of the genetic code and its consequence for protein synthesis, we then turned our attention to proteins and protein mutations. As a particularly important example of a protein that plays a crucial role in cancer initiation, and progression, we studied p53. In Chapter 4, the specific aim of our analysis was to study the frequency of the p53 protein mutations in its gene sequence across various cancer types. We focused on the somatic mutations of p53, which are reported in human cell lines, primary tissues, and fluids in the human body. We showed that some of the amino acid mutations are especially highly probable; these mutations are: RH (79%), RW (71%), RQ (73%), GS (55%), and RS (48%). We also showed that in ~84% of the cancer types, at least one of the above-mentioned mutations is the highest frequency mutation. In addition to the frequency analysis, the Shannon entropy of each mutation was calculated for all studied cancer types. We endeavored to find a quantitative relationship between the entropy of p53 mutations and the number of p53 mutations. We showed that except for lung cancer, no noticeable trend could be found. Also, we demonstrated that there is no discernable correlation between Shannon entropy and the five-year-survival rate for cancer patients. In other words, while p53 is a highly mutated functional protein in all types of cancer that loses its tumor-suppressing function due to mutations, the entropic measure of the mutation statistics does not turn out to be a prognostic factor in the survival outcome for these mutations at a cancer epidemiology level.

`

6.3.2 Future work

Thus, we subsequently focused only on the missense and silent mutations, which cover most of the p53 mutation cases in cancer. However, it might be interesting to study the other types of mutations as well as, in particular, the codon mutations in p53. Further research is still needed on the p53 protein due to its significant role in suppressing tumors. Other than the statistical approach conducted in the present study, more computational simulations need to be performed in order to improve cancer chemotherapy treatments, which aim to reactivate mutated p53 proteins. Focusing on the high-frequency ("hot spot") mutations of p53 can lead to measurable improvement of clinical outcomes by restoring the protective function of this key protein playing multiple functions in all eukaryotic cells.

6.4 Ising-Cytotoxicity Model

6.4.1 Summary

In Chapter 5, we applied the physical concept of a phase transition and the related idea of critical systems' bistability to explain the response of cancer cells to cytotoxic compounds in chemotherapy-based cancer treatments. We used the Ising spin ½ model as a powerful mathematical model developed in the theory of phase transitions and bistable systems to describe the two biological states of cancer cells (dead or alive), which are reflected in the dose-response curves. We assigned the physical spin-up and spin-down states in the Ising model to the biological dead and alive states of the cancerous cells analyzed. The concentration of the chemotherapy agent was represented by an external field similar to the magnetic field that aligns spin states of a magnet. In this case, this "cytotoxic" field promotes cell death and disfavors the state of being

`

alive. In addition, neighboring cells interact with each other in a manner similar to spin-spin interactions in magnetic systems. In cancer cell biology, this interaction between cells in the neighborhood is termed the bystander effect whereby the interacting cells tend to behave similarly whether or not they have been physically affected by the damaging agent. The effect of cytotoxic drugs on cancer (and normal) cells in the presence and absence of the bystander effect have also been investigated in this study by solving the Ising Hamiltonian with spin-spin interactions and applying the results to the biological data made available to us. The Ising cytotoxicity model was applied to numerous panels of cytotoxic compounds tested experimentally on many cancer cell lines by our collaborators in the Netherlands Translational Research Center B.V.(Oncolines). We showed that the analyzed data have a very good agreement with our model, and the results are highly consistent. In addition, our model has been demonstrated capable of predicting the value of the EC50 for each case very accurately, and the EC50 value coincided with the maximum value of one of the key characteristics of systems at criticality, namely the susceptibility function. It should be mentioned that one of the key findings about phase transitions is the singularity of the generalized susceptibility function at the critical point and hence the finding that the susceptibility function for cancer cells peaks at the EC50 value gives additional support for the use of these physical concepts in the area of cancer cell biology. Our work in this field is one of the very first that proposes the use of this methodology. Finally, the proposed model has been successfully implemented on the dose-response data of two-dimensional and three-dimensional spheroid models in the absence ($\lambda=0$), and presence ($\lambda\neq0$) of the constant interaction between cancer cells correspond to the absence and presence of the bystander effect. We also showed that spatial dimensionality does affect the cell response to the cytotoxic compounds.

6.3.2 Future work

It should be noted that in the theory phase transitions, the two parameters that determine the so-called universality class of critical systems are: the number of order parameter components and the dimensionality of physical space in which the system exists. Therefore, these two aspects merit further attention, and it requires more data, especially regarding the dimensionality of cancer cell cultures used in such studies. Therefore, this is a potential subject for further studies on the non-constant interaction factor in different spatial dimensionalities (1D, 2D, and 3D) on a larger experimental dataset and in-depth analysis to develop a mathematical model which is able to map the behavior of these complex systems. This could be of importance in finding a correlation between simplified in vitro studies in 2D cell cultures and more relevant in vivo studies with xenograft tumors in 3D grown in animal models. Such results could be of practical importance in deciding whether or not particular chemotherapy compounds should be further developed or abandoned in spite of their promising in vitro properties. Currently, such decisions are made entirely empirically, and to the best of our knowledge, no mathematical model has been implemented to the in vitro-in vivo correlation analysis of chemotherapy drugs.

`

Bibliography

[1] M. Aldana-González, G. Cocho, H. Larralde, G. Martínez-Mekler, Translocation properties of primitive molecular machines and their relevance to the structure of the genetic code, J. Theor. Biol. 220 (2003) 27–45. https://doi.org/10.1006/jtbi.2003.3108.

[2] M. Aldana, F. Cázarez-Bush, G. Cocho, G. Martínez-Mekler, Primordial synthesis machines and the origin of the genetic code, Phys. Stat. Mech. Its Appl. 257 (1998) 119–127. https://doi.org/10.1016/S0378-4371(98)00133-2.

[3] V.A. Gusev, D. Schulze-Makuch, Genetic code: Lucky chance or fundamental law of nature?, Phys. Life Rev. 1 (2004) 202–229. https://doi.org/10.1016/j.plrev.2004.11.001.

[4] E.V. Koonin, A.S. Novozhilov, Origin and evolution of the genetic code: the universal enigma, IUBMB Life. 61 (2009) 99–111. https://doi.org/10.1002/iub.146.

[5] F.H.C. Crick, L. Barnett, S. Brenner, R.J. Watts-Tobin, General Nature of the Genetic Code for Proteins, Nature. 192 (1961) 1227–1232. https://doi.org/10.1038/1921227a0.

[6] M.W. Nirenberg, O.W. Jones, P. Leder, B.F.C. Clark, W.S. Sly, S. Pestka, On the Coding of Genetic Information, Cold Spring Harb. Symp. Quant. Biol. 28 (1963) 549–557. https://doi.org/10.1101/SQB.1963.028.01.074.

[7] F.H.C. Crick, The origin of the genetic code, J. Mol. Biol. 38 (1968) 367–379. https://doi.org/10.1016/0022-2836(68)90392-6.

[8] R.T. Hinegardner, J. Engelberg, Rationale for a universal genetic code, Science. 142 (1963) 1083–1085. https://doi.org/10.1126/science.142.3595.1083.

[9] G.U. Nienhaus, ed., Protein'Ligand Interactions: Methods and Applications, Humana Press, 2005.

[10] M. Shimizu, Specific Aminoacylation of C4N Hairpin RNAs with the Cognate Aminoacyl-Adenylates in the Presence of a Dipeptide: Origin of the Genetic Code, J. Biochem. (Tokyo). 117 (1995) 23–26. https://doi.org/10.1093/oxfordjournals.jbchem.a124715.

[11] C.R. Woese, On the evolution of the genetic code, Proc. Natl. Acad. Sci. 54 (1965) 1546–1552. https://doi.org/10.1073/pnas.54.6.1546.

[12] C.R. Woese, R.T. Hinegardner, J. Engelberg, Universality in the Genetic Code, Science. 144 (1964) 1030–1031. https://doi.org/10.1126/science.144.3621.1030.

[13] C.R. Woese, D.H. Dugre, W.C. Saxinger, S.A. Dugre, The molecular basis for the genetic code, Proc. Natl. Acad. Sci. U. S. A. 55 (1966) 966–974. https://doi.org/10.1073/pnas.55.4.966.

`

[14] M.V. Vol'kenshteĭn, I.B. Rumer, [Systematics of codons], Biofizika. 12 (1967) 10–13.

[15] M. Di Giulio, The origin of the genetic code: theories and their relationships, a review, Biosystems. 80 (2005) 175–184. https://doi.org/10.1016/j.biosystems.2004.11.005.

[16] M. Hasegawa, T. Miyata, On the antisymmetry of the amino acid code table, Orig. Life. 10 (1980) 265–270. https://doi.org/10.1007/BF00928404.

[17] I.B. Rumer, [Codon systematization in the genetic code], Dokl. Akad. Nauk SSSR. 167 (1966) 1393–1394.

[18] V.R. Chechetkin, Block structure and stability of the genetic code, J. Theor. Biol. 222 (2003) 177–188. https://doi.org/10.1016/S0022-5193(03)00025-0.

[19] J.L. King, T.H. Jukes, Non-Darwinian Evolution, Science. 164 (1969) 788–798. https://doi.org/10.1126/science.164.3881.788.

[20] D. Gilis, S. Massar, N.J. Cerf, M. Rooman, Optimality of the genetic code with respect to protein stability and amino-acid frequencies, Genome Biol. 2 (2001) research0049. https://doi.org/10.1186/gb-2001-2-11-research0049.

[21] R. Wetzel, Evolution of the aminoacyl-tRNA synthetases and the origin of the genetic code, J. Mol. Evol. 40 (1995) 545–550. https://doi.org/10.1007/BF00166624.

[22] D. Haig, L.D. Hurst, A quantitative measure of error minimization in the genetic code, J. Mol. Evol. 33 (1991) 412–417. https://doi.org/10.1007/BF02103132.

[23] S.J. Freeland, T. Wu, N. Keulmann, The case for an error minimizing standard genetic code, Orig. Life Evol. Biosphere J. Int. Soc. Study Orig. Life. 33 (2003) 457–477. https://doi.org/10.1023/A:1025771327614.

[24] S. Itzkovitz, U. Alon, The genetic code is nearly optimal for allowing additional information within protein-coding sequences, Genome Res. 17 (2007) 405–412. https://doi.org/10.1101/gr.5987307.

[25] S. Matsuyama, T. Ueda, P.F. Crain, J.A. McCloskey, K. Watanabe, A novel wobble rule found in starfish mitochondria. Presence of 7-methylguanosine at the anticodon wobble position expands decoding capability of tRNA, J. Biol. Chem. 273 (1998) 3363–3368. https://doi.org/10.1074/jbc.273.6.3363.

[26] C. Allmang, A. Krol, Selenoprotein synthesis: UGA does not end the story, Biochimie. 88 (2006) 1561–1571. https://doi.org/10.1016/j.biochi.2006.04.015.

[27] J.D. Alfonzo, V. Blanc, A.M. Estévez, M.A. Rubio, L. Simpson, C to U editing of the anticodon of imported mitochondrial tRNA(Trp) allows decoding of the UGA stop codon in Leishmania tarentolae, EMBO J. 18 (1999) 7056–7062. https://doi.org/10.1093/emboj/18.24.7056.

`

[28] Á. Kun, Á. Radványi, The evolution of the genetic code: Impasses and challenges, Biosystems. 164 (2018) 217–225. https://doi.org/10.1016/j.biosystems.2017.10.006.

[29] R. Giegé, M. Sissler, C. Florentz, Universal rules and idiosyncratic features in tRNA identity, Nucleic Acids Res. 26 (1998) 5017–5035. https://doi.org/10.1093/nar/26.22.5017.

[30] R.D. Knight, S.J. Freeland, L.F. Landweber, Rewiring the keyboard: evolvability of the genetic code, Nat. Rev. Genet. 2 (2001) 49–58. https://doi.org/10.1038/35047500.

[31] D.W. Schultz, M. Yarus, Transfer RNA mutation and the malleability of the genetic code, J. Mol. Biol. 235 (1994) 1377–1380. https://doi.org/10.1006/jmbi.1994.1094.

[32] D.W. Schultz, M. Yarus, On malleability in the genetic code, J. Mol. Evol. 42 (1996) 597–601. https://doi.org/10.1007/BF02352290.

[33] S. Osawa, Evolution of the Genetic Code, Oxford University Press, 1995.

[34] S. Osawa, T.H. Jukes, K. Watanabe, A. Muto, Recent evidence for evolution of the genetic code., Microbiol. Rev. 56 (1992) 229–264.

[35] S.G. Andersson, C.G. Kurland, Genomic evolution drives the evolution of the translation system, Biochem. Cell Biol. Biochim. Biol. Cell. 73 (1995) 775–787. https://doi.org/10.1139/o95-086.

[36] S.G. Andersson, C.G. Kurland, Reductive evolution of resident genomes, Trends Microbiol. 6 (1998) 263–268. https://doi.org/10.1016/S0966-842X(98)01312-2.

[37] J.A. Krzycki, The direct genetic encoding of pyrrolysine, Curr. Opin. Microbiol. 8 (2005) 706–712. https://doi.org/10.1016/j.mib.2005.10.009.

[38] L. Wang, J. Xie, P.G. Schultz, Expanding the genetic code, Annu. Rev. Biophys. Biomol. Struct. 35 (2006) 225–249. https://doi.org/10.1146/annurev.biophys.35.101105.121507.

[39] M. Neveu, H.J. Kim, S.A. Benner, The “strong” RNA world hypothesis: fifty years old, Astrobiology. 13 (2013) 391–403. https://doi.org/10.1089/ast.2012.0868.

[40] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, Molecular Biology of the Cell, 4th ed., Garland Science, 2002.

[41] C. Saxinger, C. Ponnamperuma, C. Woese, Evidence for the interaction of nucleotides with immobilized amino-acids and its significance for the origin of the genetic code, Nature. New Biol. 234 (1971) 172–174. https://doi.org/10.1038/newbio234172a0.

[42] C.R. Woese, D.H. Dugre, S.A. Dugre, M. Kondo, W.C. Saxinger, On the fundamental nature and evolution of the genetic code, Cold Spring Harb. Symp. Quant. Biol. 31 (1966) 723–736. https://doi.org/10.1101/SQB.1966.031.01.093.

[43] S.R. Pelc, M.G. Welton, Stereochemical relationship between coding triplets and amino-acids, Nature. 209 (1966) 868–870. https://doi.org/10.1038/209868a0.

[44] P. Dunnill, Triplet nucleotide-amino-acid pairing; a stereochemical basis for the division between protein and non-protein amino-acids, Nature. 210 (1966) 1265–1267. https://doi.org/10.1038/2101267a0.

[45] A.S. Novozhilov, E.V. Koonin, Exceptional error minimization in putative primordial genetic codes, Biol. Direct. 4 (2009) 44. https://doi.org/10.1186/1745-6150-4-44.

[46] D.H. Ardell, On error minimization in a sequential origin of the standard genetic code, J. Mol. Evol. 47 (1998) 1–13. https://doi.org/10.1007/pl00006356.

[47] B. Kumar, S. Saini, Analysis of the optimality of the standard genetic code, Mol. Biosyst. 12 (2016) 2642–2651. https://doi.org/10.1039/C6MB00262E.

[48] M. Hollstein, D. Sidransky, B. Vogelstein, C.C. Harris, p53 mutations in human cancers, Science. 253 (1991) 49–53. https://doi.org/10.1126/science.1905840.

[49] M. Olivier, M. Hollstein, P. Hainaut, TP53 mutations in human cancers: origins, consequences, and clinical use, Cold Spring Harb. Perspect. Biol. 2 (2010) a001008. https://doi.org/10.1101/cshperspect.a001008.

[50] A.J. Levine, M. Oren, The first 30 years of p53: growing ever more complex, Nat. Rev. Cancer. 9 (2009) 749–758. https://doi.org/10.1038/nrc2723.

[51] A.N. Bullock, A.R. Fersht, Rescuing the function of mutant p53, Nat. Rev. Cancer. 1 (2001) 68–76. https://doi.org/10.1101/cshperspect.a001008.

[52] E.H. Baugh, H. Ke, A.J. Levine, R.A. Bonneau, C.S. Chan, Why are there hotspot mutations in the TP53 gene in human cancers?, Cell Death Differ. 25 (2018) 154–160. https://doi.org/10.1038/cdd.2017.180.

[53] F.A. Olotu, M.E.S. Soliman, From mutational inactivation to aberrant gain-of-function: Unraveling the structural basis of mutant p53 oncogenic transition, J. Cell. Biochem. 119 (2018) 2646–2652. https://doi.org/10.1002/jcb.26430.

[54] K. Sabapathy, D.P. Lane, Therapeutic targeting of p53: all mutants are equal, but some mutants are more equal than others, Nat. Rev. Clin. Oncol. 15 (2018) 13–30. https://doi.org/10.1038/nrclinonc.2017.151.

[55] T. Soussi, K.G. Wiman, TP53: an oncogene in disguise, Cell Death Differ. 22 (2015) 1239–1249. https://doi.org/10.1038/cdd.2015.53.

[56] P.G. Corrie, Cytotoxic chemotherapy: clinical aspects, Medicine (Baltimore). 39 (2011) 717–722. https://doi.org/10.1016/j.mpmed.2011.09.012.

`

[57] A.D. Wagner, N.L. Syn, M. Moehler, W. Grothe, W.P. Yong, B.-C. Tai, J. Ho, S. Unverzagt, Chemotherapy for advanced gastric cancer, Cochrane Database Syst. Rev. 8 (2017) CD004064. https://doi.org/10.1002/14651858.CD004064.pub4.

[58] M.C. Perry, The Chemotherapy source book, Williams & Wilkins, 1992.

[59] C.N. Andreassen, J. Alsner, Genetic variants and normal tissue toxicity after radiotherapy: A systematic review, Radiother. Oncol. 92 (2009) 299–309. https://doi.org/10.1016/j.radonc.2009.06.015.

[60] P.C. Davies, L. Demetrius, J.A. Tuszynski, Cancer as a dynamical phase transition, Theor. Biol. Med. Model. 8 (2011) 30. https://doi.org/10.1186/1742-4682-8-30.

[61] O.N. Vassiliev, A model of the radiation-induced bystander effect based on an analogy with ferromagnets. Application to modelling tissue response in a uniform field, Phys. A. 416 (2014) 242–251. https://doi.org/10.1016/j.physa.2014.08.052.

[62] G.G. Powathil, A.J. Munro, M.A. Chaplain, M. Swat, Bystander effects and their implications for clinical radiation therapy: Insights from multiscale in silico experiments, J. Theor. Biol. 401 (2016) 1–14. https://doi.org/10.1016/j.jtbi.2016.04.010.

[63] B.J. Blyth, P.J. Sykes, Radiation-induced bystander effects: what are they, and how relevant are they to human radiation exposures?, Radiat. Res. 176 (2011) 139–157. https://doi.org/10.1667/RR2548.1.

[64] D.J. Brenner, J.B. Little, R.K. Sachs, The bystander effect in radiation oncogenesis: II. A quantitative model, Radiat. Res. 155 (2001) 402–408. https://doi.org/10.1667/0033-7587(2001)155[0402:TBEIRO]2.0.CO;2.

[65] M.P. Little, R. Wakeford, The bystander effect in C3H 10T cells and radon-induced lung cancer, Radiat. Res. 156 (2001) 695–699. https://doi.org/10.1667/0033-7587(2001)156[0695:tbeicc]2.0.co;2.

[66] M.P. Little, J. a. N. Filipe, K.M. Prise, M. Folkard, O.V. Belyakov, A model for radiation-induced bystander effects, with allowance for spatial position and the effects of cell turnover, J. Theor. Biol. 232 (2005) 329–338. https://doi.org/10.1016/j.jtbi.2004.08.016.

[67] H. Nikjoo, I.K. Khvostunov, Biophysical model of the radiation-induced bystander effect, Int. J. Radiat. Biol. 79 (2003) 43–52. https://doi.org/10.1080/0955300021000034701.

[68] S.M. Bentzen, S.L. Tucker, Quantifying the position and steepness of radiation dose-response curves, Int. J. Radiat. Biol. 71 (1997) 531–542. https://doi.org/10.1080/095530097143860.

[69] C.Y. Chen, Y.N. Chang, P. Ryan, M. Linscott, G.J. McGarrity, Y.L. Chiang, Effect of herpes simplex virus thymidine kinase expression levels on ganciclovir-mediated cytotoxicity and the “bystander effect,” Hum. Gene Ther. 6 (1995) 1467–1476. https://doi.org/10.1089/hum.1995.6.11-1467.

`

[70] S.M. Freeman, C.N. Abboud, K.A. Whartenby, C.H. Packman, D.S. Koeplin, F.L. Moolten, G.N. Abraham, The “bystander effect”: tumor regression when a fraction of the tumor mass is genetically modified, Cancer Res. 53 (1993) 5274–5283.

[71] J. Fick, F.G. Barker, P. Dazin, E.M. Westphale, E.C. Beyer, M.A. Israel, The extent of heterocellular communication mediated by gap junctions is predictive of bystander tumor cytotoxicity in vitro, Proc. Natl. Acad. Sci. U. S. A. 92 (1995) 11071–11075. https://doi.org/10.1073/pnas.92.24.11071.

[72] C.N. Yang, The Spontaneous Magnetization of a Two-Dimensional Ising Model, Phys. Rev. 85 (1952) 808–816. https://doi.org/10.1103/PhysRev.85.808.

[73] J.-C. Tolédano, P. Tolédano, The Landau Theory of Phase Transitions: Application to Structural, Incommensurate, Magnetic and Liquid Crystal Systems, 1987. https://doi.org/10.1142/0215.

[74] R.A. Cowley, Structural phase transitions I. Landau theory, Adv. Phys. 29 (1980) 1–110. https://doi.org/10.1080/00018738000101346.

[75] J.R. Zimmerman, M.R. Foster, Standardization of N.M.R. High Resolution Spectra, J. Phys. Chem. 61 (1957) 282–289. https://doi.org/10.1021/j150549a006.

[76] J. Selinger, Introduction to the Theory of Soft Matter: From Ideal Gases to Liquid Crystals, 1st ed., Springer International Publishing, 2016.

[77] S. Torquato, Toward an Ising Model of Cancer and Beyond, Phys. Biol. 8 (2011) 015017. https://doi.org/10.1088/1478-3975/8/1/015017.

[78] J.M. Yeomans, Statistical mechanics of phase transitions, Clarendon Press, 1991.

[79] N. Goldenfeld, Lectures on phase transitions and the renormalization group, 1st ed., Westview Press, 1992.

[80] S.-K. Ma, Statistical Mechanics, World Scientific Publishing Co Inc, 1985.

[81] H.G. Katzgraber, Introduction to Monte Carlo Methods, ArXiv09051629 Cond-Mat Phys. (2009). http://arxiv.org/abs/0905.1629 (accessed August 20, 2019).

[82] S.R.A. Salinas, Introduction to statistical physics, Springer, New York :, 2001.

[83] J.C.M. Uitdehaag, J.A.D.M. de Roos, M.B.W. Prinsen, N. Willemsen-Seegers, J.R.F. de Vetter, J. Dylus, A.M. van Doornmalen, J. Kooijman, M. Sawa, S.J.C. van Gerwen, J. de Man, R.C. Buijsman, G.J.R. Zaman, Cell Panel Profiling Reveals Conserved Therapeutic Clusters and Differentiates the Mechanism of Action of Different PI3K/mTOR, Aurora Kinase and EZH2 Inhibitors, Mol. Cancer Ther. 15 (2016) 3097–3109. https://doi.org/10.1158/1535-7163.MCT-16-0403.

[84] J.C.M. Uitdehaag, J.A.D.M. de Roos, A.M. van Doornmalen, M.B.W. Prinsen, J. de Man, Y. Tanizawa, Y. Kawase, K. Yoshino, R.C. Buijsman, G.J.R. Zaman, Comparison of the cancer gene targeting and biochemical selectivities of all targeted kinase inhibitors approved for clinical use, PloS One. 9 (2014) e92146. https://doi.org/10.1371/journal.pone.0092146.

[85] R.T. Hinegardner, J. Engelberg, Rationale for a universal genetic code, Science. 142 (1963) 1083–1085. https://doi.org/10.1126/science.142.3595.1083.

[86] E.V. Koonin, A.S. Novozhilov, Origin and evolution of the genetic code: the universal enigma, IUBMB Life. 61 (2009) 99–111. https://doi.org/10/cxscpj.

[87] A.J. Griffiths, J.H. Miller, D.T. Suzuki, R.C. Lewontin, W.M. Gelbart, Universality of genetic information transfer, Introd. Genet. Anal. 7th Ed. (2000). https://www.ncbi.nlm.nih.gov/books/NBK21915/ (accessed January 1, 2020).

[88] N. Saitou, Introduction to Evolutionary Genomics, 2013 edition, Springer, London ; New York, 2014.

[89] R. Lásztity, M. Hidvegi, eds., Amino Acid Composition and Biological Value of Cereal Proteins: Proceedings of the International Association for Cereal Chemistry Symposium on Amino ... and Biological Value of Cereal Proteins, Softcover reprint of the original 1st ed. 1985 edition, Springer, Dordrecht, 2012.

[90] J.B. Reece, L.A. Urry, M.L. Cain, S.A. Wasserman, P.V. Minorsky, R.B. Jackson, Campbell Biology, 9 edition, Pearson, Boston, 2010.

[91] A. Ambrogelly, S. Palioura, D. Söll, Natural expansion of the genetic code, Nat. Chem. Biol. 3 (2007) 29–35. https://doi.org/10.1038/nchembio847.

[92] R. Owczarzy, P.M. Vallone, F.J. Gallo, T.M. Paner, M.J. Lane, A.S. Benight, Predicting sequence-dependent melting stability of short duplex DNA oligomers, Biopolymers. 44 (1997) 217–239. https://doi.org/10.1002/(SICI)1097-0282(1997)44:3<217::AID-BIP3>3.0.CO;2-Y.

[93] C.R. Woese, On the evolution of the genetic code., Proc. Natl. Acad. Sci. U. S. A. 54 (1965) 1546–1552.

[94] E. Szathmáry, Four letters in the genetic alphabet: a frozen evolutionary optimum?, Proc. Biol. Sci. 245 (1991) 91–99. https://doi.org/10.1098/rspb.1991.0093.

[95] E. Szathmáry, Why are there four letters in the genetic alphabet?, Nat. Rev. Genet. 4 (2003) 995–1001. https://doi.org/10.1038/nrg1231.

[96] A.L. Weber, S.L. Miller, Reasons for the occurrence of the twenty coded protein amino acids, J. Mol. Evol. 17 (1981) 273–284. https://doi.org/10.1007/BF01795749.

`

[97] Y. Lu, S. Freeland, On the evolution of the standard amino-acid alphabet, Genome Biol. 7 (2006) 102. https://doi.org/10.1186/gb-2006-7-1-102.

[98] M. Yčas, On earlier states of the biochemical system, J. Theor. Biol. 44 (1974) 145–160. https://doi.org/10.1016/S0022-5193(74)80035-4.

[99] M. Betts, R.B. Russell, Amino Acid Properties and Consequences of Substitutions, in: 2003. https://doi.org/10.1002/9780470059180.ch13.

[100] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G.A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H.P. Hratchian, A.F. Izmaylov, J. Bloino, G. Zheng, J.L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J.A. Montgomery, Jr., J.E. Peralta, F. Ogliaro, M. Bearpark, J.J. Heyd, E. Brothers, K.N. Kudin, V.N. Staroverov, T. Keith, R. Kobayashi, J.K.J. Normand, K. Raghavachari, A. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J.M. Millam, M. Klene, J.E. Knox, J.B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, R.L. Martin, K. Morokuma, V.G. Zakrzewski, G.A. Voth, P. Salvador, J.J. Dannenberg, S. Dapprich, A.D. Daniels, O. Farkas, J.B. Foresman, J.V. Ortiz, J. Cioslowski,, D.J. Fox, Gaussian 09 Citation Revision E.01, Inc Wallingford CT. (2013).

[101] M.S. Gordon, M.W. Schmidt, Advances in electronic structure theory: GAMESS a decade later, in: C.E. Dykstra, G. Frenking, K.S. Kim, G.E. Scuseria (Eds.), Theory Appl. Comput. Chem., Elsevier, Amsterdam, 2005: pp. 1167–1189. https://doi.org/10.1016/B978-044451719-7/50084-6.

[102] M.W. Schmidt, K.K. Baldridge, J.A. Boatz, S.T. Elbert, M.S. Gordon, J.H. Jensen, S. Koseki, N. Matsunaga, K.A. Nguyen, S. Su, T.L. Windus, M. Dupuis, J.A. Montgomery, General atomic and molecular electronic structure system, J. Comput. Chem. 14 (1993) 1347–1363. https://doi.org/10.1002/jcc.540141112.

[103] S. Shen, B. Kai, J. Ruan, J. Torin Huzil, E. Carpenter, J.A. Tuszynski, Probabilistic analysis of the frequencies of amino acid pairs within characterized protein sequences, Phys. Stat. Mech. Its Appl. 370 (2006) 651–662. https://doi.org/10.1016/j.physa.2006.03.004.

[104] T. UniProt Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 46 (2018) 2699. https://doi.org/10.1093/nar/gky092.

[105] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, M. Schneider, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, Nucleic Acids Res. 31 (2003) 365–370. https://doi.org/10.1093/nar/gkg095.

[106] National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988]

– [cited 2020]. Available from: https://www.ncbi.nlm.nih.gov/, (n.d.). https://www.ncbi.nlm.nih.gov/ (accessed March 31, 2020).

[107] NCBI Resource Coordinators, Database resources of the National Center for Biotechnology Information, Nucleic Acids Res. 46 (2018) D8–D13. https://doi.org/10.1093/nar/gkx1095.

[108] E.A.J. Mccullough, Numerical Hartree-Fock methods for diatomic molecules: a partial-wave expansion approach, Numer. Hartree-Fock Methods Diatomic Mol. Partial-Wave Expans. Approach. 4 (1986) 265–312. https://doi.org/10.1016/0167-7977(86)90020-1.

[109] J. Kobus, L. Laaksonen, D. Sundholm, A numerical Hartree-Fock program for diatomic molecules, Comput. Phys. Commun. 98 (1996) 346–358. https://doi.org/10.1016/0010-4655(96)00098-7.

[110] C.G. Darwin, Douglas Rayner Hartree. 1897-1958, Biogr. Mem. Fellows R. Soc. 4 (1958) 103–116. https://doi.org/10.1098/rsbm.1958.0010.

[111] J.J.P. Stewart, Optimization of parameters for semiempirical methods I. Method, J. Comput. Chem. 10 (1989) 209–220. https://doi.org/10.1002/jcc.540100208.

[112] J.J.P. Stewart, Optimization of parameters for semiempirical methods II. Applications, J. Comput. Chem. 10 (1989) 221–264. https://doi.org/10.1002/jcc.540100209.

[113] J.J.P. Stewart, Optimization of parameters for semiempirical methods. III Extension of PM3 to Be, Mg, Zn, Ga, Ge, As, Se, Cd, In, Sn, Sb, Te, Hg, Tl, Pb, and Bi, J. Comput. Chem. 12 (1991) 320–341. https://doi.org/10.1002/jcc.540120306.

[114] J.J.P. Stewart, Optimization of parameters for semiempirical methods IV: extension of MNDO, AM1, and PM3 to more main group elements, J. Mol. Model. 10 (2004) 155–164. https://doi.org/10.1007/s00894-004-0183-z.

[115] J. SantaLucia, A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 1460–1465. https://doi.org/10.1073/pnas.95.4.1460.

[116] E. Protozanova, P. Yakovchuk, M.D. Frank-Kamenetskii, Stacked-unstacked equilibrium at the nick site of DNA, J. Mol. Biol. 342 (2004) 775–785. https://doi.org/10.1016/j.jmb.2004.07.075.

[117] J. SantaLucia, H.T. Allawi, P.A. Seneviratne, Improved nearest-neighbor parameters for predicting DNA duplex stability, Biochemistry. 35 (1996) 3555–3562. https://doi.org/10.1021/bi951907q.

[118] Z. Ignatova, I.M. Martínez Pérez, K.-H. Zimmermann, DNA computing models, Springer, New York, NY, 2008.

[119] R.R. Sinden, DNA Structure and Function, 1 edition, Academic Press, San Diego, 1994.

`

[120] S. Arrhenius, Über die Dissociationswärme und den Einfluß der Temperatur auf den Dissociationsgrad der Elektrolyte, Z. Phys. Chem. 4 (1889) 96–116. https://doi.org/10.1515/zpch-1889-0408.

[121] S. Arrhenius, Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren, Z. Phys. Chem. 4 (1889) 226–248. https://doi.org/10.1515/zpch-1889-0416.

[122] K.A. Connors, Chemical Kinetics: The Study of Reaction Rates in Solution, John Wiley & Sons, 1990.

[123] A.D. Becke, A new mixing of Hartree–Fock and local density-functional theories, J. Chem. Phys. 98 (1993) 1372–1377. https://doi.org/10.1063/1.464304.

[124] A.D. Becke, Density-functional exchange-energy approximation with correct asymptotic behavior, Phys. Rev. A. 38 (1988) 3098–3100. https://doi.org/10.1103/PhysRevA.38.3098.

[125] C. Lee, W. Yang, R.G. Parr, Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density, Phys Rev B. 37 (1988) 785–789. https://doi.org/10.1103/PhysRevB.37.785.

[126] S.H. Vosko, L. Wilk, M. Nusair, Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis, Can. J. Phys. 58 (1980) 1200–1211. https://doi.org/10.1139/p80-159.

[127] F.J. Devlin, J.W. Finley, P.J. Stephens, M.J. Frisch, Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields: A Comparison of Local, Nonlocal, and Hybrid Density Functionals, J. Phys. Chem. 99 (1995) 16883–16902. https://doi.org/10.1021/j100046a014.

[128] R. Krishnan, J.S. Binkley, R. Seeger, J.A. Pople, Self-Consistent Molecular Orbital Methods. 20. Basis set for correlated wave-functions, J. Chem. Phys. 72 (1980) 650–654. https://doi.org/10.1063/1.438955.

[129] M.J. Frisch, J.A. Pople, J.S. Binkley, Self-consistent molecular orbital methods 25. Supplementary functions for Gaussian basis sets, J. Chem. Phys. 80 (1984) 3265–3269. https://doi.org/10.1063/1.447079.

[130] T. Clark, J. Chandrasekhar, G.W. Spitznagel, P.V.R. Schleyer, Efficient diffuse function-augmented basis sets for anion calculations. III. The 3-21+G basis set for first-row elements, Li–F, J. Comput. Chem. 4 (1983) 294–301. https://doi.org/10.1002/jcc.540040303.

[131] E. Koch, Mean-Field Theory: Hartree-Fock and BCS, (n.d.) 34.

[132] C. Froese Fischer, General Hartree-Fock program, Comput. Phys. Commun. 43 (1987) 355–365. https://doi.org/10.1016/0010-4655(87)90053-1.

`

[133] J.C. Slater, The Self Consistent Field and the Structure of Atoms, Phys. Rev. 32 (1928) 339–348. https://doi.org/10.1103/PhysRev.32.339.

[134] J.A. Gaunt, A Theory of Hartree's Atomic Fields, Math. Proc. Camb. Philos. Soc. 24 (1928) 328–342. https://doi.org/10.1017/S0305004100015851.

[135] J.C. Slater, Note on Hartree's Method, Phys. Rev. 35 (1930) 210–211. https://doi.org/10.1103/PhysRev.35.210.2.

[136] I. Friedberg, H. Margalit, Persistently conserved positions in structurally similar, sequence dissimilar proteins: Roles in preserving protein fold and function, Protein Sci. Publ. Protein Soc. 11 (2002) 350–360. https://doi.org/10.1110/ps.18602.

[137] R.A. Fisher, The genetical theory of natural selection, J. H. Bennett (A complete variorum ed.), Oxford, UK: Oxford University Press, 1999.

[138] S.J. Freeland, L.D. Hurst, The Genetic Code Is One in a Million, J. Mol. Evol. 47 (1998) 238–248. https://doi.org/10.1007/PL00006381.

[139] J.T.-F. Wong, The evolution of a universal genetic code, Proc. Natl. Acad. Sci. 73 (1976) 2336–2340. https://doi.org/10.1073/pnas.73.7.2336.

[140] K.F. Dyer, The Quiet Revolution: A New Synthesis of Biological Knowledge, J. Biol. Educ. 5 (1971) 15–24. https://doi.org/10.1080/00219266.1971.9653663.

[141] P.C.W. Davies, E. Rieper, J.A. Tuszynski, Self-organization and entropy reduction in a living cell, Biosystems. 111 (2013) 1–10. https://doi.org/10.1016/j.biosystems.2012.10.005.

[142] E. Schrödinger, What is Life? The Physical Aspect of the Living Cell., Cambridge University Press, Cambridge., Cambridge, 1967.

[143] J.M. Berg, J.L. Tymoczko, L. Stryer, J.M. Berg, J.L. Tymoczko, L. Stryer, Biochemistry, 5th ed., W H Freeman, 2002.

[144] F. Syberg, Y. Suveyzdis, C. Kötting, K. Gerwert, E. Hofmann, Time-resolved Fourier transform infrared spectroscopy of the nucleotide-binding domain from the ATP-binding Cassette transporter MsbA: ATP hydrolysis is the rate-limiting step in the catalytic cycle, J. Biol. Chem. 287 (2012) 23923–23931. https://doi.org/10.1074/jbc.M112.359208.

[145] T.V. Zharova, A.D. Vinogradov, Proton-translocating ATP-synthase of Paracoccus denitrificans: ATP-hydrolytic activity, Biochem. Biokhimiia. 68 (2003) 1101–1108. https://doi.org/10.1023/A:1026306611821.

[146] C. Bergman, Y. Kashiwaya, R.L. Veech, The effect of pH and free Mg2+ on ATP linked enzymes and the calculation of Gibbs free energy of ATP hydrolysis, J. Phys. Chem. B. 114 (2010) 16137–16146. https://doi.org/10.1021/jp105723r.

`

[147] J. Rosing, E.C. Slater, The value of G degrees for the hydrolysis of ATP, Biochim. Biophys. Acta. 267 (1972) 275–290. https://doi.org/10.1016/0005-2728(72)90116-8.

[148] S. Soboll, R. Scholz, H.W. Heldt, Subcellular metabolite concentrations. Dependence of mitochondrial and cytosolic ATP systems on the metabolic state of perfused rat liver, Eur. J. Biochem. 87 (1978) 377–390. https://doi.org/10.1111/j.1432-1033.1978.tb12387.x.

[149] H. Wackerhage, U. Hoffmann, D. Essfeld, D. Leyk, K. Mueller, J. Zange, Recovery of free ADP, Pi, and free energy of ATP hydrolysis in human skeletal muscle, J. Appl. Physiol. Bethesda Md 1985. 85 (1998) 2140–2145. https://doi.org/10.1152/jappl.1998.85.6.2140.

[150] Q.H. Tran, G. Unden, Changes in the proton potential and the cellular energetics of Escherichia coli during growth by aerobic and anaerobic respiration or by fermentation, Eur. J. Biochem. 251 (1998) 538–543. https://doi.org/10.1046/j.1432-1327.1998.2510538.x.

[151] M.J. Schnitzer, S.M. Block, Kinesin hydrolyses one ATP per 8-nm step, Nature. 388 (1997) 386–390. https://doi.org/10.1038/41111.

[152] M. Kinoshita, Importance of Translational Entropy of Water in Biological Self-Assembly Processes like Protein Folding, Int. J. Mol. Sci. 10 (2009) 1064–1080. https://doi.org/10.3390/ijms10031064.

[153] G.R. Lambert, Enzymic editing mechanisms and the origin of biological information transfer, J. Theor. Biol. 107 (1984) 387–403. https://doi.org/10.1016/S0022-5193(84)80098-3.

[154] J.T. Kim, T. Martinetz, D. Polani, Bioinformatic principles underlying the information content of transcription factor binding sites, J. Theor. Biol. 220 (2003) 529–544. https://doi.org/10.1006/jtbi.2003.3153.

[155] M.A. Savageau, Proteins of Escherichia coli come in sizes that are multiples of 14 kDa: domain concepts and evolutionary implications., Proc. Natl. Acad. Sci. U. S. A. 83 (1986) 1198–1202. https://doi.org/10.1073/pnas.83.5.1198.

[156] A.I. Su, T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M.P. Cooke, J.R. Walker, J.B. Hogenesch, A gene atlas of the mouse and human protein-encoding transcriptomes, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 6062–6067. https://doi.org/10.1073/pnas.0400782101.

[157] P. McClurg, J. Janes, C. Wu, D.L. Delano, J.R. Walker, S. Batalov, J.S. Takahashi, K. Shimomura, A. Kohsaka, J. Bass, T. Wiltshire, A.I. Su, Genomewide association analysis in diverse inbred mice: power and population structure, Genetics. 176 (2007) 675–683. https://doi.org/10.1534/genetics.106.066241.

[158] C. Wu, D.L. Delano, N. Mitro, S.V. Su, J. Janes, P. McClurg, S. Batalov, G.L. Welch, J. Zhang, A.P. Orth, J.R. Walker, R.J. Glynne, M.P. Cooke, J.S. Takahashi, K. Shimomura, A. Kohsaka, J. Bass, E. Saez, T. Wiltshire, A.I. Su, Gene set enrichment in eQTL data

identifies novel annotations and pathway regulators, PLoS Genet. 4 (2008) e1000070. https://doi.org/10.1371/journal.pgen.1000070.

[159] J.E. Lattin, K. Schroder, A.I. Su, J.R. Walker, J. Zhang, T. Wiltshire, K. Saijo, C.K. Glass, D.A. Hume, S. Kellie, M.J. Sweet, Expression analysis of G Protein-Coupled Receptors in mouse macrophages, Immunome Res. 4 (2008) 5. https://doi.org/10.1186/1745-7580-4-5.

[160] E.E. Zhang, A.C. Liu, T. Hirota, L.J. Miraglia, G. Welch, P.Y. Pongsawakul, X. Liu, A. Atwood, J.W. Huss, J. Janes, A.I. Su, J.B. Hogenesch, S.A. Kay, A genome-wide RNAi screen for modifiers of the circadian clock in human cells, Cell. 139 (2009) 199–210. https://doi.org/10.1016/j.cell.2009.08.031.

[161] C. Wu, C. Orozco, J. Boyer, M. Leglise, J. Goodale, S. Batalov, C.L. Hodge, J. Haase, J. Janes, J.W. Huss, A.I. Su, BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources, Genome Biol. 10 (2009) R130. https://doi.org/10.1186/gb-2009-10-11-r130.

[162] C. Wu, I. Macleod, A.I. Su, BioGPS and MyGene.info: organizing online, gene-centric information, Nucleic Acids Res. 41 (2013) D561-565. https://doi.org/10.1093/nar/gks1114.

[163] C. Wu, X. Jin, G. Tsueng, C. Afrasiabi, A.I. Su, BioGPS: building your own mash-up of gene annotations and expression profiles, Nucleic Acids Res. 44 (2016) D313-316. https://doi.org/10.1093/nar/gkv1104.

[164] C.E. Shannon, A Mathematical Theory of Communication, Bell Syst. Tech. J. 27 (1948) 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[165] M. Cobb, 60 years ago, Francis Crick changed the logic of biology, PLOS Biol. 15 (2017) e2003243. https://doi.org/10.1371/journal.pbio.2003243.

[166] Central dogma reversed, Nature. 226 (1970) 1198–1199. https://doi.org/10.1038/2261198a0.

[167] R. Knight, L. Landweber, M. Yarus, Tests of a Stereochemical Genetic Code, Landes Bioscience, 2013. https://www.ncbi.nlm.nih.gov/books/NBK6584/ (accessed August 15, 2019).

[168] I. Majerfeld, D. Puthenvedu, M. Yarus, RNA affinity for molecular L-histidine; genetic code origins, J. Mol. Evol. 61 (2005) 226–235. https://doi.org/10.1007/s00239-004-0360-9.

[169] I. Majerfeld, M. Yarus, Isoleucine:RNA sites with associated coding sequences, RNA N. Y. N. 4 (1998) 471–478.

[170] R.M. Turk-MacLeod, D. Puthenvedu, I. Majerfeld, M. Yarus, The Plausibility of RNA-Templated Peptides: Simultaneous RNA Affinity for Adjacent Peptide Side Chains, J. Mol. Evol. 74 (2012) 217–225. https://doi.org/10.1007/s00239-012-9501-8.

`

[171] T. Janas, J.J. Widmann, R. Knight, M. Yarus, Simple, recurring RNA binding sites for L-arginine, RNA. 16 (2010) 805–816. https://doi.org/10.1261/rna.1979410.

[172] G.J. Connell, M. Illangesekare, M. Yarus, Three small ribooligonucleotides with specific arginine sites, Biochemistry. 32 (1993) 5497–5502. https://doi.org/10.1021/bi00072a002.

[173] I. Majerfeld, M. Yarus, A diminutive and specific RNA binding site for L-tryptophan, Nucleic Acids Res. 33 (2005) 5482–5493. https://doi.org/10.1093/nar/gki861.

[174] C. Lozupone, S. Changayil, I. Majerfeld, M. Yarus, Selection of the simplest RNA that binds isoleucine, RNA. 9 (2003) 1315–1322. https://doi.org/10.1261/rna.5114503.

[175] M. Legiewicz, M. Yarus, A More Complex Isoleucine Aptamer with a Cognate Triplet, J. Biol. Chem. 280 (2005) 19815–19822. https://doi.org/10.1074/jbc.M502329200.

[176] T. Lengauer, M. Rarey, Computational methods for biomolecular docking, Curr. Opin. Struct. Biol. 6 (1996) 402–406. https://doi.org/10.1016/S0959-440X(96)80061-3.

[177] S. Izrailev, S. Stepaniants, B. Isralewitz, D. Kosztin, H. Lu, F. Molnar, W. Wriggers, K. Schulten, Steered Molecular Dynamics, in: P. Deuflhard, J. Hermans, B. Leimkuhler, A.E. Mark, S. Reich, R.D. Skeel (Eds.), Comput. Mol. Dyn. Chall. Methods Ideas, Springer Berlin Heidelberg, 1999: pp. 39–65. https://doi.org/10.1016/S0006-3495(98)77556-3.

[178] H. Grubmüller, B. Heymann, P. Tavan, Ligand Binding: Molecular Mechanics Calculation of the Streptavidin-Biotin Rupture Force, Science. 271 (1996) 997–9. https://doi.org/10.1126/science.271.5251.997.

[179] S. Stepaniants, S. Izrailev, K. Schulten, Extraction of Lipids from Phospholipid Membranes by Steered Molecular Dynamics, Mol. Model. Annu. 3 (1997) 473–475. https://doi.org/10.1007/s008940050065.

[180] S.J. Marrink, O. Berger, P. Tieleman, F. Jähnig, Adhesion forces of lipids in a phospholipid membrane studied by molecular dynamics simulations., Biophys. J. 74 (1998) 931–943. https://doi.org/10.1016/S0006-3495(98)74016-0.

[181] H. Lu, B. Isralewitz, A. Krammer, V. Vogel, K. Schulten, Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation., Biophys. J. 75 (1998) 662–671. https://doi.org/10.1016/S0006-3495(98)77556-3.

[182] D.A. Case, J.T. Berryman, R.M. Betz, D.S. Cerutti, T.E. Cheatham, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K.M. Merz, G. Monard, P. Needham, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, R. Salomon-Ferrer, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, D.M. York, P.A. Kollman, AMBER 2014, Univ. Calif. San Franc. (2015).

`

[183] H.R. Bureau, D.R.M. Jr, E. Hershkovits, S. Quirk, R. Hernandez, Constrained Unfolding of a Helical Peptide: Implicit versus Explicit Solvents, PLOS ONE. 10 (2015) e0127034. https://doi.org/10.1371/journal.pone.0127034.

[184] G. Ozer, S. Quirk, R. Hernandez, Adaptive steered molecular dynamics: validation of the selection criterion and benchmarking energetics in vacuum, J. Chem. Phys. 136 (2012) 215104. https://doi.org/10.1063/1.4725183.

[185] G. Ozer, T. Keyes, S. Quirk, R. Hernandez, Multiple branched adaptive steered molecular dynamics, J. Chem. Phys. 141 (2014) 064101. https://doi.org/10.1063/1.4891807.

[186] B. Leimkuhler, S. Reich, Simulating Hamiltonian Dynamics, Cambridge University Press, 2004.

[187] Y. Huang, H. Li, Y. Xiao, Using 3dRPC for RNA–protein complex structure prediction, Biophys. Rep. 2 (2016) 95–99. https://doi.org/10.1007/s41048-017-0034-y.

[188] Y. Huang, H. Li, Y. Xiao, 3dRPC: a web server for 3D RNA–protein structure prediction, Bioinformatics. 34 (2018) 1238–1240. https://doi.org/10.1093/bioinformatics/btx742.

[189] Y. Huang, S. Liu, D. Guo, L. Li, Y. Xiao, A novel protocol for three-dimensional structure prediction of RNA-protein complexes, Sci. Rep. 3 (2013) 1887. https://doi.org/10.1038/srep01887.

[190] H. Li, Y. Huang, Y. Xiao, A pair-conformation-dependent scoring function for evaluating 3D RNA-protein complex structures, PLOS ONE. 12 (2017) e0174662. https://doi.org/10.1371/journal.pone.0174662.

[191] F. Austin, U. Oyarbide, G. Massey, M. Grimes, S.J. Corey, Synonymous mutation in TP53 results in a cryptic splice site affecting its DNA binding site in an adolescent with two primary sarcomas, Pediatr. Blood Cancer. 64 (2017). https://doi.org/10.1002/pbc.26584.

[192] T. Hu, W. Banzhaf, Nonsynonymous to Synonymous Substitution Ratio k_a/k_s: Measurement for Rate of Evolution in Evolutionary Computation, in: G. Rudolph, T. Jansen, N. Beume, S. Lucas, C. Poloni (Eds.), Parallel Probl. Solving Nat. – PPSN X, Springer, Berlin, Heidelberg, 2008: pp. 448–457. https://doi.org/10.1007/978-3-540-87700-4_45.

[193] K. Karakostis, S. Vadivel Gnanasundram, I. López, A. Thermou, L. Wang, K. Nylander, V. Olivares-Illana, R. Fåhraeus, A single synonymous mutation determines the phosphorylation and stability of the nascent protein, J. Mol. Cell Biol. 11 (2019) 187–199. https://doi.org/10.1093/jmcb/mjy049.

[194] L. Loewe, Genetic mutation., Nat. Educ. 1 (2008) 113.

[195] A.C. Joerger, A.R. Fersht, The Tumor Suppressor p53: From Structures to Drug Discovery, Cold Spring Harb. Perspect. Biol. 2 (2010). https://doi.org/10.1101/cshperspect.a000919.

[196] S. Surget, M.P. Khoury, J.-C. Bourdon, Uncovering the role of p53 splice variants in human malignancy: a clinical perspective, OncoTargets Ther. 7 (2013) 57–68. https://doi.org/10.2147/OTT.S53876.

[197] F. Perri, S. Pisconti, G. Della Vittoria Scarpati, P53 mutations and cancer: a tight linkage, Ann. Transl. Med. 4 (2016). https://doi.org/10.21037/atm.2016.12.40.

[198] C.W. Lee, M. Arai, M.A. Martinez-Yamout, H.J. Dyson, P.E. Wright, Mapping the interactions of the p53 transactivation domain with the KIX domain of CBP, Biochemistry. 48 (2009) 2115–2124. https://doi.org/10.1021/bi802055v.

[199] C.W. Lee, M.A. Martinez-Yamout, H.J. Dyson, P.E. Wright, Structure of the p53 transactivation domain in complex with the nuclear receptor coactivator binding domain of CREB binding protein, Biochemistry. 49 (2010) 9964–9971. https://doi.org/10.1021/bi1012996.

[200] N. Raj, L.D. Attardi, The Transactivation Domains of the p53 Protein, Cold Spring Harb. Perspect. Med. 7 (2017). https://doi.org/10.1101/cshperspect.a026047.

[201] M. Ullah, P53 Mutational Signature in Cancer, Int. J. Vaccines Vaccin. Volume 4 (2017) 00070. https://doi.org/DOI: 10.15406/ijvv.2017.04.00070.

[202] K.K. Walker, A.J. Levine, Identification of a novel p53 functional domain that is necessary for efficient growth suppression, Proc. Natl. Acad. Sci. U. S. A. 93 (1996) 15335–15340. https://doi.org/10.1073/pnas.93.26.15335.

[203] Y. Cho, S. Gorina, P.D. Jeffrey, N.P. Pavletich, Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations, Science. 265 (1994) 346–355. https://doi.org/10.1126/science.8023157.

[204] R. Beckerman, C. Prives, Transcriptional regulation by p53, Cold Spring Harb. Perspect. Biol. 2 (2010) a000935. https://doi.org/10.1101/cshperspect.a000935.

[205] P. Chène, The role of tetramerization in p53 function, Oncogene. 20 (2001) 2611–2617. https://doi.org/10.1038/sj.onc.1204373.

[206] G. Gaglia, Y. Guan, J.V. Shah, G. Lahav, Activation and control of p53 tetramerization in individual living cells, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) 15497–15501. https://doi.org/10.1073/pnas.1311126110.

[207] O. Laptenko, D.R. Tong, J. Manfredi, C. Prives, The Tail That Wags the Dog: How the Disordered C-Terminal Domain Controls the Transcriptional Activities of the p53 Tumor-Suppressor Protein, Trends Biochem. Sci. 41 (2016) 1022–1034. https://doi.org/10.1016/j.tibs.2016.08.011.

[208] A.R. Blanden, X. Yu, S.N. Loh, A.J. Levine, D.R. Carpizo, Reactivating mutant p53 using small molecules as zinc metallochaperones: awakening a sleeping giant in cancer, Drug Discov. Today. 20 (2015) 1391–1397. https://doi.org/10.1016/j.drudis.2015.07.006.

[209] S. Kogan, D.R. Carpizo, Zinc Metallochaperones as Mutant p53 Reactivators: A New Paradigm in Cancer Therapeutics, Cancers. 10 (2018). https://doi.org/10.3390/cancers10060166.

[210] S.N. Loh, The missing zinc: p53 misfolding and cancer, Met. Integr. Biometal Sci. 2 (2010) 442–449. https://doi.org/10.1039/C003915B.

[211] C. Méplan, M.-J. Richard, P. Hainaut, Metalloregulation of the tumor suppressor protein p53: zinc mediates the renaturation of p53 after exposure to metal chelators in vitro and in intact cells, Oncogene. 19 (2000) 5227–5236. https://doi.org/10.1038/sj.onc.1203907.

[212] X. Yu, S. Kogan, Y. Chen, A.T. Tsang, T. Withers, H. Lin, J. Gilleran, B. Buckley, D. Moore, J. Bertino, C. Chan, S.D. Kimball, S.N. Loh, D.R. Carpizo, Zinc Metallochaperones Reactivate Mutant p53 Using an ON/OFF Switch Mechanism: A New Paradigm in Cancer Therapeutics, Clin. Cancer Res. 24 (2018) 4505–4517. https://doi.org/10.1158/1078-0432.CCR-18-0822.

[213] L. Bouaoun, D. Sonkin, M. Ardin, M. Hollstein, G. Byrnes, J. Zavadil, M. Olivier, TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data, Hum. Mutat. 37 (2016) 865–876. https://doi.org/10.1002/humu.23035.

[214] K.H. Vousden, X. Lu, Live or let die: the cell's response to p53, Nat. Rev. Cancer. 2 (2002) 594–604. https://doi.org/10.1038/nrc864.

[215] S. Madhusudan, M.R. Middleton, The emerging role of DNA repair proteins as predictive, prognostic and therapeutic targets in cancer, Cancer Treat. Rev. 31 (2005) 603–617. https://doi.org/10.1016/j.ctrv.2005.09.006.

[216] A. Comel, G. Sorrentino, V. Capaci, G.D. Sal, The cytoplasmic side of p53's oncosuppressive activities, FEBS Lett. 588 (2014) 2600–2609. https://doi.org/10.1038/sj.onc.1203413.

[217] M. Weinfeld, R.S. Mani, I. Abdou, R.D. Aceytuno, J.N.M. Glover, Tidying up loose ends: the role of polynucleotide kinase/phosphatase in DNA strand break repair, Trends Biochem. Sci. 36 (2011) 262–271. https://doi.org/10.1016/j.tibs.2011.01.006.

[218] D.W. Meek, Regulation of the p53 response and its relationship to cancer1, Biochem. J. 469 (2015) 325–346. https://doi.org/ttps://doi.org/10.1042/BJ20150517.

[219] L. Gatti, F. Zunino, Overview of tumor cell chemoresistance mechanisms, Methods Mol. Med. 111 (2005) 127–148. https://doi.org/10.1385/1-59259-889-7:127.

[220] D.W. Meek, C.W. Anderson, Posttranslational modification of p53: cooperative integrators of function, Cold Spring Harb. Perspect. Biol. 1 (2009) a000950. https://doi.org/10.1101/cshperspect.a000950.

[221] S. Kachalaki, M. Ebrahimi, L. Mohamed Khosroshahi, S. Mohammadinejad, B. Baradaran, Cancer chemoresistance; biochemical and molecular aspects: a brief overview, Eur. J.

`

Pharm. Sci. Off. J. Eur. Fed. Pharm. Sci. 89 (2016) 20–30. https://doi.org/10.1016/j.ejps.2016.03.025.

[222] C.L. Brooks, W. Gu, The impact of acetylation and deacetylation on the p53 pathway, Protein Cell. 2 (2011) 456–462. https://doi.org/10.1007/s13238-011-1063-9.

[223] K.A. Boehme, C. Blattner, Regulation of p53--insights into a complex process, Crit. Rev. Biochem. Mol. Biol. 44 (2009) 367–392. https://doi.org/10.3109/10409230903401507.

[224] I. Goldstein, V. Marcel, M. Olivier, M. Oren, V. Rotter, P. Hainaut, Understanding wild-type and mutant p53 activities in human cancer: new landmarks on the way to targeted therapies, Cancer Gene Ther. 18 (2011) 2–11. https://doi.org/10.1038/cgt.2010.63.

[225] S.-J. Wang, W. Gu, To Be, or Not to Be: Functional Dilemma of p53 metabolic regulation, Curr. Opin. Oncol. 26 (2014) 78–85. https://doi.org/10.1097/CCO.0000000000000024.

[226] S.I. Omar, J. Tuszynski, Ranking the Binding Energies of p53 Mutant Activators and Their ADMET Properties, Chem. Biol. Drug Des. 86 (2015) 163–172. https://doi.org/10.1111/cbdd.12480.

[227] S.I. Omar, M.G. Lepre, U. Morbiducci, M.A. Deriu, J.A. Tuszynski, Virtual screening using covalent docking to find activators for G245S mutant p53, PLoS ONE. 13 (2018). https://doi.org/10.1371/journal.pone.0200769.

[228] G. Chillemi, S. Kehrloesser, F. Bernassola, A. Desideri, V. Dötsch, A.J. Levine, G. Melino, Structural Evolution and Dynamics of the p53 Proteins, Cold Spring Harb. Perspect. Med. 7 (2017). https://doi.org/10.1101/cshperspect.a028308.

[229] C.D. Wassman, R. Baronio, Ö. Demir, B.D. Wallentine, C.-K. Chen, L.V. Hall, F. Salehi, D.-W. Lin, B.P. Chung, G.W. Hatfield, A. Richard Chamberlin, H. Luecke, R.H. Lathrop, P. Kaiser, R.E. Amaro, Computational identification of a transiently open L1/S3 pocket for reactivation of mutant p53, Nat. Commun. 4 (2013) 1407. https://doi.org/10.1038/ncomms2361.

[230] M.G. Lepre, S.I. Omar, G. Grasso, U. Morbiducci, M.A. Deriu, J.A. Tuszynski, Insights into the Effect of the G245S Single Point Mutation on the Structure of p53 and the Binding of the Protein to DNA, Mol. Basel Switz. 22 (2017). https://doi.org/10.3390/molecules22081358.

[231] p14 ARF links the tumour suppressors RB and p53 | Nature, (n.d.). https://www.nature.com/articles/25867 (accessed September 14, 2019).

[232] S. Bell, C. Klein, L. Müller, S. Hansen, J. Buchner, p53 Contains Large Unstructured Regions in its Native State, J. Mol. Biol. 322 (2002) 917–927. https://doi.org/10.1016/S0022-2836(02)00848-3.

[233] J.R. Bischoff, D.H. Kirn, A. Williams, C. Heise, S. Horn, M. Muna, L. Ng, J.A. Nye, A. Sampson-Johannes, A. Fattaey, F. McCormick, An adenovirus mutant that replicates

selectively in p53-deficient human tumor cells, Science. 274 (1996) 373–376. https://doi.org/10.1126/science.274.5286.373.

[234] M.V. Blagosklonny, P53: an ubiquitous target of anticancer drugs, Int. J. Cancer. 98 (2002) 161–166. https://doi.org/10.1002/ijc.10158.

[235] F. McCormick, Cancer gene therapy: fringe or cutting edge?, Nat. Rev. Cancer. 1 (2001) 130–141. https://doi.org/10.1038/35101008.

[236] T. Strachan, A.P. Read, Human molecular genetics 2, Wiley-Liss, New York, 1999.

[237] B. Vogelstein, D. Lane, A.J. Levine, Surfing the p53 network, Nature. 408 (2000) 307–310. https://doi.org/10.1038/35042675.

[238] M. Tollis, A.M. Boddy, C.C. Maley, Peto's Paradox: how has evolution solved the problem of cancer prevention?, BMC Biol. 15 (2017) 60. https://doi.org/10.1186/s12915-017-0401-7.

[239] L.M. Abegglen, A.F. Caulin, A. Chan, K. Lee, R. Robinson, M.S. Campbell, W.K. Kiso, D.L. Schmitt, P.J. Waddell, S. Bhaskara, S.T. Jensen, C.C. Maley, J.D. Schiffman, Potential Mechanisms for Cancer Resistance in Elephants and Comparative Cellular Response to DNA Damage in Humans, JAMA. 314 (2015) 1850–1860. https://doi.org/10.1001/jama.2015.13134.

[240] S.S. Mello, L.D. Attardi, Not all p53 gain-of-function mutants are created equal, Cell Death Differ. 20 (2013) 855–857. https://doi.org/10.1038/cdd.2013.53.

[241] M. Ullah, P53 Mutational Signature in Cancer, Int. J. Vaccines Vaccin. (2017). https://doi.org/10.15406/ijvv.2017.04.00070.

[242] N. Shirai, T. Tsukamoto, M. Yamamoto, T. Iidaka, H. Sakai, T. Yanai, T. Masegi, L.A. Donehower, M. Tatematsu, Elevated susceptibility of the p53 knockout mouse esophagus to methyl- N -amylnitrosamine carcinogenesis, Carcinogenesis. 23 (2002) 1541–1547. https://doi.org/10.1093/carcin/23.9.1541.

[243] M. Yamamoto, T. Tsukamoto, H. Sakai, N. Shirai, H. Ohgaki, C. Furihata, L.A. Donehower, K. Yoshida, M. Tatematsu, p53 knockout mice (-/-) are more susceptible than (+/-) or (+/+) mice to N-methyl-N-nitrosourea stomach carcinogenesis, Carcinogenesis. 21 (2000) 1891–1897. https://doi.org/10.1093/carcin/21.10.1891.

[244] A.C. Blackburn, D.J. Jerry, Knockout and transgenic mice of Trp53: what have we learned about p53 in breast cancer?, Breast Cancer Res. 4 (2002) 101–111. https://doi.org/10.1186/bcr427.

[245] S. North, F. El-Ghissassi, O. Pluquet, G. Verhaegh, P. Hainaut, The cytoprotective aminothiol WR1065 activates p21 waf-1 and down regulates cell cycle progression through a p53-dependent pathway, Oncogene. 19 (2000) 1206–1214. https://doi.org/10/c8gh7g.

`

[246] S. Emamzadah, L. Tropia, T.D. Halazonetis, Crystal structure of a multidomain human p53 tetramer bound to the natural CDKN1A (p21) p53-response element, Mol. Cancer Res. MCR. 9 (2011) 1493–1499. https://doi.org/10.1158/1541-7786.MCR-11-0351.

[247] A.C. Minella, J. Swanger, E. Bryant, M. Welcker, H. Hwang, B.E. Clurman, p53 and p21 form an inducible barrier that protects cells against cyclin E-cdk2 deregulation, Curr. Biol. CB. 12 (2002) 1817–1827. https://doi.org/10.1016/S0960-9822(02)01225-3.

[248] G. He, Z.H. Siddik, Z. Huang, R. Wang, J. Koomen, R. Kobayashi, A.R. Khokhar, J. Kuang, Induction of p21 by p53 following DNA damage inhibits both Cdk4 and Cdk2 activities, Oncogene. 24 (2005) 2929–2943. https://doi.org/10.1038/sj.onc.1208474.

[249] N.P. Pavletich, K.A. Chambers, C.O. Pabo, The DNA-binding domain of p53 contains the four conserved regions and the major mutation hot spots, Genes Dev. 7 (1993) 2556–2564. https://doi.org/10.1101/gad.7.12b.2556.

[250] D.N. Cooper, Nature Publishing Group, Nature encyclopedia of the human genome, Nature Pub. Group, London; New York, 2003.

[251] A.E. Teschendorff, S. Severini, Increased entropy of signal transduction in the cancer metastasis phenotype, BMC Syst. Biol. 4 (2010) 104. https://doi.org/10.1186/1752-0509-4-104.

[252] K. Kayser, G. Kayser, S. Eichhorn, U. Biechele, M. Altiner, H. Kaltner, F.Y. Zeng, E.V. Vlasova, N.V. Bovin, H.J. Gabius, Association of prognosis in surgically treated lung cancer patients with cytometric, histometric and ligand histochemical properties: with an emphasis on structural entropy, Anal. Quant. Cytol. Histol. 20 (1998) 313–320.

[253] U. Agrell, Draft of a general stochastic theory of cancer and its possible experimental verification with monoclonal multiplication of repairing and immunological systems, Med. Hypotheses. 20 (1986) 261–270. https://doi.org/10.1016/0306-9877(86)90042-3.

[254] J. West, G. Bianconi, S. Severini, A.E. Teschendorff, Differential network entropy reveals cancer system hallmarks, Sci. Rep. 2 (2012). https://doi.org/10.1038/srep00802.

[255] R. Berretta, P. Moscato, Cancer Biomarker Discovery: The Entropic Hallmark, PLoS ONE. 5 (2010). https://doi.org/10.1371/journal.pone.0012262.

[256] W.N. van Wieringen, A.W. van der Vaart, Statistical analysis of the cancer cell's molecular entropy using high-throughput data, Bioinforma. Oxf. Engl. 27 (2011) 556–563. https://doi.org/10.1093/bioinformatics/btq704.

[257] J.E. Dumont, S. Dremier, I. Pirson, C. Maenhaut, Cross signaling, cell specificity, and physiology, Am. J. Physiol. Cell Physiol. 283 (2002) C2-28. https://doi.org/10.1152/ajpcell.00581.2001.

[258] J.E. Riggs, Carcinogenesis, genetic instability and genomic entropy: insight derived from malignant brain tumor age specific mortality rate dynamics, J. Theor. Biol. 170 (1994) 331–338. https://doi.org/10.1006/jtbi.1994.1195.

[259] M. Tarabichi, A. Antoniou, M. Saiselet, J.M. Pita, G. Andry, J.E. Dumont, V. Detours, C. Maenhaut, Systems biology of cancer: entropy, disorder, and selection-driven evolution to independence, invasion and "swarm intelligence," Cancer Metastasis Rev. 32 (2013) 403–421. https://doi.org/10.1007/s10555-013-9431-y.

[260] K. Metze, R.L. Adam, G. Kayser, K. Kayser, Pathophysiology of Cancer and the Entropy Concept, in: L. Magnani, W. Carnielli, C. Pizzi (Eds.), Model-Based Reason. Sci. Technol., Springer Berlin Heidelberg, Berlin, Heidelberg, 2010: pp. 199–206. https://doi.org/10.1007/978-3-642-15223-8_10.

[261] K. K, K. G, M. K, The concept of structural entropy in tissue-based diagnosis., Anal. Quant. Cytol. Histol. 29 (2007) 296–308. https://doi.org/10.17629/www.diagnosticpathology.eu-2017-3:251.

[262] R.A. Gatenby, B.R. Frieden, Information dynamics in carcinogenesis and tumor growth, Mutat. Res. 568 (2004) 259–273. https://doi.org/10.1016/j.mrfmmm.2004.04.018.

[263] M.C. Daly, I.M. Paquette, Surveillance, Epidemiology, and End Results (SEER) and SEER-Medicare Databases: Use in Clinical Research for Improving Colorectal Cancer Outcomes, Clin. Colon Rectal Surg. 32 (2019) 61–68. https://doi.org/10.1055/s-0038-1673355.

[264] SEER Datasets, SEER Database. (2018). www.seer.cancer.gov/ (accessed March 17, 2020).

[265] S. Lukman, D.P. Lane, C.S. Verma, Mapping the Structural and Dynamical Features of Multiple p53 DNA Binding Domains: Insights into Loop 1 Intrinsic Dynamics, PLoS ONE. 8 (2013) e80221. https://doi.org/10.1371/journal.pone.0080221.

[266] J.N. Weiss, The Hill equation revisited: uses and misuses., FASEB J. 11 (1997) 835–841. https://doi.org/10.1096/fasebj.11.11.9285481.

[267] A.Ya. Kipnis, B.E. Yavelov, J.S. Rowlinson, Van Der Waals and Molecular Science, Oxford University Press, United Kingdom, 1996.

[268] R.A. Weinberg, The Biology of Cancer, 2nd Edition, 2nd edition, W.W. Norton & Company, New York, 2013.

[269] A.-S. Smith, Physics challenged by cells, Nat. Phys. 6 (2010) 726–729. https://doi.org/10.1038/nphys1798.

[270] S.G. Brush, History of the Lenz-Ising Model, Rev. Mod. Phys. 39 (1967) 883–893. https://doi.org/10.1103/RevModPhys.39.883.

`

[271] T. Ising, R. Folk, R. Kenna, B. Berche, Y. Holovatch, The Fate of Ernst Ising and the Fate of his Model, ArXive. 21 (2017).

[272] L. Onsager, Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition, Phys. Rev. 65 (1944) 117–149. https://doi.org/10.1103/PhysRev.65.117.

[273] H.A. Kramers, G.H. Wannier, Statistics of the Two-Dimensional Ferromagnet. Part II, Phys. Rev. 60 (1941) 263–276. https://doi.org/10.1103/PhysRev.60.263.

[274] W. Heisenberg, Zur Theorie des Ferromagnetismus, Z. Phys. 49 (1928) 619–636. https://doi.org/10.1007/BF01328601.

[275] L.D. Landau, On the theory of phase transitions, Zh.Eksp.Teor.Fiz. 7 (1937) 19–32. https://doi.org/10.1038/138840a0.

[276] C.N. Yang, The Spontaneous Magnetization of a Two-Dimensional Ising Model, Phys. Rev. 85 (1952) 808–816. https://doi.org/10.1103/PhysRev.85.808.

[277] E.M. Lifshitz, Statistical Physics : Theory of the Condensed State., Butterworth-Heinemann, San Diego, 2013.

[278] P. Sarkanych, Y. Holovatch, R. Kenna, Classical phase transitions in a one-dimensional short-range spin model, J. Phys. Math. Theor. 51 (2018) 505001. https://doi.org/10.1088/1751-8121/aaea02.

[279] D. Tahara, Y. Motome, M. Imada, Antiferromagnetic Ising Model on Inverse Perovskite Lattice, J. Phys. Soc. Jpn. 76 (2007) 013708. https://doi.org/10.1143/JPSJ.76.013708.

[280] D. I. Uzunov, Introduction to the theory of critical phenomena. Mean field, fluctuations and renormalization, 2nd ed, World Scientific Publishing Company, 2010.

[281] X. Sun, J. Bao, Y. Shao, Mathematical Modeling of Therapy-induced Cancer Drug Resistance: Connecting Cancer Mechanisms to Population Survival Rates, Sci. Rep. 6 (2016) 22498. https://doi.org/10.1038/srep22498.

[282] K.K. Rozman, J. Doull, W.J. Hayes Jr., Hayes' Handbook of Pesticide Toxicology, Chapter 1 - Dose and Time Determining, and Other Factors Influencing, Toxicity, Third, Academic Press, 2010.

[283] M. Stewart, I. Watson, Standard units for expressing drug concentrations in biological fluids., Br. J. Clin. Pharmac. (1983). https://doi.org/10.1111/j.1365-2125.1983.tb02136.x.

[284] M.E. Fisher, The susceptibility of the plane Ising model, Physica. 25 (1959) 521–524. https://doi.org/10.1016/S0031-8914(59)95411-4.

[285] Domb Cyril, Sykes M. F., Randall John Turton, On the susceptibility of a ferromagnetic above the Curie point, Proc. R. Soc. Lond. Ser. Math. Phys. Sci. 240 (1957) 214–228. https://doi.org/10.1098/rspa.1957.0078.

[286] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Res. 46 (2018) D1074–D1082. https://doi.org/10.1093/nar/gkx1037.

[287] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A.C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z.T. Dame, B. Han, Y. Zhou, D.S. Wishart, DrugBank 4.0: shedding new light on drug metabolism, Nucleic Acids Res. 42 (2014) D1091–D1097. https://doi.org/10.1093/nar/gkt1068.

[288] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A.C. Guo, D.S. Wishart, DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs, Nucleic Acids Res. 39 (2011) D1035–D1041. https://doi.org/10.1093/nar/gkq1126.

[289] D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, DrugBank: a knowledgebase for drugs, drug actions and drug targets, Nucleic Acids Res. 36 (2008) D901-906. https://doi.org/10.1093/nar/gkm958.

[290] D.S. Wishart, C. Knox, A.C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, DrugBank: a comprehensive resource for in silico drug discovery and exploration, Nucleic Acids Res. 34 (2006) D668-672. https://doi.org/10.1093/nar/gkj067.

[291] W. Senkowski, X. Zhang, M.H. Olofsson, R. Isacson, U. Höglund, M. Gustafsson, P. Nygren, S. Linder, R. Larsson, M. Fryknäs, Three-Dimensional Cell Culture-Based Screening Identifies the Anthelmintic Drug Nitazoxanide as a Candidate for Treatment of Colorectal Cancer, Mol. Cancer Ther. 14 (2015) 1504–1516. https://doi.org/10.1158/1535-7163.MCT-14-0792.

[292] C. Jarzynski, Nonequilibrium Equality for Free Energy Differences, Phys. Rev. Lett. 78 (1997) 2690–2693. https://doi.org/10.1103/PhysRevLett.78.2690.

[293] G. Ozer, S. Quirk, R. Hernandez, Thermodynamics of Decaalanine Stretching in Water Obtained by Adaptive Steered Molecular Dynamics Simulations, J. Chem. Theory Comput. 8 (2012) 4837–4844. https://doi.org/10.1021/ct300709u.

[294] P. Xiong, M. Wang, X. Zhou, T. Zhang, J. Zhang, Q. Chen, H. Liu, Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability, Nat. Commun. 5 (2014) 5330. https://doi.org/10.1038/ncomms6330.

[295] Y. Yan, D. Zhang, P. Zhou, B. Li, S.-Y. Huang, HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy, Nucleic Acids Res. 45 (2017) W365–W373. https://doi.org/10.1093/nar/gkx407.

[296] S.-Y. Huang, X. Zou, A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method, Nucleic Acids Res. 42 (2014) e55. https://doi.org/10.1093/nar/gku077.

`

[297] S.-Y. Huang, X. Zou, An iterative knowledge-based scoring function for protein-protein recognition, Proteins. 72 (2008) 557–579. https://doi.org/10.1002/prot.21949.

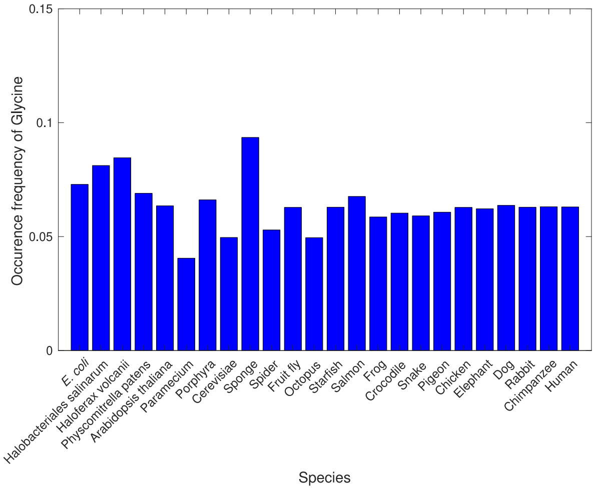
[298] Y. Yan, S. Huang, A New Pairwise Shape-Based Scoring Function to Consider Long-Range Interactions for Protein-Protein Docking, Biophys. J. 112 (2017) 470a. https://doi.org/10/ggmw5q.

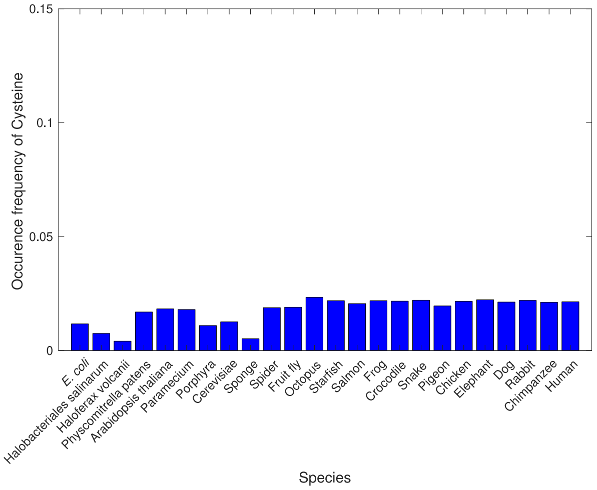
[299] S. Arbabi Moghadam, V. Rezania, S. Tuszynski, Cell Death and Survival Due to Cytotoxic Exposure Modeled as a Two-State Ising System, (2019) Dryad Digital Repository. https://doi.org/10.5061/dryad.4qrfj6q6d.

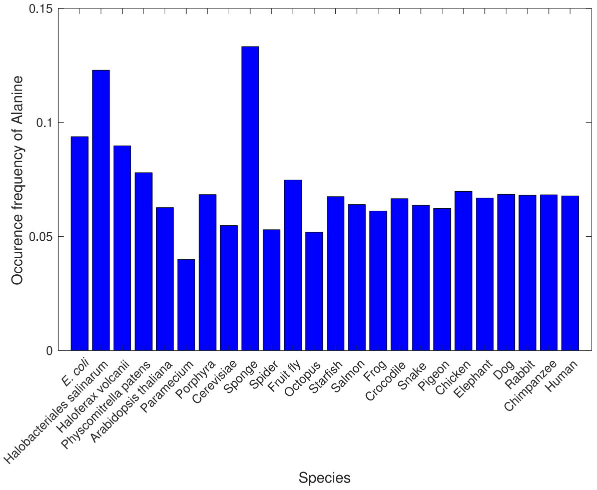
`

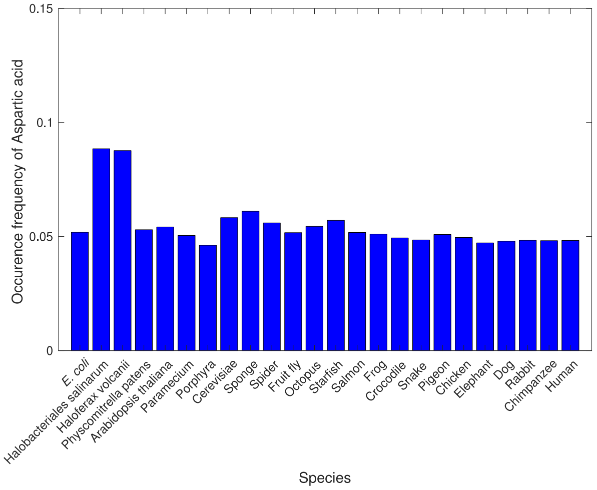
Appendix A [6]

A.1 Occurrence probability of amino acids across different species



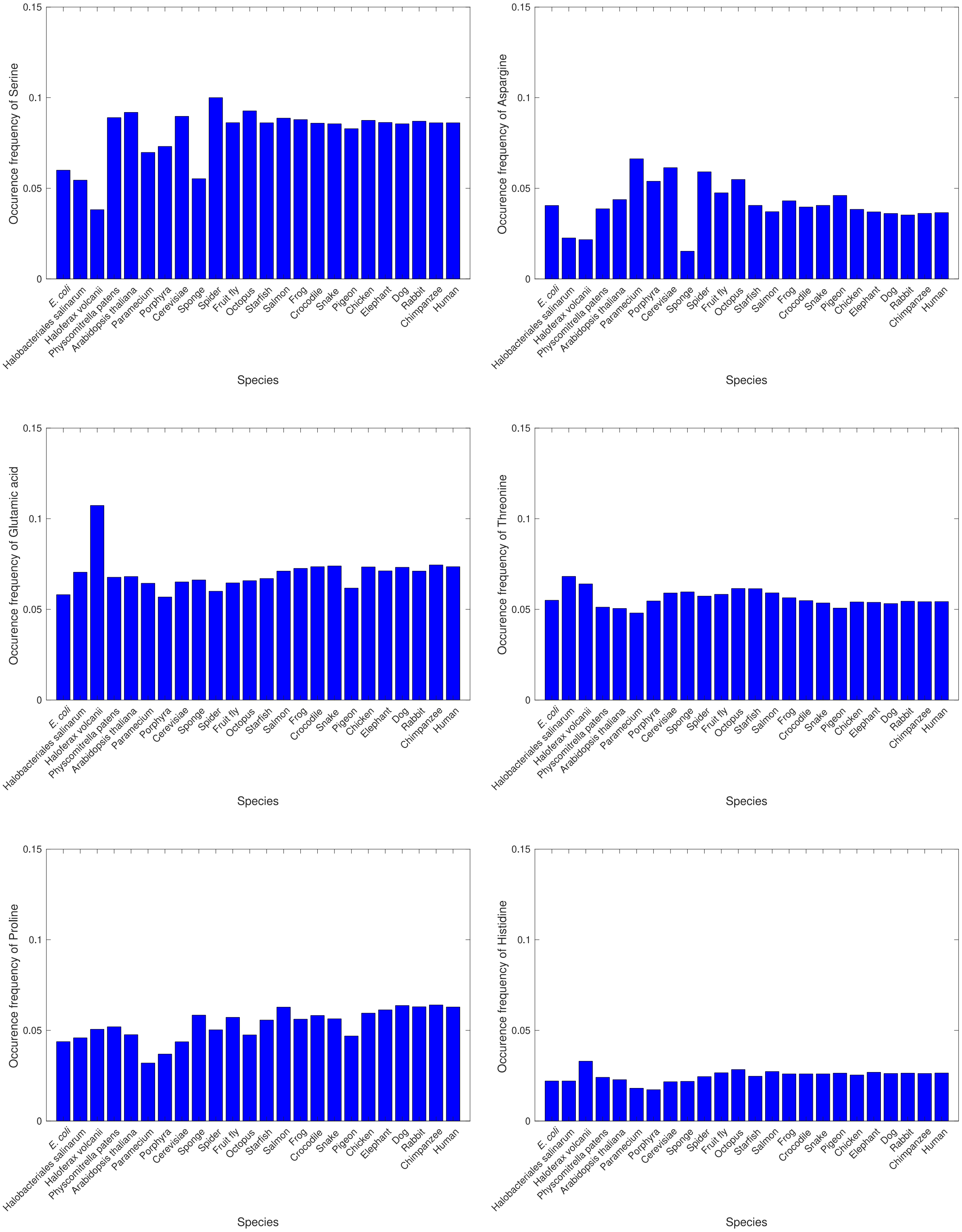




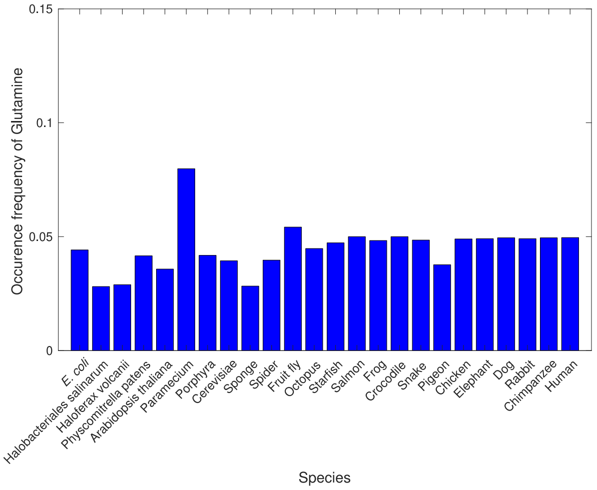


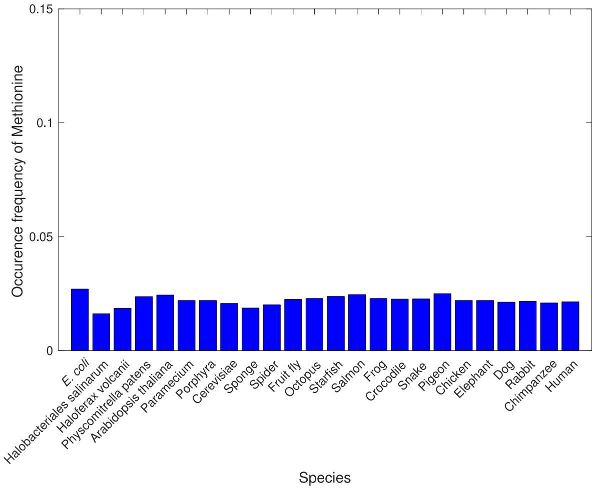
[6] This section has been accepted for publications partially as supplementary material of Arbabi Moghadam S., Klobukowski M., and Tuszynski, J.A., “A Search for the Physical Basis of the Genetic Code” in BioSystems journal.

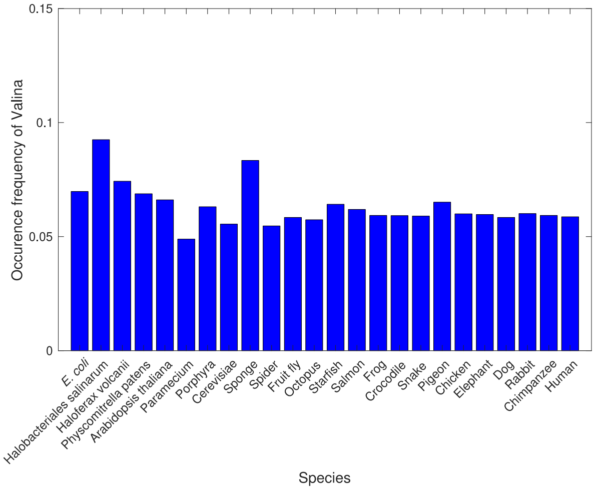
`

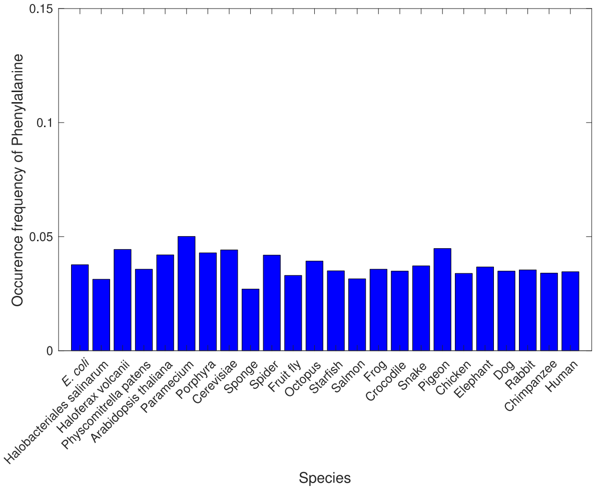

Occurence frequency of Serine
Occurence frequency of Aspargine
Occurence frequency of Glutamic acid
Occurence frequency of Threonine
Occurence frequency of Proline
Occurence frequency of Histidine
0
0.05
0.1
0.15
E. coli
Halobacteriales salinarum
Haloferax volcanii
Physcomitrella patens
Arabidopsis thaliana
Paramecium
Porphyra
Cerevisiae
Sponge
Spider
Fruit fly
Octopus
Starfish
Salmon
Frog
Crocodile
Snake
Pigeon
Chicken
Elephant
Dog
Rabbit
Chimpanzee
Human
Species

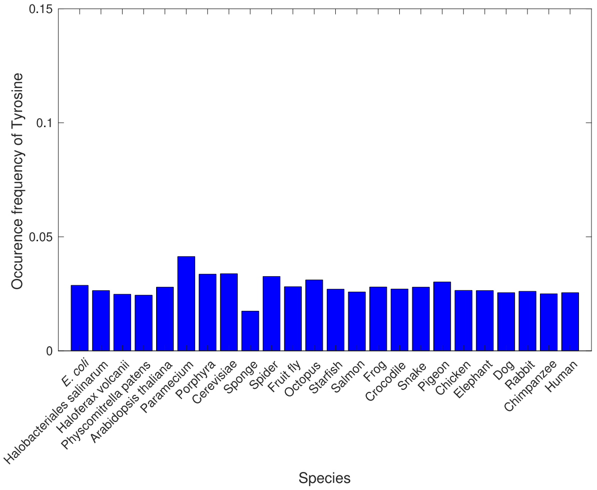
`

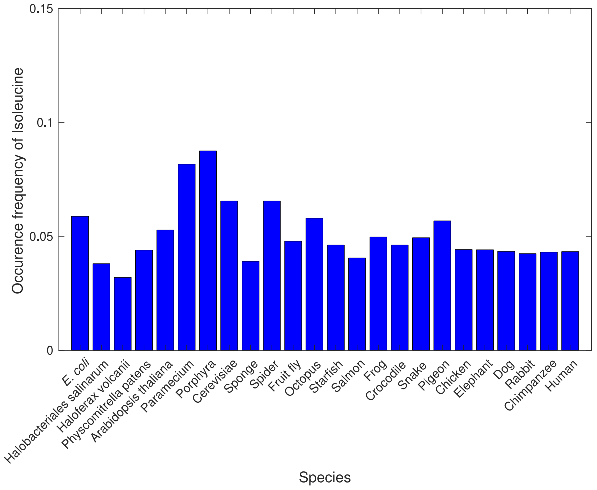

Occurence frequency of Glutamine
0.15
0.1
0.05
0
E. coli
Halobacteriales salinarum
Haloferax volcanii
Physcomitrella patens
Arabidopsis thaliana
Paramecium
Porphyra
Cerevisiae
Sponge
Spider
Fruit fly
Octopus
Starfish
Salmon
Frog
Crocodile
Snake
Pigeon
Chicken
Elephant
Dog
Rabbit
Chimpanzee
Human
Species


Occurence frequency of Methionine
0.15
0.1
0.05
0
E. coli
Halobacteriales salinarum
Haloferax volcanii
Physcomitrella patens
Arabidopsis thaliana
Paramecium
Porphyra
Cerevisiae
Sponge
Spider
Fruit fly
Octopus
Starfish
Salmon
Frog
Crocodile
Snake
Pigeon
Chicken
Elephant
Dog
Rabbit
Chimpanzee
Human
Species


Occurence frequency of Valina
0.15
0.1
0.05
0
E. coli
Halobacteriales salinarum
Haloferax volcanii
Physcomitrella patens
Arabidopsis thaliana
Paramecium
Porphyra
Cerevisiae
Sponge
Spider
Fruit fly
Octopus
Starfish
Salmon
Frog
Crocodile
Snake
Pigeon
Chicken
Elephant
Dog
Rabbit
Chimpanzee
Human
Species


Occurence frequency of Phenylalanine
0.15
0.1
0.05
0
E. coli
Halobacteriales salinarum
Haloferax volcanii
Physcomitrella patens
Arabidopsis thaliana
Paramecium
Porphyra
Cerevisiae
Sponge
Spider
Fruit fly
Octopus
Starfish
Salmon
Frog
Crocodile
Snake
Pigeon
Chicken
Elephant
Dog
Rabbit
Chimpanzee
Human
Species


Occurence frequency of Tyrosine
0.15
0.1
0.05
0
E. coli
Halobacteriales salinarum
Haloferax volcanii
Physcomitrella patens
Arabidopsis thaliana
Paramecium
Porphyra
Cerevisiae
Sponge
Spider
Fruit fly
Octopus
Starfish
Salmon
Frog
Crocodile
Snake
Pigeon
Chicken
Elephant
Dog
Rabbit
Chimpanzee
Human
Species


Occurence frequency of Isoleucine
0.15
0.1
0.05
0
E. coli
Halobacteriales salinarum
Haloferax volcanii
Physcomitrella patens
Arabidopsis thaliana
Paramecium
Porphyra
Cerevisiae
Sponge
Spider
Fruit fly
Octopus
Starfish
Salmon
Frog
Crocodile
Snake
Pigeon
Chicken
Elephant
Dog
Rabbit
Chimpanzee
Human
Species

`

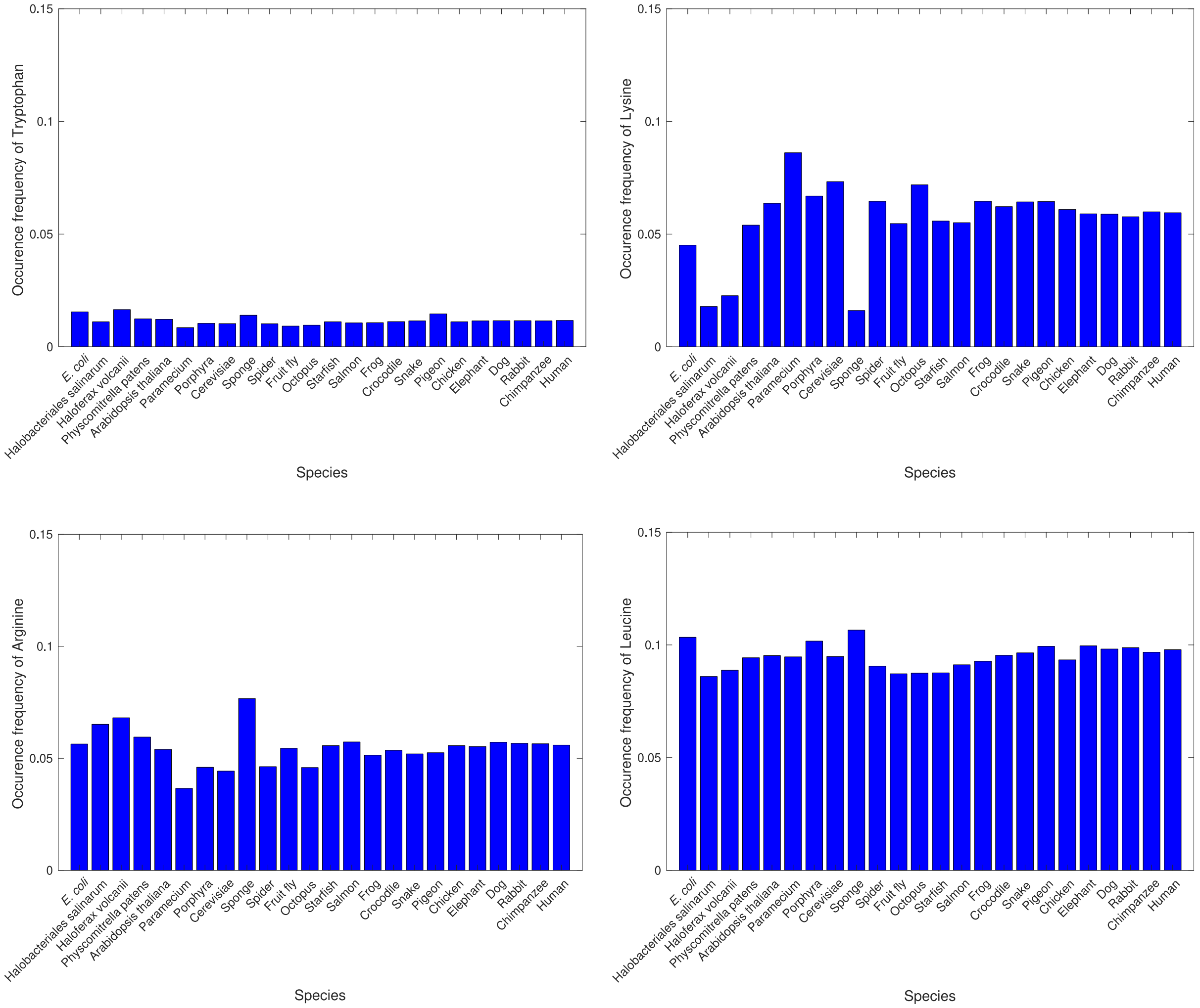
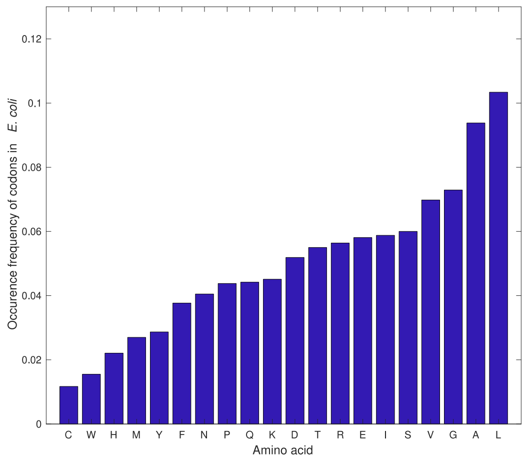
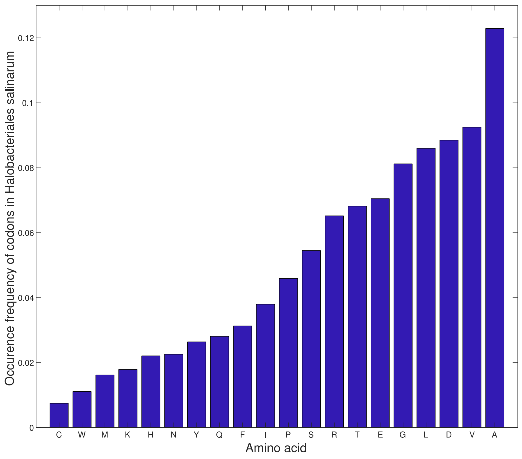


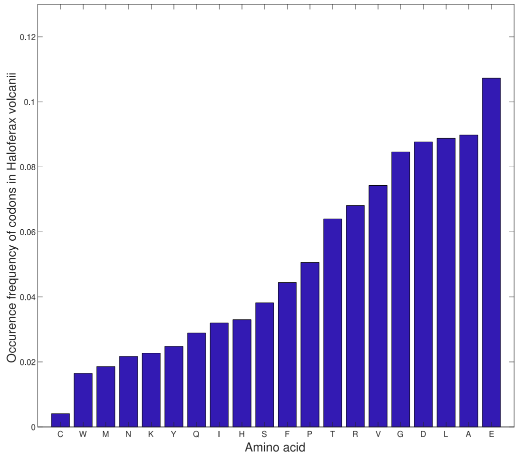
Figure A.1 Occurrence probability of amino acids across different species. The frequencies obtained from the NCBI databank, and are plotted for each amino acid across various species. The lowest standard deviation is for tryptophan ± 0.001 and the highest standard deviation is for alanine ± 0.02.

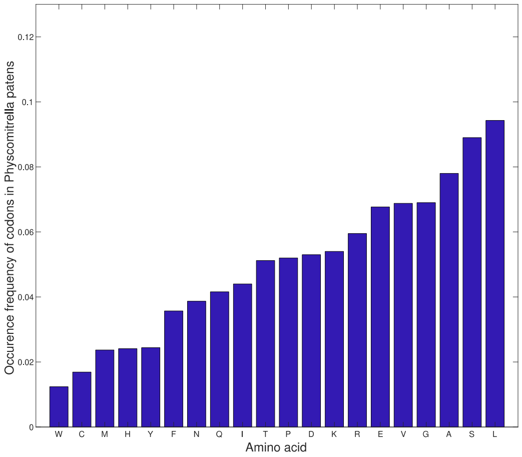
`

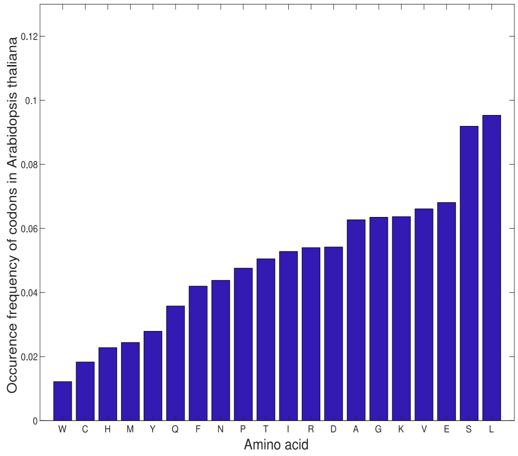
A.2 Occurrence probability of all 20 amino acids in each specie

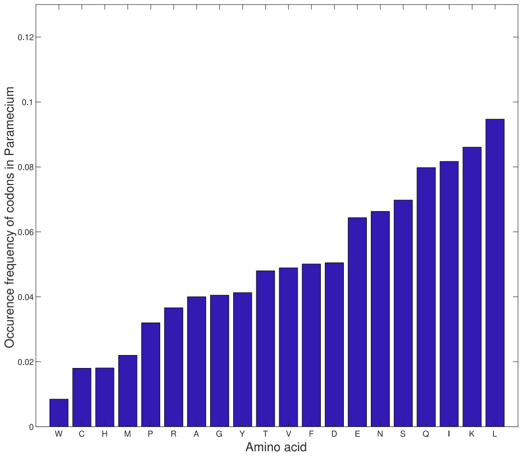




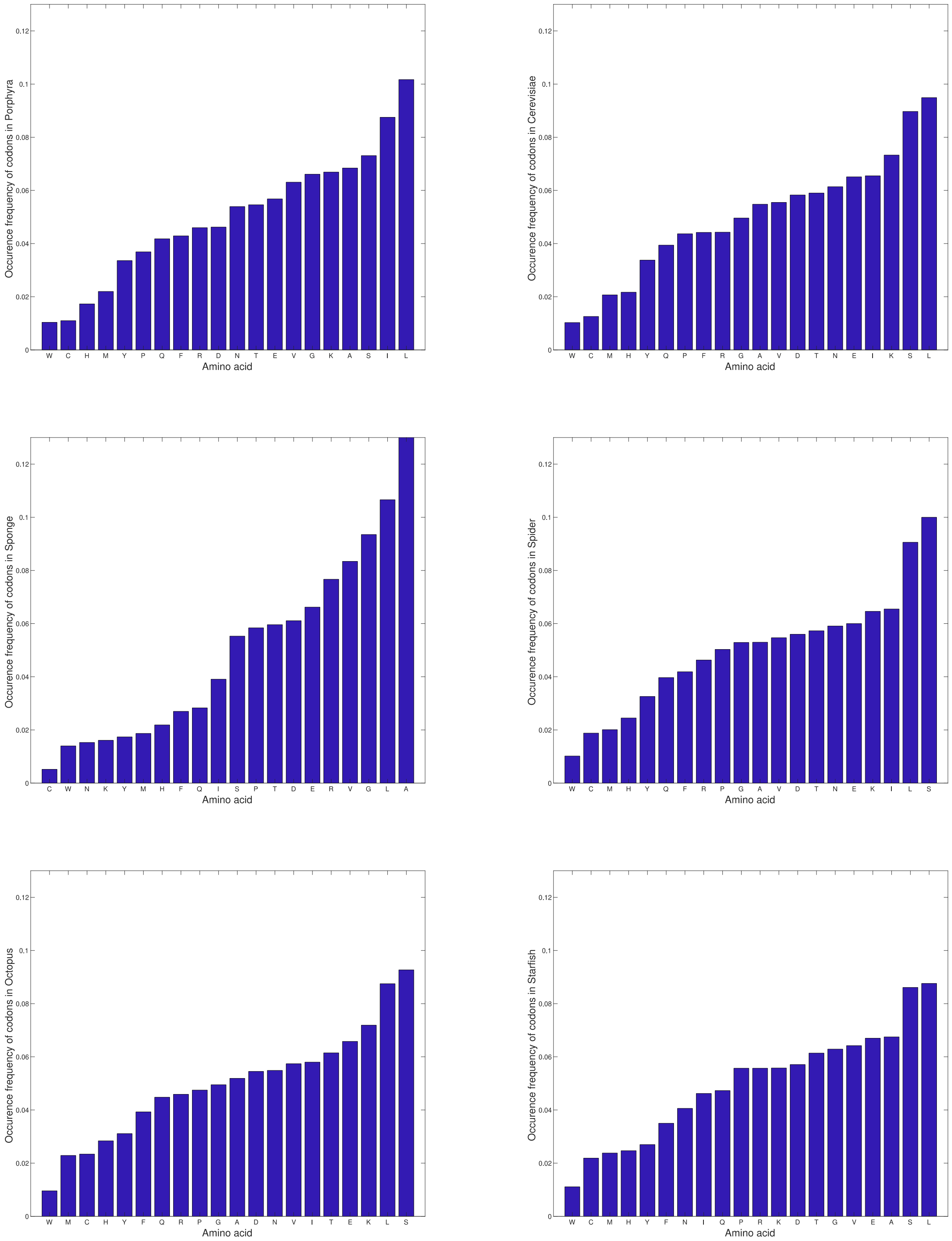


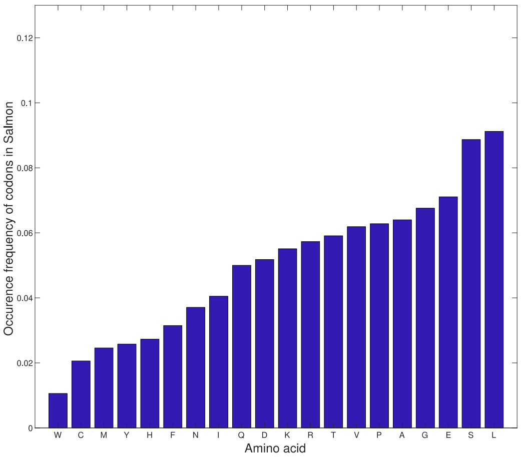


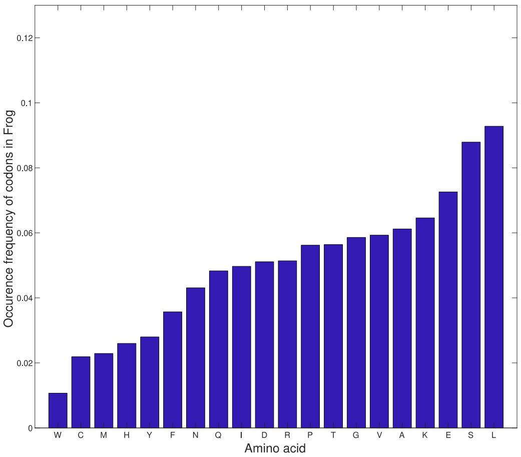


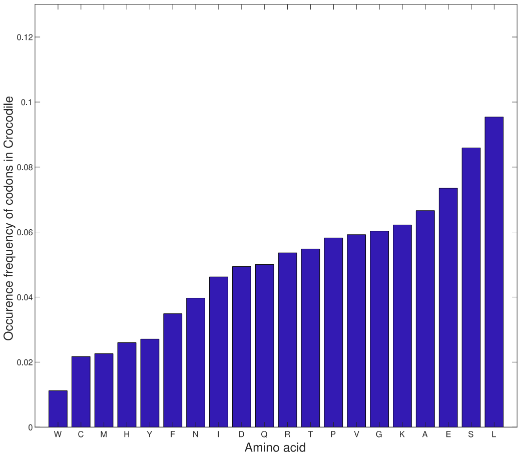


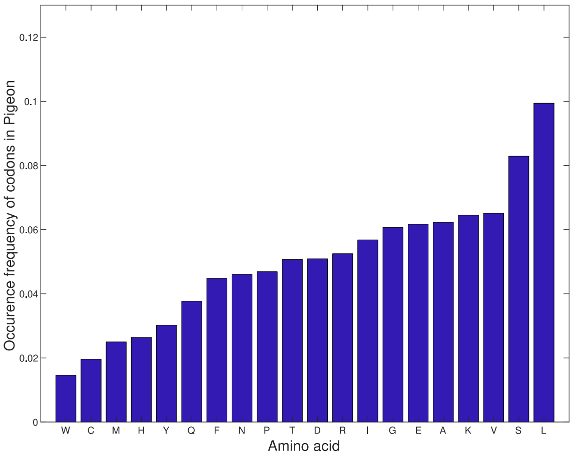
`

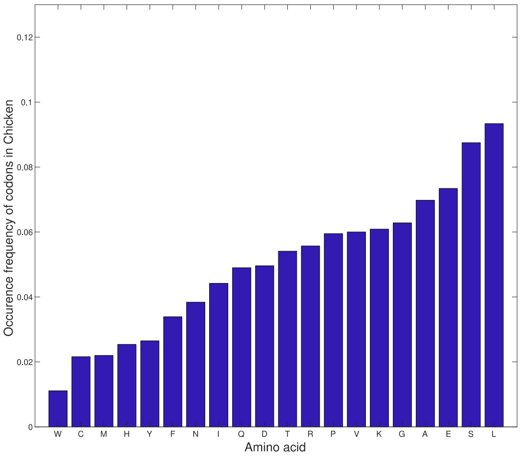

Occurence frequency of codons in Porphyra
0
0.02
0.04
0.06
0.08
0.1
0.12
W C H M Y P Q F R D N T E V G K A S I L
Amino acid
Occurence frequency of codons in Cerevisiae
0
0.02
0.04
0.06
0.08
0.1
0.12
W C M H Y Q P F R G A V D T N E I K S L
Amino acid
Occurence frequency of codons in Sponge
0
0.02
0.04
0.06
0.08
0.1
0.12
C W N K Y M H F Q I S P T D E R V G L A
Amino acid
Occurence frequency of codons in Spider
0
0.02
0.04
0.06
0.08
0.1
0.12
W C M H Y Q F R P G A V D T N E K I L S
Amino acid
Occurence frequency of codons in Octopus
0
0.02
0.04
0.06
0.08
0.1
0.12
W M C H Y F Q R P G A D N V I T E K L S
Amino acid
Occurence frequency of codons in Starfish
0
0.02
0.04
0.06
0.08
0.1
0.12
W C M H Y F N I Q P R K D T G V E A S L
Amino acid

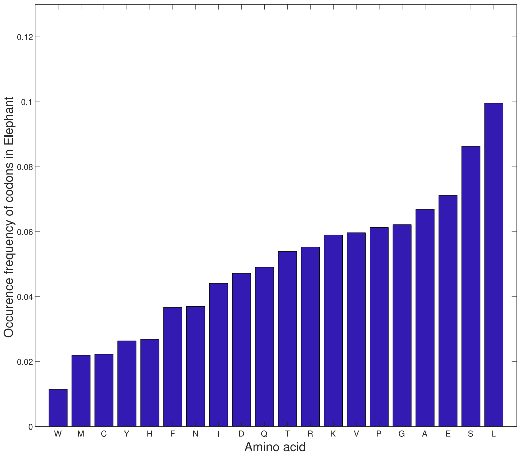

Occurence frequency of codons in Salmon
0.12
0.1
0.08
0.06
0.04
0.02
0
W C M Y H F N I Q D K R T V P A G E S L
Amino acid


Occurence frequency of codons in Frog
0.12
0.1
0.08
0.06
0.04
0.02
0
W C M H Y F N Q I D R P T G V A K E S L
Amino acid


Occurence frequency of codons in Crocodile
0.12
0.1
0.08
0.06
0.04
0.02
0
W C M H Y F N I D Q R T P V G K A E S L
Amino acid


Occurence frequency of codons in Pigeon
0.12
0.1
0.08
0.06
0.04
0.02
0
W C M H Y Q F N P T D R I G E A K V S L
Amino acid


Occurence frequency of codons in Chicken
0.12
0.1
0.08
0.06
0.04
0.02
0
W C M H Y F N I Q D T R P V K G A E S L
Amino acid


Occurence frequency of codons in Elephant
0.12
0.1
0.08
0.06
0.04
0.02
0
W M C Y H F N I D Q T R K V P G A E S L
Amino acid

`

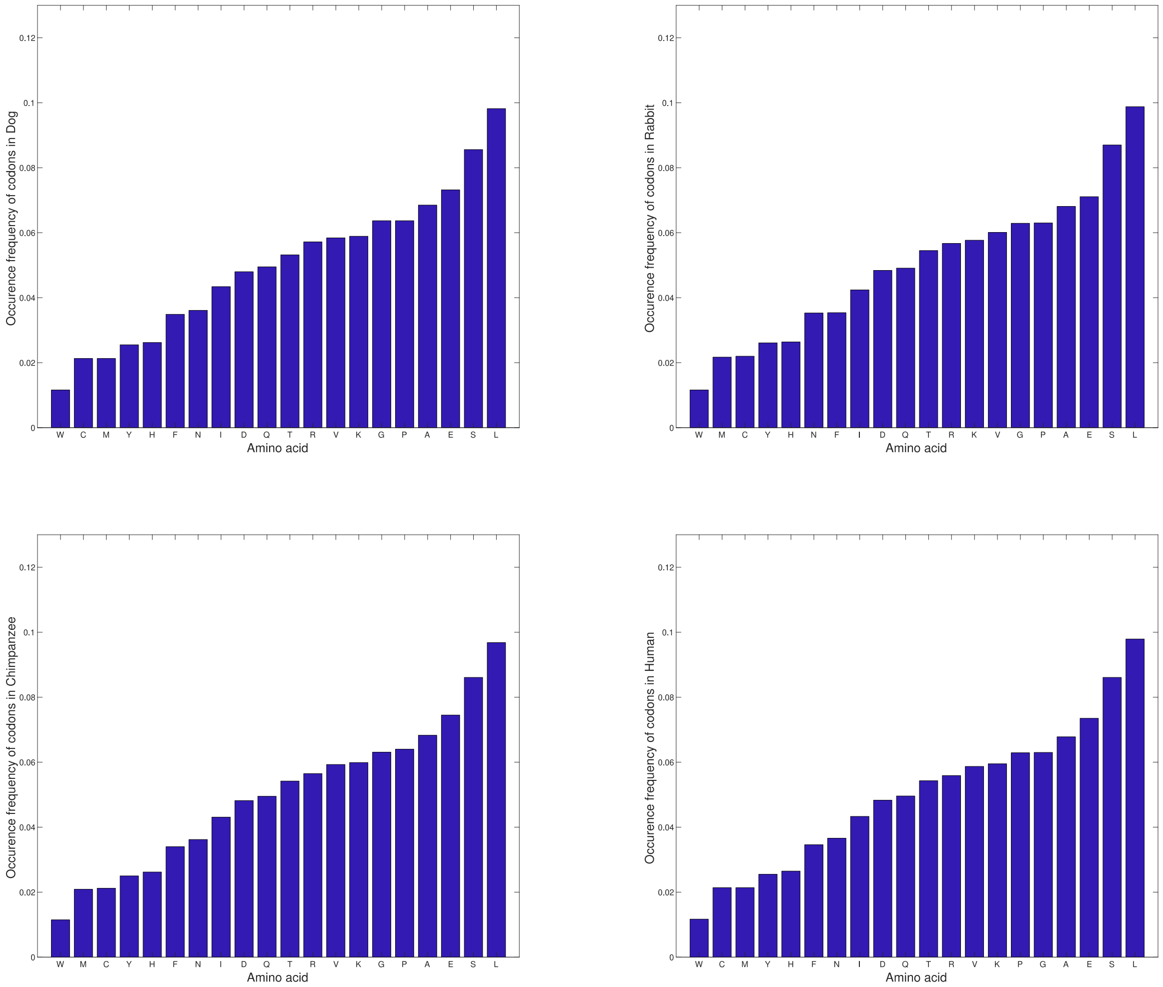


Figure A.2 Occurrence probability of all 20 amino acids in each species. The frequency is obtained from the NCBI database for each species. The lowest standard deviation is for octopus ± 0.02 and the highest standard deviations is for sponge ± 0.03.

`

A.3 The body tissues name and plots of probability

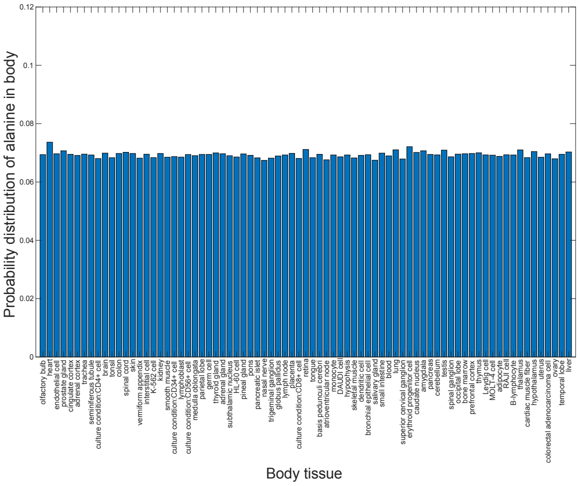
Table A.1 The complete list of studied body tissues.

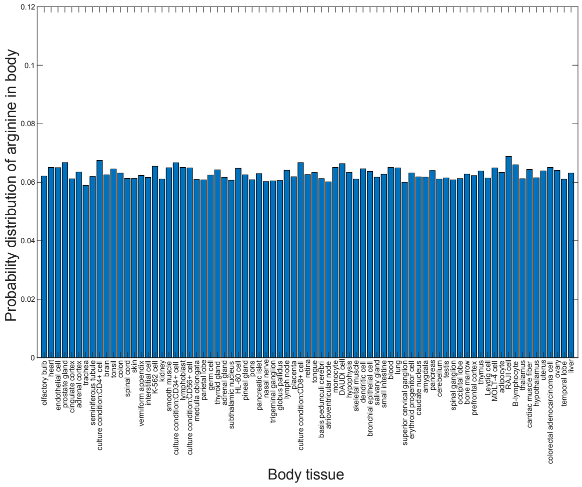
1-Olfactory bulb	40-Tongue
2-Heart	41-Basis pedunculi cerebri
3-Endothelial cell	42-Atrioventricular node
4-Prostate gland	43-Monocyte
5-Cingulate cortex	44-DAUDI cell
6-Adrenal cortex	45-Hypophysis
7-Trachea	46-Skeletal muscle
8-Seminiferous tubule	47-Dendritic cell
9-Culture condition:CD4+ cell	48-Bronchial epithelial cell
10-Brain	49-Salivary gland
11-Tonsil	50-Small intestine
12-Colon	51-Blood
13-Spinal cord	52-Lung
14-Skin	53-Superior cervical ganglion
15-Vermiform appendix	54-Erythroid progenitor cell
16-Interstitial cell	55-Caudate nucleus
17-K-562 cell	56-Amygdala
18-Kidney	57-Pancreas
19-Smooth muscle	58-Cerebellum
20-Culture condition:CD34+ cell	59-Testis
21-Lymphoblast	60-Spinal ganglion

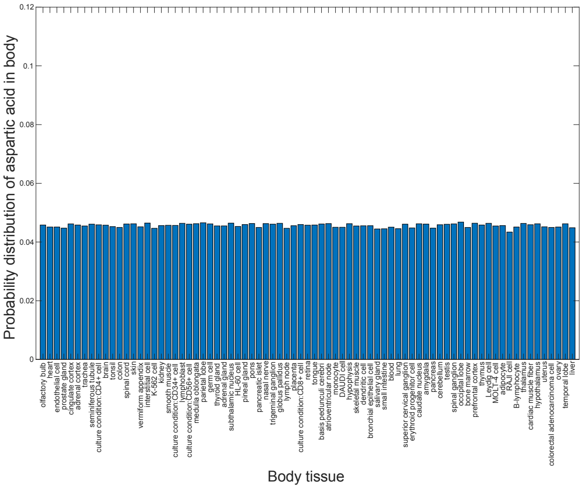
`

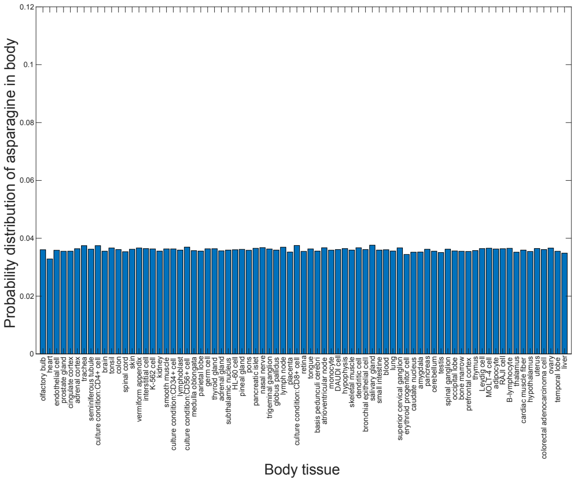
22-Culture condition:CD56+ cell	61-Occipital lobe
23-Medulla oblongata	62-Bone marrow
24-Parietal lobe	63-Prefrontal cortex
25-Germ cell	64-Thymus
26-Thyroid gland	65-Leydig cell
27-Adrenal gland	66-MOLT-4 cell
28-Subthalamic nucleus	67-Adipocyte
29-HL-60 cell	68-RAJI cell
30-Pineal gland	69-B-lymphocyte
31-Pons	70-Thalamus
32-Pancreatic islet	71-Cardiac muscle fiber
33-Nasal nerve	72-Hypothalamus
34-Trigeminal ganglion	73-Uterus
35-Globus pallidus	74-Colorectal adenocarcinoma cell
36-Lymph node	75-Ovary
37-Placenta	76-Temporal lobe
38-Culture condition:CD8+ cell	77-Liver
39-Retina	

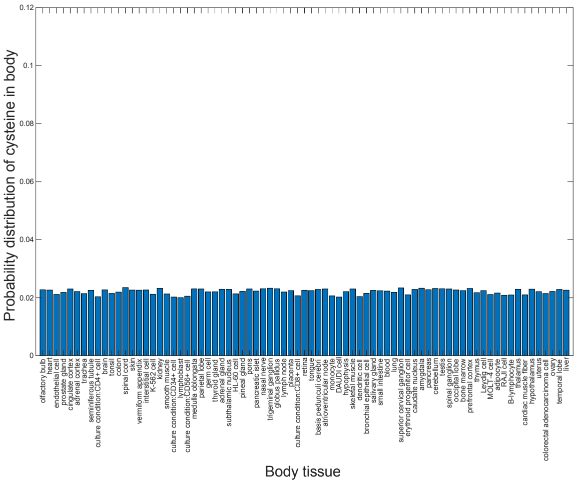
`

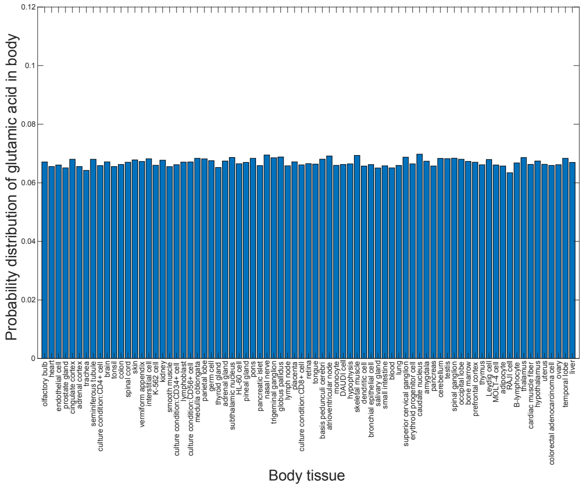

Probability distribution of alanine in body
0.12
0.1
0.08
0.06
0.04
0.02
0
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissue


Probability distribution of arginine in body
0.12
0.1
0.08
0.06
0.04
0.02
0
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissue

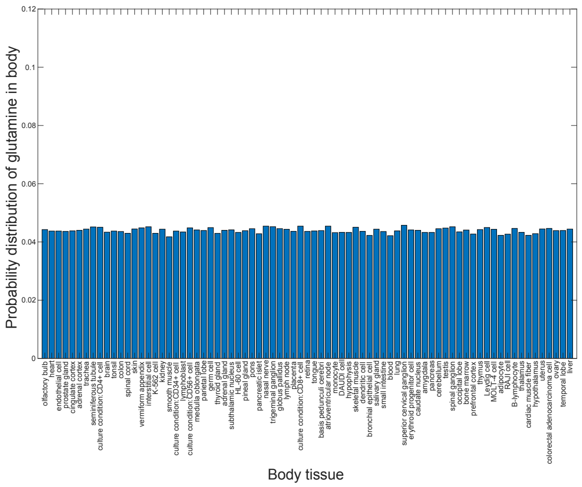

Probability distribution of aspartic acid in body
0.12
0.1
0.08
0.06
0.04
0.02
0
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissue

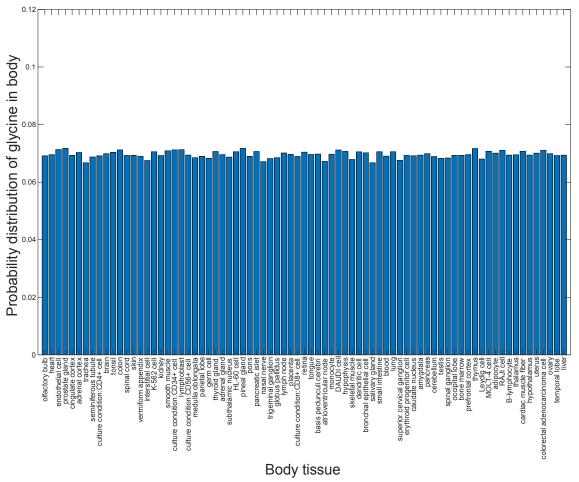

Probability distribution of asparagine in body
0.12
0.1
0.08
0.06
0.04
0.02
0
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissue

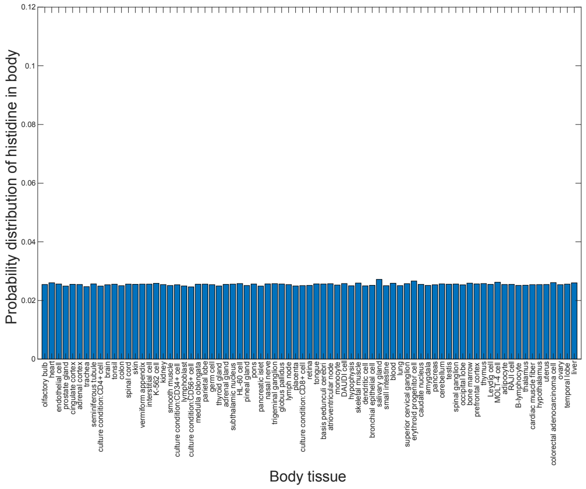

Probability distribution of cysteine in body
0.12
0.1
0.08
0.06
0.04
0.02
0
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissue

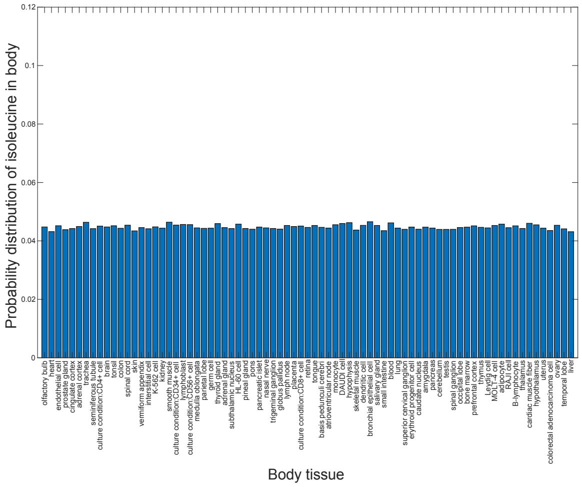

Probability distribution of glutamic acid in body
0.12
0.1
0.08
0.06
0.04
0.02
0
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissue

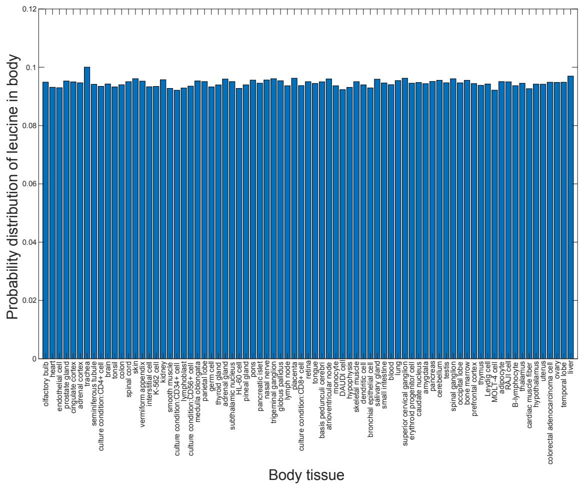
`

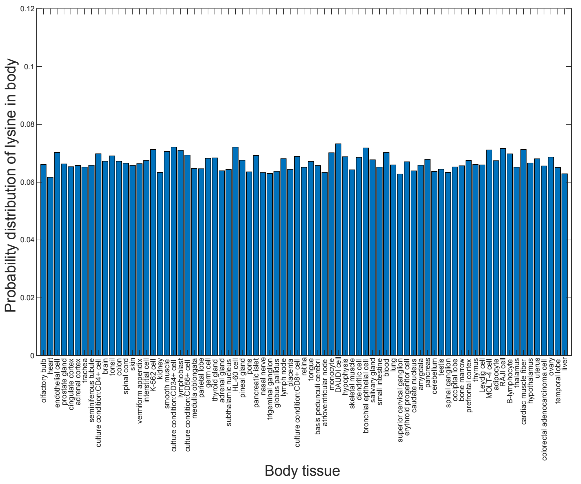

Probability distribution of glutamine in body
0
0.02
0.04
0.06
0.08
0.1
0.12
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissue


Probability distribution of glycine in body
0
0.02
0.04
0.06
0.08
0.1
0.12
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissue

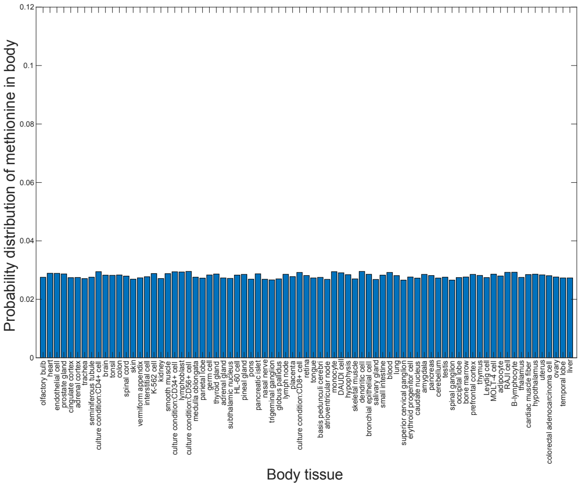

Probability distribution of histidine in body
0
0.02
0.04
0.06
0.08
0.1
0.12
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissue


Probability distribution of isoleucine in body
0
0.02
0.04
0.06
0.08
0.1
0.12
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissue

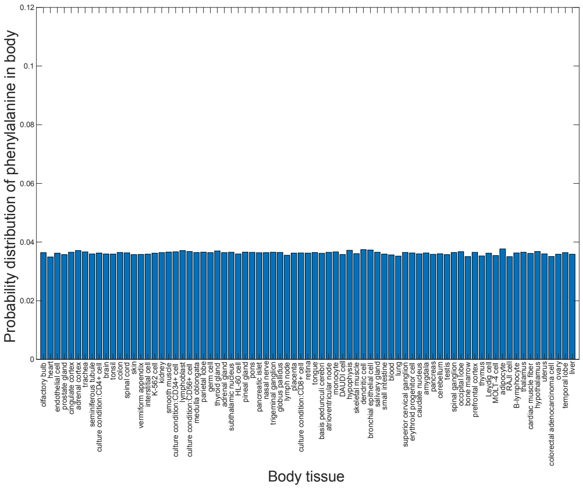

Probability distribution of leucine in body
0
0.02
0.04
0.06
0.08
0.1
0.12
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissue


Probability distribution of lysine in body
0
0.02
0.04
0.06
0.08
0.1
0.12
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissue

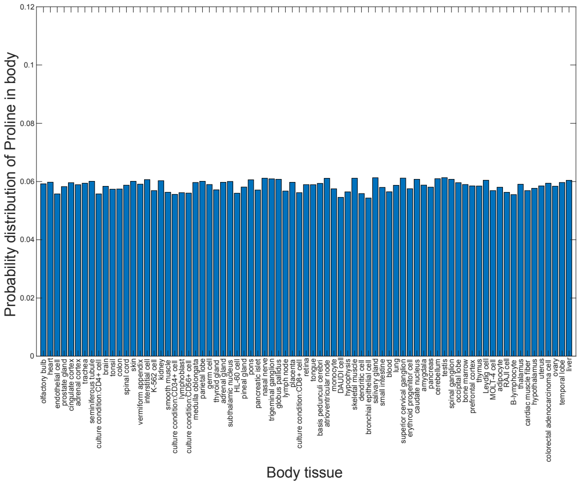
`



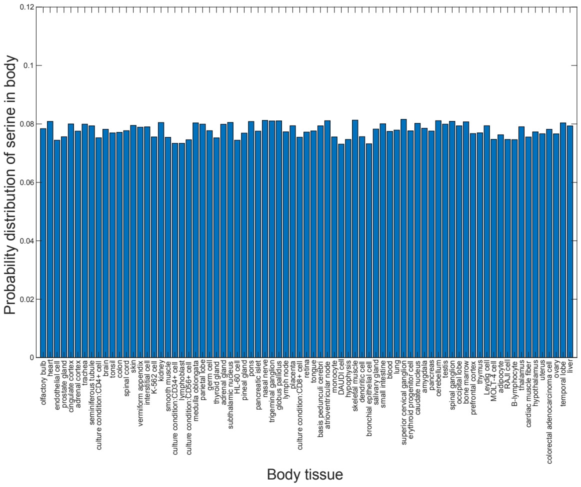
Probability distribution of methionine in body
0.12
0.1
0.08
0.06
0.04
0.02
0
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissue



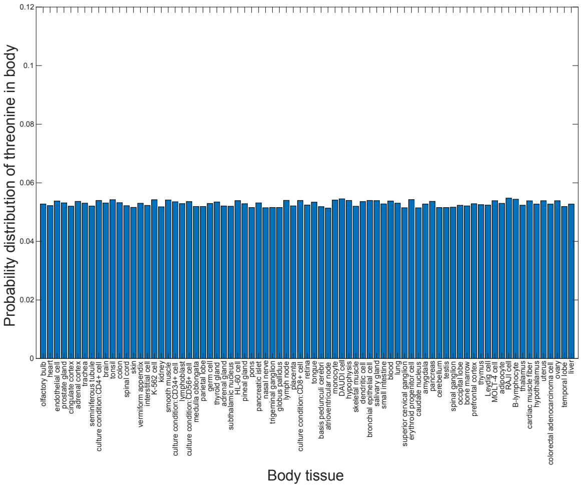
Probability distribution of phenylalanine in body
0.12
0.1
0.08
0.06
0.04
0.02
0
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissue



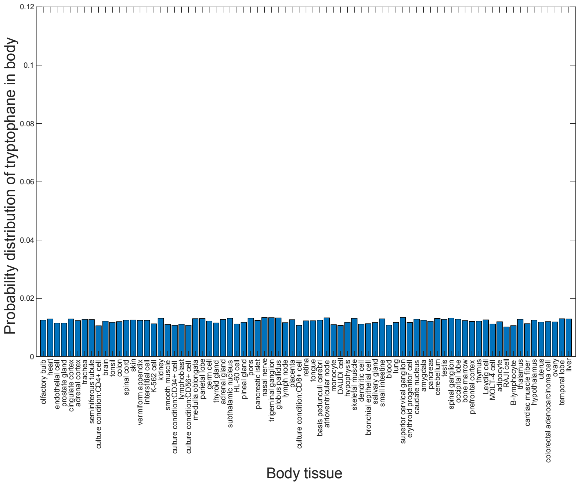
Probability distribution of Proline in body
0.12
0.1
0.08
0.06
0.04
0.02
0
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissue



Probability distribution of serine in body
0.12
0.1
0.08
0.06
0.04
0.02
0
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissue



Probability distribution of threonine in body
0.12
0.1
0.08
0.06
0.04
0.02
0
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissue



Probability distribution of tryptophane in body
0.12
0.1
0.08
0.06
0.04
0.02
0
olfactory bulb
heart
endothelial cell
prostate gland
cingulate cortex
adrenal cortex
trachea
seminiferous tubule
culture condition:CD4+ cell
brain
tonsil
colon
spinal cord
skin
vermiform appendix
interstitial cell
K-562 cell
kidney
smooth muscle
culture condition:CD34+ cell
lymphoblast
culture condition:CD56+ cell
medulla oblongata
parietal lobe
germ cell
thyroid gland
adrenal gland
subthalamic nucleus
HL-60 cell
pineal gland
pons
pancreatic islet
nasal nerve
trigeminal ganglion
globus pallidus
lymph node
placenta
culture condition:CD8+ cell
retina
tongue
basis pedunculi cerebri
atrioventricular node
monocyte
DAUDI cell
hypophysis
skeletal muscle
dendritic cell
bronchial epithelial cell
salivary gland
small intestine
blood
lung
superior cervical ganglion
erythroid progenitor cell
caudate nucleus
amygdala
pancreas
cerebellum
testis
spinal ganglion
occipital lobe
bone marrow
prefrontal cortex
thymus
Leydig cell
MOLT-4 cell
adipocyte
RAJI cell
B-lymphocyte
thalamus
cardiac muscle fiber
hypothalamus
uterus
colorectal adenocarcinoma cell
ovary
temporal lobe
liver
Body tissue

`

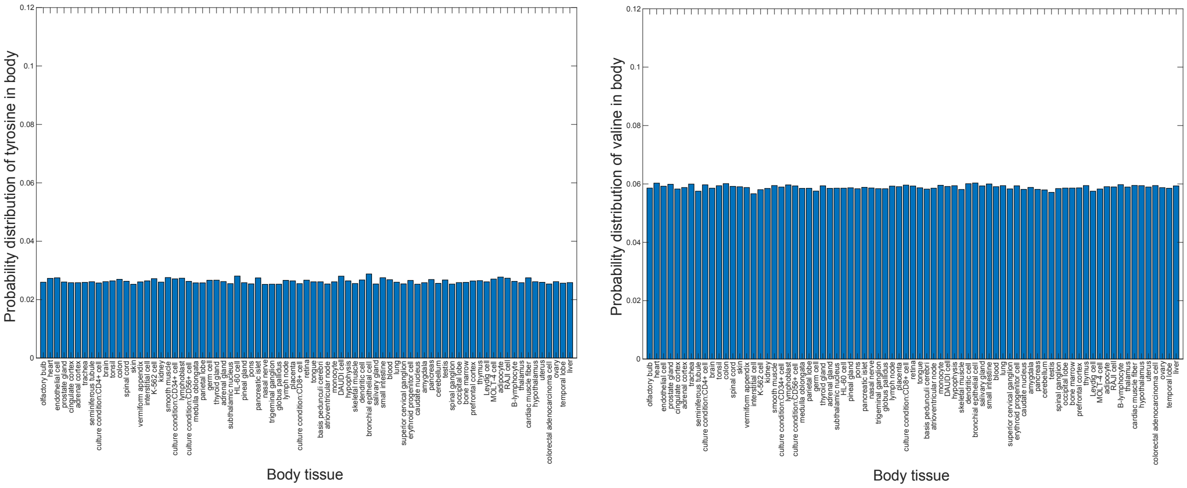


Figure A.3 Amino acid probability distribution in different human body tissues. The list of the human body tissues can be found in Table A.1. The standard deviation varies between ± 0.0004 and ± 0.002 which the lowest standard deviation. It corresponds to histidine and the highest standard deviation corresponds to lysine.

`

A.4 Frequency of animal and fungal mitochondrial proteins for each amino acids

Table A.2 Animal and fungal mitochondrial protein gene names [103] .

Fungal Gene symbol					Animal Gene symbol		
1-SAL1	105-nuo40	209-AGP2	313-CCM1	417-HNT1	1-TK2	105-SMCP	209-MSRA
2-RTC6	106-PRD1	210-MRPS9	314-BXI1	418-BUD22	2-PPP6C	106-LIG3	210-SPG7
3-YOR304C-A	107-USO1	211-ECM31	315-YPT11	419-TAP42	3-NME4	107-DAP3	211-REXO2
4-cia30	108-PGS1	212-RIM2	316-PUS4	420-COG8	4-GSTA4	108-IDH3G	212-STARD13
5-cia84	109-nuo-21	213-GPX2	317-GCV2	421-CSM3	5-KMO	109-ALDH5A1	213-MYO5A
6-peg1	110-ADK2	214-BNA4	318-TOM22	422-ARG7	6-MRPS12	110-PRKX	
7-maiA	111-RAD27	215-YBL096C	319-LYS4	423-RKR1	7-TP73	111-RIDA	
8-MDM35	112-COQ3	216-MAP2	320-PHB2	424-SOM1	8-PPM1G	112-HK2	
9-mug164	113-COQ8	217-UBP13	321-MIC26	425-KEI1	9-SLC25A20	113-MRPL12	
10-MRP10	114-ERV1	218-POA1	322-PSP2	426-MRX9	10-CFAP410	114-BLVRA	
11-bms1	115-YIM1	219-CDS1	323-UBC9	427-NDE2	11-KIF1B	115-DAPK1	
12-CYC1	116-MSP1	220-CST26	324-MIC27	428-HEM25	12-MRPS14	116-MAPK12	
13-CYC7	117-MRP17	221-TCM62	325-ALT1	429-YFH1	13-DNAJA2	117-HAP1	
14-QCR6	118-EHD3	222-FMP23	326-SCS3	430-CRD1	14-NIPSNAP2	118-HMGCS2	
15-QCR7	119-BMH1	223-PHO88	327-DOC1	431-YDL218W	15-SLC25A12	119-ALDH18A1	
16-CYB2	120-PSO2	224-AIM3	328-MTO1	432-FMP45	16-MPC2	120-NDUFV3	
17-TDH3	121-FAA1	225-IFA38	329-MTC3	433-PUF3	17-CDS2	121-MCCD1	
18-COX1	122-GPI10	226-POP7	330-MDM34	434-ISA1	18-KRT75	122-RAB2A	
19-CCP1	123-ATP7	227-FZO1	331-NCS6	435-FRA1	19-B3GALT4	123-RPS18	
20-SOD2	124-MRPL9	228-MBA1	332-MRM2	436-nuo-10.5	20-LDHA	124-PPIA	
21-CIT1	125-MSK1	229-MCX1	333-RMD9	437-PAM18	21-CYB5R3	125-UBE2I	
22-GPM1	126-GUT1	230-OM14	334-MPC1	438-LAM6	22-GSR	126-PPP2CA	
23-VAR1	127-GUT2	231-YBR238C	335-YGL069C	439-BUD20	23-SOD1	127-SLC25A3	
24-TUB2	128-MGM1	232-MIC12	336-MRH4	440-FMP25	24-ABL1	128-CDK17	
25-TUF1	129-GLN1	233-SDH8	337-PUS2	441-COQ10	25-EGFR	129-NFKB2	
26-TEF1	130-YMC1	234-YSY6	338-PKP2	442-RCL1	26-TGFB1	130-PRKCE	
27-COX4	131-OAC1	235-MGE1	339-GEP7	443-MIM1	27-APOA1	131-TOP2B	
28-CDC9	132-NDI1	236-YPT31	340-JAC1	444-MDM38	28-APOH	132-CAV1	
29-PHR1	133-KNS1	237-MSG5	341-YGR012W	445-NGL1	29-MT2A	133-P	

`

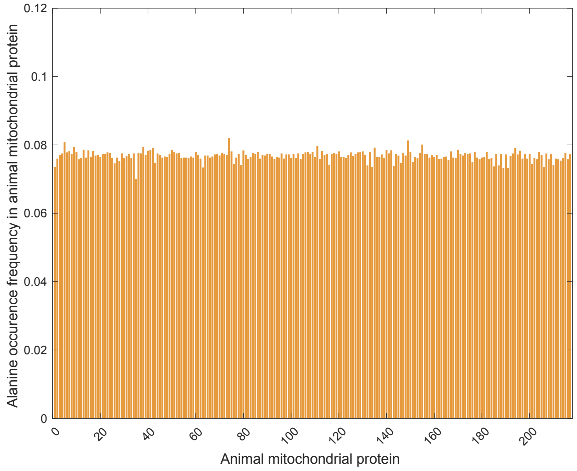
30-ATP4	134-ARP2	238-CBR1	342-EAT1	446-YOL046C	30-ESR1	134-EEF1A2	
31-ILV5	135-MRP49	239-RPS20	343-YGR021W	447-SDH5	31-MMP1	135-PTPN11	
32-CYC3	136-BIO2	240-BAT1	344-IMO32	448-AVO1	32-CAT	136-BCL2L1	
33-CYT1	137-APE2	241-HSP10	345-TIM21	449-GAS4	33-RAF1	137-RBL2	
34-TOM70	138-ACP1	242-PNT1	346-TAM41	450-ALE1	34-ANXA1	138-TP53BP2	
35-ADR1	139-DPH5	243-GGC1	347-FMP48	451-THI72	35-CAPNS1	139-AUH	
36-QCR2	140-ISF1	244-NAB3	348-PIL1	452-GEP3	36-TYMS	140-MRPL58	
37-PEP4	141-ZUO1	245-PSD1	349-TPC1	453-MET7	37-GNAI2	141-ENDOG	
38-PIF1	142-MSS1	246-XDJ1	350-PCP1	454-DGA1	38-APP	142-FASTK	
39-RAD2	143-VPH1	247-TIM17	351-GTF1	455-TUM1	39-ALDH2	143-GCKR	
40-ILV2	144-SCM4	248-FAA2	352-SHY1	456-HRK1	40-PCCA	144-GK3P	
41-CHO1	145-SUA5	249-ILV3	353-YGR164W	457-RDL2	41-PCCB	145-GK2	
42-NUC1	146-CYS4	250-CEM1	354-PUS6	458-TIM18	42-FABP3	146-STARD3	
43-ERG20	147-TOR2	251-ACO2	355-PBP1	459-MSC6	43-GSN	147-OXA1L	
44-MSS18	148-PET127	252-UBP12	356-TIM13	460-CIR2	44-GPX1	148-PEA15	
45-FLO11	149-ura3	253-YHB1	357-RSM27	461-GUP2	45-CAPN1	149-PLEC	
46-PRB1	150-SFA1	254-AFG3	358-MPC3	462-OXR1	46-HSP90AA1	150-RPS6KA2	
47-ATP5	151-FMT1	255-OXA1	359-LSC2	463-FMP40	47-ANXA6	151-TAZ	
48-YML002W	152-MGM101	256-HEM14	360-MTM1	464-rga8	48-ANXA5	152-CLPP	
49-ADH4	153-CAT2	257-RRG9	361-YGR266W	465-AIM41	49-MRPL3	153-ME3	
50-COX7	154-COX13	258-JLP2	362-TOM7	466-PPT2	50-QDPR	154-PCK2	
51-PET122	155-AIM26	259-NTA1	363-LSC1	467-VIK1	51-HMGB1	155-PTPN21	
52-MAS1	156-NFU1	260-ARG2	364-SUN4	468-GRE2	52-HMOX1	156-AGK	
53-MRS3	157-HFA1	261-YJL067W	365-SMM1	469-PUS9	53-AFG3L2	157-SLC25A53	
54-SSA1	158-LIP5	262-ATM1	366-RCF2	470-YDL157C	54-IFI6	158-EARS2	
55-PET54	159-DLD1	263-YIL161W	367-ATP23	471-DIN7	55-ARAF	159-SLC25A30	
56-MTF2	160-TIM23	264-MET18	368-MRPL50	472-REX3	56-THRA	160-THEM4	
57-SEC7	161-SKG6	265-PRM5	369-MRPS12	473-SLM3	57-HSPA5	161-SLC25A25	
58-MAS2	162-YPT7	266-PRK1	370-RSM19	474-MCP1	58-HSPA8	162-MTHFD1L	
59-nuo-12	163-SFC1	267-SYG1	371-YNR040W	475-TMS1	59-DBT	163-QSOX2	
60-MRP13	164-DUT1	268-PDR11	372-YNL285W	476-ATG9	60-BCKDHA	164-USP30	
61-MRP7	165-SDH3	269-MRPL49	373-POP3	477-MCD1	61-UNG	165-ATG9A	
62-GLN4	166-TOM6	270-HXT9	374-BOR1	478-ATP16	62-ALAS1	166-BPHL	
63-HAP3	167-MRPS5	271-nuo-24	375-GOR1	479-LEU9	63-TPT1	167-MCAT	
64-gatA	168-PTH2	272-CIN1	376-FOL1	480-RRG1	64-PRKAR2A	168-RHOT2	

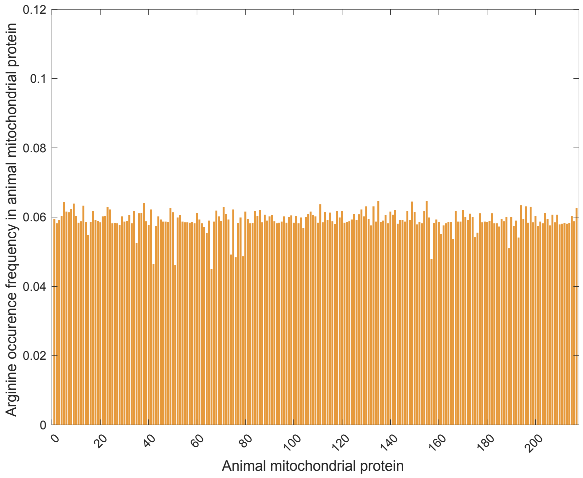
`

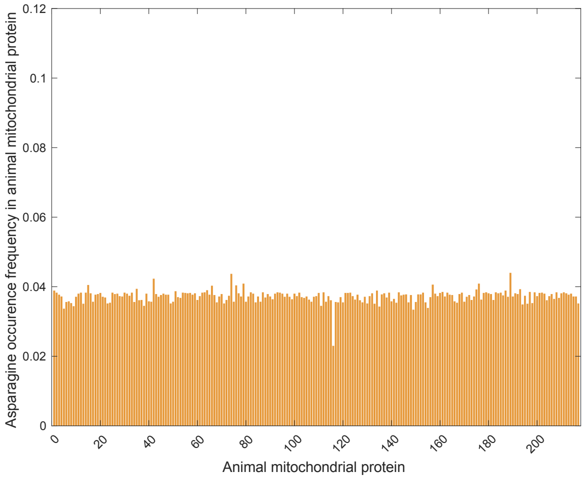
65-MRPL31	169-YBL059W	273-MSS2	377-YNL247W	481-MDM32	65-MTHFD2	169-RHOT1	
66-CBS1	170-PRX1	274-SNM1	378-MRX7	482-YOR022C	66-AKR1B1	170-SLC35F6	
67-PMS1	171-SEF1	275-IAH1	379-MRPL19	483-USB1	67-CYP11B1	171-SLC25A29	
68-CPR1	172-TOR1	276-POP1	380-MRPL22	484-PUF2	68-ST6GAL1	172-PDPR	
69-ARO3	173-TOM20	277-GPD2	381-APC1	485-YOR225W	69-COX7C	173-TRPV1	
70-CBS2	174-MDJ1	278-nuo14.8	382-ESBP6	486-MCT1	70-CREB1	174-GDAP1	
71-MTF1	175-MAC1	279-MDJ2	383-YNL122C	487-CRC1	71-HMGA1	175-PNPT1	
72-CTA1	176-CTF13	280-ZIM17	384-APJ1	488-MAM3	72-PRKCA	176-NEU4	
73-PDX1	177-TMA19	281-SLS1	385-MTQ1	489-YDR061W	73-GJA1	177-ST8SIA1	
74-PET123	178-MRP8	282-CIR1	386-SAM50	490-YPS3	74-LGALS3	178-MRPS31	
75-PDI1	179-OAR1	283-UTP10	387-ERG13	491-GLO4	75-CYP11B2	179-USP7	
76-MDM10	180-MRPL38	284-PAM16	388-TGL2	492-PST2	76-RXRA	180-WNT2B	
77-ARG8	181-RPS27A	285-QRI7	389-REX2	493-ORT1	77-BTF3	181-TBRG4	
78-RDS2	182-CBT1	286-RPN11	390-LSM3	494-GEP5	78-ALAS2	182-NEK1	
79-YAP1	183-MIA40	287-CIT3	391-ACT1	495-YOP1	79-COX7B	183-TEFM	
80-YMR31	184-YKL162C	288-ALD4	392-YAT1	496-MRX4	80-ATP5PB	184-YME1L1	
81-MRPL44	185-MCR1	289-GUS1	393-TIM10	497-TY2B-DR1	81-MPST	185-DNAJC2	
82-HOP1	186-CMC1	290-DLD2	394-alp16	498-MRPL23	82-MT3	186-MTERF1	
83-ATP15	187-HOT13	291-FMP33	395-SPBC31F10.	499-FMP16	83-tud	187-BAG1	
84-OSM1	188-YKL030W	292-MRX5	396-NUM1	500-PTC5	84-AK4	188-NIPSNAP1	
85-MRPS28	189-TCD2	293-AIM23	397-CYT2	501-nuo21.3c	85-MAPK3	189-SLC25A21	
86-SDH2	190-MIC60	294-MDV1	398-ARG5,6	502-HAA1	86-NDUFS1	190-SLC25A23	
87-MCK1	191-CAF4	295-IML2	399-nuc-2	503-RRP12	87-ATP5F1D	191-NLN	
88-RAM1	192-UTH1	296-BNA3	400-ACS1	504-TY3B-I	88-ced-4	192-AGXT2	
89-GAS1	193-FMP46	297-TIM54	401-SED1	505-B12J7.040	89-porB	193-BCO2	
90-SLY1	194-PAM17	298-AIM22	402-GCR2	506-apg-2	90-PRDX2	194-CLPB	
91-MRPL8	195-YKR070W	299-YJL045W	403-AAT1	507-dim-5	91-KIF5B	195-SLC25A31	
92-MRPL20	196-DRE2	300-YJL043W	404-MDM1	508-erp38	92-USP6	196-QRSL1	
93-MSM1	197-TRZ1	301-COX16	405-TIM44	509-cbs-1	93-NF2	197-MRPL18	
94-MRPL25	198-OMA1	302-MRX12	406-ENA2	510-DFG16	94-IDC34.5	198-GOLPH3	
95-MRS4	199-MRPL11	303-BNA2	407-YSA1	511-ODC2	95-GPX4	199-OSGEPL1	
96-MIR1	200-MRPL27	304-YJR085C	408-MRPL13	512-TY3B-G	96-ETFB	200-NOX4	
97-TOM40	201-MRPL17	305-AIM25	409-RNT1	513-PET20	97-COL18A1	201-MRPL40	
98-SCO1	202-MRPL36	306-RSM26	410-YML133C	514-ND6	98-FEN1	202-XPNPEP1	

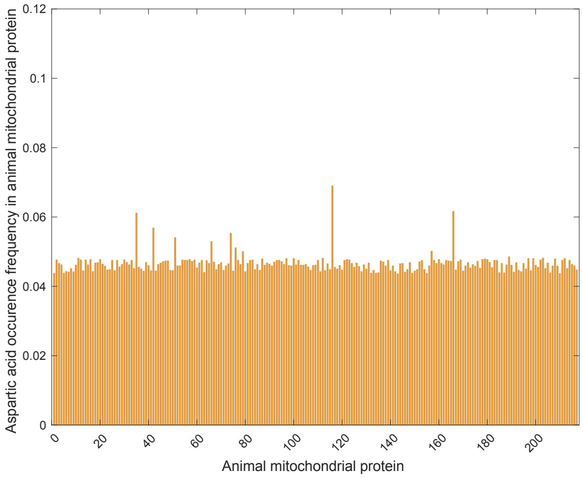
`

99-PKC1	203-MRPL37	307-YJR111C	411-YMR102C	515-SPBC1703.0	99-CSK	203-ADPRHL2	
100-nuo78	204-MRPL40	308-RSM7	412-EAR1	516-Q0010	100-ECI1	204-RAB20	
101-nuo-32	205-MRPL16	309-JHD2	413-ABZ2		101-CRAT	205-SLC25A10	
102-IFM1	206-SCO2	310-IBA57	414-YMD8		102-ATP5PO	206-APEX2	
103-MEF1	207-YMC2	311-TTI2	415-YIR020C-B		103-IDH2	207-AASS	
104-ERG6	208-GRS1	312-AAD10	416-MIX14		104-TSC2	208-WARS2	

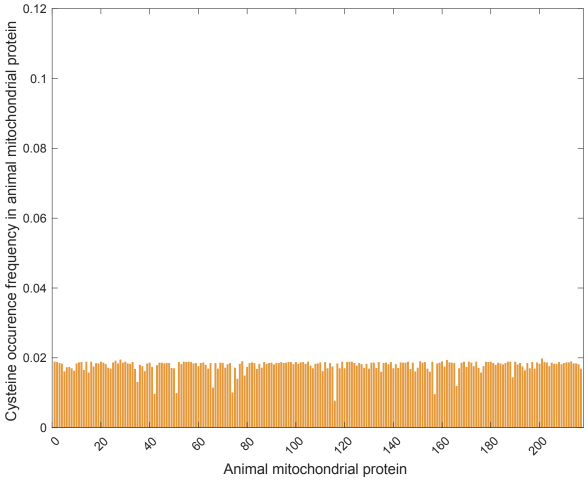


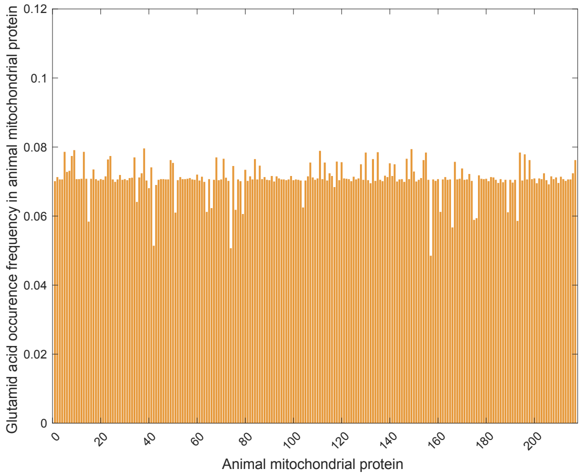


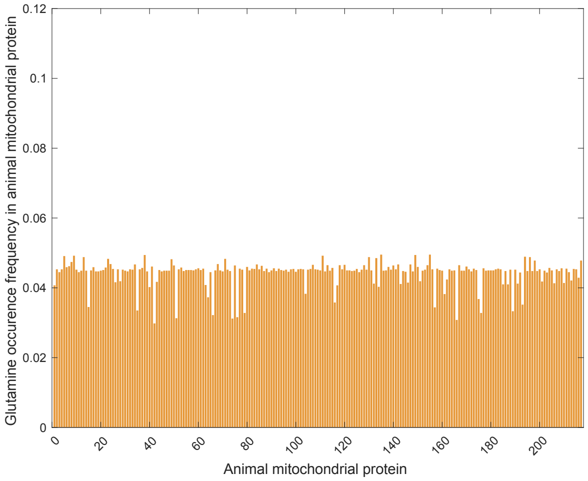


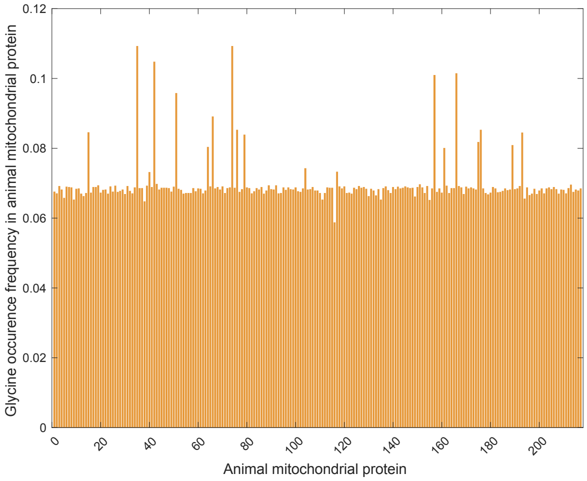


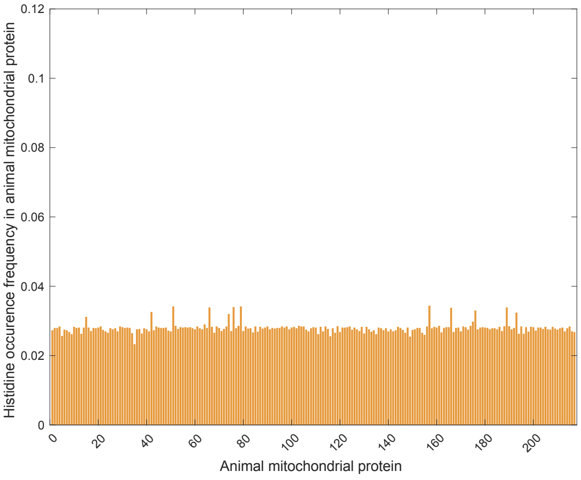
`

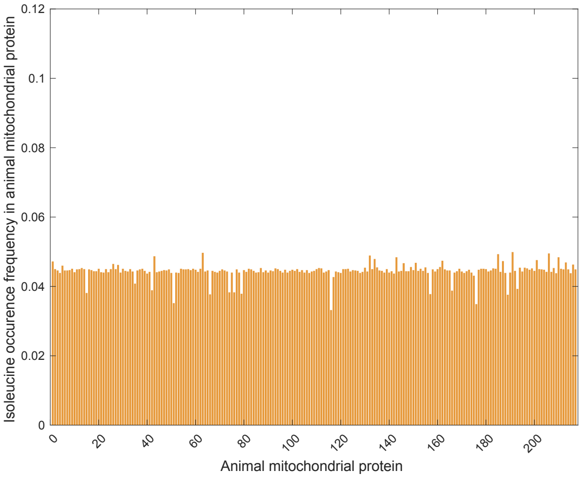

Cysteine occurence frequency in animal mitochondrial protein
0.12
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
120
140
160
180
200
Animal mitochondrial protein


Glutamid acid occurence frequency in animal mitochondrial protein
0.12
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
120
140
160
180
200
Animal mitochondrial protein

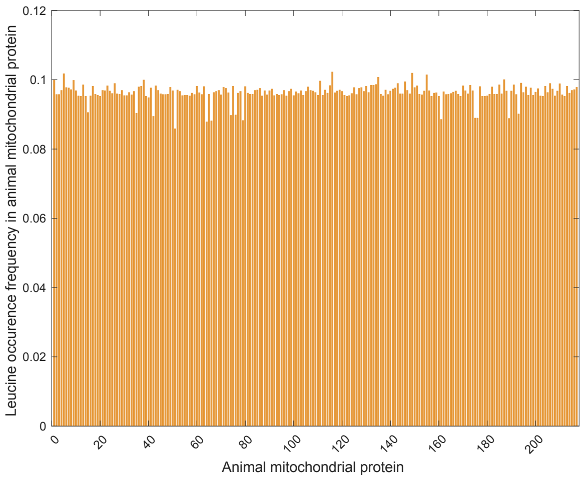

Glutamine occurence frequency in animal mitochondrial protein
0.12
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
120
140
160
180
200
Animal mitochondrial protein

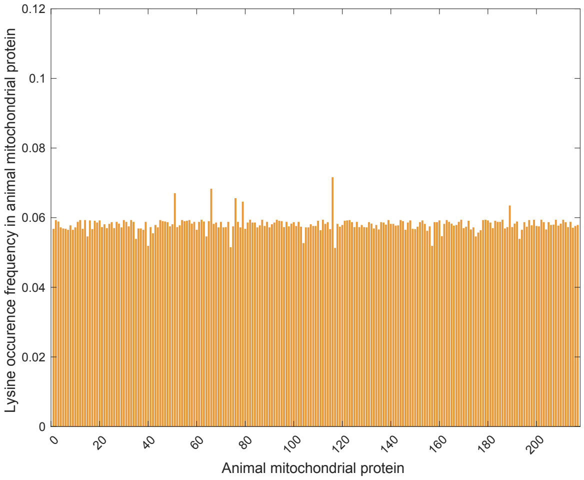

Glycine occurence frequency in animal mitochondrial protein
0.12
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
120
140
160
180
200
Animal mitochondrial protein

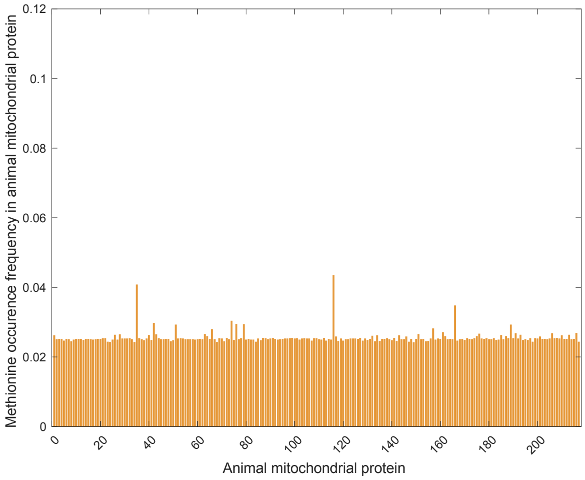

Histidine occurence frequency in animal mitochondrial protein
0.12
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
120
140
160
180
200
Animal mitochondrial protein

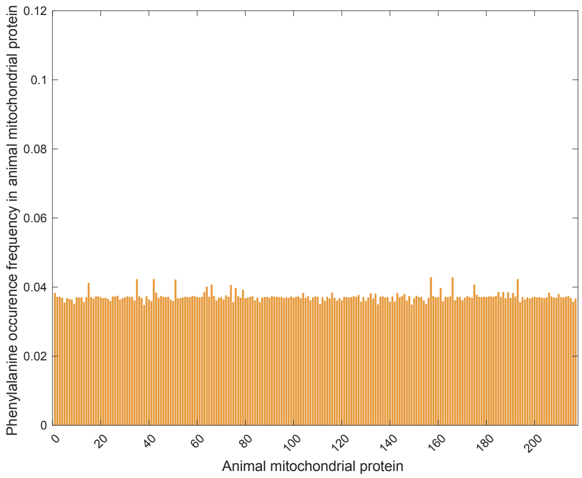

Isoleucine occurence frequency in animal mitochondrial protein
0.12
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
120
140
160
180
200
Animal mitochondrial protein

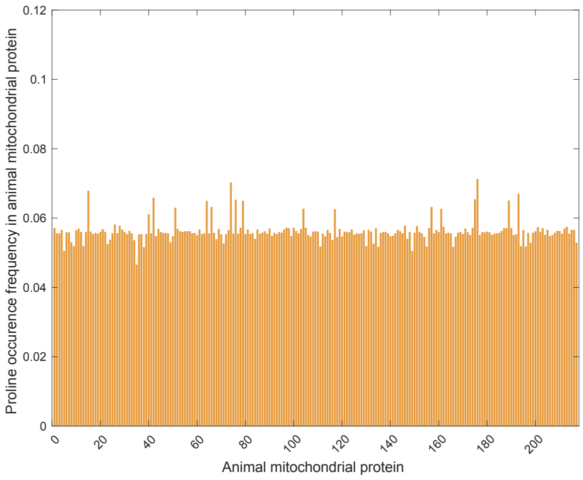
`

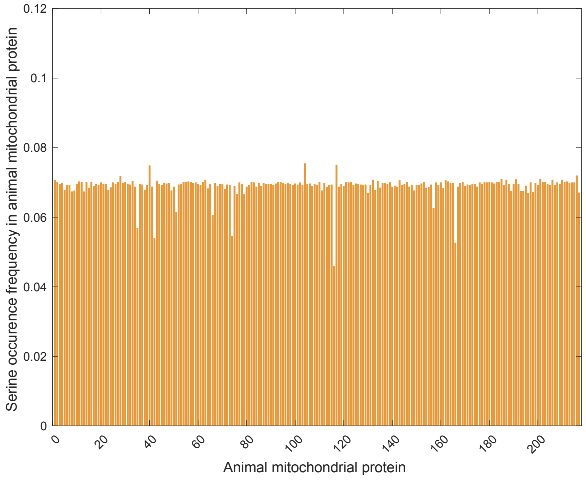

0.12
0.1
0.08
0.06
0.04
0.02
0
Leucine occurence frequency in animal mitochondrial protein
0
20
40
60
80
100
120
140
160
180
200
Animal mitochondrial protein


0.12
0.1
0.08
0.06
0.04
0.02
0
Lysine occurence frequency in animal mitochondrial protein
0
20
40
60
80
100
120
140
160
180
200
Animal mitochondrial protein


0.12
0.1
0.08
0.06
0.04
0.02
0
Methionine occurence frequency in animal mitochondrial protein
0
20
40
60
80
100
120
140
160
180
200
Animal mitochondrial protein


0.12
0.1
0.08
0.06
0.04
0.02
0
Phenylalanine occurence frequency in animal mitochondrial protein
0
20
40
60
80
100
120
140
160
180
200
Animal mitochondrial protein


0.12
0.1
0.08
0.06
0.04
0.02
0
Proline occurence frequency in animal mitochondrial protein
0
20
40
60
80
100
120
140
160
180
200
Animal mitochondrial protein


0.12
0.1
0.08
0.06
0.04
0.02
0
Serine occurence frequency in animal mitochondrial protein
0
20
40
60
80
100
120
140
160
180
200
Animal mitochondrial protein

`

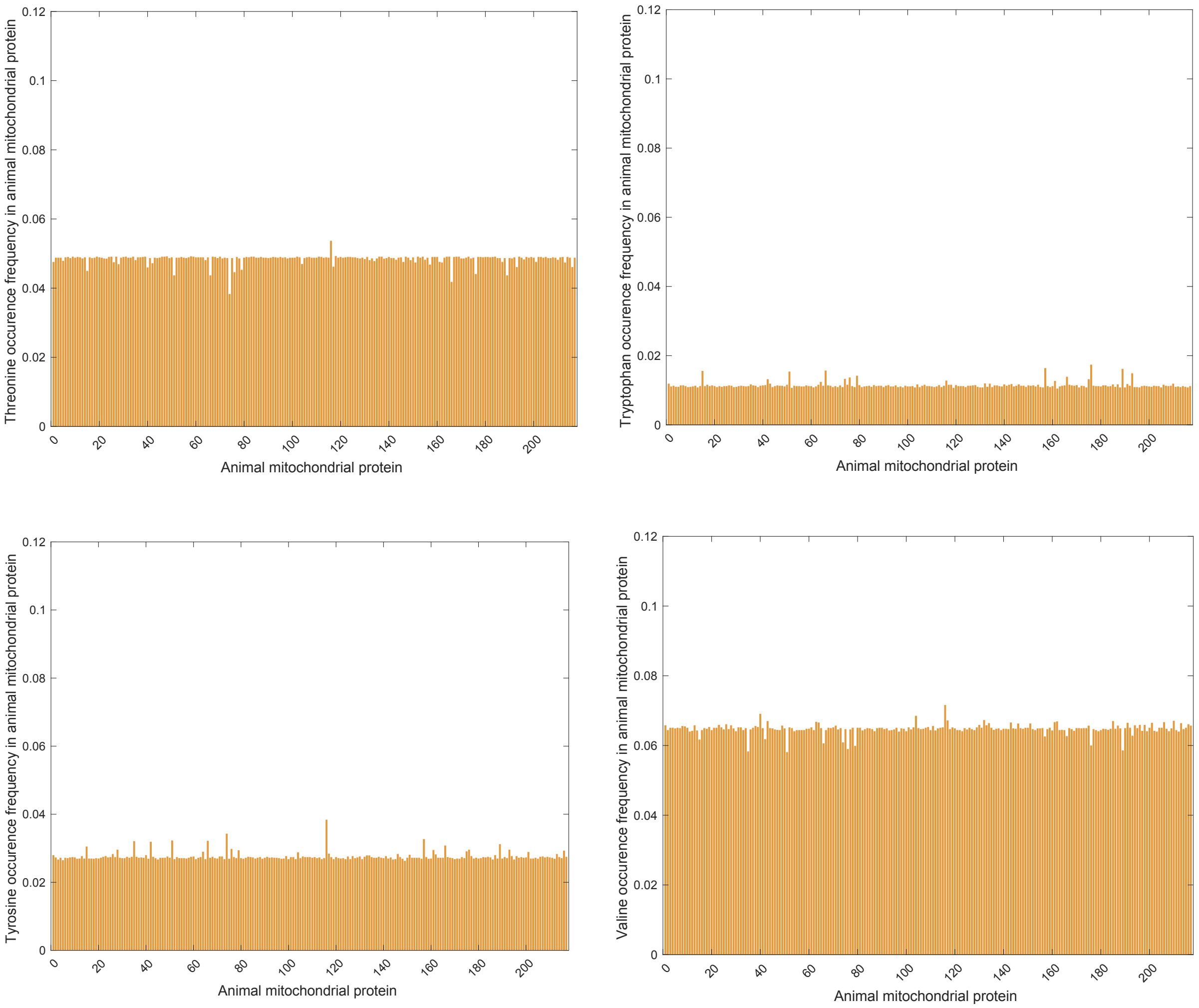
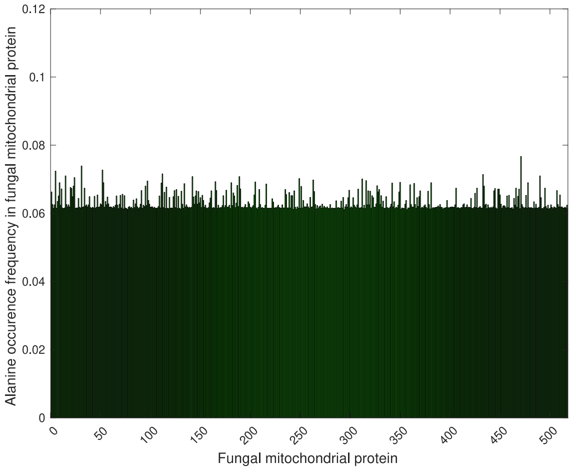
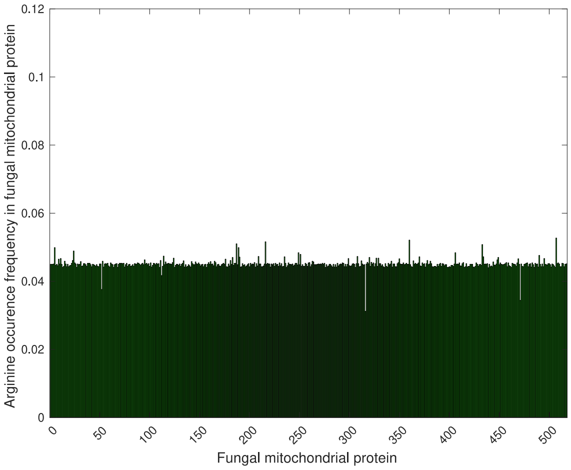
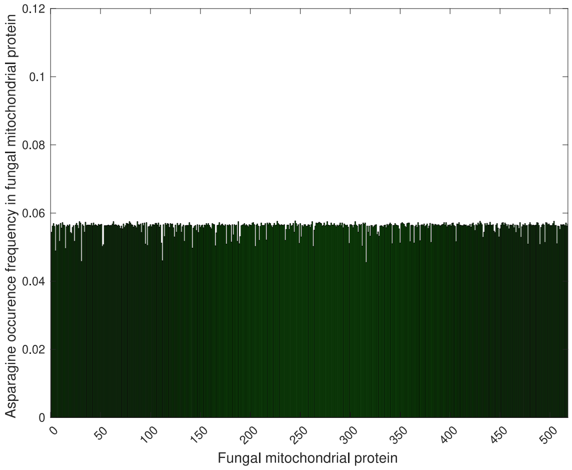


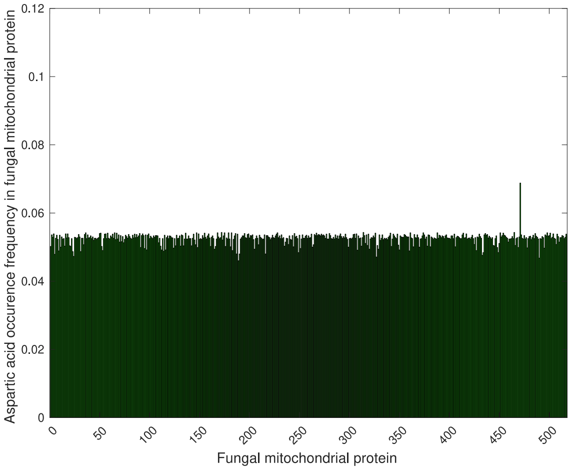
Figure A.4 Occurrence frequency of animal mitochondrial proteins for each amino acid [103]. The list of these mitochondrial proteins can be found in Table A.2. The standard deviations for these plots vary between ± 0.0009 and ± 0.007, in which the minimum standard deviation corresponds to tryptophan and the maximum standard deviation corresponds to glycine.

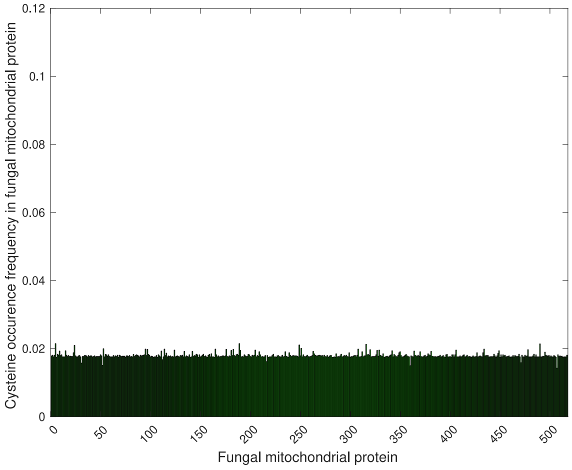
`

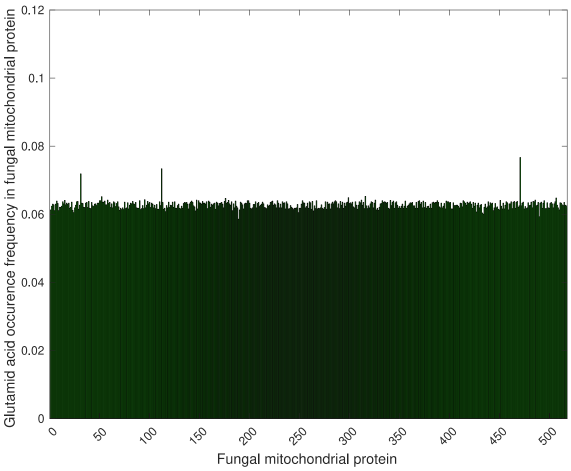

Alanine occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein

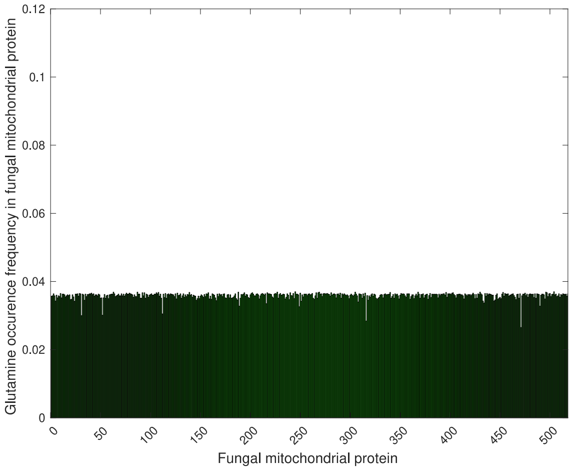

Arginine occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein

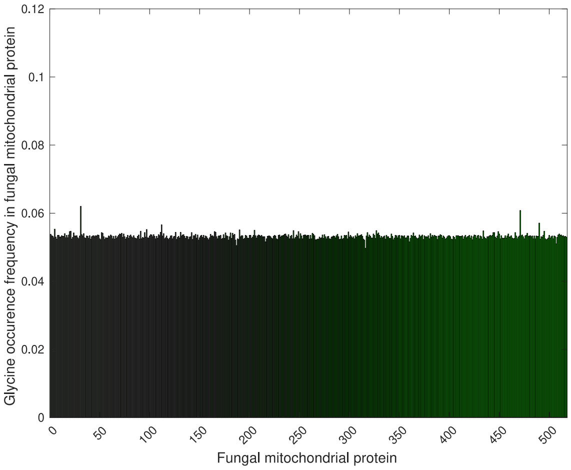

Asparagine occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein

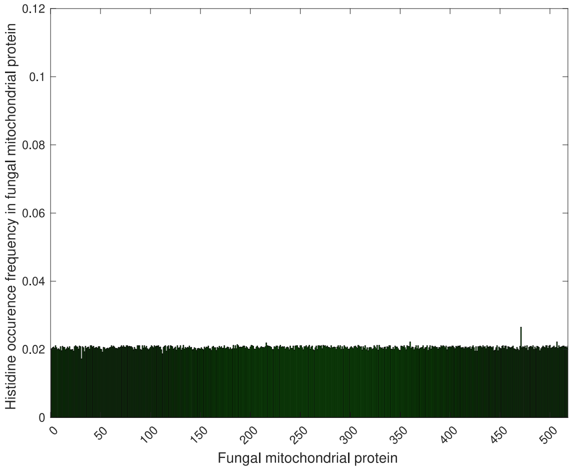

Aspartic acid occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein

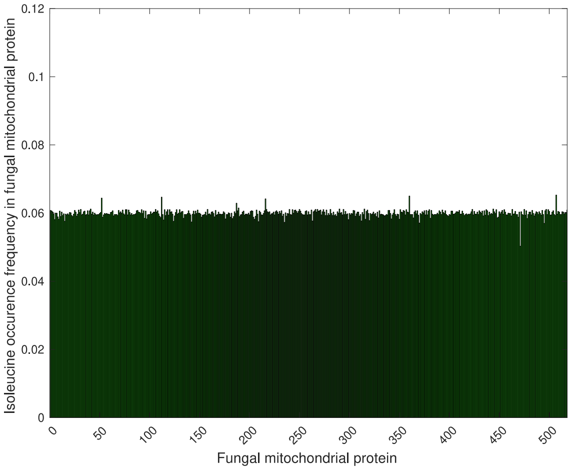

Cysteine occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein

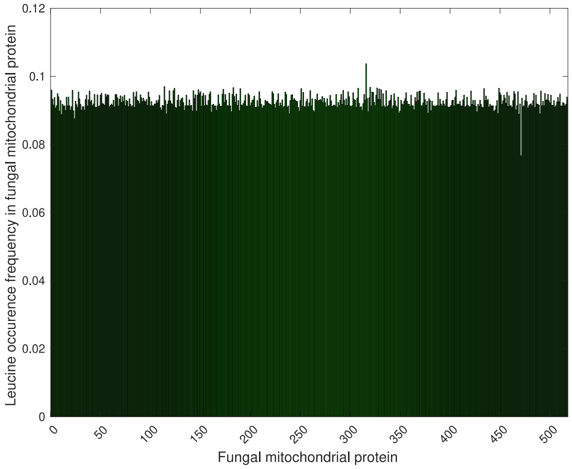

Glutamid acid occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein

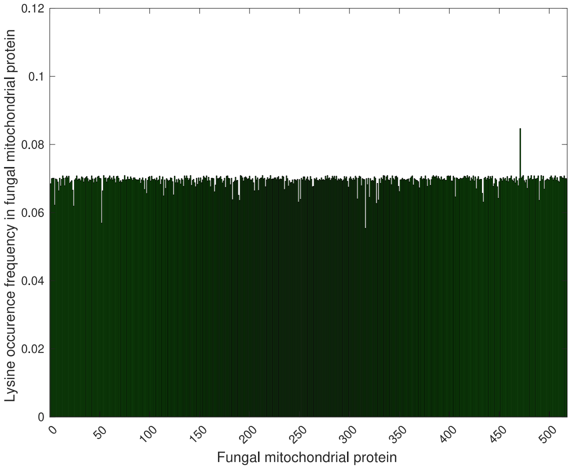

Glutamine occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein


Glycine occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein

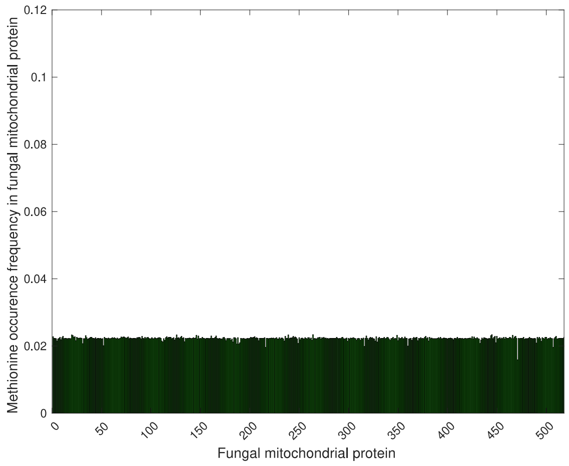

Histidine occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein

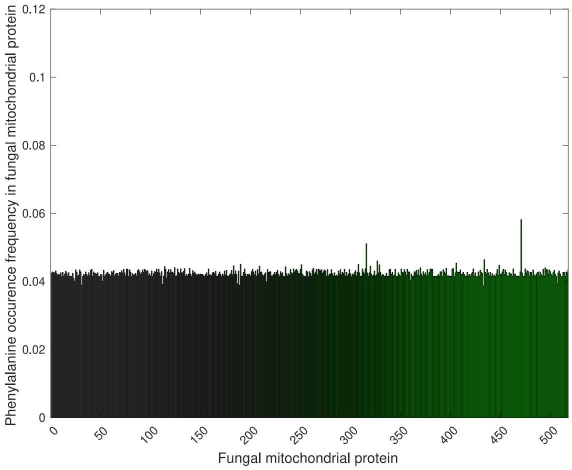

Isoleucine occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein

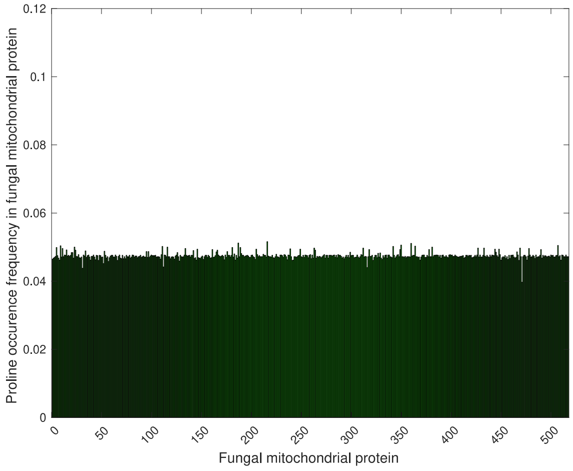

Leucine occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein

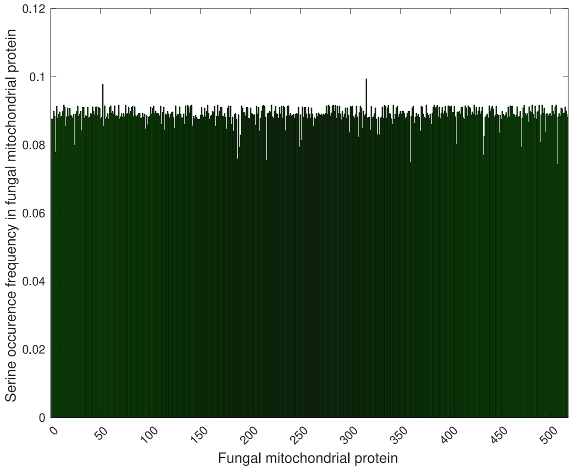

Lysine occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein

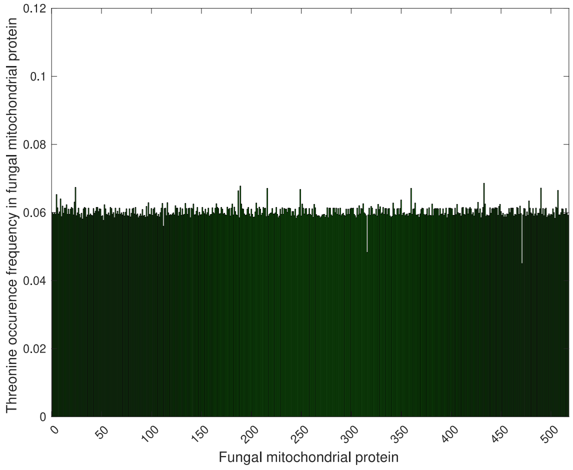
`

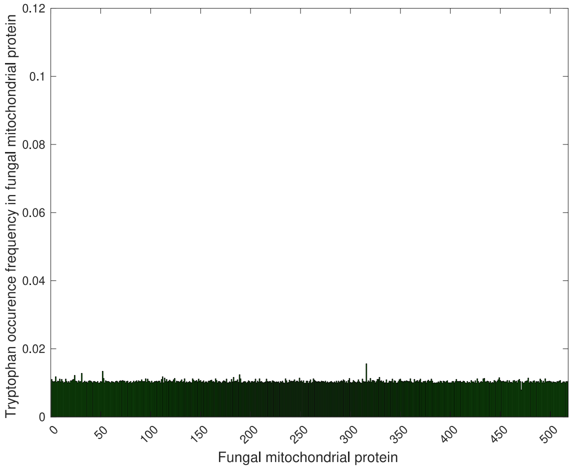

Methionine occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein


Phenylalanine occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein


Proline occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein


Serine occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein


Threonine occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein


Tryptophan occurence frequency in fungal mitochondrial protein
0
0.02
0.04
0.06
0.08
0.1
0.12
0
50
100
150
200
250
300
350
400
450
500
Fungal mitochondrial protein

`

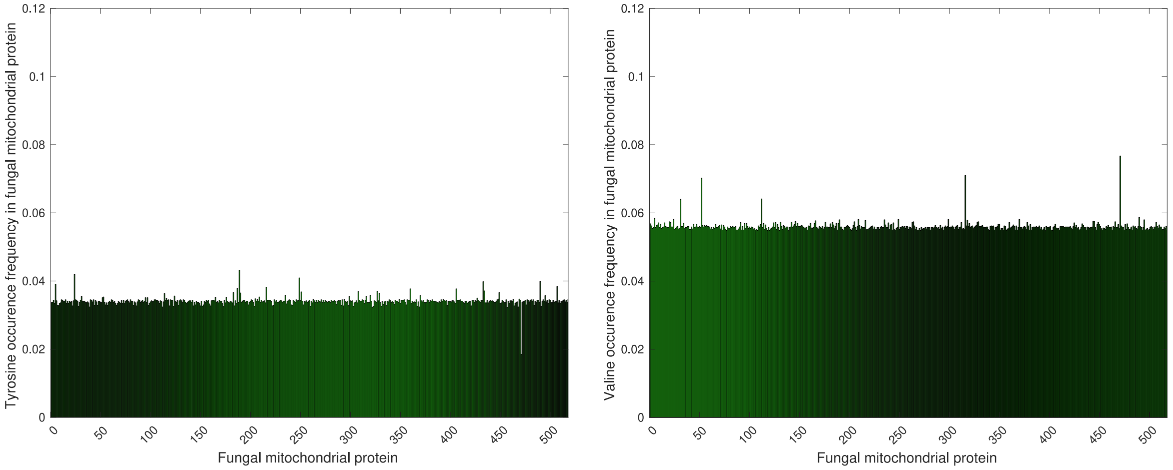
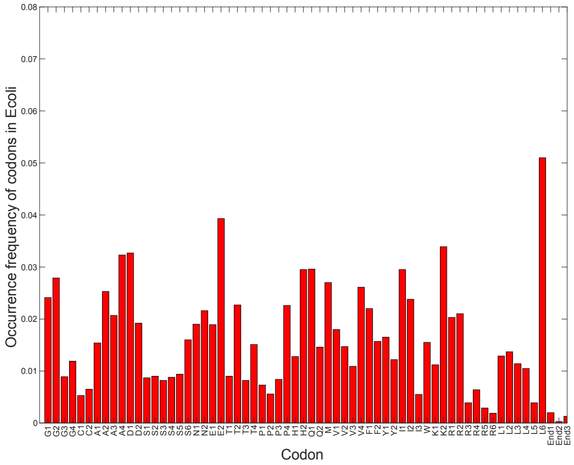
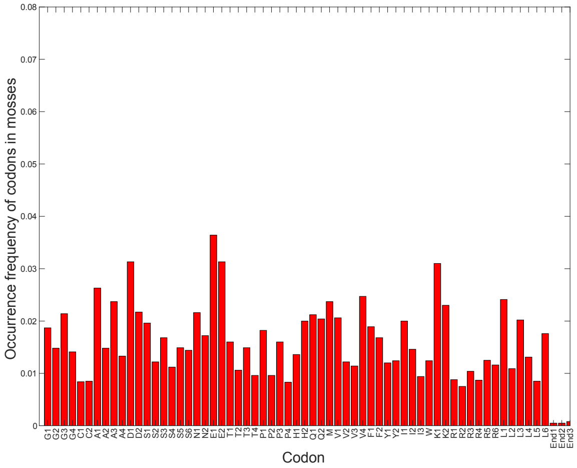


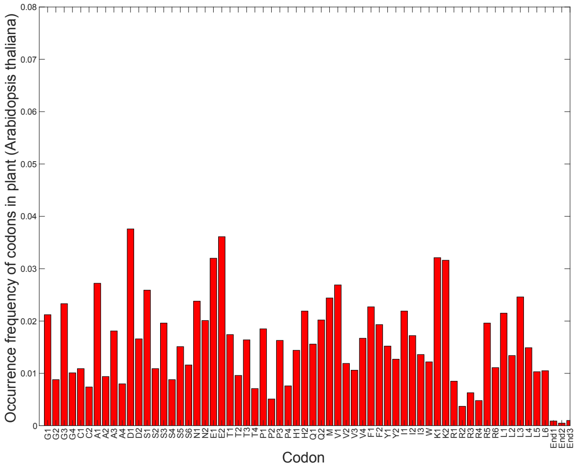
Figure A.5 Occurrence frequency of fungal mitochondrial proteins for each amino acid [103]. See Table A.2 for the list of the corresponding genes. The standard deviations vary between ±0.0004 and ±0.002, where the minimum standard deviation corresponds to tryptophan and the maximum standard deviation corresponds to serine.

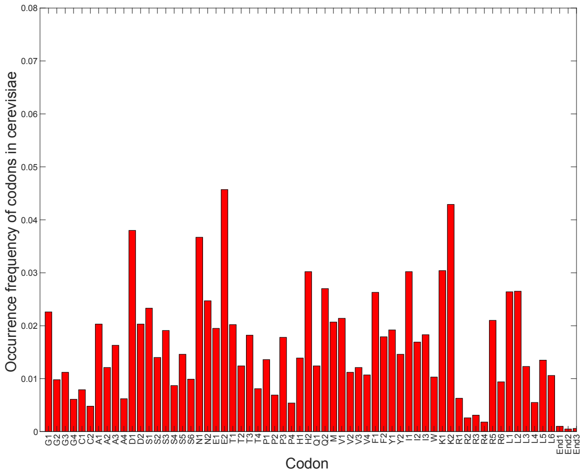
A.5 Frequency of codons in different creatures from *E. coli* to human

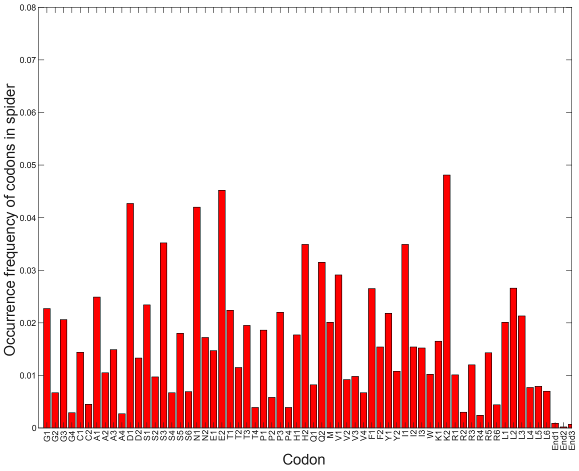


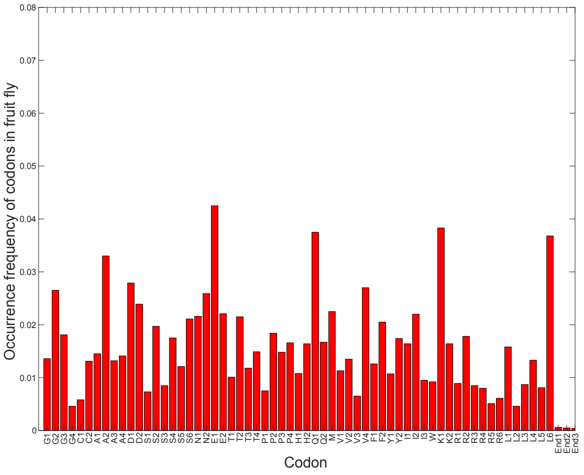


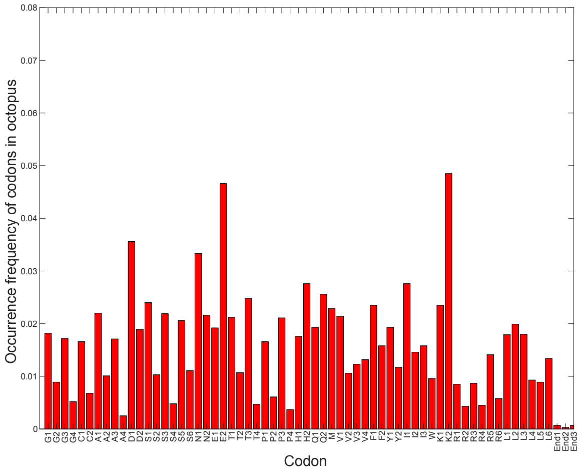
`

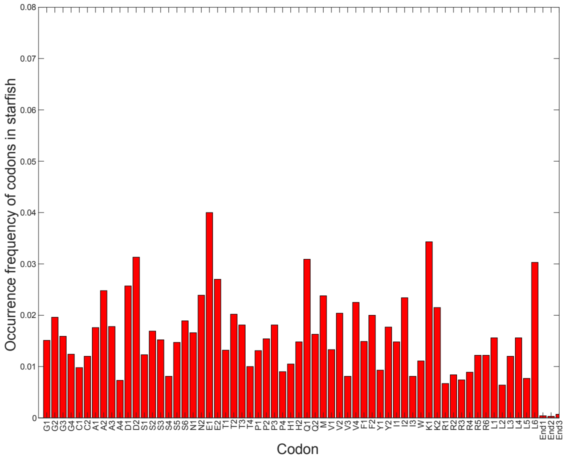

Occurrence frequency of codons in plant (Arabidopsis thaliana)
0 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08
G1 G2 G3 G4 C1 C2 A1 A2 A3 A4 D1 D2 S1 S2 S3 S4 S5 S6 N1 N2 E1 E2 T1 T2 T3 T4 P1 P2 P3 P4 H1 H2 Q1 Q2 M V1 V2 V3 V4 F1 F2 Y1 Y2 I1 I2 I3 W K1 K2 R1 R2 R3 R4 R5 R6 L1 L2 L3 L4 L5 L6 End1 End2 End3
Codon


Occurrence frequency of codons in cerevisiae
0 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08
G1 G2 G3 G4 C1 C2 A1 A2 A3 A4 D1 D2 S1 S2 S3 S4 S5 S6 N1 N2 E1 E2 T1 T2 T3 T4 P1 P2 P3 P4 H1 H2 Q1 Q2 M V1 V2 V3 V4 F1 F2 Y1 Y2 I1 I2 I3 W K1 K2 R1 R2 R3 R4 R5 R6 L1 L2 L3 L4 L5 L6 End1 End2 End3
Codon

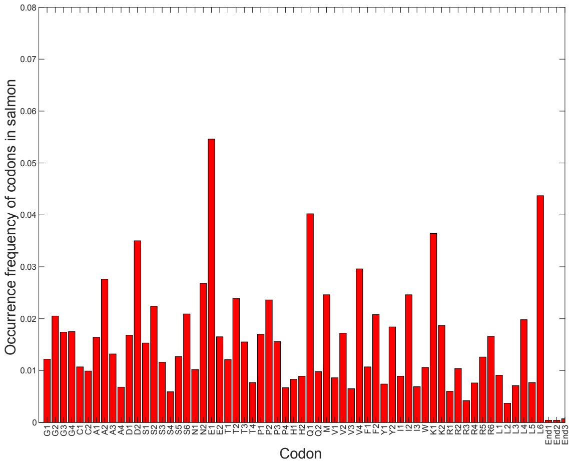

Occurrence frequency of codons in spider
0 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08
G1 G2 G3 G4 C1 C2 A1 A2 A3 A4 D1 D2 S1 S2 S3 S4 S5 S6 N1 N2 E1 E2 T1 T2 T3 T4 P1 P2 P3 P4 H1 H2 Q1 Q2 M V1 V2 V3 V4 F1 F2 Y1 Y2 I1 I2 I3 W K1 K2 R1 R2 R3 R4 R5 R6 L1 L2 L3 L4 L5 L6 End1 End2 End3
Codon

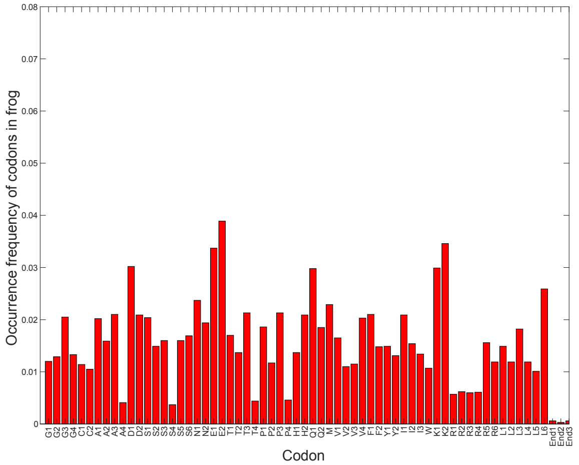

Occurrence frequency of codons in fruit fly
0 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08
G1 G2 G3 G4 C1 C2 A1 A2 A3 A4 D1 D2 S1 S2 S3 S4 S5 S6 N1 N2 E1 E2 T1 T2 T3 T4 P1 P2 P3 P4 H1 H2 Q1 Q2 M V1 V2 V3 V4 F1 F2 Y1 Y2 I1 I2 I3 W K1 K2 R1 R2 R3 R4 R5 R6 L1 L2 L3 L4 L5 L6 End1 End2 End3
Codon

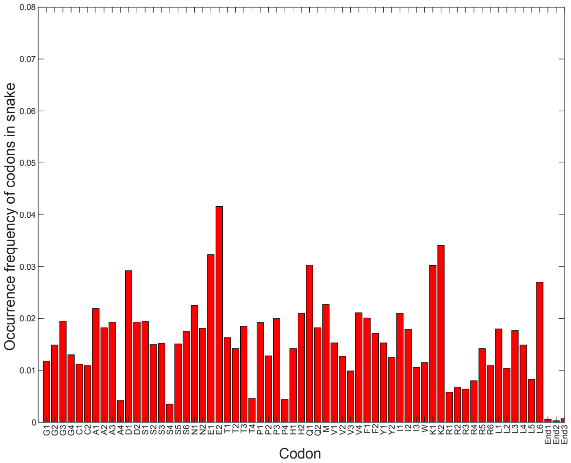

Occurrence frequency of codons in octopus
0 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08
G1 G2 G3 G4 C1 C2 A1 A2 A3 A4 D1 D2 S1 S2 S3 S4 S5 S6 N1 N2 E1 E2 T1 T2 T3 T4 P1 P2 P3 P4 H1 H2 Q1 Q2 M V1 V2 V3 V4 F1 F2 Y1 Y2 I1 I2 I3 W K1 K2 R1 R2 R3 R4 R5 R6 L1 L2 L3 L4 L5 L6 End1 End2 End3
Codon

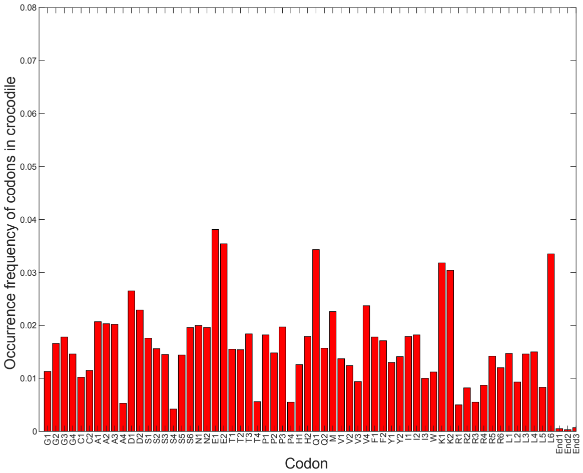

Occurrence frequency of codons in starfish
0 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08
G1 G2 G3 G4 C1 C2 A1 A2 A3 A4 D1 D2 S1 S2 S3 S4 S5 S6 N1 N2 E1 E2 T1 T2 T3 T4 P1 P2 P3 P4 H1 H2 Q1 Q2 M V1 V2 V3 V4 F1 F2 Y1 Y2 I1 I2 I3 W K1 K2 R1 R2 R3 R4 R5 R6 L1 L2 L3 L4 L5 L6 End1 End2 End3
Codon

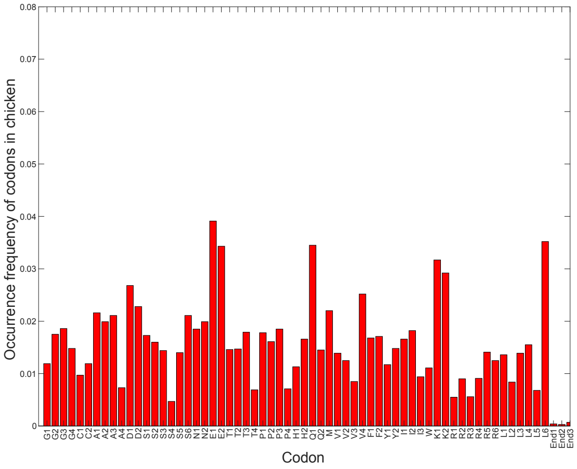
`

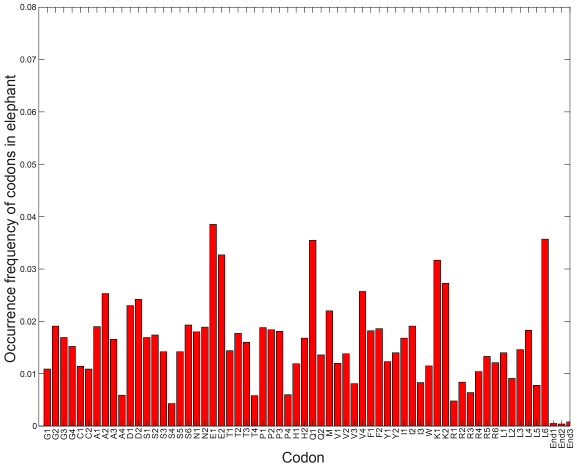

Occurrence frequency of codons in salmon
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
G1 G2 G3 G4 C1 C2 A1 A2 A3 A4 D1 D2 S1 S2 S3 S4 S5 S6 N1 N2 E1 E2 T1 T2 T3 T4 P1 P2 P3 P4 H1 H2 Q1 Q2 M V1 V2 V3 V4 F1 F2 Y1 Y2 I1 I2 I3 W K1 K2 R1 R2 R3 R4 R5 R6 L1 L2 L3 L4 L5 L6 End1 End2 End3
Codon


Occurrence frequency of codons in frog
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
G1 G2 G3 G4 C1 C2 A1 A2 A3 A4 D1 D2 S1 S2 S3 S4 S5 S6 N1 N2 E1 E2 T1 T2 T3 T4 P1 P2 P3 P4 H1 H2 Q1 Q2 M V1 V2 V3 V4 F1 F2 Y1 Y2 I1 I2 I3 W K1 K2 R1 R2 R3 R4 R5 R6 L1 L2 L3 L4 L5 L6 End1 End2 End3
Codon


Occurrence frequency of codons in snake
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
G1 G2 G3 G4 C1 C2 A1 A2 A3 A4 D1 D2 S1 S2 S3 S4 S5 S6 N1 N2 E1 E2 T1 T2 T3 T4 P1 P2 P3 P4 H1 H2 Q1 Q2 M V1 V2 V3 V4 F1 F2 Y1 Y2 I1 I2 I3 W K1 K2 R1 R2 R3 R4 R5 R6 L1 L2 L3 L4 L5 L6 End1 End2 End3
Codon


Occurrence frequency of codons in crocodile
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
G1 G2 G3 G4 C1 C2 A1 A2 A3 A4 D1 D2 S1 S2 S3 S4 S5 S6 N1 N2 E1 E2 T1 T2 T3 T4 P1 P2 P3 P4 H1 H2 Q1 Q2 M V1 V2 V3 V4 F1 F2 Y1 Y2 I1 I2 I3 W K1 K2 R1 R2 R3 R4 R5 R6 L1 L2 L3 L4 L5 L6 End1 End2 End3
Codon


Occurrence frequency of codons in chicken
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
G1 G2 G3 G4 C1 C2 A1 A2 A3 A4 D1 D2 S1 S2 S3 S4 S5 S6 N1 N2 E1 E2 T1 T2 T3 T4 P1 P2 P3 P4 H1 H2 Q1 Q2 M V1 V2 V3 V4 F1 F2 Y1 Y2 I1 I2 I3 W K1 K2 R1 R2 R3 R4 R5 R6 L1 L2 L3 L4 L5 L6 End1 End2 End3
Codon


Occurrence frequency of codons in elephant
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
G1 G2 G3 G4 C1 C2 A1 A2 A3 A4 D1 D2 S1 S2 S3 S4 S5 S6 N1 N2 E1 E2 T1 T2 T3 T4 P1 P2 P3 P4 H1 H2 Q1 Q2 M V1 V2 V3 V4 F1 F2 Y1 Y2 I1 I2 I3 W K1 K2 R1 R2 R3 R4 R5 R6 L1 L2 L3 L4 L5 L6 End1 End2 End3
Codon

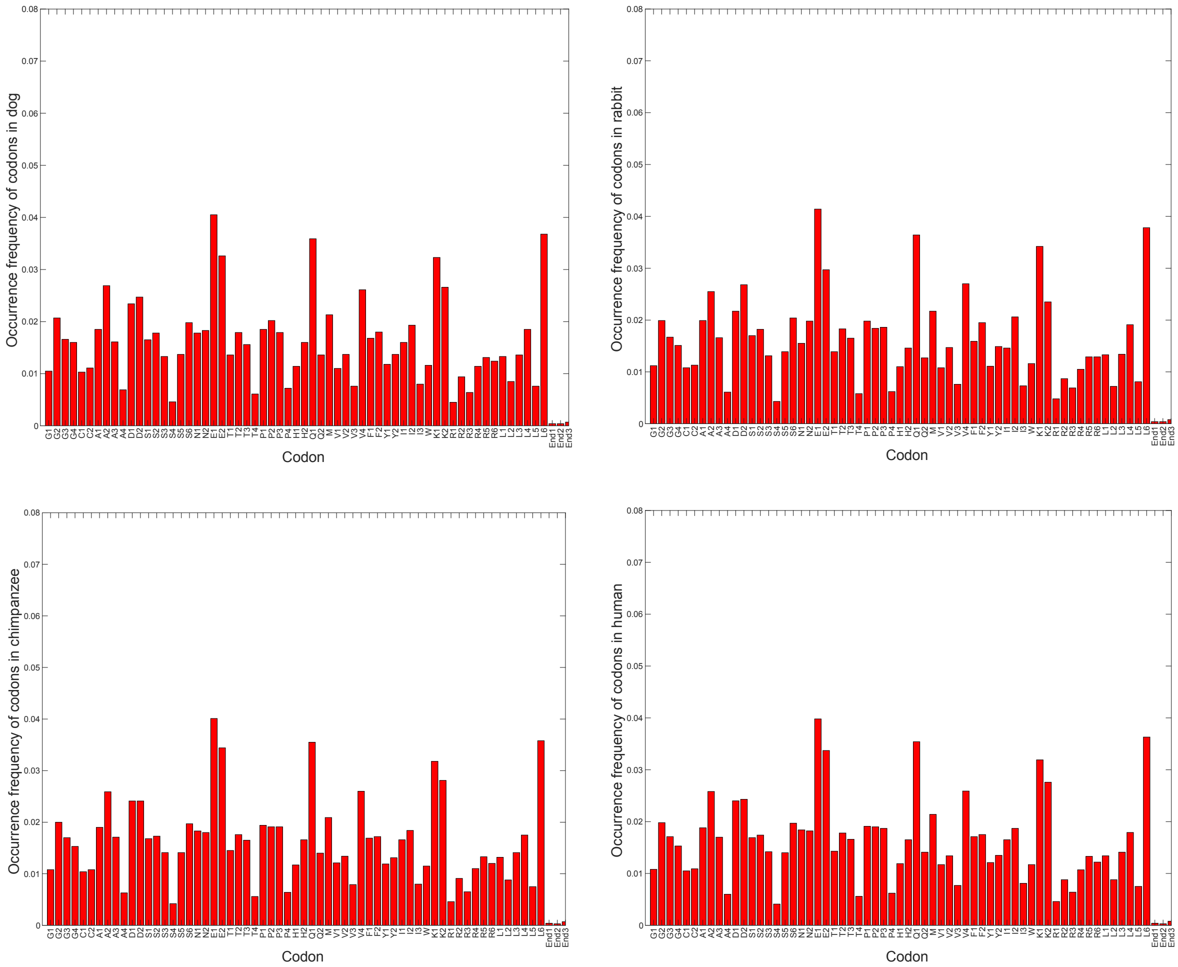


Figure A.6 Occurrence frequency of codons in different species from *E. coli* to human among synonymous codons (normalized to the highest probable codon). The frequency of each codon was obtained using the NCBI database. Along the x-axis the labels correspond to codons for each amino acid and are grouped together, i.e., glycine codons are shown as G1, G2, G3 and G4. The conversion from G# to the three letter codon can be found in Table A.3.

Table A.3 Amino acid and codon table notation.

Amino acid	**1-letter representation of codon (#)**	**3-letter representation of codon**	**Amino acid**	**1-letter representation of codon (#)**	**3-letter representation of codon**
Glycine	G1	GGT	**Glutamine**	Q1	CAG
	G2	GGC		Q2	CAA

`

	G3	GGA	**Methionine**	M	ATG
	G4	GGG	**Valin**	V1	GTT
Cysteine	C1	TGT		V2	GTC
	C2	TGC		V3	GTA
Alanine	A1	GCT		V4	GTG
	A2	GCC	**Phenylalanine**	F1	TTT
	A3	GCA		F2	TTC
	A4	GCG	**Tyrosine**	Y1	TAT
Aspartic acid	D1	GAT		Y2	TAC
	D2	GAC	**Isoleucine**	I1	ATT
Serine	S1	TCT		I2	ATC
	S2	TCC		I3	ATA
	S3	TCA	**Tryptophan**	W	TGG
	S4	TCG	**Lysine**	K1	AAA
	S5	AGT		K2	AAG
	S6	AGC	**Arginine**	R1	CGT
Asparagine	N1	AAT		R2	CGC
	N2	AAC		R3	CGA
Glutamic acid	E1	GAA		R4	CGG
	E2	GAG		R5	AGA
Threonine	T1	ACT		R6	AGG
	T2	ACC	**Leucine**	L1	TTA
	T3	ACA		L2	TTG
	T4	ACG		L3	CTT
Proline	P1	CCT		L4	CTC
	P2	CCC		L5	CTA
	P3	CCA		L6	CTG
	P4	CCG	**STOP**	STP1	TAA
Histidine	H1	CAT		STP2	TAG
	H2	CAC		STP3	TGA

`

Appendix B [7]

B.1 Steered Molecular Docking (SMD)

Using a suitable definition of the interaction potential, the basic interactions of the system are modeled by the given potential energy, which helps to predict the statistical and dynamical properties of the complex by including the intermolecular interactions. By stretching the structure from one or more ends toward the chosen degree of freedom or direction, SMD manipulates the protein/ligand structure in order to keep the ligand in the restraint point and analyze the behavior of the target protein [9,177–181]. This method allows us to model experimentally impossible situations. Hence, the implemented non-equilibrium work to the system during the steering level is associated with the difference between the free energy of the system going from one state, A, to the other state, B, by implementing the Jarzynski equality as [292]

$$G_B = G_A - \frac{1}{\beta} ln\langle e^{-\beta W_{A\to B}} \rangle_A \tag{B1}$$

where G stands for the Gibbs free energy and W is the performed work in a non-equilibrium state [183–185,293]. For the SMD simulations we used Amber14 [182]. Using the sander module, an Amber Hamiltonian has been applied to the system as a force field with the general form of

$$V(r) = \sum_{i,bound}^{n} \frac{\kappa_b}{2}\left(r_{ij} - \bar{r}_{ij}\right)^2 + \sum_{i,angles}^{n} \frac{\kappa_\theta}{2}(\theta_i - \bar{\theta}_i)^2 + \sum_{dihedrals} \frac{\kappa_\varphi}{2}[1 + \cos(n\varphi - \bar{\varphi})]^2$$

[7] This section has been submitted partially as supplementary material of Arbabi Moghadam S., Preto J., Klobukowski M., and Tuszynski J.A., "Testing amino acid-codon affinity hypothesis using molecular docking" in BioSystems journal.

`

$$+\sum_{i<j,nonb}\left(\frac{A_{ij}}{r_i^{12}}-\frac{B_{ij}}{r_i^{6}}\right)+\sum_{i<j}\left(\frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}\right) \qquad \text{(B2)}$$

where the first three terms above represent covalent interactions (bonds, angles, and dihedral angles) and the last two describe non-covalent interactions (represented by the Lennard-Jones and Coulombic potentials). Parameters $\bar{r}_{ij}$ (displacement), $\bar{\theta}_{ij}$ (angle) and $\bar{\varphi}$ (dihedral) are obtained at equilibrium and κ_b, κ_θ and κ_φ are obtained from the empirical values [182,186]. This potential can be used for applying forces to the protein-ligand complex in a solvent.

B.2 3dRPC

There are a few software packages or webservers available for protein-RNA docking; such as HDOCK and 3dRPC. We used the 3dRPC server [187–190] since HDOCK has a size limitation for ligands, and the RNA and amino-acids used in our study represent small complexes. The 3dRPC is a computational means that is being used to estimate 3D RNA-protein complex structures. The success rate of this method is higher than other RNA-protein docking methods and is comparable to the most common protein-protein docking methods that have been tested as a benchmark [189]. In order to predict the RNA-protein structural complex, this method is using a combination of the RPDOCK algorithm (explained below), for conformational sampling and the 3dRPC scoring method for choosing the most accurate and correct docked pose. The RPDOCK algorithm is constructed based on fast Fourier transform (FFT) docking algorithms, which consider the characteristics of possible interactions of any RNA-protein complex [187]. Electrostatic effects, geometric complementarities, as well as stacking interactions, are the three important factors that RPDOCK takes into account in an RNA-protein interface due to looser atom packing in an RNA-protein interface [188]. The 3dRPC-score is a potential in which nucleotide-residue conformations are considered as statistical parameters [190]. This method is also considered as a statistical potential in which the energy of each pair of the RNA-protein complex is affected by the conformation [190,294] and categorized by the RMSD (Root Mean Square Deviation) value. The 3dRPC-Score statistical potential is defined as

$$E_{ij}(C) = -\ln\left(\frac{P_{ij}(C)}{P_i P_j * P_v}\right) \tag{B3}$$

where $P_{ij}(C)$ is the probability of nucleotide type i and residue type j which are in the C-category, P_v in the ideal state, refers to the probability of C-category for nucleotide residue pairs. For all 20 amino acids and four nucleotides, there would be (4×20) combinations of possible nucleotide-amino acid pair which can be grouped into 10 classes using the K-means clustering method [190]. The success rate of the 3dRPC scoring method using RPDOCK decoys has been studied as a benchmark for 72 complexes by Huang *et al.*, who compared it to other methods such as IT-Score-PR and DECK-PR [190]. The success rate is measured by the number of correct predictions. In the first prediction, IT-Score-PR and 3dRPC-Score have a similar success rate, namely46% and DECK-PR stands on 36%. In the first ten predictions, IT-Score-PR performed better than 3dRPC-Score, although the latter performed better than DECK-PR. However, after the top ten predicted poses, 3dRPC-Score and IT-Score-PR perform in a similar way, and both had a better success rate than DECK-PR. In general, based on the IRMSD factor between unbound and native structures, the benchmarked cases are divided into three categories, easy, medium, and difficult targets $\text{IRMSD} > 2.5$ Å, $2.5 \leq \text{IRMSD} \leq 5\text{Å}$ and $\text{IRMSD} > 5\text{Å}$, respectively [187,188,190,294]. For easy cases, 3dRPC-Score and IT-Score-PR have similar performance measures but better rank than

DECK-PR. For difficult cases, all the methods fail. More details of the 3dRPC method can be found elsewhere [189]. In this work, the best-ranked poses by 3dRPC were used.

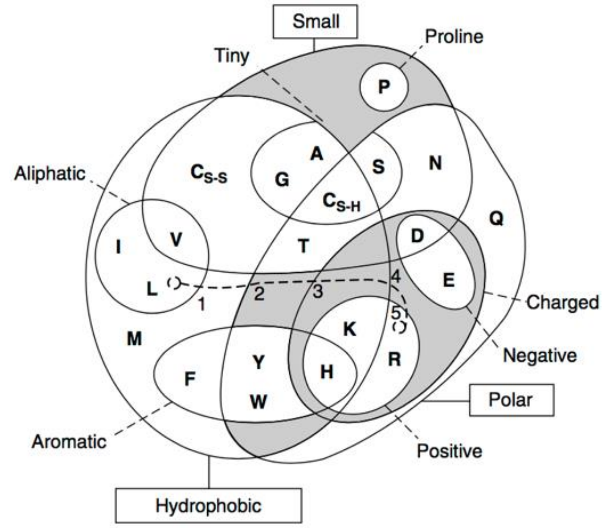


Figure B.1 Venn diagram illustrating the properties of amino acids [99].

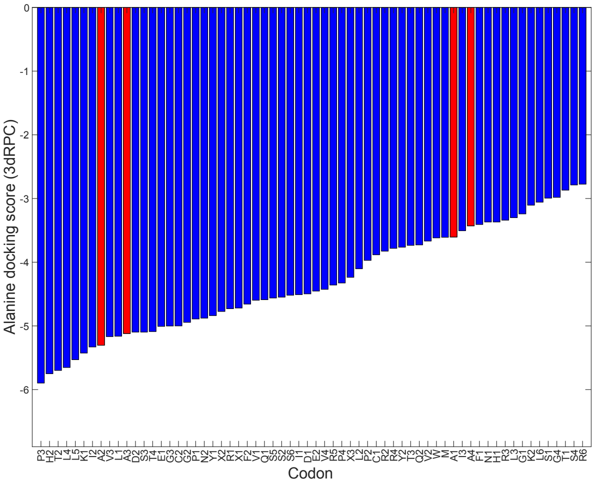
Table B.1 Amino acid and codon table notation.

Amino acid	Codon		Anticodon		Amino acid	Codon		Anticodon	
Glycine	G1	GGT	T2	ACC	**Glutamine**	Q1	CAG	L2	UUG
	G2	GGC	A2	GCC		Q2	CAA	L6	CUG
	G3	GGA	S2	UCC	**Methionine**	M	AUG	H1	CAU
	G4	GGG	P2	CCC	**Valin**	V1	GUU	N2	AAC
Cysteine	C1	UGU	T3	ACA		V2	GUC	D2	GAC
	C2	UGC	A3	GCA		V3	GUA	Y2	UAC
Alanine	A1	GCU	S6	AGC		V4	GUG	H2	CAC
	A2	GCC	G2	GGC	**Phenylalanine**	F1	UUU	K1	AAA
	A3	GCA	C2	UGC		F2	UUC	E1	GAA
	A4	GCG	R2	CGC	**Tyrosine**	Y1	UAU	I3	AUA
Aspartic acid	D1	GAU	I2	AUC		Y2	UAC	V3	GUA
	D2	GAC	V2	GUC	**Isoleucine**	I1	AUU	N1	AAU

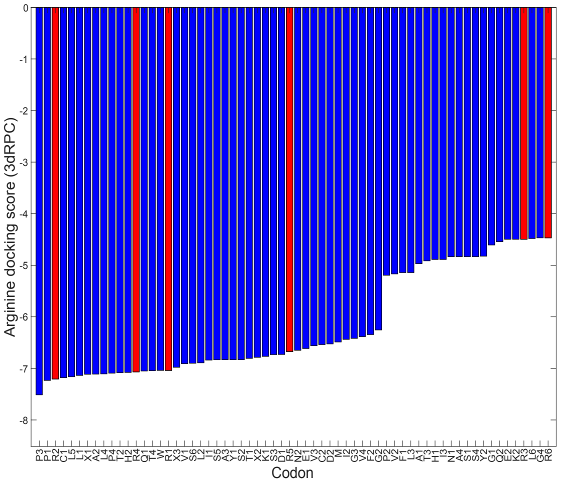
`

Serine	S1	UCU	R5	AGA		I2	AUC	D1	GAU
	S2	UCC	G3	GGA		I3	AUA	Y1	UAU
	S3	UCA	STP1	UGA	**Tryptophan**	W	UGG	P3	CCA
	S4	UCG	A3	CGA	**Lysine**	K1	AAA	F1	UUU
	S5	AGU	T1	ACU		K2	AAG	Q2	CAA
	S6	AGC	A1	GCU	**Arginine**	R1	CGU	T4	ACG
Asparagine	N1	AAU	S5	AGU		R2	CGC	A2	GCG
	N2	AAC	G1	GGU		R3	CGA	S4	UCG
Glutamic acid	E1	GAA	F2	UUC		R4	CGG	P4	CCG
	E2	GAG	L4	CUC		R5	AGA	S1	UCI
Threonine	T1	ACU	S5	AGU		R6	AGG	P1	CCU
	T2	ACC	G1	GGU	**Leucine**	L1	UUA	STP1	UAA
	T3	ACA	C1	UGT		L2	UUG	Q2	CAA
	T4	ACG	R1	CGU		L3	CUU	K2	AAG
Proline	P1	CCU	R6	AGG		L4	CUC	E2	GAG
	P2	CCC	G4	GGG		L5	CUA	STP2	UAG
	P3	CCA	W	UGG		L6	CUG	R3	CAG
	P4	CCG	R4	CGG	**STOP**	STP1	UAA	L1	UUA
Histidine	H1	CAU	M	AUG		STP2	UAG	L5	CUA
	H2	CAC	V4	GUG		STP3	UGA	S3	UCA

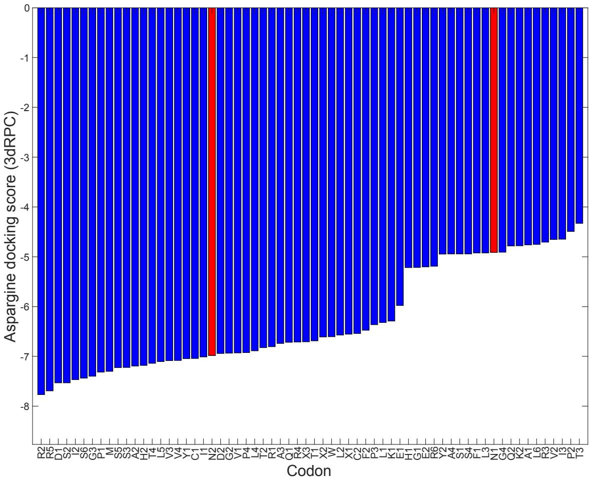
`



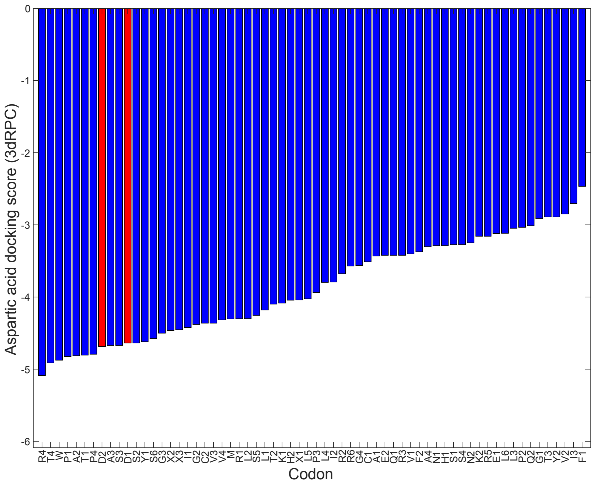
Alanine docking score (3dRPC)
0
-1
-2
-3
-4
-5
-6
Codon
P3
H2
T2
L4
L5
K1
I2
A2
V3
L1
A3
D2
S3
T4
E1
G3
C2
G2
P1
N2
Y1
X2
R1
X1
F2
V1
Q1
S5
S2
S6
I1
D1
E2
V4
R5
P4
X3
L2
P2
C1
R2
R4
Y2
T3
Q2
V2
W
M
A1
I3
A4
F1
N1
H1
R3
L3
G1
K2
L6
S1
G4
T1
S4
R6



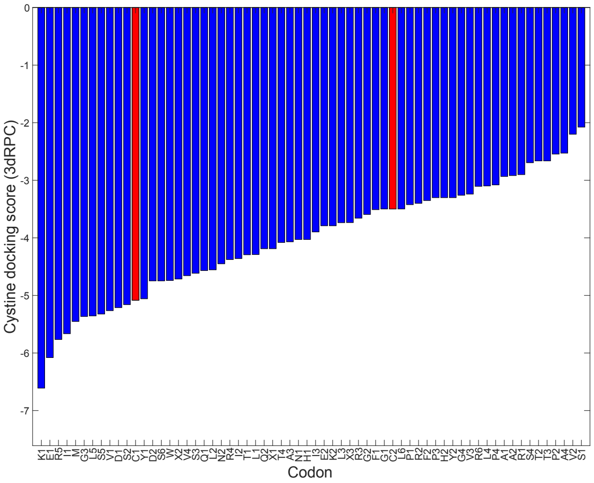
Arginine docking score (3dRPC)
0
-1
-2
-3
-4
-5
-6
-7
-8
Codon
P3
P1
R2
C1
L5
L1
X1
A2
L4
P4
T2
H2
R4
Q1
T4
W
R1
X3
V1
S6
L2
I1
S5
A3
Y1
S2
T1
X2
K1
S3
D1
R5
N2
E1
V3
C2
D2
M
I2
G3
V4
F2
G2
P2
V2
F1
L3
A1
T3
H1
I3
N1
A4
S1
S4
Y2
G1
Q2
E2
K2
R3
L6
G4
R6



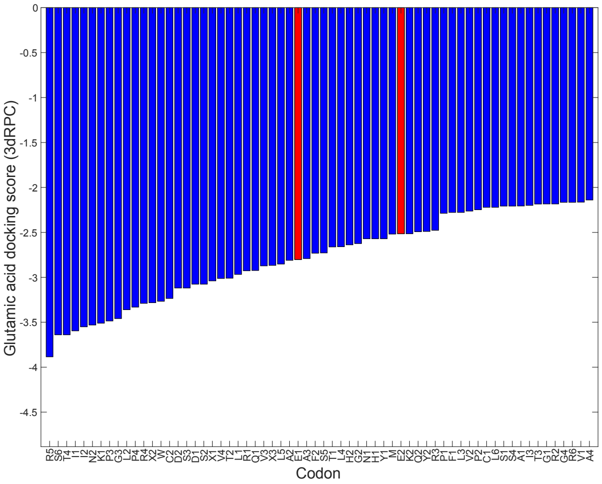
Aspargine docking score (3dRPC)
0
-1
-2
-3
-4
-5
-6
-7
-8
Codon
R2
R5
D1
S2
I2
S6
G3
P1
M
S5
S3
A2
H2
T4
L5
V3
V4
Y1
C1
I1
N2
D2
G2
V1
P4
L4
T2
R1
A3
Q1
R4
X3
T1
X2
W
L2
X1
C2
F2
P3
L1
K1
E1
H1
G1
E2
R6
Y2
A4
S1
S4
F1
L3
N1
G4
Q2
K2
A1
L6
R3
V2
I3
P2
T3



Aspartic acid docking score (3dRPC)
0
-1
-2
-3
-4
-5
-6
Codon
R4
T4
W
P1
A2
T1
P4
D2
A3
S3
D1
S2
Y1
S6
G3
X2
X3
I1
G2
C2
V3
V4
M
R1
L2
S5
L1
T2
K1
H2
X1
L5
P3
L4
I2
R2
R6
G4
C1
A1
E2
Q1
R3
V1
F2
A4
N1
H1
S1
S4
N2
K2
R5
E1
L6
L3
P2
Q2
G1
T3
Y2
V2
I3
F1

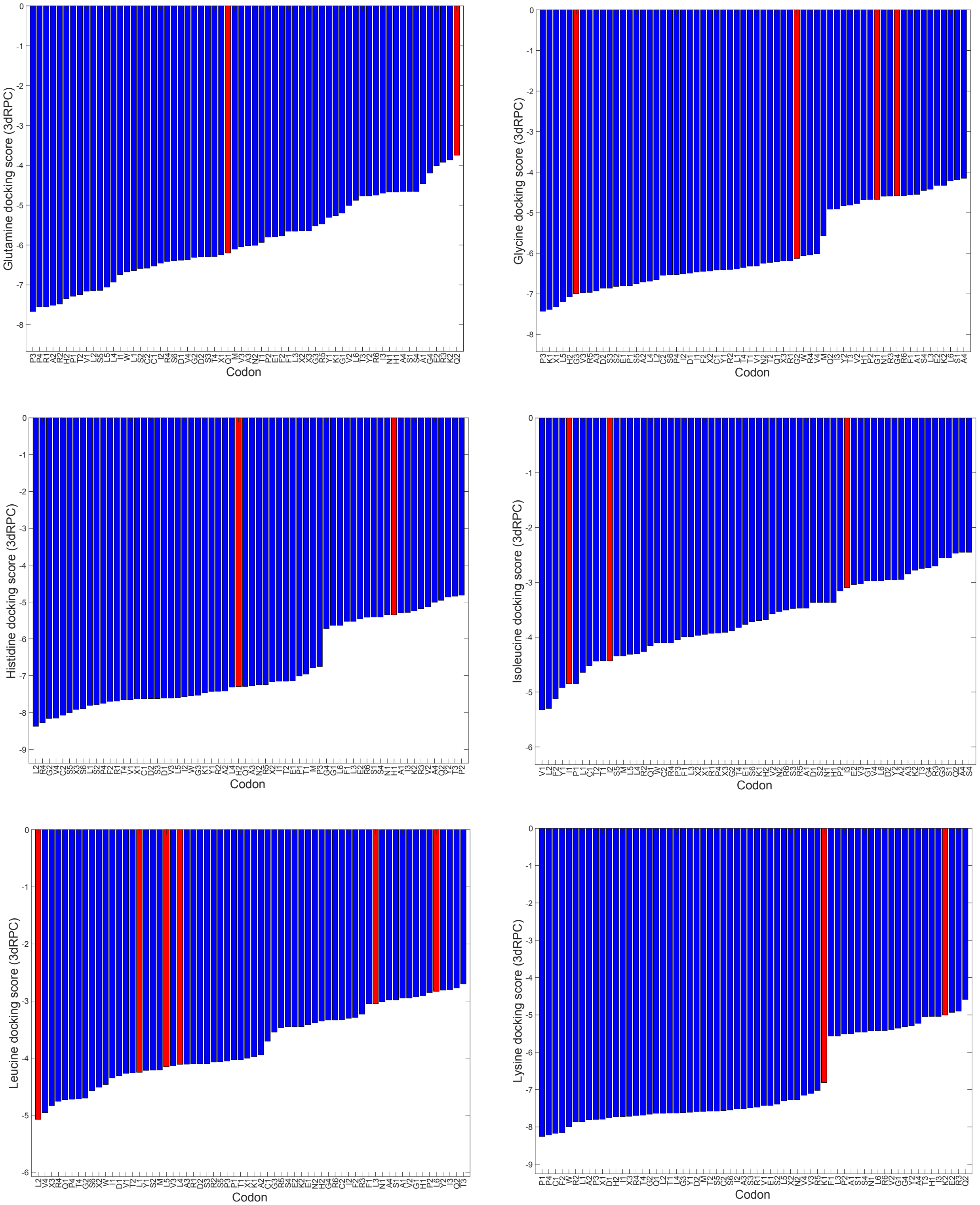


Cystine docking score (3dRPC)
0
-1
-2
-3
-4
-5
-6
-7
Codon
K1
E1
R5
I1
M
G3
L5
S5
V1
D1
S2
C1
Y1
D2
S6
W
X2
V4
S3
Q1
L2
N2
R4
I2
T1
L1
Q2
X1
T4
A3
N1
H1
I3
E2
K2
L3
X3
R3
G2
F1
G1
C2
L6
P1
R2
F2
P3
H2
Y2
G4
V3
R6
L4
P4
A1
A2
R1
S4
T2
T3
P2
A4
V2
S1

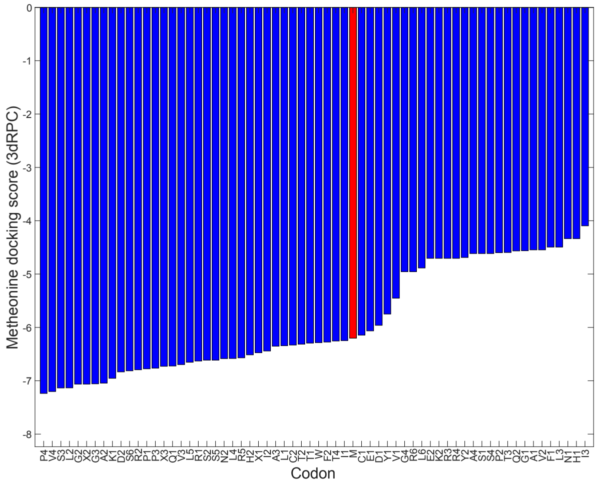


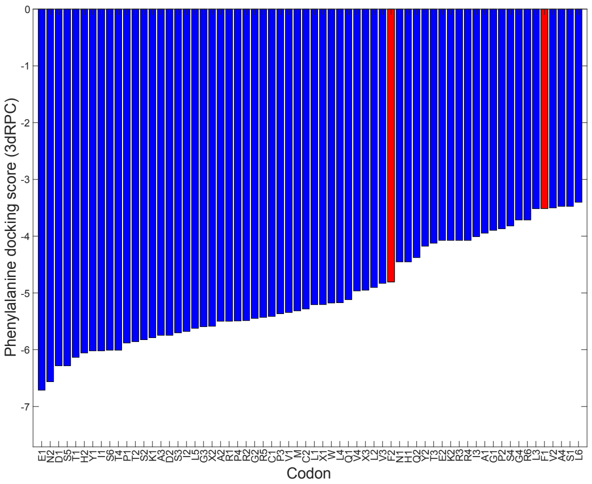
Glutamic acid docking score (3dRPC)
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
-4.5
Codon
R5
S6
T4
I1
I2
N2
K1
P3
G3
L2
P4
R4
X2
W
C2
D2
S3
D1
S2
X1
V4
T2
L1
R1
Q1
V3
X3
L5
A2
E1
A3
F2
S5
T1
L4
H2
G2
N1
H1
Y1
M
E2
K2
Q2
Y2
R3
P1
F1
L3
V2
P2
C1
L6
S1
S4
A1
I3
T3
G1
R2
G4
R6
V1
A4

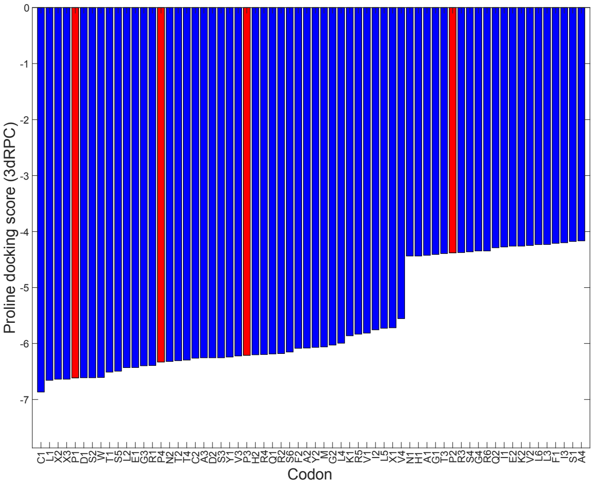
`

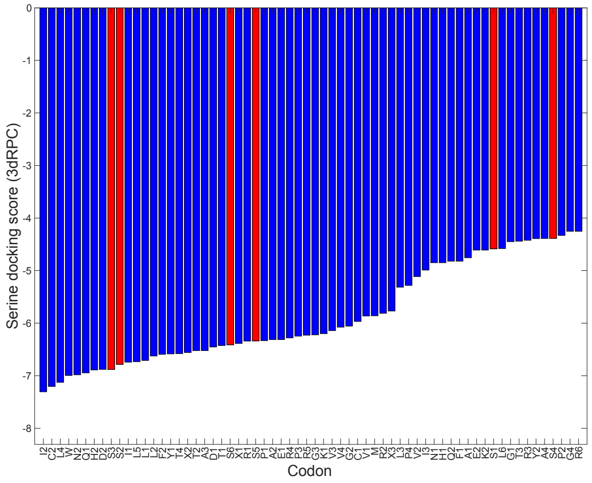

Glutamine docking score (3dRPC)
0
-1
-2
-3
-4
-5
-6
-7
-8
P3 P4 R1 A2 R2 H2 P1 T2 V1 L2 S5 L5 L4 I1 W L1 S2 C2 C1 I2 R4 S6 D1 V4 G2 D2 S3 T4 X1 Q1 M V3 A3 N2 T1 P2 E1 F2 F1 L3 X2 X3 G3 R5 Y1 K1 G1 V2 L6 T3 Y2 R6 I3 N1 H1 A4 S1 S4 A1 G4 E2 R3 K2 Q2
Codon
Glycine docking score (3dRPC)
0
-1
-2
-3
-4
-5
-6
-7
-8
P3 K1 X1 L5 H2 G3 V3 R5 A3 D2 S3 S2 E1 P1 S5 A2 L4 L2 C2 S6 P4 I2 D1 I1 F2 X2 C1 Y1 R2 L1 T4 T1 V1 N2 T2 Q1 X3 R1 G2 W R4 V4 M Q2 I3 Y2 T3 V2 H1 P2 G1 N1 R3 G4 R6 F1 A1 S4 L3 E2 K2 L6 S1 A4
Codon
Histidine docking score (3dRPC)
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
L2 R4 G2 V4 C2 S5 X3 S6 L1 S2 P4 F2 R1 T4 V1 X1 C1 D2 S3 D1 V3 L5 I2 W G3 K1 Y1 R2 A2 L4 H2 Q1 A3 N2 R5 X2 I1 T2 E1 P1 T1 M P3 G4 G1 L6 F1 L3 E2 R6 S1 S4 N1 H1 A1 I3 K2 R3 V2 A4 Q2 Y2 T3 P2
Codon
Isoleucine docking score (3dRPC)
0
-1
-2
-3
-4
-5
-6
V1 L2 F2 Y1 I1 P1 L1 C1 T2 T1 I2 S5 M L5 L4 R2 Q1 W C2 R4 P3 F1 L3 X2 X1 R1 P4 X3 G2 T4 E1 S6 K1 H2 V2 N2 R6 S3 R5 A1 D1 S2 N1 H1 P2 I3 E2 V3 G1 V4 L6 D2 Y2 A2 A3 K2 T3 G4 R3 G3 S1 Q2 A4 S4
Codon
Leucine docking score (3dRPC)
0
-1
-2
-3
-4
-5
-6
L2 V4 X3 R4 Q1 P4 T4 G2 S6 X2 W I1 D1 V1 T2 L1 Y1 S2 M L5 V3 L4 A3 R1 D2 S3 R2 S5 P3 P1 T1 X1 K1 A2 C1 G3 R5 S4 E2 K2 E1 N2 H2 G4 R6 C2 I2 F2 R3 F1 L3 N1 A4 S1 A1 V2 G1 H1 P2 L6 Y2 I3 Q2 T3
Codon
Lysine docking score (3dRPC)
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
P1 P4 C1 F2 W R2 L1 A2 P3 X1 D1 H2 I1 X3 R4 T4 G2 Q1 L2 T1 L4 G3 Y1 D2 M T2 S5 C2 S6 I2 A3 S3 R1 V1 E1 S2 L5 X2 N2 V4 V3 R5 K1 F1 L3 P2 A1 S1 S4 N1 L6 R6 V2 G1 G4 Y2 A4 T3 H1 I3 K2 E2 R3 Q2
Codon

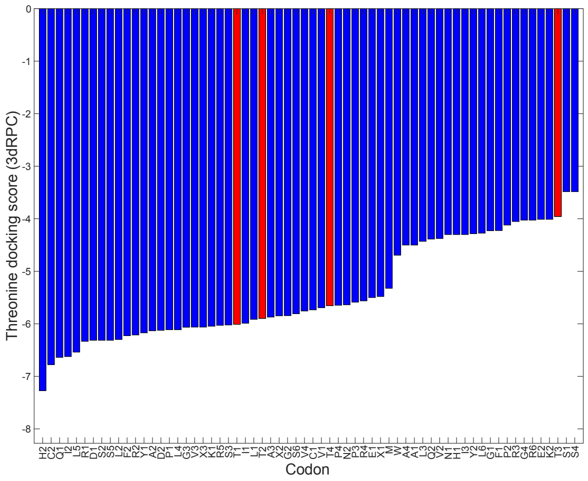
`

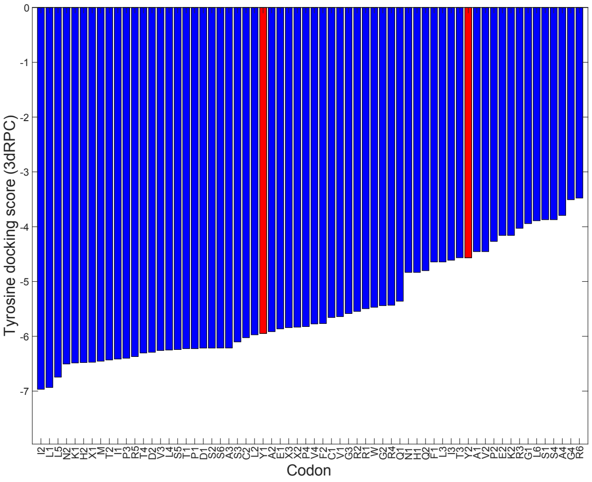

Metheonine docking score (3dRPC)
0
-1
-2
-3
-4
-5
-6
-7
-8
P4 V4 S3 L2 G2 X2 G3 A2 K1 D2 S6 R2 P1 P3 X3 Q1 V3 L5 R1 S2 S5 N2 L4 R5 H2 X1 I2 A3 L1 C2 T2 T1 W F2 T4 I1 M C1 E1 D1 Y1 V1 G4 R6 L6 E2 K2 R3 R4 Y2 A4 S1 S4 P2 T3 Q2 G1 A1 V2 F1 L3 N1 H1 I3
Codon


Phenylalanine docking score (3dRPC)
0
-1
-2
-3
-4
-5
-6
-7
E1 N2 D1 S5 T1 H2 Y1 I1 S6 T4 P1 T2 S2 K1 A3 D2 S3 I2 L5 G3 X2 A2 R1 P4 R2 G2 R5 C1 P3 V1 M C2 L1 X1 W L4 Q1 V4 X3 L2 V3 F2 N1 H1 Q2 Y2 T3 E2 K2 R3 R4 I3 A1 G1 P2 S4 G4 R6 L3 F1 V2 A4 S1 L6
Codon


Proline docking score (3dRPC)
0
-1
-2
-3
-4
-5
-6
-7
C1 L1 X2 X3 P1 D1 S2 W T1 S5 L2 E1 G3 R1 P4 N2 T2 T4 C2 A3 D2 S3 Y1 V3 P3 H2 R4 Q1 R2 S6 F2 A2 Y2 M G2 L4 K1 R5 V1 I2 L5 X1 V4 N1 H1 A1 G1 T3 P2 R3 S4 G4 R6 Q2 I1 E2 K2 V2 L6 L3 F1 I3 S1 A4
Codon


Serine docking score (3dRPC)
0
-1
-2
-3
-4
-5
-6
-7
-8
I2 C2 L4 W N2 Q1 H2 D2 S3 S2 I1 L5 L1 L2 F2 Y1 T4 X2 T2 A3 D1 T1 S6 X1 R1 S5 P1 A2 E1 R4 P3 R5 G3 K1 V3 V4 G2 C1 V1 M R2 X3 L3 P4 V2 I3 N1 H1 Q2 F1 A1 E2 K2 S1 L6 G1 T3 R3 Y2 A4 S4 P2 G4 R6
Codon


Threonine docking score (3dRPC)
0
-1
-2
-3
-4
-5
-6
-7
-8
H2 C2 Q1 I2 L5 R1 D1 S2 S5 L2 F2 R2 Y1 A2 D2 P1 L4 G3 V3 X3 K1 R5 S3 T1 I1 L1 T2 A3 X2 G2 S6 V4 C1 V1 T4 P4 N2 P3 R4 E1 X1 M W A4 A1 L3 Q2 V2 N1 H1 I3 Y2 L6 G1 F1 P2 R3 G4 R6 E2 K2 T3 S1 S4
Codon


Tyrosine docking score (3dRPC)
0
-1
-2
-3
-4
-5
-6
-7
I2 L1 L5 N2 K1 H2 X1 M T2 I1 P3 R5 T4 D2 V3 L4 S5 T1 P1 D1 S2 S6 A3 S3 C2 L2 Y1 A2 E1 X3 X2 P4 V4 F2 C1 V1 G3 R2 R1 W G2 R4 Q1 N1 H1 Q2 F1 L3 I3 T3 Y2 A1 V2 P2 E2 K2 R3 G1 L6 S1 S4 A4 G4 R6
Codon

`

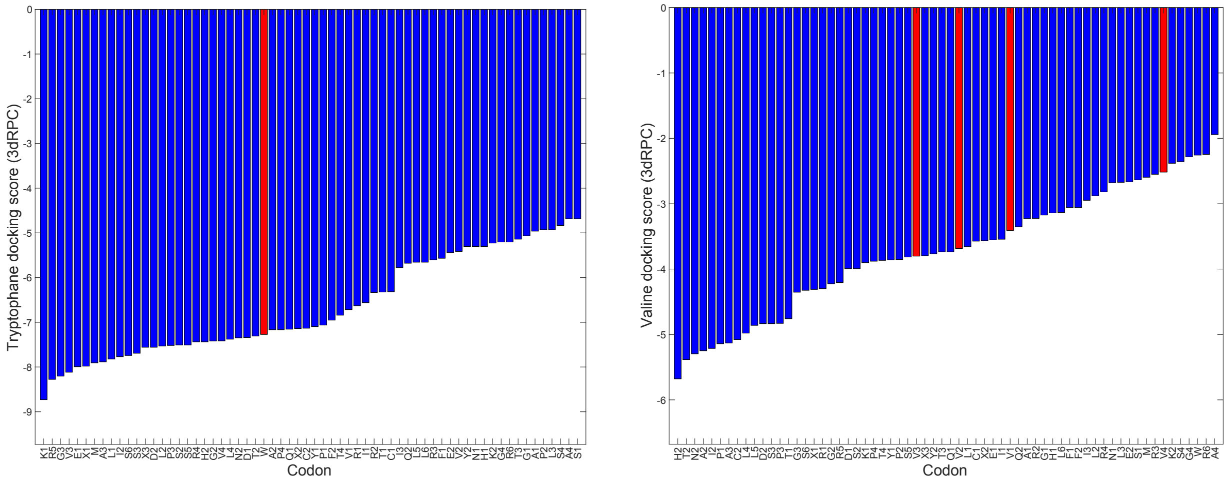


Figure B.2 Amino acid-codon docking results using the 3dRPC docking method. Red bars represent the corresponding codon for each amino acid.

Docking comparison: In order to confirm the docking results using the 3dRPC method, we used another software called HDOCK. This software is used for protein-protein or protein-RNA/DNA docking. The HDOCK has been designed based on a hybrid algorithm of *ab initio* docking and template-based modeling. This program is an FFT-based program developed for molecular docking calculations [295–297]. The scoring function is based on a modified long-range shape-based function to calculate the best pose binding affinity [295,298]. Using HDOCK, we calculated the docking score of glycine to all 64 possible codons. Figure B.3 illustrates the glycine scores for all 64 codons *via* the 3dRPC and HDOC methods.

`

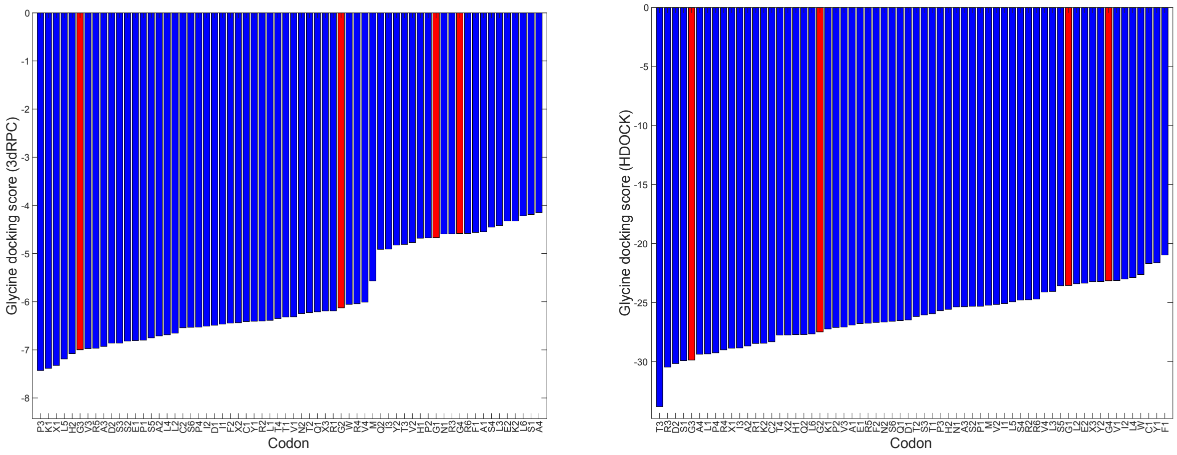
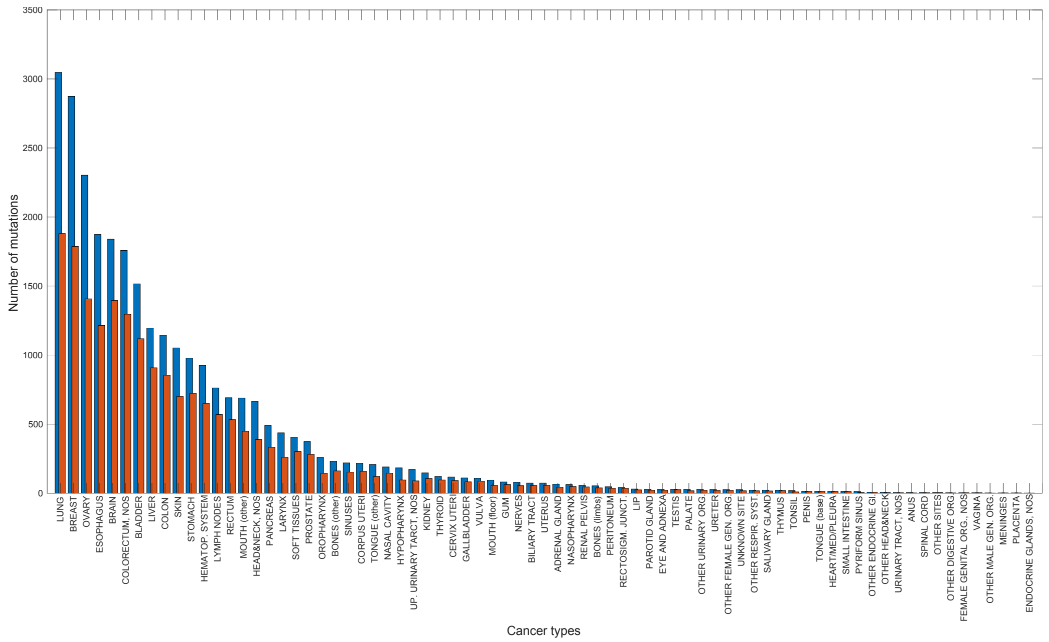


Figure B.3 Docking score comparison between 3dRPC and HDOCK method for glycine docked to all 64 codons. Red bars represents the 3dRPC score for the corresponding codons for each amino acid.

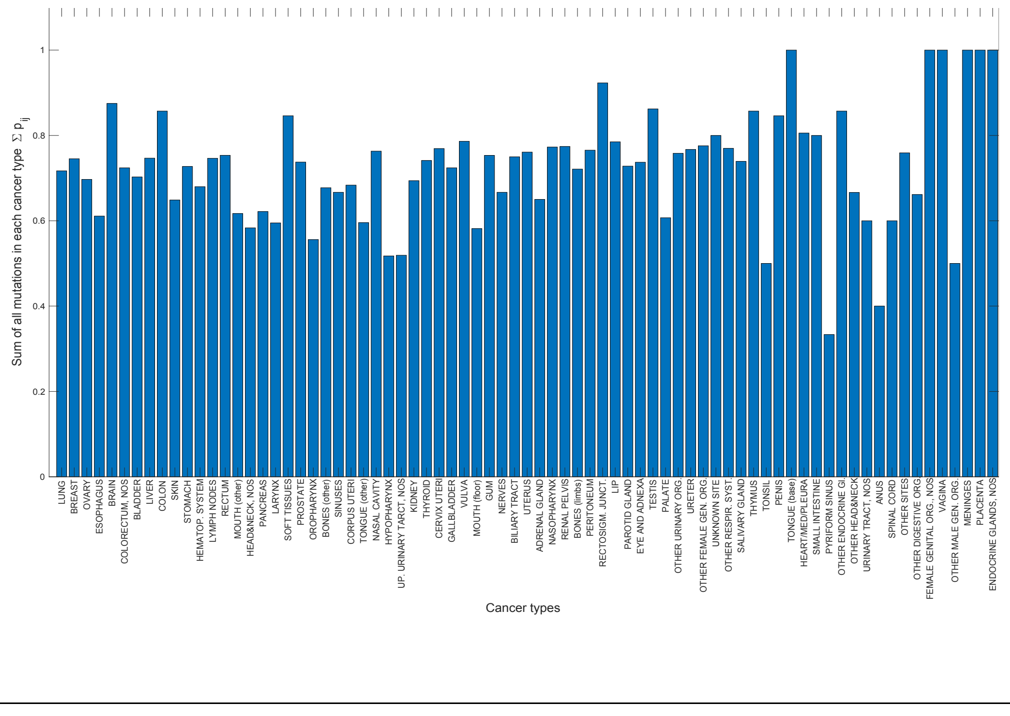
`

Appendix C [8]

(a)



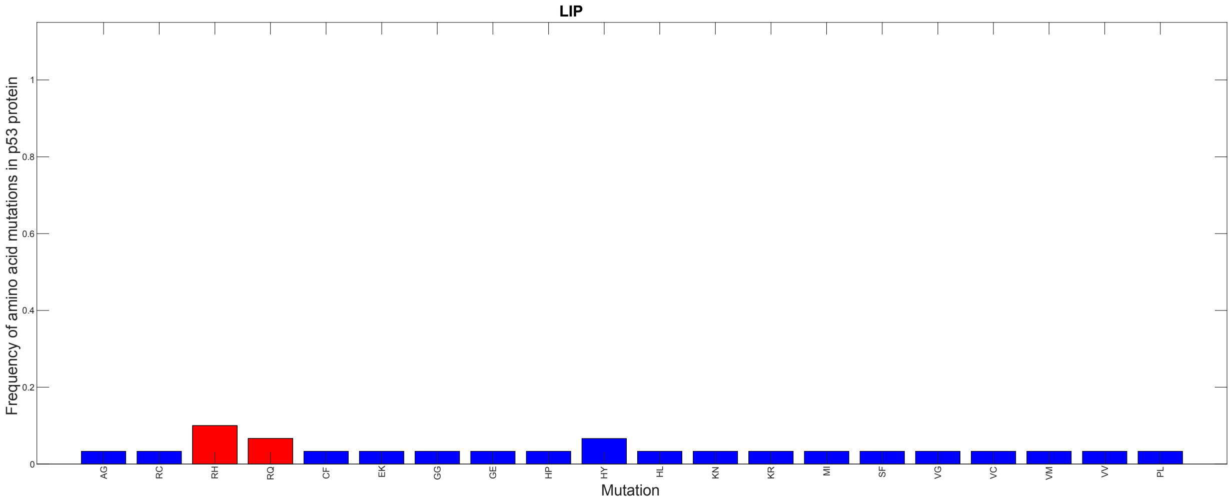
(b)

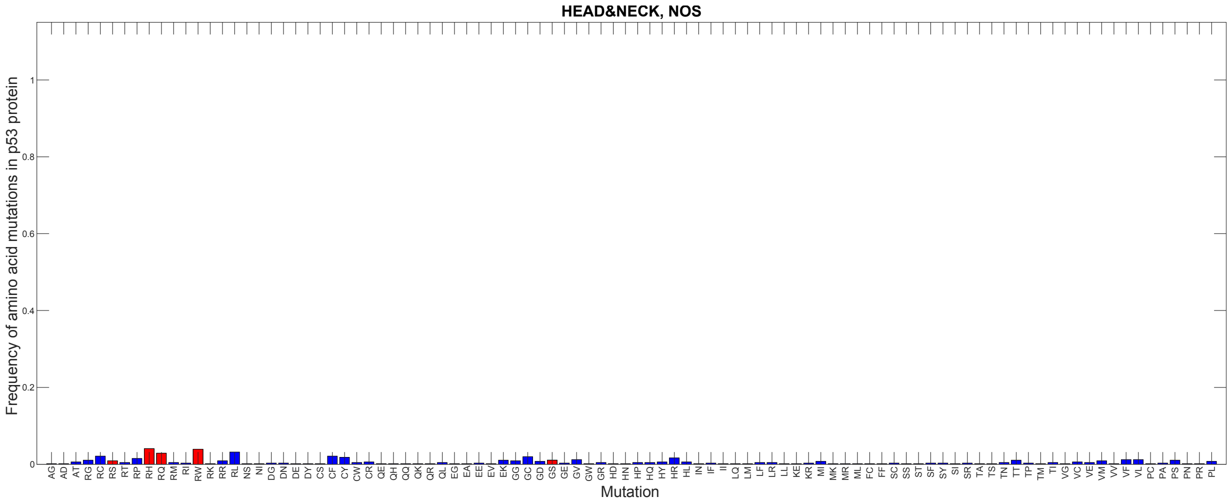


[8] This section has been submitted partially as supplementary material of Arbabi Moghadam S., Omar S. I., and Tuszynski J.A., “Probability Distributions of p53 Mutations and their Corresponding Shannon Entropies in Different Cancer Cell Types” in Theoretical Biology and Medical Modeling journal.

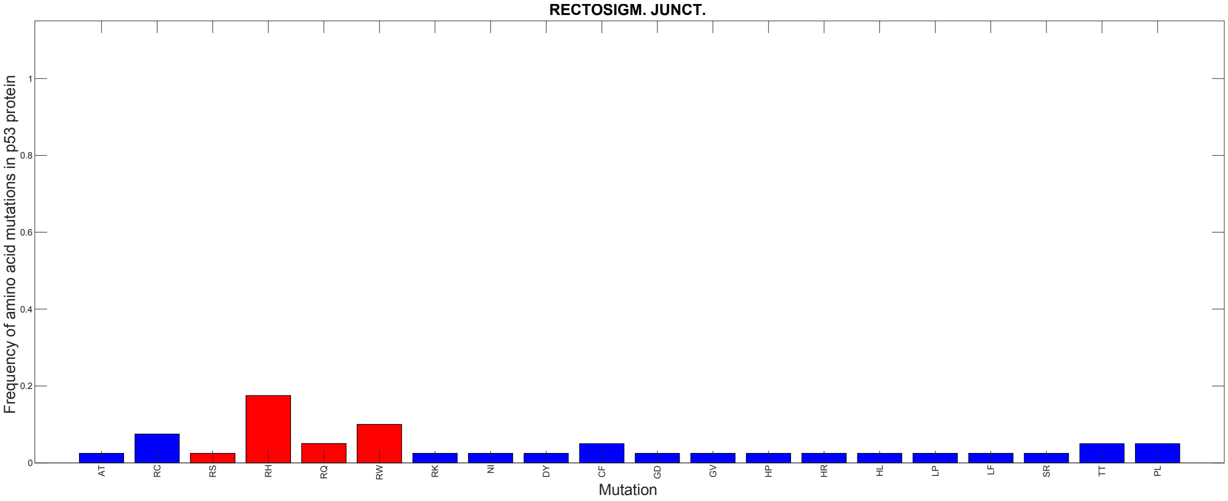
`

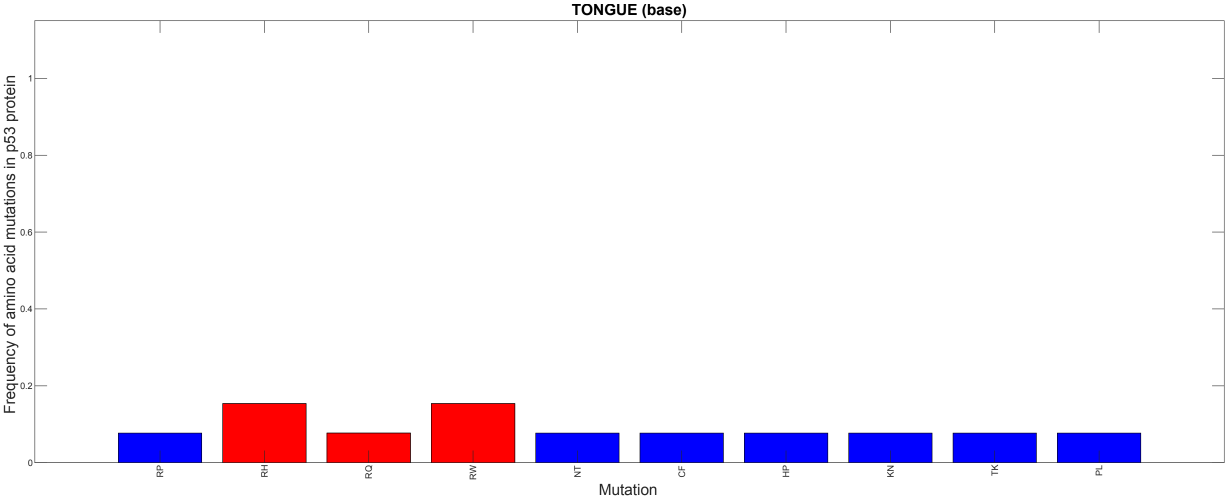
Figure C.1 (a) Numbers of mutations for each cancer type shown in decreasing order. **(b)** Sum over all the amino acid occurrence frequencies in each cancer type in Eq. (1.2) in the same order of appearing in plot **(a)**.

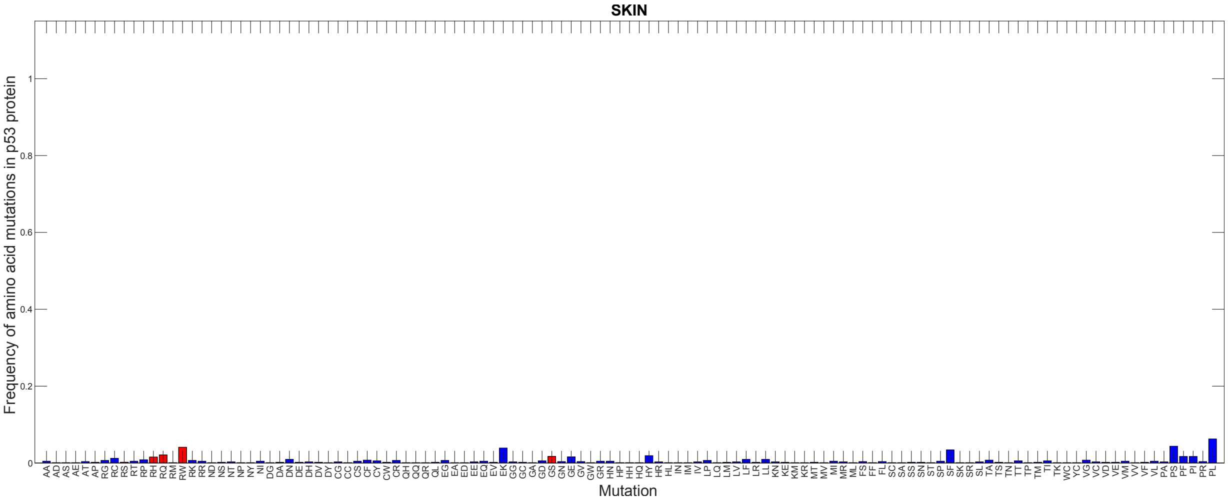




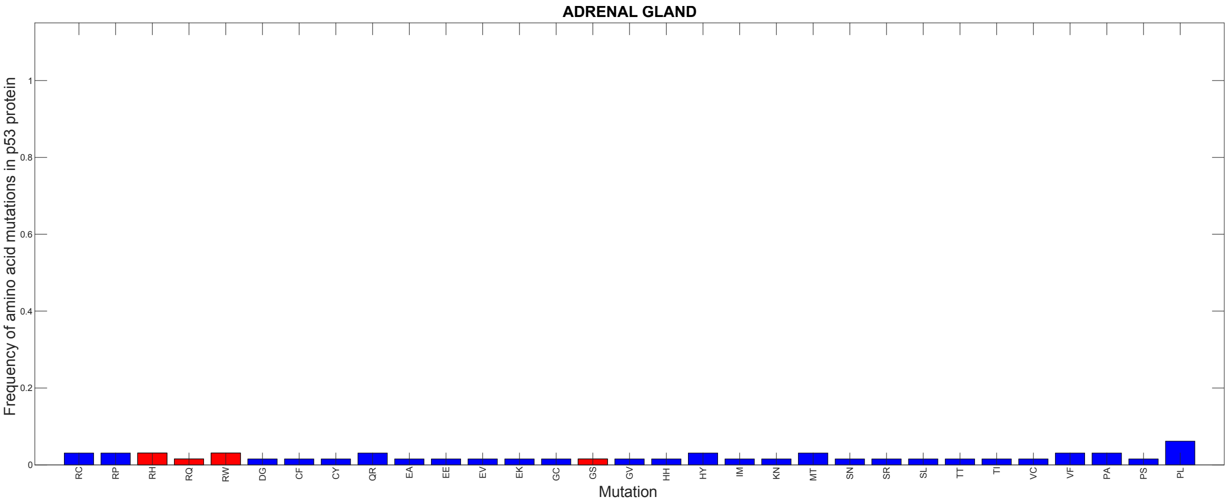
`

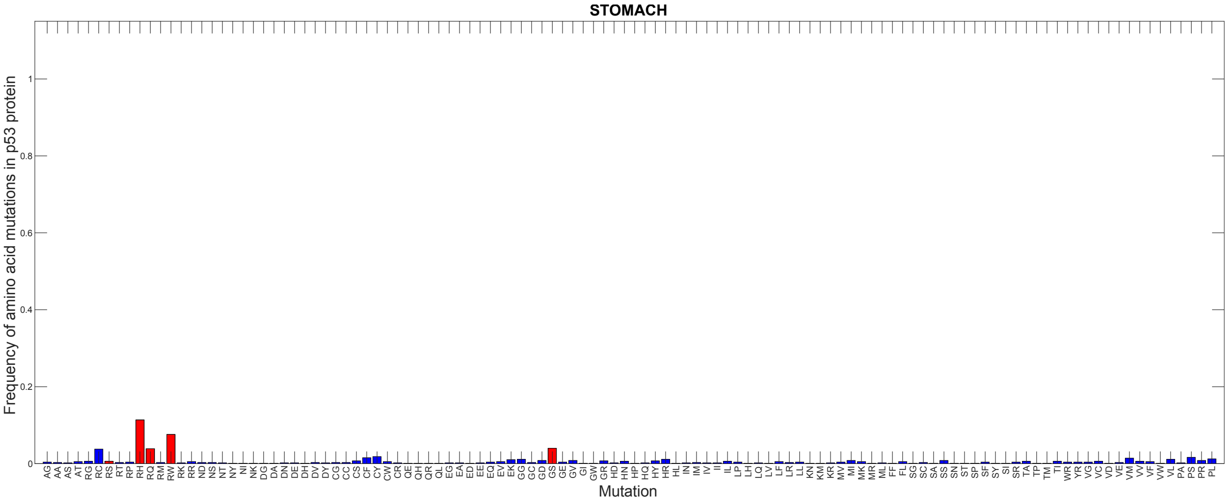

RECTOSIGM. JUNCT.
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AT
RC
RS
RH
RQ
RW
RK
NI
DY
CF
GD
GV
HP
HR
HL
LP
LF
SR
TT
PL
Mutation

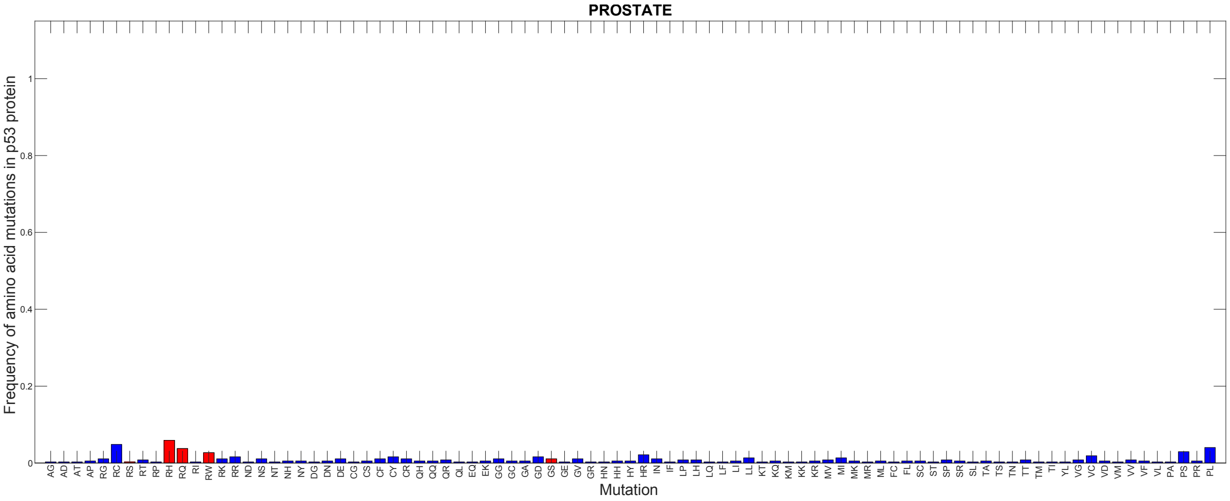

TONGUE (base)
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
RP
RH
RQ
RW
NT
CF
HP
KN
TK
PL
Mutation

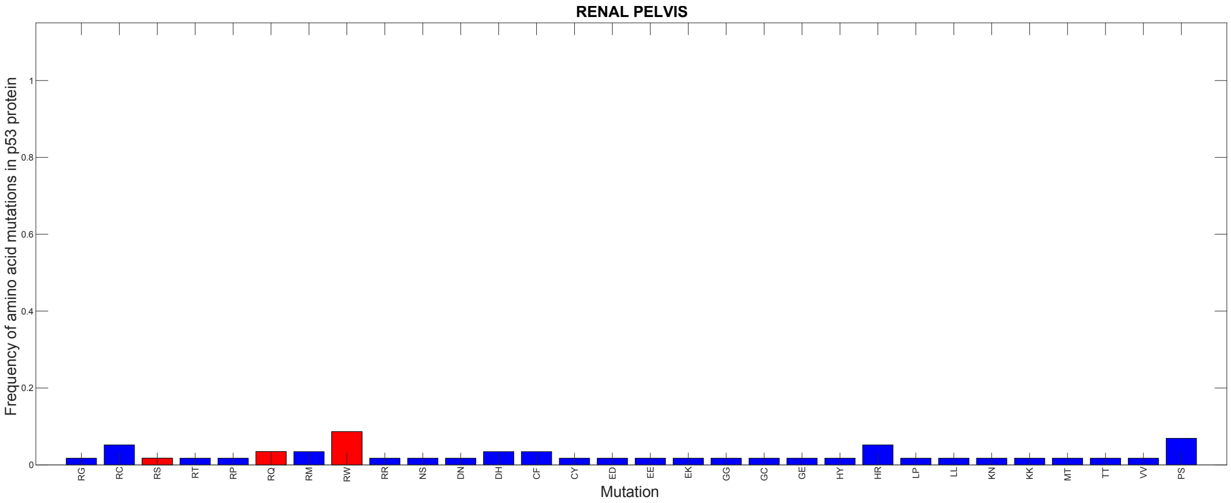

SKIN
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AA
AD
AS
AE
AT
AP
RG
RC
RS
RT
RP
RH
RQ
RM
RW
RK
RR
ND
NS
NT
NP
NY
NI
DG
DA
DN
DE
DH
DV
DY
CG
CC
CS
CF
CY
CW
CR
QH
QQ
QR
QL
EG
EA
ED
EE
EQ
EV
EK
GG
GC
GA
GD
GS
GN
GE
GV
GW
GR
HN
HP
HH
HQ
HY
HR
HL
IN
IM
IV
LP
LQ
LM
LV
LF
LR
LL
KN
KE
KM
KR
MT
MV
MI
MR
ML
FS
FF
FL
SC
SA
SS
SN
ST
SP
SF
SK
SR
SL
TA
TS
TN
TT
TP
TM
TI
TK
WC
YC
VG
VC
VD
VE
VM
VV
VF
VL
PA
PS
PF
PI
PR
PL
Mutation

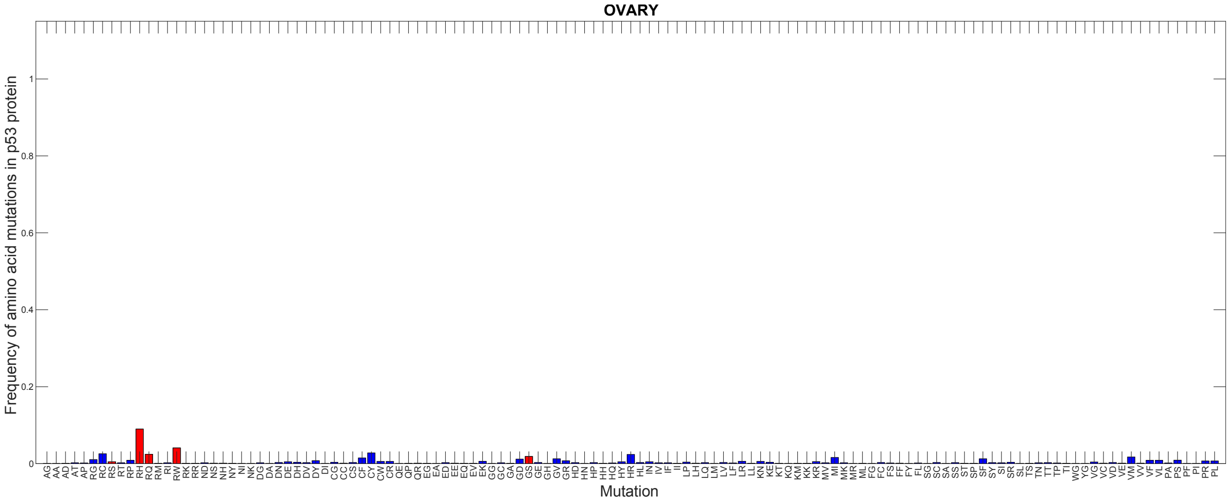
`

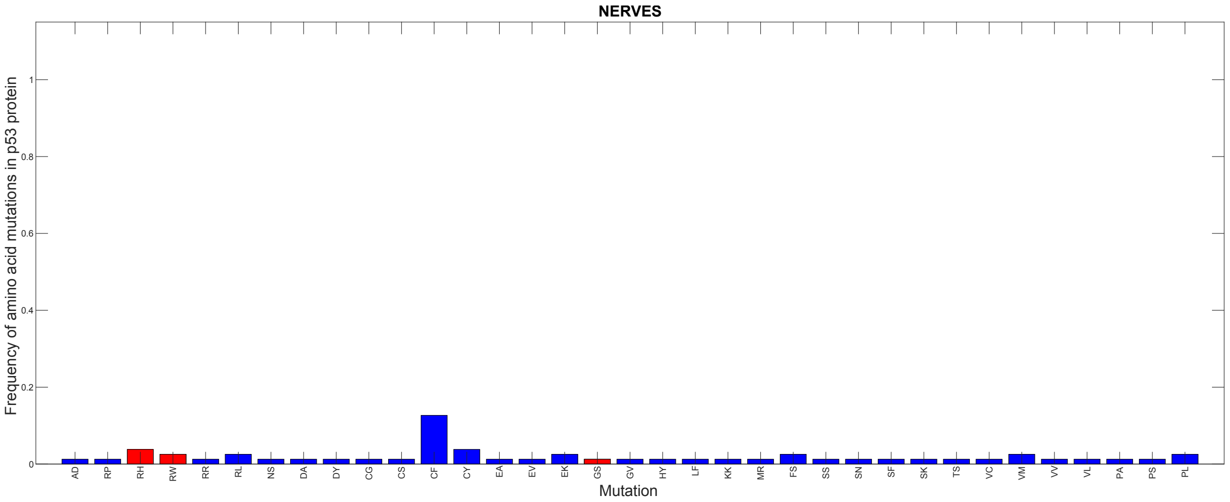

ADRENAL GLAND
Frequency of amino acid mutations in p53 protein
1
0.8
0.6
0.4
0.2
0
RC
RP
RH
RQ
RW
DG
CF
CY
QR
EA
EE
EV
EK
GC
GS
GV
HH
HY
IM
KN
MT
SN
SR
SL
TT
TI
VC
VF
PA
PS
PL
Mutation

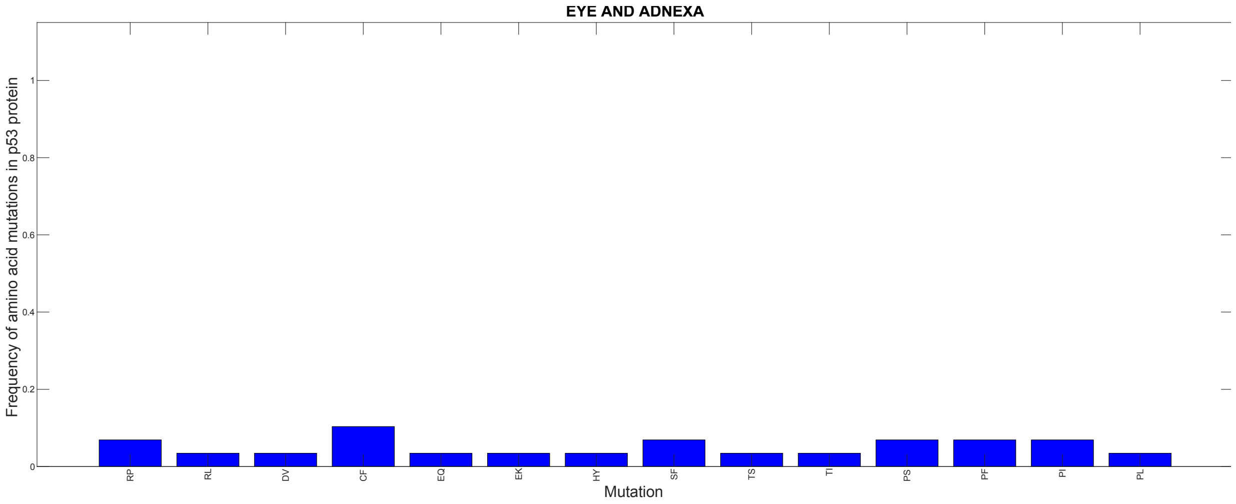

STOMACH
Frequency of amino acid mutations in p53 protein
1
0.8
0.6
0.4
0.2
0
AG
AA
AS
AT
RG
RC
RS
RT
RP
RH
RQ
RM
RW
RK
RR
ND
NS
NT
NY
NI
NK
DG
DA
DN
DE
DH
DV
DY
CG
CC
CS
CF
CY
CW
CR
QE
QH
QR
QL
EG
EA
ED
EE
EQ
EV
EK
GG
GC
GD
GS
GE
GV
GI
GW
GR
HD
HN
HP
HQ
HY
HR
HL
IN
IM
IV
II
IL
LP
LH
LQ
LV
LF
LR
LL
KN
KM
KR
MV
MI
MK
MR
ML
FF
FL
SG
SC
SA
SS
SN
ST
SP
SF
SY
SI
SR
TA
TP
TM
TI
WR
YR
VG
VC
VD
VE
VM
VV
VF
VW
VL
PA
PS
PR
PL
Mutation

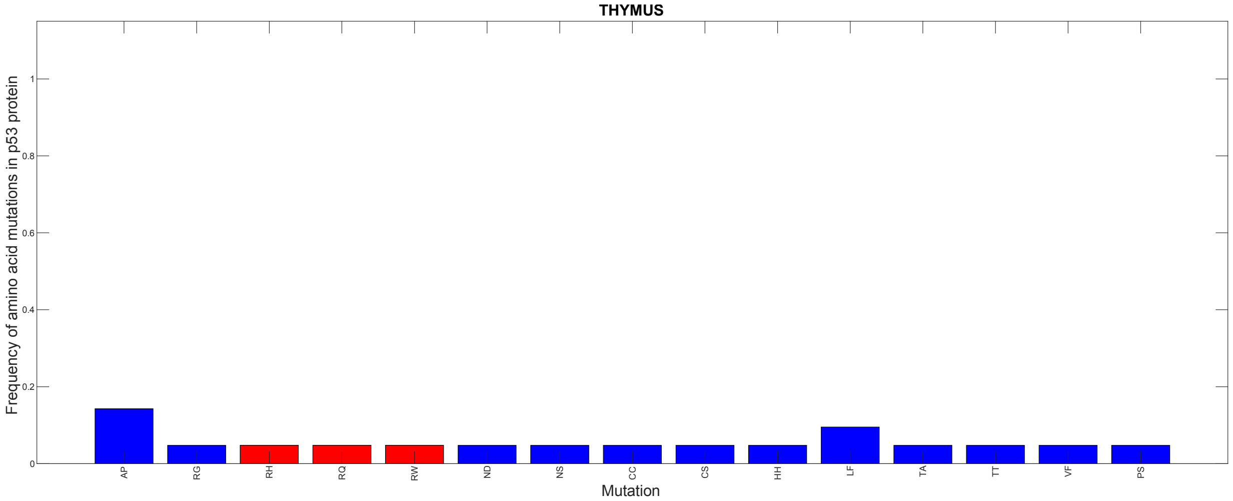

PROSTATE
Frequency of amino acid mutations in p53 protein
1
0.8
0.6
0.4
0.2
0
AG
AD
AT
AP
RG
RC
RS
RT
RP
RH
RQ
RI
RW
RK
RR
ND
NS
NT
NH
NY
DG
DN
DE
CG
CS
CF
CY
CR
QH
QQ
QR
QL
EQ
EK
GG
GC
GA
GD
GS
GE
GV
GR
HN
HH
HY
HR
IN
IF
LP
LH
LQ
LF
LI
LL
KT
KQ
KM
KK
KR
MV
MI
MK
MR
ML
FC
FL
SC
ST
SP
SR
SL
TA
TS
TN
TT
TM
TI
YL
VG
VC
VD
VM
VV
VF
VL
PA
PS
PR
PL
Mutation

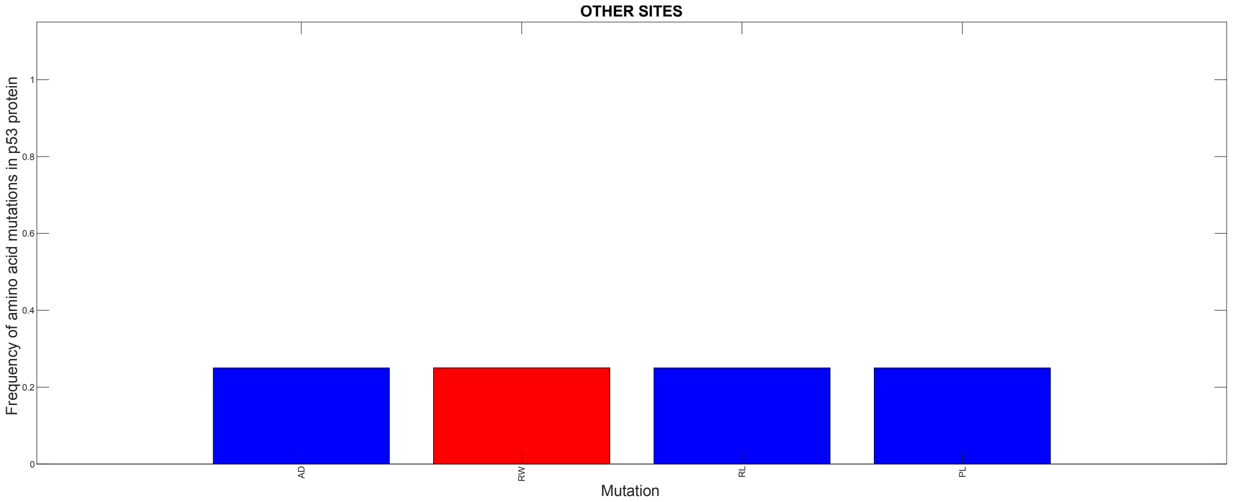

RENAL PELVIS
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
RG
RC
RS
RT
RP
RQ
RM
RW
RR
NS
DN
DH
CF
CY
ED
EE
EK
GG
GC
GE
HY
HR
LP
LL
KN
KK
MT
TT
VV
PS
Mutation

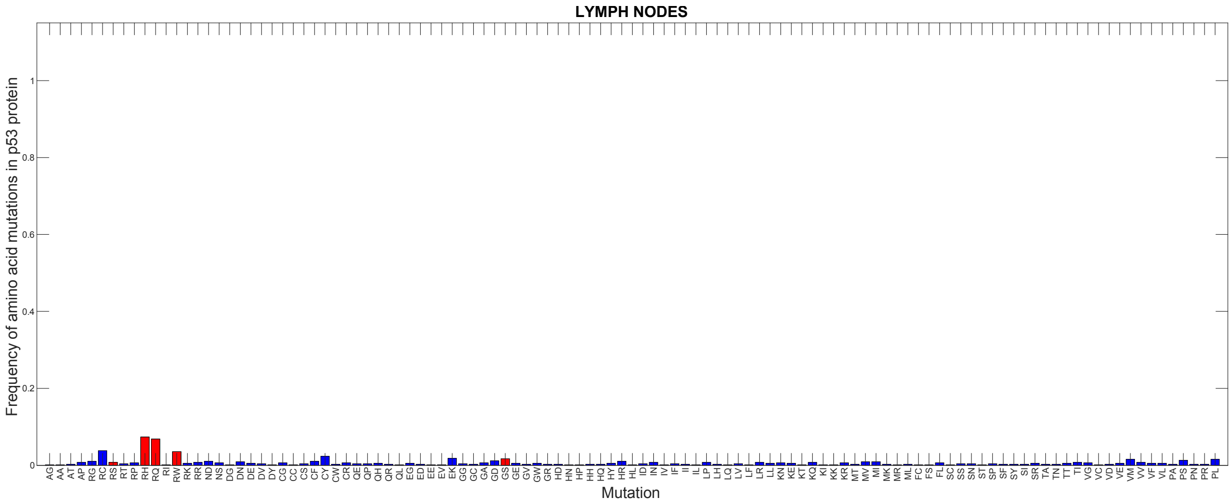

OVARY
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AG
AA
AD
AT
AP
RG
RC
RS
RT
RP
RH
RQ
RM
RI
RW
RK
RR
ND
NS
NH
NY
NI
NK
DG
DA
DN
DE
DH
DV
DY
DI
CG
CC
CS
CF
CY
CW
CR
QE
QP
QR
EG
EA
ED
EE
EQ
EV
EK
GG
GC
GA
GD
GS
GE
GH
GV
GR
HD
HN
HP
HH
HQ
HY
HR
HL
IN
IV
IF
II
LP
LH
LQ
LM
LV
LF
LR
LL
KN
KE
KT
KQ
KM
KK
KR
MV
MI
MK
MR
ML
FG
FC
FS
FF
FY
FL
SG
SC
SA
SS
ST
SP
SF
SY
SI
SR
SL
TS
TN
TT
TP
TI
WG
YG
VG
VC
VD
VE
VM
VV
VF
VL
PA
PS
PF
PI
PR
PL
Mutation

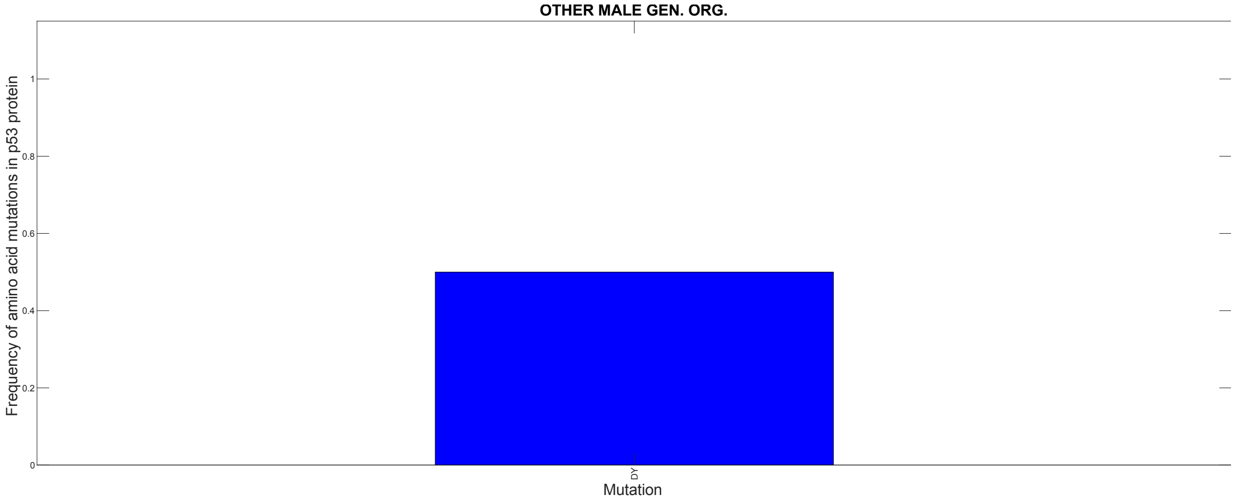

NERVES
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AD
RP
RH
RW
RR
RL
NS
DA
DY
CG
CS
CF
CY
EA
EV
EK
GS
GV
HY
LF
KK
MR
FS
SS
SN
SF
SK
TS
VC
VM
VV
VL
PA
PS
PL
Mutation

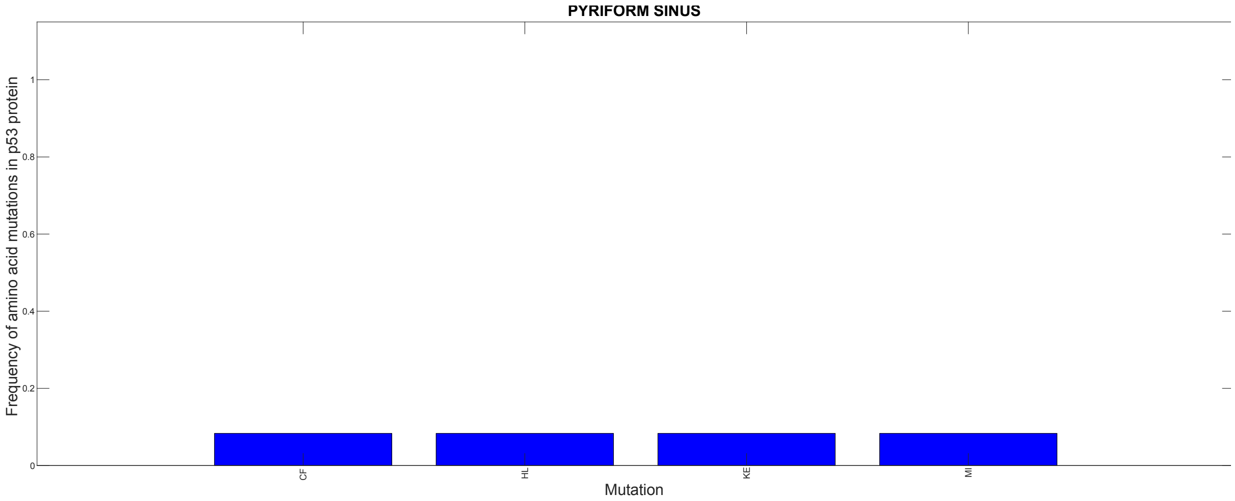

EYE AND ADNEXA
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
RP
RL
DV
CF
EQ
EK
HY
SF
TS
TI
PS
PF
PI
PL
Mutation


THYMUS
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AP
RG
RH
RQ
RW
ND
NS
CC
CS
HH
LF
TA
TT
VF
PS
Mutation

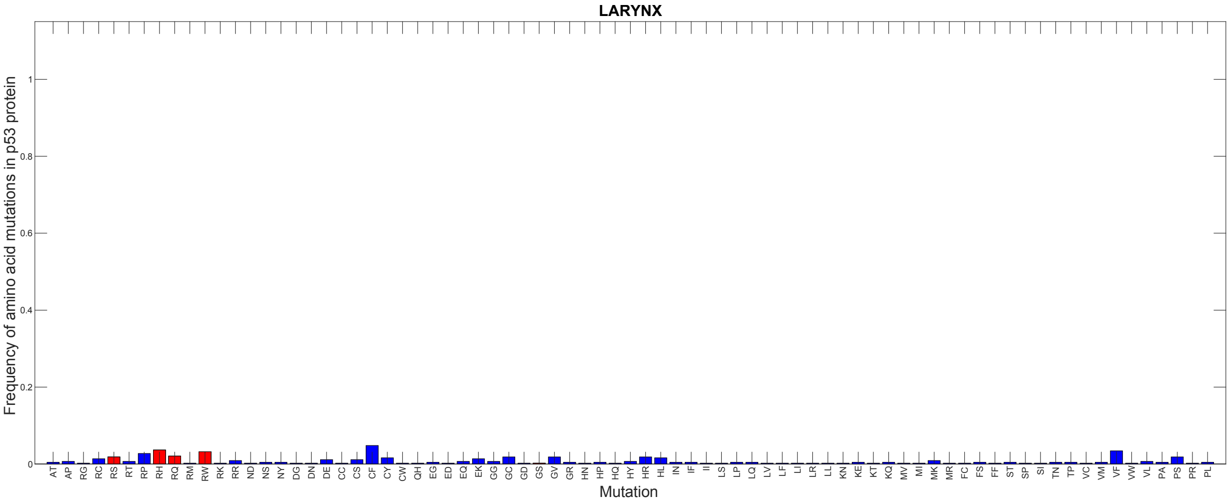

OTHER SITES
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AD
RW
RL
PL
Mutation

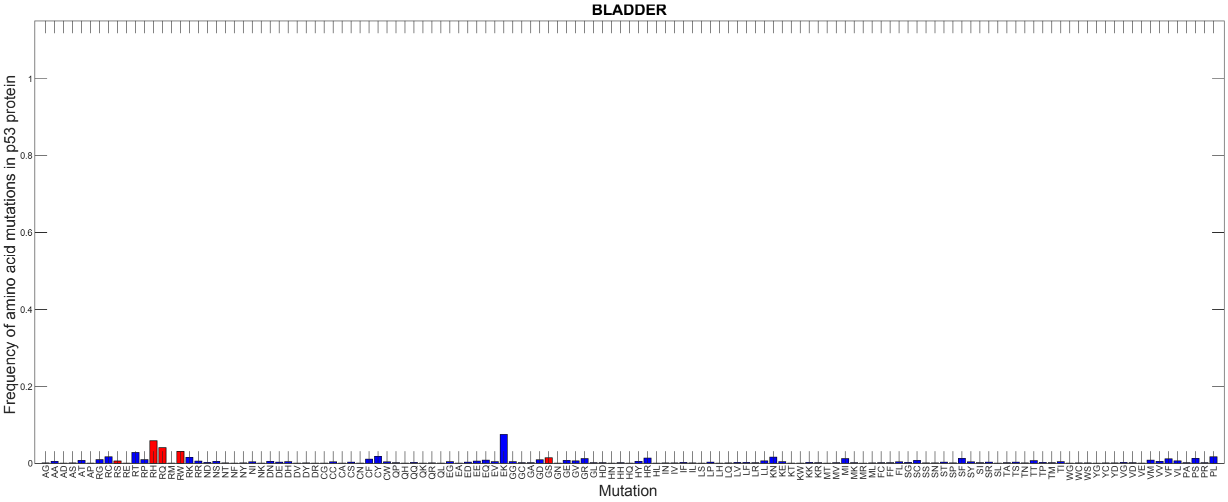

LYMPH NODES
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AG AA AT AP RG RC RS RT RP RH RQ RI RW RK RR ND NS DG DN DE DV DY CG CC CS CF CY CW CR QE QP QH QR QL EG ED EE EV EK GG GC GA GD GS GE GV GW GR HD HN HP HH HQ HY HR HL ID IN IV IF II IL LP LH LQ LV LF LR LL KN KE KT KQ KI KK KR MT MV MI MK MR ML FC FS FL SC SS SN ST SP SF SY SI SR TA TN TT TI VG VC VD VE VM VV VF VL PA PS PN PR PL
Mutation

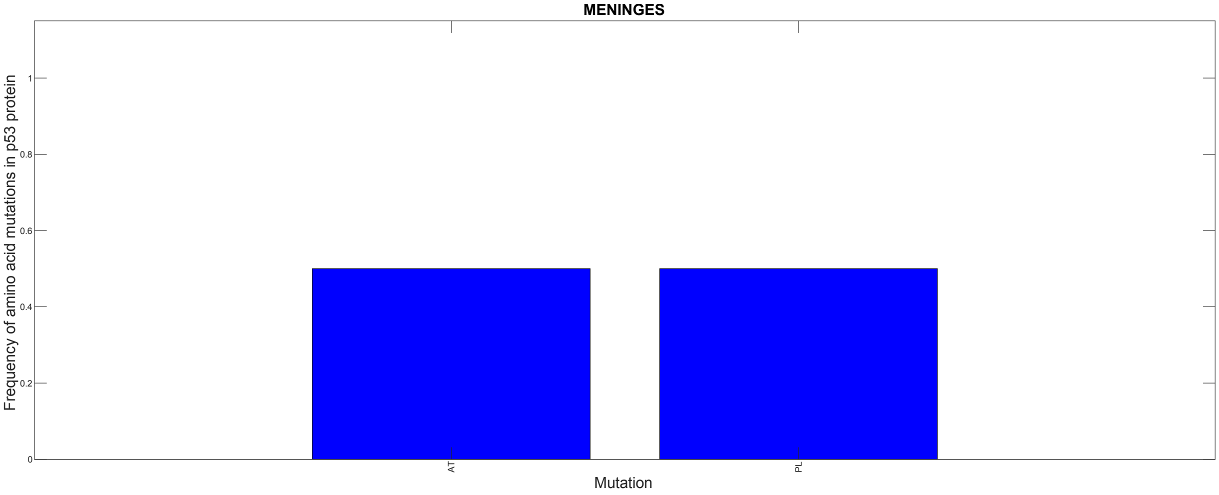

OTHER MALE GEN. ORG.
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
DY
Mutation


PYRIFORM SINUS
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
CF
HL
KE
MI
Mutation

`

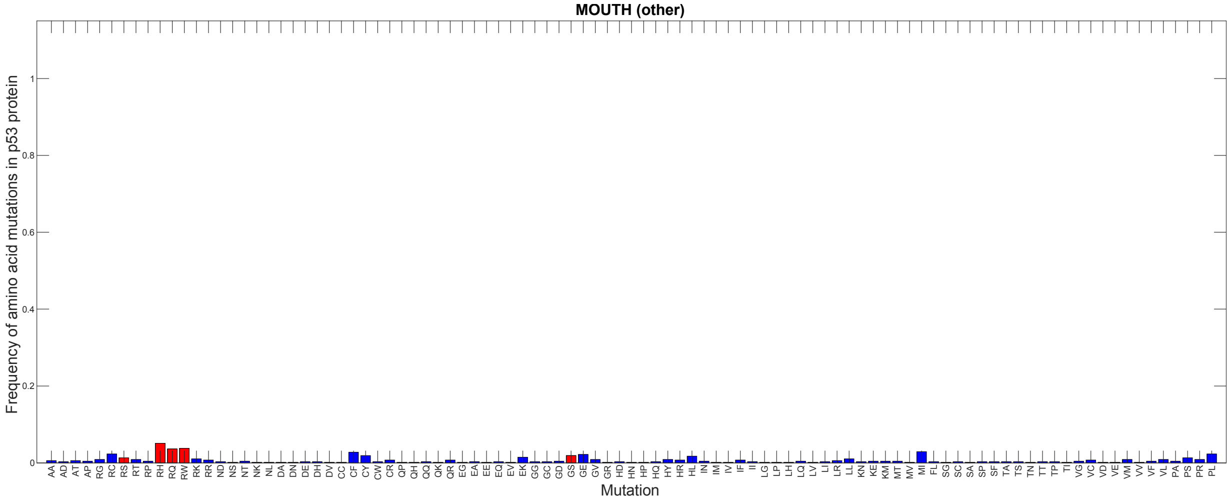

LARYNX
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
Mutation
AT
AP
RG
RC
RS
RT
RP
RH
RQ
RM
RW
RK
RR
ND
NS
NY
DG
DN
DE
CC
CS
CF
CY
CW
QH
EG
ED
EQ
EK
GG
GC
GD
GS
GV
GR
HN
HP
HQ
HY
HR
HL
IN
IF
II
LS
LP
LQ
LV
LF
LI
LR
LL
KN
KE
KT
KQ
MV
MI
MK
MR
FC
FS
FF
ST
SP
SI
TN
TP
VC
VM
VF
VW
VL
PA
PS
PR
PL


BLADDER
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
Mutation
AG
AA
AD
AS
AT
AP
RG
RC
RS
RE
RT
RP
RH
RQ
RM
RW
RK
RR
ND
NS
NT
NF
NY
NI
NK
DN
DE
DH
DV
DY
DR
CG
CC
CA
CS
CN
CF
CY
CW
QP
QH
QQ
QK
QR
QL
EG
EA
ED
EE
EQ
EV
EK
GG
GC
GA
GD
GS
GN
GE
GV
GR
GL
HD
HN
HH
HQ
HY
HR
HL
IN
IV
IF
IL
LS
LP
LH
LQ
LV
LF
LR
LL
KN
KE
KT
KW
KK
KR
MT
MV
MI
MK
MR
ML
FC
FF
FL
SG
SC
SS
SN
ST
SP
SF
SY
SI
SR
SL
TA
TS
TN
TT
TP
TM
TI
WG
WC
WS
YG
YC
YD
VG
VD
VE
VM
VV
VF
VL
PA
PS
PR
PL

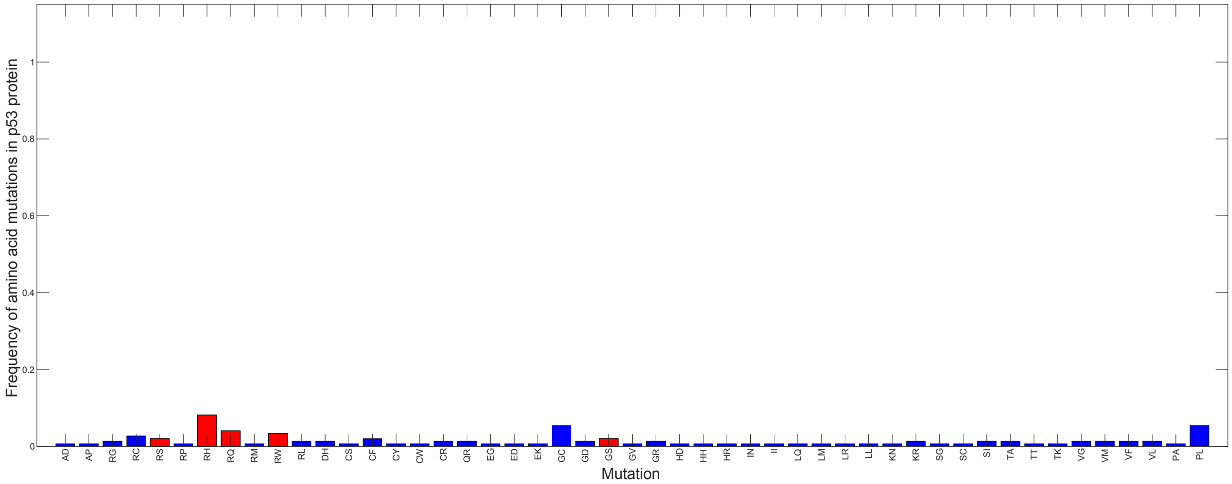

MENINGES
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
Mutation
AT
PL

`

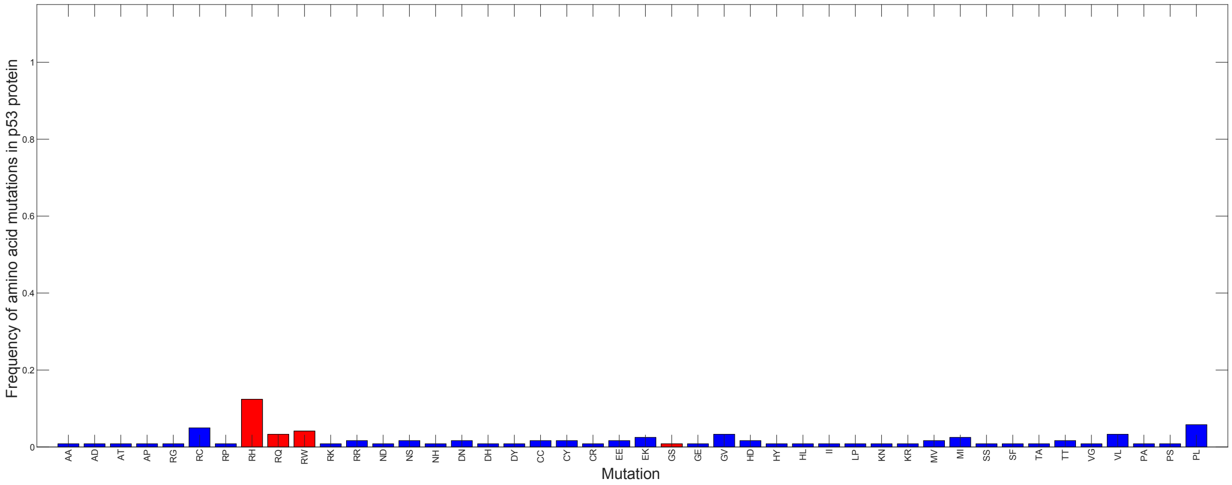
MOUTH (other)



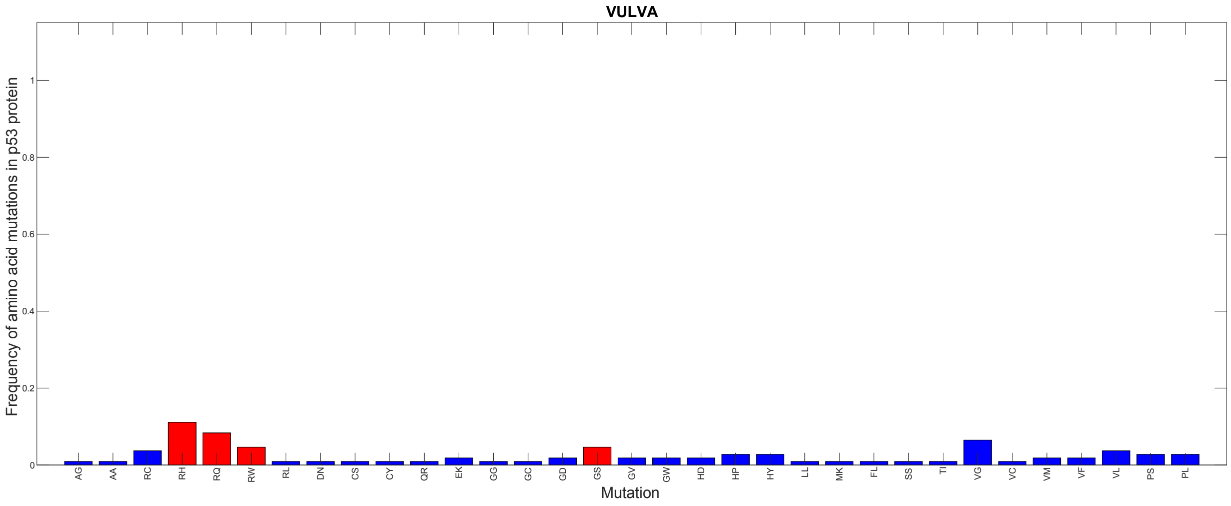
KIDNEY

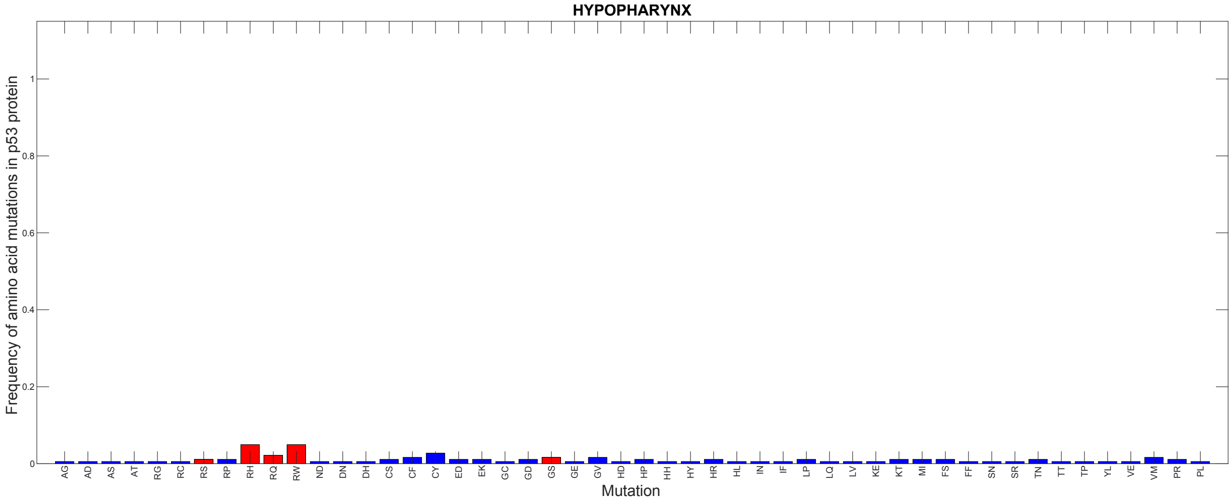


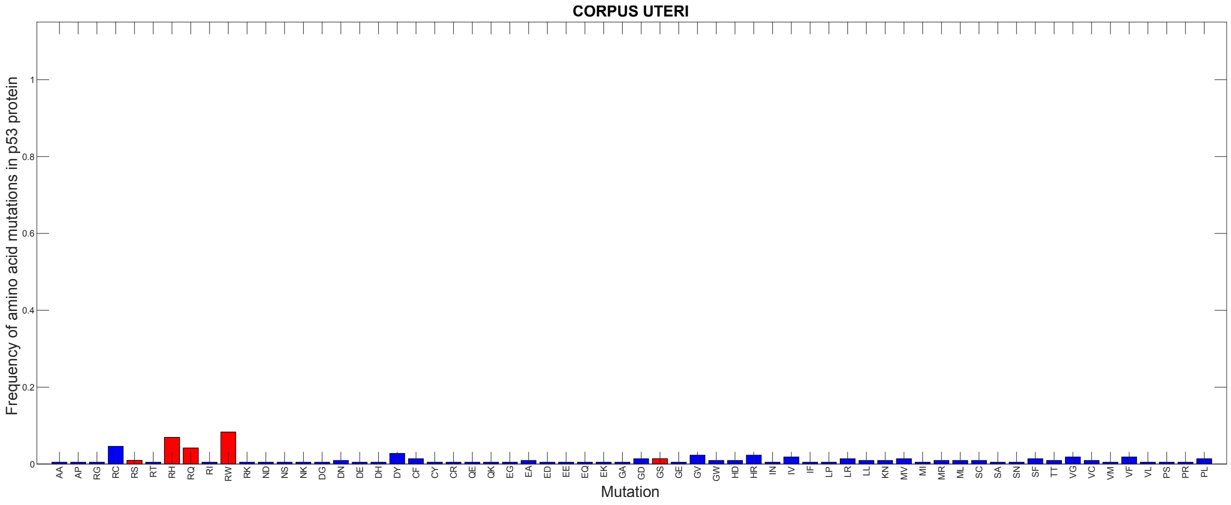
THYROID



`

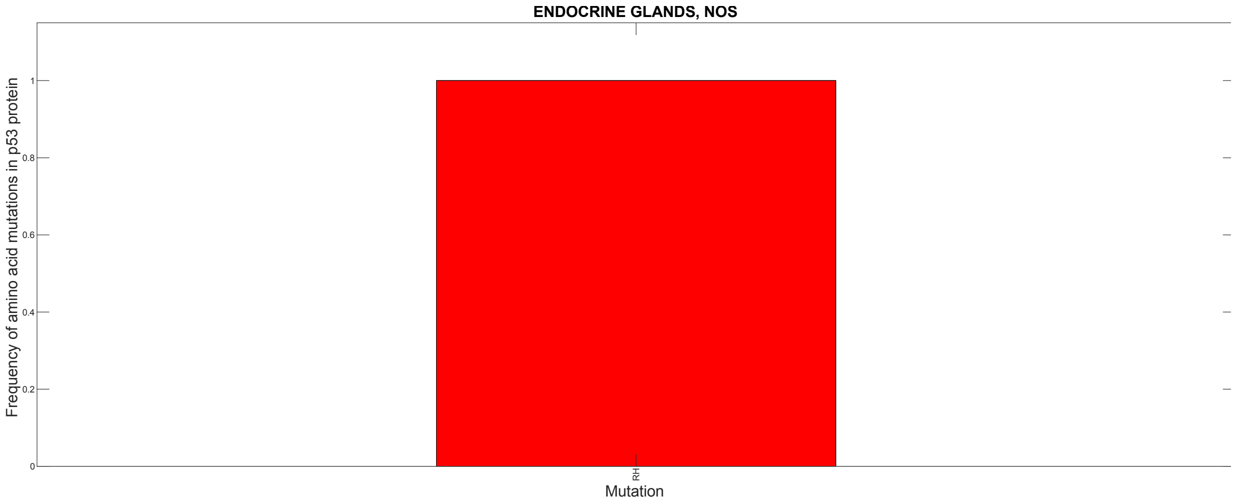

VULVA
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AG
AA
RC
RH
RQ
RW
RL
DN
CS
CY
QR
EK
GG
GC
GD
GS
GV
GW
HD
HP
HY
LL
MK
FL
SS
TI
VG
VC
VM
VF
VL
PS
PL
Mutation


HYPOPHARYNX
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AG
AD
AS
AT
RG
RC
RS
RP
RH
RQ
RW
ND
DN
DH
CS
CF
CY
ED
EK
GC
GD
GS
GE
GV
HD
HP
HH
HY
HR
HL
IN
IF
LP
LQ
LV
KE
KT
MI
FS
FF
SN
SR
TN
TT
TP
YL
VE
VM
PR
PL
Mutation

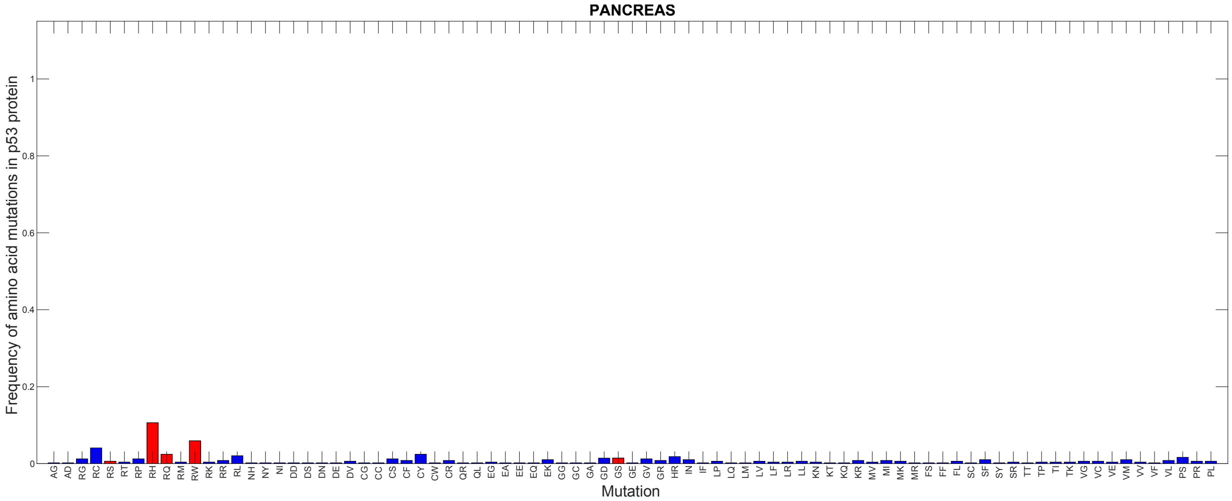

CORPUS UTERI
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AA
AP
RG
RC
RS
RT
RH
RQ
RI
RW
RK
ND
NS
NK
DG
DN
DE
DH
DY
CF
CY
CR
QE
QK
EG
EA
ED
EE
EQ
EK
GA
GD
GS
GE
GV
GW
HD
HR
IN
IV
IF
LP
LR
LL
KN
MV
MI
MR
ML
SC
SA
SN
SF
TT
VG
VC
VM
VF
VL
PS
PR
PL
Mutation

`

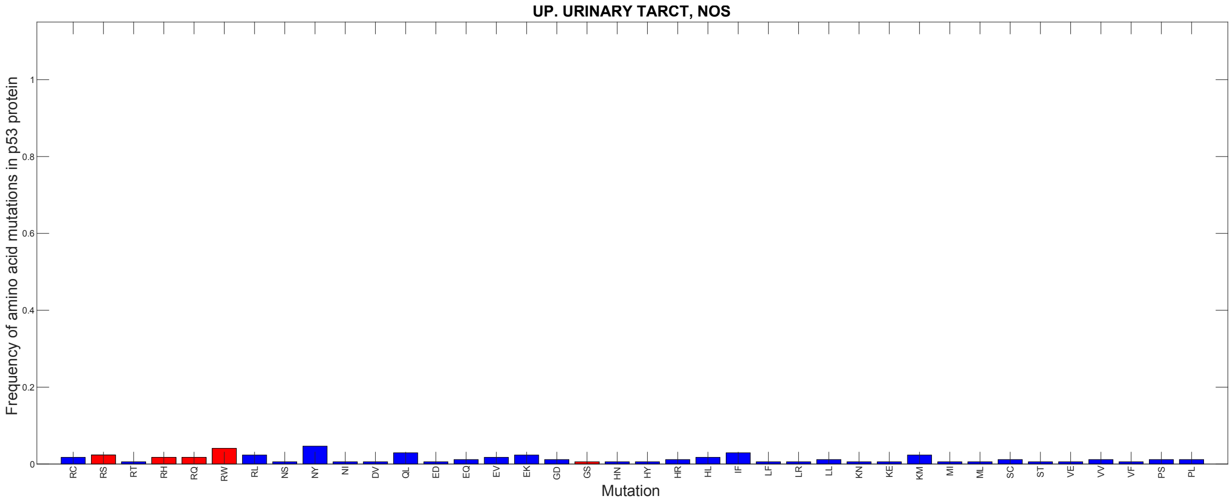
ENDOCRINE GLANDS, NOS



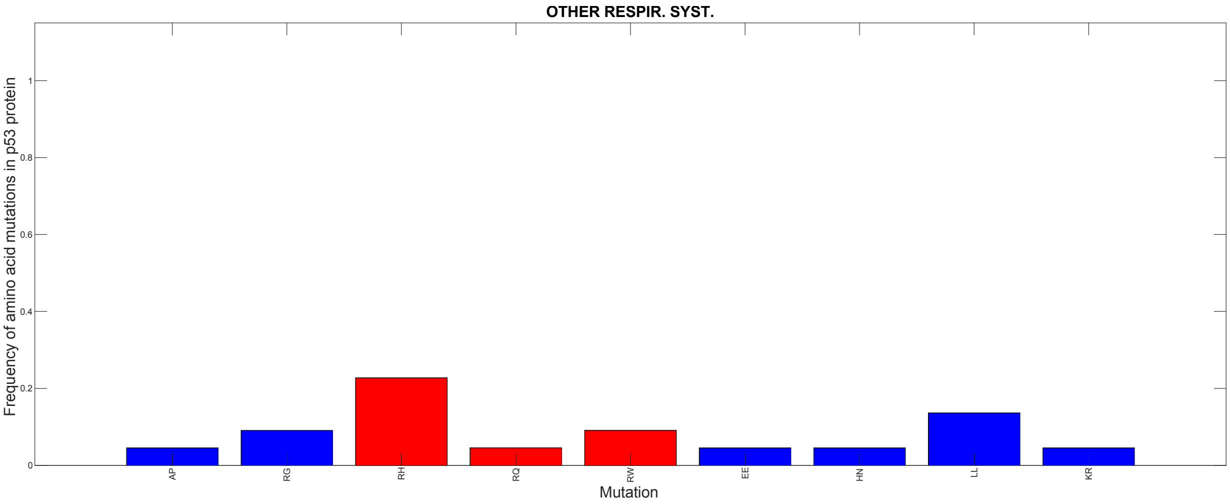
PANCREAS

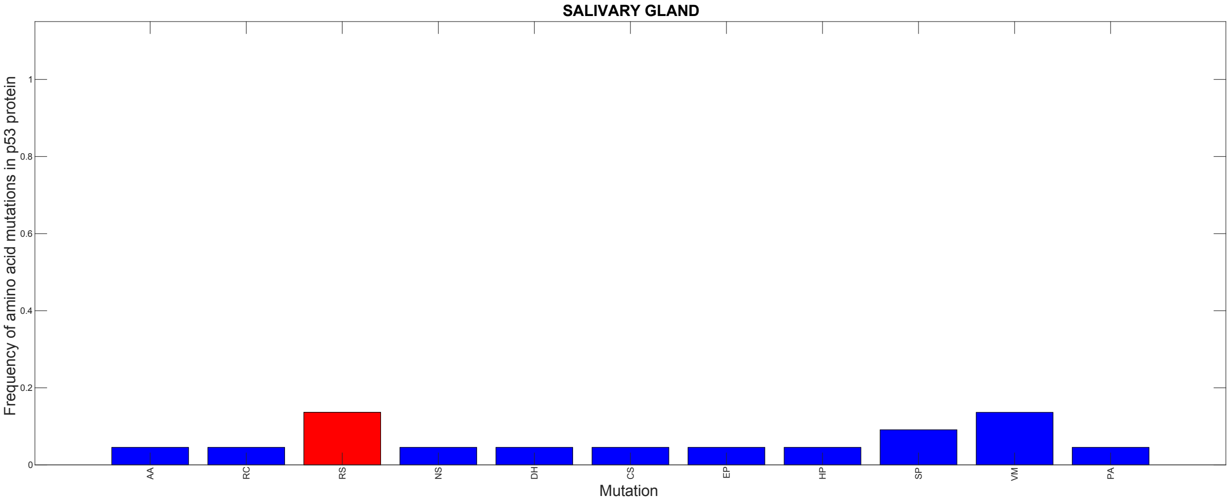


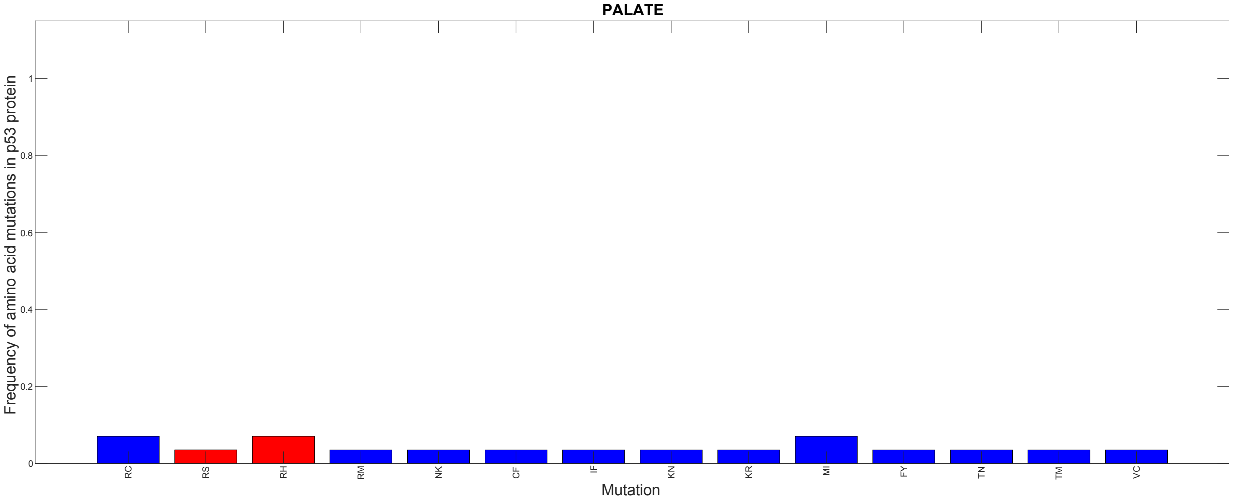
UP. URINARY TARCT, NOS



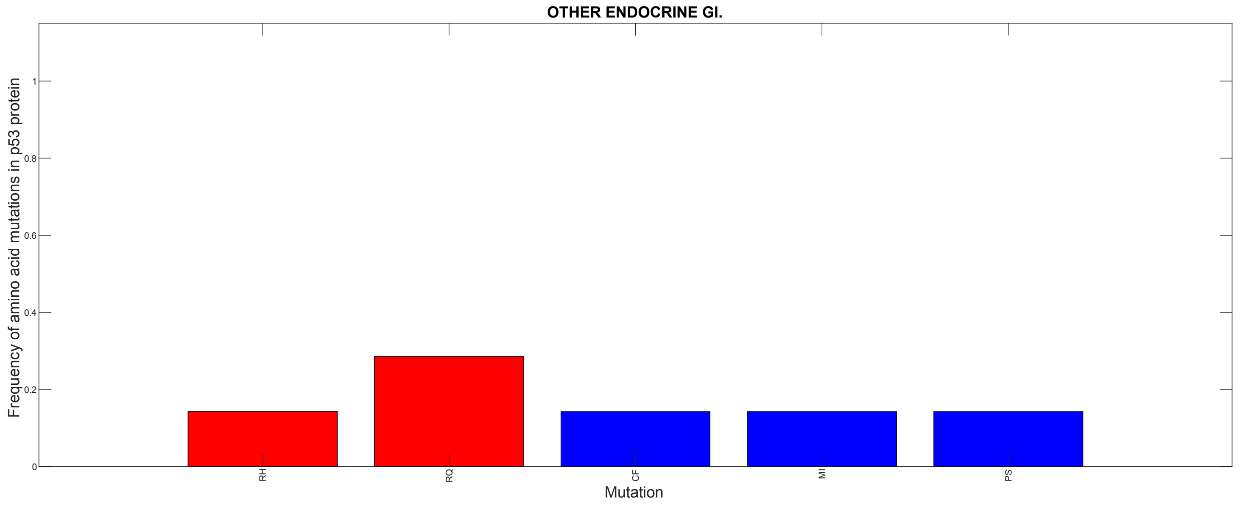
`

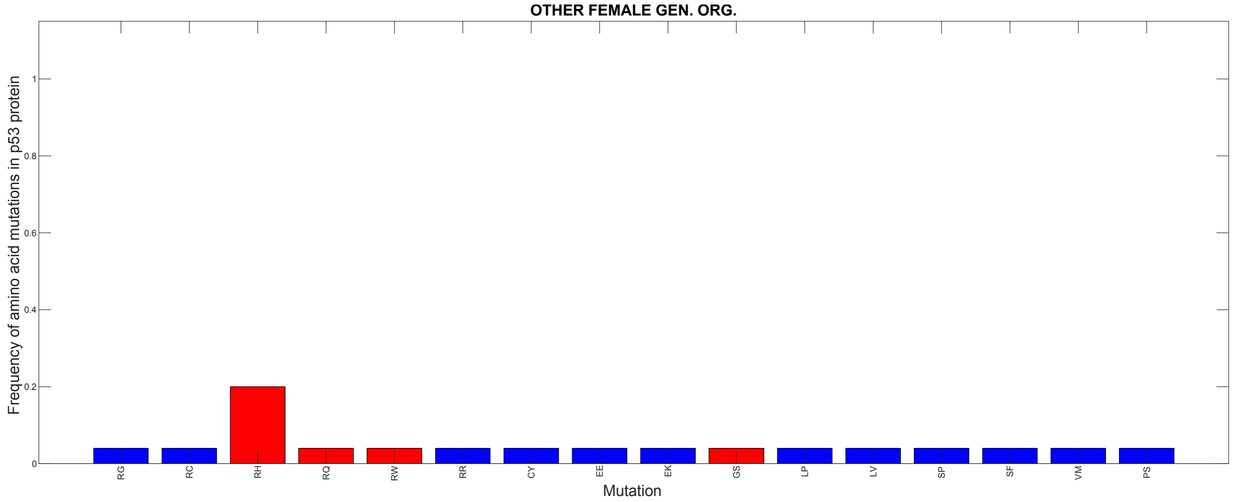

OTHER RESPIR. SYST.
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AP
RG
RH
RQ
RW
EE
HN
LL
KR
Mutation

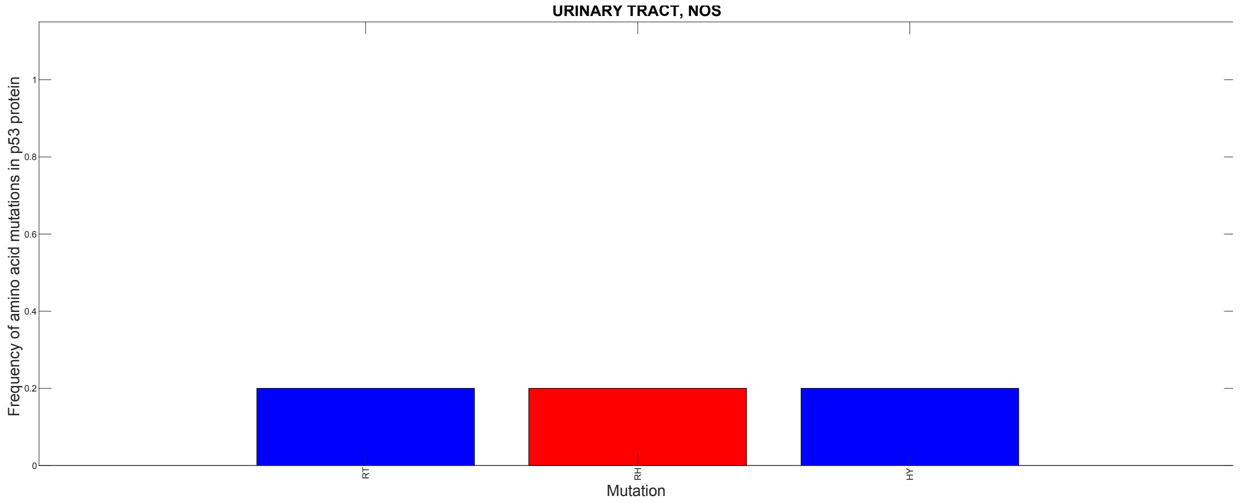

SALIVARY GLAND
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AA
RC
RS
NS
DH
CS
EP
HP
SP
VM
PA
Mutation


PALATE
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
RC
RS
RH
RM
NK
CF
IF
KN
KR
MI
FY
TN
TM
VC
Mutation

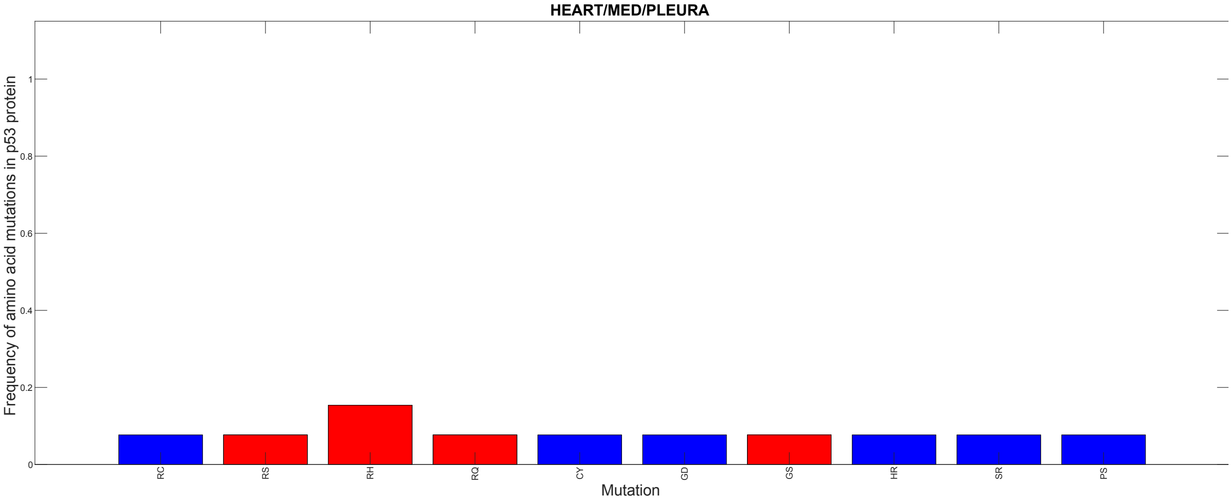
`

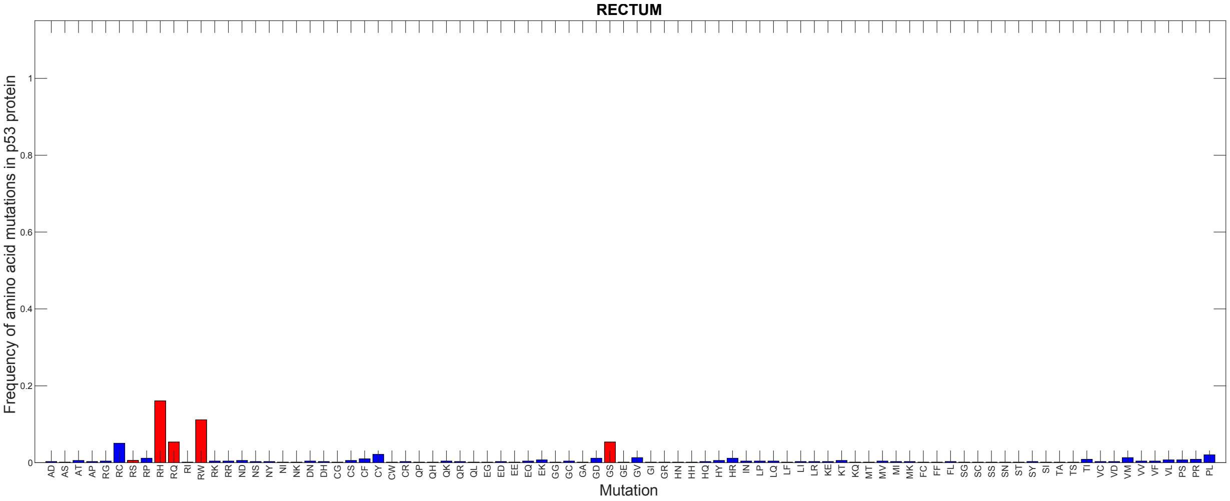

OTHER ENDOCRINE Gl.

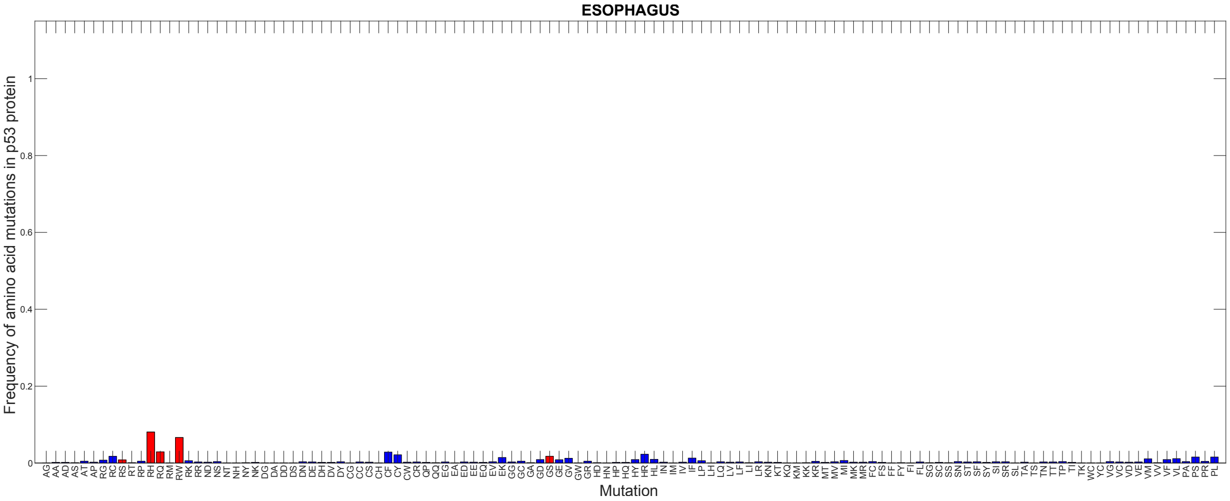

OTHER FEMALE GEN. ORG.


URINARY TRACT, NOS

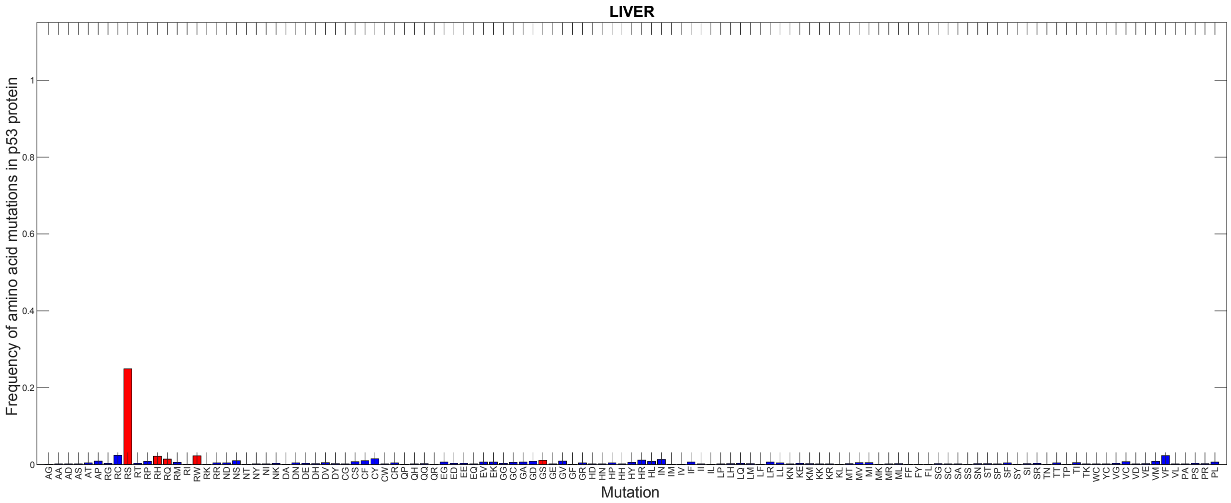
`

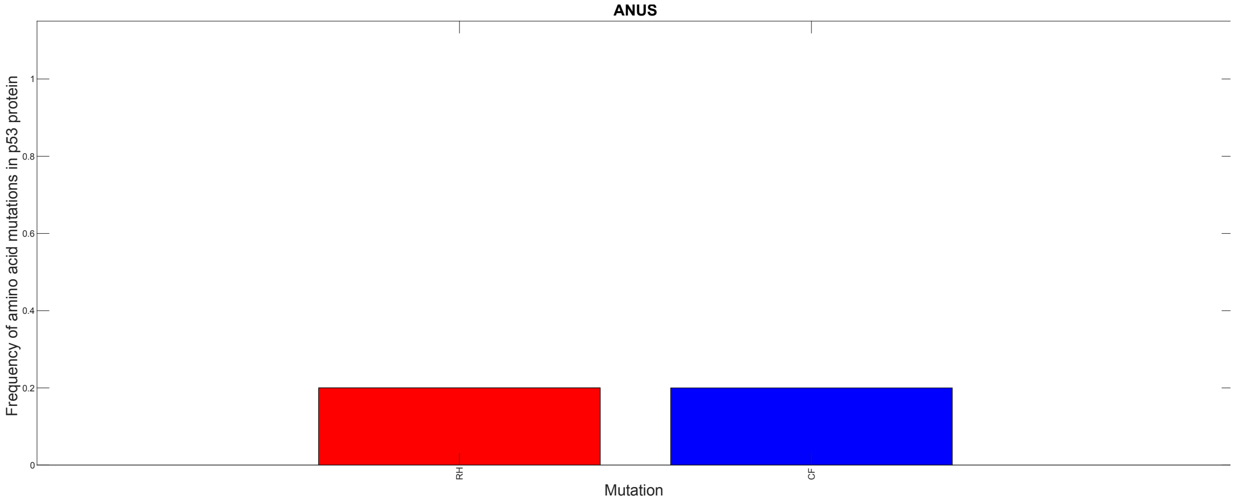

HEART/MED/PLEURA
Frequency of amino acid mutations in p53 protein
1
0.8
0.6
0.4
0.2
0
RC
RS
RH
RQ
CY
GD
GS
HR
SR
PS
Mutation

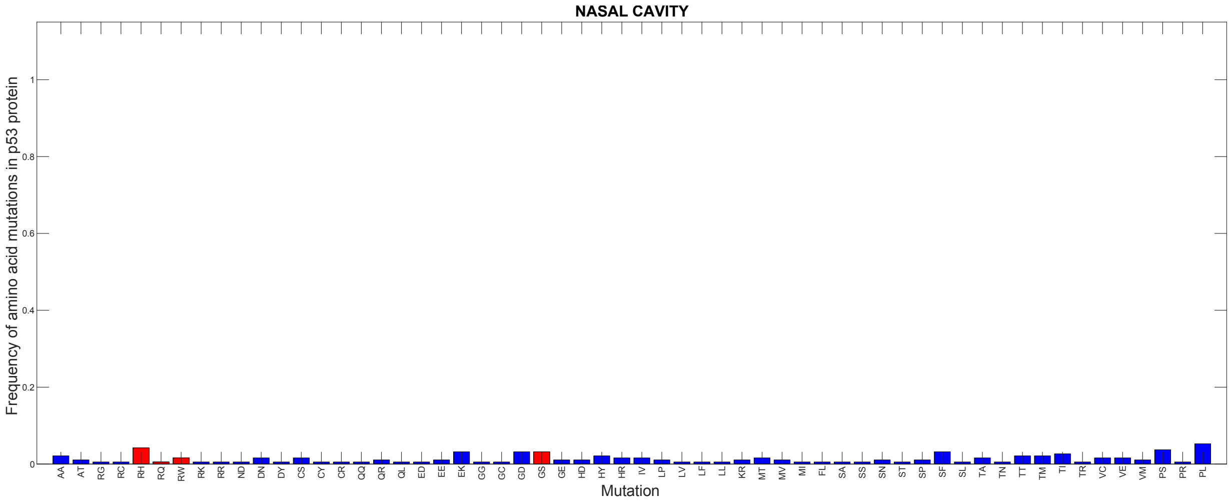

RECTUM
Frequency of amino acid mutations in p53 protein
1
0.8
0.6
0.4
0.2
0
AD AS AT AP RG RC RS RP RH RQ RI RW RK RR ND NS NY NI NK DN DH CG CS CF CY CW CR QP QH QK QR QL EG ED EE EQ EK GG GC GA GD GS GE GV GI GR HN HH HQ HY HR IN LP LQ LF LI LR KE KT KQ MT MV MI MK FC FF FL SG SC SS SN ST SY SI TA TS TI VC VD VM VV VF VL PS PR PL
Mutation


ESOPHAGUS
Frequency of amino acid mutations in p53 protein
1
0.8
0.6
0.4
0.2
0
AG AA AD AS AT AP RG RC RS RT RP RH RQ RM RW RK RR ND NS NT NH NY NK DG DA DD DS DN DE DH DV DY CG CC CS CH CF CY CW CR QP QQ EG EA ED EE EQ EV EK GG GC GA GD GS GE GV GW GR HD HN HP HQ HY HR HL IN IM IV IF LP LH LQ LV LF LI LR KN KT KQ KM KK KR MT MV MI MK MR FC FS FF FY FI FL SG SC SS SN ST SF SY SI SR SL TA TS TN TT TP TI TK WC YC VG VC VD VE VM VV VF VL PA PS PR PL
Mutation

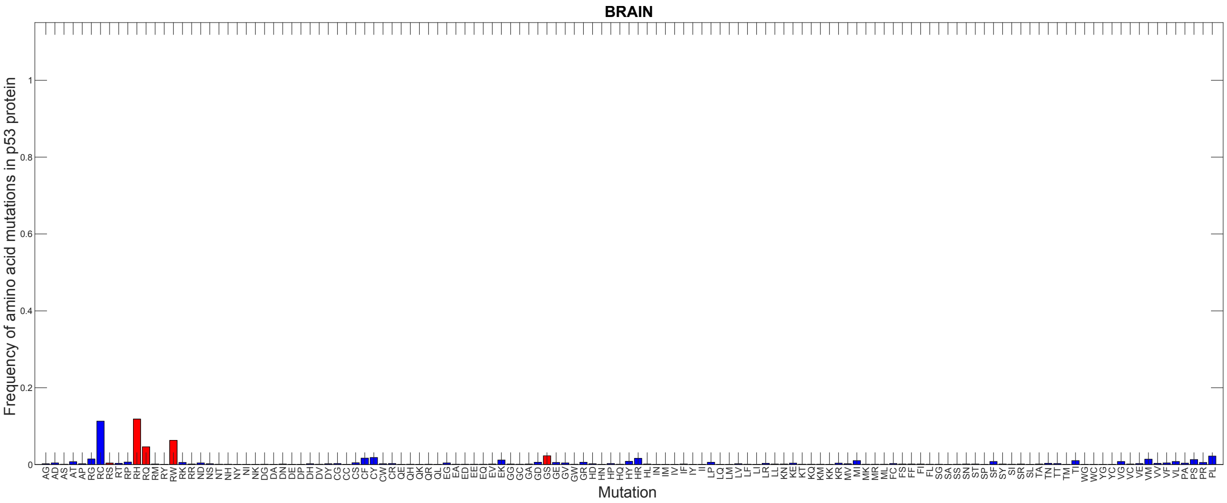
`

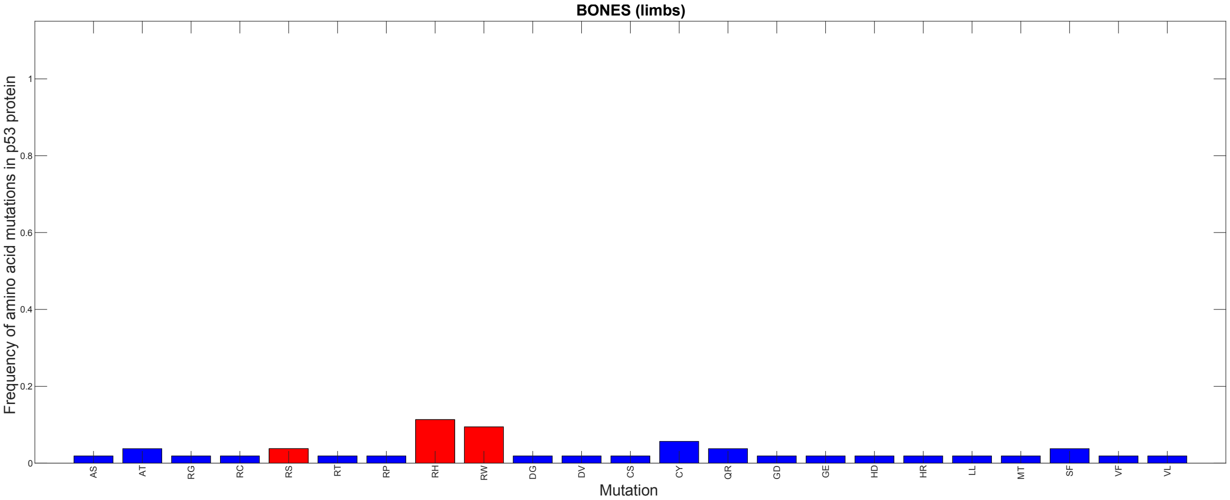

LIVER
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
Mutation
AG
AA
AD
AS
AT
AP
RG
RC
RS
RT
RP
RH
RQ
RM
RI
RW
RK
RR
ND
NS
NT
NY
NI
NK
DA
DN
DE
DH
DV
DY
CG
CS
CF
CY
CW
CR
QP
QH
QQ
QR
EG
ED
EE
EQ
EV
EK
GG
GC
GA
GD
GS
GE
GV
GF
GR
HD
HN
HP
HH
HY
HR
HL
IN
IM
IV
IF
II
IL
LP
LH
LQ
LM
LF
LR
LL
KN
KE
KM
KK
KR
KL
MT
MV
MI
MK
MR
ML
FF
FY
FL
SG
SC
SA
SS
SN
ST
SP
SF
SY
SI
SR
TN
TT
TP
TI
TK
WC
YC
VG
VC
VD
VE
VM
VF
VL
PA
PS
PR
PL

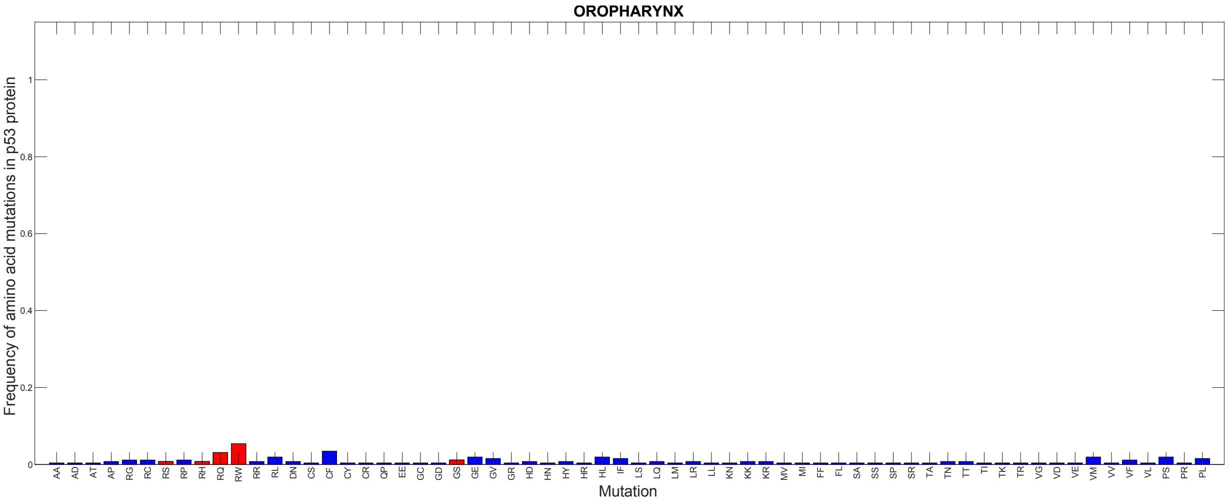

ANUS
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
RH
CF
Mutation


NASAL CAVITY
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AA
AT
RG
RC
RH
RQ
RW
RK
RR
ND
DN
DY
CS
CY
CR
QQ
QR
QL
ED
EE
EK
GG
GC
GD
GS
GE
HD
HY
HR
IV
LP
LV
LF
LL
KR
MT
MV
MI
FL
SA
SS
SN
ST
SP
SF
SL
TA
TN
TT
TM
TI
TR
VC
VE
VM
PS
PR
PL
Mutation

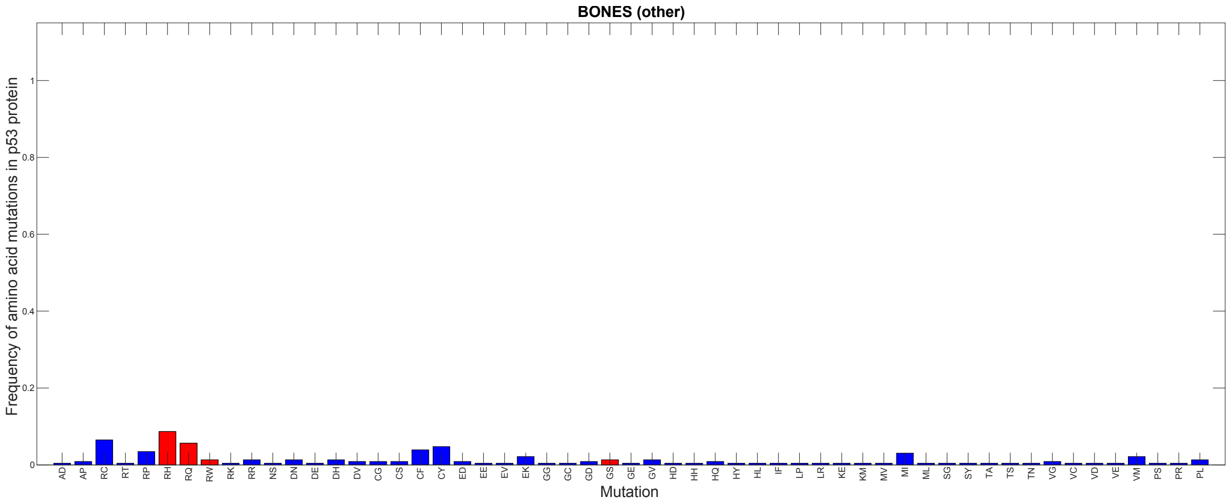
`

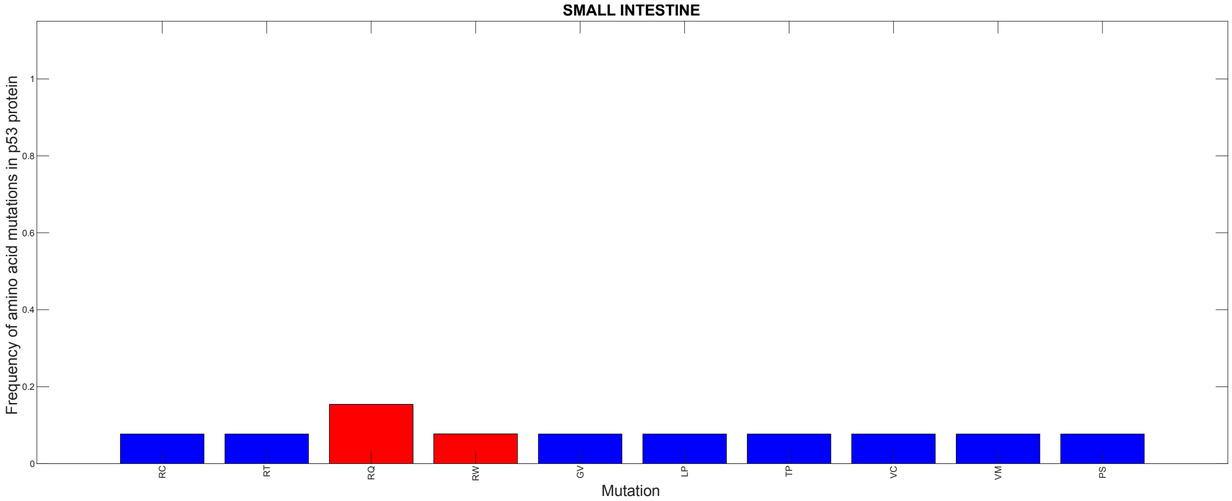

BRAIN
Frequency of amino acid mutations in p53 protein
1
0.8
0.6
0.4
0.2
0
AG AD AS AT AP RG RC RS RT RP RH RQ RM RY RW RK RR ND NS NT NH NY NI NK DG DA DN DE DP DH DV DY CG CC CS CF CY CW CR QE QH QK QR QL EG EA ED EE EQ EV EK GG GC GA GD GS GE GV GW GR HD HN HP HQ HY HR HL IN IM IV IF IY II LP LQ LM LV LF LI LR LL KN KE KT KQ KM KK KR MV MI MK MR ML FC FS FF FI FL SG SA SS SN ST SP SF SY SI SR SL TA TN TT TM TI WG WC YG YC VG VC VE VM VV VF VL PA PS PR PL
Mutation

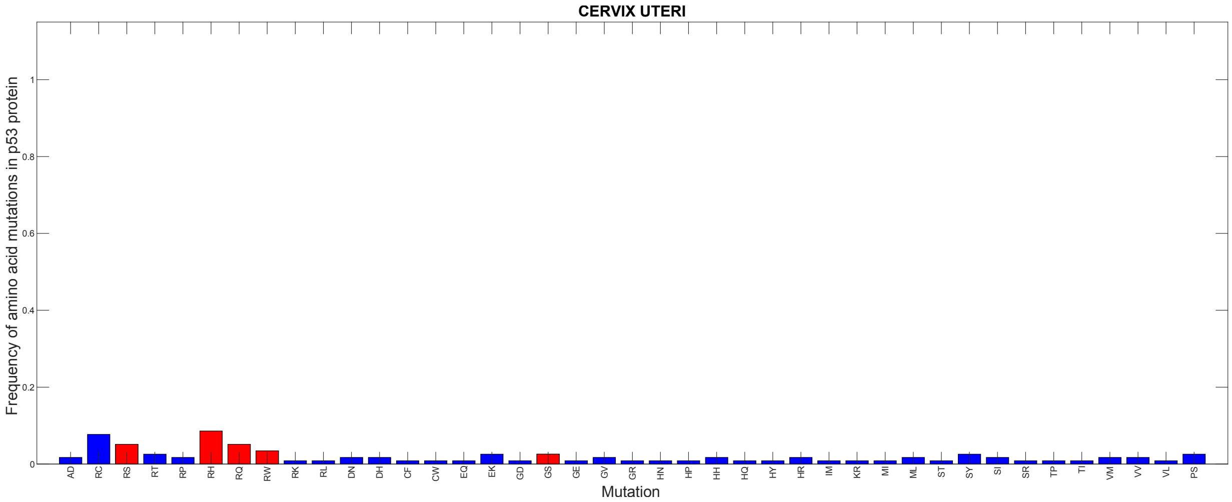

BONES (limbs)
Frequency of amino acid mutations in p53 protein
1
0.8
0.6
0.4
0.2
0
AS AT RG RC RS RT RP RH RW DG DV CS CY QR GD GE HD HR LL MT SF VF VL
Mutation


OROPHARYNX
Frequency of amino acid mutations in p53 protein
1
0.8
0.6
0.4
0.2
0
AA AD AT AP RG RC RS RP RH RQ RW RR RL DN CS CF CY CR QP EE GC GD GS GE GV GR HD HN HY HR HL IF LS LQ LM LR LL KN KK KR MV MI FF FL SA SS SP SR TA TN TT TI TK TR VG VD VE VM VV VF VL PS PR PL
Mutation

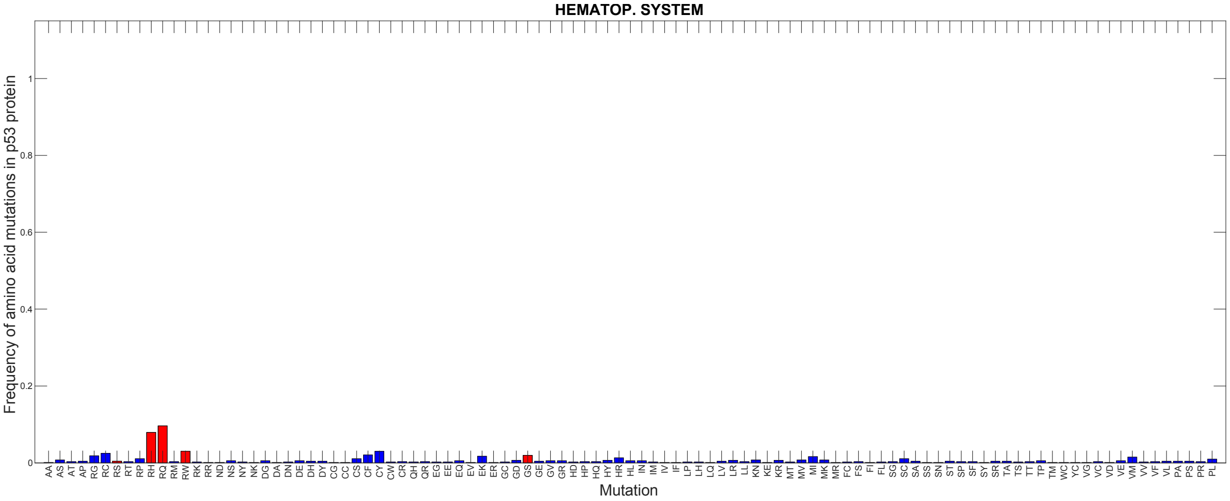
`

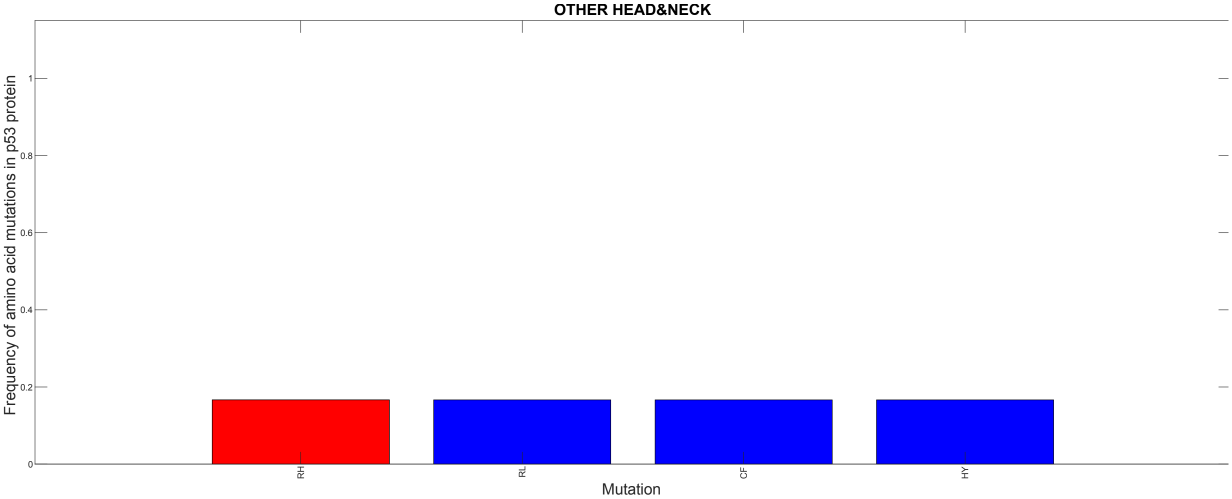

BONES (other)
Frequency of amino acid mutations in p53 protein
1
0.8
0.6
0.4
0.2
0
AD
AP
RC
RT
RP
RH
RQ
RW
RK
RR
NS
DN
DE
DH
DV
CG
CS
CF
CY
ED
EE
EV
EK
GG
GC
GD
GS
GE
GV
HD
HH
HQ
HY
HL
IF
LP
LR
KE
KM
MV
MI
ML
SG
SY
TA
TS
TN
VG
VC
VD
VE
VM
PS
PR
PL
Mutation

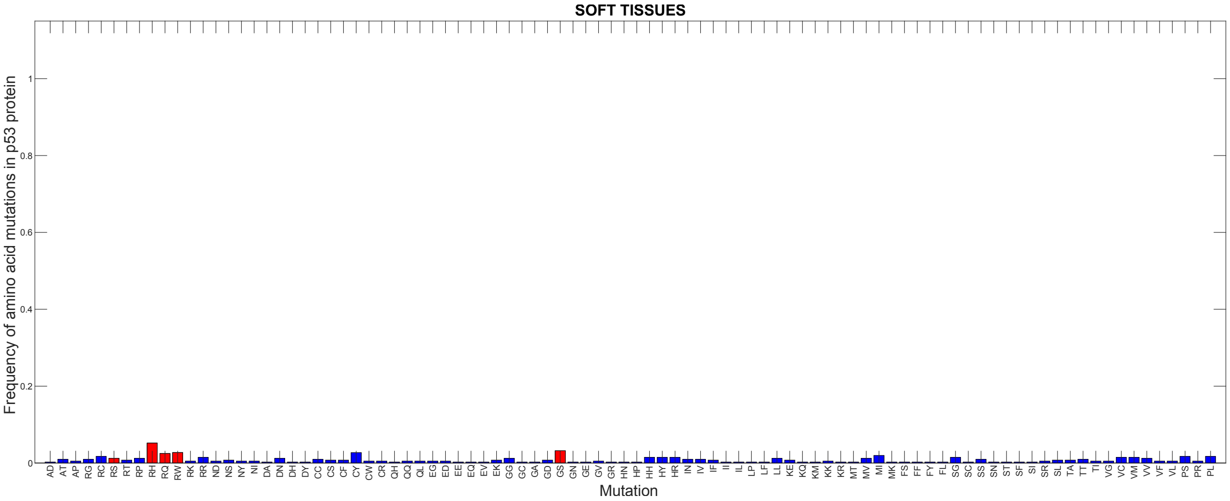

SMALL INTESTINE
Frequency of amino acid mutations in p53 protein
1
0.8
0.6
0.4
0.2
0
RC
RT
RQ
RW
GV
LP
TP
VC
VM
PS
Mutation


CERVIX UTERI
Frequency of amino acid mutations in p53 protein
1
0.8
0.6
0.4
0.2
0
AD
RC
RS
RT
RP
RH
RQ
RW
RK
RL
DN
DH
CF
CW
EQ
EK
GD
GS
GE
GV
GR
HN
HP
HH
HQ
HY
HR
IM
KR
MI
ML
ST
SY
SI
SR
TP
TI
VM
VV
VL
PS
Mutation

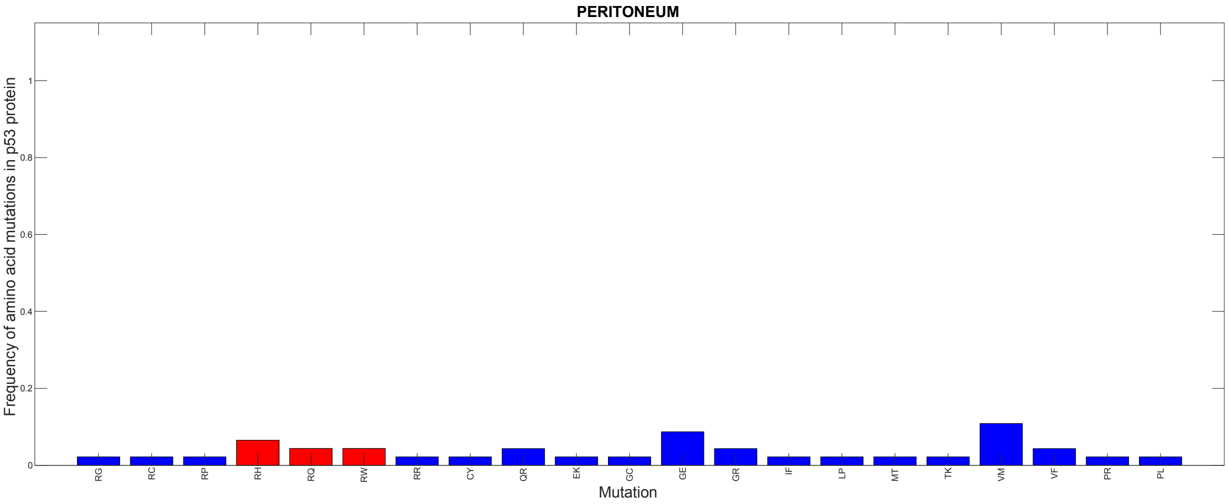
`

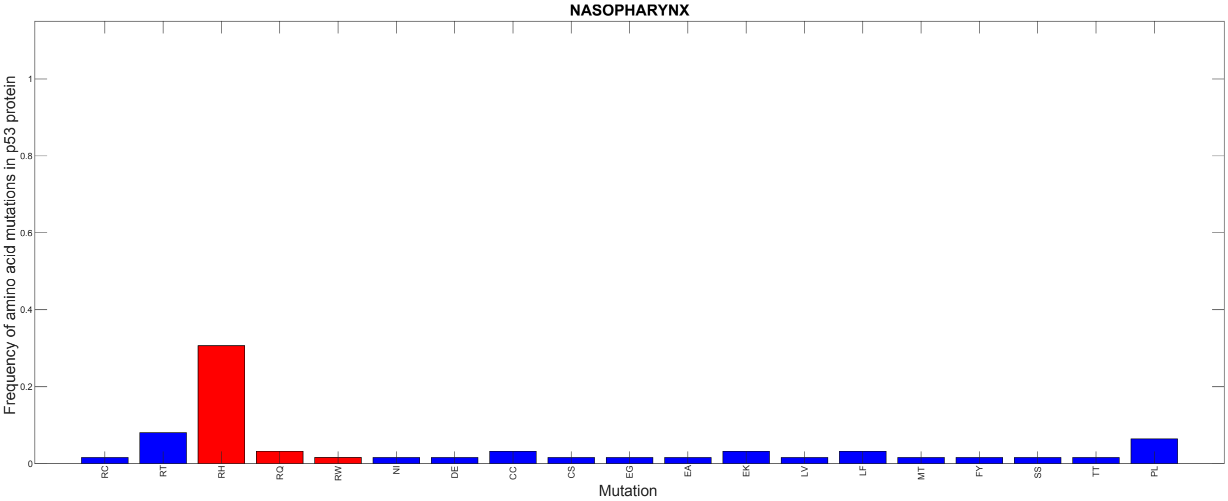

HEMATOP. SYSTEM
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AA AS AT AP RG RC RS RT RP RH RQ RM RW RK RR ND NS NY NK DG DA DN DE DH DY CG CC CS CF CY CW CR QH QR EG EE EQ EV EK ER GC GD GS GE GV GR HD HP HQ HY HR HL IN IM IV IF LP LH LQ LV LR LL KN KE KR MT MV MI MK MR FC FS FI FL SG SC SA SS SN ST SP SF SY SR TA TS TT TP TM WC YC VG VC VD VE VM VV VF VL PA PS PR PL
Mutation

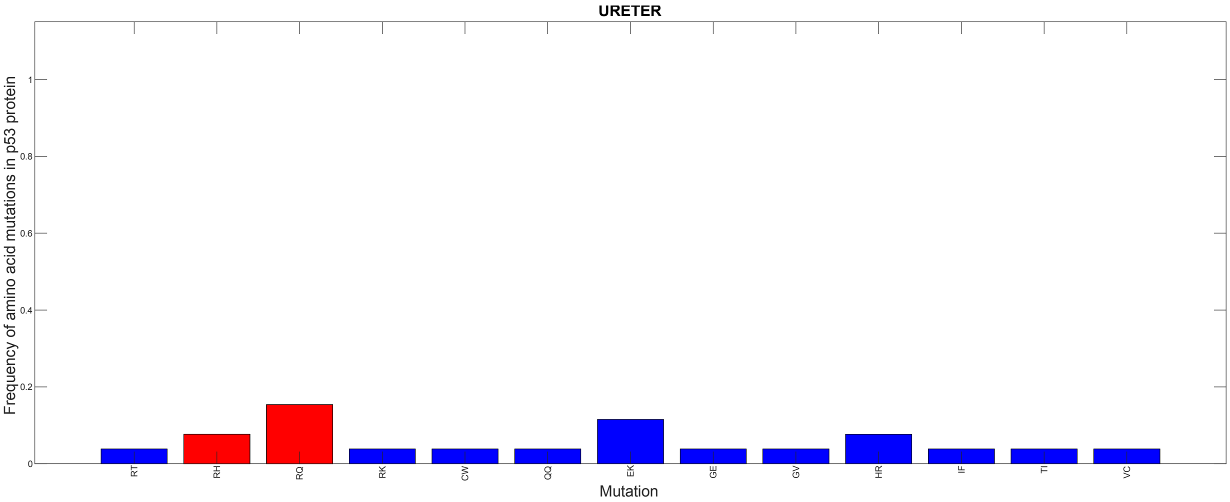

OTHER HEAD&NECK
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
RH
RL
CF
HY
Mutation


SOFT TISSUES
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AD AT AP RG RC RS RT RP RH RQ RW RK RR ND NS NY NI DA DN DH DY CC CS CF CY CW CR QH QQ QL EG ED EE EQ EV EK GG GC GA GD GS GN GE GV GR HN HP HH HY HR IN IV IF II IL LP LF LL KE KQ KM KK KR MT MV MI MK FS FF FY FL SG SC SS SN ST SF SI SR SL TA TT TI VG VC VM VV VF VL PS PR PL
Mutation

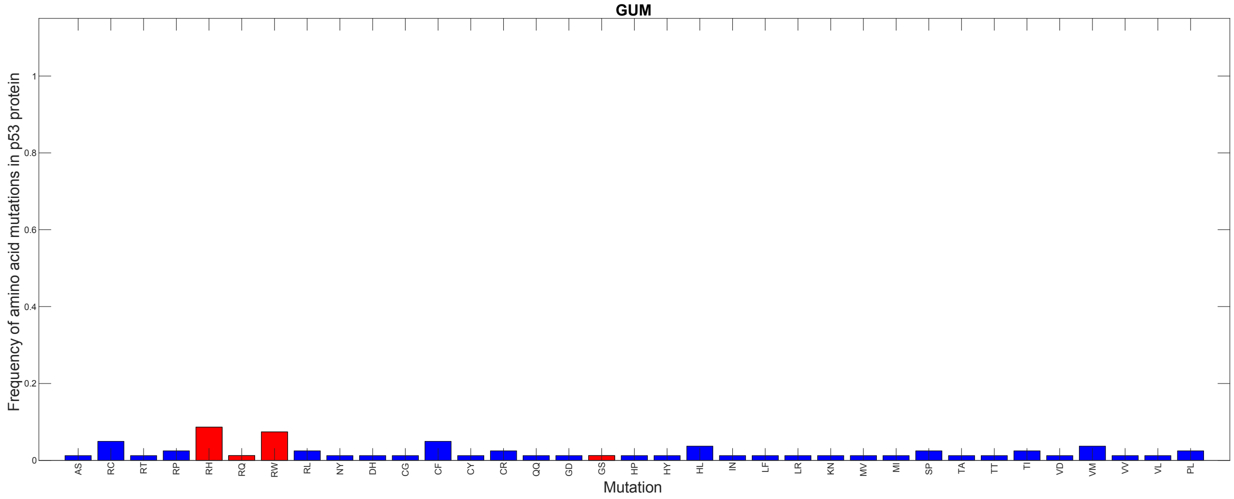
`


PERITONEUM
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
RG
RC
RP
RH
RQ
RW
RR
CY
QR
EK
GC
GE
GR
IF
LP
MT
TK
VM
VF
PR
PL
Mutation

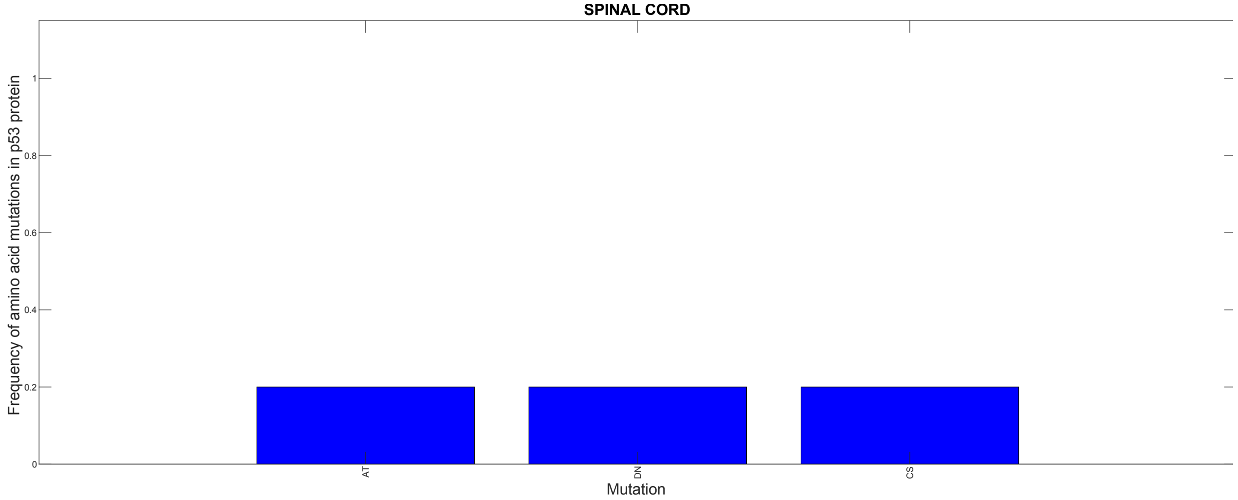

NASOPHARYNX
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
RC
RT
RH
RQ
RW
NI
DE
CC
CS
EG
EA
EK
LV
LF
MT
FY
SS
TT
PL
Mutation


URETER
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
RT
RH
RQ
RK
CW
QQ
EK
GE
GV
HR
IF
TI
VC
Mutation

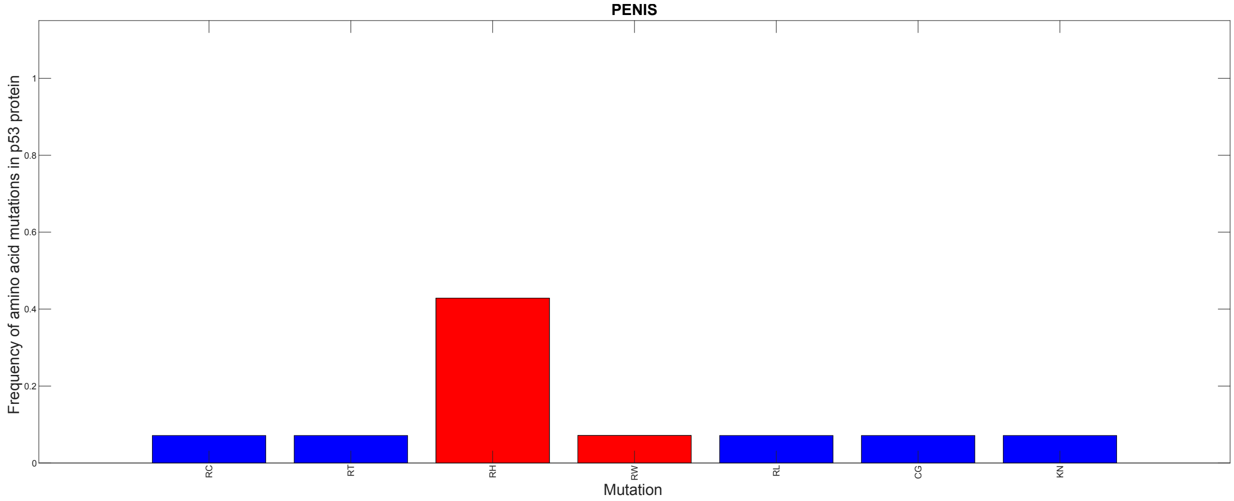
`



GUM

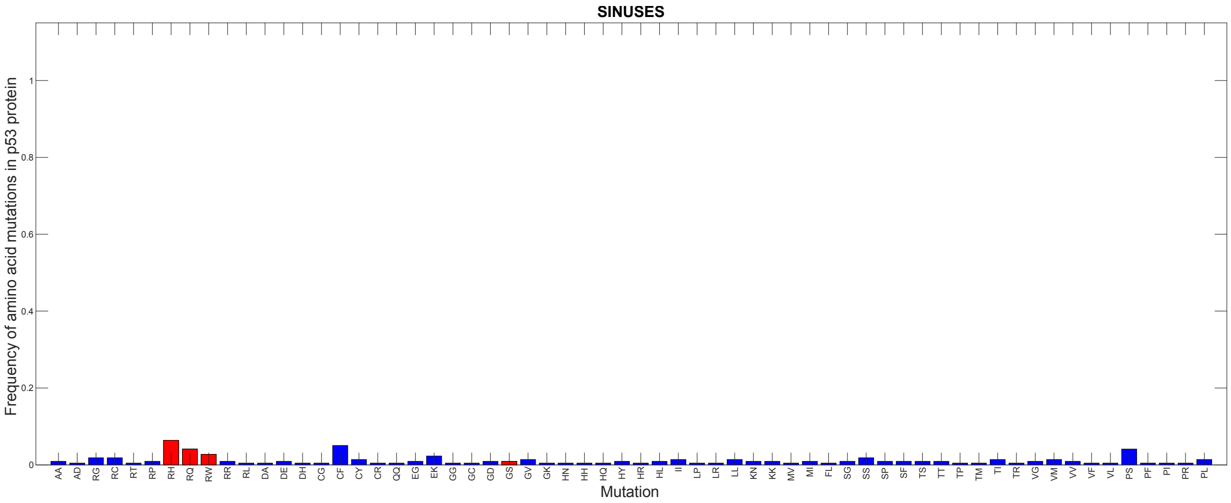


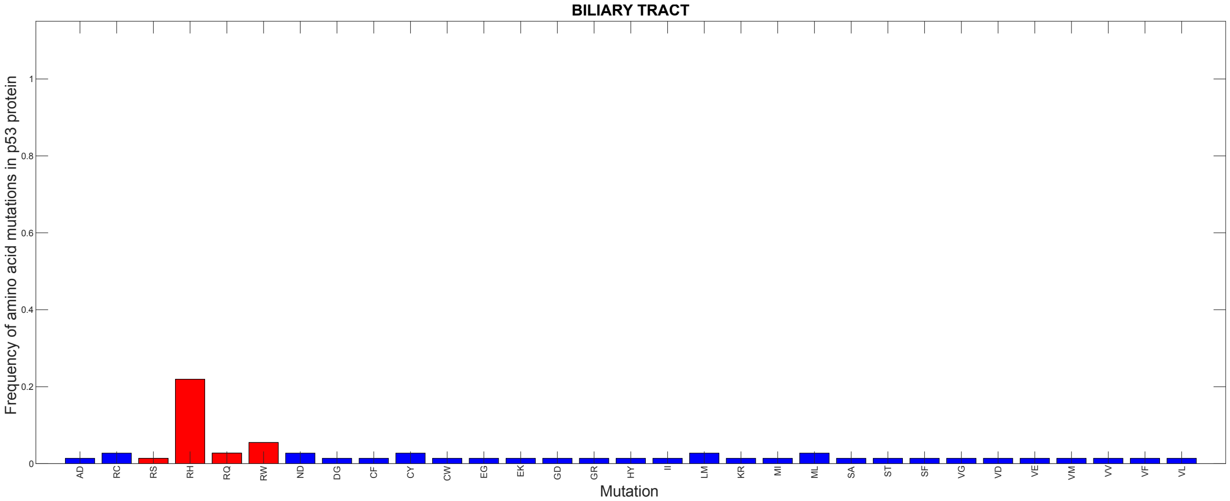
SPINAL CORD

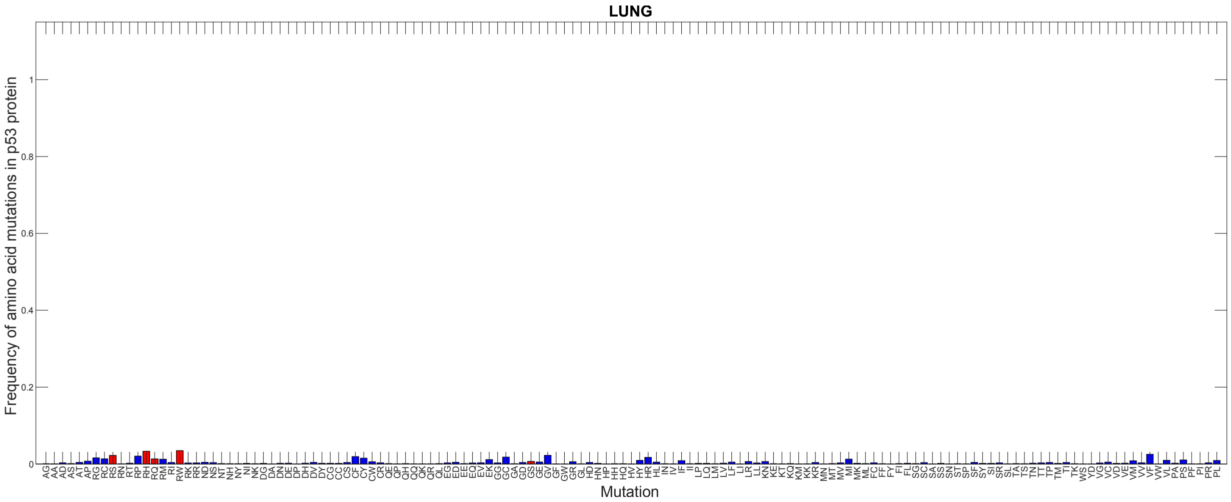


PENIS

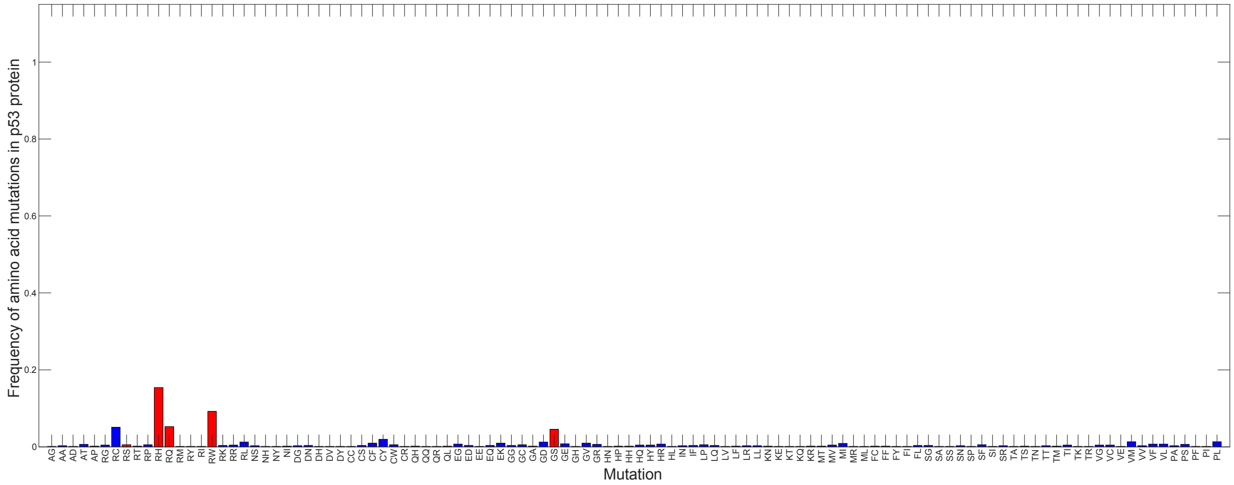
`


SINUSES
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AA AD RG RC RT RP RH RQ RW RR RL DA DE DH CG CF CY CR QQ EG EK GG GC GD GS GV GK HN HH HQ HY HR HL II LP LR LL KN KK MV MI FL SG SS SP SF TS TT TP TM TI TR VG VM VV VF VL PS PF PI PR PL
Mutation

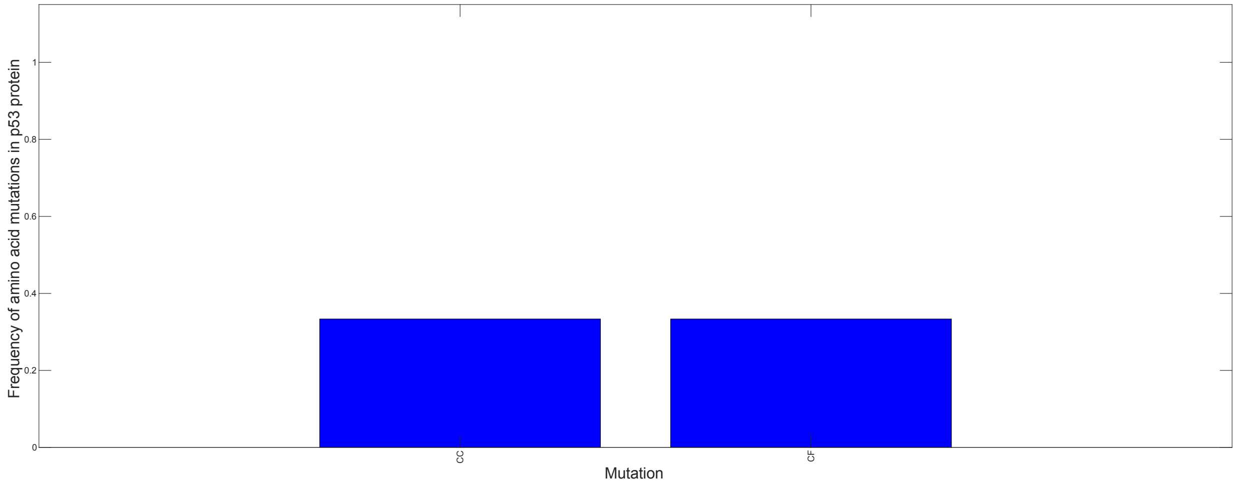

BILIARY TRACT
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AD RC RS RH RQ RW ND DG CF CY CW EG EK GD GR HY II LM KR MI ML SA ST SF VG VD VE VM VV VF VL
Mutation


LUNG
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AG AA AD AS AT AP RG RC RS RN RT RP RH RQ RM RI RW RK RR ND NS NT NH NY NI NK DG DA DN DE DP DH DV DY CG CC CS CF CY CW CR QE QP QH QQ QK QR QL EG ED EE EQ EV EK GG GC GA GD GS GE GV GF GW GR GL HD HN HP HH HQ HV HY HR HL IN IV IF II LP LQ LM LV LF LI LR LL KN KE KT KQ KM KK KR MN MT MV MI MK ML FC FF FY FI FL SG SC SA SS SN ST SP SF SY SI SR SL TA TS TN TT TP TM TI TK WS YD VG VC VD VE VM VV VF VW VL PA PS PF PI PR PL
Mutation

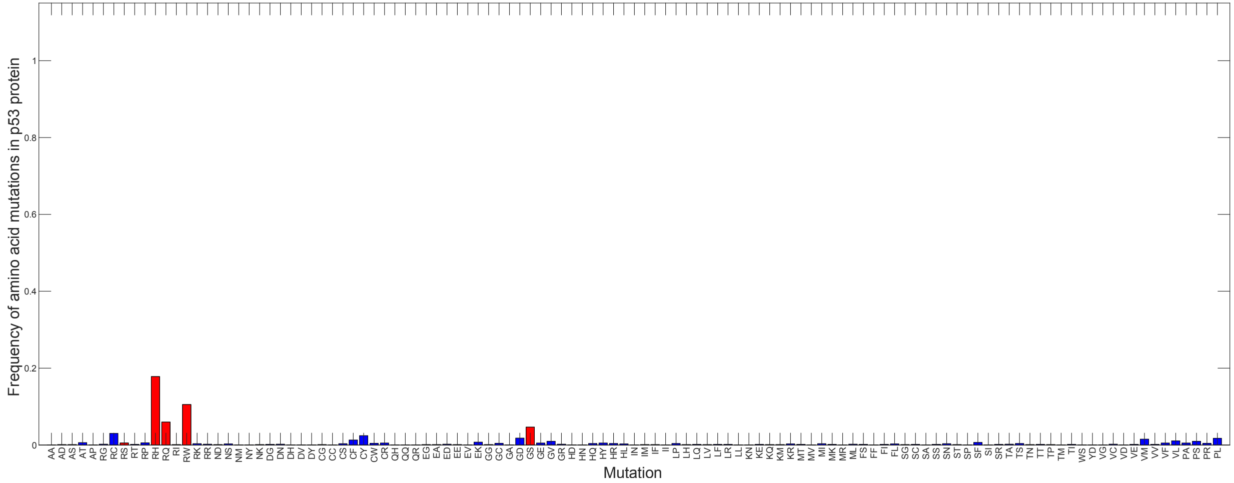
`



COLON

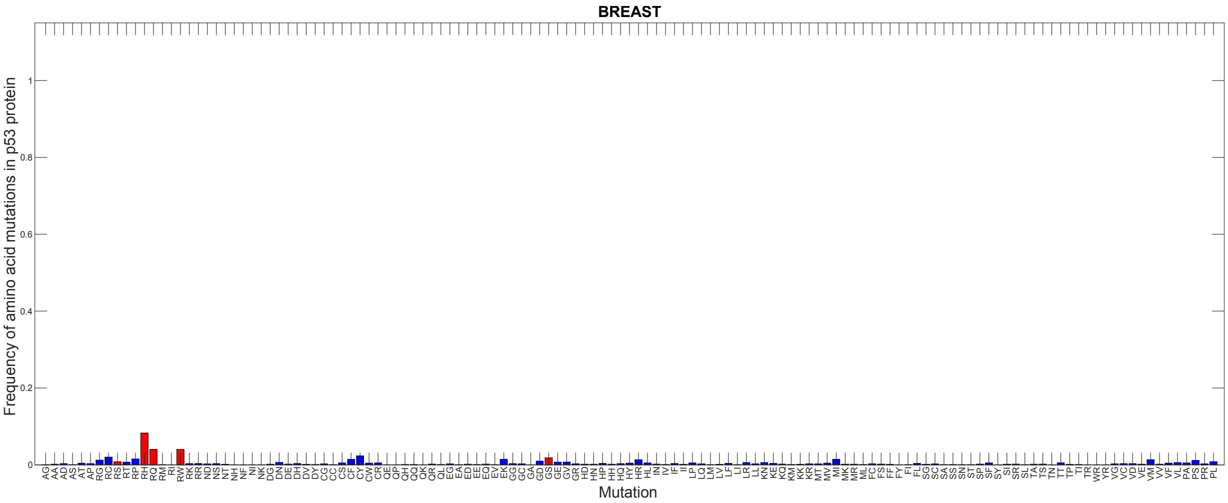


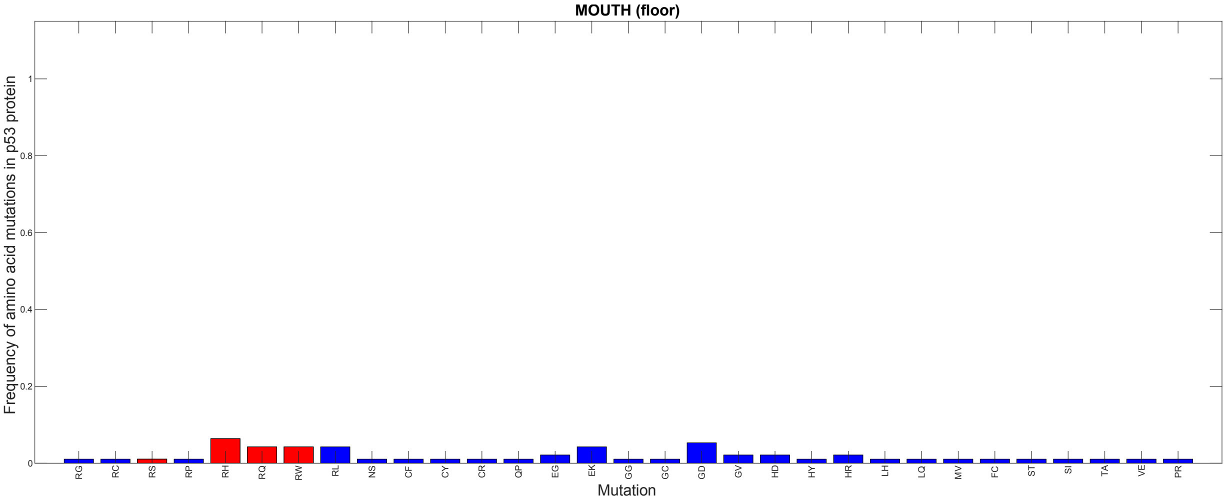
OTHER DIGESTIVE ORG.

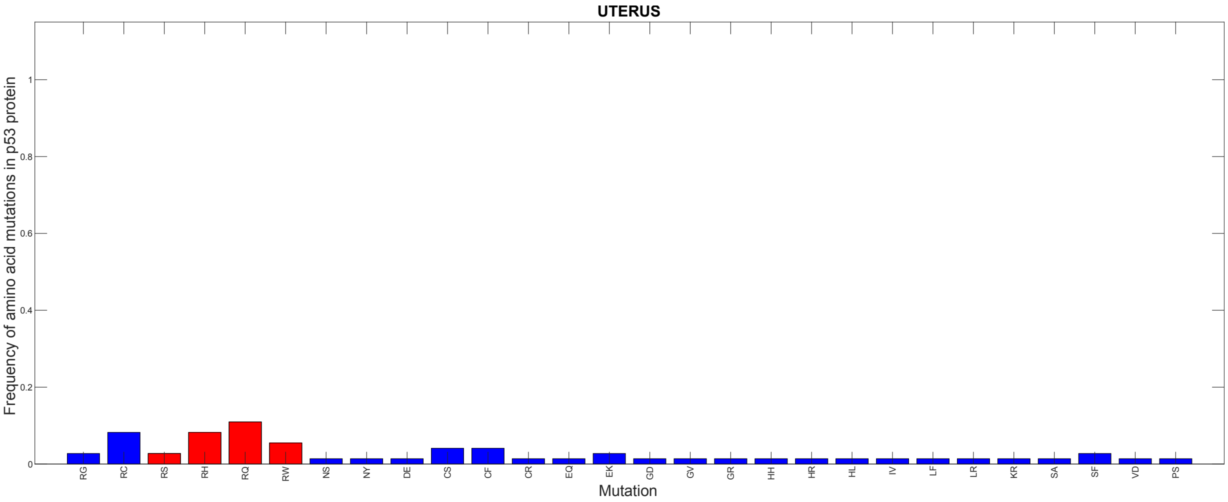


COLORECTUM, NOS

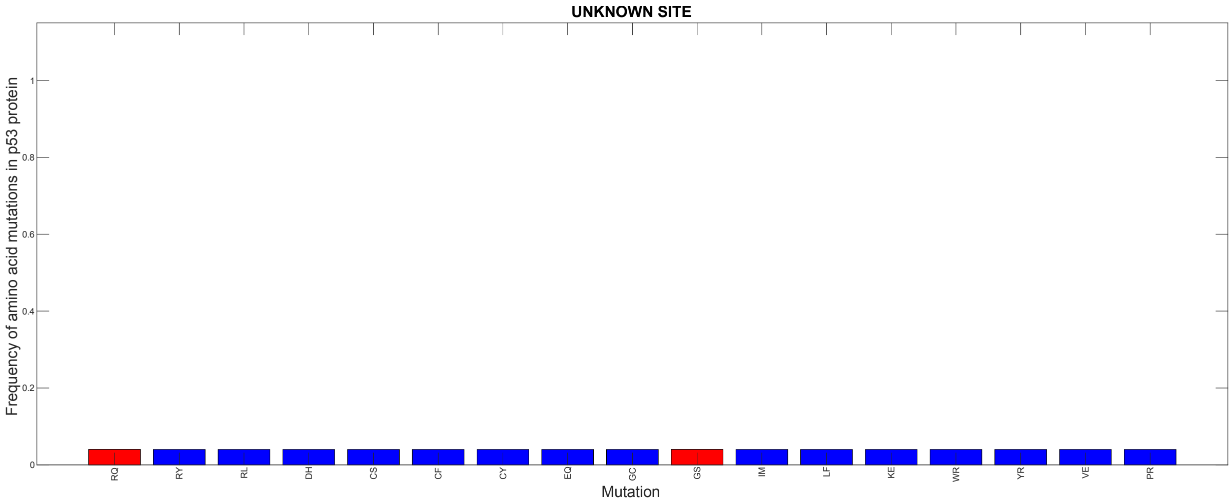
`

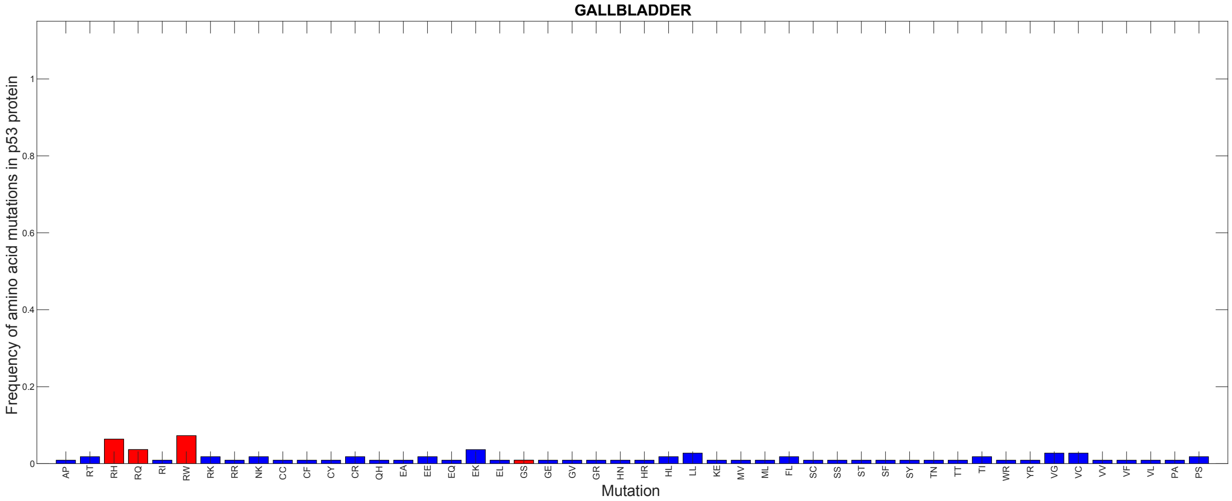

BREAST
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AG AA AD AS AT AP RG RC RS RT RP RH RQ RM RI RW RK RR ND NS NT NH NF NI NK DG DN DE DH DV DY CG CC CS CF CY CW CR QE QP QH QQ QK QR QL EG EA ED EE EQ EV EK GG GC GA GD GS GE GV GR HD HN HP HH HQ HY HR HL IN IV IF II LP LQ LM LV LF LI LR LL KN KE KQ KM KK KR MT MV MI MK MR ML FC FS FF FY FI FL SG SC SA SS SN ST SP SF SY SI SR SL TA TS TN TT TP TI TR WR YR VG VC VD VE VM VV VF VL PA PS PR PL
Mutation

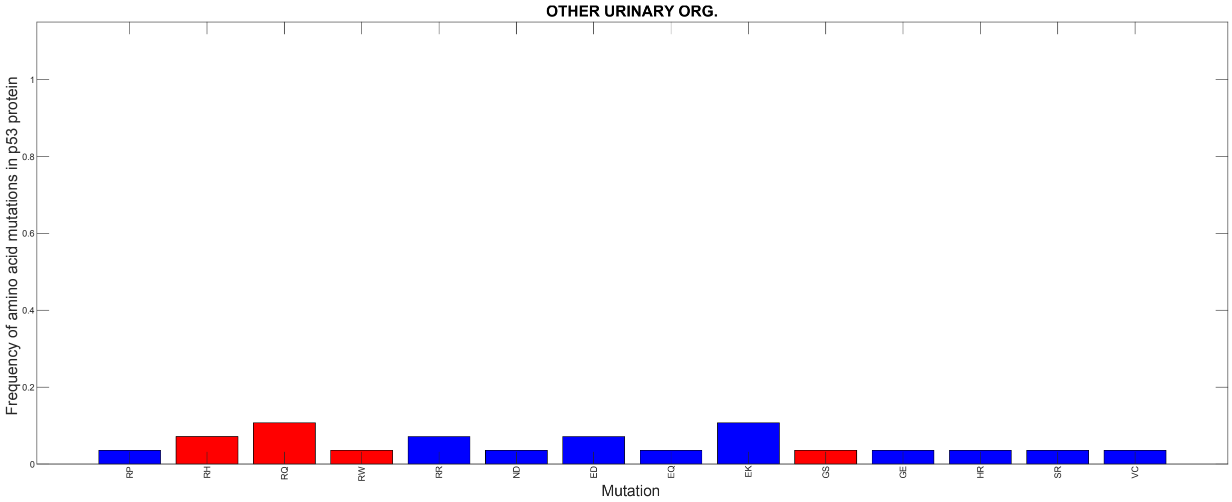

MOUTH (floor)
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
RG RC RS RP RH RQ RW RL NS CF CY CR QP EG EK GG GC GD GV HD HY HR LH LQ MV FC ST SI TA VE PR
Mutation


UTERUS
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
RG RC RS RH RQ RW NS NY DE CS CF CR EQ EK GD GV GR HH HR HL IV LF LR KR SA SF VD PS
Mutation

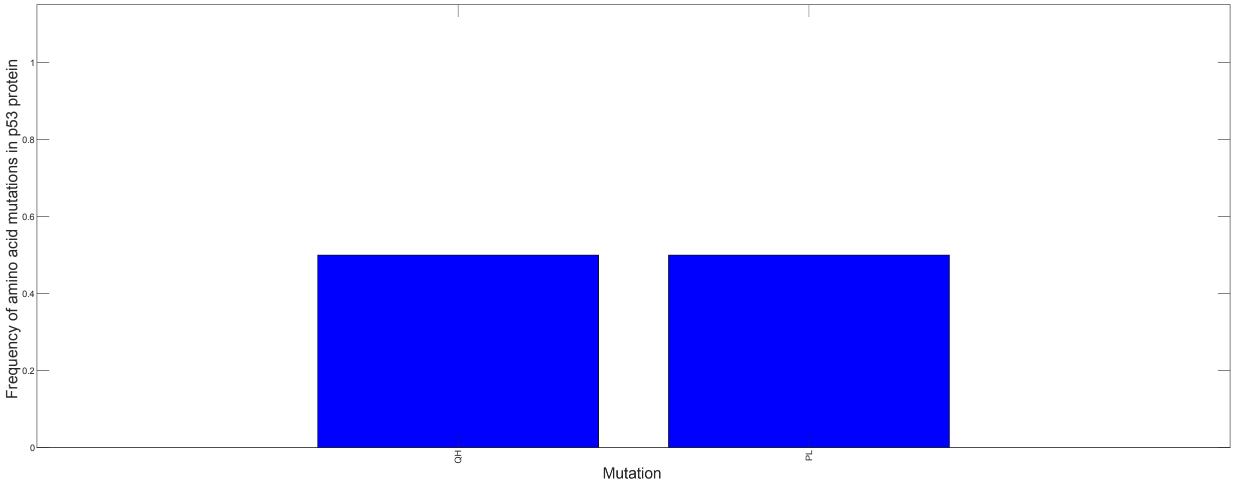
`

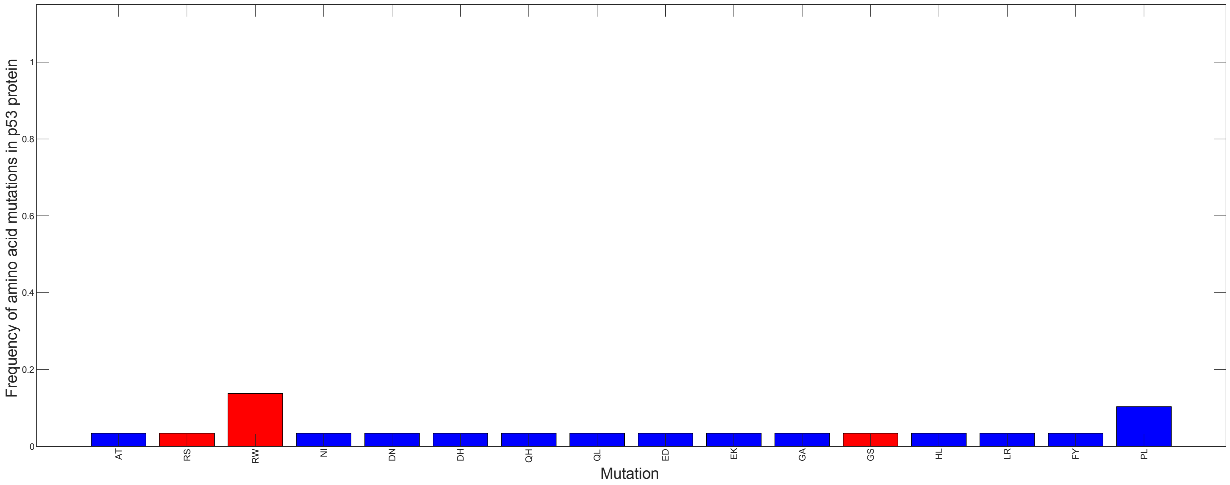

UNKNOWN SITE
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
RQ
RY
RL
DH
CS
CF
CY
EQ
GC
GS
IM
LF
KE
WR
YR
VE
PR
Mutation

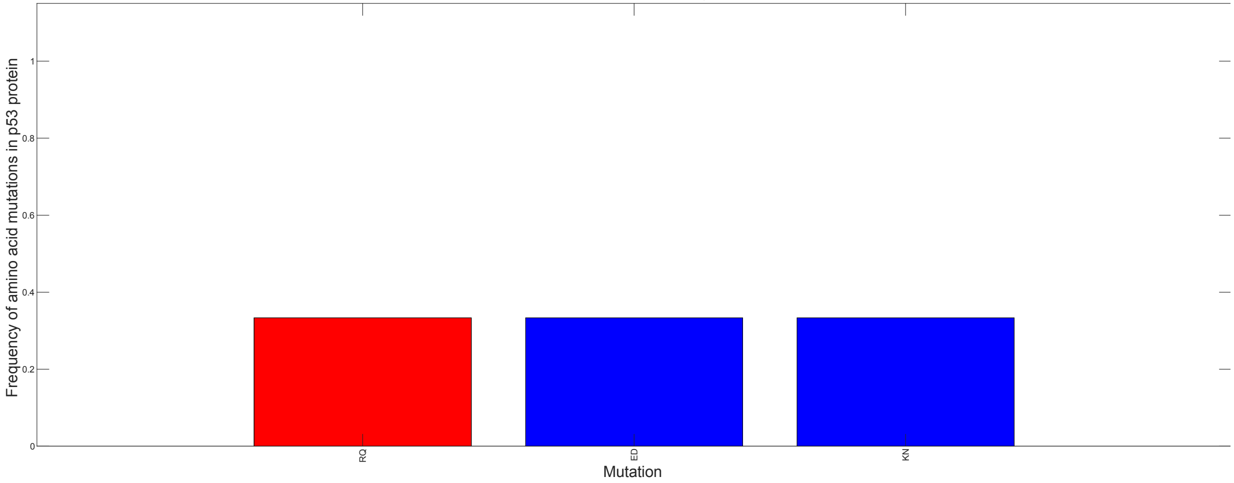

GALLBLADDER
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
AP
RT
RH
RQ
RI
RW
RK
RR
NK
CC
CF
CY
CR
QH
EA
EE
EQ
EK
EL
GS
GE
GV
GR
HN
HR
HL
LL
KE
MV
ML
FL
SC
SS
ST
SF
SY
TN
TT
TI
WR
YR
VG
VC
VV
VF
VL
PA
PS
Mutation


OTHER URINARY ORG.
Frequency of amino acid mutations in p53 protein
0
0.2
0.4
0.6
0.8
1
RP
RH
RQ
RW
RR
ND
ED
EQ
EK
GS
GE
HR
SR
VC
Mutation

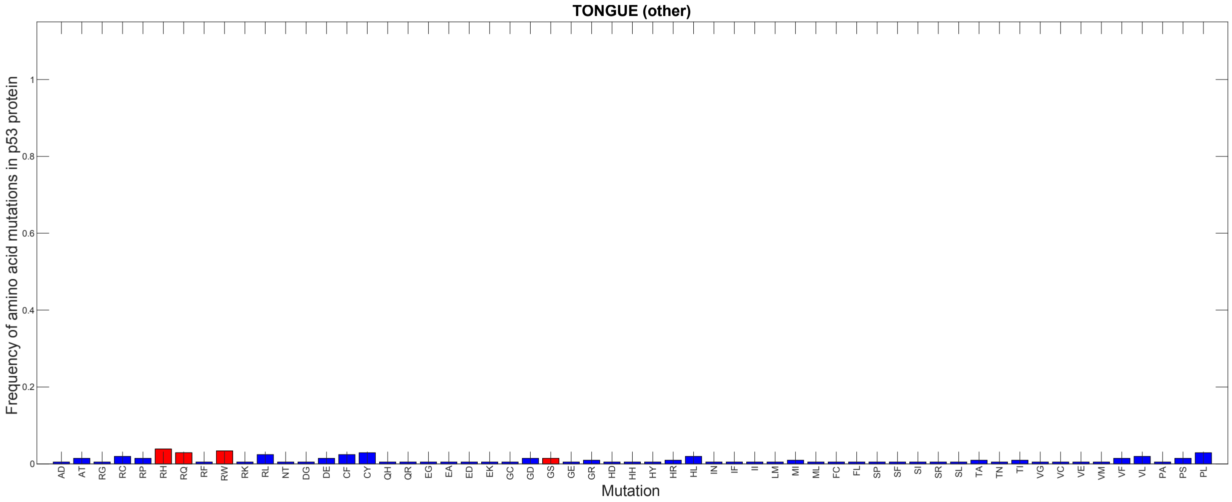
`

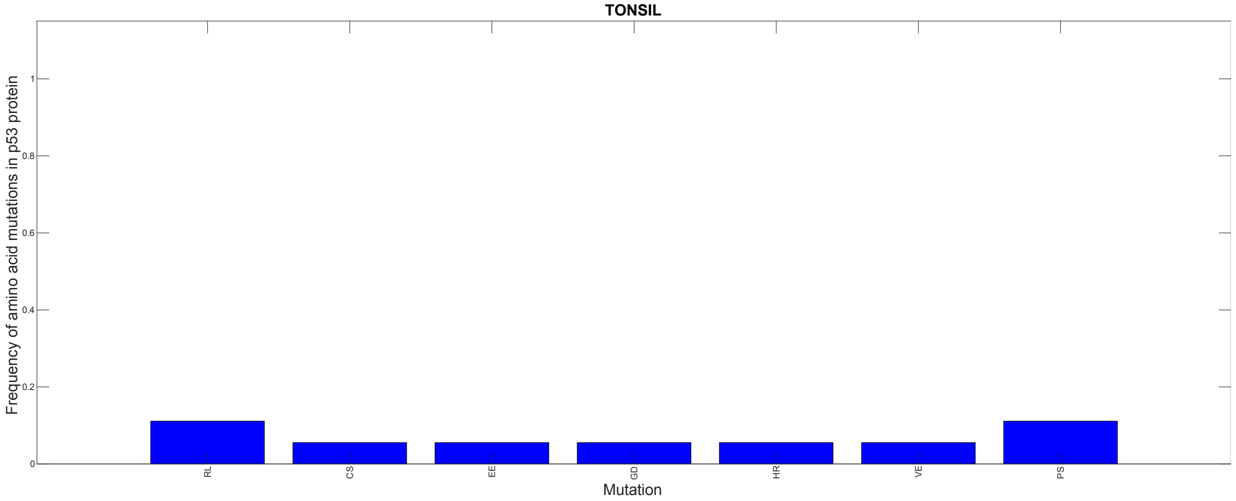

PLACENTA

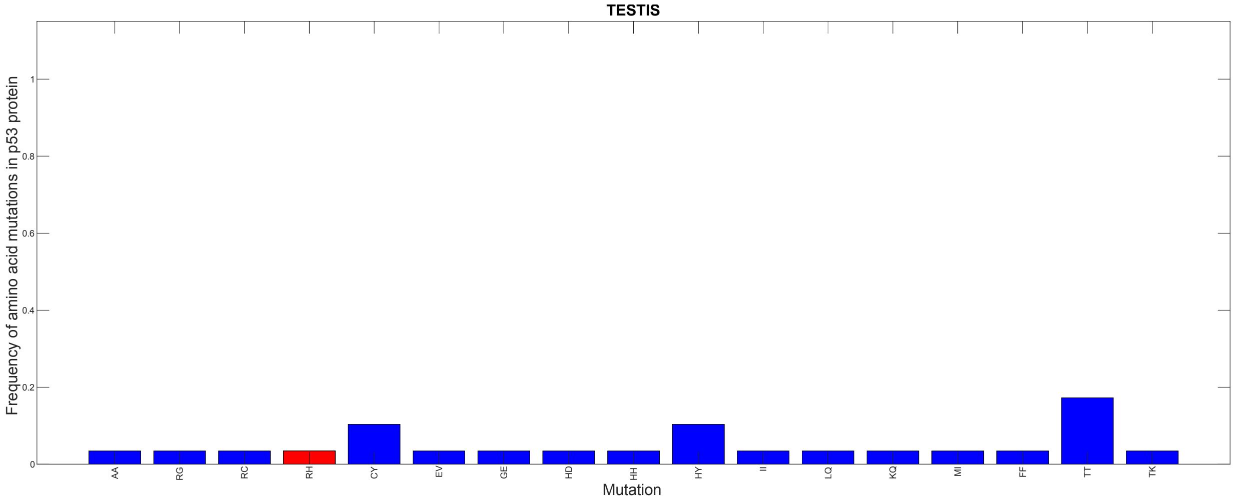

PAROTID GLAND


FEMALE GENITAL ORG., NOS

`


TONGUE (other)


TONSIL


TESTIS

`

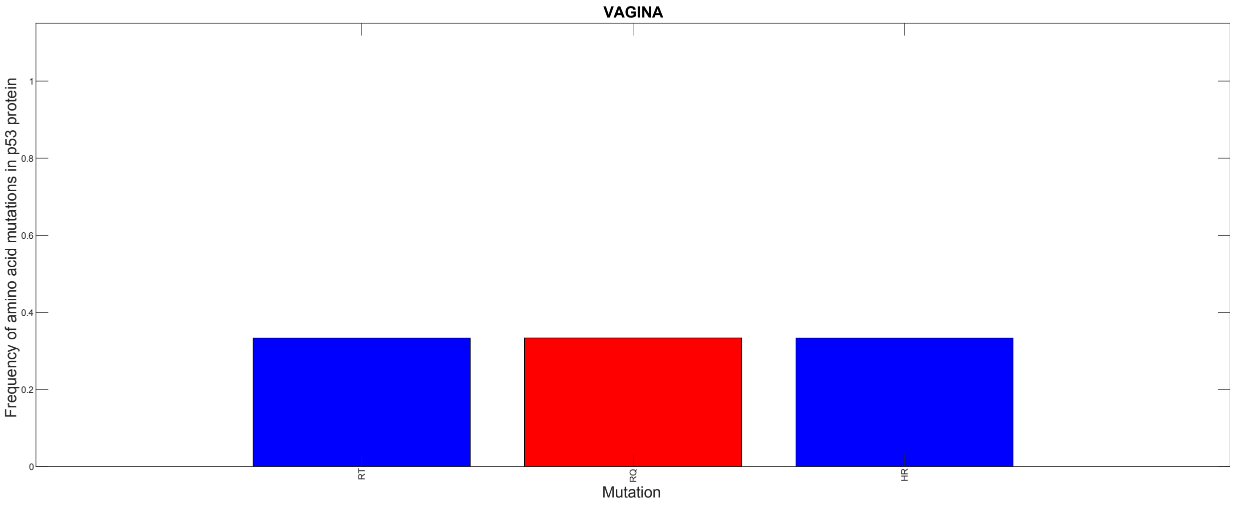


Figure C.2 Frequency of amino acid mutations in the p53 protein in different cancers. The data extracted using IARC database. Each plot shows the mutations of the p53 protein in specific cancer types. The red bar shows the p53 hotspot mutations. In almost all of the cancer types at least one hotspot mutation exists and it is one the highest frequency mutations (in almost 84% of the studied cases).

`

Appendix D [9]

D.1 Landau theory of phase transitions

Landau theory of phase transitions is commonly used to study phase transitions in magnetic materials. Unlike the Ising model, Landau theory doesn't directly and explicitly rely on the interactions between individual spins. Instead, it approximates their interactions by introducing a single average effect. These systems are described by a free energy function, $F(M)$, a power series of the magnetization, M, as an order parameter. To guarantee equal spin orientation in the absence of the external magnetic field, the even powers in the summation are only allowed, so that the free energy of Landau theory would read as

$$F(M,T) = F_0 - hM + \frac{A}{2}M^2 + \frac{B}{4}M^4 + \mathcal{O}(M^6) \tag{D.1}$$

where $A = a(T - T_C)$ and h is the external magnetic field as a control parameter, T is temperature and T_C is the critical temperature for this system that marks the phase transition point [73,74,275,277]. Since the free energy is expected to be at a minimum at thermodynamic equilibrium, three different situations should be considered. The first one is when $A < 0$, in which the free energy will have two minima. The second case is when $A > 0$, hence the free energy is characterized by only one minimum taking place at the origin ($M = 0$), and finally the third case when $A = 0$ which will be a transition between the first and second case. Figure D.1.a depicts all the three conditions of the Landau free energy in the absence of an external field [73,74,275,277].

[9] This section has been published partially as supplementary material of Arbabi Moghadam S., Rezania V., and Tuszynski J. A., "Cell death and survival due to cytotoxic exposure modeled as a two-state Ising system", in Royal Society of Open Science (*R. Soc. Open Sci)*, **7**: 191578*,* doi: https://doi.org/10.1098/rsos.191578].

`

Solving the Landau model for equilibrium conditions gives the following results for magnetization, external field and the susceptibility, χ, near critical temperature

$$\frac{\partial F(M)}{\partial M}=0 \quad \text{and} \quad h=0, \quad M=\begin{cases} \pm\left(\frac{a(T_c-T)}{B}\right)^{1/2} & T<T_c \\ 0 & T>T_c \end{cases}, \qquad (\beta=\frac{1}{2})$$

$$\frac{\partial F(M)}{\partial M}=0 \quad (\text{Close to } T_c), \qquad h=BM^3, \qquad (\delta=3) \tag{D.2}$$

$$\chi^{-1}=\frac{\partial^2 F}{\partial M^2}>0, \qquad \chi^{-1}=\begin{cases} a(T-T_c) & T>T_c \\ -2a(T_c-T) & T<T_c \end{cases}, \qquad (\gamma=1)$$

So, close to T_c magnetic field and the magnetization follow $h \approx M^3$, in our model, T_c can be related to the EC50 value and we showed in figure D.1.b that the model is working close to T_c and Landau theory of phase transition and Ising model for cytotoxicity are correlating.

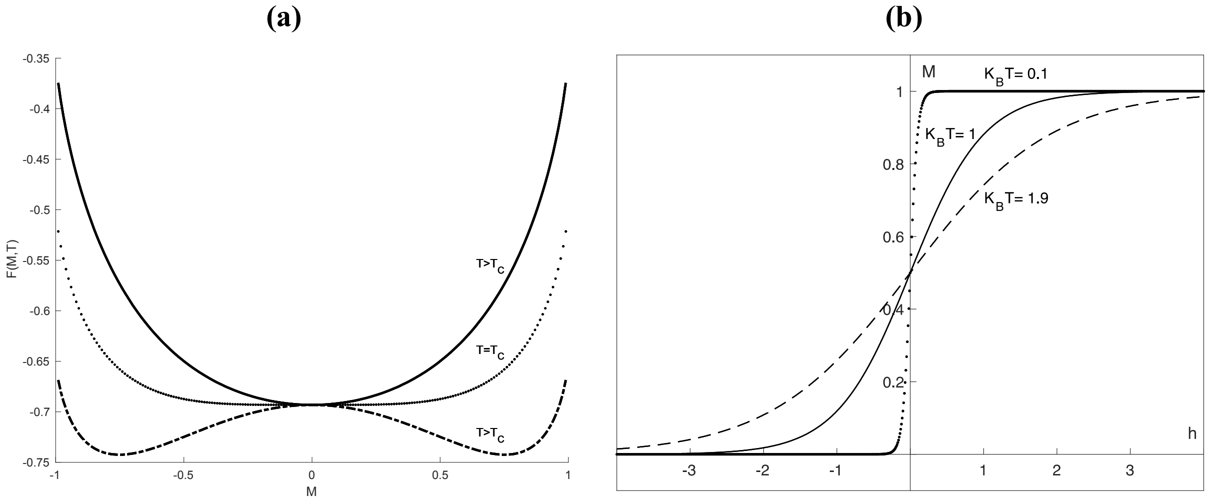


Figure D.1 (a) Free energy of Landau theory for different conditions in temperature, $(h=0)$, **(b)** Plot of magnetization as a function of external field when $(q=0)$ for different temperature limits $T<T_C$ $(k_BT=0.1)$, $T=T_C(k_BT=1)$ and $T>T_C$ $(k_BT=1.9)$.

`

D.2 Experimental data

All these 13 cytotoxic drugs have a defined principal molecular target; they belong to a specific cytotoxic classification and they are used against a particular cancer type in clinical trial experimental assays [286–290]. In addition, information about all the 66 cell lines has been obtained from the American Type Culture Collection (ATCC). The cell lines were cultured in ATCC-recommended media. Using the method of ATPlite 1step, the cell proliferation assays were extracted [83,84]. The exposure time was 72 hours for all compounds except for Vincristine and Tazemetostat where it was 120 hours. After 72 hours, inhibition of growth in the presence of these compounds was determined. The experiments were also carried out for the same cell lines without adding any compound but only by adding vehicle (DMSO) to the cells in order to provide controls for comparison. Between the two untreated profiles, growth inhibition is given for an increasing concentration of drug every 72 hours. Each experimental assay was repeated twice while increasing the concentration and the data between two untreated cells were measured four times. For most of the compounds the following quantities are reported: IC50 (note that the IC50 is the same as C_M in equation (5), i.e. the inhibition concentration where the response is reduced by half; GI50, the growth inhibition that denotes the drug concentration at which it causes 50% reduction in cancer cells growth; and LD50, the lethal dose that represents the amount of drug which kills 50% of a test sample. The maximum concentration tested for the compounds was 31.6 μM and no further increases in the concentration were made [83,84].

`

D.3 Data pre-processing

In the Table D.1 the experimental IC50 values of the studied cell lines, collecting by the collaborators in the Netherlands Translational Research Center B.V. (Oncolines), are listed [83,84]. The empty cells shows that the IC50 was higher than 31.6 μM and they were not measured experimentally. For analysing the data, the death rate is obtained using the cell survival rates, $R(C) = 1 - S(C)$. Using the cell response-Ising model, the death rates has been fitted with the Eq. (5.20). For obtaining the fitting parameters in the interacting and non-interacting cells, Python and MATLAB software have been used and the codes can be found in the Dryad repository [299].

Table D.1 The anti-cancer compounds used in experimental assays [286–290].

Name	Main target and Classification of cytostatic	Clinical trial use
Busulfan	DNA alkylating (Alkylating agents)	bone marrow transplantation, especially in chronic myelogenous leukemia (CML) and other leukemias, lymphomas, and myeloproliferative disorders
Methotrexate	folate synthesis (Antimetabolites)	breast cancer, leukemia, lymphoma, lung cancer, and osteosarcoma
Paclitaxel	Tubulin (Anti-microtubule agents-Taxanes)	ovarian cancer, breast cancer, Kaposi sarcoma, lung cancer, cervical cancer, and pancreatic cancer
Vincristine	Tubulin (Anti-microtubule agents-Vinca alkaloid)	acute lymphocytic leukemia, acute myeloid leukemia, Hodgkin's disease, neuroblastoma, and small cell lung cancer
Doxorubicin	topoisomerase II (Antitumor antibiotics)	breast cancer, bladder cancer, Kaposi's sarcoma, lymphoma, and acute lymphocytic leukemia
Cisplatin	DNA damage (Others/platin-like)	testicular cancer, ovarian cancer, breast cancer, bladder cancer, cervical cancer, head and neck cancer, esophageal cancer, lung cancer, mesothelioma, brain tumors and neuroblastoma

Irinotecan	topoisomerase I (Topoisomerase inhibitors)	Treat colon cancer and small cell lung cancer
Bortezomib	proteasome	multiple myeloma and mantle cell lymphoma
Tazemetostat	EZH2	lymphoma (non-Hodgkin lymphoma adult patients with certain genetically defined solid tumors, including INI1-negative tumors and synovial sarcoma, and patients with mesothelioma characterized by BAP loss of function)
Specific kinase inhibitors		
Afatinib	EGFR	non-small cell lung carcinoma (NSCLC) with common epidermal growth factor receptor (EGFR) mutation
Idelalisib	PI3K	hematological malignancies
Palbociclib	CDK 4/6	ER-positive and HER2-negative breast cancer developed
Trametinib	MEK	advanced malignant melanoma

D.4 Cell lines agreement to the Ising model

In Table D.2, the studied cell lines are listed for 13 drugs in an increasing order of the correlation coefficient with the Ising cytotoxicity model. Cell lines having good agreement with the model are shown in orange, filtering by $R^2 > 0.5$ and the susceptibility $\chi < 10$, while cell lines with poor agreement are shown in red labeled by a cross mark.

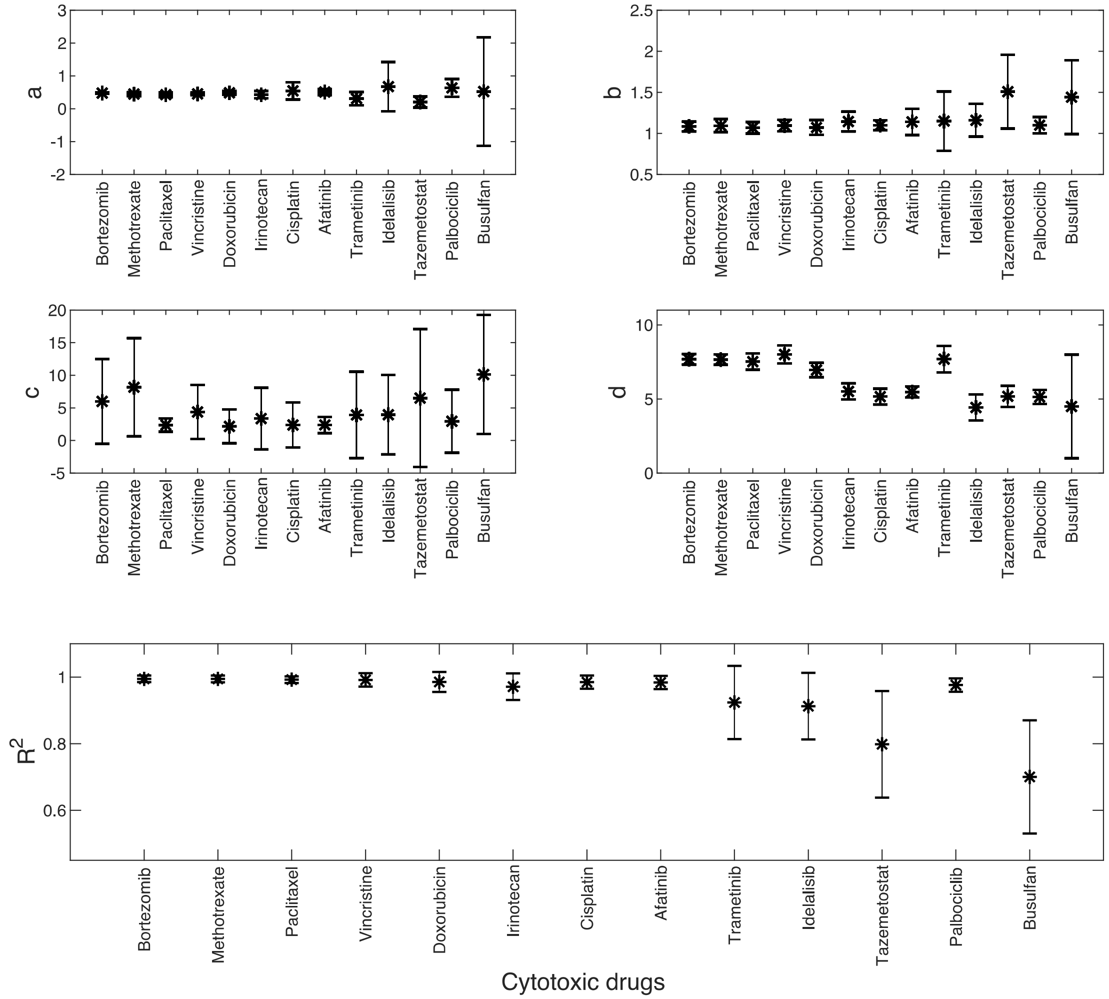


Figure D.2 Fitting parameters for equation (5.20), a, b, c, d and R^2 for the cytotoxic drugs tested. The average value of the parameter corresponds to each compound shown in black star. Error bars show the standard deviation to the average value for each case.

Table D.2 Behavior of the cell lines with respect to the $R^2 < 0.5$ and the susceptibility $\chi > 10$. The red cells represent the outlier to the model.

Cell line	Bortezomib	Vincristine	Doxorubicin	Methotrexate	Paclitaxel	Irinotecan	Cisplatin	Afatinib	Tremetinib	Idelalisib	Tazemetostat	Palbociclib	Busulfan	Cell line	Bortezomib	Vincristine	Doxorubicin	Methotrexate	Paclitaxel	Irinotecan	Cisplatin	Afatinib	Tremetinib	Idelalisib	Tazemetostat	Palbociclib	Busulfan
769-P											X			**MCF7**												X	
786-O									X	X	X			**C-33 A**								X		X	X		X
A-498											X			**DoTc2 4510**								X		X	X		X
A-704									X				X	**CAL 27**													X

`

ACHN								X			X			**COLO 205**											X		X
A-172										X			X	**DLD-1**									X				
CCRF-CEM											X			**HCT 116**													X
Jurkat E6.1									X					**HCT-15**											X		X
KU812										X		X		**LS 174T**													
SUP-T1									X		X	X	X	**LoVo**										X	X		
SR										X				**RKO**													X
MOLT-4														**SW48**												X	
K-562										X	X			**SW480**										X			X
A-204							X		X		X	X		**SW620**												X	
SJCRH30											X		X	**SW948**							X				X		X
A375										X	X			**SNU-C2B**								X		X			X
COLO 829											X	X	X	**T24**										X	X	X	
MeWo							X							**RT4**												X	
RPMI-7951												X	X	**J82**												X	X
A388								X					X	**Daoy**									X				
A-427												X		**U-87 MG**									X		X	X	X
A-549											X			**T98G**												X	X
NCI-H460														**SK-N-AS**											X	X	
SHP-77			X			X			X	X			X	**SK-N-FI**			X									X	
NCI-H82									X	X			X	**MG-63**												X	X
AN3 CA						X				X		X		**U-2 OS**									X	X			X
AsPC-1								X				X	X	**VA-ES-BJ**												X	
BxPC-3									X			X	X	**DU 145**								X	X				
MIA PaCa-2														**LNCaP FGC**									X			X	X
AU-565									X		X			**TT**							X					X	X
BT-20										X	X	X	X	**FaDu**								X		X		X	X
BT-549									X		X	X		**OVCAR-3**												X	
Hs 578T						X			X					**PA-1**									X	X	X		X

D.5 Interaction parameter, λ

Figure D.3 compares the fitting parameter averaged over all 66 cell lines, a, b, c, d and λ for each cytotoxic drug in two cases of interacting and non-interacting cells, respectively, followed by the average and error bars. It can be seen that the two cases are fairly consistent in the values of a, b, c, and d, and the important interaction parameter, λ, changes between 2.1 to 6.1 on average.

`

However, considering plots 5.1.a to c, and the large value of c, the cell-cell interactions are not significant.

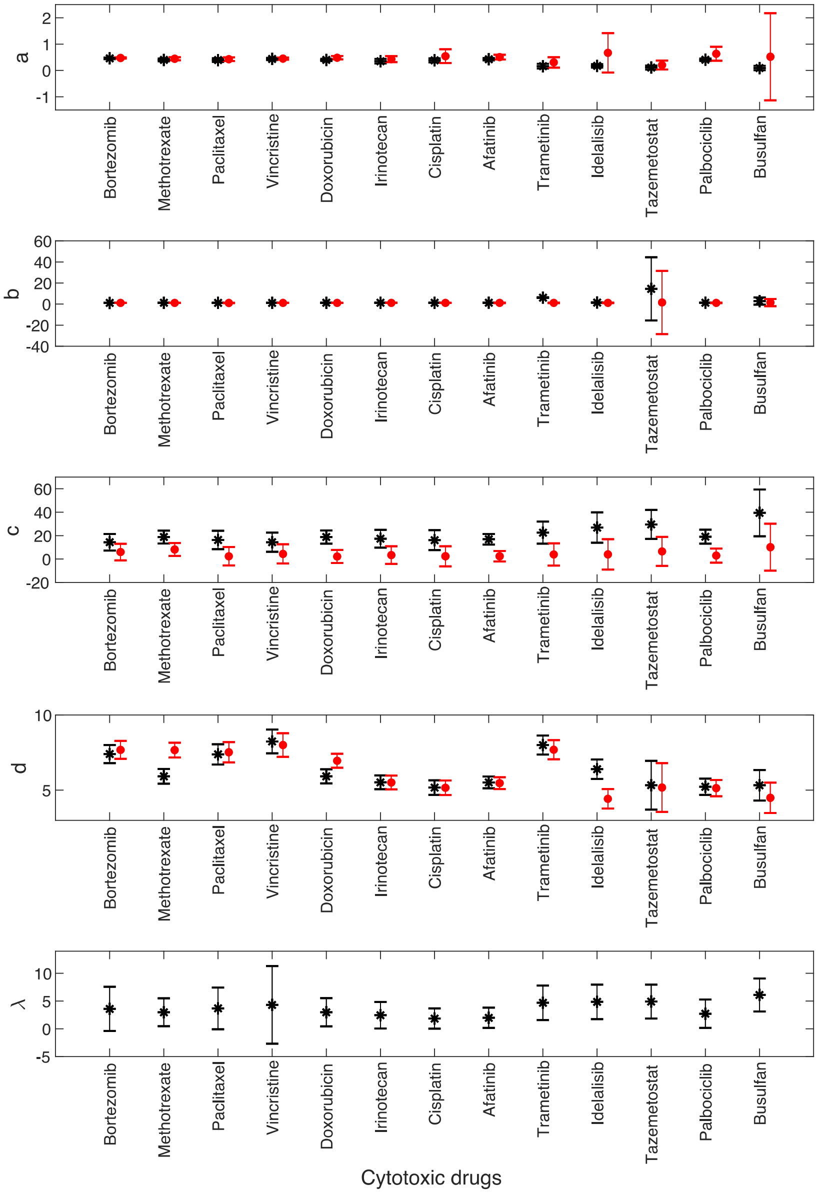


Figure D.3 Ising cytotoxic model's best-fit parameters for the cases with interactions using the equation $R = 0.5(1 + \tanh(1.15(\log(C) - \log(\mathrm{EC50})) + \lambda R)$ (in black) and the non-interacting cases described by the equation $R = 0.5(1 + \tanh(1.15(\log(C) - \log(\mathrm{EC50})))$ (in red) in all cancer cells exposed to chemotherapy drugs. The standard deviations from the average values are shown in each plot.