EVOLUTION OF DIMERIC PROTEIN INTERFACES AFTER GENE DUPLICATION

EVOLUTION OF DIMERIC PROTEINS AFTER GENE DUPLICATION

By ARMIN HODAEI, M.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Science

McMaster University MASTER OF SCIENCE (2020) Hamilton, Ontario (Physics)

TITLE: Evolution of dimeric proteins after gene duplication

AUTHOR: Armin Hodaei, M.Sc. (Koc University)

SUPERVISOR: Professor P. Higgs

NUMBER OF PAGES: xii, 62#pgs

LAY ABSTRACT:

A large fraction of proteins are found to exist as dimers composed to two identical subunits. If the gene for the single subunit is duplicated, three types of dimers can emerge, two homodimer structures and a heterodimer structure. Gene duplication is a major driving force of evolution as it can allow the proteins to perform new tasks. Here we define a model to understand the evolution of dimeric proteins as they undergo mutations in their interface, changing their stickiness to each other.

We find that evolution favours the dimers to either be homodimer or heterodimer, but not both at the same time. When there are two homodimers, one of them can acquire a new function (which is known as neofunctionalization). When there is a heterodimer, both genes are now required to do the orginal job of a single gene (which is known as subfunctionalization). These mechanism provide two possible reasons why the duplicate gene cannot subsequently be deleted from the genome.

ABSTRACT:

A significant number of proteins function as multimeric structures, most commonly as dimers. One of the primary mechanisms by which proteins evolve is through gene duplication and mutations of the resulting duplicated gene. The evolution of dimeric proteins after gene duplication is of interest because it can form three types of dimer: two homodimers and a heterodimer. Point mutations that occur in the interface of dimers would affect their binding strength and might change their path in the evolution.

Here we designed an evolutionary model for protein dimerization after gene duplication. In this work, we have used dimers' PDB structures to construct the network of contacts between amino acids in the interface. Several pairwise energy contact matrices were examined to find reasonable interface binding energies. Using the population genetics theory, we defined a selection criteria based on dimer interface strength and let them evolve as the mutations happen. We observed that the dimer structures are bound to be in the mostly homodimer state or mostly heterodimer state, and there are few occasions that we have all three types of structures as strong dimers.

We anticipate three fates for the dimer protein's evolution after gene duplication, neofunctionalization, subfunctionalization, and loss of the gene. A loss of function in homodimer structures might eventually lead to a subfunctionalization since the two interfaces are different. On the other hand, if a heterodimer loss happens, we would have two strong homodimer structures so both neofunctionalization and subfunctionalization might still happen. In the first case, one could gain a new function while the other homodimer performs the protein's old function. In the latter case, the two separate homodimers could each assume different parts of the full function of the original gene (which is the definition of subfunctionalization).

Acknowledgements

First and foremost, I would like to express my deepest appreciation to my advisor Dr. Paul Higgs, who has guided me throughout my studies at McMaster University. Weekly meetings and his valuable discussions have helped me a lot in gaining new insights. I am very grateful to him for being very patient and for all his time explaining the various subjects and going through the numerous versions of this dissertation.

I am grateful to the Physics and Astronomy department for providing the funding, which allowed me to undertake this research.

I wish to thank all the people whose assistance was a milestone in completing this project, especially Andrew and Vismay, for their countless pieces of advice and for sharing valuable experience.

Above all, I would like to thank my wife, Fatemeh, who has been extremely supportive of me throughout this journey and has made many sacrifices to help me get to this point. But most of all, thank you for being my best friend. I owe you everything.

Table of Contents

Chapter 1: Introduction **1**
1.1 Protein Structure 3
1.1.1 Multimer Proteins 4
1.1.2 Protein interfaces 4
1.2 Gene Duplications 9
1.2.1 Evolutionary fates after gene duplication 13
1.3 Duplication of a dimeric protein 15
1.3.1 Theory for fixation of a mutation in a population 17
1.3.2 Fixation probabilities 19
Chapter 2: Single gene evolution **21**
2.1 Methods 21
2.1.1 Single gene 21
2.2 Energy Models 23
2.3 Testing the free energy model with real interface data 25
2.4 Results 30
2.4.1 Interface propensities of amino acids 34
Chapter 3: Gene Duplication **38**
3.1 Results 40
3.2 Deletions from the duplicated state 46
3.3 Loss of function 47
3.4 Subfunctionalization 51
3.5 Neofunctionalization 52
Chapter 4: Conclusions **53**
Chapter 5: Appendix **56**
References: **58**

List of Figures and Tables

Figures

Figure 1.1 Levels of Protein structure 3

Figure 1.2 Homodimer vs. heterodimer structures. The structure on the left is a homodimer called Glucose-6-phosphate isomerase and the one on the right is Tubulin which is a heterodimer structure..... 5

Figure 1.3 Tetramer structures. The cyclic structure(Shaker Potassium Channel, PDB id: 1a68) has one four-fold rotational axis, and the dihedral structure(Electrophorus electricus acetylcholinesterase, PDB id: 1C2B) has three two-fold rotational axis 6

Figure 1.4 Surface amino acids are shown in a $\mathbf{4} \times \mathbf{4}$ array 7

Figure 1.5 The interface representation for a homodimer. The left figure, is the three dimensional interface of the 2PRZ dimer structure. Chains are in purple and green, the interfaces are in pink and yellow respectively. The right figure is the graph of the contact network for this structure. The interface is isologous, and all the interactions are appearing twice. For the homodimer cases we have the same residues in contact on both interface, this can be observed in the network graph 8

Figure 1.6 The interface representation for a heterodimer (pdb id: 3SGB) 9

Figure 1.7 Evolutionary fates of gene duplication 13

Figure 1.8 This is an example of Neofunctionalization after gene duplication in the Yeast. (pdb id: 1D6V)..... 15

Figure 1.9 Human Hemoglobin structure(PDB id: 1A3N). An example of heteromers... 16

Figure 1.10 Illustration of the spread of new mutations in a population. Each circle represents a gene copy and each row represents a generation. Every lower row represents the offspring of the upper generation. Bold lines show the lines of descent of genes in the current generation. This figure depicts how a sample mutation can get fixed in the population (Higgs & Attwood, 2005)..... 18

Figure 1.11 Graph of fixation probability in a population of $N = 200$ as a function of selection coefficient (Higgs & Attwood, 2005). There is advantageous and deleterious mutations; for small selection both mutations are acting like neutral mutations ($^1/_N = 0.005$) 20

Figure 2.1 The scatterplot of the data using the Miyazawa and Jernigan matrix..... 27

Figure 2.2 Relation of contact numbers with the binding free energy of heterodimer list of structures. Every dot represents a PDB structure 28

Figure 2.3 Relation of contact numbers with the binding free energy of homodimer structures from the Yeast. In the second figure, the amino acid/residue number in contact to the chain versus the binding free energy is illustrated. Every dot represents one PDB structure 29

Figure 2.4 Three dimensional Representation of 1SMS homodimer structure..... 30

Figure 2.5 Energy spectrum of the homodimer(1SMS) for different selection strengths. 31

Figure 2.6 Dimer fractions of homodimer 1SMS with different selection strengths..... 32

Figure 2.7 Amino acid propensities for the Yeast protein 3PYM(Homodimer)..... 35

Figure 2.8 Mean propensities for heterodimer structures..... 35

Figure 2.9 Mean propensities in ranking (heterodimer evolution)..... 36

Figure 2.10 Mean propensities in ranking using the Betancourt pairwise energy matrix. 37

Figure 3.1 Two monomers that can build three different dimers..... 38

Figure 3.2 Evolution of a random system by mutations with a strong selection strength of $\sigma = 0.1$ and a population of $N_e = 200$. Dii is the dimer fractions and $Dtot$ is the total dimer fraction. The first half is the evolution of single gene dimerization. The second half shows after alteration it reaches to mostly heterodimer state 41

Figure 3.3 Evolution of a random system by mutations with a strong selection strength of $\sigma = 0.1$ and a population of $N_e = 200$. The second half shows after alteration it reaches to mostly homodimer state 41

Figure 3.4 Change of dimer fractions with loss of homodimer binding interface after gene duplication. The purple curve shows the heterodimer fraction, and the x-axis is the energy of homodimer structure 42

Figure 3.5 Change of dimer fractions with loss of heterodimer binding interface after gene duplication 43

Figure 3.6 Three dimensional structure of 3PYM homodimer 44

Figure 3.7 Evolution of the 3PYM structure within a time span; with $N_e = 200$ and selection strength of $\sigma = 0.1$. We see that dimer fractions are alternating between mostly homodimer and mostly heterodimer state 44

Figure 3.8 The U-shaped distribution of a few homodimers from the Yeast. In here "Dxx" is the total homodimer fraction ($Dxx = D11 + D22$) 45

Figure 3.9 Total dimer fraction and heterodimer fractions of 3PYM structure 45

Figure 3.10 Dimer fractions of homo1 and homo22, when they lost function 48

Figure 3.11 Energy spectrum of each dimer when homodimers lose function 48

Figure 3.12 Heterodimer fraction when homodimers lose function 48

Figure 3.13 Energy spectrum of each dimer for loss of heterodimer function 49

Figure 3.14 Homodimer and heterodimer fraction when the heterodimer loses function 49

Figure 3.15 Energy spectrum of each dimer when homo22 and hetero12 loses function 50

Figure 3.16 Heterodimer fraction when homo22 and hetero12 loses function 50

Figure 4.1 Duplication pathway 53

Tables

Table 2.1 The cutoff vs error table for the two energy matrices. It is seen that the Miyazawa-Jernigan pairwise potential matrix shows better correlation 26

DECLARATION OF ACADEMIC ACHIEVEMENT

I, Armin Hodaei, am the sole author of this document and I declare this thesis to be my work. This thesis has not yet been published or submitted for publication or a higher degree at another institution. All simulations and results in this thesis were carried out by me personally, with the exception of the tests of correlation of the model free energies with the measured DDGs in figure 2.1 and Table 2.1, which were performed by Dr. Paul Higgs. To the best of my knowledge, this document's content does not infringe on anyone's copyright.

My supervisor, Dr. Paul Higgs, has provided guidance and support at all stages of this project. I completed all of the research work.

Chapter 1: Introduction

Life, as we know it today, depends on protein function. Evolution acts upon proteins through genetic mutations and natural selection. To conceive how it works, it is essential to know how proteins evolve. In a changing environment and under competition between organisms, proteins emerge with new functions and determine how successfully an individual can reproduce. It is now clear that one of the primary mechanisms by which proteins evolve is through gene duplication (Ohno et al.1970; Zhang et al. 2003; Crow et al. 2005), and many pairs of related genes are found in genomes that have arisen by duplication – these are known as paralogs (Van Zee et al. 2016). Proteins have a property to form complex structures consisting of two or more subunits (Levy et al.2013). Those with two subunits are called dimers. Homodimers, in which the two proteins are the same, constitute the majority of the dimer proteins (Marsh et al.2015). Mutations of amino acids at protein-protein interfaces are known to have large effects on human health because they affect protein complexes' formation. Many research groups are trying to find a theory to explain how dimer structures form and change as they undergo evolution. The evolution of dimeric proteins after gene duplicaton is of interest because if there are two gene copies there are three types of dimers - two homodimers and a heterodimer. When the two proteins are identical, all three types of dimers can form, but if mutations occur which decrease the binding strength of one or more of the dimer interfaces, then some of the dimers will cease to form.

Marchant et al. (2019) looked at dimeric proteins that have been duplicated in the genome of Yeast (saccharomyces cerevisiae). They have studied the whole genome duplications on the Yeast protein structures. They tracked the protein paralogs that emerged from gene duplication and counted how many structures have formed homodimer or heterodimer structures. They concluded that maintaining the homomers after gene duplication can indirectly lead to preserving the paralogous heterodimers. In a different study, Hochberg et al. (2018) analyzed the evolution of heat-shock proteins and pointed out that after gene duplication the two genes can create a mixture of structures selecting to self-assemble (homomers) or co-assemble (heteromers). They concluded that the homomers are more favorable, and the structures tend to self-assemble, creating homomeric paralogs.

In this work, we assume a simple model that allows both the physical and evolutionary aspects of protein dimerization to be addressed. This work connects the free energies of interaction of the protein surface residues to the evolution of protein dimers after gene duplication by using population genetics theory and subsequently maintaining sequences generated by mutation and natural selection. We define fitness to be dependent on the energies of the surface interactions. We consider a protein with one interacting surface. First, we use a monomer with an interacting interface, we let it construct dimers and evolve through mutations. Then, we start the gene duplication event having two possible isologous interfaces forming two homodimers and one heterologous interface which forms a heterodimer. Then we use a pairwise interaction energy table to calculate the surface binding energies between each amino acid. We investigate how the interface binding energies are connected to the evolution. To find an effective energy matrix, and have a better understanding of the protein interfaces and their role in evolution of dimeric protein structures.

1.1. Protein Structures

Proteins are molecular machines, building blocks, and arms of a living cell. Their importance arises from the remarkable diversity of their functional roles inside the living body. Proteins have structure and function, whereas most random peptides have no stable three-dimensional structure. In a protein, amino acids are connected by peptide bonds between the carboxyl group and amino groups of the adjacent amino acid residues. Every protein has a primary structure: the amino acid sequence, a secondary structure, and a tertiary and quaternary structure. The primary structure of every protein has polypeptide covalent bonds. The secondary structure consists of alpha-helices and beta-strands. The tertiary and quaternary structures are the three dimensional conformation of the protein. A quaternary structure is formed when several separate proteins aggregate to form a multimer. (Volkert et al. 2009).

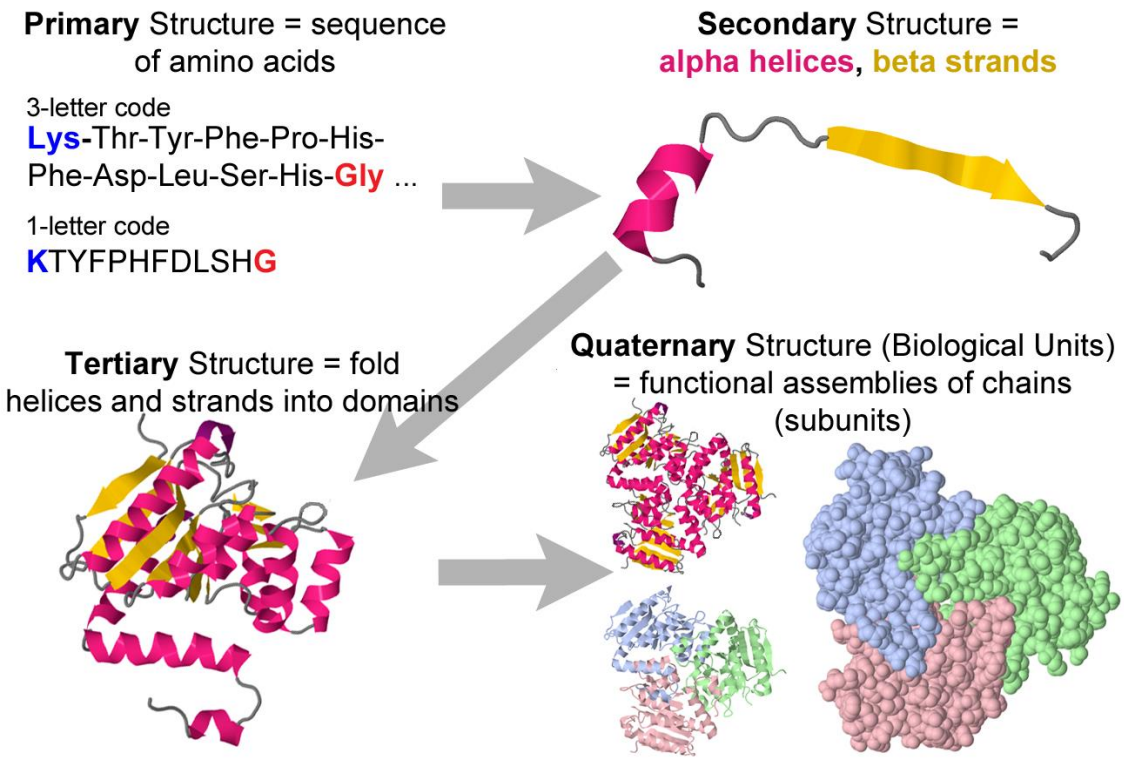


Fig. 1.1. Levels of protein structure (Volkert et al. 2009).

1.1.1. Multimer proteins

Observing different protein structures and utilizing many experiments that have been done over the past few decades, we understand that most of the proteins form symmetrical oligomeric complexes consisting of two or more identical subunits. Symmetry seems to be a critical ingredient in grasping the protein structures and functions. Most proteins function as symmetrical higher-order complexes consisting of subunits encoded by the same locus. According to numbers of subunits, these complexes are called homodimers, homotrimers, homotetramers, etc. Understanding the appearance of such liaisons is a critical issue in the field of evolutionary biology (Lynch 2013). The preference of proteins to self-associate is a feature well known to structural biologists. Analyzing protein surfaces demonstrates that they have statistically higher affinity for self-attraction in contrast to the propensity of attraction between different proteins. These statistical propensities are prone to produce self-self or similar interfaces with very low affinity. However, it can be reasonably presumed that any such interfaces that result in a functional advantage to an organism may evolve into higher affinity interfaces that settle specific oligomer formation. Homodimer interfaces have a higher degree of conservation in protein evolution than heterodimers (Lukatsky 2007).

Another critical factor that allows the evolvability of interactions in homomers is their symmetry. Even in a random pool of protein complexes with low energy binding, significantly symmetric interfaces are observed (Andre et al. 2008). Structural symmetry permits a single mutation to have a two-fold influence, thus a symmetrical, head-to-head interface is statistically more likely to appear (Lukatsky et al. 2007). Cyclic proteins have one rotational axis of symmetry. The proteins in this group specialize in functions that require directionality or sidedness, such as forming a hollow tube or chamber, therefore they have heterologous interfaces (i.e. head-to-tail). Dihedral proteins have an additional perpendicular axis of two-fold symmetry via isologous interfaces (i.e. head-to-head). The latter provides a greater variety of interfaces that leads to more stability. In the figure below, there are two examples of isologous (Glucose-6-phosphate isomerase protein) and heterologous (Tubulin protein) structures representing head-to-head and head-to-tail examples respectively.

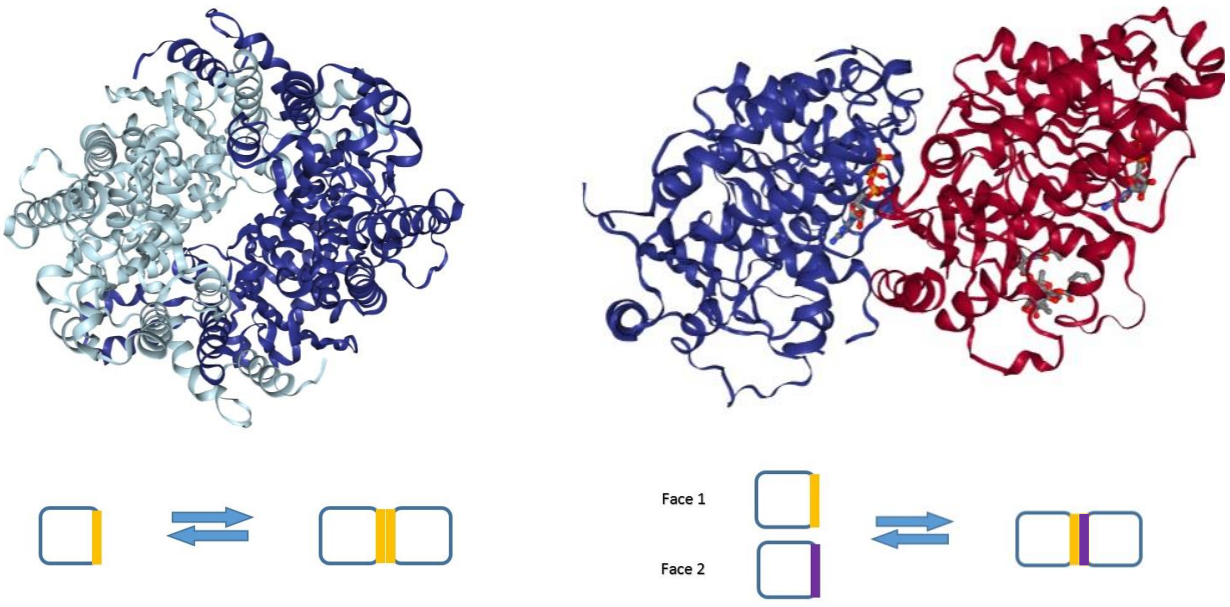


Fig. 1.2 Homodimer vs. heterodimer structures. The structure on the left is a homodimer called Glucose-6-phosphate isomerase and the one on the right is Tubulin which is a heterodimer structure.

A homomeric complex can have both types of interfaces, and many dihedral complexes can be described as stacks of cyclic complexes. Dihedral complexes mostly have isologous interfaces, but heterologous interfaces can also be present depending on the mode of assembly. For instance, a dimer of dimers structure will have both isologous and heterologous interfaces. Since symmetric interfaces are more likely to evolve than asymmetric ones in the first place and are selected for several functional reasons, dihedral complexes are more abundant than cyclic complexes (Nido et al. 2012). Some examples of dimer proteins are presented below; in the fig. 1.3 two homotetramer structures are also demonstrated. If we consider the proteins in cubic form, then there are two interacting interfaces for the tetramer structures. Therefore, in this figure it is shown how dihedral and cyclic tetramers could form.

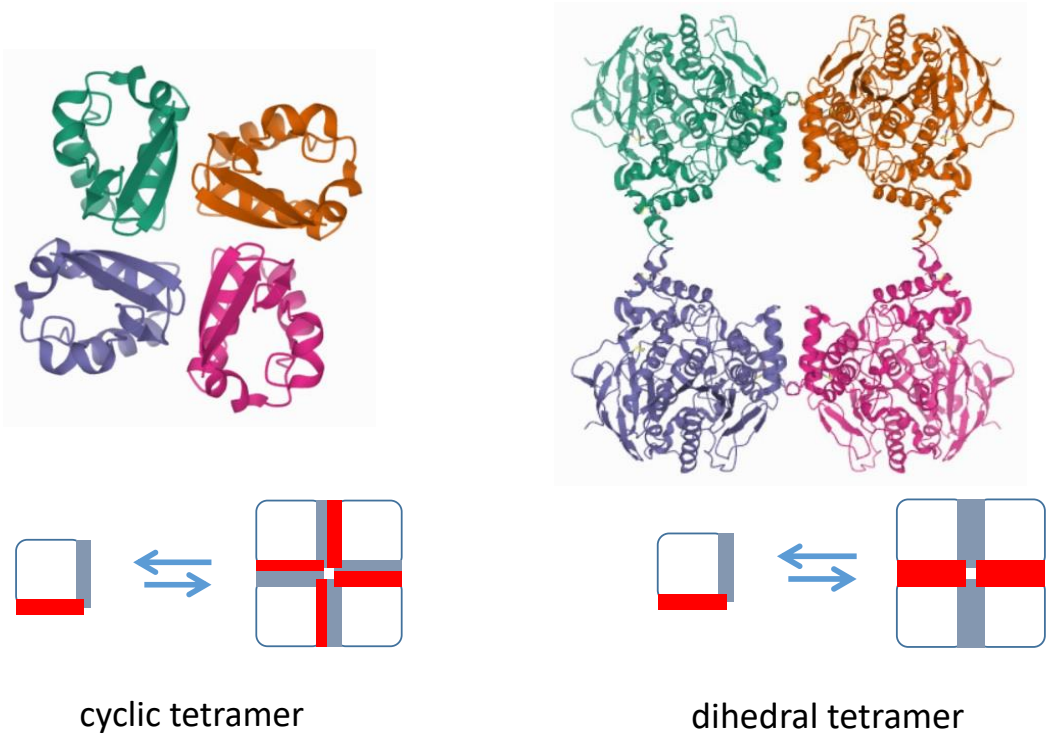


Fig. 1.3 Tetramer structures. The cyclic structure (Shaker Potassium Channel, PDB id: 1a68) has one four-fold rotational axis, and the dihedral structure (Electrophorus electricus acetylcholinesterase, PDB id: 1C2B) has three two-fold rotational axis.

1.1.2. Protein interfaces

In previous work, proteins were modeled as a cube, and the surface of proteins are assumed to be a 4×4 array of amino acids. In this model, to calculate the potential binding energies between two opposing faces (isologous interface, i.e., face A) of the protein we sum the 16 pairwise interactions of the surface amino acids (Fig. 1). Several different pairwise potentials were used in this analysis. The figure shows the square model used in prior work (Zabel et al. 2019).

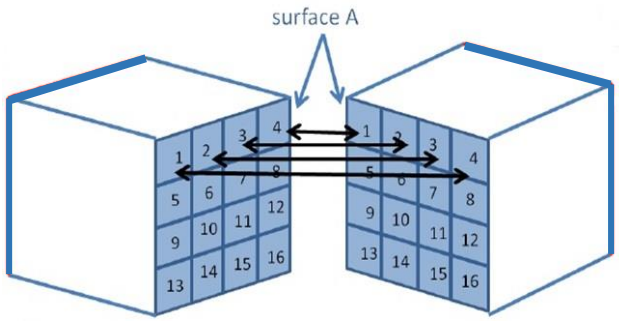


Fig. 1.4. Surface amino acids are shown in a 4×4 array.

The main idea of Zabel et al.'s work was to investigate the competition between forming closed dimers and open chains (fibrils) using the cubic model. They allowed proteins to aggregate as either closed dimers or open fibrils of all lengths.

However, in this thesis we did a comprehensive comparison using real PDB structures. We constructed the interfaces using the three dimensional coordinates of each protein. Then using the known free energies after a single mutation we found the best pairwise potential.

Interfaces are defined as the set of residues presenting a region through which two protein chains bind to each other through non-covalent interactions. This set consisted of contacting residues between the two chains (i.e., interacting residues). Two residues from the opposite chains were marked as interacting if there was at least a pair of atoms, one from each residue, at a distance smaller than the threshold distance. We consider a full atomistic analysis, consisting all the atomic distances to identify residues in contact.

For instance, one of the Yeast homodimers ("S. cerevisiae orotate phosphoribosyltransferase" with the pdb id of 2PRZ) in our analysis is represented in Fig.1.5. In the right figure, there is the network of the contacts on the interface. The colour scheme is red for attractive, blue for repulsive, and grey for in the middle.

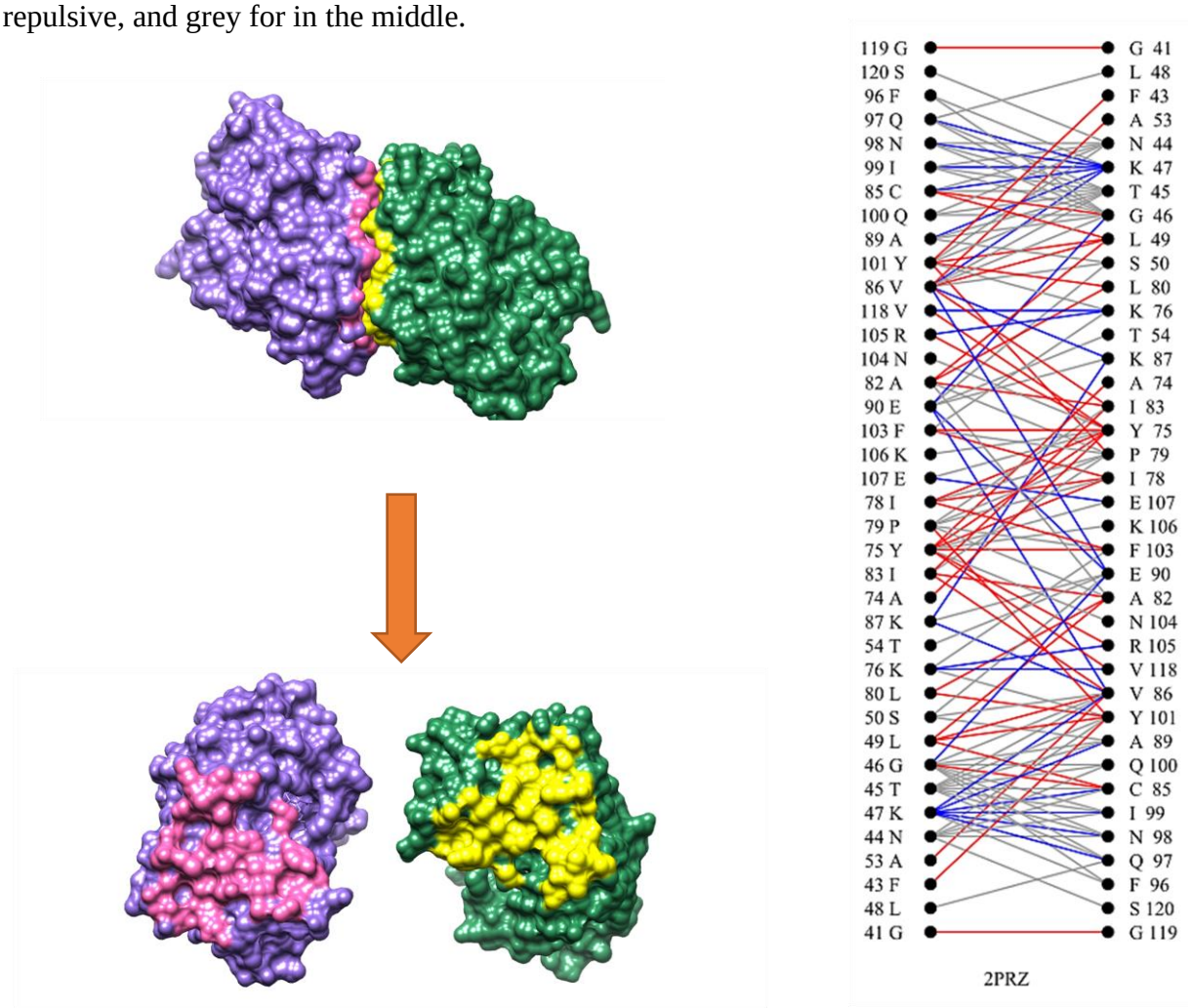


Fig. 1.5. The interface representation for a homodimer. The left figure, is the three dimensional interface of the 2PRZ dimer structure. Chains are in purple and green, the interfaces are in pink and yellow respectively. The right figure is the graph of the contact network for this structure. The interface is isologous, and all the interactions are appearing twice. For the homodimer cases we have the same residues in contact on both interface, this can be observed in the network graph.

As another example, below, is the three dimensional picture of a heterodimer ("Streptomyces griseus protease" with pdb id of 3SGB) used in our simulations.

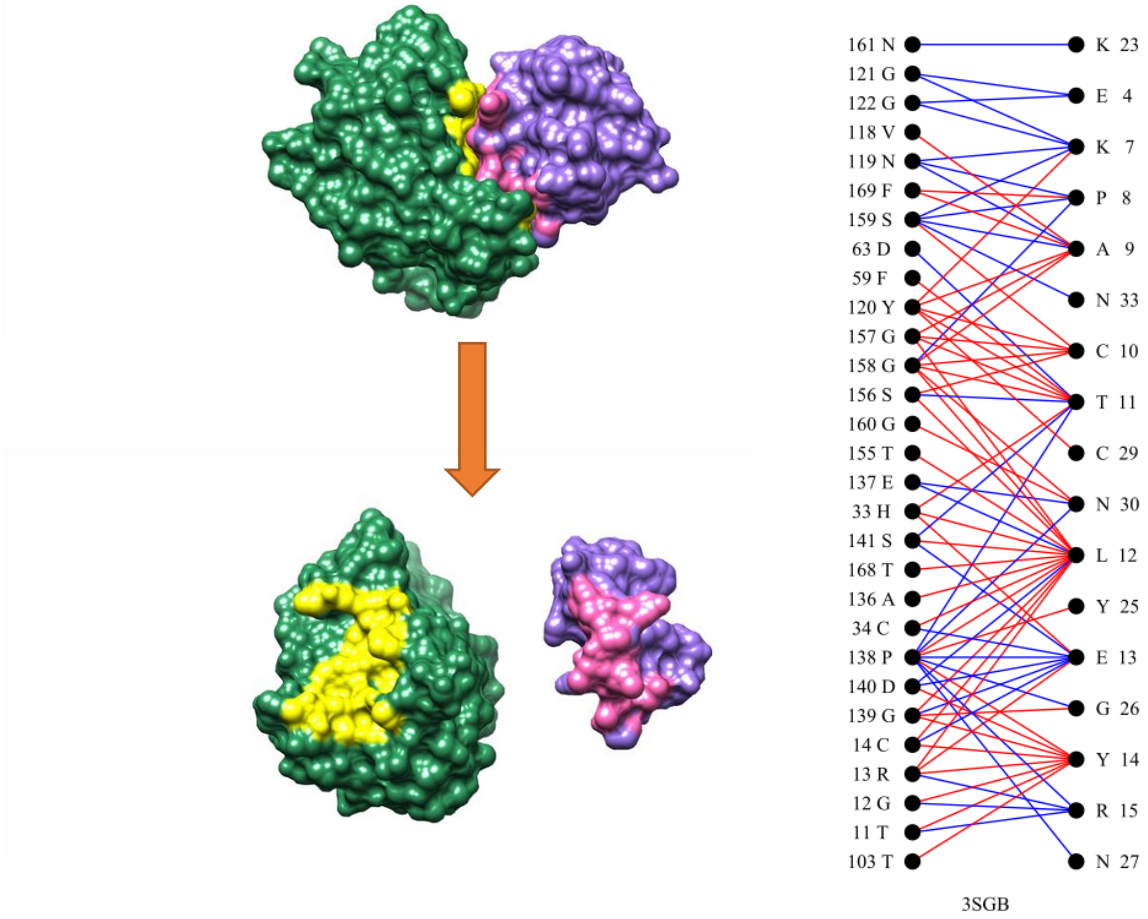


Fig. 1.6. The interface representation for a heterodimer (pdb id: 3SGB).

The left figure, is the three dimensional interface of the 3SGB dimer structure. Chains are in purple and green, the interfaces are in pink and yellow respectively. The right figure, is the graph of the contact network for this structure. These are the contacts used in our simulations.

Having the real structures for each interface on the proteins enables us to have a more realistic model to count the exact number of the amino acids on two chains in contact. Therefore, we could determine the interfacial binding energy of two chains using the pairwise energy matrices. The energy matrices help us to measure the interaction strength of two chains sticking together.

We are also interested in investigating the frequencies of amino acid in the interface. It is well known that there are differences in the distribution of amino acids between the interior and exterior of a protein (Chakrabarti 2002). Furthermore, studies revealed that the frequencies of amino acids in a protein-protein interface are different from that of the rest of the protein surface (Jones 1996, Bordner 2005).

In our analysis, we calculate the interface energies using the contacts on the interface, so we would be able to determine the relation between the number of contacts on each residue with the binding energies. It is known that the number of interface residues is proportional to the interface area (Chakrabarti et al. 2002, Brinda et al. 2002). Furthermore, it was observed that stronger protein subunit associations were generally associated with larger interface areas (Jones et al. 1995).

1.2. Gene Duplication

Evolution through the duplication of genes was postulated to be an essential process facilitating the change in function and creating diversity in the organisms long before entering the genome sequencing era (Roth et al. 2007). Gene duplication is a mutational event that occurs in a single individual within a population. Gene duplication can happen by several mechanisms, including whole-genome duplication (WGD) and single gene duplication. WGD (polyploidization) duplicates all of an organism's genes at once and generates a considerable duplicated genes. Ongoing research on this topic spans scientific subjects from bioinformatics to organismal biology and is related to different aspects of gene duplication, ranging from molecular mechanics of the duplication to the duplicate genes' fate.

Gene duplication creates a redundant gene copy and therefore releases one or both copies from the negative selective pressure. There are several different models for maintaining the duplicate pair and these make different evolutionary fates for the duplicate pair. Among all theories on how gene duplications evolve and finally become fixed is a combination of ideas that support their argument on the concept of functional redundancy (Fisher, 1935; Ohno, 1970; Kimura and King, 1979; Wagner, 1998; Force et al., 1999; Lynch and Force, 2000). The main idea of the functional redundancy of gene duplications can be explained in the following framework. Assume that a single gene does all the possible functions that are required. In this instance, by duplicating this gene, it cannot gain any extra functionality, and it must be redundant from selective and functional pressure. This assumption asserts that the individual's fitness with one original copy of the gene is precisely equal to an individual with two or more copies.

The primary definition of the functional redundancy (Ohno, 1970) suggested that natural selection cannot differentiate between the ancestral gene copy and the new gene copy. We are assuming that there is no change in the protein expression level because of the addition of a new gene copy and that there is no cost to the organism of having extra DNA. This new copy was free to wander in genotype space, which could undergo random mutations and acquire a new function. Such a random acquisition of a new function is called neofunctionalization (Li, 1997; Hughes, 1999; Force et al., 1999). The opposite of this process is the loss of new gene copies to degenerate mutations that do not decrease fitness but are harmful to this gene copy. Loss of a functional copy of the gene, which is called nonfunctionalization (Li, 1997; Hughes, 1999; Force et al., 1999), is the anticipated outcome of neutral fixation of degenerate mutations, and as the duplicated gene copies diverge its probability will increase (Walsh, 1995; Wagner, 1998).

1.2.1. Evolutionary fates after gene duplication

When whole genome data became available, it became evident that the number of gene duplications in genomes is more than the number we can rationally anticipate to be retained by acquiring new functions (Force et al., 1999; Lynch and Conery, 2000). Consequently, there emerged a necessity to explain how a large number of gene duplicates could be retained by natural selection. Claiming that natural selection is not able to differentiate between the old and the new gene copy, Force (1999) altered the premise of genetic redundancy of gene duplicates by suggesting that at the moment of duplication mutations in both copies of the gene are neutral since the other gene is still performing the function.

This adjustment resulted in the proposal of a new term called subfunctionalization, a process by which redundant gene copies can evolve to be retained by selection. Subfunctionalization happens when both gene copies undergo slightly degenerate mutations, mutations that are neutral at the time of their occurrence but are detrimental to the function of one of the gene copies. Figure 1.7 illustrates the three common fates of the gene duplication.

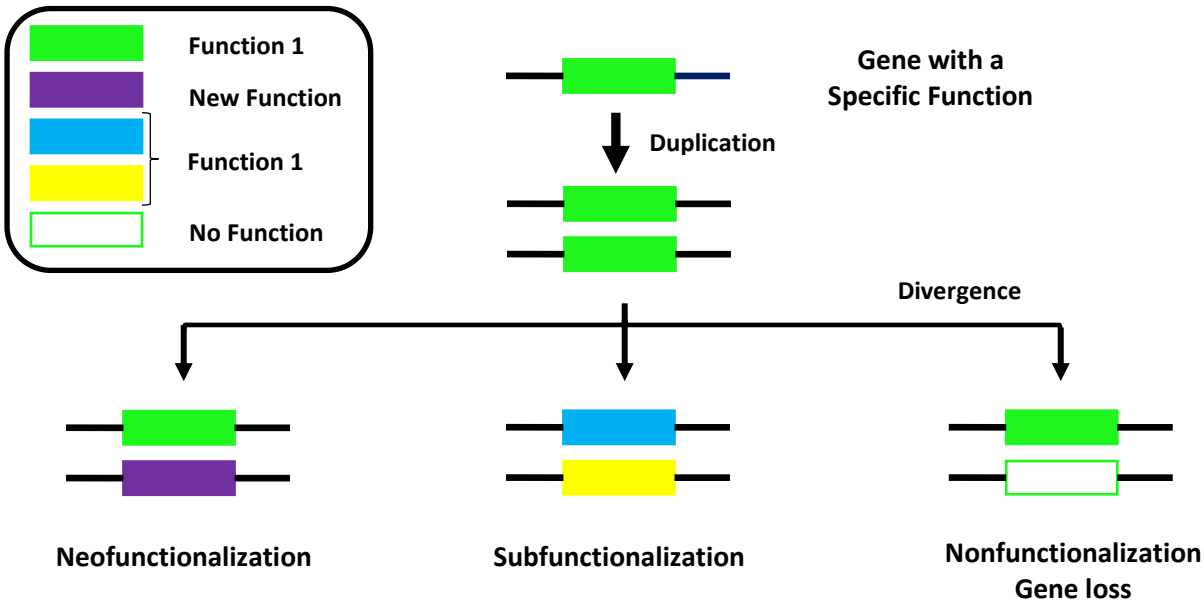


Fig. 1.7. Evolutionary fates of gene duplication.

If both the new and the old copies accumulate adequate complementary degenerate mutations, then both copies will be essential to fulfill the entirety of the original function and to be retained by natural selection (Force et al., 1999; Lynch and Force, 2000). Initial gene duplication events occur in a single individual. The population size and the degree of selection can vary the fixation rate in a population that has undergone gene duplication. It can be treated as neutral selection (Force et al., 1999; Lynch and Force, 2000) or with a positive selection (Perry et al., 2007); in this paper, they have generated a genome-wide map of copy number variation in humans and chimpanzees. Utilizing population genetic analyses for use with copy number data, they spotted functional categories of genes that have probably evolved under positive selection for copy number changes. They concluded that genes' duplications and deletions might have been fixed by positive selection and involved in humans and chimpanzees' adaptive phenotypic differentiation. Different lineage in the tree of life demonstrates different propensities to tolerate gene duplication; for instance, mammals of small effective population size differ from plants of the same effective population size. Even in a specific species, different genes and gene functions can be retained in a completely different way after duplication (Hanada et al., 2008), even though this paper does not yet clarify the role of selection in the initial duplication event. In reality, these events happen in a single individual and are dependent on population-level processes simultaneous to divergence. If the addition of a gene is neutral, the probability that it goes to fixation in the population is equal to $1/N_e$, where "N_e" is the effective population size. If the gene addition is advantageous or deleterious, the probability of fixation is higher or lower than this.

This means that selective processes are standard in the organism of larger effective population size. The mutation rate can also alter the process of evolution; a higher mutation rate would provide a more significant sampling of changes to access those of adaptive effect.

1.3. Duplication of a dimeric protein

The duplication of genes would create new complexes by a single duplication event. As a result of duplication, homomeric proteins will lead to the formation of homodimers and heterodimers of paralogs. Some paralogs form homomers and have lost the ability to form heteromers through evolution, for instance, duplicated histidine kinases (Ashenberg et al., 2011) and many heat-shock proteins (Hochberg et al., 2018). Hochberg's paper illustrates that many proteins linked by gene duplication of an oligomeric ancestor have evolved into a strong homomers and the heteromer interfaces were weaker, which corresponds with gaining distinct functions. They showed how two oligomeric small heat-shock protein paralogs avoid heteromerization.

We are eager to know how the evolutionary forces act on the two genes after a gene duplication event and whether that would lead to the emergence of paralogous proteins or obligate heterodimers. Forming paralogous homodimers may be considered as an example of neofunctionalization, and forming heterodimers may be considered as an example of subfunctionalization. To split the ancestor's function, we need to have hetero structures; this means each chain should be different to perform separate functions. Having that in mind, for a protein to gain a new function besides the ancestor's function, we need to have the homodimer since one gene can still perform its old function, and the other could undergo mutations and gain a new function. In Figure 1.8, there is an example of neofunctionalization after gene duplication, the homodimer shown is a paralog structure in Saccharomyces cerevisiae (Yeast); and the paralogous genes are TSA1 and TSA2 (stands for Thiol-specific Antioxidant).

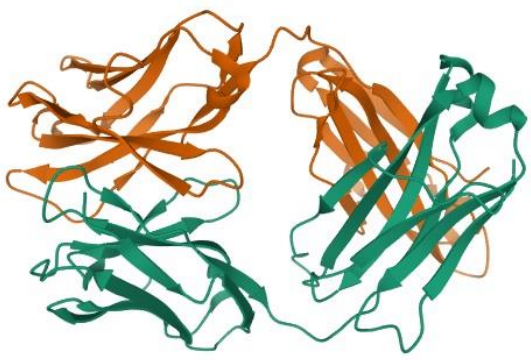


Fig. 1.8. This is an example of Neofunctionalization after gene duplication in the Yeast. (pdb id: 1D6V).

Human hemoglobin provides a good example of subfunctionalization after gene duplication. For instance, the hemoglobin alpha and beta proteins shown in fig 1.9 are paralogs. The gene for hemoglobin α-chain is derived from a duplicate copy of hemoglobin β-chain.

The aim of this thesis is to consider the evolution of two duplicate genes that code for dimeric proteins. After the duplication there are two identical genes A and B and three types of dimeric proteins AA, BB and AB. As mutations accumulate in the two genes, the strengths of the interfaces in the three dimers will change. We will show that it is likely that the genes will diverge in such a way that either the two homodimers remain and the AB dimers are lost, or the heterodimer will remain and the homodimers will disappear. We would like to know how often these two alternatives occur.

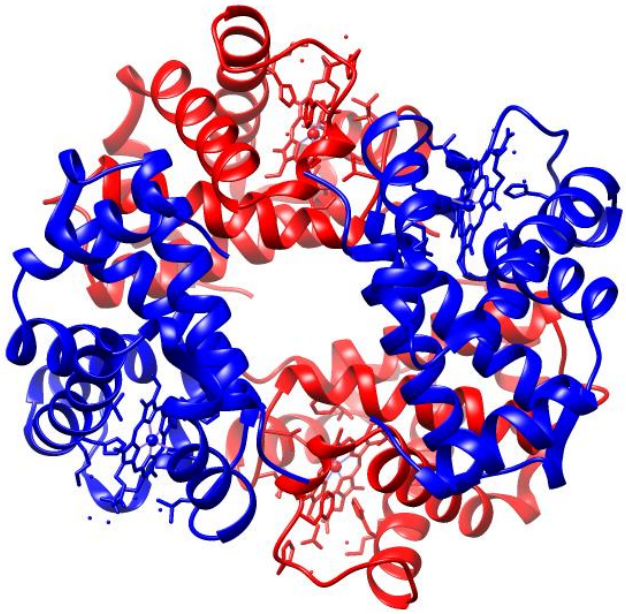


Fig. 1.9. Human Hemoglobin structure(PDB id: 1A3N). An example of heteromers.

1.3.1. Theory for fixation of a mutation in a population

In this thesis we will carry out simulations of evolution of proteins that form dimers. We will consider point mutations that occur in the interface region of the protein structure. These mutations affect the free energies of the interfaces and hence alter the probabilities of formation of the homodimers and heterodimers. The fitness change resulting from the change in the dimer concentrations will be described in the next chapter. In this section, we consider the general population genetics theory for fixation of a mutation. A mutation is a change in a gene sequence that can be passed on to offspring. Mutations might occur as a consequence of an error in replication or damage to the DNA or RNA molecule. The most basic mutation is a point mutation, where another base of a different type replaces a single nucleotide base in the DNA chain. A point mutation may occur anywhere over the length of the DNA (or RNA) sequence. When a mutation happens within a protein-coding region, it can change the amino acid sequence of the protein that will be synthesized from the gene. Since there is redundancy in the genetic code, not every mutation can cause a change in the protein sequence such mutations are termed silent mutations. Some mutations in a codon can lead to a change in the amino acid produced, which could potentially lead to a change in the function of the protein. However, most functional mutations are deleterious, i.e., their fitness would be less compared to the original sequence. Other mutations may have little effect on the gene's function that would have no apparent consequences for the individual. However, small selective effects are significant on the evolutionary time scale. Natural selection, acting over many generations, tends to keep the frequencies of slightly deleterious mutations to low levels. Indeed, not all mutations are deleterious. Some will be neutral, and there must be occasional advantageous mutations that increase the fitness of the sequence.

If there were no advantageous mutations, then the gene sequence could never have evolved to be functional in the first place. Advantageous mutations are thought to be rare, so the majority of the mutations in a population are deleterious. Gene sequences in a population are related to one another by descent from common ancestors. In a current population, if we follow lines of descent of a gene from two individuals, we will see that by tracing back in time, they will coalesce; this means that they had a common ancestor. In a population size of N, the typical time back to the coalescence point will be of the order N generations. After a new mutation takes place in a population, there will initially be only one copy. Selection and drift will effectively act and increase/decrease the number of these new copies in the population. The majority of the new mutations will be eliminated from the population as there is a substantial probability of them being lost by chance since the copy number is small, even though they are selectively advantageous. Once in a while, a mutation will become fixed in the population, this means the mutation will take over the population by reaching a high frequency in the population.

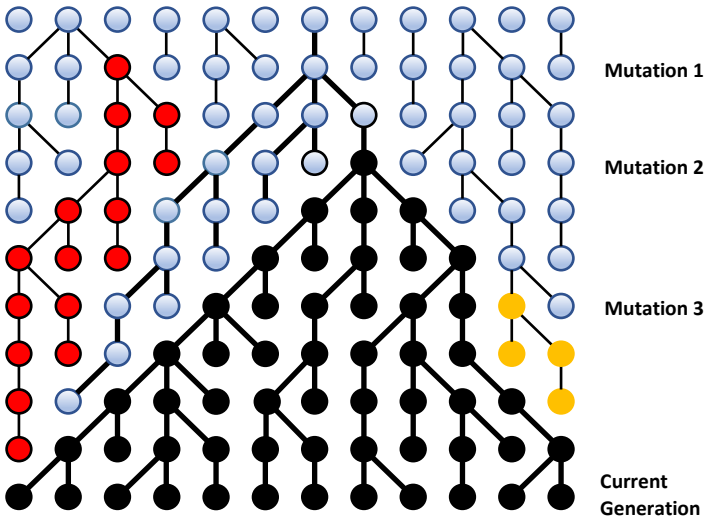


Fig. 1.10. Illustration of the spread of new mutations in a population. Each circle represents a gene copy and each row represents a generation. Every lower row represents the offspring of the upper generation. Bold lines show the lines of descent of genes in the current generation. This figure depicts how a sample mutation can get fixed in the population (Higgs & Attwood, 2005).

Figure 1.11 demonstrates how new mutations occur and might spread through the population until it gets fixed. It can be seen that three different mutations happen. Finally, the black circle generation took over the population and got fixed. We recognize that most mutations will not be fixed, no matter if they are advantageous or deleterious; however, occasionally, some of the mutations can be fixed.

1.3.2. Fixation probabilities

The probability that a mutation will eventually be the ancestor of all other individuals in a population is called fixation probability. In this thesis we used Kimura's fixation probability, which can be written as:

The old fitness of the gene: $w_{old} = 1 + s_1$

The new fitness of the gene: $w_{new} = 1 + s_2$

The difference of the two fitness will be: $\Delta s = s_2 - s_1$

$$P_{fix}(\Delta s) = \frac{1 - e^{-2\Delta s}}{1 - e^{-2N\Delta s}}$$

Where "Δs" is the change in fitness, and "N" is the population size. Using the equation for a neutral mutation, the probability of becoming fixed is $1/N$. This probability for an advantageous mutation with the fitness $1 + \Delta s$ has a more substantial probability of becoming fixed if $\Delta s > 1/N$. On the other hand, for a deleterious mutation, there is a tiny chance of becoming fixed if $\Delta s < 1/N$. In the case of $\Delta s \ll 1$:

$$P_{fix}(\Delta s) \sim \frac{2\Delta s}{(1 - e^{-2N\Delta s})}$$

Both advantageous and deleterious mutations are classified as nearly neutral cases. This implies that random drift is more significant than the selection in determining their fate, and their probability of fixation is too close to that of a neutral mutation. We can also find the average time for a successful mutant to go from a single individual to becoming fixed in the population. The mean time required for this single allele to take over the population is $< t >= 2N$.

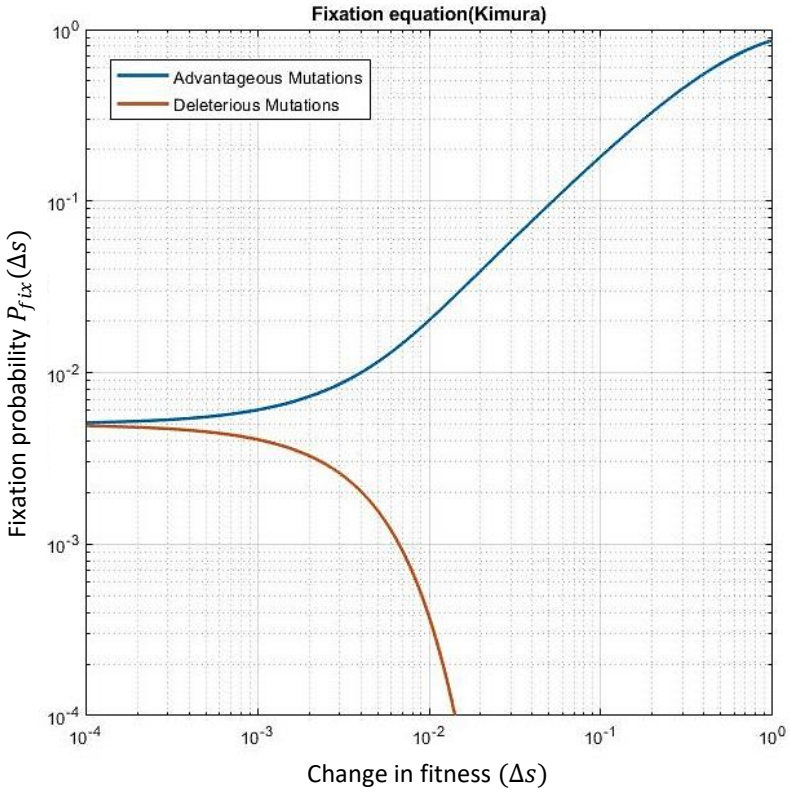


Fig. 1.11. Graph of Kimura's fixation probability in a population of $N = 200$ as a function of selection coefficient (Higgs & Attwood, 2005). There is advantageous and deleterious mutations; for small selection both mutations are acting like neutral mutations ($^{1}/_{N} = 0.005$)

We have assumed that the population is dominated by one sequence variant at a time – i.e. we ignore polymorphisms. The rate at which a population switches from sequence i to sequence j will be:

$$r_{ij} = uP_{fix}(s_i - s_j)$$

where u is the rate of occurrence of the mutation and P_{fix} is the probability that it spreads to fixation. We want to know the probability of finding a population in any sequence i after a long simulation. In the steady state, for any two sequences i and j we must have:

$$p_i u P_{fix}(s_j - s_i) = p_j u P_{fix}(s_i - s_j) \quad \rightarrow \quad \frac{p_i}{p_j} = e^{2N_e(s_i - s_j)}$$

Since this applies for every i and j, it follows that

$$p_i = constant \times e^{2N_e(s_i)}$$

Another way to get the same steady state distribution is to use the Metropolis method. In this case, we accept or reject each proposed mutation with a probability:

$P_{acc} = 1$, if $s_2 - s_1 > 0$

$P_{acc} = e^{2N_e(s_2 - s_1)}$, if $s_2 - s_1 < 0$

We obtain

$$\frac{p_i}{p_j} = \frac{uP_{acc}(s_2 - s_1)}{uP_{acc}(s_1 - s_2)} = e^{2N_e 2(s_1 - s_2)} \rightarrow p_i = constant \times e^{2N_e(s_i)}$$

the same as with the Kimura fixation probabilities.

Using the metropolis method allows a much more rapid simulation because a majority of mutations are accepted in this case, whereas in the Kimura case only roughly 1/Ne mutations are accepted. The metropolis method gives the correct steady state distrbution but not the timescale of the dynamics.

Chapter 2: Single gene evolution

2.1. Methods

In the earlier chapter, I showed how we used the real three-dimensional protein structures to define the interfaces and find the contacts. To do this, we built the contact matrix for each protein complex, in our case, dimers. We designed a method to define which amino acids are in contact; we specified a threshold and counted the number of contacts in the interfaces. In this method, every amino acid can have multiple contacts with other amino acids on the other chain of interaction. Therefore, with only a single mutation we would have several changes in the pairwise energies having a more significant effect on the whole binding energy of two interfaces (Bonvin et al. 2005).

2.1.1 Single gene

The fitness of the gene depends on the dimer concentration, which is calculated in the following way. The concentration of dimers depends on the total concentration of the protein produced from the gene, which we define as 2ϕ.

$$2\phi = c + 2d$$

where c is monomer concentration and d is dimer concentration, both at equilibrium. The maximum dimer concentration for strongly interacting dimers is ϕ.

The free energy $\Delta G$of the interface is calculated using the pairwise contacts between the amino acids in the interface, as explained in the following section. The dissociation constant for dimer formation is

$$K = \exp\left(\frac{\Delta G}{RT}\right)$$

At equilibrium $d = c^2/K$, therefore

$$2\phi = c + 2c^2/K$$

Solving this, we obtain:

$$c = \frac{K}{4}\left(-1 + \sqrt{1 + 16\phi/K}\right)$$

$$d = \frac{K}{8}\left(1 + 8\phi/K - \sqrt{1 + 16\phi/K}\right)$$

and the dimer fraction is:

$$D_{sing} = \frac{d}{\phi}$$

The subscript *sing* emphasizes that this is the single gene case. Suppose the fitness depends linearly on the dimer fraction

$$w_{sing} = 1 + \sigma D_{sing}$$

When a mutation occurs, the dimer fraction changes. The change in fitness is

$$\Delta s = \sigma \Delta D_{sing}$$

When then accept or reject the proposed mutation with either the Kimura fixation probability or the metropolis acceptance probability, according to which method is being used.We are selecting for higher dimer fractions, using the Kimura's relations we decide whether the mutations are accepted or not.

2.2. Energy Models

To investigate the residue-residue energies, we compared several different pairwise contact energy matrices. These matrices provide the interaction energies between any pair of amino acids in each binding surface on the protein chains.

An assumption underlying the estimation of contact energies is that, for a large sample, the effects of particular amino acid sequences will average out. Thus the numbers of residue-residue contacts seen in a high number of protein crystals will depend on the mean values of the inter-residue contact energies. This hypothesis is consistent with the ‘‘principle of structural consistency’’ that was initially proposed by Go (1983) and also called the ‘‘principle of minimal frustration’’ for the energy landscape view of proteins (Bryngelson & Wolynes 1987) because the present assumption is similar to the assumption that on average the inherent contact interactions are those compatible with the high stability of native structures. This assumption is also equal to the assumption that the distribution of the numbers of contacts in protein structures is a ‘‘self-averaging property’’ (Bryngelson et al. 1995);

Boltzmann-like statistics observed in protein structures are a general feature of the stable structures of heteropolymer chains and that the ‘‘temperature’’ in these statistics is not the usual temperature of the medium but a ‘‘selective temperature’’, at which the native structure is ‘‘frozen out’’ from an extensive set of structures (Gutin et al. 1992).

Not many computational resources are needed to calculate the statistical potentials. However, several conceptual difficulties remain. The probability that two residues A and B are within a cut-off distance r_{cut} is called $P_{contact}(A,B)$, which is computed from a carefully selected set of structures.The likelihood is compared with the reference distribution, $P(A)P(B)$, the probabilities observing both A and B if they were independent. A "mean force" pair potential is then defined, which is also called a statistical potential:

$$V_{AB} = -k_B T \ln[P(A,B)/(P(A)P(B))]$$

To correctly estimate these probabilities, the structures that are used to compute the frequencies of the contacts must be enumerated. For instance, they should not include any homologous proteins. The presence of highly similar proteins may overcount some types of contacts. Moreover, misfolded structures cannot be used to extract the correct frequencies because their weight in the computations of the distances is unknown (Tobi et. al. 2000).

It is evident that a proper pairwise potential is an essential ingredient of an attempt to find protein interfaces. There is a need for a comparison between pairwise potentials, which were extracted using different methods. It is critical to know how different the potentials are and what information is considered necessary to make a reasonable pairwise potential.

In this regard, Miyazawa and Jernigen (Miyazawa & Jernigen, 1996) constructed an effective pair potential matrix for amino acid contacts using globular protein structures. This matrix has been used in many coarse-grained simulations (Kim & Hummer, 2008.Dignon et al. 2018). Betancourt and Thirumalai (1999) showed that by shifting the elements of Miyazawa's matrix relative to threonine, it could be accounted for interactions between a solvent and amino acid. This transformed matrix, unlike the original matrix, which just had negative values, has positive and negative elements, which better shows the interactions of hydrophobic amino acids.

2.3. Testing the free energy model with real interface data

In order to determine how well the model for the interface can predict the free energies of real protein interfaces, we made use of the data of Xiong et al (2017), which tabulates the change in free energy $\Delta\Delta G$ for the binding of protein dimers when a mutation is made in an interface residue. We took 12 well-studied protein dimers for which free energy data for at least 20 single-point mutations were available. We considered only single-point mutations from these data. The pdb files for the original protein structures and the numbers of mutations available in each structure are: 3sgb (190), 1ppf (171), 1r0r (152), 1a22 (58), 1jtg (34), 1emv (32), 1brs (28), 1a4y (27), 1iar (25), 2g2u (24), 2wpt (23), 1ktz (20), making a total of 784 mutations.

The matrix of amino acid contacts was obtained for each of these protein dimers using the structural information in the pdb files. An amino acid was considered to be part of the interface if it was in contact with at least one amino acid from the other protein in the dimer. Amino acids were considered to be in contact if the distance between the closest two atoms from the two amino acids was less than or equal to a cut-off threshold d_{cut}. Values of d_{cut} between 5Å and 8Å were considered.

We calculated the interface free energy ΔM for each of the wild type protein pairs by summing the contact energies from the pairwise energy matrix. We then determined the change in free energy $\Delta\Delta M$ for each of the point mutations. Table 1 shows the Pearson correlation coefficients between the measured $\Delta\Delta G$'s and the predicted $\Delta\Delta M$'s for the set of 784 mutations. We used the pairwise matrices from Miyazawa and Jernigan (1999) and from Betancourt and Thirumalai (1999). The former was found to have a higher correlation with the data, and this correlation was highest (r = 0.430) when d_{cut} = 6Å. The scatterplot of the data using the Miyazawa and Jernigan matrix and 6Å cutoff is shown in Fig 2.1. The correlation is highly significant ($p \sim 8 \times 10^{-37}$) but the scatter is very large, so the pairwise matrix is not a very accurate predictor of the observed $\Delta\Delta G$.

We tried to improve this correlation by varying the way in which amino acid contacts were defined. Instead of taking the closest distance between all pairs of atoms, we tried using distances between α carbons, distances between β carbons, and distances between side-chain atoms only. However, none of these improved on the correlation obtained from using the distance between the closest pair of atoms.

DISTANCE CUTOFF (ANGSTROMS)	MIYAZAWA AND JERNIGAN (1999)	BETANCOURT AND THIRUMALAI (1999)
5.000	0.404	0.273
6.000	0.430	0.313
7.000	0.430	0.307
8.000	0.383	0.276

Table 2.1. The cutoff vs error table for the two energy matrices. It is seen that the Miyazawa-Jernigan pairwise potential matrix shows better correlation.

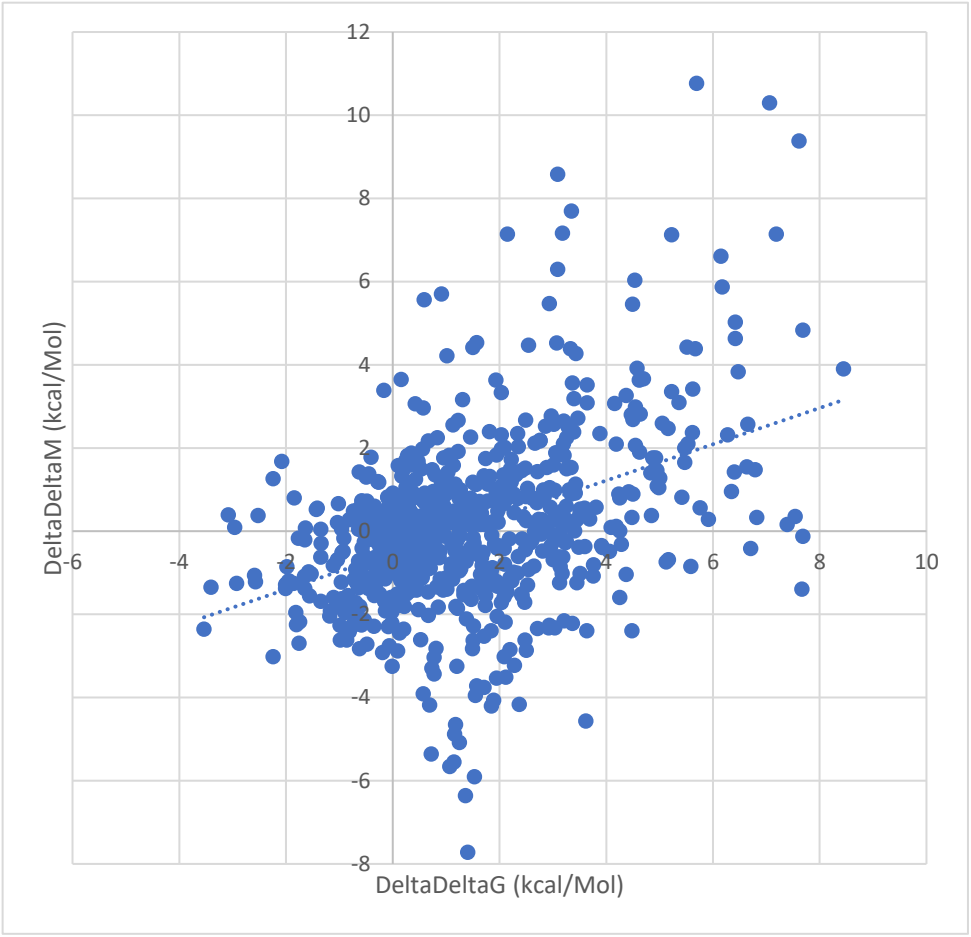


Fig. 2.1. The scatterplot of the data using the Miyazawa and Jernigan matrix.

When carrying out the evolutionary simulations we begin with wild type sequences of known structure. The positions of the interface contacts are known in the wild-type structure. The free energy for mutant sequences that arise in the simulation depends on the amino acids at the contact positions. We assume that the positions of these contacts remain the same as in the wild type structure, but the amino acids at those positions change due to point mutations. We want to set the free energy of the interface in our model to be equal to the free energy of the real wild-type protein, which we will call ΔG_{wt}, when the sequence is the wild-type sequence. In some cases, ΔG_{wt} is measured in experiments. In other cases we can estimate this using the method of Bonvin et al.(2005) which gives an approximate prediction of ΔG from the interface contacts.

Let ΔM be the free energy of any sequence calculated by summing the pairwise contact energies for that sequence, and let ΔM_{wt} be this value for the wild-type sequence.

We set the constant ΔG_0 to be

$$\Delta G_0 = \Delta G_{wt} - \Delta M_{wt}$$

and then for any mutant sequence, we set

$$\Delta G = \Delta G_0 + \Delta M$$

which ensures that $\Delta G = \Delta G_{wt}$ or the wild-type sequence.

We also observed that the binding free energy of the interfaces will increase by the contact number. Figure below belongs to the list of heterodimers from the Bonvin et al (2005) paper.

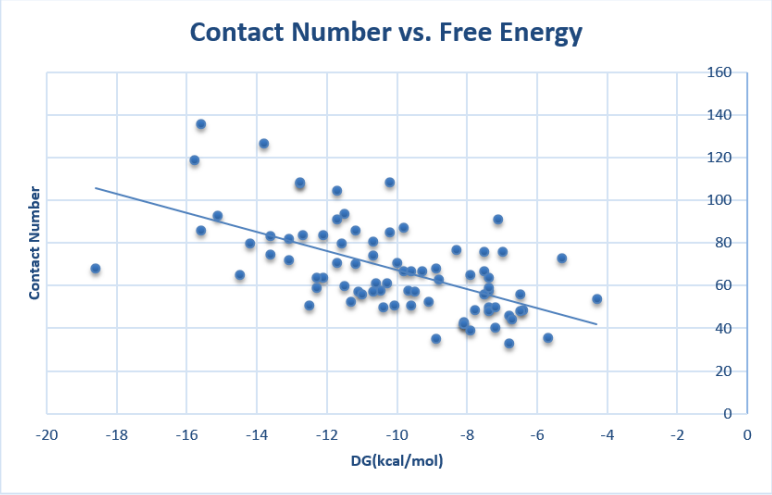


Fig. 2.2. Relation of contact numbers with the experimental binding free energy of heterodimer list of structures. Every dot represents a PDB structure.(data from Bonvin et al.(2005))

We also did the same analysis for the homodimer list used in Marchant et al. (2019) since they studied gene duplication evolution in their work and we want compare their results with our method. Using the method of Bonvin, we calculated the free energy for each structure.

By increasing the contact number, the interface energy decreases. This means we have stronger bindings for bigger interface and also demonstrates typical range of free energies for the interfaces of homodimer structures. Since in our analysis some residues have more than one contact with the other chain, we also created a graph showing the behaviour of binding energy versus the residues in contact. Every residue has a mean of 3.5 contacts with the residues on the other chain.

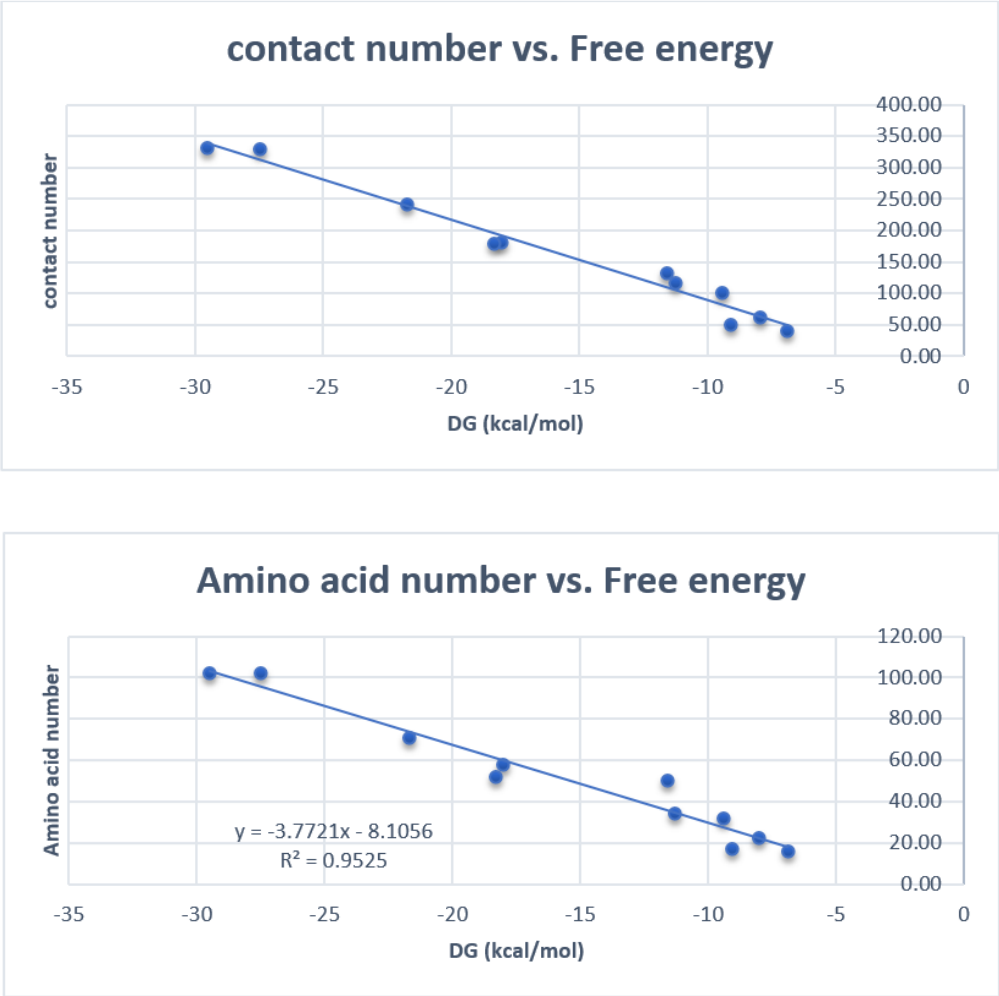


Fig. 2.3. Relation of contact numbers with the binding free energy of selected homodimer structures from Marchant paper. In the second figure, the amino acid/residue number in contact to the chain versus the binding free energy is illustrated. Every dot represents one PDB structure. The pdb files of structures are: 3pym, 2hjh, 1jeh, 1v59, 3cq0, 3d8x, 2prz, 4g9k, 3rkk, 1afw, 1pxt.

2.4. Results

We begin these simulations with a single gene. We start evolving the initial sequence of the interface through mutations. For instance, one of the homodimer structures from the Yeast is "Ribonucleotide Reductase Rnr4" with the PDB id of 1SMS is illustrated below. Using the method of Bonvin et al.(2005), we are able to determine the binding affinity of the real structure. That would also give us the ΔG for each structure. We need to fix the initial dimer fraction. To do this, we use the K_d(binding affinity) of the structure to define the concentration, setting the initial conditions to get the dimer fractions to be as high as 80% as the initial point of simulations since we start evolution with a real structure. In order to see how the selection would affect the evolution, we ran several simulations with different selection coefficients.

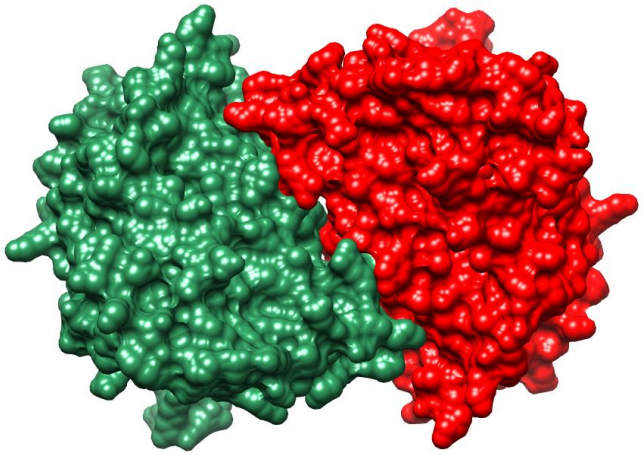
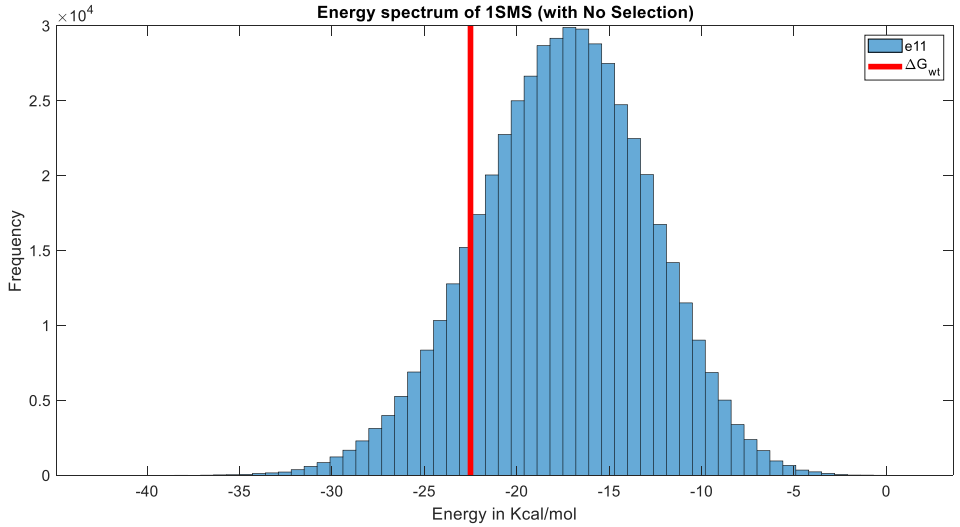


Fig. 2.4. Three dimensional Representation of 1SMS homodimer structure

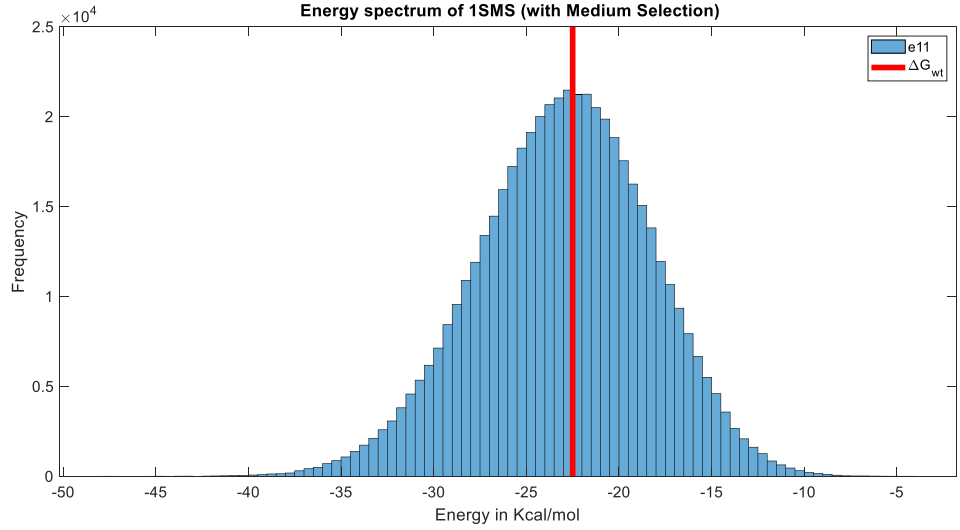
First, we run simulations with zero selection strength. We expect to lose interface in that case.

Then we run simulations for medium and high selection as depicted below.



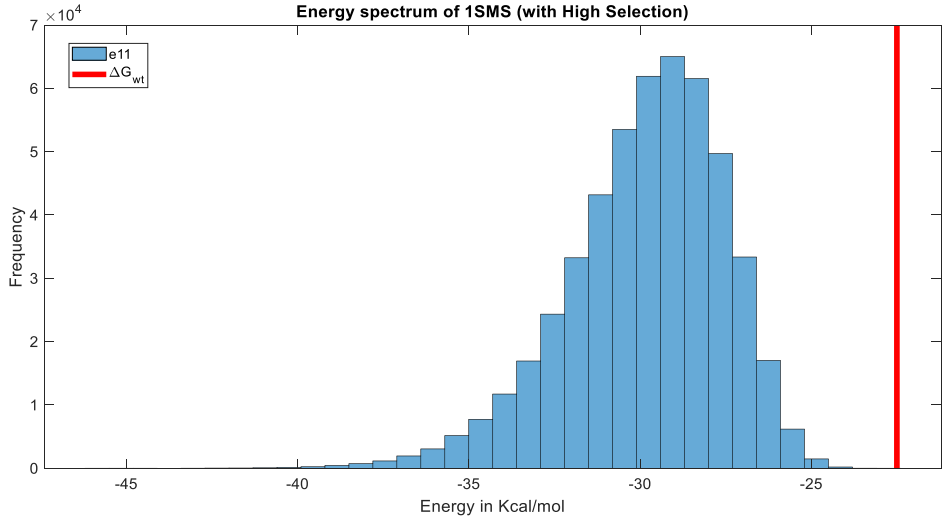
Population: $N_e = 200$

Selection strength: $\sigma = 0$



Population: $N_e = 200$

Selection strength: $\sigma = 0.005$



Population: $N_e = 200$

Selection strength: $\sigma = 0.1$

Fig. 2.5. Energy spectrum of the homodimer (1SMS) for different selection strengths.

We can conclude that the mean energy remains close to ΔG_{wt} for medium selection. Also for the case with strong selection we expected to reach to lower energies. For high selection almost all of the structures would have energies less than the wild type which means they have stronger binding energies. On the other hand, when we have no selection acting on the dimers, we still have negative energies, but the total dimer fractions will stay too low. This is shown in the Fig.2.6.

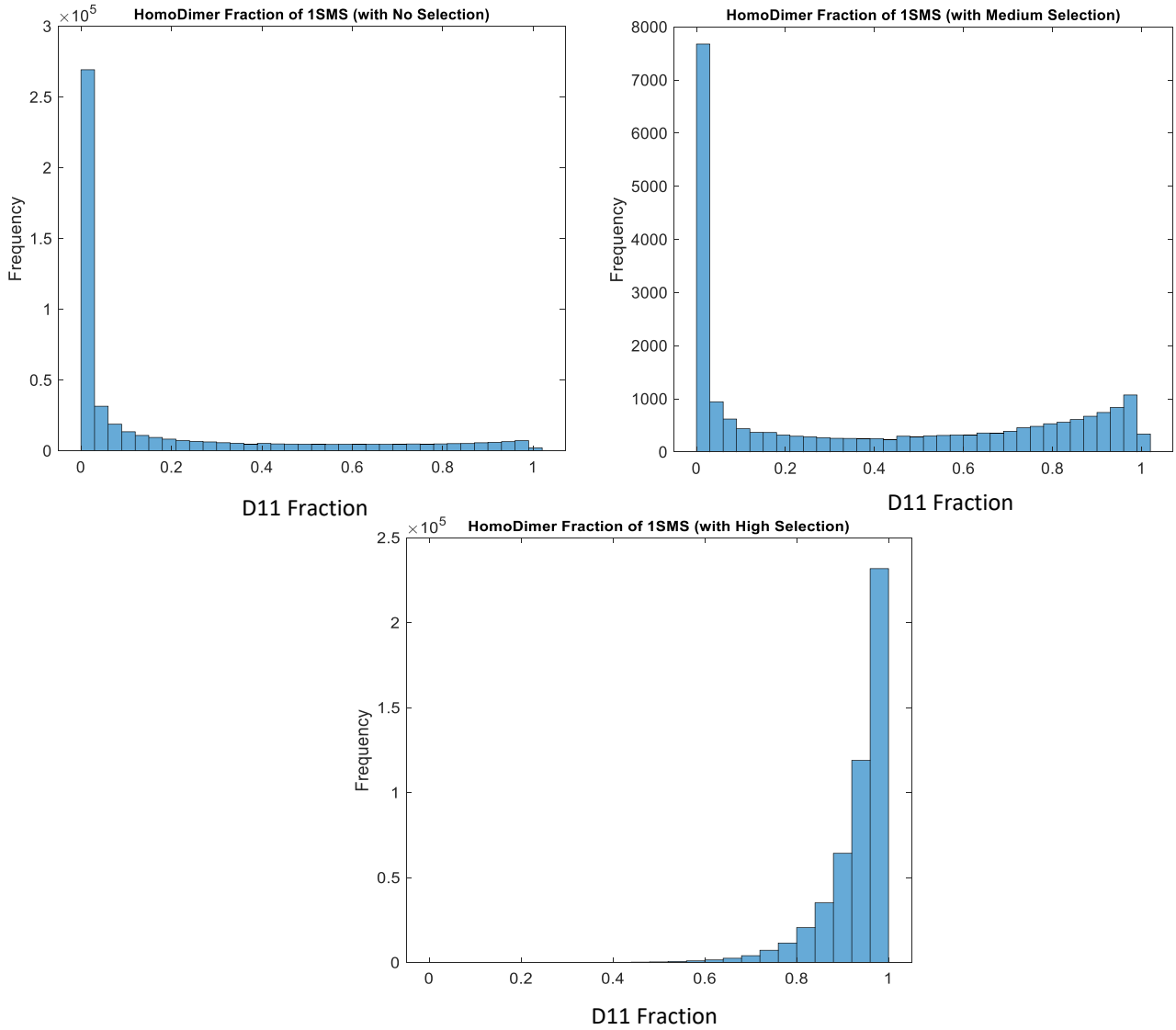


Fig. 2.6. Dimer fractions of homodimer 1SMS with different selection strengths.

Figure 2.6 shows us how by changing the selection coefficient, the dimer fractions would change. It can be observed that for zero selection, we are losing the interface, and there is very little dimer fractions present. On the other hand, for strong selections, we anticipated to see high dimer fractions. For the medium case, we can see that we can reach to high dimer fractions, however, the probability to have strong dimers are still low. Since we set the dimer fraction of the wild type to be 0.8, we can see that for the medium case we also have some cases that we reach higher fractions than the wild type.

2.4.1 Interface propensities of amino acids

It is known that the frequencies of amino acids occurring at interface positions are different from the average frequencies at other positions in the structure. In the papers by Jones et al.1997 and Levy et al.2012 they have defined interface propensity scores using large datasets of real protein structures. In previous work in our group (Zabel et al. 2019) we calculated interface propensities in the model for the flat cubic interface and showed that these were similar to the observed values. We therefore wished to look at interface propensities in the current model. In our simulations, we are using point mutations of amino acids, assuming that each of the 20 amino acids has a frequency of 0.05. Therefore the interface propensity P_i is defined as:

$$P_i = p_i/0.05$$

where p_i is the observed frequency of amino acid *i* in the simulations.

This model enables us to measure the propensities of the amino acid occurrence in real protein structures. In the Figure 2.7, there are propensities for a long simulation with high selection (selection strength is $\sigma = 0.1$) and the population is $N = 200$.

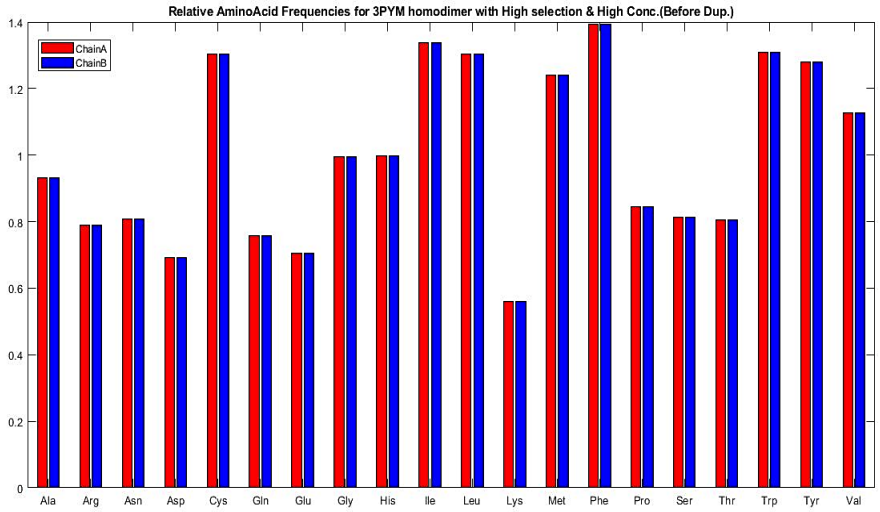


Fig. 2.7. Amino acid propensities for the Yeast protein 3PYM (Homodimer)

It can be seen that most of the hydrophobic amino acids have high propensities. We also selected a list of heterodimers from Bonvin et al.(2005) to compare the propensities with the homodimer case. In figure 2.8 we used the mean propensities for several different heterodimer structures. The PDB structures are, 3SGB, 1PPF, 1R0R, 1A22, 1EMV, 1BRS, 1A4Y, 1IAR, 2G2U, 2WPT, 1KTZ and 1JTG.

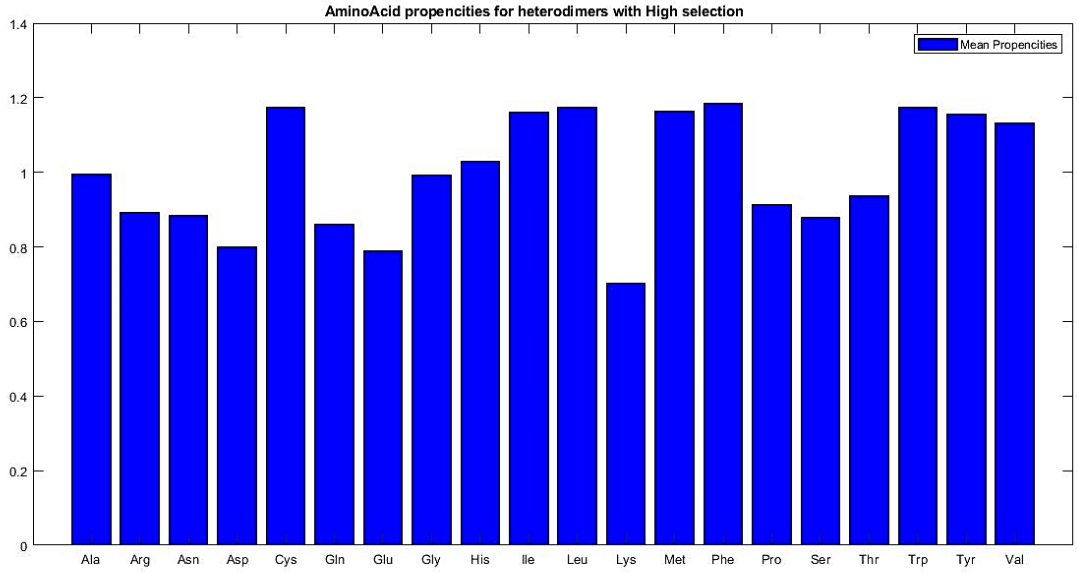


Fig. 2.8. Mean propensities for heterodimer structures

To see the ranking of amino acids the next figure is a better representation:

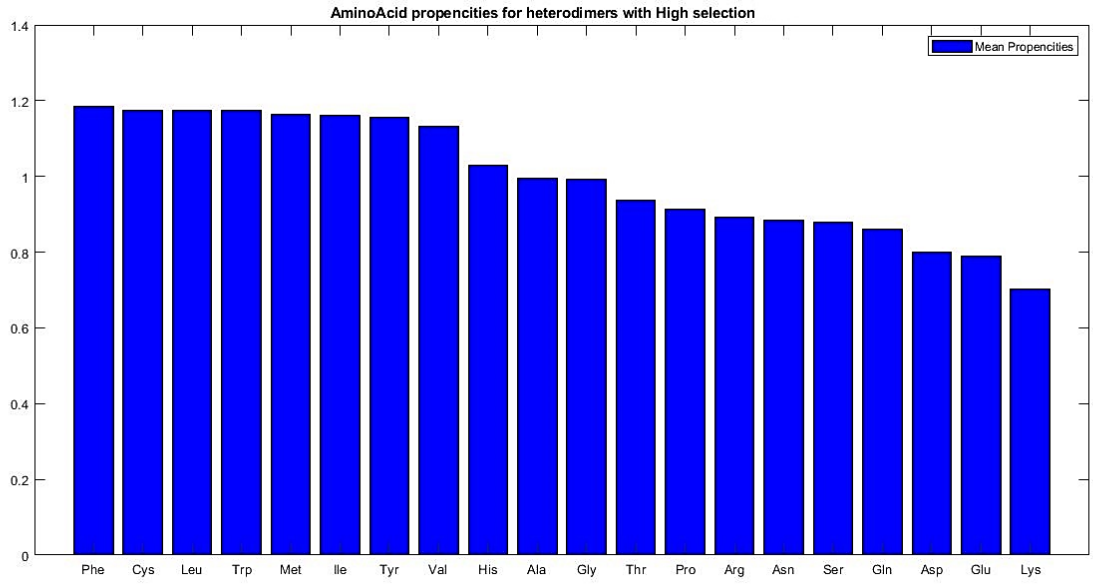


Fig. 2.9. Mean propensities in ranking (heterodimer evolution).

We observed that the ranking of amino acid frequencies for the heterodimer list is the same as the evolution of homodimers.

All the measured relative frequencies rankings agree with the experimental data on "relative interface propensity" from Jones et al.1997, and "stickiness" is the interface propensity scale from Levy et al.2012. Both of these scales are derived from the experimental frequencies of amino acids at protein-protein interfaces relative to their frequencies at non-interacting surfaces. For stickiness the ranking is: Phe, Ile ,Cys, Met, Leu, Tyr, Trp, Val, Ser, His, Thr, Ala, Arg, Gly, Pro, Asn, Gln, Asp, Glu, Lys. Using Spearman's correlation we have a ρ of 0.924. This indicates a strong positive relationship between the ranks of amino acod propencities obtained in the our method and experimental results.

We also tried to find the propencities of the amino acids using the Betancourt matrix. Using Spearman's correlation we have a ρ of 0.665. This indicates a positive relationship between the ranks of amino acid propencities obtained in our method using Betancort matrix and experimental results. These results also demonstrates that Miyazawa's matrix has a better correlation than the Betancourt matrix with the experimental data.

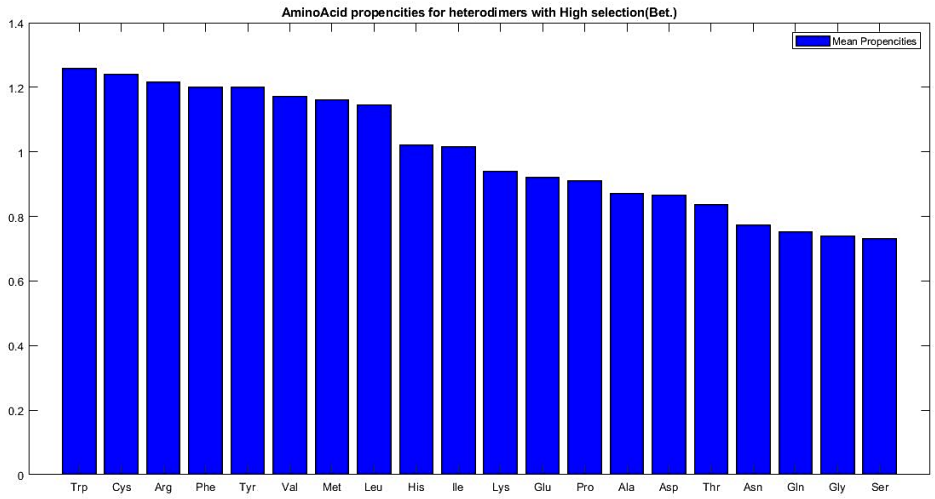


Fig. 2.10 Mean propensities in ranking using the Betancourt pairwise energy matrix

Chapter 3: Gene duplication

In this chapter we consider the evolution of two paralogous genes after a gene duplication event. We begin these simulations with a single gene, as described in the previous chapter. Then we duplicate the gene to make a second copy which is identical to the first.

We name each of them protein 1 and protein 2, which would lead to three different dimers (11, 22, and 12).

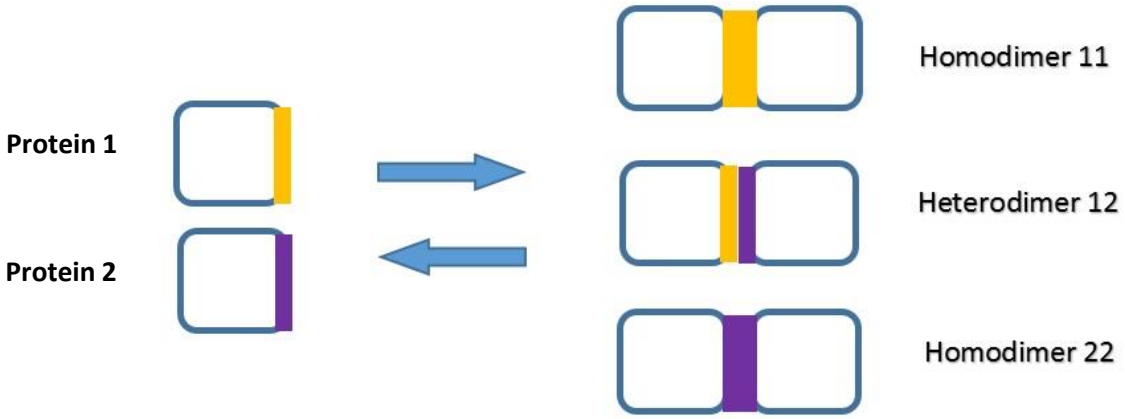


Fig. 3.1. Two monomers that can build three different dimers

Initially the three free energies of the three interfaces will be the same, but as mutations accumulate and they will all become different. There are thus three different dissociation constants K_{11}, K_{22} and K_{12} and three different equilibria:

$$d_{11} = c_1^2/K_{11}$$

$$d_{12} = 2c_1c_2/K_{12}$$

$$d_{22} = c_{2/K_{22}}^2$$

Total concentrations:

$$\phi_1 = c_1 + 2c_1^2/K_{11} + 2c_1c_2/K_{12}$$

$$\phi_2 = c_2 + 2c_2^2/K_{22} + 2c_1c_2/K_{12}$$

Suppose that the level of protein expression from each gene is reduced by half, so that the total concentration remains the same. $\phi_1 = \phi, \phi_2 = \phi$.

Solve numerically for c_1 and c_2. Define the three dimer fractions as

$D_{11} = \frac{d_{11}}{\phi}$, $D_{22} = \frac{d_{22}}{\phi}$, $D_{12} = \frac{d_{12}}{\phi}$.

We assume that the fitness simply depends on the total dimer fraction because all types of dimers perform the same function.

$$w_{dup} = 1 + \sigma(D_{11} + D_{12} + D_{22})$$

3.1 Results

In the figure 3.2, the evolution of a dimer by mutations is shown. We have the dimer reaching an equilibrium with high dimer fractions for the first half of the simulations. Immediately after the gene duplication, the three *K*'s are the same. The total dimer fraction after duplication is the same as the dimer fraction D_{sing} before duplication. Thus, the duplication event is neutral. The ratio of the three dimer types is

$$D_{11}: D_{12}: D_{22} = 1/4: 1/2: 1/4$$

If selection for dimers is strong (*i.e.* $\sigma N > 1$), then the total dimer fraction is close to 1.

The two surfaces evolve by mutations. The three *K*'s change. There is selection for maintaining the total dimer faction. If K_{12} becomes high, the heterodimer interface becomes weak, and we reach a state of mostly homodimers, where $D_{11}: D_{12}: D_{22} \approx 1/2: 0: 1/2$. If K_{11} and K_{22} become high, the homodimer interfaces become weak, and we reach a state of mostly heterodimers, where $D_{11}: D_{12}: D_{22} \approx 0: 1: 0$. We have shown that this can occur with very little reduction in the total dimer concentration, because one type of dimer compensates for the loss of the other.

We present the results for evolution of a random system by mutations with a strong selection strength of $\sigma = 0.1$ and a population of $N_e = 200$ on the next page. These two graphs are demonstrating how dimer fractions of three different dimer structures would change after a gene duplication event. In each figure, a different final state has been shown. We chose 3PYM protein, using this protein's contact network we start our simulations by assigning random amino acids for each residue. In this case, the initial $\Delta G_{mut} = \Delta M_{mut}$ since we are not starting the simulations using the real structure's amino acids.

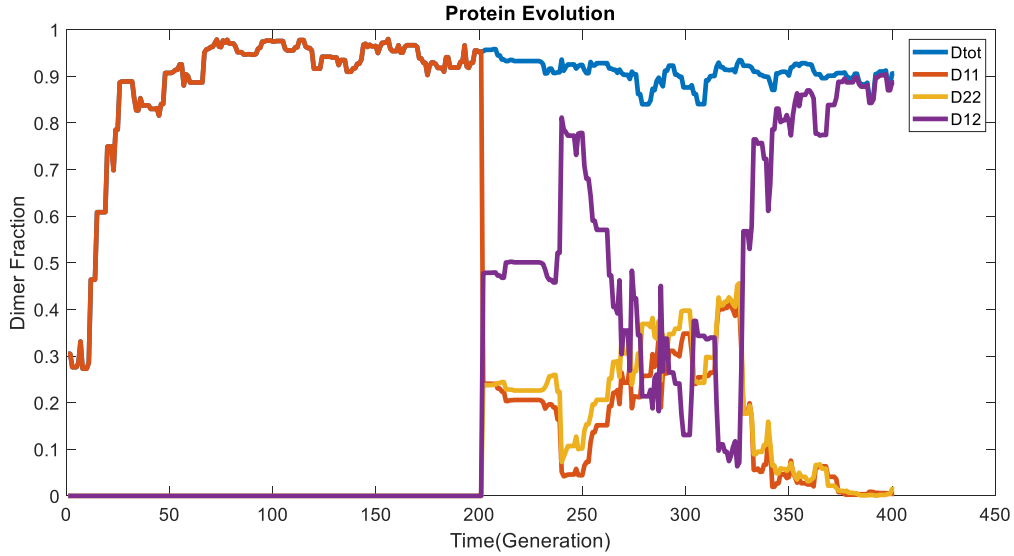


Fig. 3.2. Evolution of a 3PYM with random sequence by mutations with a strong selection strength of $\sigma = 0.1$ and a population of $N_e = 200$. Dii is the dimer fractions and $Dtot$ is the total dimer fraction. The first half is the evolution of single gene dimerization. The second half shows after alteration it reaches to mostly heterodimer state.

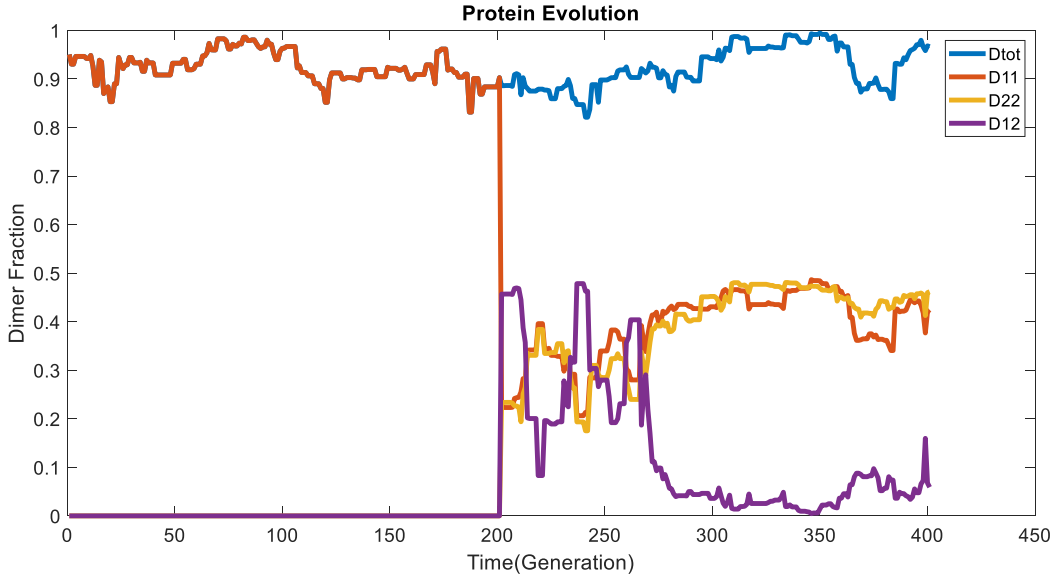


Fig. 3.3. Evolution of 3PYM with a random sequence by mutations with a strong selection strength of $\sigma = 0.1$ and a population of $N_e = 200$. The second half shows after alteration it reaches to mostly homodimer state.

The following argument explains why we see this behaviour. In the figure below, we demonstrate that if a loss of the binding interface happens for either structure, the total dimer fractions will compensate, and we will see that by losing the binding interface of either homodimer or heterodimer structure, the other one would increase in fraction, keeping the total dimer fractions high since the fitness is still dependent on $Dtot$.

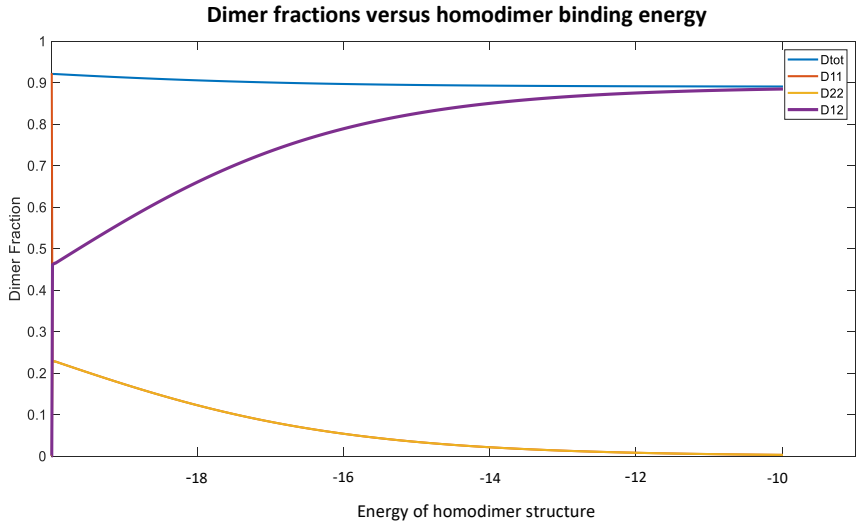


Fig. 3.4. Change of dimer fractions with loss of homodimer binding interface after gene duplication. The purple curve shows the heterodimer fraction, and the x-axis is the energy of homodimer structure.

In Figure 3.4, we have calculated the dimer fractions as a function of the three ΔGs. On the left of each figure all three ΔGs are strong ($\Delta G = -20\ kcal/mol$). Here we have increased ΔG of the homodimers while the heterodimer remains fixed. We demonstrate that by increasing the homodimer structure's energy and keeping the heterodimer energy constant, the heterodimer structure's fractions will increase. Since we defined our selection criteria to have high dimer fractions, the homodimer fractions decrease but the heterodimer fraction increases, in such a way that the total dimer fraction remains close to 1; and we see that we will end up having more heterodimer structures in this scenario.

The same would happen if we have the loss of binding interface for heterodimer structure, keeping the homodimers energy constant. In the figure below we have increased ΔG of the heterodimer while the homodimer remains fixed. We observe that both homodimer fractions stay around 0.5 in compensation of the decrease in the heterodimer structure fractions, keeping the total dimer fraction high. These examples explains why we anticipated this behaviour in our simulations to have the dimers choosing either extreme, being mostly at homodimer or heterodimer state. Having this in mind, we also expect to observe a U-shaped distribution for the dimer fractions.

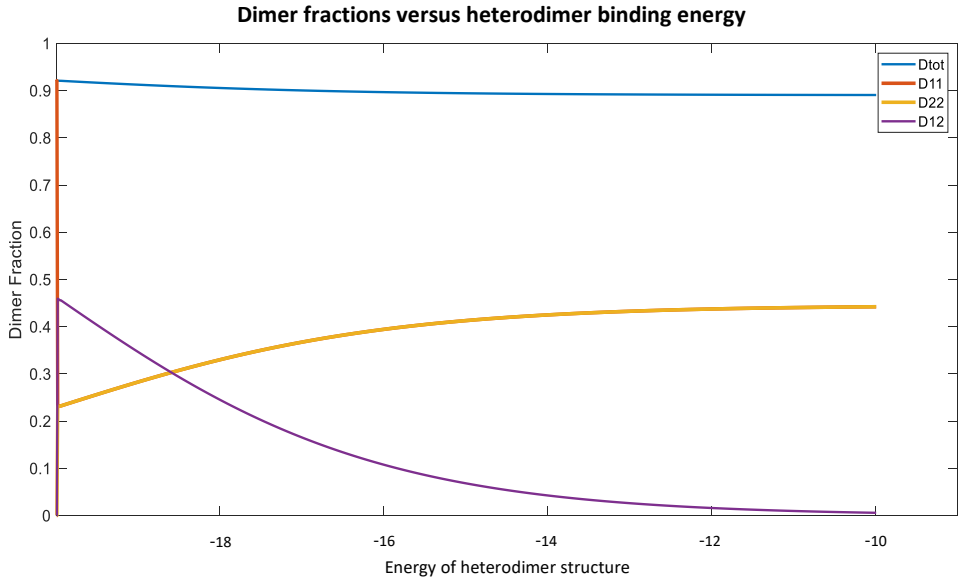


Fig. 3.5. Change of dimer fractions with loss of heterodimer binding interface after gene duplication.

It is possible to lose either the homodimer interfaces or the heterodimer interface without reducing the total dimer fraction very much. There are frequent mutations that will cause the reduction in binding strength of either the homomers or the heteromer. As there is very little selection acting against these mutations (because Dtot remains high) then we expect such mutations to accumulate. Thus the system will evolve either to the state with two homodimers or to the heterodimer state, and it is unlikely to remain in a balanced state with all interfaces strong and a 1:2:1 ratio of concentrations.

Simulations of the selected protein list from Sacharomyces cerevisiae are shown in the figure below. First, we ran long simulations using the Metropolis accept/reject method. In fig.3.7, the frequencies of the homodimer fraction are illustrated. These figures show that there is a higher probability close to the homodimer end of the distribution, and thus the rusults predict that there is a higher probability of finding two homodimers than a heterodimer.

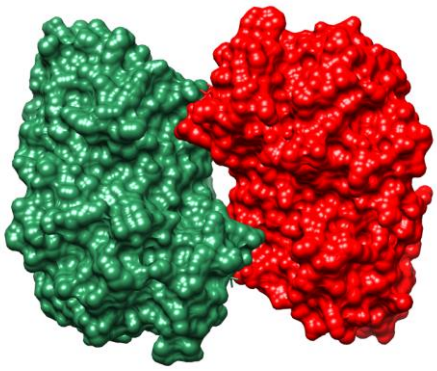


Fig. 3.6. Three dimensional structure of 3PYM homodimer.

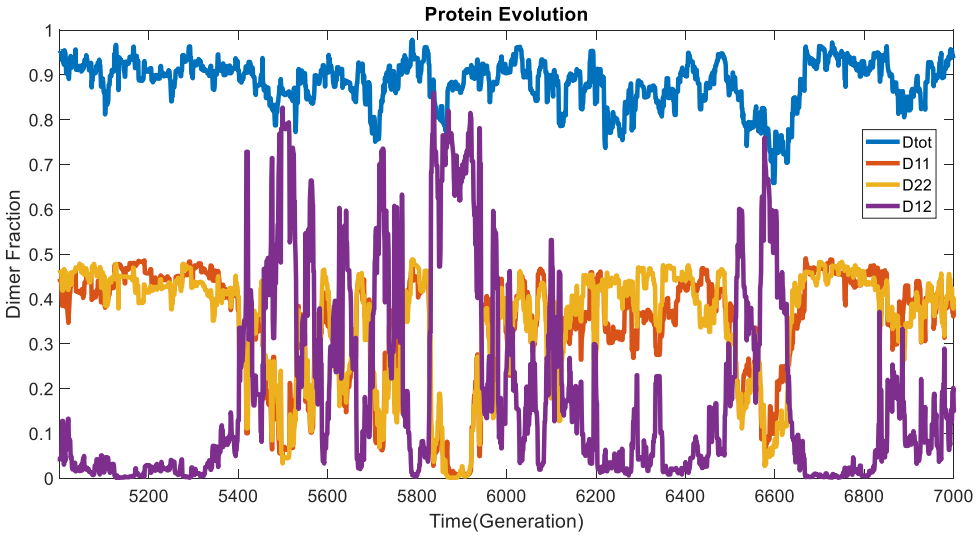


Fig. 3.7. Evolution of the 3PYM structure within a time span; with $N_e = 200$ and selection strength of $\sigma = 0.1$. We see that dimer fractions are alternating between mostly homodimer and mostly heterodimer state.

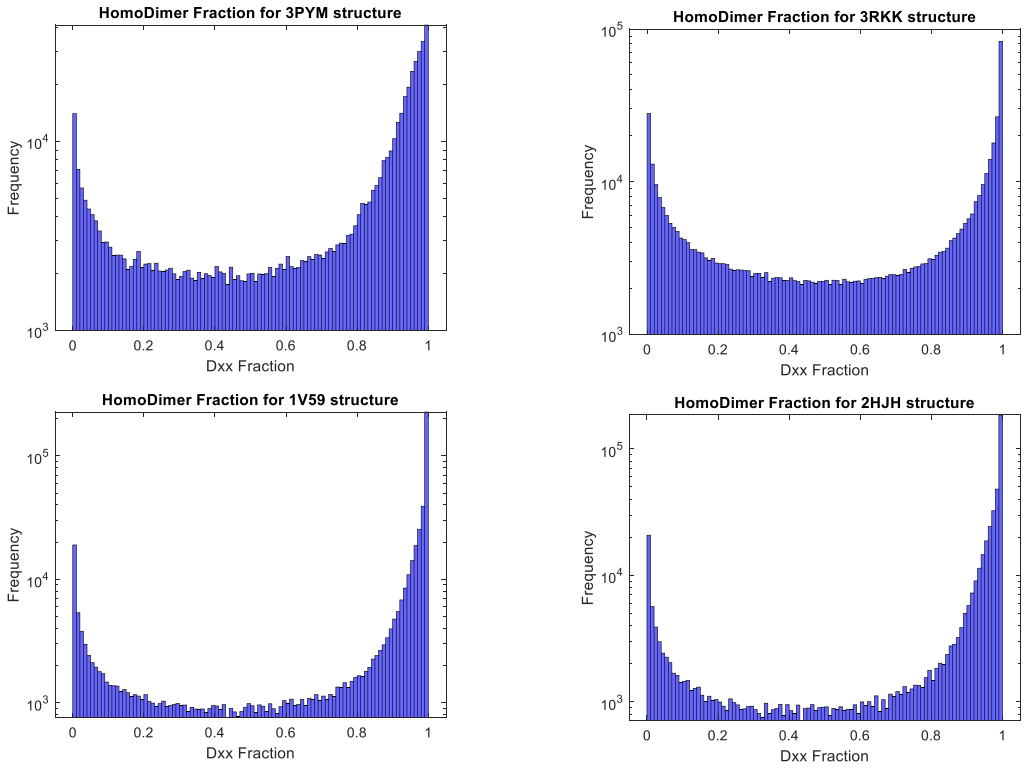


Fig. 3.8. The U-shaped distribution of a few homodimers from the Yeast. In here "Dxx" is the total homodimer fraction ($Dxx = D11 + D22$).

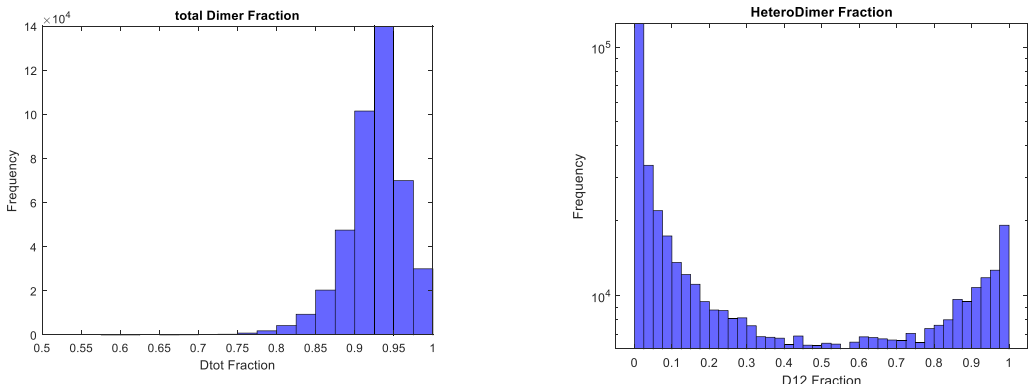


Fig. 3.9. Total dimer fraction and heterodimer fractions of 3PYM structure.

This shows that the dimers prefer to be mostly at one extreme, either the homodimer state or the heterodimer state. In the previous cubic model, the system alternates between states of mostly homodimers and mostly heterodimers due to the occurrence of mutations in the interfaces. States with all three dimers having strong interfaces will be rare. However, using the real structures of PDBs, we also observe the same alteration between the two states. The only difference between these two would be that real structures prefer staying on the homodimer state more than the heterodimer state.

As the next step in our Project, we are interested to see how the simulations would lead us to an evolutionary fate of the duplicated genes (Lynch and Force, 2000). We have the results showing us the intermediary state where after gene duplication the dimer is going back and forth from being in mostly heterodimer state to mostly homodimer state. Therefore we investigate further for probability of deletion, subfunctionalization and neofunctionalization.

3.2 Deletions from the duplicated state

Assume that either one of the two genes can be deleted with some probability. Before the gene duplication we defined the concentration of the single protein as 2ϕ and the maximum dimer concentration as ϕ. We assumed that when the gene was duplicated, there was some kind of regulation of the total amount of protein produced, so that the concentration of protein produced from each gene was ϕ, and the total dimer concentration still had a maxmum of ϕ. This means that the duplication is neutral. Now, if a gene is deleted, we assume that the concentration of protein produced from the remaining gene is increased to its original value of 2ϕ. The dimer fraction is then given by D_{sing} once again. Therefore the change in fitness upon deletion of a gene is:

$$\Delta s = \sigma(D_{sing} - D_{11} - D_{12} - D_{22})$$

In the mostly homodimer state, $D_{11} \approx D_{22} \approx 1/2$, $D_{12} \approx 0$. After the deletion, either gene 1 or gene 2 becomes the single gene, and $D_{sing} \approx 1$, because both the 11 and 22 interfaces are strong. Hence, the fitness change on deletion from the mostly-homodimer state is close to zero. Fixation of the deletion can occur at the neutral deletion rate - which we expect to be relatively fast.

In the mostly heterodimer case, $D_{12} \approx 1$, and $D_{11} \approx D_{22} \approx 0$. If either gene is deleted, the dimer fraction after deletion is very small ($D_{sing} \approx 0$), even if the concentration of the remaining gene goes back to 2ϕ, because the 11 and 22 interfaces are weak. Therefore $\Delta s = -\sigma$.

Deletion of a gene is substantially deleterious. The prediction is that the system will be observed more frequently in the mostly-heterodimer state than the mostly-homodimer state because deletions from the mostly-homodimer state will return the system to the single-gene state.

3.3 Loss of function

Assume that function at the active site can be lost by mutational events that do not involve the interface amino acids. Assume that the heterodimers and homodimers can lose function independently of each other.

If the heterodimer loses function, the fitness is

$$w_{homo} = 1 + \sigma(D_{11} + D_{22})$$

The change in fitness due to the loss of function is w_{homo} - $w_{dup.}$ If the loss of function occurs when D_{12} is high, then this fitness change is substantially deleterious, and is unlikely to be fixed. If it occurs in the mostly-homodimer case, when $D_{12} \approx 0$, then the fitness change on loss of function is almost zero. Thus loss of function of heterodimers can occur easily if the heterodimer interface is already weak.

If the homodimers lose function, the fitness is

$$w_{hetero} = 1 + \sigma D_{12}$$

This can occur almost neutrally, if it occurs in the mostly-heterodimers state, where D_{11} and D_{22} are very small. If loss of function occurs in either of these ways, this stops the drift between the mostly-heterodimer and mostly-homodimer state, because only one of these will be fit.

To test this, we started running several simulations. Figure 3.9 illustrates the case that the homodimers loses their function.

The First case is when he homodimer structures lose their function:

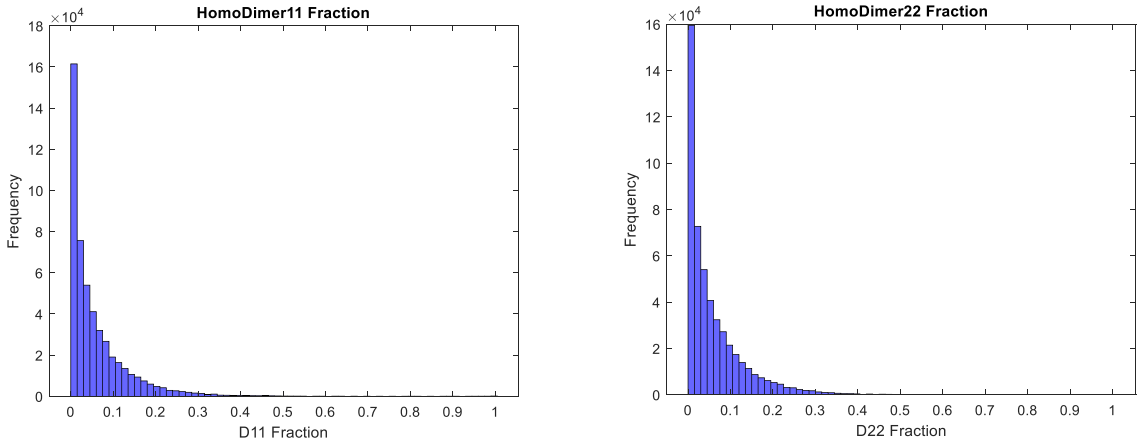


Fig. 3.10. Dimer fractions of homo11 and homo22, when they lost function.

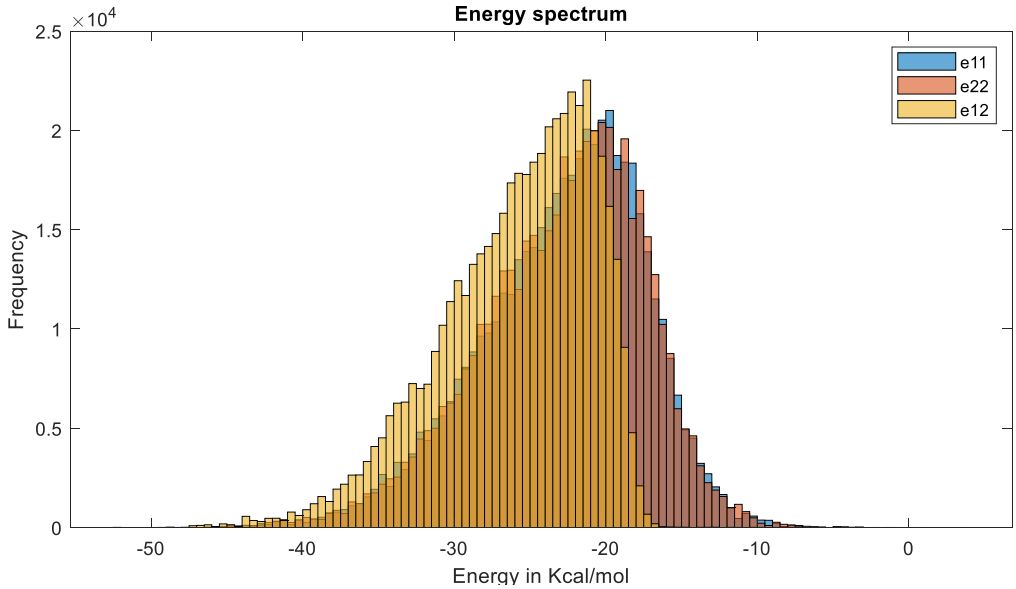


Fig. 3.11. Energy spectrum of each dimer when homodimers lose function.

For this case, we see that the mean energy for hetero12 is lower than the other two. The heterodimer structure is more stable than the homo cases. We see that heterodimer fractions are quite high when we select for them. Both homodimer structures have low concentrations, however, the total dimer concentrations are high.

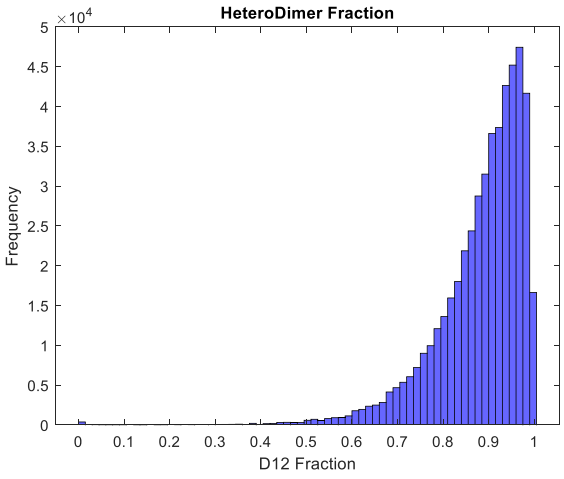


Fig. 3.12. Heterodimer fraction when homodimers lose function.

The second case is when the heterodimer would lose its function:

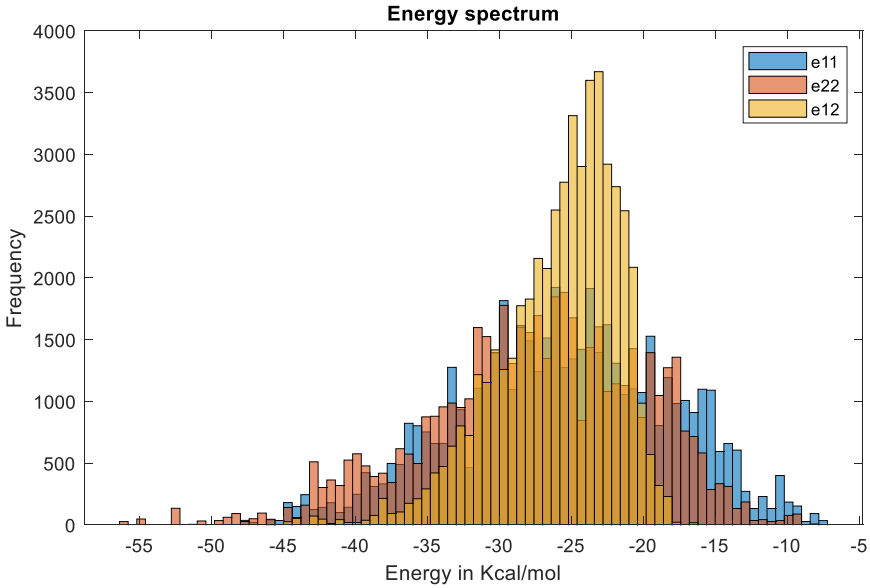


Fig. 3.13. Energy spectrum of each dimer for loss of heterodimer function.

In this scenario, the mean energy for all structures are in a close vicinity, however, energy range is wider for homo cases, being more frequent in lower energies. For the total homodimer cases we see a peak on around 1.

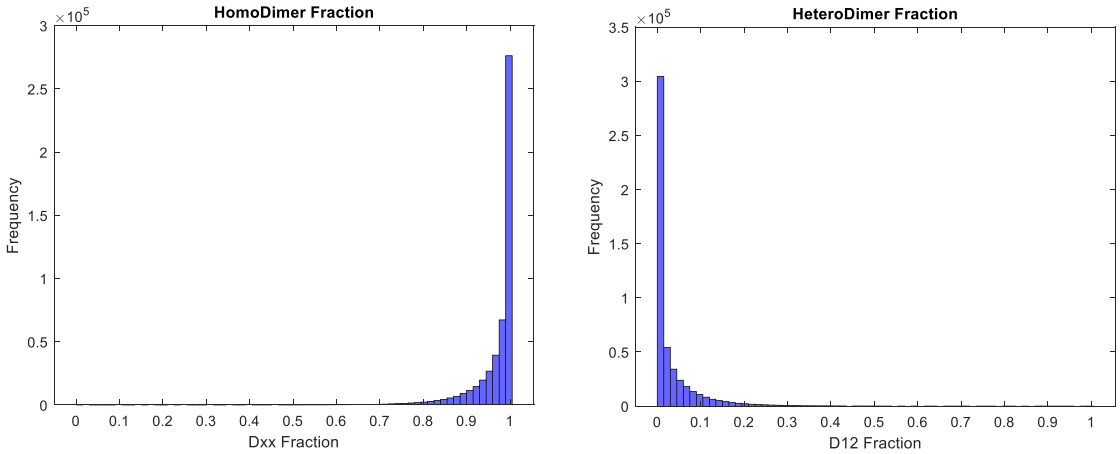


Fig. 3.14. Homodimer and heterodimer fraction when the heterodimer loses function.

In the case when we have function loss for the homodimer22 and the heterodimer, the energy spectrum will look like figure below.

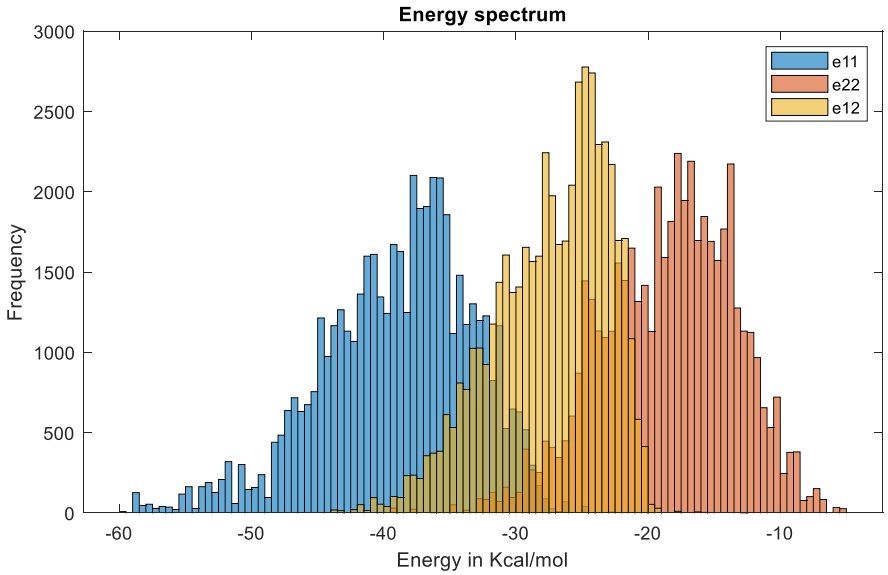


Fig. 3.15. Energy spectrum of each dimer when homo22 loses function.

In this scenario, we see that the mean energy for this structure is lower than the other two. We also observed that the heteromer energy is in between of the two homomers. One reason for maintaining heteromers might be that if homomer11 has stronger binding energy, then the homodimer 22 is still partly sticky for the other protein.

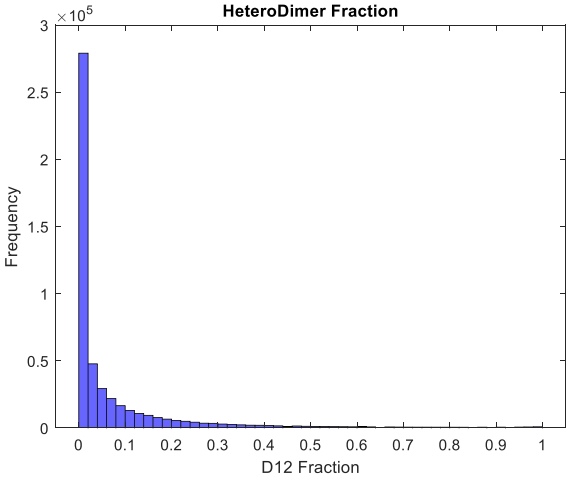


Fig. 3.16. Heterodimer fraction when homo22 loses function.

3.4 Subfunctionalization

Subfunctionalization means that two genes are required to carry out the job of one original gene. If there is loss of function of the homodimers, this is already true. The fitness w_{hetero} can be considered as the fitness of the subfunctionalized state. Deletion of either gene is very deleterious from this state because neither of the homodimers is functional, and because the homodimer interfaces are weak, so the homodimers do not form in any case. Therefore subfunctionalization can occur simply by loss of function of the homodimers.

It is possible that there is greater flexibility for evolution of the dimer, once it becomes an obligate heterodimer. Thus it may be that a higher fitness is possible in the subfunctionalized state:

$$w_{sub} = 1 + \sigma_{sub} D_{12}$$

where $\sigma_{sub} > \sigma$. If this occurs, this will be a further reason to retain both genes. However, this increase in function is not necessary to explain the subfunctionalization, because deletion of the genes is already deleterious after the loss of function of the homodimers.

3.5 Neofunctionalization

If the loss of function of the heterodimers occurs, then we have two independent genes doing the same function. However, this is not enough to constitute neofuctionalization, because deletion of either of these genes gets us back to the single gene state (assuming that the expression level of the remaining gene goes back to 2ϕ). If gene 2 acquires a new function, and is no-longer performing the original function, we will write the fitness as

$$w_{neo} = 1 + \sigma D_{11} + \sigma_{neo} D_{22}$$

We assume that both genes 1 and 2 return to the high level of expression because they are now regulated separately ($\phi_1 = 2\phi, \phi_2 = 2\phi$). Since are regulated separately then the max D_{11} and D_{22} are both 1, so the max fitness is $w_{neo} = 1 + \sigma + \sigma_{neo}$. Therefore, it will be deleterious to delete either of these genes because the corresponding fitness terms in w_{neo} will be lost.

Neofunctionalization really does require a new function. Loss of function of the heterodimers is not enough, because if there is no additional function, neutral deletion of one of the genes is possible. This contrasts with subfunctionalization, where loss of function of the homodimers is sufficient to stabilize the heterodimer state without any increase in fitness of the heterodimers.

Chapter 4: Conclusion

In this thesis we have designed and investigated an evolution theory based on interface energies to form dimeric proteins. Using the three-dimensional dimer protein structures of the protein data bank, from both homodimer and heterodimer proteins, we established a contact network calculating the binding energies of interfaces using an appropriate pairwise energy matrix so that strong binding energies would have stable structures, leading to higher dimer fractions. Then by defining the structures' fitness dependent on the total dimer fractions, we concluded that the structures are always bound to go one way or another towards a mostly homodimer state or a mostly heterodimer state. In our simulations, we demonstrated that as the mutations occur, the probability distributions of the homodimer fraction and the heterodimer fraction are both U-shaped graphs. This illustrate that there are many ways to have homodimer or heterodimer structures, but there are very few ways to maintain them both as strong dimer interfaces. In order to enlighten this results, we designed a pathway in Fig.4.1 to demonstrate all the states that protein might go through after gene duplication.

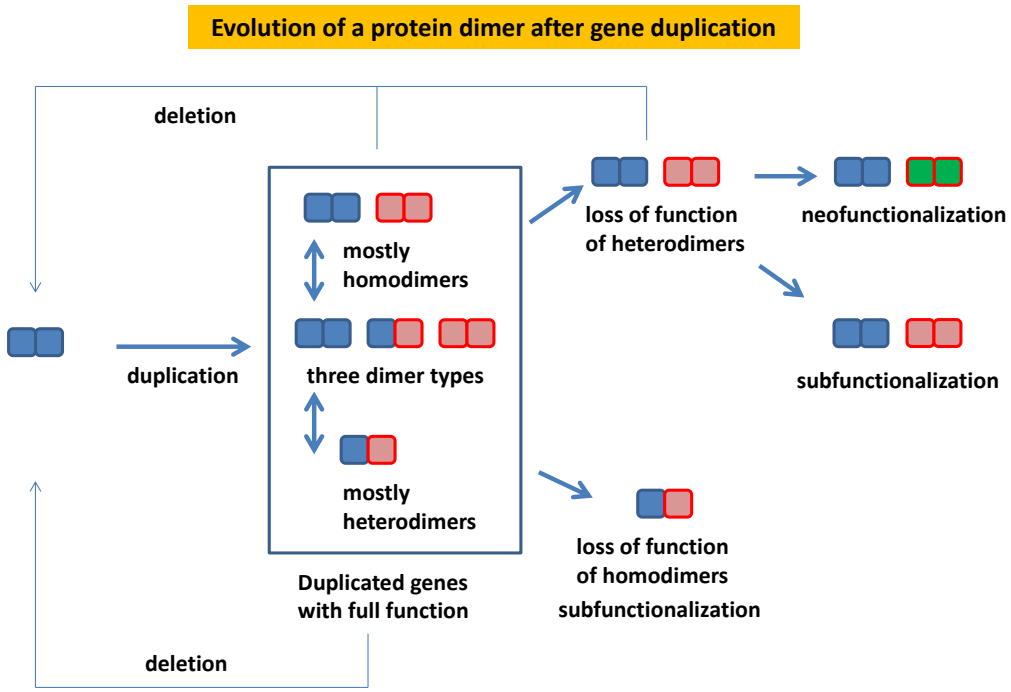


Fig. 4.1. Duplication pathway

The duplication pathway graph, demonstrates that after the gene duplication event, if we define the fitness to have high dimer fractions then we would end up in the box in the middle. This box is the state of alternating between being mostly homodimer or heterodimer. As we showed in the results section, if we increase ΔG of either homodimers or heterodimers while the other structure's energy remains fixed, then the structure with the fixed energy will have an increase in dimer fractions to compensate for the loss of interface energy of the other structure in such a way that the total dimer fraction remains close to 1. This demonstrates that we are bound to go one way or another.

In our simulations, we pointed out what would happen to the structures after alternating in the box. The deletion and loss of function of either structure could happen next. In the case of loss of function for homodimers, we would end up having heterodimers compensate for it. Furthermore, in order for the heterodimers to compensate, the heterodimer's free energy does not need to get stronger. It can merely stay the same because as was explained in the previous paragraph, the heterodimer fraction goes up automatically when the homodimer fractions go down. The same would happen if we have the loss of function for the heterodimer case. The only difference was that in the case of loss of heterodimer structure, the mean energy spectrum for all three structures were in the close vicinity; however, the range of the energy spectrum was much wider for the homodimer cases than the heterodimer structure.

In the end, if we continue from the box in the middle, we reach the final fate of the duplicated gene. We would expect three fates, subfunctionalization, neofunctionalization, or loss of the gene. We suggest that a loss of function in homodimer structures might eventually lead to a subfunctionalization fate since the two interfaces are different. For the other scenario, if a loss in heterodimer structure occurs, we would have two homodimers that one could gain a new function while the other is performing the proteins old function, which is called neofunctionalization. Furthermore, in this case the two separate homodimers could each assume different parts of the full function of the original gene,which would be subfunctionalization. In other words, once we have two independent homodimers they can be maintained in the genome by either neofunctionalization or subfunctionalization, as would occur if we have a duplication of a monomeric protein. Finally, the last case is deletion of the gene, which is the most probable fate of the duplicated genes.

5. Appendix

In this part we have written the derivation of the Kimura's fixation probability. Imagine there is a population with p individuals. Population would change by a random amount Δp on the next generation. To begin with, we can write the decomposition of the fixation probability,

$$\pi(p) = E_{\Delta p}\{\pi(p + \Delta p)\}$$

Where $E_{\Delta p}$ is the probability of a particular change in p, and $\pi(p + \Delta p)$ is the fixation probability for the new frequency. The sum is over all possible changes in p. Using Taylor series we can write

$$\pi(p + \Delta p) = \pi(p) + \pi'(p)\Delta p + \frac{1}{2}\pi''(p)(\Delta p)^2 + \cdots$$

$$\pi(p) = E_{\Delta p}\left\{\pi(p) + \pi'(p)\Delta p + \frac{1}{2}\pi''(p)(\Delta p)^2 + \cdots\right\}$$

$$\pi(p) \approx \pi(p) + \pi'(p)E_{\Delta p}\{\Delta p\} + \frac{1}{2}\pi''(p)E_{\Delta p}\{(\Delta p)^2\}$$

$$\frac{1}{2}\pi''(p)E_{\Delta p}\{(\Delta p)^2\} + \pi'(p)E_{\Delta p}\{\Delta p\} = 0$$

We define, $V(p) = E_{\Delta p}\{(\Delta p)^2\}$ as the variance of change in p, and $m(p) = E_{\Delta p}\{\Delta p\}$ as the mean change in p. Finally we get Kolmogorov backward equation

$$\frac{1}{2}V(p)\pi''(p) + m(p)\pi'(p) = 0$$

To solve this we change second order differential equation to a first order

$$f(p) = \pi'(p) \quad \rightarrow \quad f'(p) + \frac{2m(p)}{V(p)}f(p) = 0$$

Using boundary conditions $\pi(0) = 0$ & $\pi(1) = 1$

First boundary condition means if the p of the initial copy is zero, then the probability of fixation will be zero. Second boundary condition says if p is equal to one, then definitely we have fixation.

$$B(p) = 2\int_0^p \frac{m(x)}{V(x)}dx \quad \rightarrow \quad f(p) = \pi'(p) = c_1 e^{-B(p)}$$

$$\pi(p) = c_1 \int_0^p e^{-B(p)}\, dy + c_2 \quad \rightarrow \quad b.c. \quad \rightarrow \quad c_2 = 0 \;\&\; {c_1}^{-1} = \int_0^1 e^{-B(y)}dy$$

$$\pi(p) = \frac{\int_0^p e^{-B(y)}dy}{\int_0^1 e^{-B(y)}dy} \;\rightarrow\; m(p) = pq\frac{s}{2} \;\&\; V(p) = \frac{pq}{2N} \;\rightarrow\; B(y) = 2\int_0^y \frac{pq\frac{s}{2}}{\frac{pq}{2N}}dx$$

Finally for the Kimura's fixation probability we get:

$$\pi(p) = \frac{1 - e^{-2Nsp}}{1 - e^{-2Ns}}$$

The probability of emerging a new mutation in the population is $p = 1/N$, therefore fixation probability will be:

$$P_{fix}(s) = \pi\left(\frac{1}{N}\right) = \frac{1 - e^{-2s}}{1 - e^{-2Ns}}$$

References

André, I., Strauss, C., Kaplan, D., Bradley, P., Baker, B. (2008) Emergence of symmetry in homooligomeric biological assemblies. Proceedings of the National Academy of Sciences Oct 2008, DOI: 10.1073/pnas.0807576105

Bajaj, M., & Blundell, T. (1984). Evolution and the tertiary structure of proteins. Annual review of biophysics and bioengineering, 13, 453–492. https://doi.org/10.1146/annurev.bb.13.060184.002321

Betancourt, M. R., & Thirumalai, D. (1999). Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein science : a publication of the Protein Society, 8(2), 361–369. https://doi.org/10.1110/ps.8.2.361

Bordner, A. J., & Abagyan, R. (2005). Statistical analysis and prediction of protein-protein interfaces. Proteins, 60(3), 353–366. https://doi.org/10.1002/prot.20433

Brinda, K. V., Kannan, V., Vishveshwara, V.(2002) Analysis of homodimeric protein interfaces by graph-spectral methods, Protein Engineering, Design and Selection, Volume 15, Issue 4, April 2002, Pages 265–277, https://doi.org/10.1093/protein/15.4.265

Bryngelson, J. D., & Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of protein folding. Proceedings of the National Academy of Sciences of the United States of America, 84(21), 7524–7528. https://doi.org/10.1073/pnas.84.21.7524

Bryngelson, J. D., Onuchic, J. N., Socci, N. D., & Wolynes, P. G. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins, 21(3), 167–195. https://doi.org/10.1002/prot.340210302

Chakrabarti, P. and Janin, J. (2002), Dissecting protein–protein recognition sites. Proteins, 47: 334-343. https://doi.org/10.1002/prot.10085

Crow, K. D., Wagner, G. P.,(2006) What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity?, Molecular Biology and Evolution, Volume 23, Issue 5 https://doi.org/10.1093/molbev/msj083

Dignon, G. L., Zheng, W., Best, R. B., Kim, Y. C., & Mittal, J. (2018). Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. Proceedings of the National Academy of Sciences of the United States of America, 115(40), 9929–9934. https://doi.org/10.1073/pnas.1804177115

Dignon, G. L., Zheng, W., Kim, Y. C., Best, R. B., & Mittal, J. (2018). Sequence determinants of protein phase behavior from a coarse-grained model. PLoS computational biology, 14(1), e1005941. https://doi.org/10.1371/journal.pcbi.1005941

Hanada, K., Zou, C., Lehti-Shiu, M. D., Shinozaki, K., & Shiu, S. H. (2008). Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant physiology, 148(2), 993–1003. https://doi.org/10.1104/pp.108.122457

Higgs, Paul G., and Teresa K. Attwood. Bioinformatics and Molecular Evolution. Blackwell Publishing, 2005.

Hochberg, G., Shepherd, D. A., Marklund, E. G., Santhanagoplan, I., Degiacomi, M. T., Laganowsky, A., Allison, T. M., Basha, E., Marty, M. T., Galpin, M. R., Struwe, W. B., Baldwin, A. J., Vierling, E., & Benesch, J. (2018). Structural principles that enable oligomeric small heat-shock protein paralogs to evolve distinct functions. Science (New York, N.Y.), 359(6378), 930–935. https://doi.org/10.1126/science.aam7229

Jones, S., & Thornton, J. M. (1996). Principles of protein-protein interactions. Proceedings of the National Academy of Sciences of the United States of America, 93(1), 13–20. https://doi.org/10.1073/pnas.93.1.13

Kim, Y. C., & Hummer, G. (2008). Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding. Journal of molecular biology, 375(5), 1416–1433. https://doi.org/10.1016/j.jmb.2007.11.063

Kimura M. (1979). The neutral theory of molecular evolution. Scientific American, 241(5) https://doi.org/10.1038/scientificamerican1179-98

Konrad, A., Teufel, A. I., Grahnen, J. A., & Liberles, D. A. (2011). Toward a general model for the evolutionary dynamics of gene duplicates. Genome biology and evolution, 3, 1197–1209. https://doi.org/10.1093/gbe/evr093

Levy ED, Teichmann S. Structural, evolutionary, and assembly principles of protein oligomerization. Prog Mol Biol Transl Sci. 2013;117:25-51. doi: 10.1016/B978-0-12-386931-9.00002-7. PMID: 23663964.

Lukatsky, D. B., Shakhnovich, B. E., Mintseris, J., & Shakhnovich, E. I. (2007). Structural similarity enhances interaction propensity of proteins. Journal of molecular biology, 365(5), 1596–1606. https://doi.org/10.1016/j.jmb.2006.11.020

Lynch M. (2013). Evolutionary diversification of the multimeric states of proteins. Proceedings of the National Academy of Sciences of the United States of America, 110(30), E2821–E2828. https://doi.org/10.1073/pnas.1310980110

Lynch, M. and Force, A. (2019). The Probability of Duplicate Gene Preservation by Subfunctionalization. [online] Genetics. Available at: https://www.genetics.org/content/154/1/459

Marchant, A., Cisneros, A. F., Dubé, A. K., Gagnon-Arsenault, I., Ascencio, D., Jain, H., Aubé, S., Eberlein, C., Evans-Yamamoto, D., Yachie, N., & Landry, C. R. (2019). The role of structural pleiotropy and regulatory evolution in the retention of heteromers of paralogs. eLife, 8, e46754. https://doi.org/10.7554/eLife.46754

Marsh, J. A., & Teichmann, S. A. (2015). Structure, dynamics, assembly, and evolution of protein complexes. Annual review of biochemistry, 84, 551–575. https://doi.org/10.1146/annurev-biochem-060614-034142

Miyazawa, S., & Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. Journal of molecular biology, 256(3), 623–644. https://doi.org/10.1006/jmbi.1996.0114

Nido, G. S., Méndez, R., Pascual-García, A., Abia, D., & Bastolla, U. (2012). Protein disorder in the centrosome correlates with complexity in cell types number. Molecular bioSystems, 8(1), 353–367. https://doi.org/10.1039/c1mb05199g

Roth, C., Rastogi, S., Arvestad, L., Dittmar, K., Light, S., Ekman, D., & Liberles, D. A. (2007). Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. Journal of experimental zoology. Part B, Molecular and developmental evolution, 308(1), 58–73.

Shakhnovich, E. I., & Gutin, A. M. (1993). Engineering of stable and fast-folding sequences of model proteins. Proceedings of the National Academy of Sciences of the United States of America, 90(15), 7195–7199. https://doi.org/10.1073/pnas.90.15.7195

Tobi, D., Shafran, G., Linial, N., & Elber, R. (2000). On the design and analysis of protein folding potentials. Proteins, 40(1), 71–85. https://doi.org/10.1002/(sici)1097-0134(20000701)40:1<71::aid-prot90>3.0.co;2-3

Van Zee, J.P., Schlueter, J.A., Schlueter, S. *et al.* Paralog analyses reveal gene duplication events and genes under positive selection in *Ixodes scapularis* and other ixodid ticks. *BMC Genomics* **17,** 241 (2016). https://doi.org/10.1186/s12864-015-2350-2

Vangone, A., & Bonvin, A. M. (2015). Contacts-based prediction of binding affinity in protein-protein complexes. eLife, 4, e07454. https://doi.org/10.7554/eLife.07454

Xiong, P., Zhang, C., Zheng, W., & Zhang, Y. (2017). BindProfX: Assessing Mutation-Induced Binding Affinity Change by Protein Interface Profiles with Pseudo-Counts. Journal of molecular biology, 429(3), 426–434. https://doi.org/10.1016/j.jmb.2016.11.022

Zabel, W. J., Hagner, K. P., Livesey, B. J., Marsh, J. A., Setayeshgar, S., Lynch, M., & Higgs, P. G. (2019). Evolution of protein interfaces in multimers and fibrils. The Journal of chemical physics, 150(22), 225102. https://doi.org/10.1063/1.5086042

Zhang, H., Li, X. J., Martin, D. B., & Aebersold, R. (2003). Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. Nature biotechnology, 21(6), 660–666. https://doi.org/10.1038/nbt827