The Sequence Space of Natural Proteins

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Laura Weidmann-Krebs

aus Karlsruhe

Tübingen

2020

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	12.05.2020
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Andrei N. Lupas
2. Berichterstatter:	Prof. Dr. Oliver Kohlbacher

dedicated to my family

Abstract

Proteins carry out the majority of functions at the molecular level of all organisms. They are composed of amino acid sequences that upon folding assume specific structures, which are essential to perform their function. In contrast, the great majority of randomly generated amino acid sequences fails to fold into a defined structure and is not functional. In order to better understand functional proteins, the aim of this thesis is to determine general features of natural protein sequences by contrasting them to random sequence models. For this, three different approaches are applied.

The first approach focuses on sequence features that are shared among all proteins, resulting in a global consideration of natural proteins. For this, the pairwise similarity between sequence fragments derived from a large data set of bacterial proteomes is analyzed. These similarities are interpreted as distances, indicative of how sequences are distributed over the space of all possible sequences. The results show that the great majority of distances between natural sequences coincide with those between random sequences of the same amino acid composition. The global occupation of sequence space by natural proteins is thus almost random, an observation that contrasts with the widespread concept of sequences organized into dense clusters defined by common descent. In fact, most related sequences share a similarity that is expected from the random sequence model. They are thus not more similar than random sequences, resulting in their wide distribution across sequence space. Most distances between natural sequences that remained unaccounted for by the random sequence model, can be associated with the different use of amino acids in individual proteins. Only few distances are found to be affected by common sequence motifs in non-related proteins. With this, the amino acid composition of individual proteins is demonstrated to be the most distinctive feature that characterizes natural protein sequences globally. Furthermore, common descent and divergent evolution are demonstrated to have no impact on the global occupation of sequence space, while convergent evolution is responsible for specific sequence motifs that are common in natural proteins.

The second approach analyzes the range of sequence similarities that is associated with common descent. In contrast to the first approach that studies the global occupation of sequence space, here, the local one is of interest. For this, sequences in close proximity to individual query sequences are studied. With increasing distance to the query, the likelihood of common descent decreases, becoming uncertain at a range that has been coined the ‘twilight zone’. Previous studies validated common descent by structural similarity

in order to estimate the boundaries of the twilight zone. The approach applied in this thesis determines these boundaries from the statistical significance of sequence similarity, thereby refining its definition.

With the third approach, the characteristic amino acid composition of individual proteins was further studied at a local level. Given that proteins are generally composed of distinct structural and functional parts, their amino acid composition along the entire sequence was expected to fluctuate accordingly. However, the results of a random model based on the amino acid composition of domain-sized fragments are comparable to those of the model based on the composition of proteins. In contrast to the initial expectation, this finding suggests a homogeneous amino acid composition along individual protein sequences. Different reasons for this homogeneity are considered such as fold-specific recombination, topology and genomic context, which could not be associated to this finding. By analyzing the codon composition of protein domains it becomes clear that this homogeneity of amino acids is correlated to a homogeneous usage of codons. This suggests that amino acid composition may be modulated by codon bias, an effect that has been associated with expression level and translation efficiency in other studies. With this approach, structural constraints on amino acid composition could be contrasted with constraints that cause codon bias, two features of proteins that have been analyzed extensively before and are studied here jointly.

Kurzfassung

Die meisten Funktionen aller Organismen werden auf molekularer Ebene von Proteinen ausgeführt. Diese bestehen aus einer Aneinanderreihung von Aminosäuren, welche sich zu spezifischen Strukturen zusammenfalten, die essentiell für die Ausführung der Proteinfunktion ist. Im Gegensatz dazu ist die große Mehrheit an zufällig generierten Proteinsequenzen nicht in der Lage sich zu falten und ist somit nicht funktional. Ziel dieser Arbeit ist es, generelle Eigenschaften natürlicher Sequenzen herauszuarbeiten, um funktionale Proteine besser zu verstehen. Dafür werden in drei verschiedenen Ansätzen natürlich vorkommende Proteinsequenzen mit verschiedenen Zufallsmodellen verglichen.

Der erste Ansatz konzentriert sich auf die globale Betrachtung von Gemeinsamkeiten zwischen Proteinen. Hierzu werden aus einem großen Datensatz, der aus bakteriellen Proteomen besteht, Sequenzfragmente paarweise auf ihre Ähnlichkeit untersucht. Diese Ähnlichkeiten werden als Distanzen interpretiert und sind bezeichnend für die Verteilung der Sequenzen im Raum aller möglichen Sequenzen. Das Ergebnis dieses Ansatzes zeigt, dass fast alle Distanzen zwischen natürlichen Sequenzen mit denen zwischen Zufallssequenzen derselben Aminosäuren-Zusammensetzung modelliert werden können. Die globale Anordnung natürlicher Proteinsequenzen im Raum entspricht somit fast vollständig einer durch ein Zufallsmodell generierten Anordnung. Diese Beobachtung widerspricht der weitverbreiteten Ansicht, dass natürliche Sequenzen zu dichten Gruppen im Raum organisiert sind, die jeweils aus verwandten Sequenzen bestehen. Tatsächlich haben die meisten verwandten Sequenzen eine Ähnlichkeit, die man auch in einem Zufallsmodells erwartet. Somit sind selbst verwandte Sequenzen weit verteilt über den gesamten Raum. Die meisten Distanzen zwischen natürlichen Sequenzen, die nicht modelliert werden konnten, stehen im Zusammenhang mit der spezifischen Aminosäuren-Zusammensetzung einzelner Proteine. Nur wenige Distanzen konnten mit häufig verwendeten Sequenzmotiven in Proteinen in Verbindung gebracht werden. Mit dieser Untersuchung wurde die Aminosäuren-Zusammensetzung auf Proteinebene als das markanteste Merkmal natürlicher Proteinsequenzen herausgearbeitet. Verwandtschaft unter Proteinen hat nachweislich keine Auswirkung auf die globale Anordnung von Sequenzen im Raum, wohingegen Konvergenz zu bestimmten Sequenzmotiven führt, die typisch für natürliche Proteine sind.

Mit dem zweiten Ansatz wird der Zusammenhang zwischen Sequenzähnlichkeit und Verwandtschaft näher untersucht. Dafür werden alle Sequenzen betrachtet, die sich im

nahen Umfeld einer Referenzsequenz befinden. Somit wird im Gegensatz zum ersten Ansatz nicht die Anordnung der Sequenzen im globalen, sondern im lokalen Raum untersucht. Je weiter entfernt sich Sequenzen von der Referenz befinden, desto geringer ist die Wahrscheinlichkeit, dass sie verwandt sind. Ab einer Distanz, bei der die Verwandtschaft unwahrscheinlich wird, spricht man von der 'twilight zone'. In vorherigen Studien wurde Strukturähnlichkeit verwendet, um auf Verwandtschaft zu schließen, und so die Grenzen der 'twilight zone' zu bestimmen. Der Ansatz dieser Arbeit bestimmt diese Grenzen anhand statistischer Signifikanz von Sequenzähnlichkeiten, wodurch diese Definition konkretisiert wurde.

Der dritte Ansatz beschäftigt sich mit der Aminosäuren-Zusammensetzung auf Proteinebene aus einer lokalen Perspektive. Dadurch, dass Proteine aus ungleichen strukturellen und funktionalen Teilen zusammengesetzt sind, ist es naheliegend, dass ihre Aminosäuren-Zusammensetzung mit dieser Heterogenität variiert. Die Ergebnisse eines Zufallsmodells, das die Zusammensetzung von Sequenzfragmenten reflektiert, waren jedoch vergleichbar mit denen eines Modells, das die Zusammensetzung ganzer Proteine reflektiert. Entgegen der ursprünglichen Annahme, wurde somit eine homogene Aminosäuren-Zusammensetzung in ganzen Proteinsequenzen festgestellt. Durch die Betrachtung der Codon-Zusammensetzung strukturierter Proteindomänen wird deutlich, dass diese Homogenität im Zusammenhang mit einer homogenen Codon-Zusammensetzung steht. Diese Beobachtung legt nahe, dass die Aminosäuren-Zusammensetzung von Proteinen unter anderem durch die Codon-Zusammensetzung reguliert wird, ein Effekt, der mit Expressionslevel und Translationseffizienz assoziiert wird. Mit dieser Untersuchung konnten somit Einschränkungen der Aminosäuren-Zusammensetzung durch Proteinstruktur und Codon-Verwendung gegenübergestellt werden, zwei ausführlich untersuchte Eigenschaften, die hier im Zusammenhang analysiert werden.

Acknowledgements

The first to be mentioned is Andrei Lupas, who has encouraged me to deeply understand protein evolution and to go, where no-one has gone before. In my odyssey through sequence space, he has guided my thoughts with his seemingly limitless knowledge and excitement for proteins. We have challenged each others perspectives many times and have, nevertheless, always found common ground - island or not. His constant support and trust have allowed me to do science completely free and to unfold my ideas. For all of this, I am very thankful to you.
In many ways, Oliver Kohlbacher was an asset to my doctoral studies. When I was wandering about, his ability to grab a problem by its roots made it possible to redirect and sort my thoughts within few minutes. His detailed questions of 'how I do things' and 'why I do things the way I do' often got me to my limits and finally made my approach to scientific questions a better one. He was always kind and understanding, which made the exchange with him even more priceless. I thank you for this.
For all their time dedicated to this thesis, I want to acknowledge Patrick Müller and Andreas Dräger. Their efforts and input allowed me to gain a view onto my work from a different perspective. I thank you for all your support and having joined my examination committee.
Whenever I needed help, I could always turn to Tjeerd Dijkstra. I remember countless white board discussions, late-night snack-times and stacks of his revisions on my work. For your open door and the cheerful times, I thank you.
Among my colleagues, Mohammad ElGamacy was always there, even if he had no time or a gel was running. He not only believed in my abilities but also challenged me to make the best of it. I am deeply thankful to you.
The constant exchange with Joana Pereira is essentially how I imagine science to be done. With her clear and open mind she could jump into all sorts of questions, being only one turn of chairs away. I thank you for your inspiration.
It was Jens Bassler, who told me that I will get there. His critical assessments and the many coffee breaks on the terrace with him had made a great impact. Thank you for this.

Finally, I want to express my gratitude to my family, who stands by my side in any situation of my life. Especially to my parents, Florina and Rolf Weidmann, for all their care and love; and to my husband, Andreas Krebs, for being my support in times of struggle and success. I cannot thank you enough.

Contents

Chapter 1

Introduction

One of the most fundamental questions is the question about life itself. How did it emerge and evolve? What mechanisms have led to us being able to breath, think and love? While the answer to this fundamental question is still far out of reach, it is known that all living things are based on a complex system of tiny components at the molecular level. Among these, proteins represent the class of molecules that is responsible for most of the functional diversity. They are essential for brain activity, muscle contraction, digestion, reproduction, development, and most other mechanisms that sustain life. Understanding key features of proteins and their evolution can thus help to better understand life.

Proteins have been thoroughly analyzed at their sequence, structure, and function level. Many key aspects that define natural proteins have been characterized, such as their general composition, basic structural elements and features responsible for their activity. What is more, huge efforts of analyzing natural sequence and structure data have made it possible to comprehend the variety of proteins as a collection of similar units. Grouping proteins into subsets according to their relatedness results into about 10,000 protein families. When further grouping them by structural similarity, only about 2,300 unique protein domains have been observed. The great diversity in natural organisms is thus built around few structural components that may possess different individual characters but are basically the same.

Protein design efforts have demonstrated that there are functional structures that have not been explored by nature. There may be different reasons for why these structures have not evolved, ranging from the conservative progression of evolution to reasons that can be inferred from theoretical descriptions of the possible structure space. Applied strategies that aim to explore the limits to proteins are often brute-force and sample sequence space almost randomly to test for possible structure and function. It is essentially unclear how this space that is useful to proteins is structured and how much diversity has not yet been revealed. A better understanding of natural proteins can help to extrapolate knowledge from existing to possible proteins. This would help to shed light onto evolution and also to help design new proteins with a more rational approach.

With this thesis, I contribute to the efforts of studying proteins at the level of their amino acid sequence. After a brief outline of the biochemical, evolutionary and methodologi-

cal background, current results focused around features of natural protein sequences are reviewed in Section 1.3 to give an overview onto the achievements and open questions in the field. The results are then presented in three chapters, which are further outlined in Section 1.4.

1.1 Background

1.1.1 Protein sequences and structure

DNA to proteins

Life on Earth is built around a diversity of macromolecules. Above all, nucleic acids and proteins are responsible for organizing all molecules to act together as a system. While sequences of *deoxyribonucleic acids* (DNA) store the information of how proteins

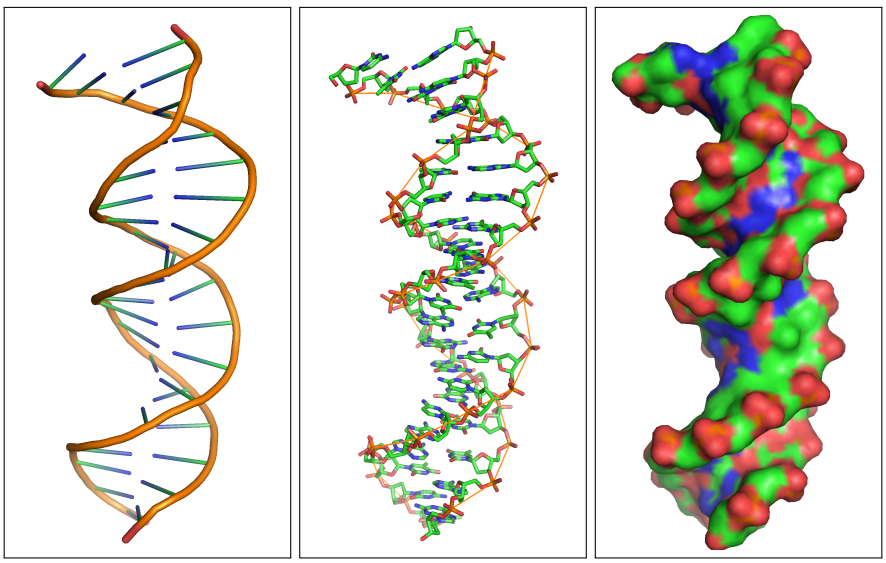

Figure 1.1: DNA structure in three representations: a schematic representation of the double helix and the base pairs to the left, a representation of atoms and atom bonds in the middle and a representation of the surface of DNA. The structure was obtained from the Protein Data Bank (PDBID 1W0T).

are built in the majority of all organisms, the latter are responsible to perform most processes, mechanisms and functions that sustain life.
DNA is composed of two linear chains of consecutive *nucleotides*, that wind around each other and form a double-helix [Watson and Crick, 1953]. Nucleotides are small molecules composed of a deoxyribose, a phosphate group and a nucleobase, referred to as adenine, guanine, cytosine, and thymine, depending on their base. Across the

two chains bases are paired in a complementary way, forming pairs between guanine and cytosine or thymine and adenine. The double helix of nucleotide sequences and the base pairing across the two chains is presented in Figure 1.1. The order in which the nucleotides are arranged along the chain determines the genetic information that encodes the blueprints of the carrier's proteins. DNA can be transcribed into its complementary *ribonucleic acid* (RNA). The transcript of protein coding regions, so-called *genes*, is referred to as *messenger RNA* (mRNA). It is further translated into protein sequences by ribosomes, which are large RNA-protein complexes. This translation process is the connection between the RNA and protein world and has the *genetic code* at its core [Martin et al., 1961]. It represents the encoding of specific *amino acids* by nucleotide-triplets, the so-called *codons*.

Primary structure of proteins

Amino acids are the basic building blocks of proteins. They are composed of an amine group connected to a carbon acid group, that has a side chain attached to one of its carbons. This side chain can be derived from different types of acids that determine the main characteristics of each amino acid. Commonly, there are 20 different types of proteogenic amino acids with different biochemical properties that range from hydrophobic to hydrophilic, polar to neutral, rigid to flexible or small to large. Their chemical structures are presented in Figure 1.2.
Amino acids occur with different frequencies in nature. While abundant amino acids such as leucine and alanine have a frequency of almost 10% in natural proteins, rare amino acids such as tryptophan and cysteine have a frequency of around 1%. The natural amino acid composition of proteins deviates significantly form an uniform frequency of 5% for each amino acid.
Amino acids can be covalently bound to each other through peptide bonds between an amine group of one and a carbon acid group of another amino acid. This way, linear sequences of amino acids are formed. In Figure 1.3 two consecutive amino acid are represented where R_1 and R_2 represent their respective side chains. The order of specific amino acids determines the so-called *primary structure* of a protein and is also referred to as protein sequence. Commonly, it is represented by a string of 20 letters, where each letter stands for a specific amino acid, as shown in Figure 1.2.

Secondary structure

Depending on the order of amino acids in a chain, different biochemical interactions will occur. In order to minimize the free energy, interactions that stabilize the structure are favored. Proteins usually assume a local structure over small sequence stretches, the so-called *secondary structure*, which occurs due to specific backbone interactions, so-called *hydrogen bonds*. These bonds are electrostatic interactions between a hydrogen atom, which is connected to a more electronegative atom such as a carbon, nitrogen or

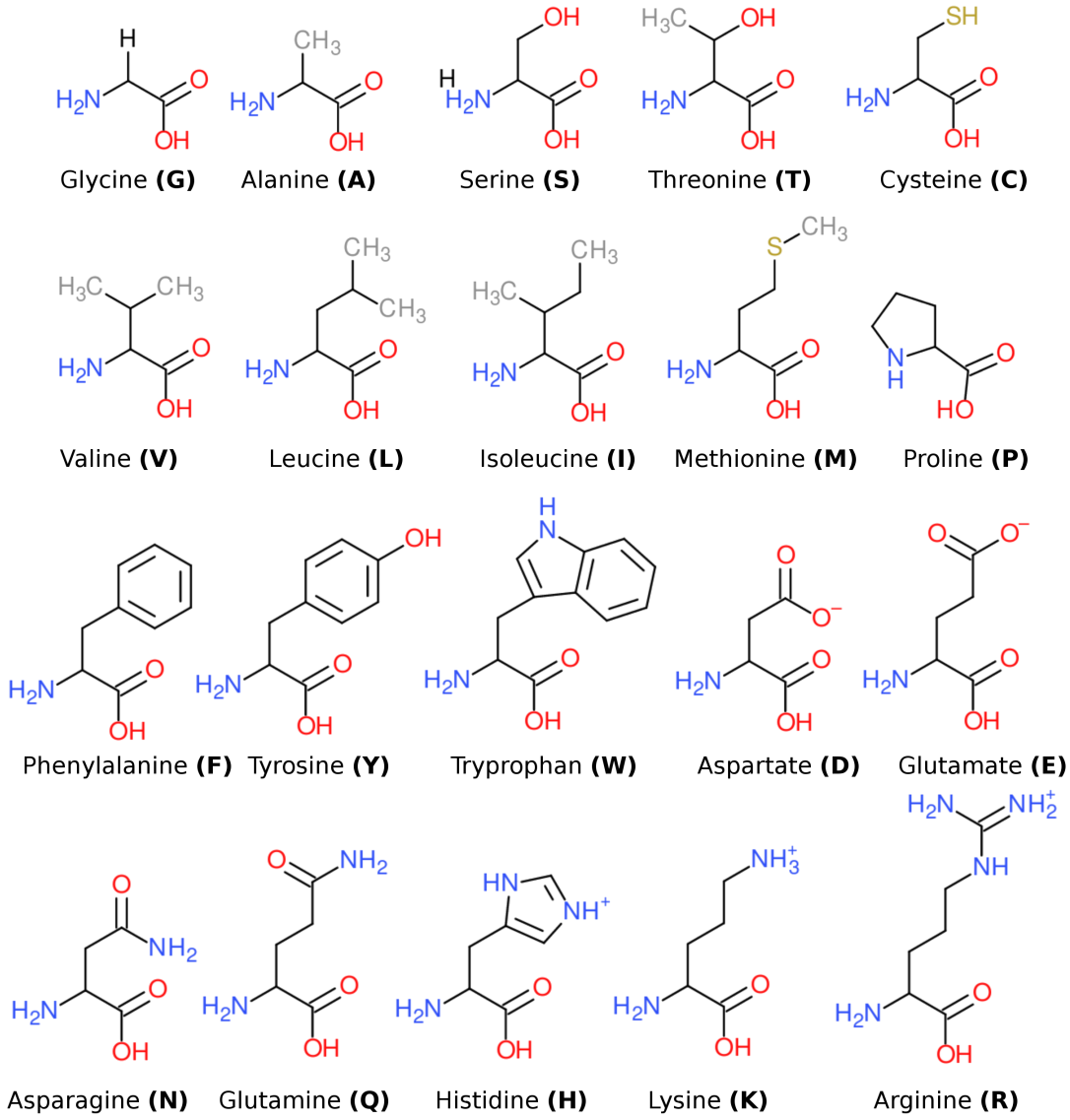

Figure 1.2: Chemical structure of the 20 proteogenic amino acids. The letter in parentheses is commonly used to refer to the respective amino acid in protein sequences.

oxygen, and an electronegative atom with at least one electron pair. The hydrogen atom is partially positively charged and attracts the partially negative electronegative atom to form a hydrogen bond.

There are two common types of secondary structure elements. One secondary structure element is the *alpha helix*, which forms a spiral by hydrogen-bonds between the backbone of the n-th with that of the (n+4)-th amino acid. The other common element is the *beta strand*, which forms sheet-like structures through hydrogen bonds between the

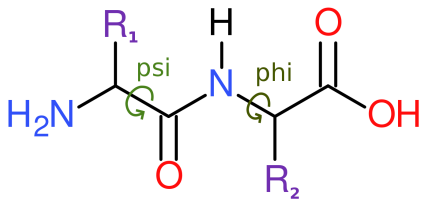

Figure 1.3: Peptide bond between two amino acids and backbone angles psi and phi.

backbones of two distinct beta-strands. In Figure 1.4, the backbone structure of two proteins is depicted that are composed to a large proportion of either alpha helices or beta strands. In alpha helices, all side chains point to the outside of the spiral whereas in beta-strands, the side chains point alternating to one of the sides of the beta-sheet. The structure of alpha helices and beta strands is characterized by specific rotation angles around the axis of atom bonds that form the back bone. These are commonly referred to as phi- and psi-angles and are indicated in Figure 1.3. Apart from these two most frequent forms of secondary structure, proteins contain less common or unspecific structures such as loops that connect secondary structure elements or unstructured regions.

Tertiary structure

The free energy is further reduced as secondary structure elements come together to form the *tertiary structure*, the overall arrangement of an amino acid sequence in three dimensional space. Generally, there are three types of molecular interactions that determine the tertiary structure of protein sequences: First, similar to the backbone interactions that determine secondary structure, hydrogen bonds can also occur between pairs of side chains or between the backbone and side chains. Second, another kind of electrostatic interactions are the *van der Waals interactions*, which occur from spontaneously formed shifts of electrons. Such shifts lead to partially positively and negatively charged atoms, which attract each other. Third, if a reaction between sulfur groups of the side chains of two cysteines occurs, the linear protein sequence is cross-linked by a *disulfide bridge*. The result of this folding is that water is excluded from the structure and most interactions become intra-molecular.
Finally, the *quaternary structure* of proteins is defined by the arrangement of individual protein chains relative to each other. Contacts between two proteins are characterized by specific, mostly hydrophobic interactions between side chains. This allows to exclude water at the interface. Surface exposed residues are rather hydrophilic, thereby being able to interact with water [Chothia and Janin, 1975].

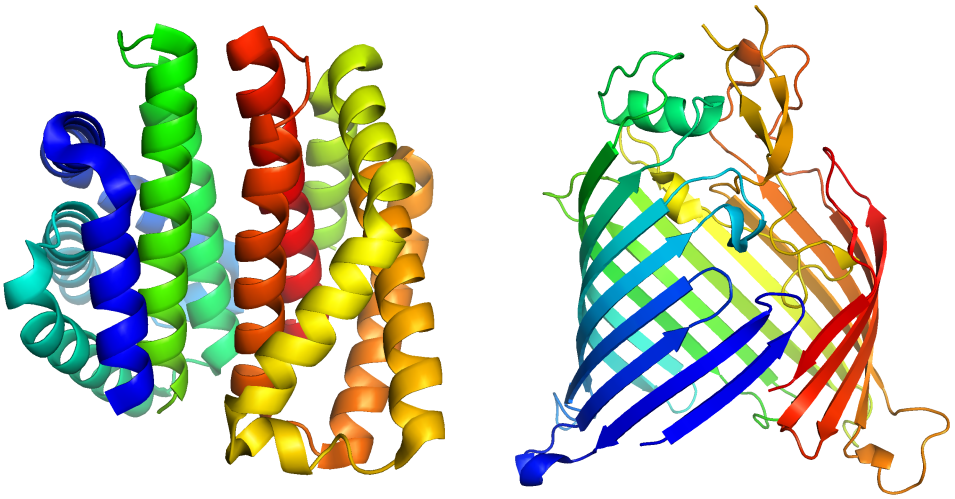

Figure 1.4: Examples of proteins with high alpha helix or beta strand content. The all-alpha structure on the left side (PDBID 4ZVA) represents a globin from *Escherichia coli*. The beta-barrel structure on the right side (PDBID 6QGW) represents a structure of BamA.

Structure determination

The folding process of a protein sequence is a highly complicated and not completely understood process [Šali et al., 1994]. Protein structure predictions aim to determine the structure of a protein from its sequence. Most of the folding information is stored in the primary structure of a sequence, allowing it to fold independently of other factors into its native structure and therefore mostly by interactions between its own amino acids. External factors such as the environment where the folding process takes place or translation speed may facilitate or hinder a protein to assume its native structure. In theory, there should thus be a way to determine the structure given the protein sequence.
The most common methods to predict the 3D-structure from protein sequences are based on the concept of common descent (see Section 1.1.2). They consider related sequences of which the structure is known and map the query sequence onto existing structures. This method is based on the principle that sequences diversify faster than structure [Rost, 2001; Chothia and Lesk, 1986], as for most proteins their arrangements in space is essential to function. However, methods relying on common descent cannot be applied in cases where no structures or related sequences are available. Hybrid or *in silico* methods can overcome this problem.
One of the most common *in silico* methods is Rosetta [Rohl et al., 2004], which maps existing conformations of short peptides onto the query sequence in an iterative fashion. While this approach works well for common sequences and structures, it performs worse for rare cases. So far, there is no method that works reliably for any given sequence. The complexity of the conformational space is gigantic [Ngo and Marks, 1992], even

for small peptides, and it is unclear how to determine which interactions are stronger than others. In the competition of the Critical Assessment of protein Structure Prediction (CASP), this problem is bi-annually addressed, showing that structure determination methods improve but still fail in many cases of special structures [Kryshtafovych et al., 2018].

Domains as the units of protein structures

A feature that most proteins share is their composition of independently folding units, referred to as *domains*. Domain recombination is a common process in protein evolution that has given rise to many combinations of distinct or alike domains within one protein sequence [Chothia et al., 2003]. Determining regions that can be considered as domains, can help to understand a protein structure and function better as a whole.
The length of domains varies between 40 and several hundred amino acids; an average length of a domain is roughly 100 residues [Lobry and Gautier, 1994; Shen et al., 2005; Vinogradov, 2004]. Domains can be classified into a hierarchy based on structure and sequence similarity. Different classification methods are presented in Subsection 1.2.3.

1.1.2 The role of evolution

There are many evolutionary mechanisms and aspects that together have shaped the protein world we see today. Here, a few key aspects of protein evolution are illustrated, that are essential to comprehend strategies and arguments used in this thesis.

Heredity and duplication

The evolution of present-day proteins proceeds to a great amount by duplication and diversification [Zhang, 2003]. Sequences that have been duplicated and therefore share a common ancestor are referred to as *homologs*. Homologs are further specified into two distinct classes, depending on the mechanism that caused the duplication. The first most obvious mechanism is heredity. Through reproduction, whole genomes are being copied and passed on to the next generation. Genes of the same origin that occur in different individuals are referred to as *orthologs*. These diversify independently and generally maintain the same function. Other mechanisms that modify the genome itself, lead to duplication of gene material within a genome. These inner-genome duplications result into *paralogs*. Duplicates within one genome can possess the same basic function, which often diversifies over time.
In many cases, the relatedness of protein sequences is of great importance for understanding coherences between distinct proteins. When studying evolution and the origin of life, knowing which sequences have descended from a common ancestor is crucial to establish relationships between proteins and whole organisms.

Diversification

All existing sequences diversify over time and previously identical duplicates may not resemble each other after a long time. Mutations on the DNA-level lead to this diversification of the original sequences. These mutations can be of different kinds.
A *single-nucleotide polymorphism* (SNP) occurs if one of the four DNA-nucleotides is replaced by another one. Human genomes, for example, contain one SNP in roughly 100-300 nucleotide bases [Shifman et al., 2002]. SNPs are not always coupled to the protein level, given that some codons translate to the same amino acid. If they are visible in the protein sequence, they are referred to it as a *point mutation* on the protein level. SNPs occur frequently and are often tolerated, as their changes at the protein level are not always severe. Especially substitutions with amino acids of similar biophysical properties are often tolerated (see Section 1.2.2).
Modification can also occur by *insertion* and *deletion* (InDel) of short or even large chunks of sequence. These do not always occur in the same frame of the respective gene if they are of a length that is not divisible by the codon length of three nucleotides. Thus, translation can be shifted into different reading frames, which often causes major damage to the original gene.

Fitness, natural selection and evolutionary pressure

The general concept of an organism's *fitness* is associated with its ability to survive and reproduce. This concept can be transferred to the fitness of individual proteins according to their contribution to the overall fitness. The fitness of a protein sequence is first of all dependent on its *function*. In order to fulfill its specific biochemical task, a protein must function correctly. Through diversification, sequences change along with their fitness, a dependency that is studied on high-dimensional *fitness landscapes*. Therein, neighboring sequences are inter-reachable by common mutations and the height in this landscape is interpreted as fitness. Natural sequences diversify and traverse over this fitness landscape. However, not all possible sequences are viable and only those with an adequate fitness may survive [Kondrashov and Kondrashov, 2015]. The famous term

> *"Survival of the fittest"*
> [Spencer, 1864]

summarizes this concept of *natural selection*. However, this term evokes an image of an active choosing process concerning which sequence will survive. Actually, natural selection acts on the other end of the fitness spectrum, by removing or pruning the unfit. While there is no active mechanism that changes the gene pool by the existence of successful individuals (indirectly by enhanced reproduction perhaps), it is the removal of deleterious or unfit individuals that impacts the gene pool. Even though after the removal the remaining fitter part of a population may be surviving, it is the dying of the unfit that has seemingly "selected" the fit.

Furthermore, natural selection is not directed and does not optimize proteins to be perfectly fit, given that mutations follow a rather random process. Indeed, most natural proteins are on the edge of unfoldedness and are easy to break with few random mutations. They are fragile, "just" functional machines, far away from *in silico* designed hyperstable protein structures [Elgamacy et al., 2018]. This mechanism is well expressed with Francois Jacobs words:

> *"Evolution is a tinkerer, not an engineer"*
> [Jacob, 1977]

Therefore, even though better sequences may exist for one specific task, they are not actively sought out by evolution. Natural sequences rather tend to linger in a neutral zone [Kimura, 1983] by maintaining their basic function without noticable change. Randomly traversing over the fitness landscape at the expense of possibly deleterious mutations only occurs if necessary [Aguirre et al., 2018].
Exhaustive mutagenesis occurs mostly in systems that are in existential need to adapt. In case of near-death, some bacteria are known to start randomly mutating, testing many usually non-expressed protein sequences with the aim to rescue themselves [Foster, 2005]. Another example are viruses that need to escape the host's immune system by constantly mutating [Kamp and Bornholdt, 2002]. This need to change in order to survive and reproduce successfully is referred to as *evolutionary pressure*. Under more standard conditions the greatest evolutionary pressure lies on sequences that are essential to the organism. Residues that fulfill the essential function need to be conserved, allowing to estimate evolutionary pressure according to sequence conservation.

Contingency and epistasis

The chance of a major evolutionary event, such as a random but advantageous mutation, is only very small [Bershtein et al., 2017]. However unlikely such an event may be, if it does occur and gains acceptance, every following event will build upon it. Hence, evolution is a contingent outcome and possibly many other life forms could have arisen if other mutations had occurred [Starr et al., 2017].
This strong dependence on existing sequences can be clearly seen on a small scale, when investigating interacting sites within a protein. Residues that are in close proximity after the folding process influence each other. Only specific pairs of amino acids are preferred for such interactions, which depend on their biophysical properties. If one of the residues mutates to an unfavorable amino acid, there is an evolutionary pressure on both sides for a compensating mutation. Such correlated sites are *evolutionary coupled*, as their evolution depends on each other.
This dependence of follow-up mutations on previous mutations is associated with the concept of *epistasis*. A sequence that traverses through sequence space by accumulating mutations (for example in its natural evolution) may find some mutation paths to be accessible while others are not, dependent on certain previous mutations. Hence, even

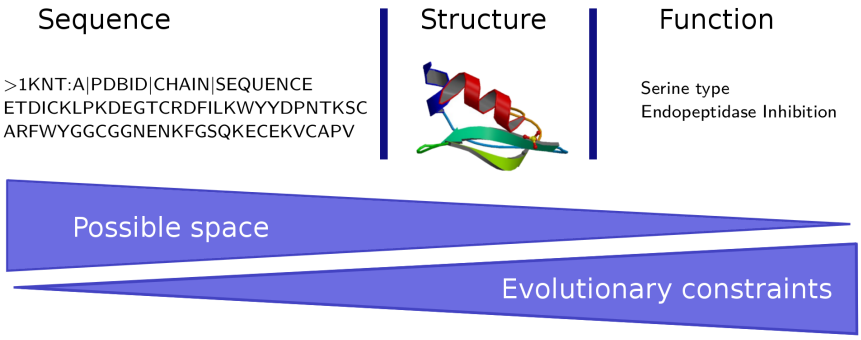

Figure 1.5: Relationship between sequence, structure and function. Function experiences most constraints as its change or loss is often deleterious. The same function can, however, be accommodated by a variety of possible structures, leading to less constraints on the structure level. The number of possible protein sequences for a specific structure is generally of a large magnitude.

if two proteins fulfill the same function, dependent on where in sequence space they are located, different opportunities for mutation and evolution are available to them. Therefore, the fitness landscape is rather referred to as an *adaptive landscape* [Wright, 1932], since the landscape may change with each mutation of the overall system. Sometimes the mutation of a key position opens doors to many other similar sequences or it entrenches a sequence in a small subarea with many constraints. The amount of freedom a sequence experiences to evolve has been termed *evolvability*, which may be regarded as the opposite of general constraints on a sequence.

Conservation and relatedness

The dependency on specific functions of a protein results in evolutionary pressure, which reduces the evolvability of a protein. Mutations of residues carrying this essential function tend to be deleterious. Homologs often demonstrate sequence conservation corresponding to these residues with essential function. Relatedness can thus often be inferred from conservation and increased *sequence similarity*. Several studies classify proteins or domains according to their conserved residues (see Subsection 1.2.3).
Except for key residues, sequence similarity often degenerates over time, leaving no significant overlap behind. In [Rost, 1997] it is noted, that only few residues (3-4%) are actually important for maintaining the essential function. These can in general be associated with self-maintenance of a protein such as folding, solubility and stability or to the interaction with other molecules resulting in its catalytic activity, ligand binding, conformational changes and interface formation. While the function of proteins emerges from the interaction with other molecules, the purpose of maintaining a folded structure is a necessity to fulfill this higher-order function. That is why function is more important than structure and is therefore also more constrained.

Structure space

Evolutionary constraints act strongest on the functional, then on the structural and only last on the sequence level, as depicted in Figure 1.5. This effect correlates inversely with the number of possible sequences, as many different sequences can fold into the same structure and different structures can accommodate the same function. Also, among related proteins, structure is more conserved than sequence, given that function is often very dependent on the shape of protein. Even after having accumulated many mutations on the sequence level, structure tends to persist [Rost, 2001; Chothia and Lesk, 1986]. Homology is thus often inferred from *structure similarity* of a certain degree [Schneider et al., 1997].
Furthermore, relatedness is not the only reason, for similar structures. Given that there are only two commonly used local structures, namely alpha helices and beta strands, the number of observed topologies of consecutive elements is limited compared to sequence space. Structure space is frequently represented by four main classes (all-alpha, all-beta, alpha/beta, alpha+beta) according to their secondary structure content (see Subsection 1.2.3). It can thus rather easily be discretized according to secondary structure.
Decades of bioinformatic work have further classified the variety of domain structures, into so-called *folds*, where not only the frequency and order of secondary structure elements play a role but also the overall structure of the domains. Although the amount of sequence and structure data still increases exponentially over time, the number of observed folds has stagnated to increase [Söding and Lupas, 2003]. Only about 2,300 folds are registered in the ECOD database [Cheng et al., 2014], a number that varies between classification methods but not substantially (see Subsection 1.2.3). It is essentially unclear why the folds we observe today have emerged and how many other folds could potentially emerge from unsampled sequences.

1.1.3 The role of convergence

Functional convergence

The emergence of similarities from unrelated sequences is referred to as *convergence*. There are several possible structures and even more sequences that can fulfill the same function, allowing functional convergence to occur. For example, there are at least five known peptides with completely different structures, that all bind RNA [Alva et al., 2015]. Very specific functions that impose strong constraints on sequences are globally less frequent and their convergence is also less probable.

Structural convergence

The number of possible domain-sized sequences is of the order 20^{100}, corresponding to the 20 proteogenic amino acids and an average domain size of 100 residues. Given the constraints of the backbone angles phi and psi of a protein sequence, the number of

possible structures (with an upper boundary of 2^{100}) is relatively small compared to the size of the possible sequence space, leaving room for structural convergence.
On the scale of secondary structure elements, it is undisputed that alpha helices and beta sheets have emerged convergently. However, the larger a structure of interest, the less probable is a convergent emergence by chance events. Like the dominant secondary structure elements, there are some larger and more complex structures, that are predicted to have emerged multiple times. Special forms of beta barrels are discussed to have arisen convergently [Franklin et al., 2018].

Biophysical constraints

There are elements in protein sequences that can be used to extract structural information, as proteins face specific *biophysical constraints* that force them to use a restricted set of amino acids at certain positions. Given that proteins are macromolecules composed of atoms and bonds that act according to physical laws, recurrent phenomena in proteins can also be traced in the protein sequence.
The amino acid sequence of a folded protein, for example, requires to possess a distinct free energy minimum upon folding and also to have this minimum being reachable by accessible transition states, i.e., to have a folding path without high energy barriers [Šali et al., 1994]. While the folding process has so far not been successfully decoded from protein sequences in general, sequence patterns that lead to a locally low energy state have been studied. Some of the major sequence-structure relationships are reviewed in Subsection 1.3.2.

1.1.4 Complexity of the sequence space

The combinatorial space of amino acid sequences is gigantic. Any analysis of such a large space struggles with the problems that are examined in the following paragraphs.

A space of astronomical size

Characterizing natural sequences is a problem of great complexity. The sequence space spanned out by sequences of a specific length N has a size of 20^N, which is growing exponentially with sequence length, as depicted in Figure 1.6. Each residue in a sequence represents a dimension in the space of possible sequences, that can assume one of 20 possible values. Thus, longer sequences belong to a space of higher dimensionality. Even for short peptides the number of possible sequences assumes a value of astronomical order [Kondrashov and Kondrashov, 2015; Grigoryan and Degrado, 2011]. Already the volume of the sequence space of 20mers, which is in the order of 10^{26}, assumes a size that is not possible to handle efficiently.

Common sequence length

Generally, protein sequences do not have the same length. Even if they are dissected into their constituent units of folding, their domains, sequence lengths still mostly range between 30 and 150 residues. Sequences of different lengths cannot be mapped onto the same sequence space and must thus be truncated to a common length. As the average domain length is about 100 residues (see Section 1.1.1) this length is used predominately as sequence length when investigating relationships of domain-sized sequences in this thesis. For this, whole protein sequences are dissected into fragments of the respective length and further analyzed. In this context, partial protein sequences are referred to as *fragments*.

Sparsity of the occupied sequence space

All the time that has elapsed since the big bang is not sufficient for nature to have explored all possible 100mers. Hence, there is a sparsity of the natural occupied sequence space simply due to the small amount of time that does not allow to have explored the entire space [Strait and Dewey, 1996]. In the following the relationship between sparsity and fragment length is demonstrated for a large natural data set comprising 1,307 bacterial genomes (see Subsection 2.2.1). The progression of sparsity (Figure 1.6: A) with fragment length can be indicated by the number of distinct sequence fragments that occur in the natural sequence data. Starting with fragment length 6, the number of distinct natural fragments is smaller than the size of the respective *k*mer space. This progression of sparsity with increasing fragment length can be plotted relative to the size of the possible space (Figure 1.6: B), indicating the fraction of occupied space by natural sequences. As noted before, sequence space of fragments longer than five residues is not entirely covered by natural fragments.
This sparsity becomes apparent when distribution all observed 100mers in nature over the possible space: The RefSeq database [Pruitt et al., 2007] of the National Center for Biotechnology Information (NCBI) consists of more than $2.1 \cdot 10^8$ protein sequences [NCBI, accessed 2019-11-07]. Assuming an upper limit of 10,000 residues per entry, an upper limit of $2.1 \cdot 10^{12}$ 100mers in natural sequence data can be assumed. Further assuming that all these 100mers had different sequences, they would cover $1.7 \cdot 10^{-116}\%$ of the possible space, implying that sequence space of 100mers is only very sparsely occupied.
This effect of sparsity in mainly dependent on the amount of used data. The progression of sparsity for random sequences is almost identical to that of natural sequences as depicted in Figure 1.6.

Error caused by sparsity

The sparsity of the sequence space occupation results into a stochastic error concerning the observed peptide frequencies. The sparser the occupation of the entire space by

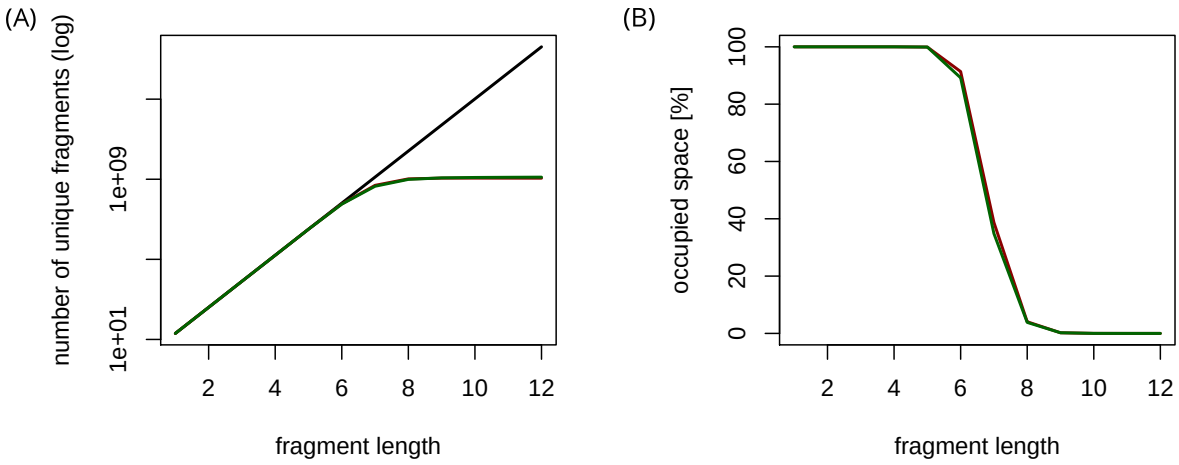

Figure 1.6: Sparsity in sequence space as a function of fragment length. With increasing fragment length (dimensions of sequence space) the amount of occupied space by natural sequences (green) decreases relative to the possible space. Starting with 6mers not all possible fragments have been observed in the used natural data set of 1,307 bacterial genomes. The sparsity progresses almost identically for random sequences (red) of a data set with the same size.

a given data set, the greater is the error. This error is caused by the finite sampling of the data relative to the size of the space. As a result, neither natural nor random sequence data sets can reproduce peptide frequencies accurately. This implies that the 100mer frequencies of half of the RefSeq database does not result into the same 100mer frequencies of the other half of the data base.

The transition from exhaustive occupation to sparsity of sequence space at a fragment length of six marks the point, where the stochastic error significantly affects the observed peptide frequencies.
For a fragment length of five or less, the stochastic error is negligible, allowing to extract differences in the peptide frequencies between natural and random sequences according to Equation 1.1.

$$D(f) = P_{\text{nat}}(f) - P_{\text{rand}}(f) \tag{1.1}$$

Studies of the exhaustive sequence space occupation have been performed for fragments up to a length of five residues [Poznański et al., 2018; Lavelle and Pearson, 2009]. Therein, the authors focus on characterizing fragments whose frequencies deviate from their expected frequency under random conditions. Some 5mers could be revealed to be over-represented others to be under-represented in nature, where over-represented fragment occur in functional sites [Poznański et al., 2018] or are associated to alpha helical structure [Lavelle and Pearson, 2009]. Due to the here illustrated stochastic error for longer fragments, the approaches were not applicable to longer fragments.

The amount of stochastic error cannot be direly inferred from the natural data set as it is mixed with these deviations arriving from natural constraints, which are of main interest. However, not only the natural data set is affected by stochastic error; it applies equally to random sequences and can thus be extracted from those. The amount of overall stochastic error E can be extracted as the cumulative difference between the observed $P_{\text{obs}}(f)$ and expected frequency $P_{\text{exp}}(f)$ according to a closed form model of peptide frequency (see Section 1.2.1). This sum is derived over all fragments f with a length of l residues.

$$E(l) = \sum_{f \in A^l} |P_{\text{obs}}(f) - P_{\text{exp}}(f)| \tag{1.2}$$

With this, a sample of random sequences with the same size as the natural data set can be investigated for its deviation from an expected frequency as defined by a closed form model. The resulting value of $E(l)$ corresponds to the stochastic error (Figure 1.7: red).

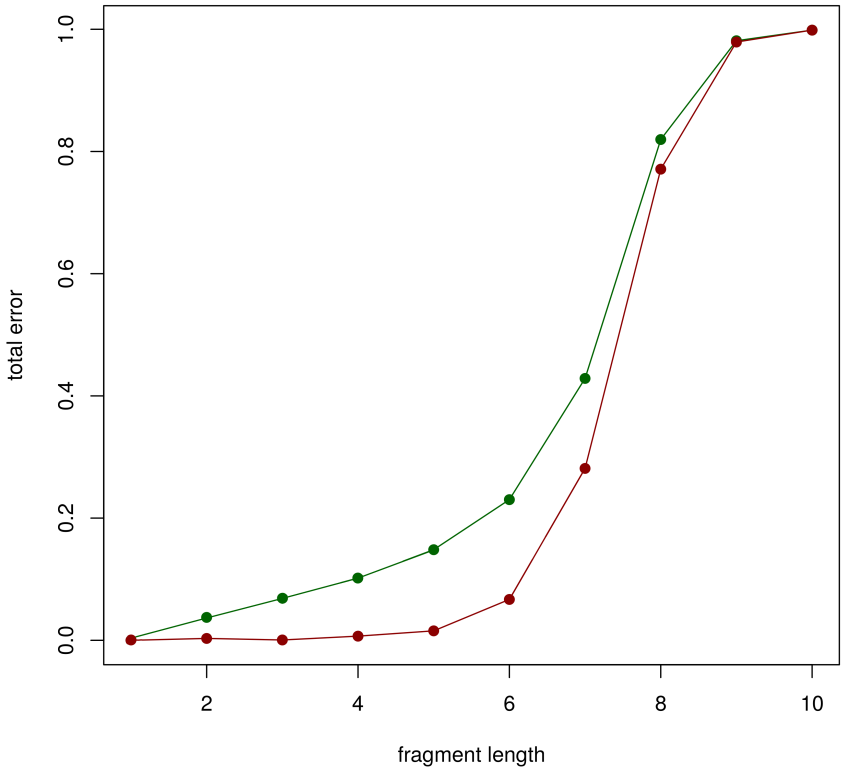

Figure 1.7: Accummulated error over peptide frequency due to finite sampling. Until 5mers, the difference between the natural and random peptide frequencies is purely related to natural biases (green). For longer fragments the stochastic error (red) increases, thereby leading to errors in the assessment of relative peptide frequency between natural and random sequences. For fragments of length 10, this stochastic error is reaching almost 100%.

The main difference between the sampled random sequences and the closed form random model is, that sampled fragments have an abundance that is a natural number such as 0, 1 or more. In the closed form model, a fragment can be expected to occur 0.5 times, a criterion that cannot be met by the sampled random sequences. With increasing fragment length but constant data size, more fragments will be expected to occur with a frequency below 1. This expected frequency converges towards 0 for longer fragments, causing the stochastic error to reach almost 100% for a fragment length of ten or above.
The deviation of the natural peptide frequency to the closed form model as presented in Equation 1.2 is plotted in green next to the stochastic error. For a fragment length below six residues, the stochastic error is almost zero, while the natural deviation is reaching 18%, a value that can be associated to natural constraints. For 6mers up to 8mers, the natural deviation is still notably higher than the stochastic error, however for 10mers it is also reaching an error of almost 100% as the sample of random sequences.
Peptide frequency is thus not a comprehensive measure of how natural protein sequences differ from random sequences for sequence lengths above five residues. To overcome this problem of sparsity, a distance-based approach is presented in Chapter 2.

1.2 Methodology and Materials

1.2.1 Random sequence models

Randomness and information content

In general, any form of information that possesses no repetitive structure is referred to as being *random*. Information in the context of protein sequences occurs at each position in the sequence by assuming specific states (amino acids), indicated by distinct letters. The *information content* $i(f)$ of a sequence f indicates the amount of information it possesses relative to all other sequences. It is derived as the logarithm of the inverse probability $P(f)$ to occur in the respective data set:

$$i(f) = \log_2 \left(\frac{1}{P(f)} \right) \tag{1.3}$$

The information content can be interpreted as the value of a certain piece of information. If there is sun in 99 of 100 days in Tübingen and it rains only in 1 out of 100 days, the information that it has rained in Tübingen yesterday is more decisive than the information that there was sun. Similarly, a sequence that occurs frequently (such as the P-loop motif), has a lower information content than a less abundant sequence (e.g., a specific linker between two domains) or a cysteine-rich sequence.
In colloquial language frequent events are often referred to as informative, given that more data exists for them, due to their frequent occurrence. However, it is more correct to refer to their observed *redundancy*, given that their information content is in fact

decreased. This redundancy is often quantified by the or *Shannon entropy H* of an occurrence. It corresponds to the average number of bits that are needed to encode a character a in a given message.

$$\begin{aligned} H &= \sum_a P(a) \cdot i(a) \\ &= -\sum_a P(a) \cdot \log_2(P(a)) \end{aligned} \tag{1.4}$$

The smaller the entropy the higher is the redundancy, which indicates a repetitive structure in the underlying data. A higher entropy in random than in natural sequences is thus an indication of redundancies specifically occurring in natural sequences.

Random sequence models

The most basal model considers random sequences of the 20 proteinogenic amino acids, in which each occurs with an equal probability of 5%. This model is here referred to as the E-model. It is known to approximate natural sequences only poorly [Strait and Dewey, 1996].
This is hardly surprising as natural amino acid frequencies in fact range between 1% and 10%, a bias which is associated with metabolic pathways, bio-availability, and codon frequency. Depending on the used natural protein sequence data, the frequency of different amino acids fluctuates. Generally, leucine, alanine and glycine are abundant amino acids with a frequency of 8-10% whereas tryptophan, methionine, histidine, and cysteine are rare amino acids with a frequency of 1-3%. The remaining 13 amino acid possess an intermediate frequency. In order to account for the naturally observed amino acid frequency, the overall frequency in the respective data set is used as background amino acid frequency in the random sequence model, which here is referred to as the A-model. However, the overall composition does not account for compositional fluctuations coming from more specific origins. Therefore, other models are used that factor in composition at increasingly local levels. The composition of different genomes, for example, varies with GC-content and environmental influences [Fukuchi and Nishikawa, 2001; Fukuchi et al., 2003]. This effect can be accounted for by using random sequences with the composition of the individual genomes, here referred to as the G-model.
With an increasingly local focus, compositional bias can also be accounted for at the level of proteins [Lee et al., 2006; Cedano et al., 1997]. Due to the different environments, functional constraints and constraints from differential codon usage, proteins have different amino acid compositions that deviate from the composition at the genome level. When accounting for the composition of individual proteins, the random model is here referred to as the P-model.
An even more local consideration of amino acid composition is achieved when using the composition of individual domains. Such a model accounts for inner-protein fluctuation of composition after domain recombination for example. The generation of such a model is not straight-forward as often the boundaries of domains are not well-defined and the handling of sequences that are not part of a folded domain is unclear. Instead, the natural

Table 1.1: Random sequence models. The most basic E-, A, and T-models, incorporate features that are spread homogeneously over the whole random data set. The E-model uses an equal propensity for each amino acid, the A-model is based on the observed natural amino acid composition and the T-model incorporates the overall dipeptide frequency of the natural data set. The more advanced G-, P-, L and D-models are based on the context-specific composition in natural sequences on the level of genomes, proteins, domain-sized fragments or real domains. In order to compare the homogeneity of composition between two domains, the D_2-model is used to reflect the composition of two combined domains. The L1-, L2- and L3-models are used to include the contributions of homologous and convergent sequence similarity.

model	natural feature	class of feature
E	natural amino acid alphabet, equal propensity for each letter	single, overall
A	overall amino acid composition	features
T	overall dipeptide frequency	
G	composition of individual genomes	
P	composition of proteins	context-specific
L	composition of domain-sized fragments	composition
D	composition of domains	
D_2	combined composition of two domains	
L1	L-model + homology sequence bias	mixed models that
L2	L-model + analogy sequence bias	incorporate
L3	L-model + homology and analogy sequence bias	sequence bias

composition of domain-sized fragments has been incorporated into the random model as presented in Chapter 2. This model is further referred to as the L-model and it can be applied to arbitrary sequences. In Chapter 4, a different approach based on isolated domain sequences is presented. Random sequences that are obtained with the premise to reflect the natural composition of domains, are referred to as the D-model. A specific derivation of the combined composition of two domains is therein used and referred to as the D_2-model.
All random models used in this thesis are listed in Table 1.1. In Subsection 2.4.2 specific models that incorporate sequence and composition information of natural protein sequences are used to distinguish between the effects of homology and convergence. They are presented in that context.

Constructing samples of random models

Due to the sparsity caused by finite sampling, natural sequences comprise a certain amount of stochastic error (see Section 1.1.4). In several analyses performed in this thesis (e.g., Subsection 4.3.1 and Section B.2) this error was hindering the analysis and an equivalent sample of random sequences was required to account for the magnitude of

this stochastic error. Samples of random sequence models were reproduced by generating data sets with the identical size and sequence lengths as the original, natural data. Thereby, the samples of random sequence data comprised the same amount of stochastic error than the natural data set.
The A-model is based on the underlying amino acid composition of a given data set. Randomized data for this model was obtained by randomly shuffling all amino acids of the natural data. Thereby, protein lengths were maintained and the number of amino acids stayed exactly the same. For these permutations, the Mersenne Twister algorithm mt19337 of the C++ 14 std library with the standard seed was used.
For the E-model, was produced the same way as the A-model. The only difference is that the natural data set was replaced by writing over all valid amino acids with the 20 possible amino acids in lexicographical order.
To account for genome or protein composition, amino acids were shuffled within the context of genomes or proteins. For the G-model, valid amino acids within each genome were permuted. For the P-model, those within each protein were permuted.
For the L-model, natural fragments of length 100 were shuffled. In contrast to the previous random models, generating a single randomly shuffled data set is not computationally convenient since storing an instance of all shuffled 100mers would increase the data size approximately 100-fold. Therefore, 100mers were shuffled on the fly. The implementation of the more specific random models (D_2 and L1-3) is presented along with the analysis and results.

Closed form A-model

In several parts of this thesis, I refer to the closed form of the A-model to infer the expected peptide frequency. This expected frequency of a sequence fragment f is derived under the assumption that all positions in the sequence are independent from each other. Through the multiplication of the observed amino acid frequencies in the whole data $P(a)$ for each amino acid in the fragment, the closed form for the expected frequency is given by:

$$P(f) = \prod_i P(f(i)) \tag{1.5}$$

As this expected frequency assumes independence of all positions in the sequence, contrasting it with the frequencies of natural sequences can identify short-range dependencies between positions. The expected abundance of a sequence fragment f can be derived from this closed form by multiplying the probability by the number of all fragments N of length $|f|$ in the respective data set.

$$E(f) = P(f) \cdot N \tag{1.6}$$

This formula is used in word count analysis (see Section B.2) or to generally to normalize for the expected occurrence under the assumption of independence.

1.2.2 Protein sequence alignment

In many cases, protein sequences are aligned to each other in order to study their similarity. In an alignment, residues of one sequence are assigned to those of the other sequence in a consecutive way, such that their similarity, as captured by a *similarity score*, is maximized under a set of constraints. The most common application for alignments is to search for related sequences. Other examples may be the comparison of secondary structure elements or convergent features [Lee et al., 2006].
In Figure 1.8 an example of a sequence alignment is depicted. Therein, the sequences of two helical structures (PDBID: 2LFR and 4KP4) are aligned using HHpred from the MPI Bioinformatics toolkit [Zimmermann et al., 2018]. The aligned protein sequences

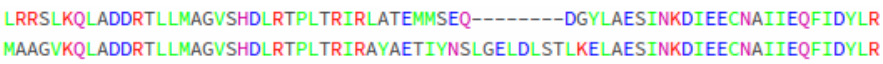

Figure 1.8: Example for a sequence alignment.

are colored according to the biochemical properties of the corresponding amino acids. Residues in the same column are referred to as being aligned. In the middle of the upper sequence, there is a gap, indicated by dash-characters. Only the right and left part of the bottom sequence is aligned to the upper sequence. The strategy to find the alignment that maximizes the sequence alignment score is described elsewhere [Altschul and Erickson, 1986]. Here, I focus on the parameters for alignments as these are more relevant to understand the generation of distances between sequences as used in this thesis.
Ideally, alignment constraints embody the frequency of specific occurrences among natural sequences. Depending on the sequences being compared and the underlying reason of the alignment, it may be reasonable to use different constraints.

Amino acid similarity

Some amino acid substitutions are more often selected for than others. Conserved sequence patterns across homologs preserve such specificity and can be used to score amino acid substitutions in sequence alignments. The BLOSUM scoring matrix [Henikoff and Henikoff, 2000] accounts for pairwise amino acid substitutions by summing over the replacement frequencies P_{ab} of amino acids a and b in conserved blocks of aligned sequences and normalizing it by the expected replacement frequency of their individual occurrence P_a and P_b.

$$B(a,b) = 2 \cdot \log_2 \frac{P_{ab}}{P_a \cdot P_b} \tag{1.7}$$

For this, the conserved blocks were derived from alignments at different levels of conservation, resulting into a series of BLOSUM matrices. In an alignment where a BLOSUM matrix is used, the underlying hypothesis is that the aligned sequences come from a conserved region of the respective conservation level.

Gap penalties

The simplest alignment would be to align residues position-wise. Thereby, the n-th position of the first sequence is aligned to the n-th position of the second sequence. When sequences have different lengths, the terminating residues of the longer sequences would not be aligned to anything. This requires to align a *gap* to a residue. Gaps are penalized with a negative score, which is set according to observations in natural sequences. In this thesis, I use the model of an affine gap cost:
To avoid gaps occurring due to technical issues when using sequences with unequal lengths, gaps at the beginning and end of sequences can be chosen to not be penalized. This allows to shift the sorter sequence within the range of the longer sequence for example.
Gaps are accounting for a common mechanism in the evolution of natural sequences, which is the insertion or deletion of sequence material (see Section 1.1.2). Depending on the rate of such InDels in natural sequences, alignments are penalized with a *gap opening penalty* every time a new gap occurs of aligned residues in one sequence and gaps in the other sequence. InDels occur frequently in natural sequences and their lengths is estimated to be generally shorter than four base pairs [Bhangale et al., 2005]. The distribution of InDels longer than four base pairs is roughly linear. This linearity can be reflected by using a constant small cost for all consecutive gaps after the opening gap, referred to as the *gap extension penalty*.

Global and local alignment

In a global sequence alignment all residues of both sequences are aligned and result into an overall score the similarity between the compared sequences. The most common algorithm to derive the maximal score in this setting is the Needleman-Wunsch algorithm [Needleman and Wunsch, 1970]. When comparing protein sequences of the same superfamily, this approach is reasonable to use, as major divergence is generally not expected among closely related sequences. A global alignment is also applicable in cases, where the overall similarity between sequences is aimed for.
The overall similarity between two sequences is not always a productive measure, as in some cases only shorter regions, such as domains, may be related to each other. Comparing unrelated parts of the sequences would only decrease the score. In this case, it is reasonable to extract a score over a local stretch of both sequences and to neglect terminating sequences. The highest score over all local regions is derived and chosen as the score of the alignment. The most common approach for extracting such local similarity is the Smith-Waterman algorithm [Smith et al., 1981].
In this thesis both local and global alignments are used and contrasted to each other. A global alignment compares the overall similarity of sequences. In terms of sequence space and the approach of using fragments of the same length, a global alignment enforces an alignment that includes the entire sequences. In terms of evolution and the

task to identify related sequences, a local alignment is more sensible as it finds the most pronounced similarities anywhere in the compared fragments.

Combinatorics

Sequence alignments are indicative for sequence relationships in many cases. However, from a modeling perspective, the resulting distribution of scores contain many artifacts that are not reproducible by simple random models. This is due to several technical aspects that relate to the combinatorics of possible alignments and their respective scores [NCBI, accessed 2019-11-17]. Among these are edge effects of natural sequences, the use of different amino acids in the beginning and end of protein sequences, and the combinatorics of gapped, especially local alignments.

Used tools

For efficient alignments, the open source C++ library SeqAn [Reinert et al., 2017] was deployed. The presented results were derived using the 2.4 release of SeqAn. This enabled fast and parallel alignment of many sequences.
In cases where the homology between sequences needed to be established, tools of the HH-suite [Söding, 2005] were used. Through the generation (HHblits) and comparison (HHalign) of Hidden Markov models, confidently homologous sequences could be revealed. Version 3.0.3 of all HH-suite tools was used to derive the results in this thesis.

1.2.3 Protein databases and classification methods

There are several protein databases that use different criteria and data to classify protein sequences according to homology and structure. Here, some databases are outlined in detail and an overview of the classification results of other common databases is given in Table 1.2.

Structural Classification of Proteins - SCOP

According to the Structural Classification of Proteins database (SCOP) [Murzin et al., 1995], there are four major structural classes of proteins: (a) all-alpha (b) all-beta (c) alpha+beta (d) alpha/beta. These classes are based on basic structural elements and can be characterized by distinct amino acid compositions [Ofran and Margalit, 2006; Rost and Sander, 1993; Wang and Yuan, 2000; Chou and Zhang, 1995]. The SCOP database is further ordered hierarchically into folds of similar structures, superfamilies that comprise functionally and structurally similar families together, which contain confidently related proteins. In this hierarchy the focus pivots from from structural similarity to evolutionary relatedness. The currently most updated SCOP2 database contains 5089 families and 1386 folds [UK MRC, accessed 2020-01-19].

Table 1.2: Statistics of different protein classifications. The number of protein families is limited to less than 20,000 in each classification method. On the structural level of folds, no more than 2,500 classes have been reported.

database	families	folds	update	source
SCOP2	5134	1398	Jan 2020	[UK MRC, accessed 2020-01-19]
Pfam	17,929	(clans) 628	Sep 2018	[EMBL-EBI, accessed 2020-01-19]
ECOD	13,896	2345	Jan 2020	[Grishin lab, accessed 2020-01-19]
CATH	6631	-	Jan 2020	[CATH database, accessed 2020-01-19]
SMART	-	1302	Sep 2017	[Letunic and Bork, 2018]

Protein family database - Pfam

Pfam is a protein family database [Punta et al., 2012] that classifies proteins into groups of related proteins using a sequence-based method based on profile Hidden Markov Models. The profiles are generated from multiple sequence alignments of sequences from the UniProt Knowledgebase [The UniProt Consortium, 2016] and represent non-overlapping clusters of natural protein sequences. Sequences matching one profile (with a sensible score) are assigned to the corresponding Pfam entry. All sequences belonging to one entry are assumed to be related to each other. These are further grouped into clans if an evolutionary relatedness can be detected. Most such relationships could be established using the ECOD database, which is based on known domain structures. In total, release 32.0 contains 17,929 Pfam entries and 628 clans.

ECOD

The Evolutionary Classification Of protein Domains database (ECOD) [Cheng et al., 2014] classifies protein domains into a hierarchy according to a sequence-based strategy. Domains that cannot be classified by sequence are analyzed with a structure-based strategy to assign the remaining sequences. ECOD is frequently updated, adding newly published structures of the PDB database into its classification. It is the current gold standard in domain assignments and, at more than 13,000 families, provides a structural basis for most known domains. Currently, ECOD comprises 13,896 protein families and 2,345 fold classes.

1.2.4 Domain assignment

In order to assign domains to a given sequence, HHsearch [Remmert et al., 2012] was used to search for domains within the ECOD database that match with the query. HHsearch assigns Hidden Markov Model-profiles (HMM), which reflect the sequence variability among sequences of the same ECOD domain to substrings of a query sequence.

Significant matches are considered to correspond to domains, which were then assigned to the query sequence according to the boundaries of the match. Of all matches for an ECOD domain the best-scoring (highest probability), non-overlapping hits are assigned as predicted domains to the sequence. For this domain assignment, a conservative threshold of minimum 90% probability to be correct according to HHsearch is chosen. Thus, some parts of the sequence, which are actually real domains, may not be assigned to a domain. However, the assigned domains are confident matches and only few false positive hits are assumed from this procedure. This assignment is specifically used in Subsection 2.3.3 and Section 4.3.

1.2.5 The power law

The power law captures a specific relationship between two variables that can describe many phenomena. Therein, the variable $f(x)$ is proportional to the reciprocal of x to the power of k:

$$f(x) = a \cdot x^{-k} \tag{1.8}$$

A famous example of the power law describes the relationship between the frequency of specific words $P(n)$ and its respective rank n among all word frequencies, which is referred to as [Zipf, 2013]:

$$P(n) = a \cdot n^{-1} \tag{1.9}$$

Zipf's law is closely related to the word count frequencies, which has been performed on protein sequence data in Section B.2. Therein, the abundance of words occurring f-times is proportional to the reciprocal of f to the power of k:

$$A(f) = a \cdot f^{-k} \tag{1.10}$$

Scale invariance

The power law is scale invariant, implying that a relative change of one variable results into an accordingly reciprocal change (to the power of k) of the other:

$$\begin{aligned} x_2 &:= x_1 \cdot c \Rightarrow \\ f(x_2) = f(x_1 \cdot c) &= a \cdot (x_1 \cdot c)^{-k} = f(x_1) \cdot c^{-k} \end{aligned} \tag{1.11}$$

Interpretation of exponent

In many cases, the exponent of the power law can be interpreted as basal property of the mechanism that generated the relationship between $f(x)$ and x. For cases where $k = 1$, the relative change of one variable x results into the inverse change of the other $f(x)$. Thus, increasing one parameter by 2-fold leads to a 2-fold smaller value of the other parameter. In this case, the parameters are at balance.

For cases where $k > 1$ the second variable decreases faster than the first increases. This leads to a steeper functional relationship. Concerning the word count frequency, a $k > 1$ will lead to many more words with frequency ≈ 1 and less words that occur very often. For cases where $k < 1$ the opposite effect occurs, where the second variable decreases slower than the first increases, causing the functional relationship to be less steep. In the word count distribution, this tendency leads to more words that occur more frequently and less words that only occur once. It is thus representative of a redundancy of the same words.
In Subsection 3.2.3 the observed $k < 1$ is interpreted as an indicator of redundancy in the given sequence data. Therein, the node degree distribution is studied within a large dataset of multiple genomes and other data sets comprising only one genome.

1.3 Research focused around sequence space

1.3.1 Evolutionary perspective on sequence space

The field of evolutionary biology is focused on the origin of life and on the paths that evolution has taken. A central question is thus the relatedness between organisms, proteins, and genes. Categorizing these into clusters of related entities through phylogenies and hierarchies is a major challenge in the field.
Decades of bioinformatic work has succeeded to map out an increasingly comprehensive description of sequences assigned to protein families, based on the detection of ever more remote homology [Cheng et al., 2014; Punta et al., 2012; Tatusov et al., 2000]. In Subsection 1.2.3 details about common approaches that classify proteins and domains into families are provided. The number of families that comprise confidently homologous sequences does currently not exceed 20,000 in any of these approaches. At a less conservative level of presumably non-related origin and structural similarity less than 2,500 folds are known. Clustering sequences by their similarity allows researchers to group sequences belonging to the same folds [Alva et al., 2009].
The majority of sequences can be assigned to a fold class by sequence similarity or by additional structure information. It seems that the clusters that define natural fold space are the major hubs around which natural sequences are scattered. From this perspective it is apparent that these hubs have a substantial role, not only in shaping the local, but also the global structure of sequence space occupation, corresponding to the image of islands formed by natural sequences within the global sea of possibilities [Lupas and Koretke, 2008].

Proteins from fragments

These central hubs of domain folds could be the starting point of the emergence of the protein world [Lupas et al., 2001]. A common hypothesis of early evolution is the exis-

tence of an RNA-world prior to the present-day mostly protein-mediated world [Alberts et al., 2002]. RNA at this stage is assumed to have had the ability to replicate and perform basic catalytic reactions by itself. With the emergence of small peptides, which served initially as co-factors for the RNA-functionality, proteins started to evolve. At this stage, evolution could possibly explore peptide space up to the length of supersecondary structures in an exhaustive manner. That is because evolutionary constraints on these peptides were supposedly weak, as the main functionality was performed by RNA. Of the emerged peptides, those gained acceptance that could fulfill essential functions such as RNA-binding or metal binding. Indeed, in the study of [Alva et al., 2015] primordial peptides have been revealed of which a large fraction possesses high RNA-affinity. Through accretion, the selected fragments may have become longer and evolved to domains and proteins. It is standing to reason, that sequence space of longer peptides has been explored starting from these ancient peptides. The idea of how sequence space might be occupied according to this theory is illustrated in Figure 1.9.

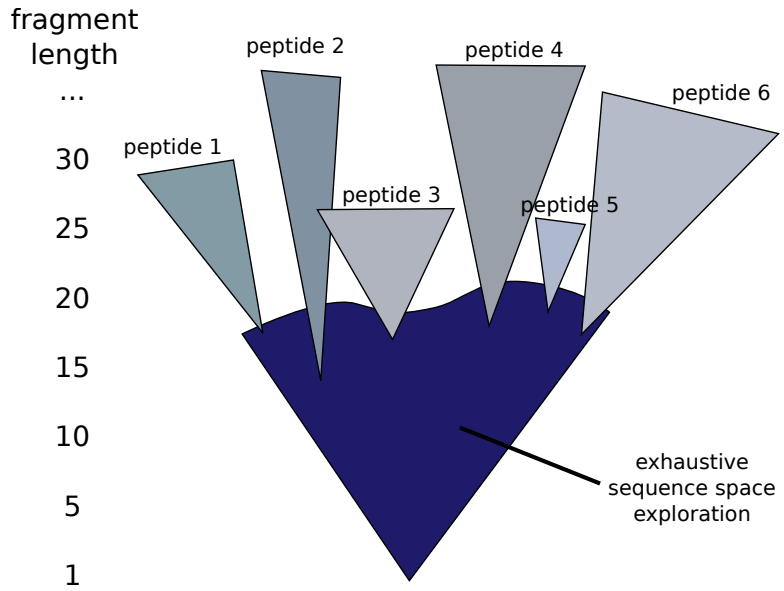

Figure 1.9: Sequence space occupation according to proteins from peptides. At the time of the RNA-peptide world, evolutionary constraints of peptides were only weak. They could evolve freely and exhaust the sequence space of supersecondary structure of a length about 15-20 residues. At this length, they may have possessed functionality that has led to their fixation and the evolution of proteins.

Exhaustive enumeration approaches

For a global description of natural sequences, ultimately, the abundance of all sequences in the overall space can be assayed. Relative to an expected value of abundance, derived from a random model, it is possible to determine whether natural sequences are over- or under-represented, in order to estimate their relevance to natural proteins. Due to the sparsity of sequence data relative to the size of sequences space (see Section 1.1.4), exhaustive enumeration approaches have focused on peptides of length five or smaller. In [Poznański et al., 2018] the authors demonstrate that within a redundant data set, the 5mer frequency of the majority of fragments deviates from the expected abundance of random sequences. They found an enrichment of peptides that comprised rare amino acids and related this finding to a global feature of functional and thus conserved sequences. The *k*mer frequencies in bacterial genomes could also be associated with phylogenetic relationships [Osmanbeyoglu and Ganapathiraju, 2011] and pathogenicity [Grzymski and Marsh, 2014]. They can thus be interpreted as a fingerprint for general features of the genome.

Local expansion

In order to extend from single sequences as used in the exhaustive enumeration approaches towards a better understanding of the relationships between sequences, the local space around existing sequences has been studied to reveal possible alternative sequences. In [Starr et al., 2017] the effects of epistasis to the chosen evolutionary trajectory through sequence space are analyzed, illustrating how the chosen trajectories determine results of evolution. Anything that has emerged and became essential to an organism experiences constraints that lock in this contingent outcome of many successive chance events.
Multiple functional variants of natural sequences could be revealed in [Harms and Thornton, 2014] and [Urlinger et al., 2000] by exhaustive mutation of functional sites. Some of these alternatives demonstrated enhanced activity or were entrenched in areas of sequence space that are not reachable from the original sequence by consecutive viable intermediates. These approaches of local expansion of the existing sequences show, that even on the local level, sequences experience many constraints that influence their evolvability and also, that local alternatives exist, that have probably not been explored by evolution.

Cluster analysis

A way to investigate longer sequences is taken through cluster analysis. Clustering natural sequences that share a significant similarity and can thus be transformed into each other by few mutations, can help to understand how evolution traverses through sequence space and how natural sequences are positioned relative to each other. Several studies

have published results of clustering approaches that have shed light onto different features of the sequence space that has been explored by nature.
Among other studies, in [Buchholz et al., 2018] clusters of related sequences were shown to comprise few hub sequences that are inter-connected to many other sequences in the same cluster. These hub sequences are more robust to mutations than those that are only loosely connected, given than many mutations lead to similar and valid sequences. This observation is revealed by analyzing the node degree between related sequences. For networks of related sequences, this node degree is power law-distributed, resulting in a scale-free network of natural sequences (see Subsection 1.2.5). Many natural phenomena follow this scale-free behavior [Koonin et al., 2002], which is different from that of random networks, where the node degree is Poisson-distributed [Erdős and Rényi, 1960].
Similar results were found in [Dokholyan et al., 2002], where the authors studied the structure space of natural protein domains. They showed it to be organized like a scale-free network, where some structures act like hubs that occur frequently in sequence space, while others occur only once. This capacity of distinct sequences for the same structure, referred to as *sequence capacity*, has been demonstrated to correlate with evolutionary age [Tian and Best, 2017]. Those structures that have been explored by a variety of sequences may be more easy to reach through a random walk, which also infers increased evolvability, or they were longer explored by evolution, leading to more diversed sequences. A decoupling of these two aspects is assayed over the entire structure space, free from effects of the natural exploration of structure space, and is presented below in Subsection 1.3.3.
In all these studies each cluster comprises sequences with a certain commonality, either common descent or structural similarity. However, these commonalities can be spread across the sequence space and such sequences that are positioned in distant locations share a randomly expected similarity. This effect is presented in [Rost, 1997] for common descent and in [Tian and Best, 2017] for structure space. Thus, although they belong evolutionary or structurally to the same cluster, sequences may be far apart in sequence space. The main cause for this is the high-dimensionaltiy of sequence space, a phenomenon that is further discussed in Subsection 5.2.1.

1.3.2 Convergence due to biophysical constraints

As presented above, the way evolution has traversed sequence space has a great impact on which sequences we observe today. This evolutionary perspective is in contrast to a biochemical perspective, that focuses on the constraints defined by folding, structure and biophysical interactions [Harms and Thornton, 2013]. Such constraints can be common among proteins and also have an impact on the selection of natural protein sequences. They lead to *sequence-structure relationships* that are laid out in the following.

Protein stability

The importance of protein stability as structural constraint is suggested in many studies to be under natural selection [Bastolla et al., 2017; Bershtein et al., 2017; Dasmeh et al., 2013]. Stability is therein described to be one of the strongest constraints causing intermolecular epistasis and consequently having an influence at the sequence level. Protein stability depends on many factors that all can lead to a biophysical imprint on natural sequences.

Hydrophobicity patterns

Stability and folding of a protein are largely defined by hydrophobic interactions [Dill, 1985]. Patterns of hydrophobic amino acids are therefore indicative of specific interactions. On the genomic scale, an asymmetry in the distribution of hydrophobicity across proteins is documented in [Lobry and Gautier, 1994]. This finding is therein associated to proteins with transmembrane regions, which are generally more hydrophobic than soluble proteins.
The technique of hydrophobic cluster analysis [Gaboriaud et al., 1987] of individual protein sequences is a common approach to compare natural and random protein sequences. It is based on the detection of hydrophobic clusters by patterns occurring from hydrophobic surfaces of alpha helices. In [Bitard-Feildel et al., 2015] it is used to identify orphan genes and to estimate their evolutionary age.
Furthermore, a general asymmetry in hydrophobicity, the ability to form hydrogen bonds and polarity are illustrated in [Pande et al., 1994]. The authors find that along a protein's sequence, these characteristics follow an uneven distribution. They relate this finding to a biophysical driven selection.
In contrast to these results, it is stated in [White and Jacobs, 1990] that the statistical distribution of hydrophobic residues within proteins follows a random pattern. The authors conclude, that natural sequences may have evolved from random sequences and also that sequence is only weakly constrained in natural proteins.

Secondary structure

In an analysis of global 5mer-space, unrelated sequences have been reported to have a frequency that is only slightly deviating from the expected frequency of random sequences, implying a global almost random structure of the 5mer-space [Lavelle and Pearson, 2009]. The most outstanding signal relating to the deviation from random behavior in this study was a statistically detectable bias in the dipeptide frequency. Deviations were associated to secondary structures formation of alpha helices. Secondary structure formation can, therefore, constrain the global occupation of sequence space.
According to [Yu et al., 2016], the lengths of predicted secondary structure elements are the same for natural protein sequences and randomized sequences of the same compositions with a minimal increase of helical content in natural sequences. Furthermore,

the authors notice a bias towards more intrinsic disorder in natural sequences. A similar result about the secondary structure content of random sequences (uniform amino acid composition) was obtained in [Minervini et al., 2009]. Therein, the structure of random sequences was predicted by Rosetta [Rohl et al., 2004]. In [de Lucrezia et al., 2012] secondary structure was likewise found to be a characteristic of completely random sequences. The authors state, however, that overall, structural characteristics in natural sequences are deviating significantly from random sequences. Moreover, they claim to be able to classify natural from random sequences.

1.3.3 Theoretical perspective

In [Bornberg-Bauer, 1997] the folding of all sequences of an alphabet comprising two types of amino acids (hydrophobic, hydrophilic) was computed and the sequence capacity of all valid structures was analyzed. By modeling the folding in a HP-model [Dill, 1985] all possible sequences were analyzed, thereby exhausting sequence space. The results imply that the abundance of sequences that fit to a specific structure is distributed according to Zipf's law (see Subsection 1.2.5) and that sequence space is occupied nearly randomly with no obvious higher-order clustering.
From a more theoretical perspective, natural protein sequences have been analyzed with methods from the field of information theory. The complexity of natural protein sequences is found to be 99% of that expected by random sequences with the same composition, as demonstrated in [Weiss et al., 2000]. Therein, the authors analyzed the entropy of word frequencies. Even when reducing the alphabet to biophysically meaningful amino acid classes the high information content of natural sequences could not be reduced. Similar results are achieved in [Strait and Dewey, 1996] where the authors performed a study of the conditional entropy among subsequent amino acids. These results are also in line with early assumptions on the randomness of protein sequences [Ptitsyn, 1985], where natural sequences are referred to as only 'slightly edited' random sequences.
Based on these findings of the mostly random nature of protein sequences, the ability to discriminate naturally folding protein sequences from random sequences is stated to not be possible in [Shakhnovich and Gutin, 1993].

1.3.4 Sources of opposing views

Natural protein sequences have been studied extensively from many different perspectives. Some attempts present excellent research with outstanding and convincing results. However, the discrepancies among these results demonstrate that the question if, by how much and why natural sequences stand out from randomness is largely unresolved. Here, I review several aspects that have led to opposing results.

The choice of natural data

Several discrepancies are associated to different settings of the studies. One difference between studies is that they use different natural sequence data sets. Using data of one genome, of genomes from different phyla, of sequences from one superfamily, of structured or disordered sequences, will in often lead to different results. That is because all of these data sets comprise distinct natural biases, resulting into different sequence features to be more or less pronounced.
An example of different data sets leading to different conclusions are two studies of 5mer frequencies of [Poznański et al., 2018] and [Lavelle and Pearson, 2009]. In [Poznański et al., 2018] the 5mer space is shown to be differently populated compared to random sequences, while in [Lavelle and Pearson, 2009] only minor deviations were reported. One of the reasons for these different conclusions may be related to the usage of different data sets. In the first study, a data set of related sequences, comprising overlapping conserved sites was used, in the second study a non-redundant data set of unrelated sequences was used. Using a data set without related sequences leads to a globally random view.

The choice of random sequence model

The random sequence model is crucial for the statement by how much and for what reasons natural sequences differ from randomness. In the presented studies the term 'random sequences' was used to refer to different kinds of random sequences, ranging from sequences with a uniform amino acid composition (E-model), sequences that are biased by the overall genome composition (G-model), sequences biased by the composition of natural proteins (P-model) and even those biased by the composition of natural sub-domain sized fragments (L-, D-model).
Using uniformly random sequences as a model, natural sequences could be discriminated from random sequences in [de Lucrezia et al., 2012]. As the amino acid compositions of these sequences are very distinct from that of natural protein sequences, this approach is likely to detect compositional differences instead of differences in sequence. The statement in [Shakhnovich and Gutin, 1993] that natural sequences cannot be discriminated from random sequences is based on a different random sequence model that incorporates the natural amino acid composition.
In [Lobry and Gautier, 1994] the amino acid composition of the *Escherichia coli* genome was incorporated into the model of random sequences (G-model) and the hydrophobicity distribution across all proteins within the genome was studied, finding it to deviate from the randomly expected distribution. In contrast, in [White and Jacobs, 1990] the composition of proteins (P-model) was used to normalize for fluctuations between different proteins. Therein, hydrophobicity was found to be randomly distributed within single proteins.
A similar model incorporating the compositional fluctuation between proteins (P-model) was used in [Pande et al., 1994], where the authors found the opposite, that hydropho-

bicity is not randomly distributed in a group of many proteins. This discrepancy of the perception if the inner-protein fluctuation is random as described in [White and Jacobs, 1990] or not may be related to the different data sizes of natural sequence data (single sequences or group of sequences) being compared.

The finite sampling problem

In [Pande et al., 1994] the authors summarize the local fluctuations of hydrophobicity over many protein sequences. This averaged observation over many proteins is not enough to discriminate single sequences to be natural or random but is indicative of a certain trend.
In contrast, the fluctuation of hydrophobicity in very short protein sequences may not be pronounced, relative to the random model, as shown in [White and Jacobs, 1990]. The judgment about random or nonrandom behavior thus depends on the used data sizes in a comparison, which are related to the finite sampling problem (see Section 1.1.4). For longer sequences, the stochastic error due to sampling is smaller, allowing to detect minor differences. Averaging over many sequences, accumulates minor deviations that on the level of single sequences are not detectable.
Due to this problem of data insufficiency, many studies proceed by averaging effects over many sequences, aiming to detect general tendencies in natural protein sequences, rather than to classify single sequences to behave naturally or randomly.

The global-local problem

Another recurring aspect that leads to different perspectives concerning the randomness of natural protein sequences is associated with the local or global representation of them. From a local perspective, the accumulation of natural sequences around conserved sequence motifs is a nonrandom phenomenon related to heredity, conservation and partially convergence. This local increase of sequence space occupation around existing sequences is often imagined as 'islands' in sequences space. A variety of studies referring to this concept of islands are presented in Subsection 5.2.2, where I discuss this view in detail in the light of the insights gained in this thesis.
This local view neglects relationships between all other sequences, which put local sequence clusters into the global context of sequence space. Theoretical approaches in [Strait and Dewey, 1996] and [Weiss et al., 2000] that focus on the overall complexity of natural sequences have shown, that the occurrence of similar, related sequences does not impact the global occupation of sequence space. Performing these studies on subsets of natural sequences, i.e., for those of one superfamily, which is a local subset of natural sequences, may lead to different results.

1.4 Positioning of this thesis

This thesis focuses on characterizing the sequence space occupied by natural proteins with the aim to determine general features of functional sequences and also to associate these features to their biological origin. As laid out in Section 1.3, many settings and possible perspectives have led to different results concerning the characteristics of natural protein sequences and their differences relative to random ones. Here, many considerations are taken together to relate the impact of divergent evolution, biophysical constraints and phenomena on the DNA-level, thereby understanding protein sequences as a result of many pressures with different effects.

1.4.1 Outline of this thesis

The second chapter is dedicated to the global occupation of protein sequence space by nature. Using a distance-based method, it was possible to overcome sequence space sparsity and to study sequences longer than five residues, which was not possible in previous enumeration approaches. In line with other studies, this method revealed that the occupation of sequence space by natural proteins is mostly random. This observation is here discussed in the light of evolution. The used method allowed me to weigh different biological biases as reflected in each random sequence model by contrasting their accuracy to reproduce distances between natural sequences. Minor deviations between random sequence models and natural sequences on the global scale could mostly be associated with specific amino acid compositions, foremost the overall composition and that of individual protein sequences. Having accounted for the amino acid composition of natural proteins, the sequence bias free from any compositional bias could be extracted, a footprint of biochemically preferred amino acid patterns. For this, sequence similarities arriving from divergent evolution were dissociated from those from convergent evolution.

In the third chapter the focus changes from a global to a local assessment of sequence space. This space is often characterized by an enhanced occupation due to the divergence of related sequences. Previous studies have focused on the sequence similarities where homology cannot directly be inferred. This range has been coined the *twilight zone* and has previously been derived by validating common descent with structural similarity. Here, the twilight zone is revisited and a different derivation is presented by using statistical significance to derive the boundaries of the twilight zone.

In the forth chapter, compositional fluctuations within whole protein sequences are assessed. Domain recombination and fold-specific constraints lead to proteins being composed of structurally and functionally distinct parts, resulting in the assumption that this heterogeneity should also be reflected in the different use of amino acids along a protein chain. The finding of a supposedly homogeneous amino acid composition of bacterial proteins, as presented in the first chapter, was therefore counter-intuitive. A comparative

analysis revealed that similarity between amino acid compositions of domains within the same protein is detached from structure-specific recombination, recombination bias in proximate genomic regions, and presumably also of protein topology. The observed homogeneity of proteins is here shown to be correlated to the usage of identical codons along the protein chain, a phenomenon that has previously been associated to the expression level, translation efficiency, tRNA abundance, and other DNA-related constraints. With the presented comparison a more detailed insight into the dependencies between DNA and protein evolution has been achieved.

In the last chapter, the discussion, I review key aspects that concern comprehensive studies of global features in protein sequences, given the high-dimensionality of sequence space. Therein, the common metaphor of sequence space being occupied by dense clusters that are imagined as islands in a vast sea of all possibilities is discussed.

Chapter 2

Protein sequences on a global scale

2.1 Motivation

2.1.1 Global features of natural protein sequences

Divergent evolution

Natural proteins form the backbone of the complicated biochemical network that has given rise to the great variety of life on Earth. This highly interwoven framework of reactions seems impossible to have arisen by chance, simply because the great majority of random protein sequences fails to form a specific structure, let alone possess chemical activity. Finding general features of natural sequences, that determine their success can help to better understand proteins, both in order to understand protein evolution [Shah et al., 2015; Luigi Luisi, 2003] and to guide the design of new proteins [Woolfson et al., 2015; Pande et al., 1994]. More specifically, we were interested in the question if natural sequences can globally be characterized by constraints of divergent evolution.

Convergence and randomness in the global scale

A partial answer to this question is presented in [Poznański et al., 2018] where the authors found that over-represented pentamers tend to be characterized by an accumulation of rare amino acids such as cysteine, methionine, tryprothan and histidine in a redundant set of sequences. They relate this finding to evolutionary constraints of functional sites (which often comprise rare amino acids for their activity), resulting in more, conserved pentamers containing rare amino acids. This finding occurs to be less pronounced when investigating a non-redundant set of protein sequences, as has been shown in [Lavelle and Pearson, 2009]. Therein, the authors associate the most pronounced over-representation of pentamers with short-range sequence correlations of mainly alpha-helical structures, related to biophysical pressures and not to divergent evolution.
Compared to evolutionary constraints, biophysical constraints follow more universal patterns. Given that proteins live in an environment defined by biophysical laws, they must obey these constraints at all times. The need to form a hydrophobic core for globular proteins, the hydrophobic surfaces between coiled coils or protein-protein interactions,

hydrophilic surfaces for the interaction with water, secondary structure formation and the ability to fold rapidly, are some examples of biophysical principles that proteins are facing. Literature focusing on the effects of these constraints is based on sequence-structure relationships, aiming to detect signatures of structure that are imposed onto sequence (see Subsection 1.3.2). These kind of relationships are caused by convergent evolution. However, sequences that are not related to each other and that are not exposed to a great amount of similar biophysical constraints share a rather random similarity. Aligning the sequence of a transmembrane receptor to that of a hemoglobin will give little insights into their common ability to fold, possess a defined structure and function. These kind of random relationships dominate the all-to-all comparisons among natural protein sequences, resulting into natural sequences to look globally mostly random [Weiss et al., 2000; Strait and Dewey, 1996]. In the field of theoretical biology, it has been demonstrated, that natural sequences are indeed very similar to random sequences from a global point of view (see Subsection 1.3.3).

2.1.2 Content of this chapter

In this study, the focus lies on analyzing the global occupation of sequence space by natural proteins, to extract existing deviations from a randomly expected occupation and to determine reasons for the observed deviations. Using a distance-based approach by interpreting sequence similarity as distance in sequence space, it was possible to extract and to characterize existing deviations between the natural and random occupation of the global sequence space for sequence fragments of domain size.
In Section 2.3, the results are presented when comparing the distribution of distances, as defined by their similarity in an alignment. These were derived from a diverse bacterial data set, as representative of natural protein sequences, and compared to those derived from a model, which accounts for the overall amino acid composition (A-model). Further compositional differences are accounted for by including the natural amino acid compositions of genomes (G-model), proteins (P-model) and subdomain-sized fragments (L-model) into the random model. With the consideration of more local amino acid composition, the natural distance distribution can be better approximated than by the more general A-model. The most local consideration of composition has presumably accounted for all influences that impact the amino acid composition of protein sequence. The remaining deviations between the distance distribution of the natural data set and that of the L-model are thus related to sequence relationships that are further interpreted in the light of common descent and convergence. In Section 2.4, a decomposition of sequence relationships into confidently homologous, analogous and unknown is presented. This procedure was able to distinguish between sequence effects arriving from these different mechanisms. With this, it was possible to demonstrate that similarities caused by divergent evolution are not effecting the natural distance distribution. The remaining discrepancies could instead be associated to non-related structures, implying that global sequence similarities are determined by convergence.

Table 2.1: Contribution to performed research presented in Chapter 2. Title of Paper: "Where Natural Protein Sequences Stand out From Randomness", Status in publication process: published in bioRxiv.

Author	Author position	Scientific ideas%	Data generation %	Analysis and interpretation %	Paper writing %
Laura Weidmann	first	50%	90%	40%	45%
Tjeerd Dijkstra		15%	10%	30%	10%
Oliver Kohlbacher		10%	-	10%	-
Andrei N. Lupas	last	25%	-	20%	45%

Statement of contributions

In this chapter, most of the intellectual and analytical work was performed in continuous exchange with Andrei Lupas, Tjeerd Dijkstra and Oliver Kohlbacher. The joint efforts have led to a manuscript that has been uploaded to the preprint server bioRxiv [Weidmann et al., 2019]. In Table 2.1, the authors are listed according to their contributions and the respective area of the performed research. The presented work in this chapter, figures and text is largely overlapping with this manuscript. In the context of this joint work, the 1st person plural is used rhetorically as the active person.

2.2 Materials and methods

2.2.1 Bacterial diversity as natural data set

For an adequate data set that reflects the natural protein sequence space, we aimed to achieve a reasonable coverage of deep phylogenetic branches with complete and well-annotated proteomes. Given that the genome coverage for the archaeal and eukaryotic lineages is still sparser than for bacteria and that particularly eukaryotic genomes are affected by issues of assembly, gene detection, and intron-exon boundaries, we built our database from the derived bacterial proteomes collected in UniProt [Apweiler, 2009]. To control for redundancy, we selected only one genome per genus and filtered each for identical open reading frames and low-complexity regions. In total, our data set comprises 1,307 genomes, $4.7 \cdot 10^6$ proteins, and $1.2 \cdot 10^9$ residues. We simplified complexities arising from the use of modified versions of the 20 proteinogenic amino acids, which occurred in a few hundred cases, by converting these to their unmodified precursors, thus maintaining an alphabet of 20 characters throughout. The amino acid composition of this data set is provided in Table 2.2.

Table 2.2: Amino acid composition of the bacterial data set.

amino acid	letter	count	frequency [%]
alanine	A	108055041	9,28
cysteine	C	11236906	0,96
aspartate	D	67465102	5,79
glutamate	E	74062415	6,36
phenylalanine	F	46949016	4,03
glycine	G	88841861	7,63
histidine	H	25272724	2,17
isoleucine	I	69564413	5,97
lysine	K	54872067	4,71
leucine	L	114820546	9,86
methionine	M	24373940	2,09
asparagine	N	44133479	3,79
proline	P	54073615	4,64
glutamine	Q	44030893	3,78
arginine	R	68559528	5,89
serine	S	68209111	5,86
threonine	T	64140047	5,51
valine	V	83563375	7,18
tryptophan	W	15279906	1,31
tyrosine	Y	36293197	3,11
total		1163797182	100

Genome curation

Apart from redundancy at the genome level, we control for recent gene duplication events. For each genome, we cluster its proteins using cd-hit [Li et al., 2001] (version 4.6 with 99% sequence identity and 90% coverage). A representative protein sequence, as defined by cd-hit, was selected for each cluster; all other proteins were discarded.

Low complexity filtering

Low-complexity regions (LCRs) are a well-known features of natural sequences that do not occur as frequently in random sequences. We first analyzed our data including LCRs and found that they majorly contribute to the differences between natural sequences and our models (data not shown). Therefore, we pruned LCRs of our data set using seg-masker [Wootton and Federhen, 1996] (version 2.3.0+ with the standard settings), to obtain differences between natural and random sequences that are not due to this well-

known feature. This pruning of LCRs leads to sequences of slightly higher complexity than expected for short peptides (data not shown). The introduced bias due to this pruning plays an insignificant role, especially for longer sequences, which are of most interest in our study. Since, N-terminal methionines were sometimes included, we stripped them to standardize our sequences.

Sequence adjustments

To simplify our analysis, we changed a couple of hundred cases of letters referring to uncommon or ambiguous amino acids to their most similar proteinogenic amino acid. Additionally, we removed the invalid amino acid X by replacing it with an end-of-line-character, effectively dividing a protein sequence into multiple parts. In order to use the exact same data set for all sequence lengths, we pruned our data set of sequences shorter than 100. This approach reduced edge effects to a certain amount.

Table 2.3: Residues referring to ambiguous amino acids or rare modified versions were assumed as invalid letters and were replaced with a suitable alternative.

amino acid	letter	replaced with	count
aspartate or arginine	B	D	1
glutamate or glutamine	Z	E	1
pyrrolysine	O	K	6
selenocysteine	U	C	445
unidentified	X	end-of-line-character	102,840,390

2.2.2 Representing global sequence space occupation by distances

The main difference of our study to the most common approaches to global sequence space is the aim to step away from specific locations in sequence space. Instead, we interpret the occupation of sequence space by distances between observed sequences. This allows to overcome the sparsity issue, which hinders location-specific studies to analyze sequences of a length above 5 residues (see Section 1.1.4). Studying the layout of space through pairwise distances is common in other fields, such as protein structure determination [Wüthrich, 1986], spatial statistics [Diggle, 2014] and economics [Duranton and Overman, 2005]. In several studies of protein sequences, similar statistics are being used [Rost, 1997; Buchholz et al., 2018], however, a connection to the global occupation of sequence space was not indicated.

Distance distributions

Our approach is built on the probability mass function of pairwise distances between sequences of the same length, in the following referred to as *distance distribution*. A distance distribution illustrates how often sequences are positioned at a certain distance to each other. We use it to study the way sequences are spread across the possible space and build distance distributions for the natural data set and also for each data set of random sequences derived from specific models. By using lengths of up to 100 residues, our sequences thus reach domain size [Wheelan et al., 2000].
As a metric for distance, we focus on the *normalized local alignment score* of a Smith-Waterman alignment [Smith et al., 1981], since this metric is commonly used to capture similarities between natural sequences [Rost, 1999; Schneider et al., 1997]. Additionally, we present the results of a global Needleman-Wunsch alignment [Needleman and Wunsch, 1970], a Shift metric without internal gaps and the Hamming distance. The distance metrics are presented in detail in Section 2.2.2.
The choice of distance metric is not of great relevance for the main implications of our study; relative to each other, the distance distributions of the random models deviate similarly from that of natural sequences irrespective of the chosen metric. In this context, it is important to note that our method differs from common approaches, as it only considers the pairwise similarity between two sequences and thus their actual distance in the sequence space. In contrast, many bioinformatic methods that compare sequences to each other scale distances according to their statistical significance and in many cases iterate comparisons in order to extract patterns of conserved residues, as indicators of homologous relationships. These approaches result in distances that reflect evolutionary relationships, visualized as islands of higher density in sequence cluster maps [Alva et al., 2009; Nepomnyachiy et al., 2014]. These distinct approaches to sequence space are discussed in Subsection 5.2.2.
Distance-based approaches do not preserve information about specific positions of data points in space, but rather characterize their global distribution, which includes *global clustering and dispersion*. A corollary of this is that distinct data sets become comparable through their distance distribution, even if they do not share any specific data points.

Residual and total residual

For the comparison of the natural to a random distance distribution, we first subtract the fraction of distances observed in the random data set from that observed in the natural data set for each alignment score. We refer to this difference as the *residual*. Over all alignment scores, residuals sum up to zero and may have values that are either positive (more natural distances) or negative (more random distances). In order to obtain an overall measure of how different two distance distributions are, we derive the *total residual*, which is the variational distance between two distance distributions. More precisely, the total residual is the sum over the absolute residuals, normalized to a range between 0%

and 100%.
If the two distance distributions are completely non-overlapping, the total residual assumes the maximal value of 100%, indicating that no distance between natural fragments can be modeled with the underlying random sequences. If they are identical, the total residual assumes a value of 0%, indicating that for pairs of natural sequences, there is a corresponding pair of random sequences with the same distance. Thereby, the total residual represents the fraction of natural distances that are not accounted for by the distance distribution of a random model.

Metrics

For the calculation of distances between protein sequences, we use four types of distance metrics. In order to capture local similarities between sequences, we use the Smith-Waterman alignment and a Shift alignment, where the first allows for internal gaps while the second does not. For global similarities, we use the Needleman-Wunsch algorithm and the Hamming distance, which also differ in the permission of internal gaps.
Gap penalties were used according to the model of affine gap cost (see Section 1.2.2) with a score of -3 for gap opening and -0.1 for gap extension according to the values used in [Schneider et al., 1997]. To consider only sequence identities, the identity matrix was used to score amino acid substitutions, resulting in a score of +1 for amino acid matches. All distances were derived using the C++ software package SeqAn [Reinert et al., 2017], which allowed a fast computation of many sequence alignments in parallel.

2.2.3 Separating homology from convergence

Sequence similarities can be found between sequences of common descent (homology), non-related sequences that experience similar structural of biophysical constraints (convergence) or chance. For the classification of sequence similarities into homologous or analogous origin, we use tools of the HHsuite [Söding, 2005] to detect homologous relationships. To detect analogous relationships, we assign domains of the ECOD database [Cheng et al., 2014] to these used sequences and check if they comprise only non-related domains, inferring an analogous relationship.

Sampling sequence relationships

The generation of Hidden Markov models (HMM) with HHblits [Söding, 2005] for individual sequences is time consuming, as it can take up to a couple of minutes. Given that the number of relationships between fragments is of quadratic size relative to the number of fragments, a small number of fragments already suffices to sample a large number of relationships. For this, we systematically sampled our bacterial data set in steps of 1 million consecutive amino acids, starting from different indices of the data set. At each

sampled position a fragment of 100 amino acids was extracted. If the fragment was overlapping with the beginning or end of a sequence, it was rejected. With this procedure, we extracted 10 independent sets of natural 100mers that are equally distributed over our data set, each containing approximately 650 fragments.
We sampled relationships between these sets of fragments by aligning all fragments in one set to all of those in another set and vice versa. In total there are 90 possible combinations of the fragment sets, of which we chose 10 as representative sets of pairwise relationships. Thereby, every set of fragment was combined to two other sets for the sam-

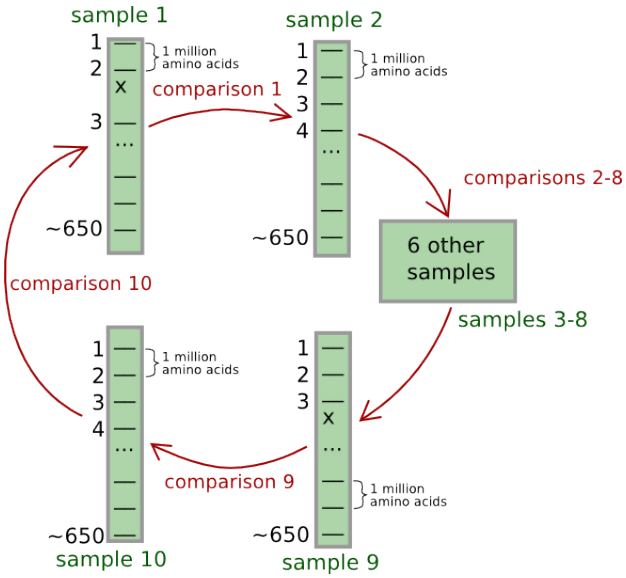

Figure 2.1: Sampling procedure of fragment pairs and relationships. 10 sets of 100mers were systematically sampled of the original data set, comprising about 650 fragments each. For the sampling of relationships, all fragments in one sample were compared with all fragments of the next sample, resulting into 2 million comparisons.

pling of relationships. This resulted in 2 million pairwise fragment comparisons divided into 10 disjunctive sets. Figure Figure 2.1 illustrates this sampling procedure.

Classification approach

Using HHblits, we generated HMMs with the standard settings for each of the sampled fragments with two iterations, using uniclust30 [Mirdita et al., 2016] as underlying database (version August 2018). For the sampled relationships, we pairwise aligned the generated HMMs with HHalign, in order to estimate whether two fragments are homologous or not. For this, we chose a conservative threshold of a probability above 90% to be homologous according to the HHalign prediction.
The remaining fragment pairs were analyzed for certain analogy. For this, we assigned

ECOD [Cheng et al., 2014] domains to the sequences of interest (see Subsection 1.2.4) and consider sequences that contain only unlike domain folds as unrelated. Domain fold classes are represented in ECOD by the X-group. The X-group is the highest level at which homology still needs to be considered as a possibility. Sequences that were assigned to domains of only distinct X-levels can thus be considered as confidently analogous.

2.3 Approximation by natural amino acid composition

2.3.1 Overall composition bias

We start our analysis by assessing to what extent the global amino acid composition, as captured in the A-model, can account for similarities between natural sequences. For this, we compare the distance distributions of the natural to that of the random data sets for fragment lengths up to 100 residues, in increments of 10. At all fragment lengths, the results are closely comparable. We show the results of 100mers as representative for domain-sized sequences in Figure 2.2 using a Smith-Waterman alignment as metric. The distance distributions of natural and A-model data overlap extensively. Both are unimodal with a peak at a low normalized alignment score of 11%. The minor difference between the natural and random distance distribution becomes apparent, when the y-axis is plotted in a logarithmic scale (Figure 2.2: B) or when their residuals are considered (Figure 2.3).
Using a logarithmic scale, a long tail of unexpected similarities becomes visible, which we associate to sequence similarities between homologs. This tail has only a minor weight compared to the majority of all similarities and we further discuss its contribution to the global sequence space in Section 2.4.
The residuals between the natural and A-model distance distribution take the shape of a wave, with two crests at alignment scores of 9% and 15% (reflecting an over-representation of the corresponding similarities in the natural data set), and a trough at 11% (reflecting an under-representation). The over-representation of both high and low similarity scores in the natural data set, suggests that natural sequences are not overall more similar to each other but also display a great amount of heterogeneity compared to their general composition, as captured by the A-model.
We rationalize this effect with the observation that natural sequences possess characteristic compositions for specific folds or for different location in the cell such as membrane, nucleus or cytoplasm. Another reason for this compositional diversity of proteins is related to codon bias, as presented in Subsection 4.3.5. These shifts in amino acid composition towards an accumulation of specific amino acids of similar biochemical features lead to a *lower compositional complexity* than generally observed (see Section A.2). Thus, this diversification in natural composition will lead to sequences with more diverse compositions. Similarities of sequences with more distinct compositions will thus

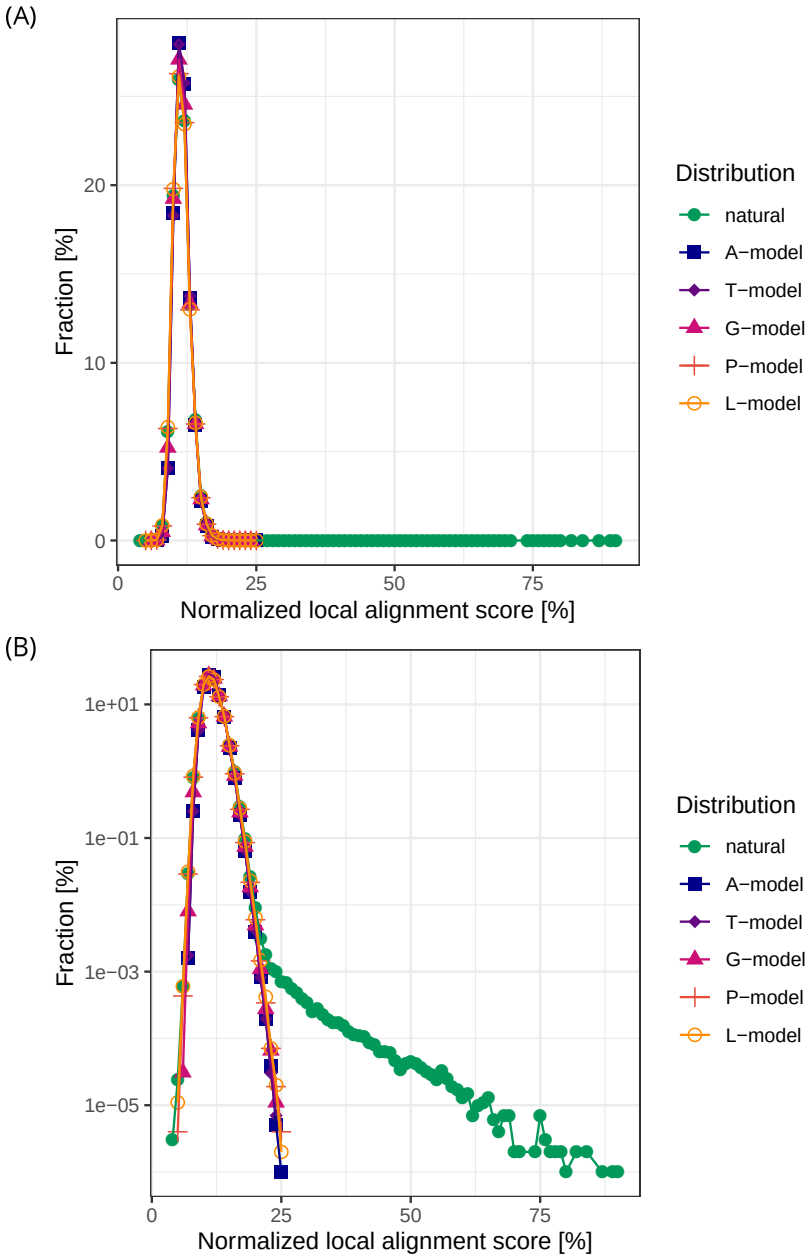

Figure 2.2: Distance distribution of natural data and random models. (A) Distance distributions indicate the frequency of specific normalized local alignment scores among natural 100mers and those among 100mers from different random sequence models. They are mostly overlapping, ranging around a low score of 11%. Values of zero were omitted. (B) When plotting the y-axis in logarithmic scale, a long tail of unexpectedly high scores becomes apparent. These similarities are presumably of homologous origin.

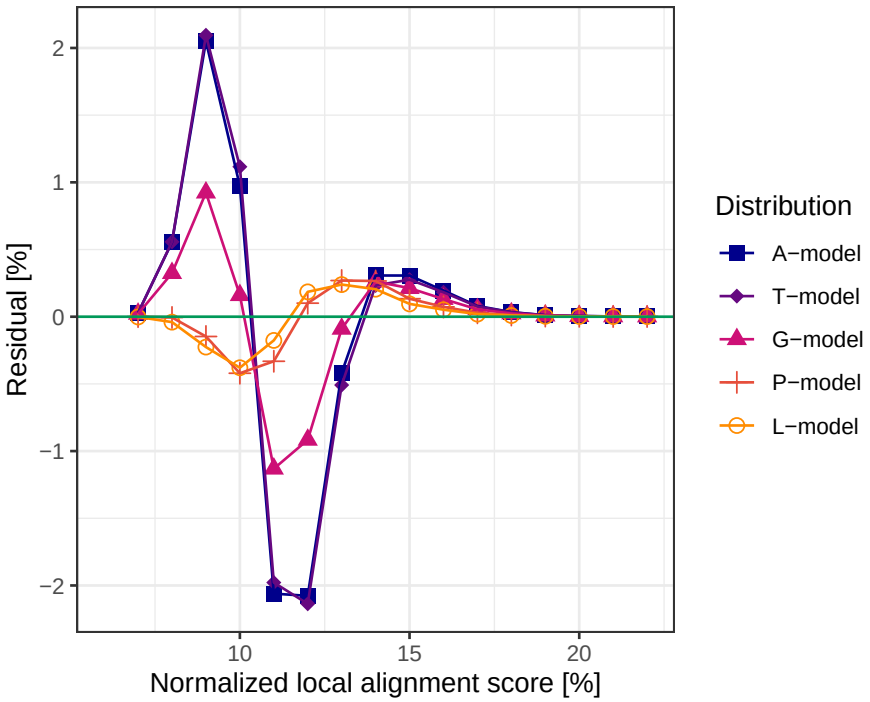

Figure 2.3: Residuals of natural data and random models. The differences between natural and random distance distribution is captured in the residuals. These are generated from 500 million distances between fragments of length 100 for each model as well as for the natural sequence data.

score poorly.
We note, however, that this discrepancy between natural sequences and the A-model is not very pronounced, as the total residual has a value of only 4.6% for 100mers (Figure 2.4: A). It is even less pronounced at smaller fragment lengths, reaching 0.4% for 10mers. We conclude that the A-model becomes less accurate in describing the distances between natural sequences at lengths that are biologically relevant, but that it already achieves considerably higher accuracy than the completely random model (E-model). The distance distribution of the E-model deviates significantly for that of the natural data (Figure 2.5: A), with a total residual of 30.4% (Figure 2.5: B) for 100mers.

Dipeptide bias

We evaluated whether adding sequence information to the unified compositional bias of the A-model could further improve its fit to the natural distance distribution. Since nature favors certain amino acid combinations as neighboring residues, a model that reflects the natural dipeptide frequency (T-model), has been proposed to represent natural sequences better than the A-model [Lavelle and Pearson, 2009].

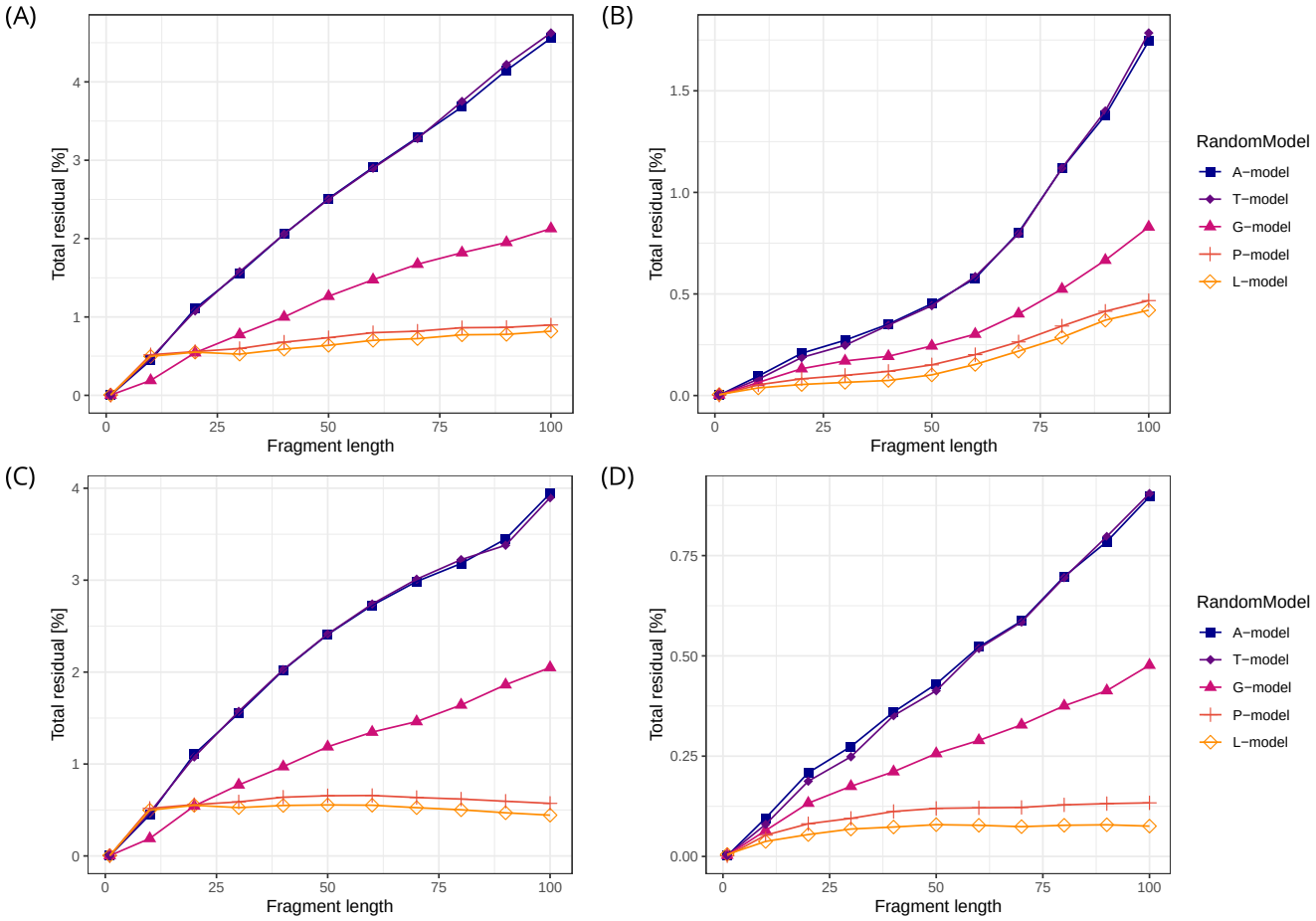

Figure 2.4: Total residuals as a function of fragment length. (A) Smith-Waterman alignment (B) Needleman-Wunsch alignment (C) Shift (D) Hamming distance.

We implemented the T-model by extracting the dipeptide frequencies from our natural data set and using them to generate random sequences with a Markov Chain Model. For all fragment lengths, we derived the distance distribution of the T-model, its residuals and the total residual.

By all these measures the T- and the A-model yielded essentially identical results in modeling the natural distance distribution. This outcome was somewhat surprising, as the addition of dipeptide frequencies to the A-model did produce a measurable improvement in the enumeration study of 5mers [Lavelle and Pearson, 2009]. This may be due to the different methodology in that study, which collated exact 5mer frequencies, corresponding to a position-wise Hamming distance of zero, and thus being close to a global, not to a local alignment as used in our study.

In fact, when using the Hamming distance as metric, the T-model achieves a slightly better accuracy over the A-model for sequences of 50 or less residues (Figure 2.4: D).

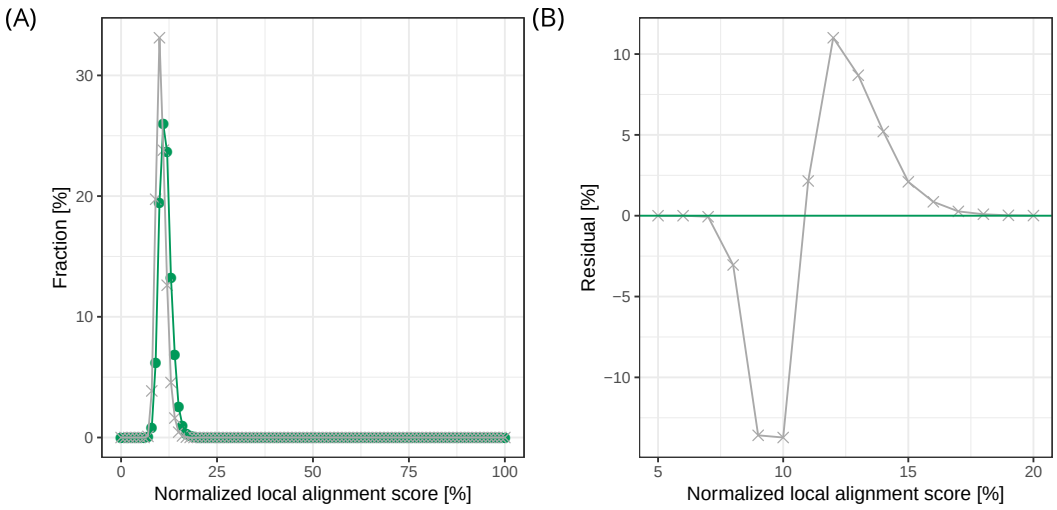

Figure 2.5: E-model compared to natural data set.

From the results obtained with the A- and T-models, we conclude that global measures of composition and sequence bias already approximate the similarities between natural sequences fairly accurately, but that this accuracy decreases with sequence length. Especially for longer fragments, we expect further improvement by including local compositional biases.

2.3.2 Context-specific composition

In order to capture context-dependent features, we investigated the effects of naturally occurring local amino acid compositions. As a first step, we considered a model that accounts for genome diversity (G-model). Therein, the random data set is produced by shuffling residues of the natural data set within the boundaries of each genome. Given that our natural data set comprises 1,307 genomes, the derived sequences are sampled from 1,307 distinct compositions. Further locality was achieved by accounting for the composition of individual proteins (P-model). Here, the random data set is produced by shuffling residues within each natural protein, corresponding to $4.7 \cdot 10^6$ compositions. Since proteins are generally composed of domains, which are usually autonomous in structure and also often in function, the next level of locality would be achieved by accounting for the compositional biases of individual domains. Producing such a model is however not straightforward, as roughly 30% (see Subsection 2.3.3) of all residues in our data set cannot be assigned to a domain structure. We conclude that in order to use a model based on the composition of domains, the data set should be selected accordingly with individual domains as sequence entries. A study based on this approach is presented in Section 4.3. In this specific case, we decided to consistently use the same data and instead use a local composition model, referred to as the L-model. It considers the amino

acid composition of domain-sized fragment comprising 100 residues. Correspondingly, we considered all natural sequences, whether or not they are part of a structured domain and thus included linker sequences and intrinsically unstructured regions.
Comparing the G-model to the A- and T-models over the bacterial data set shows a dampened wave for the residuals, with the same shape, but a decreased amplitude (Figure 2.3). The total residual is correspondingly smaller by a factor of about 2 for all fragment lengths (Figure 2.4: A), implying that controlling for genome composition provides a substantial improvement in modeling similarities among natural sequences. A further improvement is clearly achieved with the P-model, even though, at sequence lengths below 20 residues, it produces minor inconsistencies in its total residuals relative to the A-, T-, and G-models (Figure 2.4: A). We suspect that this is an artifact of using local alignments (Figure 2.4: A, C) and, indeed, the effect disappears when using a global alignment as distance metric over the same data set (Figure 2.4: B, D). As for the A-, T-, and G-models, the residuals of the P-model also have a wave shape, which is however qualitatively different from the shapes for the less specific random models, as it has only one crest at an alignment score of 13%. The crest for the unexplained long-range distances is gone, which we attribute to the fact that accounting for composition at the level of individual proteins has introduced the compositional heterogeneity of natural sequences into the random model. For 100mers the total residual of the P-model is 0.9%, a value that is not improved remarkably by an even greater locality: The residuals of the L-model have the same wave shape as those of the P-model and a comparable amplitude, providing a minor improvement with a total residual of 0.8%.
The resemblance between the P- and L-model was somewhat surprising, as it is well established that many proteins are composed of disparate parts such as domains of distinct fold classes, intrinsically unstructured regions or fibrous parts. These parts are known to be characterized by different residue compositions [Dubchak et al., 1995; Ofran and Margalit, 2006]. The composition of proteins that are composed of heterogeneous parts should thus be scrambled in the P-model and preserved in the L-model. We therefore expected that the L-model would provide a clearer improvement over the P-model.

2.3.3 Similar results of L- and P-models are associated to data set

We see two reasons why the total residuals of the L- and the P-models are almost identical. One is a technical reason, namely that there is no room for fluctuation of local residue composition in our bacterial data set, as it may comprise a large number of short and single-domain proteins. The other is a potential qualitative characteristic of our data set, namely that in long bacterial proteins the local residue composition does not fluctuate remarkably.
In order to distinguish how these two reasons contribute to the comparable total residuals of the L- and P-models, we added two eukaryotic data sets for comparison to the following analysis. We retrieved the highly curated proteomes of *Homo sapiens* and *Arabidopsis thaliana* from UniRef [Apweiler, 2009] and pruned them according to the

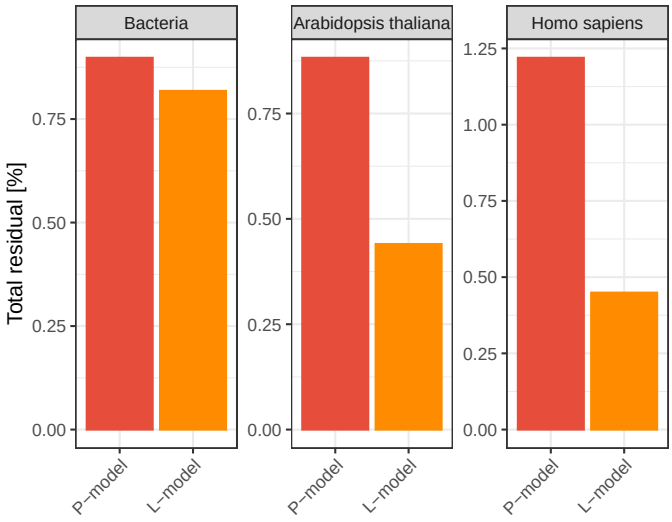

Figure 2.6: Total residuals for bacterial dataset, the proteome of *Arabidopsis thalia* and *Homo sapiens* of the P- and L-models. Relative to the total residual of the P-model, the total residuals of the L-models differ in the three presented data sets. In bacteria, the total residual of both models is almost identical, whereas for the eukaryotic data sets the total residual of the L-model decreases more than 2-fold relative to that of the P-model.

procedure used for our bacterial data set. Comparisons of total residuals between the bacterial and eukaryotic data sets show that, whereas the total residuals of the P- and L-models for the bacterial data set are essentially equivalent, the total residual of the L-model in the eukaryotic data sets is more than 2-fold smaller than those of the P-models (Figure 2.7: A), and thus closer to our expectation.

Correlations to sequence lengths, domain count and structure content

In order to evaluate the first, technical reason, we analyzed sequence lengths in all three data sets and estimated the number of single- and multi-domain proteins. The bacterial data set has the shortest proteins with a median length of 315 residues. The *Arabidopsis thaliana* data set has a median length of 400 residues and the *Homo sapiens* data set comprises the longest proteins with a median length of 550 residues (see Figure 2.7: A). The overall length distribution thus correlates with the ratio between the total residuals of the P- and L-models, and potentially contributes to the observed effect. To estimate the number of single and multi-domain proteins, we randomly sampled each of the three data sets and used HHpred [Söding, 2005] for their domain annotation against the ECOD database [Cheng et al., 2014], as presented in Subsection 1.2.4. We considered proteins multi-domain if they had at least 2 domains assigned to them, otherwise we considered them as single-domain proteins. The predicted fraction of multi-domain proteins in our

bacterial dataset is 30%, which is smaller in *Arabidopsis thaliana* (25%) and greater in *Homo Sapiens* (35%). The number of domains per protein does not not correlate to the observed ratio between the total residual of the P- and L-model.
In order to evaluate the qualitative reason, namely that sequences of distinct composition are combined within proteins, we assessed the fraction of structured and unstructured regions in the used proteins.
To that end, we estimated the fraction of structured regions for each protein with HHpred against the ECOD database (Figure 2.7: B). For the bacterial dataset, 40% of all sampled proteins are predicted to be structured over $\geq$90% of their sequence, a fraction that is smaller in *Arabidopsis thaliana* (15%) and *Homo sapiens* (13%). The structure content of proteins thus also correlates with the ratio between the total residuals of the P- and L-models (Figure 2.7: A), possibly because scrambling between structured and unstructured regions leads to greater compositional disturbance than scrambling within these regions. In total, the fraction of structured residues is 70% in the case of the bacterial data set.
We conclude that the L-model approximates the natural distance distribution better than the P-model in all cases, however in a more pronounced way for data sets containing heterogeneous mixtures of long sequences combining structured with unstructured regions. In our analysis, these effects were more pronounced in eukayotic than in bacterial proteins. The hypothesis of an unexpected homogeneous amino acid composition within proteins is further investigated in Chapter 4. Therein, compositional fluctuations between domains within the same protein could be substantiated, indicating that within structured regions, compositional fluctuations occur to be minimized.

2.4 Impact of homology and convergence

Having accounted for compositional effects at increasingly local level, the remaining discrepancy between the distance distribution of the L-model and that of the natural data set can be related to the actual sequence of amino acids. This discrepancy arises supposedly either through divergence from a common ancestor (homology) or convergence as a result of structural constraints, particularly secondary structure formation (analogy). In order to evaluate the relative contribution of these two mechanisms to similarities among natural sequences the proportion of pairwise alignments that can be assigned confidently to either homologous or analogous relationships needs to be identified to then evaluate their contribution to the overall similarities.

2.4.1 Decomposition based on distance assignment

A set of fragment pairs was analyzed and categorized into homologous, analogous and unknown relationships. These three groups of fragment pairs display different pairwise

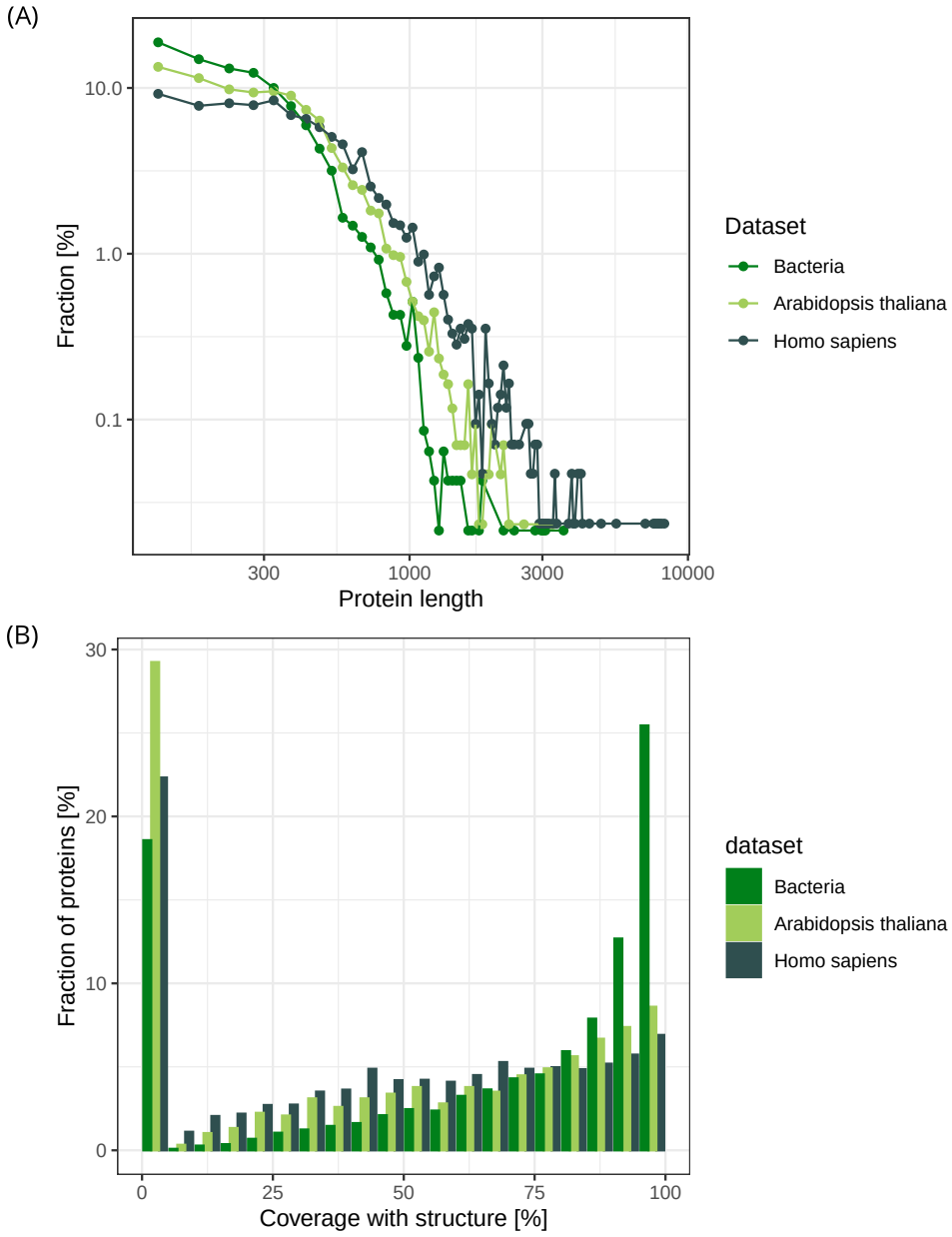

Figure 2.7: Contrasting the bacterial data set with two eukaryotic proteomes. (A) Distribution of protein length. The median protein length is smallest for bacteria with 315 residues, 400 residues in the *Arabidopsis thaliana* dataset and 550 residues in the *Homo sapiens* data set. The increase of median protein length correlates with the decrease in the total residual of the D-model relative to the P-model. (B) Coverage of proteins by structured domains. For each protein in the three datasets, an estimate of the coverage by structured domains was obtained by assigning ECOD families to regions in the protein. The fraction of residues within assigned domains compared to the protein length was obtained and plotted as a histogram over all sampled proteins. In bacteria 40% of the sampled proteins are almost completely structured (coverage of $\geq$90%), a fraction that is greater compared to that in *Arabidopsis thaliana* (15%) and *Homo sapiens* (13%).

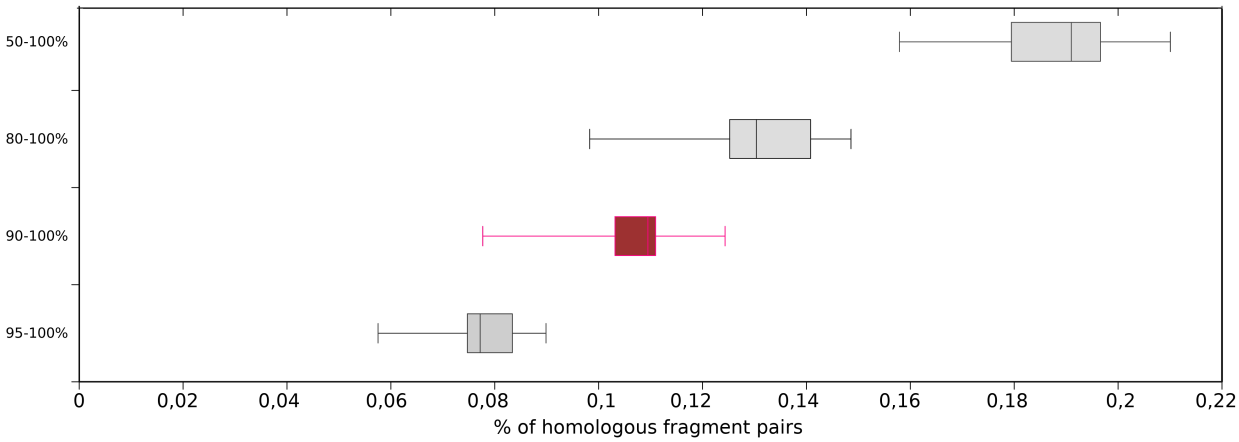

Figure 2.8: Assigned homologous relationships dependent on threshold for HHpred. The boxplots indicate the distribution over the 10 generated sets of relationships. The used value for a threshold of 90% is plotted in red.

sequence similarities and their contribution to the observed natural distance distribution was analyzed by accounting for their individual sequence bias.

Sufficiency of sampling

The detection of homologous relationships requires advanced approaches, which are computationally much more expensive than simple sequence alignments. We therefore only considered a small subset of our sequences and their relationships within this subset, which could be derived computationally in a reasonable amount of time; the exact sampling procedure is described in Section 2.2.3. With this sampling, the total residual of the L-model could be recovered down to double digit precision. We further demonstrate, that it was also sufficient to estimate the number of confident homologs and analogs in the following.

Fraction of homologous, analogous and unknown relationships

The approach to classify pairs of fragments into homologous or analogous relationships is described in Subsection 2.2.3. The remaining unclassified pairs, are annotated to have an unknown relationship.
With this approach, more than 4900 distances were identified as confidently homologous, for a probability of 90% or more according to HHalign. This corresponds to 0.11% of all (2 million) sampled relationships. Comparing results across the independent samples of relationships has led to a standard error of the mean (SEM) of 0.0043%, implying an single digit precision. In Figure 2.8, the fraction of predicted homologous fragment pairs

is plotted for different thresholds. Going down to a threshold of 50% probability would increase the magnitude of presumably homologous fragment pairs by only 1.7-fold to 0.19%.
The fraction of analogous relationships was identified to be 51.17% with an SEM of 0.64%. All sequence pairs that could not be confidently assigned to either group were considered to be of unknown relationship, amounting to 48.72% of the total with an SEM of 0.64%. The sampling is thus sufficient and it results in an accurate reflection of the true fraction of confidently homologous and analogous relationships according to the used methods.
The number of confidently analogous pairs exceeds the number of homologous pairs by more than 2 orders of magnitude. This indicates that the influence of homology on the global distance distribution in natural sequences is dwarfed by analogy. Considering the great number of presumably non-related domain folds (see Subsection 1.2.3), this finding is not surprising given that a cross-family comparison is more expected than a within-family comparison.

2.4.2 Sequence bias related to homology and convergence

Having decomposed sequence pairs into confident homologous and analogous relationships, we analyzed to what extent the remaining total residual of the L-model can be explained by incorporating corresponding sequence biases into our L-model. Therefore, we generated three new hybrid models in the following way: we omitted either homologous pairs, or analogous pairs, or both from our set of assigned relationships, generated an L-model for the remaining fragment pairs through the same shuffling procedure as used previously, and then added back the omitted pairs without shuffling. In the following we refer to the hybrid model that adds the sequence bias of homologs to the domain composition as the L1-model, the one that adds the sequence bias of analogs as the L2-model, and the one that adds both biases as the L3-model.
The residuals of these three models are compared to that of the L-model (Figure 2.10: A). Due to the reduced sampling over only 2 million fragment pairs, instead of 500 million, the distance distribution of the L-model in this analysis is not exactly the same as that obtained over the entire data set. However, rounding to two digits leads in both cases to a residual of 0.82% (Figure 2.10: B).
Relative to this distance distribution of the purely compositional L-model, the L1-model, which includes homologous sequence effects, is only minimally better (total residual reduced by 0.016%) in approximating the natural distance distribution (Figure 2.10: B). We assume that two reasons are mainly responsible for this only minor improvement: First, the proportion of homologous relationships is only 0.11%, giving them little leverage. Second, the distance distribution of homologs (Figure 2.9: B, yellow) differs only to a small extent from the distance distribution of the natural data set. It has been recognized previously that most homologous sequences share no significant similarity [Rost, 1997].

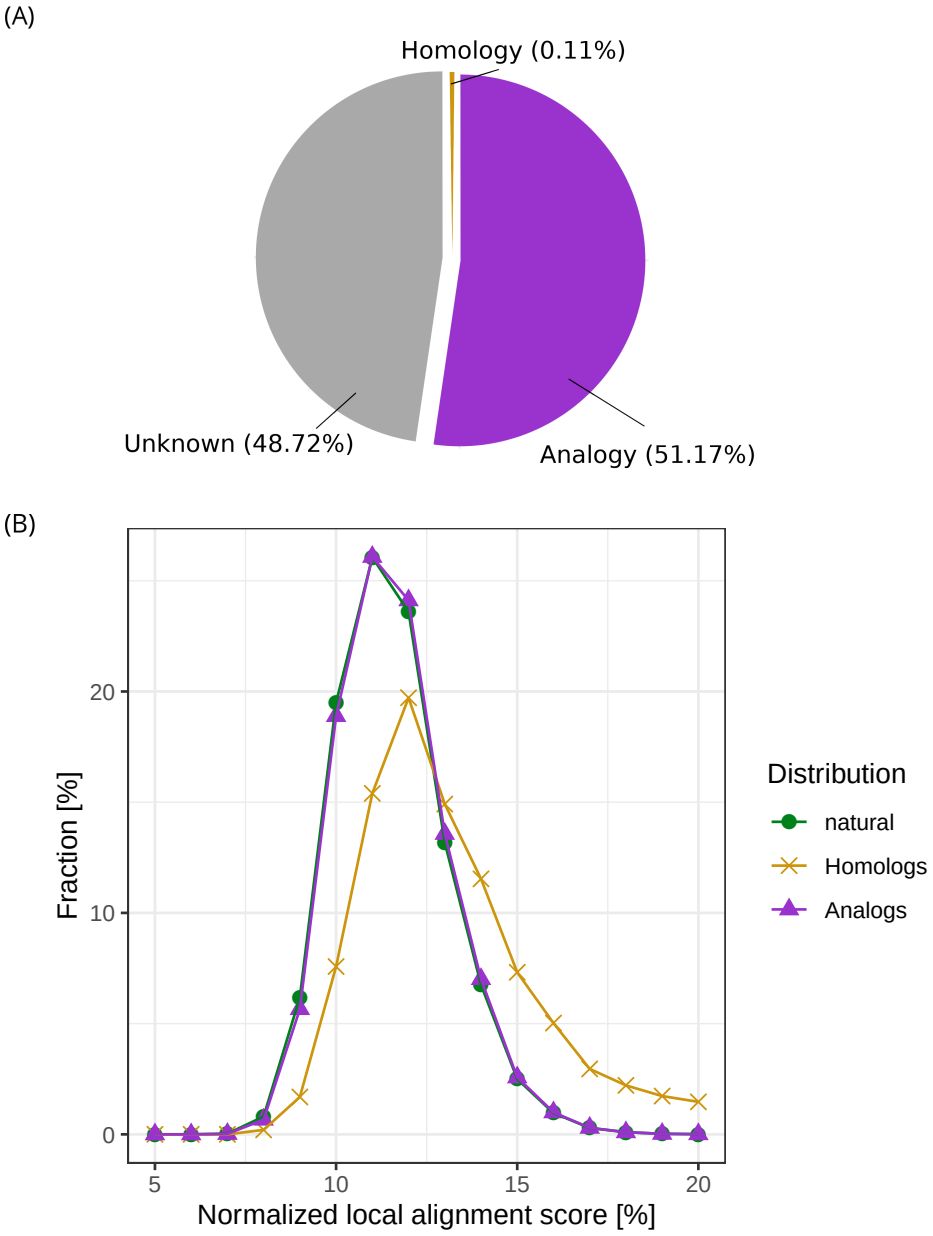

Figure 2.9: Contribution of homology and analogy to the natural distance distribution. (A) Decomposition of fragment pairs into their origins. We sampled 2 million fragments pairs and analyzed if their relationship is confidently homologous or analogous. The fraction of analogous relationships was determined to be 51.17%, homologous relationships only 0.11% and the remaining fraction is labeled as unknown origin. (B) Distance distribution between homologs and analogs contrasted with the natural distance distribution. The qualitative difference between the distance distribution of analogs and that of all fragments is relatively small. Compared to this, the distance distribution of homologs displays a tendency towards a higher sequence identity score. It nevertheless has a major overlap with the natural distribution.

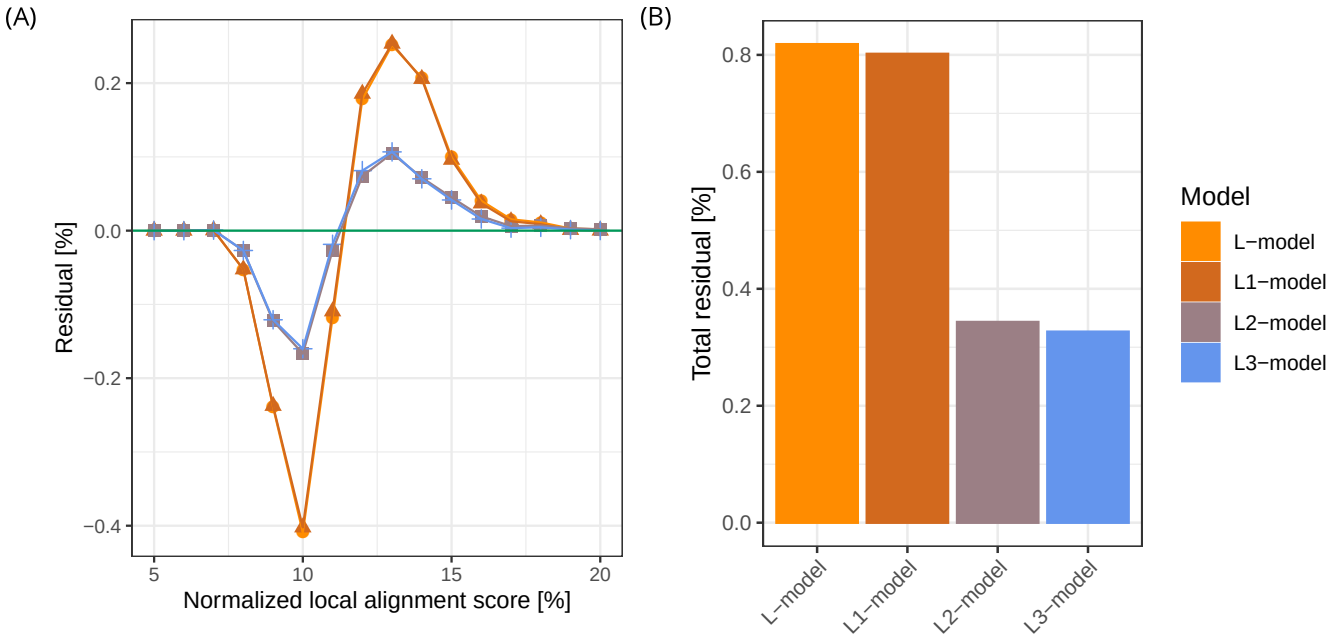

Figure 2.10: Models incorporating sequence bias of homology and analogy. (A) Residuals of the models incorporating the sequence bias of homology and analogy. We generated mixed models, that include the sequence bias of homology (L1-model), analogy (L2-model) and both (L3-model) into the L-model. The L1-model hat almost the same residuals as the purely composition-based L-model. The residuals of the 2-model deviate severely from that of the L-model and the L3-model yields similar results as the L2-model. (B) The total residuals behave according to the residuals. The total residual of the L-model over the sampled fragment pairs is 0.82% and is in accordance with the total residual of the L-model over the entire dataset. The L1-model displays an only minor improvement in the total residual of 0.016%. The L2-model reaches a total residual of 0.34% and is more than 2-fold more accurate than the L-model. Adding the homology bias to the L2-model to obtain the L3-model has almost no effect.

In contrast, the total residual of the L2-model (0.34%), which includes analogous sequence effects, is decreased about 2.4-fold relative to the L-model (0.82%). Thus, although analogs have a similarity distribution that is very similar to the natural (Figure 2.9: B, purple and green), their leverage is 2 orders of magnitude higher than that of homologs, causing these small differences to improve substantially the fit of the L2-model to the natural distance distribution. Most sequences in our natural data set share the ability to form secondary structures (see Subsection 2.3.3), resulting in a sequence bias that is not fully captured by residue composition [Pande et al., 1994; Lavelle and Pearson, 2009]. As expected from the L1-model, adding the homologous sequence bias to the L2-model did not really improve its ability to approximate the natural similarity distribution.

2.5 Discussion and Outlook

2.5.1 Summery

Similarities among natural protein sequences are globally well represented by a random sequence model, that accounts for the overall natural amino acid composition. This representation can be refined by including the compositional bias of genomes and furthermore that of individual proteins. The remaining sequence effects could be associated to similarities among analogous sequences, and are thus a result of convergent evolution. The global effects of divergent evolution are thus negligible, implying that natural sequences do not globally share significant overlaps due to common ancestry. Evolutionary constraints are supposedly very specific for individual sequence clusters. They are not generally applicable to all natural sequences and do not shape the global occupation of sequence space.

2.5.2 Novelty of this study

The global occupation of sequence space has successfully been described for sequences up to a length of 5 residues [Poznański et al., 2018; Lavelle and Pearson, 2009]. Methods used in these studies were not applicable to longer sequences due to the increasing sparsity of sequence data with fragment length. With the distance-based approach used in this thesis, this problem could be circumvented by considering relationships between sequences instead of sequences themselves, allowing to study the relative position of sequences of arbitrary length.
The here presented approach of comparing distance distributions between natural and random sequences is used to interpret characteristics of the occupation of sequence space. Although several studies use similar distributions to explore features of natural sequences [Rost, 1997; Buchholz et al., 2018], they have not been interpreted them in the light of

the global structure of sequence space occupation.
A large corpus of literature focuses on the categorization of proteins sequences into cluster of related entities. This has led to a wide-spread image of protein sequence space to be populated by functional islands in a huge sea of possibilities. Our results here challenges this image of islands, implying that even these posses a rather random global shape. This view is further discussed in Subsection 5.2.2.
The observation of the mostly random structure of protein sequence space occupation is in accordance with studies that have demonstrated that globally, natural protein sequences behave mostly like random sequences [Lavelle and Pearson, 2009; Weiss et al., 2000; Strait and Dewey, 1996]. However, minor deviations exist, which is also undoubted in these studies.
Having accounted for local composition bias of genomes, proteins and domain-sized fragments, we were able to differentiate between deviations arriving from compositional effects at either level. Comprehensive studies that use a diversity of random models to contrast natural protein sequences are rare as most focus on specific deviations relative to one chosen model. Our approach can be used to estimate compositional heterogeneity at different levels and the effect to sequence comparisons within a heterogeneous data set, also as a function of sequence length. With this, it was possible to demonstrate that the use of the commonly used A-model approximates similarity between natural sequences worse, when considering longer sequences.
Furthermore, we could demonstrate that compositional effects contribute the most to the deviations between natural and random sequences and that they are greater than sequence effects. In Section A.3, the remaining sequence bias is captured, free from compositional biases. The usage of compositional effects at the protein level has previously been acknowledged and is part of the BLAST program [Schaffer, 2002]. Similar to our finding that L- and P-models produced comparable total residuals, the authors find that the composition at an even lower level does not significantly contribute to an improvement of the alignment accuracy.

Chapter 3

The transition between global and local sequence space

3.1 Motivation

On the global scale, differences between natural and random sequences are only hardly detectable. In Chapter 2, this observation has been studied using an approach that studies the occupation of sequence space through pairwise distances between observed sequences. However, sequences often share similarities [Pearson, 2013] and demonstrate clustering in local proximity through common ancestry, which deviate from random behavior. In many cases, function and structure can be extrapolated from these related sequences, enabling to associate unstudied sequences with knowledge about existing, studied sequences [Koonin et al., 1995].

3.1.1 Content of this chapter

In Section 3.2, an approach that focuses specifically on the local sequence space occupation due to common ancestry is presented. It is based on the approach presented in Chapter 2 and analyzes all distances of one to all other sequences, thereby detecting the abundance of sequences in local proximity. Characteristics in the occupation of the local sequence space around individual sequences can be interpreted as evolutionary footprints, indicating abundance and presumably time of duplication events. With this approach specific sequences can be analyzed in the context of a given data set.
In Section 3.3, the transition from local sequence space, as defined by certain homologous sequence similarity, to the global mostly randomly occupied structured sequence space is being analyzed. This transition has been studied with a focus on the twilight zone and the homology indicating sequence threshold of structurally homologous sequences [Schneider et al., 1997; Rost, 1999]. My work builds upon these studies by using new and more sequence data to study this transition in detail. It extends from previous ideas by the search for a sequence-specific threshold.

3.2 Exploring the local sequence space

The population of the local sequence space around a specific sequence can be determined by contrasting one sequence against all others in a respective data set and selecting those sequences that are in close proximity. Similarly to the all against all heuristic of distances among sequences presented in Figure 2.2, here, distances are investigated between one to all other sequences of the respective data set. This type of distance distributions are further referred to as *local distance distributions*.
For this study, the bacterial data set (see Subsection 2.2.1) was used and distances were derived using the Hamming distance a metric (see Section 2.2.2). Using the Hamming distance, gradual diversification through point mutations can be captured, while insertions or deletions are not accounted for. In Figure 3.1 four examples of local distance distributions are presented. The query sequence is indicated in the upper part of the plots and the local distance distributions are illustrated in green. The random distribution, indicated in red, is identical to the closed form A-model, which accounts for the compositional bias of the overall data set.
In both cases, the overall distance distribution (Figure 2.2: B) and these local distance distributions (Figure 3.1), the natural distribution starts to deviate from the random at approximately 20% sequence identity. Irrespective of the chosen metric, this transition marks the transition from an equal abundance of distances to smaller distances that are over-represented in the natural data. This transition can be interpreted as a junction between local to global sequence space, where smaller distances can no longer be accounted for by the random model. The most prominent difference is the irregular behavior of the local distributions for a sequence identity above this threshold, which decays in an almost perfectly exponential way in the overall distribution. Therein, all local distributions are summarized. This irregular progression illustrates the specific occupation of the local sequence space around the query sequences. It is indicative of the visible evolution of sequences around the query sequences and can be viewed as an evolutionary diffusion footprint in the local sequence space.

3.2.1 Interpretation of evolutionary footprints

The query sequence in this one against all approach marks the reference position in sequence space. Analyzing the population of sequence space regions with 100% to 0% sequence identity to the query sequence, corresponds to a radial inspection of the multidimensional sequence space around the query sequence at an increasing radius. In Figure 3.2 this concept of a radial inspection is depicted in a two-dimensional plot. At the center, the query sequence is represented as a red circle. The black circles represent the border of the local sequence space with increasing point mutation distance. The blue circles represent sequences at a certain location in sequence space. These 2D representations are not dimensionally equivalent to the actual spacial distribution of sequences and are just for illustrative purposes.

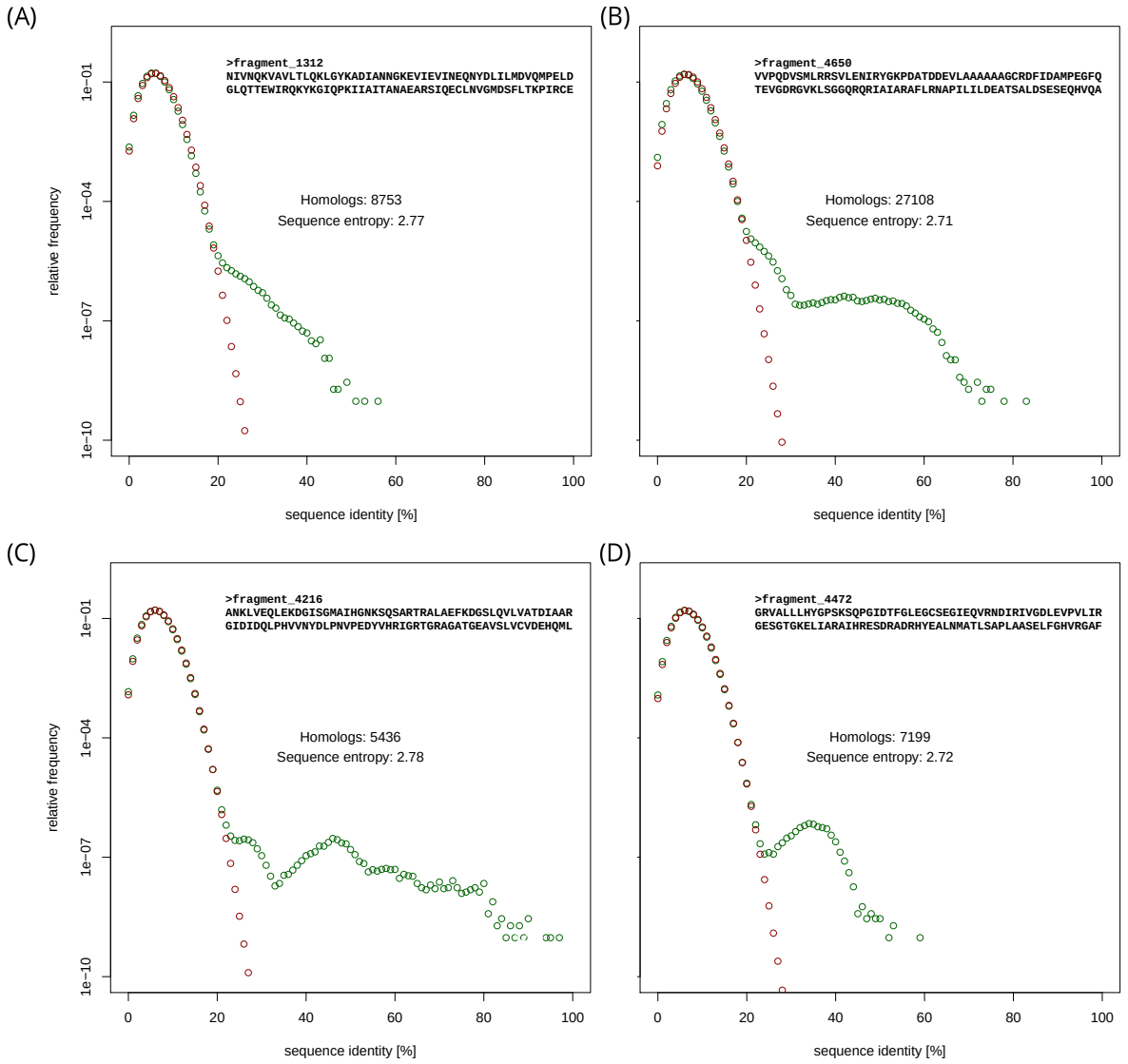

Figure 3.1: Examples of the local sequence space. (A) local sequence space around the query is populated in a sequence identity area of 30-50%, suggesting that no truly recent duplication events of these sequence have occurred within the given dataset as the closes sequence has an identity of 50%. (B) there are two bumps, one rather broad increase ranging in sequence identity area of 30-80% and a sharper one below 30%. This sharp bump can be imagined to possess a continuation into the region of even less sequence identity. Possibly it can be explained by old duplication events were sequence identity is already only noticeable and merging into an area of expected similarity between random sequences. (C) sequence similarities of all degrees are occurring, suggesting that probably this sequence is still being duplicated and occurs in multiple species in a very conserved manner. (D) an increased abundance of distances with 20-50% sequence identity implies a relatively abundant and old duplication.

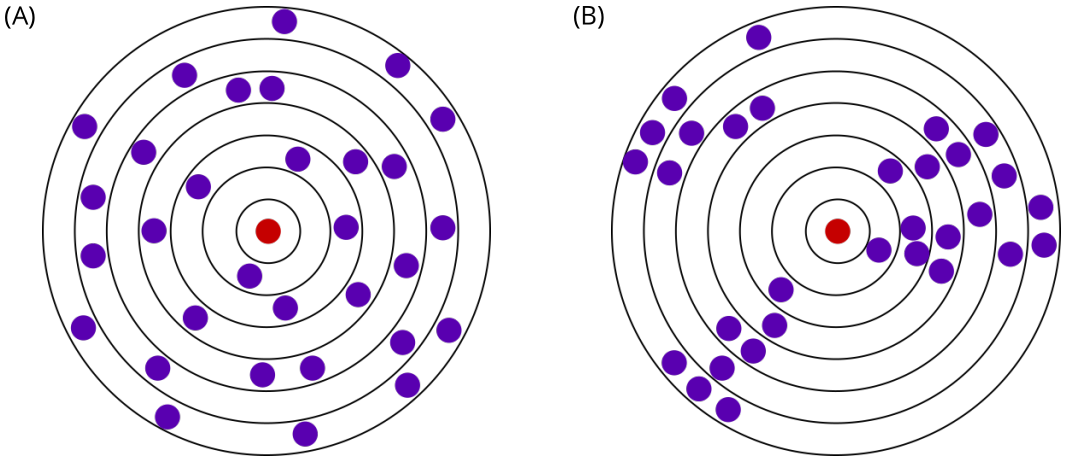

Figure 3.2: Sketch of local sequence space occupation in 2D. (A) the local sequence space around the query sequence is populated, in a rather random fashion, distributed into all directions. (B) there are three distinct local regions that are populated. These may emerge if specific, different residues are preserved.

The closer a sequence is to the query sequence, the more likely is their common descent. This suggests that at some point in time, a sequence was being duplicated and the two sequences (original and copy) have diverged into the query and the observed sequence in the local vicinity of the query's sequence space. It is suggestive that relationships with a higher sequence identity correspond to a recent duplication event, given that the sequences have not diverged far from each other. They could also be exposed to greater evolutionary pressure, which enforces high conservation. In Figure 3.1: C, sequences with a high sequence identity to the query sequence exist.
A sequence that is further away to the query sequence, suggests that (if they are of common descent) they have diverged more and probably the corresponding duplication has occurred a longer time ago. The observed divergence may have affected both sequences to different extends. It is thus unclear which sequence is closest to the common ancestor.

3.2.2 Iterative expansion

Distances in the local sequence space are suggestive of the amount and time of duplication events that are related to the query sequence. However, using distances, the direction of divergence is blended out, which can comprise more specific features about the occupation of the local sequence space. In Figure 3.2, the number of sequences at a certain distance to the query sequence is the same, corresponding to the same local distance distribution. However, the local sequence space is differently occupied. While in the first sketch sequences are scatted in a rather random fashion over the whole space, they are clustered into three groups in the second. Such clustering in the high-dimensional space

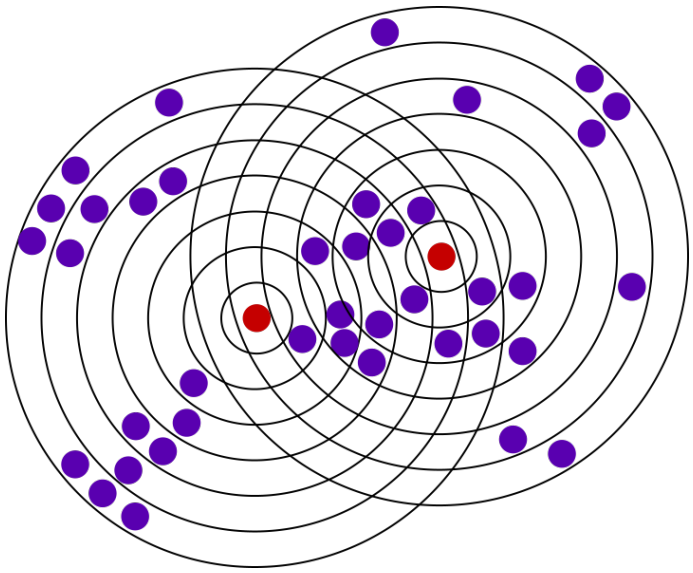

Figure 3.3: Expanding the view of local sequence space occupation iteratively. In order to analyze if an accumulation of specific distances is due to a locally accumulated occupation, the local sequence space around sequences at the respective distance can be derived. This iterative approach can also be used to generate connected components via significant sequence similarities.

of sequences can occur, if few positions are identical among the respective sequences. This kind of clustering cannot be extracted from the local distance distributions alone and needs further investigation to be revealed.
An increased abundance of a certain distance can be indicative of a clustering of similar sequences at this distance, as it corresponds to duplicate sequences that have a similar distance to the query sequence. In Figure 3.1: D, for example, distances of about 35% sequence identity occur more frequently than those with 5% higher or lower sequence identity.
Using sequences at this specific distance as new query sequences, can shed light onto the question whether local sequence space of these sequences is densely occupied. This approach is illustrated in Figure 3.3. In the case of a random occupation at this distance, the local sequence space of the corresponding sequences will be less populated.

Aiming to collect all sequences that are inter-connected by significant similarities, this approach can be used in an iterative way. All found sequences can be summarized into connected components, and further analyzed for inter-connectivity, functional divergence or cluster sizes. These components are likely to cross the boundaries of local sequence space, where sequences share a significant, pairwise similarity, as dissimilar sequences may be connected through other sequences that share significant similarities. An exhaustive study on connected components has been performed by joining sequences with

a one-point-mutation distance and is presented in Section B.1. Therein, large clusters were found to percolate sequence space. Their sizes are distributed according to the power-law.

3.2.3 Node degree distribution of homologous neighbors

The evolutionary footprints in the local sequence space can be used to estimate the abundance of significant sequence relationships that supposedly emerged through common descent. The number of presumably homologous neighbors H is in the following derived as the sum of the difference between the natural DD_{nat} and random distance distribution DD_{rand} over a sequence identity $s \geq 7\%$ multiplied by the data base size $|DB|$. A sequence identity of 7% marks the maximum of the random distance distribution, therefore, differences of distances smaller than expected on average are neglected.

$$H = \sum_{s \in 7-100\%} (DD_{\text{nat}}(s) - DD_{\text{rand}}(s)) \cdot |DB| \tag{3.1}$$

With this, not only the very significant relationships are accounted for but also the over abundance of less significant sequences. When constructing a network of sequences, that are connected if they are assumed to be related, this number of homologous neighbors corresponds to the node degree of the query sequence. For similarities with an identity of $\geq 28\%$, no random sequences share such a high similarity, hence all natural relationships with a minimum identity of 28% can be assumed to be confidently homologous and a network between these can be constructed. For similarities between 7-27% there is a certain ratio of natural to randomly expected similarities and the relevant, truly homologous connections are not obvious from this comparison alone. Although it is unclear, to which sequences the distances with less significant similarity belong, they can be included into the estimate of the node degree, as their presence is not reproducible by the random model. The possibility of convergence is further discussed in Section 3.3 and is not exluded in this approach.
This heuristic attempt differs from many standard network theory approaches, that usually construct connected components or clusters according to given thresholds [Buchholz et al., 2018; Nepomnyachiy et al., 2017]. Thereby, a low threshold of significant sequence identity will lead to false positive connections in the network that are randomly expected. A high threshold will lead to the exclusion of connections between truly homologous sequences. The choice of the threshold thus entails a trade off between the inclusion of false connections and the negligence of true connections. Approaches that depend on such thresholds lack a correction according to the estimated number of true neighbors. Other approaches circumvent this problem by connecting all sequences and weighing edges differently according to their significance [Alva et al., 2009]. However, in both cases the node degree is not directly corrected to obtain the estimated value as derived from the difference between natural and random local distance distribution, as

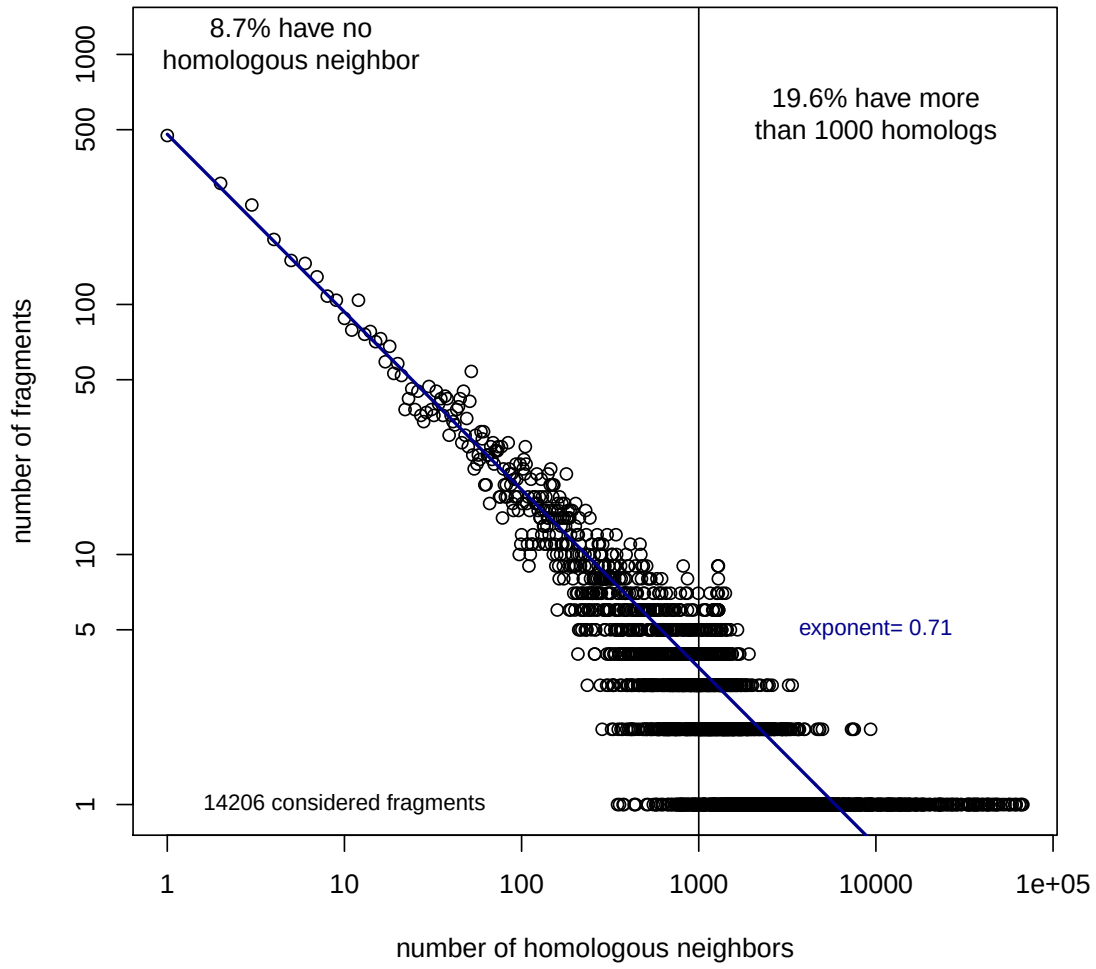

Figure 3.4: Power law-distributed node degree of homologs. Within the bacterial data set, the number of homologs for a given 100mer as defined by Equation 3.1 is distributed according to the power law with an exponent of 0.71. This low value indicates a great redundancy, caused by the use of multiple genomes. In total, 8.7% of all 100mers has no identified homologous neighbor and 19.6% had more than 1000 homologs.

presented here.
To further study the connectedness through homology, the node degree can be investigated over a large set of sequences. Its distribution follows the power law, a behavior previously described in many other studies of natural sequences [Buchholz et al., 2018; Dokholyan et al., 2002; Deeds et al., 2003; Koonin et al., 1995]. Networks that possess a node degree distribution following the power law are scale-free, implying that no specific cluster sizes exist (see Subsection 1.2.5). This implies that with growing amount of data, the largest clusters become larger than smaller ones. In Figure 3.4 the frequency of node

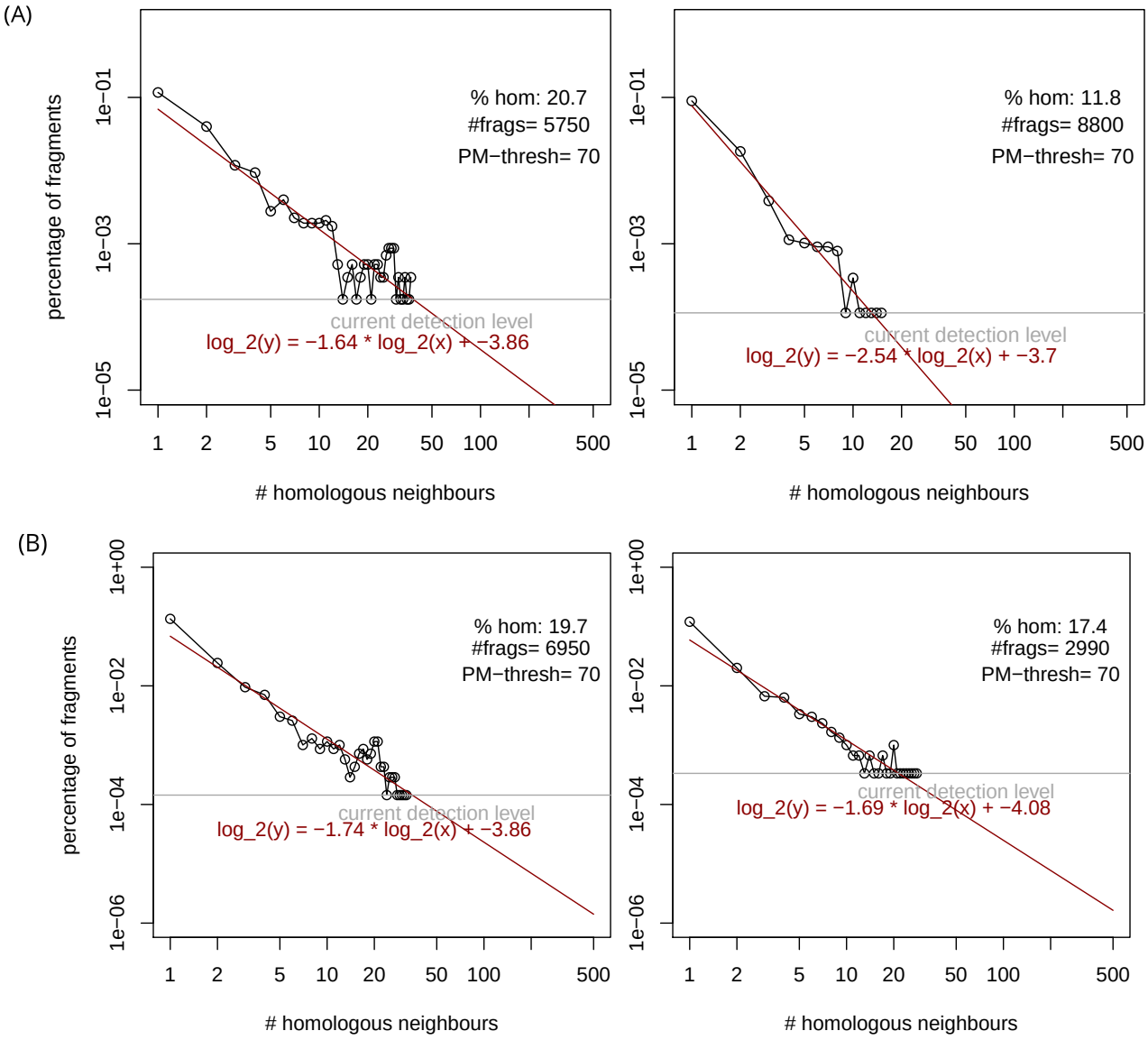

Figure 3.5: Paralogous and orthologous reuse. (A) Within a genome, the paralogous reuse of sequences is power law-distributed. The exponent depends on the used data set. (B) Comparing homologs across two genomes leads to similar results.

degree over the bacterial data set is plotted. The exponent of the fit to the power-law is 0.71, which is compared to many other networks relatively low as it generally ranges between 1 and 3 [Koonin et al., 1995]. This implies that there is a tendency of more sequences to have more neighbors.
This fact probably relates to the use of a redundant set of 1,307 bacterial genomes. Or-

thologous reuse of the same sequences in different organisms may cause the exponent of the power law to be this low.

In a small-scale analysis of paralogous reuse within individual genomes of the same data set, the exponent of the power law was relatively increased, ranging between 1 and 4. In this analysis, homology was assumed between 100mers that share at least 30% sequence identity. Therefore a different definition of homology was used to infer the node degree than provided in the above. In the upper panel of Figure 3.5, three examples of inner genome reuse are depicted. Similar results were achieved in a cross-genome comparison of two genomes. Three such examples are depicted in the bottom panel of Figure 3.5. In each plot the number of investigated fragments is indicated as well as the percentage of these fragments that possessed at least one homolog.

3.3 The twilight zone of sequence similarity

The term twilight zone is coined by [Russell F. Doolittle, 1981] and further expanded by [Rost, 1999] considering structural homology. It refers to the sequence identity area of 20-35%, where sequence analysis often fails to correctly distinguish homologous from non-homologous relationships. Due to advances in bioinformatic methods and the growing amount of sequence data, homology detection has become more sensible since then. The exact borders of the twilight zone in this definition are likely to have shifted. What has remained, is the fact most homologs share a pairwise sequence identity below 20% (see Section 2.4), an area populated by randomly expected similarities [Krause, 2000; Russell F. Doolittle, 1981; Rost, 1999], that is beyond the twilight zone.
In the following, I will refer to the area, where the significance of sequence identity relative to a random model is ambiguous as the twilight zone. This area is strongly dependent on the considered sequence length, as has been studied in [Schneider et al., 1997] and [Rost, 1999]. In this study, these results could be reproduced with the slightly different definition of homology by significant sequence similarity not structural similarity. Furthermore, I present an idea, aiming to refine the definition of the twilight zone.
Homology can be assumed between sequences with a similarity that is not expected under random condition and is not caused by convergence. Similarities that approach the region of randomly expected similarity are less significant, having no significance when reaching a similarity that is expected between random sequences. A transition between this confident similarity into this twilight zone can be extrapolated from the abundance of naturally observed similarities, contrasted to the randomly expected abundance.
A natural distance distribution that with an logarithmic scale on the y-axis, demonstrates a tail distribution in the significant area of sequence identity. This tail transitions into the randomly expected behavior at about 22% sequence identity, as depicted in Figure 3.6. The interval between the natural and random distance distribution indicates how over-represented the corresponding identity is between natural sequences. It thus reflects the

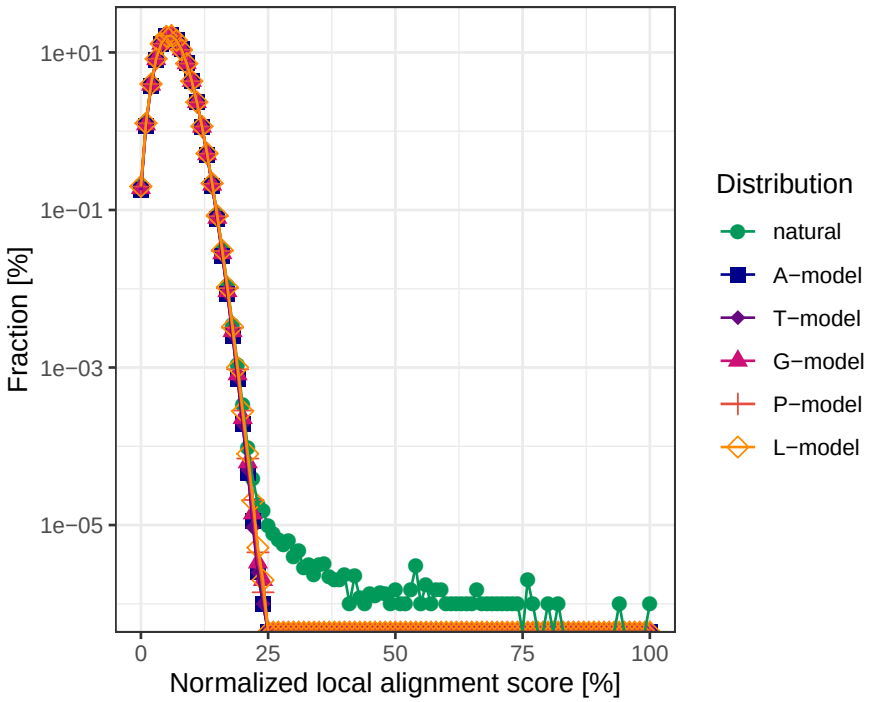

Figure 3.6: Logarithmic depiction of distance distributions using Hamming distance. The tail of the natural distance distributions deviates from the random models at a sequence identity of about 22%.

significance of this distance.
This significance contrasts all phenomena causing enhanced similarity between natural sequences against those in a random model. However, convergent similarities due to composition are common among natural sequences in the twilight zone (see Figure 2.3). It is likely that the remaining differences due to sequence preferences of common structures cause convergent similarities that have not been accounted for here. In order to study the contribution of convergent similarities to the twilight zone, a decomposition of the natural distance distribution can be applied.

In an early attempt of such a decomposition of the natural distance distribution, I proceeded based on the hypothesis, that the natural distance distribution comprises a large random part ε, as captured by the random distance distribution. The remaining part was assumed to be composed of either a distribution among homologous sequences or a biased distribution, associated to convergent effects among natural sequences.

$$DD_{\text{nat}} = \varepsilon \cdot DD_{\text{rand}} + h \cdot DD_{\text{homologous}} + (1 - h - \varepsilon) \cdot DD_{\text{biased}} \tag{3.2}$$

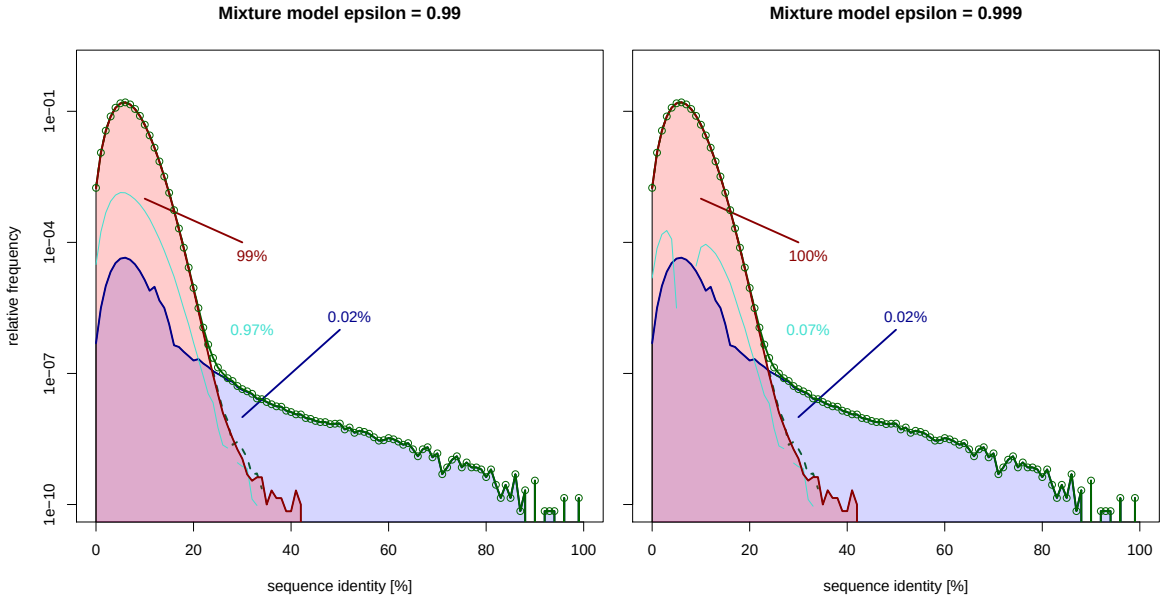

Figure 3.7: Mixture model. The natural distance distribution can be decomposed into distributions associated with distinct origins. A large part of it can be captured by the distribution is a random model. For the long tail in the high sequence identity area, homologous descent is responsible. The remaining, biased distances may be due to convergence.

For this attempt, I used the Hamming distance as metric and the A-model as random sequence model. The homologous distribution $DD_{\text{homologous}}$ was derived by sampling fragment pairs with each sequence identity and applying HHalign to the raw sequences, an approach, which has been improved in the final decomposition presented in Subsection 2.2.3. The scaling factor of the homologous distance distribution $h = 0.02\%$ was chosen to fit the tail of the natural distribution at sequence identity ≥ 40 %. The scaling factor ε of the random, A-model distribution was adjusted such that the remaining distribution of the biased distribution DD_{biased} still possessed a bell-shaped form.
The first panel of Figure 3.7 illustrates the predicted decomposition for $\varepsilon = 99\%$. Therein, the weight of the biased distribution indicated in cyan accounts for 0.97% of the whole natural distribution. The second panel illustrates an undesired fit, there the remaining biased distribution possessed a negative faction due to the over-estimation of the random distribution of $\varepsilon = 99.9\%$.
With this approach, there is no obvious way to judge, whether the chosen ε correctly reflects the random fraction of the natural distance distribution. Choosing $\varepsilon = 99\%$ as it is the largest value that still results into a smooth distribution of the remaining biased part, may be an arbitrary choice. A different, more advanced way to decompose the natural distance distribution was taken in Subsection 2.2.3, which also distinguishes between

confidently analogous relationships.
In order to better distinguish homologous, convergent and randomly expected sequence identity, the decomposition can be transformed into a representation that reflects the ratio between these causes along a certain sequence identity. In Figure 3.8 this representation

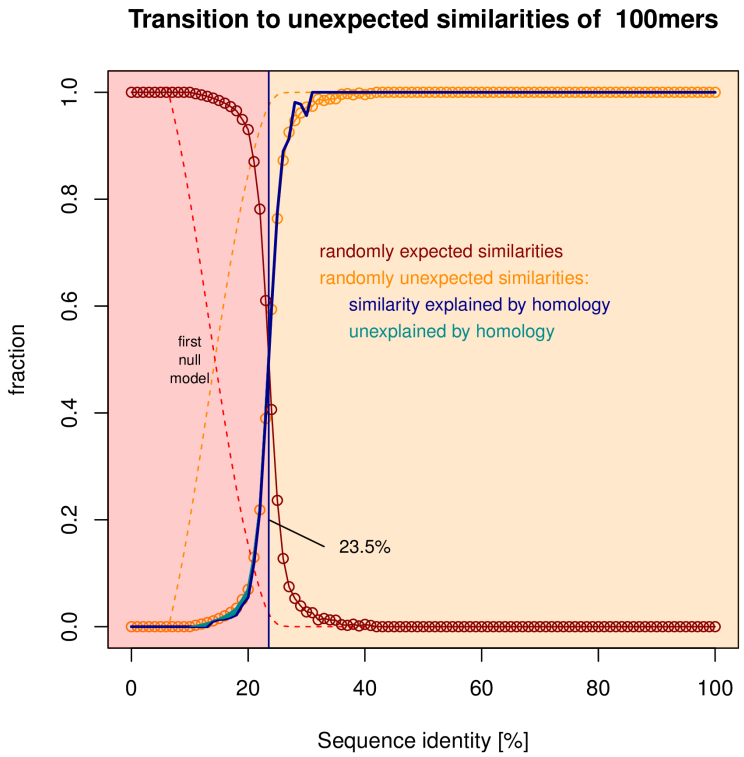

Figure 3.8: Transition from expected to unexpected similarity. The major proportion of similarities switches from expected to unexpected for a sequence identity above 23.5%. Distances above this threshold are more often unexpected by the random model and belong mostly to homologous similarities. For shorter distances, unexpected similarities could be assigned to convergent features.

of 100mers is depicted. The randomly expected fraction of similarities is colored in red and the fraction of randomly unexpected similarities is colored in orange. The region where sequence similarity is not clearly assignable to either randomly expected or unexpected ranges around 15-30%. The sequence identity of 23.5% marks the transition point, where randomly expected or unexpected similarities are equally represented. A range of 15-30% sequence identity is slightly smaller than that reported as the twilight zone, which may be related to the strict comparison metric of the Hamming distance, used in this study.
The blue line corresponds to the fraction of homologs predicted by HHalign, the turquoise

area relates to the remaining similarities that could not be explained by the A-model or homology. Similarities of a higher identity than the transition point of the twilight zone have been associated to a homologous relationship. Only for some cases in an area of 15-23.5% sequence identity, similarities are caused by convergence. According to the results in Chapter 2 most of these may be related to the natural composition of genomes and proteins. The usage of the P-model may thus be more appropriate in order to account for similarities due to sequence not composition.

3.3.1 Twilight zone with fragment length

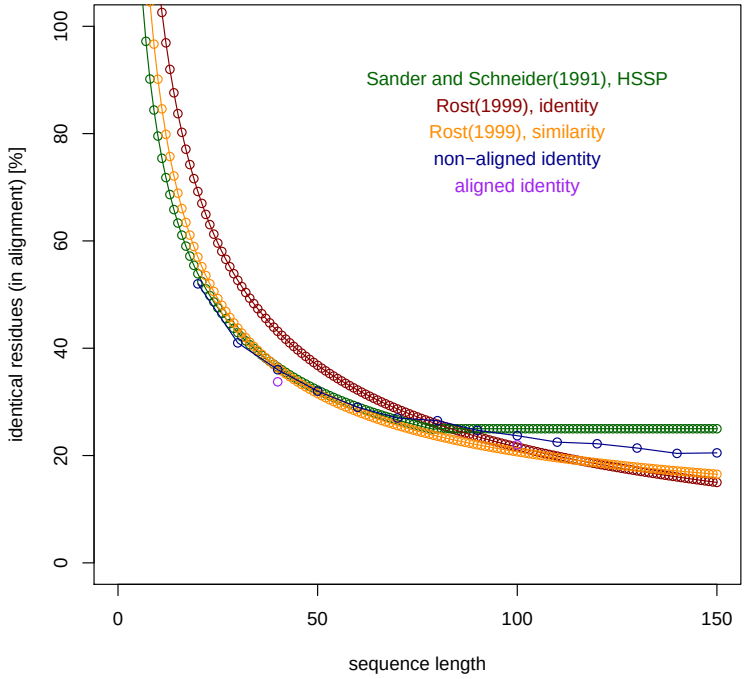

Figure 3.9: Twilight zone shifts with fragment length. With increasing fragment length, the sequence identity of the transition between expected and unexpected similarities decreases. The here derived transition coincides well with definitions of previous publications.

The twilight zone crucially depends on sequence length [Schneider et al., 1997; Rost, 1999]. For example, 100% identical 5mers are often reoccurring in a large random data set, while identical 100mers are not expected to occur by random chance. In order to compare the results of this study to existing results of the transition between expected and unexpected similarity, I derived the transition points between expected and unexpected similarities for sequence length 10 to 100 in steps of 10. In Figure 3.9, all results are presented next to each other. It is worthwhile mentioning, that in [Schneider et al.,

1997; Rost, 1999] the fragment length (x-axis) corresponds to the alignment length of a local alignment between longer protein sequences. Their distributions thus include the locally best matches over whole protein sequences, while in this study proteins were dissected into overlapping fragments of length 100 and all fragments were aligned against all other non-overlapping fragments. The presented distributions are thus not containing equivalent comparisons, also due to the fact that completely different data sets were being used. It is thus rather surprising how accurately all functions align to each other. This may be due to the fact that the effects of the exponentially growing sequence space with alignment length are simply larger than all mentioned effects in the above. This effect is discussed in detail in Subsection 5.2.1.

3.3.2 Sequence-specific twilight zone

The studies of the evolutionary footprints of individual sequences and the twilight zone as a general concept of the transition between significant and insignificant similarity, can be combined to better estimate the sequence-specific similarity. Considering a specific sequence, the threshold of the generally derived twilight zone may vary due to the different baseline probabilities of sequences. The sequence identity between a cysteine-cluster and some other sequence for example is more significant compared to the same sequence identity between a leucine-rich repeat and some other sequence. This suggests that the transition between expected and unexpected similarities will be moved towards higher sequence identity for less probable sequences and towards lower sequence identity for more probable sequences.
In a small-scale study of few arbitrary 100mers (using a Smith-Waterman alignment) I have tested this hypothesis. Instead of using the A-model, the query sequence was compared to random sequences of the P-model. The transition point varied only marginally in the range of 1% sequence identity. This may be due to the fact that the exponentially growing sequence space or due to the decreased variance of the fragment probability depends on longer fragments length. For shorter fragments the sequence identity of the twilight zone increases (Figure 3.9) and varies more among specific sequences.

Rational expansion

As noted in Subsection 3.2.2, it is possible to investigate if the increased abundance of specific distances are occurring within a local area of sequence space. It would be expected that sequences with the same distance to the query sequence are rather randomly scattered over the sequence space. A biased local occupation of sequence space by these sequences indicates that the reoccurring sequence pattern is under evolutionary pressure. Distances in the twilight zone can be investigated for such occurrences. If the query sequence overlaps with the locally conserved pattern, the relationship gains in relevance.
This idea is based on the principle that the likelihood of two sequences a and b to share some similarity is larger than the likelihood of them to share specific similarities $b_i = k$,

that are defined by the conserved region in the local sequence space.

$$P(\exists i : a_i = b_i) > P(a_i = b_i = k) \tag{3.3}$$

Proceeding with this approach can establish relevant relationships between sequences of similarities in the twilight zone. This may be relevant in cases, where sequences have diverged beyond recognition and pairwise sequence comparisons do not succeed to find any homologs.

Integration of sequence similarity

As pointed out in [Rost, 1999], using sequence similarity instead of sequence identity can lead to a more accurate detection of homologous relationships. In this specific case, further sequence similarity between non-identical amino acids can be captured and evaluated. It is also possible to use a different distance metric from the beginning that includes sequence similarity in the alignment of sequences.

3.4 Discussion and Outlook

3.4.1 Remarks about the methods used in this section

This study was performed as a side project in parallel to writing the paper about the main study of my thesis. The results presented here are thus not refined and a continuation of this project in my postdoctoral studies should include several aspects, which I address in the following.
The usage of the A-model is not optimal and should be refined, using a sequence-specific P- or L-model. Distances should therein be constructed by aligning the query sequence with shuffled protein sequences or shuffled domain-sized fragments. These model, serving the purpose of capturing convergent compositional effects, may distinguish between expected and unexpected similarities better.
The main work of this section is based on distance distribution derived by using the Hamming distance. More sophisticated alignment metrics can lead to different, more in-depth results as those presented here. However, the usage of a local alignment combined with the consideration of all overlapping fragments will necessarily lead to a smear of the evolutionary footprint, as overlapping fragment pairs will lead to more pronounced incrementally diversifying distances. Proceeding the same way as [Rost, 1999] and [Schneider et al., 1997], by aligning protein sequences with a local alignment metric and taking the alignment length as respective fragment length is conceivable. Using the same sequence lengths for the random model, the combinatorial effects of the local alignment should be accounted for.

3.4.2 Local, as defined by natural evolution and sparsity

The definition of local sequence space by its transition from naturally caused similarity through homology and convergence to random similarity is an abstract concept that only applies since natural evolution mostly proceeds in small steps through sequence space and major changes are often pruned. Actually, there may be many things to be found in the local sequence space. A study of [Alexander et al., 2007] proved that two unrelated folds can be transformed into each other by mutating only 7 amino acids. The designed sequences shared 88% sequence identity and are fully functional. Sequences close in sequence space can thus fold into completely different structures and also possess specific functions. This again demonstrates that the vastness of sequence space contains many things that have not been reached by evolution and that the concept of local and global sequence space is mostly relevant in the light of evolution, not structure or function [Valas et al., 2009].
The fact that we can assume that significant similarities (when compared to a random sequence model such as the P-model) arrive from common descent is merely due to the sparse occupation of sequence space - in both natural sequence data and the random model. Directed design efforts that explore sequence space specifically as they are not bound to evolutionary exploration of sequence space (as assumed by the random sequence models). Instead, they can tap into specific areas and find potent sequences in the local sequence space of existing natural sequences that are of distinct structure of possess different functions.

Chapter 4

Amino acid and codon composition of domains

4.1 Motivation

4.1.1 Amino acid composition of proteins and domains

In Chapter 2 compositional heterogeneity among natural protein sequences has been found to possess the greatest impact to their overall diversity. This effect is greater than any effects arriving from the actual sequence of amino acids. The importance of composition to protein structure and function has been highlighted several times before in different contexts.
Intrinsically unstructured proteins can be classified from structured proteins through differences in their amino acid composition [Dosztányi et al., 2005]. Their secondary structure content or aggregation propensity showed no characteristic sequence; instead amino acid composition alone was found to be the key role for these properties [Vymětal et al., 2019]. More specific structure could also be associated to amino acid composition, as unrelated proteins with the same folds were found to possess a similar composition [Ofran and Margalit, 2006; Dubchak et al., 1995]. In general, folds of the same structural classes tend to cluster together by their composition [Alva et al., 2009]. This correlation between structure and composition may be caused the respective secondary structure contents or to their specific packing, which has been shown to have an impact on amino acid composition [Fleming and Richards, 2000]. Furthermore, composition has been associated to the cellular location of proteins [Chou, 2001]. The therein used approach included the description of hyrophobicity, hydrophilicity and mass patterns along the sequence to the ordinary description of amino acid composition. Accounting for compositional fluctuations is part of the BLAST program [Schaffer, 2002] and allows to better separate similarities caused by structural or functional convergence from truly homologous similarities. Therein, the composition of natural proteins is accounted for, similar to the herein used P-model.
On a more local level, the L-model accounts for the natural amino acid composition of domain-sized fragments with a length of 100 residues. The distance distribution of the

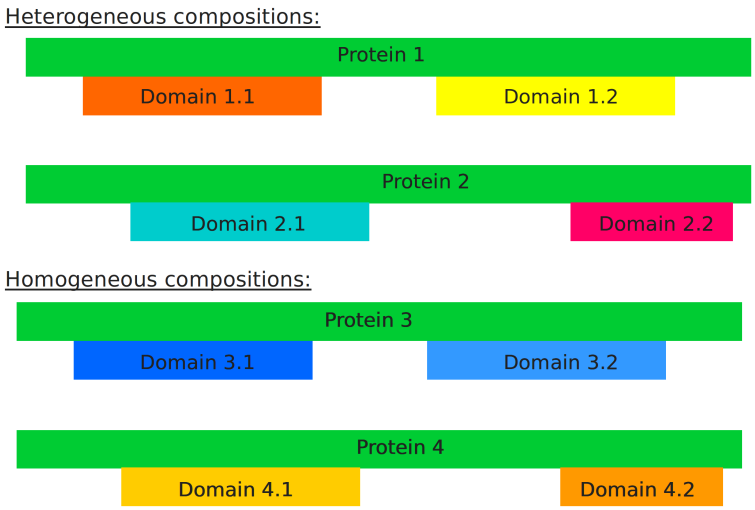

Figure 4.1: The concept of heterogeneous and homogeneous compositions. The amino acid composition of domains within the same protein can possibly be heterogeneous (indicated by domains of distinct colors in protein 1 and 2) or homogeneous (indicated by shades of blue and orange in protein 3 and 4). General tendencies in natural proteins and reasons for either observation are presented in this chapter.

P- and the L-model resembled each other closely (see Subsection 2.3.2), suggesting that within a protein sequence, there is no significantly heterogeneous use of amino acids. This finding appeared counter intuitive as domain recombination of different folds is a well-known mechanism [Apic et al., 2001; Chothia et al., 2003] and, as outlined in the above, folds can be characterized by their distinct amino acid compositions. Hence, without further investigation it was unclear how to interpret this observation of a homogeneous amino acid composition within entire protein sequences. This homogeneity within proteins has been associated to a short protein length as well as with the proportion of structured sequences in Subsection 2.3.3. However, it was not possible to exclude the fact that the amino acid compositions of domains within the same protein are for some other reason adjusted to each other, leading to the question if co-occurrence in the same protein constrains a domain's amino acid composition more than structural constraints.
Within the context of the genome, a protein is locally encoded and it is known that genomic regions demonstrate certain compositional biases due to shifts in the GC-content, pyrimidine/purine ratio, codon usage, strand-dependencies [Quax et al., 2015; Cebrat and Dudek, 1998; Plotkin and Kudla, 2011; Novoa and Ribas de Pouplana, 2012]. Also, the expression level of a protein is mediated by codon bias [dos Reis et al., 2003]. These effects, detectable on the DNA-level, may thus have an influence that leads to the observation of an intra-protein homogeneity of amino acid composition.

4.1.2 Content of this chapter

In this chapter, the amino acid compositions of domains within the same protein are analyzed, aiming to characterize and explain their observed homogeneity. These results are contrasted with the randomly expected homogeneity within proteins, with comparison of arbitrary domains across the genome, with comparisons of domains in neighboring proteins in genomic proximity and also with comparisons of domains of the same fold. For this, a selection of 12 genomes is used with representatives of the three kingdoms archaea, bacteria and eukaryotes. The results indicate that there is a significant correlation between the amino acid composition of domains within the same protein, which is more pronounced in archaea and bacteria and less in eukaryotes. This homogeneous amino acid usage of domains within a protein is more heterogeneous between arbitrary domains of the same genome, domains occurring in genomically adjacent proteins and domains of the same fold.
In order to investigate the contributions of codon usage to this effect, the compositional homogeneity of codons between domains of the same proteins is studied, which is found to be more pronounced than that of amino acids. By demonstrating the coupling between similar codon and amino acid usage, it was possible to directly associate a similar amino acid usage to a similar codon usage along whole protein chains. This finding is in line with the literature on the enhanced translation efficiency caused by codon redundancies [Quax et al., 2015; Novoa and Ribas de Pouplana, 2012]. It supports studies that demonstrate how amino acid composition is under the influence of translational effects [Lobry and Gautier, 1994] and, for the first time, contrasts constraints of translation and codon bias to those of protein structure.

4.2 Methods and Materials

4.2.1 Defining compositions

There are several ways to represent the amino acid composition of protein sequences. The most standard representation is a vector of length 20, where each entry represents the count of a specific amino acid a within a domain g:

$$C_{\text{abs}}(a) = |\{i \mid g(i) = a\}|. \tag{4.1}$$

This absolute count can be transformed into a relative composition C by normalizing it by the length of the sequence:

$$C(a) = \frac{C_{\text{abs}}(a)}{|g|}. \tag{4.2}$$

In the following, any definition of a relative composition C is referred to as a *composition vector*. This definition is also used in the context of codon composition, where the vector has a length of 61 entries for each amino acid coding codon.

Comparing compositions

Composition vectors can be compared using different metrics. The difference of composition vectors of length n can be derived by the Manhattan or Euclidean distance:

$$d_m(C_1,C_2) = \sum_{0\leq i<n} |C_1(i) - C_2(i)| \tag{4.3}$$

$$d_e(C_1,C_2) = \sqrt{\sum_{0\leq i<n} (C_1(i) - C_2(i))^2} \tag{4.4}$$

These metrics give more emphasis to common amino acids and less to uncommon amino acids such as cysteine and tryptothan. Uncommon amino acids may be especially important for a specific functional task. In order to balance the weight of amino acid specific differences in favor of less frequent amino acids, the differences can be scaled according to the general frequency f of amino acids.

$$d_m^f(C_1,C_2) = \sum_{0\leq i<n} \frac{1}{f(i)} |C_1(i) - C_2(i)| \tag{4.5}$$

$$d_e^f(C_1,C_2) = \sqrt{\sum_{0\leq i<n} \frac{1}{f(i)} (C_1(i) - C_2(i))^2} \tag{4.6}$$

A similar metric to d_e^f (see Equation 4.6) has been used in [Lobry and Gautier, 1994] for example. Here, I only use compositional distances according to Equation 4.3 and Equation 4.4. They are commonly referred to as the L1-norm and L2-norm in linear algebra. Other studies successfully use a metric based on the entropy of compositions [Ofran and Margalit, 2006], which however does not directly capture amino acid-specific differences.

Compositions and the finite sampling problem

When comparing compositions, it is crucial to acknowledge that short sequences are subjects to the finite sampling problem. Changing the composition slightly, by adding one amino acid for example, can change the previously determined composition majorly. Extending a sequence of length m to a length of $m+1$ by adding an amino acid b, will lead to a new composition vector C'

$$C'(a) = \begin{cases} ((C(a)\cdot m)+1)\cdot \frac{m}{m+1} & \text{if } a = b \\ C(a)\cdot \frac{m}{m+1} & \text{if } a \neq b \end{cases}$$

The smaller m, the greater is the difference between 1 and $\frac{m}{m+1}$, hence the more different is the new composition $C'(a)$ from the previous $C(a)$. In other words, while the over-

all composition of a large data set converges towards a constant value, compositions of small sequences fluctuate simply because of their size. Therefore, if two composition vectors of short sequences are very dissimilar (large d_m or d_e), this does not necessarily imply that this observation is significant. For this reason, I account for the randomly expected difference between two given natural sequences (see Subsection 4.2.3 and Subsection 4.2.4).

4.2.2 Natural sequence data sets

Data sets of 12 whole genomes (DNA sequences of predicted coding regions) were acquired from the National Center of Biotechnology Information. They were chosen according to their quality in the assembly and genomes with more proteins were preferred over smaller proteomes. Details of all used genomes are provided in Table 4.1. The size of the eukaryotic genomes was too large to investigate exhaustively and proteins were sampled randomly across the genome. One characteristic of the retrieved genomic data is that the sequences are ordered by their occurrence in the genome, allowing to compare neighboring proteins in the same genomic context. The results of neighboring proteins was not performed for eukaryotes due to a random sampling of proteins. The DNA data was translated into protein sequences and terminating non-triplets in a sequence were discard together with the starting methionine.

For each genome, domains were assigned to the analyzed protein sequences. The HH-suite was used for this assignment and the ECOD database was used as reference for domains. A detail description of this approach is provided in Subsection 1.2.4).
For a consistent data set, proteins with less than 80% of all their residues being assigned to a domain were discard. With this procedure, effects coming from potentially unstructured parts or unassigned domains were presumably removed. In the primary study, only double-domain proteins were considered, in order to compare proteins with the same domain topology. In Subsection 4.3.6 proteins of different topologies were analyzed, in order to test if topology plays a role in the compositional differences between domains.

4.2.3 Random sequences with the composition of natural domains

Previously, natural peptides of length 100 were used to account for the local amino acid composition of natural proteins and random sequences were generated by permutating residues within these peptides (L-model). Thereby, real domain boundaries were neglected and unstructured sequences, connector sequences between domains and partial domains were considered as reference.
In this study, a variation of the D-model is used, which accounts for the composition of real, structured domain sequences. From a technical perspective, the D-model is similar to the P-model, which is based on amino acid sequences shuffled in the context of proteins, thereby preserving the composition of natural proteins. As query sequences

Table 4.1: Genomes used to study composition of domains. Three archaeal genomes were used, five bacterial genomes and four eukaryotic genomes. The eukaryotic genomes contain splicing alternatives and were randomly sampled. Of the analyzed proteins, the number of double-domain proteins is indicated. In cases where more domains were analyzed, the number of proteins with 3 or 4 assigned domains and that of proteins with more than 4 domains is given.

organism	predicted cds in genome	analyzed proteins	double-domain proteins	3/4 domains	>4 domains
Haloterrigena turkmenica	5,167	all	470	384	79
Methanobacterium veterum	3,208	all	325	216	N/A
Saccharolobus solfataricus	3,204	all	314	230	N/A
Escherichia coli	4,357	all	672	520	168
Granulicella mallensis	4,704	all	454	409	N/A
Bacillus simplex	5,195	all	686	449	N/A
Amycolatopsis mediterranei	9,228	all	1,311	831	N/A
Faecalibacterium prausnitzii	2,956	all	342	N/A	N/A
Homo sapiens	119,511	11,659	478	499	299
Orcinus orca	27,925	14,372	753	793	N/A
Solanum tuberosum	37,966	7,849	420	296	N/A
Arabidopsis thaliana	48,265	5,290	262	195	N/A

are chosen to correspond to domain instead of protein sequences, shuffling within their individual boundaries leads effectively to sequences that possess the same composition of natural domains. In order to compare two domains to each other under the consideration of their homogeneity, this procedure was slightly altered by permutating residues between these domains, thereby preserving the combined composition, further referred to as the D_2-model. For this, amino acids occurring in the query domain sequences g_1 and g_2 are assigned to arbitrary positions in the random sequences of equal lengths, as illustrated in Figure 4.2. Following properties hold for these random sequences:

$$|g_1| = |r_1| \wedge |g_2| = |r_2|$$

$$C_{abs}(g_1) + C_{abs}(g_2) = C_{abs}(r_1) + C_{abs}(r_2)$$

The derived composition vectors of these random sequences can be used to calculate the compositional difference that is expected from the combined composition of the given domain sequences, independent of the sequential order. The greater the compositional difference between the natural domains compared to that of the random sequences of the D_2-model, the more suggestive is the heterogeneous use of amino acids across the query domains (see Figure 4.1).
Given the outlined effects of the finite sampling problem (see Section 4.2.1), the variance

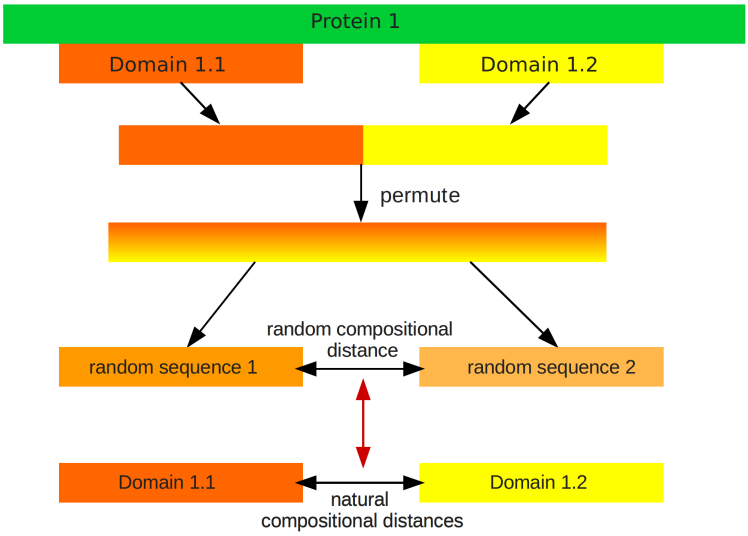

Figure 4.2: Generating permutated sequences of the D_2-model. The composition of two combined sequences, which are generally belonging to domains, is used as background composition to sample random sequences of the same lengths as the given query sequences. The compositional distance between sequences of the D_2-model is then compared to that of the natural sequences.

of the compositional difference between instances of randomly generated sequences r_1 and r_2 will vary substantially. Generating one set of random sequences for a given query and comparing its compositional distance to that between the natural sequences is hence not enough to estimate if the naturally observed compositions are particularly similar or dissimilar compared to an expected value.

4.2.4 Significance of a compositional differences

When deriving a distribution of compositional distances by generating multiple (1000 throughout this study) random sequence pairs for each pair of domains, it is possible to estimate the significance of a particular distance between natural compositions. The probability density function (PDF) summarizing the distribution of randomly expected compositional differences is in the following referred to as $R(d)$.

$$R(d) = PDF(d(C(r_1), C(r_2))) \tag{4.7}$$

It is depicted in Figure 4.3 for an example sequence of *Escherichia coli*. Note, that 90% of the compositional distances (according to the L2-norm) range between 0.08 and 0.14, reflecting a great variation among the expected distances. Depending on the distance between the natural compositions $d(C(g_1), C(g_2))$ in this distribution, the percentile rank PR can be derived as a measurement of how significant the natural distance is relative to

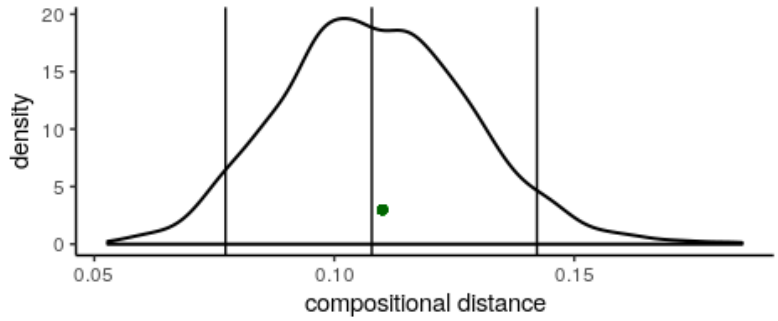

Figure 4.3: Distribution of compositional distances of the D_2-model. The used sequence to derive this distribution of compositional distances $R(d)$ is the third protein in the *Escherichia coli* genome. Random sequence samples of the combined composition of the two domains in this protein lead to a distribution of the euclidean distances between the derived compositions ranging between 0.05 and 0.18. The vertical bars indicate the 5th, 50th and 95th percentile, corresponding to a distance of 0.08, 0.11, and 0.14. The compositional distance derived from the natural domains is 0.11 (indicated by green circle), which is slightly more dissimilar than the median compositional distances among the permutated sequences.

the expected value.

$$PR(g) = \sum_i R(i) \cdot 100 : i < d(C(g_1), C(g_2)) \tag{4.8}$$

This percentile rank indicates the percentage of randomly simulated distances that are smaller than the observed natural distance. For a significance level of $\alpha = 0.05$ a percentile rank of ≤ 5 indicates a significantly similar composition, while a percentile rank of ≥ 95 indicates a significantly dissimilar composition. In Figure 4.3 the natural distance is 0.11, represented by a green circle, which is located to the right of the median expected compositional distance. The natural compositions are thus slightly more dissimilar relative to the randomly expected distances, an observation that is not significant. In the case of a pair of random sequences, i.e., with no internal sequence or composition correlations, the percentile rank *PR* will be distributed according to the generated distribution $R(d)$. In the following, this distribution is referred to as the *percentile rank distribution* (*PRD*), which is a central function in this chapter. It is used as the primary method to compare compositional differences on a broad scale with each other. By the definition of the percentile rank, the *PRD* is an equal distribution for random sequences and only fluctuates due to finite sampling. Note, that thus 5% of the sampled random distances are significantly similar and 5% are significantly dissimilar for a significance level of $\alpha = 0.05$. When investigating the *PRD* over a set of natural sequences, the percentile rank may be distributed differently from those expected by random sequences, which follow an equal distribution. It is derived as the probability density function of

natural percentile rank *PR* according to Equation 4.8.

$$PRD(i) = PDF(PR_{\text{nat}}) \tag{4.9}$$

The median of a percentile rank distribution indicates a general tendency towards more similar or more dissimilar compositions relative to the random sequences of the combined composition. In order to derive if two percentile rank distributions differ significantly, a standard statistical test described in the following is used.

Significance of compositional distances in two sets

To test if two sets of sequence pairs differ in their compositional distances the *two-sample Kolmogorov–Smirnov test* is applied. This test is used also in another study of compositional differences to assess their significance relative to random differences [Ofran and Margalit, 2006]. Therein, compositional differences are assessed by entropy differences, which differs from this metric, as it does not include amino acid specific differences.
This statistical test requires to represent the sampled sets s as empirical distribution functions F.

$$F_s(x) = \frac{1}{|s|} \sum_{s_i \in s : s_i \leq x} 1$$

The Kolmogorov–Smirnov statistic (TKS) is defined based on the two derived empirical distribution function of a set u and v as:

$$TKS(u,v) = \arg\max_x |F_u(x) - F_v(x)|$$

The null hypothesis, that both samples come from the same distribution, is rejected for a significance level α if :

$$TKS(u,v) > \sqrt{-\frac{1}{2}\ln(\alpha)} \cdot \sqrt{\frac{|u|+|v|}{|u|\cdot|v|}} = T$$

In the presented plots, the statistic TKS and the minimal α for which the hypothesis can be accepted are depicted. The values that α can assume are defined as $\{10^s \mid s \in \{[-10,-1] \cap \mathbb{Z}\}\} \cup \{[0,100] \cap \mathbb{N}\}$. This representation was chosen to give a better understanding on how distinct two distributions are, as α represents the amount of *risk* for having falsely rejected the null hypothesis. A large α therefore indicates that the distributions are quite similar, while a small α indicates a deviation between the distributions. Typically α is chosen before performing the test, leading to a binary assessment of the two samples to be derived from the same distribution. It is generally set to 0.01 or 0.05 to minimize the risk of a false rejection of the null hypothesis. In line with this common approach, two sampled sets of *PR* are referred to as likely to be from the same distributions if $\alpha < 0.05$ and unlikely to be from the same distribution if $\alpha \geq 0.05$.

4.3 Composition of domains and proteins

4.3.1 Composition of domains from the same protein

The homogeneity of the amino acid compositions of domains from the same protein was established as a first step in this study. To analyze this aspect on a broad scale, 12 data sets of distinct genomes were derived as described in Subsection 4.2.2. For each data set double-domain proteins with a domain coverage above 80% are used to avoid artifacts arriving from multi-domain arrangements and to use consistently structured proteins. Results on other topologies are provided in Subsection 4.3.6. Throughout this study, the Euclidean distance (L2-norm) is primarily used to compare compositions (see Equation 4.4). For comparison, it is contrasted to the results derived when using the Manhattan distance (L1-norm). Only in Subsection 4.4.3 the Manhattan distance was used as primary metric, which is further discussed therein.
In this section, the results of *Haloterrigena turkmenica*, *Escherichia coli*, and *Homo sapiens* are presented and discussed, as representatives for archaea, bacteria and eukaryotes. The compositional distance between domains within the same protein were derived along with their percentile rank PR_{within}, which indicates the significance of this distance according to the randomly expected value of sequences with a composition of the combined domains (see Subsection 4.2.4). This measure allows to estimate if the derived distance is significantly small ($PR \leq 5$) or large ($PR \geq 95$). In order to obtain a general view over many natural sequences, the percentile rank distribution PRD_{within} over all compositional distances of double-domain proteins was derived (Figure 4.4: red).

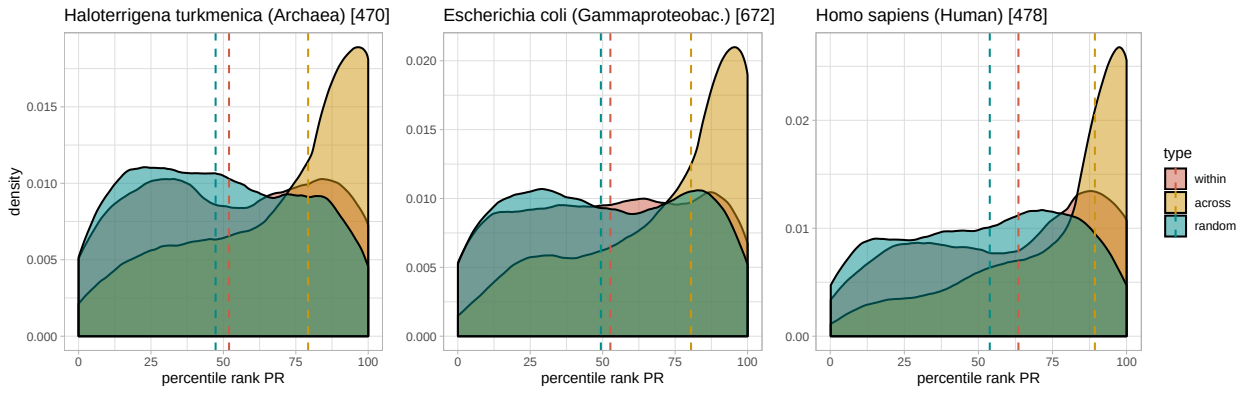

Figure 4.4: Distributions of compositional differences. The percentile rank distributions indicate how the significance of compositional similarity is distributed relative to the expected distances of the D_2-model. A low PR indicates an enhanced compositional homogeneity, a high PR an enhanced heterogeneity. The median PR (dashed line) can indicate the tendency towards either side. Three types of comparisons are contrasted: comparisons of domains within the same double-domain protein (red), comparison of arbitrary domains across proteins (yellow) and comparisons of the D_2-model corresponding to the within-protein comparisons.

The median percentile rank ranges between 51 and 71 (see Table 4.2) for all genomes, which already indicates a slightly increased heterogeneity in the amino acid composition of domains within the same protein relative to their combined composition. This tendency was further investigated by conducting the two-sample Kolmogorov–Smirnov test (see Section 4.2.4) to derive the significance of this finding. For this, PRD_{within} was compared to a percentile rank distribution of the D_2-model, further referred to as PRD_{random} (Figure 4.4: blue). In the case of *Haloterrigena turkmenica*, the TKS statistic is 0.087, which is larger than T for $\alpha = 0.03$ (Figure 4.5: upper panel, left). With a risk of 3%, as indicated by α, it is not possible to reject the hypothesis that the percentile ranks of the compositional distances within a protein differ from those of their combined composition. Thus, the compositions cannot be assumed to be completely homogeneous in

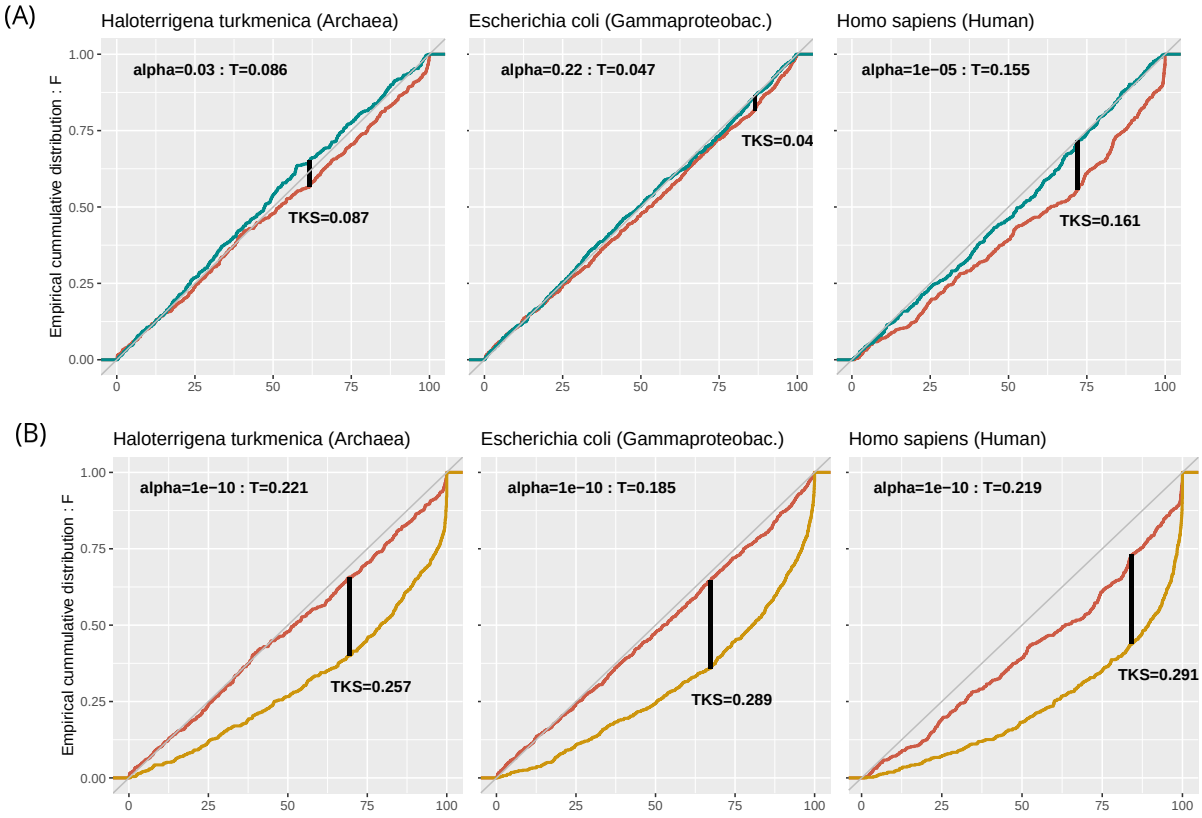

Figure 4.5: Significance tests of distinct percentile rank distribuitons. (A) The PRD of the within-protein comparisons (red) in Figure 4.4 are tested for being from the same distribution as expected from sequences with their combined composition as defined by the D_2-model (random). An acceptance of the null hypothesis ($\alpha \leq 0.05$) indicates a homogeneous composition of domains from the same protein that is indistinguishable from the random model. (B) Test for being from same distribution of within-protein (red) and across-protein (blue) comparisons. In all three cases, there is almost no risk to assume the values are sampled from two different distriutions.

domains from the same protein in this genome. As layed out in Section 4.2.4, I choose $\alpha = 0.05$ as threshold to decide on significance. Larger values of α result into a too high risk to conclude that the two samples are from different distributions, leading to the conclusion that they are from the same. Smaller values allow to accept the hypothesis that the two samples are from different distributions. Using the Manhattan distance to compare the compositions, leads to a higher risk of 17% (see Table 4.2) and would lead to the conclusion of significantly similar compositions. The risk is larger in the case of the *Escherichia coli* genome, assuming a value of 22% (Figure 4.5: A). This implies that the amino acid composition of double-domain proteins in *Escherichia coli* is very homogeneous. Similar results were achieved for all other analyzed genomes from archaea and bacteria, as presented in Table 4.2.
The results differ from those of *Homo sapiens*, where the median percentile rank is increased to 71. The hypothesis of the compositional distances being from the same distribution as that of permutated sequences can be rejected at a significance level of $\alpha = 0.00001$, with a risk of only 0.001% to have falsely rejected the hypothesis (Figure 4.5: A). Thus, within the human genome, the amino acid composition of double-domain proteins is significantly heterogeneous compared to the combined composition of the respective domains. This tendency could be confirmed for all other used eukaryotic genomes.
In summary, these results confirm the observation in Subsection 2.3.2, that the composition of domains within one protein is generally significantly homogeneous in bacteria. The relationship between this and the previous work is further discussed in Section 4.5.1. In some cases, this homogeneity cannot be disassociated from the homogeneity expected from the combined composition of the two domains. Domain recombination may thus be biased towards combinations of similar amino acid compositions. A random recombination would lead to arbitrary combinations of domains. In the following, I analyze whether a random recombination of natural domains leads to the observed compositional diversity among natural proteins.

organism	proteins	GC content	median PR within (L2)	median PR across (L2)	risk (L2)	median PR within (L1)	median PR across (L1)	risk (L1)
Haloterrigena turkmenica	470	66%	52	79	3%	52	81	**13%**
Methanobacterium veterum	325	36%	57	79	1%	57	78	**7%**
Saccharolobus solfataricus	314	36%	51	82	**28%**	50	80	**59%**
Escherichia coli	672	52%	53	80	**22%**	53	81	2%
Granulicella mallensis	454	61%	54	85	3%	55	84	**5%**
Bacillus simplex	686	41%	52	77	**10%**	51	76	2%
Amycolatopsis mediterranei	1311	72%	54	80	1%	56	81	0.1%
Faecalibacterium prausnitzii	342	59%	52	82	**17%**	52	83	**19%**
Homo sapiens	478	50%	64	89	0%	70	89	0%
Orcinus orca	753	52%	71	89	0%	63	89	0.01%
Solanum tuberosum	420	42%	68	89	0%	65	90	0%
Arabidopsis thaliana	262	45%	66	90	0.1%	64	91	0.01%

Table 4.2: Comparing amino acid compositions of domains within and across double-domain proteins with Euclidean distance (L2) and Manhattan distance (L1). The number of used proteins is indicated together with their GC-content. The median percentile rank of the within comparisons is lower in archarea and bacteria than in eukaryotes. In the last column, the risk of falsely rejecting the hypothesis of the within and random (D_2-model) comparisons are from the same distribution is indicated. In cases where the compositional divergence of domains within proteins is not distinguishable from their combined composition, the risk is highlighted.
Same data using Manhattan distance to compare amino acid compositions. The median percentile ranks are comparable to those derived using the Euclidean distance. The risks are most affected from a change in the distance metric. A trend of the compositions of domains in archaeal proteins to be more homogeneous compared to bacteria, who are more homogeneous than in eukaryotes becomes apparent.

4.3.2 Heterogeneous composition of arbitrary domain recombinations

Under the assumption of random recombination, domains from distinct proteins are investigated for their compositional distances and their percentile rank PR_{across}. For this, arbitrary pairs of domains from distinct double-domain proteins are derived along with their percentile rank distribution PRD_{across}. The median percentile rank is 77-90, depending on the chosen genome (see Table 4.2) and, therefore, higher compared to the median percentile rank of compositions within the same protein, as presented in Subsection 4.3.1. The two-sample Kolmogorov–Smirnov test was applied to compare these sets of percentile ranks. For all genomes the null hypothesis that the percentile ranks of within- and across-protein comparisons come from the same distribution can be rejected for a significance level below 10^{-6}, implying a negligible amount of risk for this rejection (Figure 4.5 : B). Thus, random domain recombination leads to the combination of significantly distinct compositions, which is more pronounced than that of observed domain recombinations.
In conclusion, there is a tendency of a compositional homogeneity between naturally recombined domains relative to a random recombination of domains from different proteins of the same genome. In the following, I will use the term *harmonization* to refer to this increased compositional similarity between the compositions of domains within the same protein relative to a random domain recombination. Note, that the term harmonization is typically used to describe the biased usage of synonymous codon, which has a decreased entropy [Mignon et al., 2018; Fisher et al., 2011]. There may be different reasons for the harmonization of amino acid usage, which are outlined in the following.

Structural bias Recombination may not be random but favor domains of similar compositions due to some unknown constraints. A compositional comparison across all used domains may thus not reflect natural recombinations and selective recombination of domains with similar structures may cause the observed harmonization. Therefore, the compositions of domains within the same protein were contrasted with those of domains from distinct proteins with the same folds. The analysis of fold-specific compositional correlations is presented in Subsection 4.3.3.

Genomic context Transposition and duplication events are known to occur more often between genomic areas in local proximity. Given that the GC-content influences the amino acid composition, a locally similar amino acid composition may be caused by similar GC-contents. The correlation between genomic context and compositional similarity is analyzed by two approaches: (1) compositional differences of domains from distinct proteins that are adjacent to each other in the genomic context, further referred to as 'neighbor' comparison (2) correlation of the GC-content between domains in adjacent proteins. The analyses are presented in Subsection 4.3.4.

Natural assimilation over time In a scenario where domain recombination was random and not dependent on the initial composition of the domains, the remaining most obvious conclusion would be that compositions of domains in the same protein become more similar to each other over time. This assumption suggests that an evolutionary pressure exists that acts on the whole protein and overwrites the evolutionary pressures to composition that individual domains are exposed to.
From a functional point of view, the benefits of a homogeneous composition may be due to the fact that a protein (an exception are for example transmembrane domains) lives in one particular part of the cell, thereby being exposed to the same environment. It has been reported before that the cellular location affects the amino acid composition [Cedano et al., 1997]. This hypothesis has been pursuit without positive results and is not presented here.

On the protein level, irrespective of protein structure and function, there is a well-known bias related to the DNA level. In the translation process of mRNA to amino acid chains, the codon composition plays a major role [dos Reis et al., 2004]. Foremost, translation efficiency and its relationship to codon usage has previously been studied to an exhaustive extend. Depending on the genome, number of tRNA genes and the expression level, different codons and combinations of codons are used for different protein sequences. This influence is so strong, that codon usage not only adapts DNA to the amino acid sequence, as defined by structural and functional constraints on the protein level, but also actively changes the amino acid sequence itself [Quax et al., 2015; Plotkin and Kudla, 2011; Lobry and Gautier, 1994]. To check for the effects of codon usage to the observed amino acid usage, I performed the same study of compositional differences in codon usage, thereby translating the DNA into a composition vector of length 61, corresponding to all amino acid encoding codons. The analyses are presented in Subsection 4.3.5 and considered in the larger context of general codon bias research in Section 4.4.

4.3.3 Fold-specific compositions

To account for fold class differences, domains are grouped by their fold class, corresponding to approximately 2300 possible X-groups defined by ECOD. The 12 most abundant fold-combinations observed in double-domain proteins were analyzed. For each of these fold-combinations, 1000 domain combinations of the same fold from distinct proteins are collected. In Figure 4.6 each plot refers to a X-fold combination as indicated in the respective title. The green number indicates how often the specific combination occurs within the assayed double-domain proteins. The percentile rank distribution of comparisons within and across proteins are depicted as box-plots. In all presented cases, the median percentile rank of the comparisons within proteins is lower than that of the comparison across proteins. This observation is true for all investigated genomes and applies to most fold-combinations. Against this general tendency towards harmonization within proteins, there are proteins that are counterexamples, of which one is presented

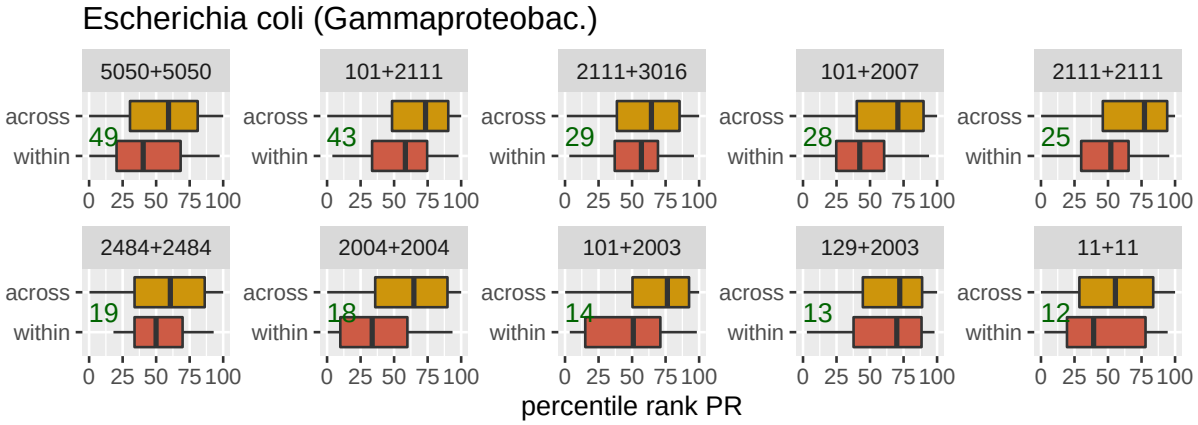

Figure 4.6: Fold-specific comparison of compositions. The ten most abundant fold recombinations in double-domain proteins are investigated. The recombined fold classes as defined by the X-group in the ECOD hierarchy are indicated in the header of the individual plots. Irrespective of the recombined folds, compositions within the same protein are more similar than the same folds of arbitrary proteins.

and discussed in Section 4.3.6
Comparing folds of the same X-group within the same protein is a biased comparison, given that this recombination has probably occurred due to a duplication event. Thus, the compositions are likely to be similar, which is here presented for X-groups 5050, 2111, 2484 and 11.
The specific combination of folds is thus generally not associated to the similar compositions of recombined domains. Structural constraints of naturally recombined domains are therefore not responsible for the observed harmonization.

4.3.4 Correlations to adjacent proteins in the genomic context

The harmonization within proteins may be related to genomic regions, leading to the hypothesis, that neighboring proteins may also be harmonized. To check if domains in the same genomic regions possess similar compositions, the domains in each double-domain protein were compared to those in the consecutive protein. In this case, not only domains of well-structured double-domain proteins were considered. This was necessary as not all double-domain proteins have a double-domain protein as genomic neighbor, leading to a small set of comparisons.
Due to technical reasons as laid out in Subsection 4.2.2, the genomic context in eukaryotic genomes was not preserved. Results concerning the genomic neighborhood are only derived for archaea and bacteria.
As a first step, the correlation of the GC-content in the genomic neighborhood is studied.

In Figure 4.7 the GC-content of domains in adjacent proteins is moderately correlated (purple dots). In *Haloterrigena turkmenica*, the Pearson Correlation Coefficient (PCC) is 0.35, in *Escherichia coli* it is 0.4. All other genomes demonstrate the same tendency. This correlation is weaker than that from correlating the GC-content of domains in the

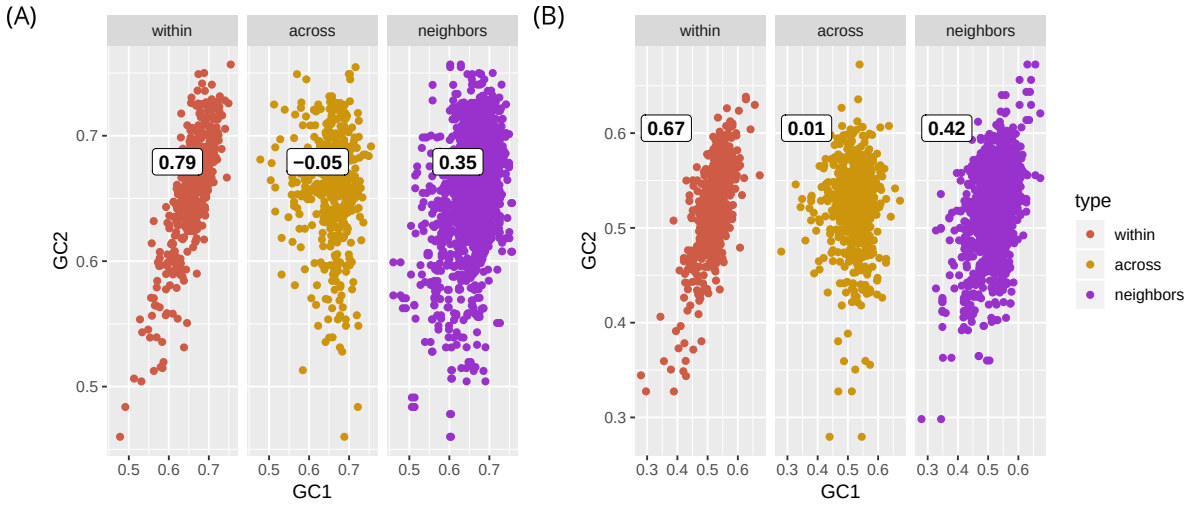

Figure 4.7: Correlations of GC-content in (A) *Haloterrigena turkmenica* and (B) *Escherichia coli*. The GC-content of domains within the same protein correlates strongly (red). Arcross arbitrary proteins, the GC-content does not correlate (yellow). Domains of neighboring proteins in the genomic context also correlate but to a weaker extent (purple).

same protein (Figure 4.7: red). The PCCs are 0.79 in *Haloterrigena turkmenica* and 0.67 in *Escherichia coli*. Between arbitrary proteins, the PCC ranges around 0.
This is in accordance with studies reporting this correlation of the GC-content in genomic proximity and fluctuations across the genome [Karlin et al., 1998]. Although compared to the correlations of the GC-content of domains within the same protein the GC-content in neighboring proteins is less correlated, it can possibly result into a harmonization.
In Figure 4.8 the percentile rank distribution of the compositions of domains in neighboring proteins is depicted in purple. This distribution resembles the distribution of comparisons across proteins (yellow) with a slightly decreased median percentile rank. It is very different from the comparisons within proteins and the two-sample Kolmogorov–Smirnov test confirms that there is almost no risk (below $\alpha = 10^{-6}$ for all genomes) in assuming that these percentile rank distributions are from distinct distributions. Thus, the amino acid composition of a genomic context is not homogeneous and is not from the same distribution as the comparison within proteins.
Concerning the comparisons across proteins, the two-sample Kolmogorov–Smirnov test leads to a higher minimal risk of $\alpha = 0.1$ (*Haloterrigena turkmenica*) and $\alpha = 0.01$ (*Escherichia coli*). These values are below 0.05, leading to the conclusion that the compositional comparisons of domains across arbitrary and of neighboring proteins are not

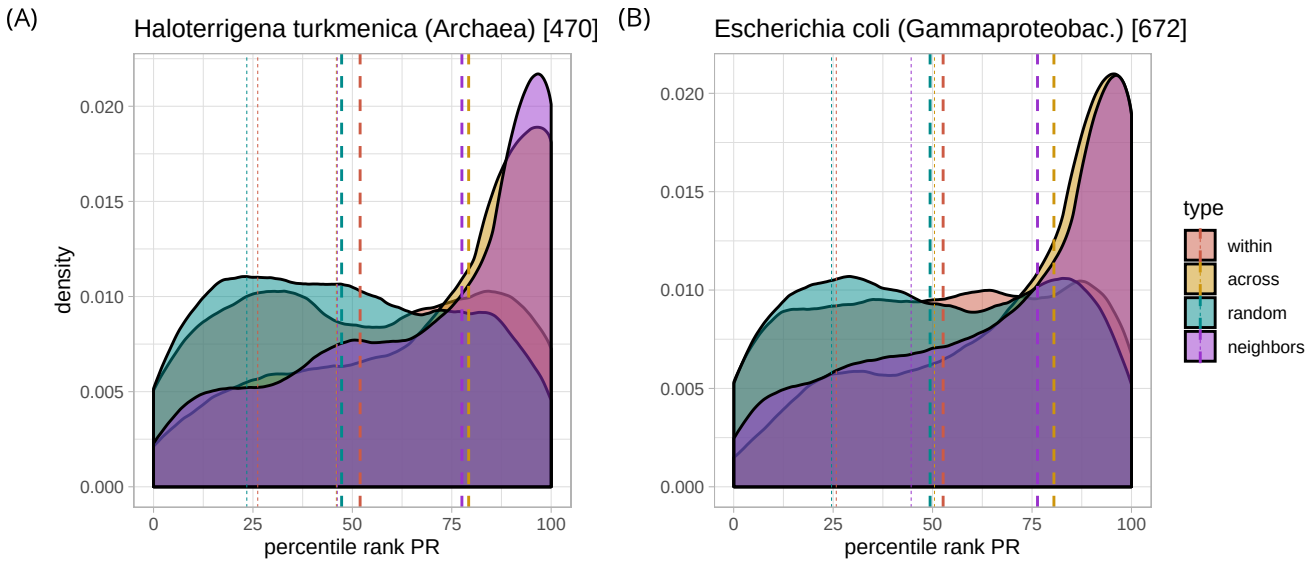

Figure 4.8: Compositional distances of domains in adjacent proteins in the genomic context of (A) *Haloterrigena turkmenica* and (B) *Escherichia coli*. The percentile rank distribution of the comparison of domains in adjacent proteins (purple) resembles the results of the across-protein comparisons. It is very distrinct from the within-protein comparisons.

the same but similar. Given that the GC-content of neighboring proteins is correlated and uncorrelated across arbitrary proteins, this minor deviation may be an artifact arriving from the local GC-content.
The analysis of other genomes lead to similar results (data not shown). With this, it becomes clear that the harmonization indicates an evolutionary pressure that is specific for each individual protein and does not correlate to genomic context in archaea and bacteria.

4.3.5 Similar codon usage of domains in the same protein

A well-known evolutionary pressure that acts on individual proteins is associated to the DNA level. It is caused by constraints of the transcription and translation process and can be analyzed on the codon level.
Codon bias has been studied extensively and has been related to translation efficiency, expression level, mRNA structure, tRNA abundance and many other complex constraints. Given this complexity, and the fact that codon bias acts on entire protein sequences, it has mostly been studied in the context of whole protein sequences. The harmonization of codons within a protein is here studied with respect to domain boundaries and their structures, allowing to contrast the constraints on codon bias and protein structure. For this, the similarities codon compositions of domains from the same protein are investigated and evaluated if it is coupled to the harmonization on the amino acid level.

Correlations between codon and amino acid compositional differences

To investigate similarity of codon composition, I performed the same analysis as presented for the amino acid composition on compositional vectors of length 61, for all amino acid coding codons. In Figure 4.9, the correlation between distances of the amino

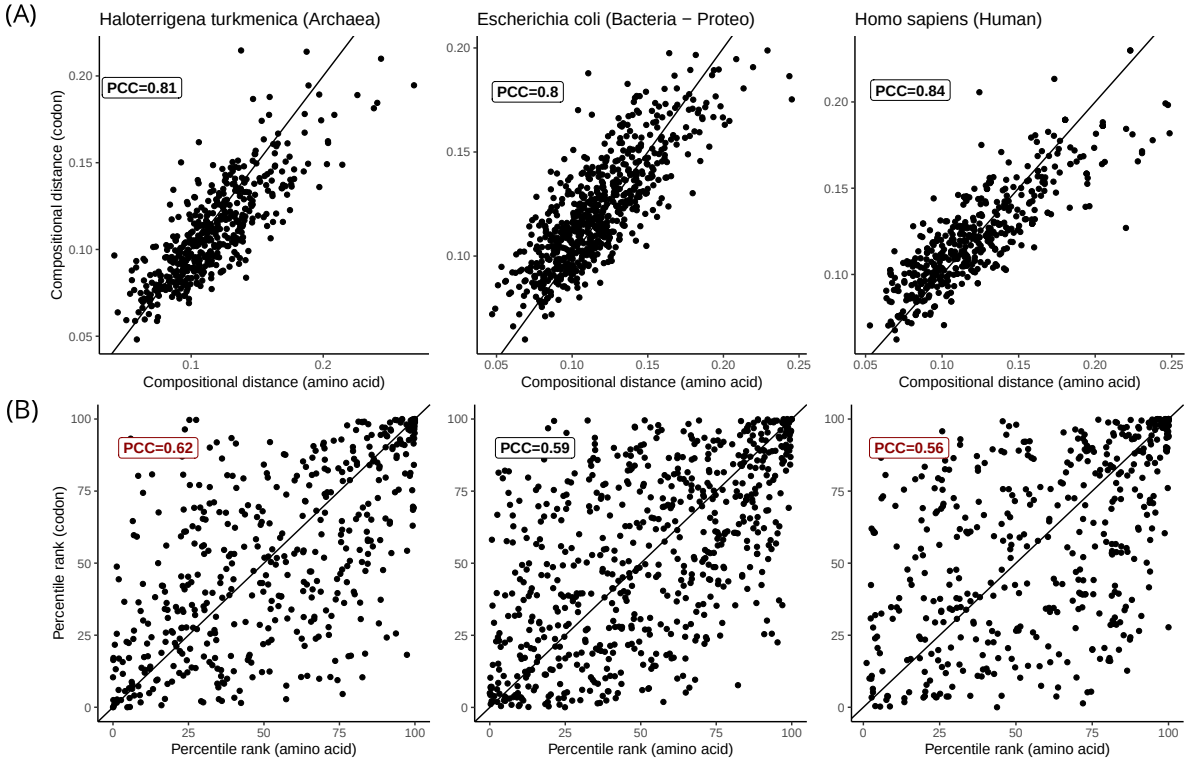

Figure 4.9: Correlations between amino acid and codon harmonization of *Haloterrigena turkmenica*, *Escherichia coli* and *Homo sapiens*. (A) distances between amino acid compositions of domains within the same protein (x-axis) are correlated to those derived from the respective codon compositions (y-axis). They correlate well with a PCC around 0.8. (B) the percentile rank, as a metric of the significance of these compositional distances, are also correlated with a PCC around 0.6.

acid and codon composition vectors is depicted for domains within the same protein. These distances correlate strongly with a PCC ranging around 0.8 for all three example genomes. This correlation was expected as codons are always translated into a definite amino acid.
However, the degeneracy of the genetic code allows for flexibility for homogeneous amino acid usage but heterogeneous codon usage. This would be the case if one codon of a specific amino acid is used mostly in one domain and another codon for the same amino acid in the other domain. A finding like this would indicate that amino acid composition harmonizes rather than codon composition and would indicate a directionality

of this correlation. From the obtained correlation between distances such a trend is not apparent, given that there is no accumulation of proteins in the upper left corner.

Percentile rank distribution of codon composition

To further investigate the significance of these distances, the percentile rank is derived and also correlated between comparisons of amino acid and codon compositions, resulting in a PCC ranging around 0.6 for all three example genomes (Figure 4.9: A). This

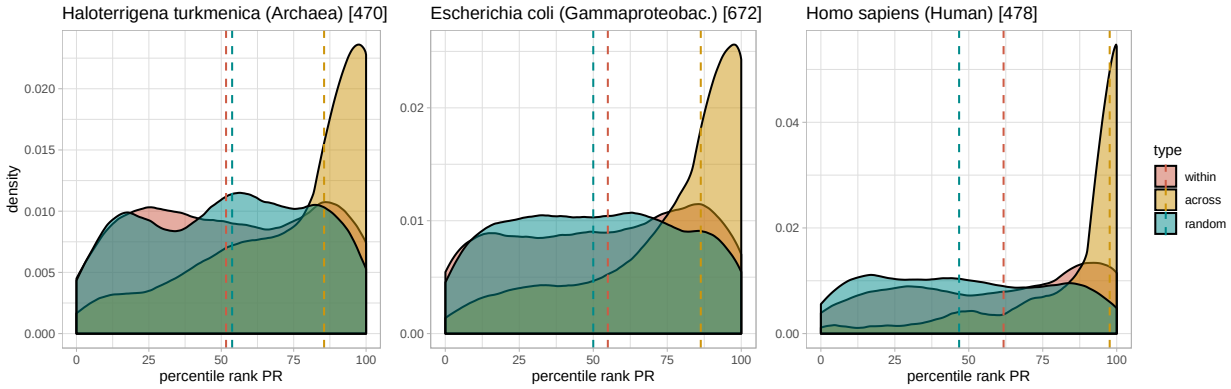

Figure 4.10: Percentile rank distributions of codon composition. Analogous to the percentile rank distributions of amino acid compositions in Figure 4.4, here, the distributions of the same data are derived for codon composition. The median percentile ranks of the within-protein comparisons resemble those derived from the amino acid composition. In the across-protein comparison, the median is relatively increased, especially in the case of the human genome. Codon usage is thus more heterogeneous across proteins than amino acid composition relative to their respective random model.

coefficient is smaller than that obtained from distances, which may be related to the fact that significance scales differently for vectors with different lengths. A trend of heterogeneous codon but homogeneous amino acid composition does also not become apparent from these correlations.
The percentile rank distributions derived from codon composition are plotted in Figure 4.10. The median for the within-protein comparisons and the random D_2-model is almost the same in both amino acid and codon comparisons (compare to Figure 4.4), implying that both amino acids and codons are almost equally harmonized relative to their random model. In contrast, the median in the comparison across proteins increases for codon composition. This indicates that amino acid usage is more homogeneous compared to codon usage between arbitrary proteins. There is thus a general constraints on proteins to possess a certain amino acid composition, that can be realized by a more diverse codon usage. This observation is in line with the fact that proteins experience structural pressure to possess a certain amino acid composition, that due to the degener-

acy of the code can be assumed by distinct codons.
This, together with the correlation studies, suggests that the observed similarity between amino acid compositions of domains in the same protein is coupled to a similar composition of codon usage. Given that harmonization has not been pin-pointed to specific amino acids and their respective codons, this hypothesis of a coupling needs further investigation and an approach to study the amino-acid specific coupling of harmonization is presented in Subsection 4.4.3.

4.3.6 Multi-domain topologies

In order to contrast the results derived from strictly double-domain proteins to those with more domains, the same analysis was performed for proteins with 3 or 4 domains and at least 80% structure content. Using this kind of data, there are 3 or 6 domain combinations within each protein that are assessed. Proteins with more than 4 domains were excluded to ensure a similar topology among all used proteins. The results are

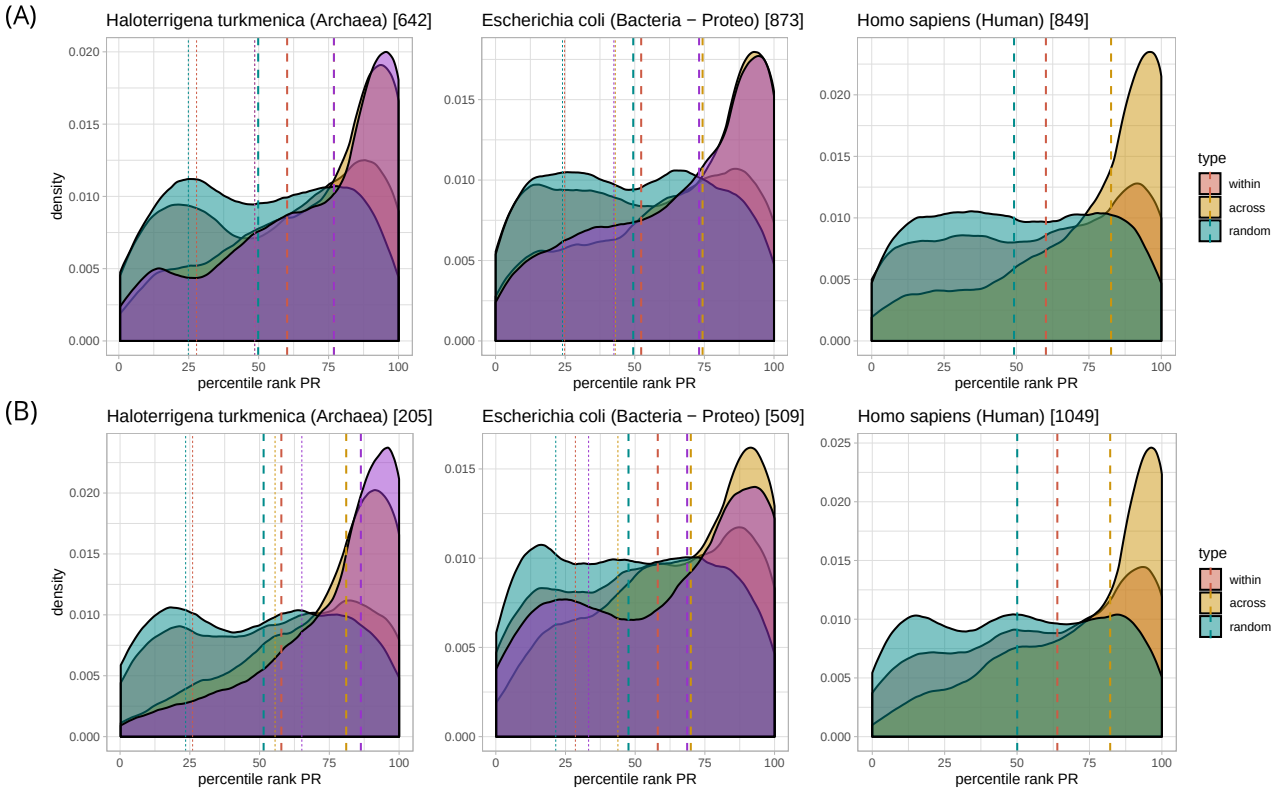

Figure 4.11: Percentile rank distributions of proteins composed of more than 2 domains. (A) comparisons of the amino acid composition of proteins with 3 or 4 domains. (B) comparison of amino acid composition of domains in proteins with more than 4 domains. Analogous to the percentile rank distributions of double-domain proteins in Figure 4.4.

closely related to those derived from double-domain proteins: (a) The median percentile rank of comparisons within proteins is ranging around 50 for *Haloterrigena turkmenica* and *Escherichia coli* and is increased in *Homo sapiens* (see Figure 4.11: A, red). Similar results were derived for the codon composition (data not shown). (b) In the comparison across proteins (Figure 4.11: A, yellow) the median percentile rank is increased relative to the comparisons within proteins. It is even more increased for codon composition (data not shown). (c) The correlation between distances and percentile rank as presented in Section 4.3.5 are closely comparable to the results of double-domain proteins (data not shown). (d) Comparisons in the genomic neighborhood (Figure 4.11: A, purple) are also more similar to the across-protein than to the within-protein comparisons. (e) Fold-specific comparisons also demonstrate a generally lower percentile rank for within than across comparisons (Figure 4.12).

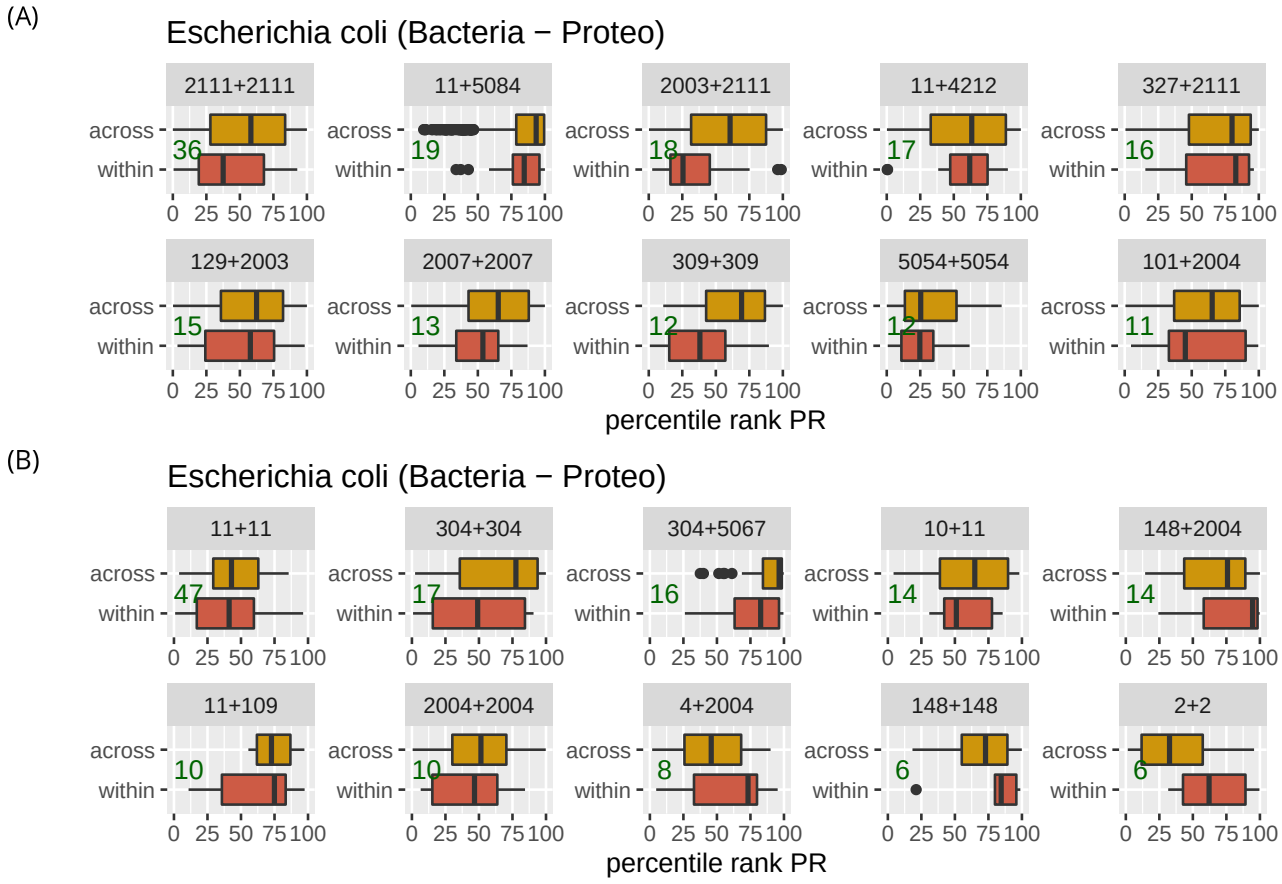

Figure 4.12: Fold-specific recombinations in multi-domain proteins on amino acid level. (A) comparisons of amino acid composition of domains of proteins with 3 or 4 domains (B) of proteins with more than 4 domains. These plots are analogous to the fold-specific recombinations in Figure 4.6.

Counterexample demonstrating compositional divergence within a protein

In order to analyze all proteins, those with more than 4 domains were investigated in a final batch (see Figure 4.11: B). The general tendencies (a-d) also apply to these multi-domain proteins (data not shown). However, among the most frequent X-fold recombinations, 4 cases of fold combinations occurred (148+2004, 4+2004, 148+148 and 2+2), that demonstrate a compositional heterogeneity when co-occurring in the same proteins relative to occurrences across different proteins (Figure 4.12 : B). The X-fold recombination 2+2 was further investigated, which corresponds to OB-fold recombinations.
All 6 comparisons between X-group 2 map back to the same protein, which is a Ribonuclease (NCBI Reference id: WP_061349297.1), composed of 4 OB-folds and an HTH-fold. The dissimilarity of the composition among these OB-folds was further detectable at the codon level. This finding suggests that the general tendency of harmonization does not apply to this protein. It is reasonable to assume that the OB-folds experience an evolutionary pressure to diversify, which may be binding of distinct nucleotides for example. Even though frequent fold-recombinations that demonstrate a compositional divergence within proteins have only been detected in multi-domain proteins of more than 4 domains in *Escherichia coli*, such examples are likely to exist in all topologies. These examples may occur in less frequent fold-recombinations, that have been excluded to avoid stochastic errors. For each individual protein that comprises domains with heterogeneous compositions, a compositional divergence is suggestive. Finding reasonable evolutionary pressures that support the hypothesis of a compositional divergence can help to detached this phenomenon from a chance event.
In conclusion, topology plays probably only a minor role in the harmonization or divergence of compositions within the same protein in the analyzed genome of *Escherichia coli*. This hypothesis is further discussed in Section 4.5.1. If topology plays no major role in harmonization, the assumption that pressures on the overall sequence dominate structural preferences of individual domains becomes even more evident.

4.4 Tracing harmonization

Several constraints such as fold-specific recombination, structure, topology and genomic context have mostly been excluded from the possibility of causing the observed enhanced similarities of the amino acid compositions of domains within the same protein. This finding has been hypothesized to be caused on the DNA level and specific codon usage as harmonization at both levels correlate.
The origin of codon bias is one of the most controversially discussed topics in molecular evolution. Here, few aspects that relate common research of codon bias to my work are outlined in order to put it into context with other results. In Subsection 4.4.3 the linkage between harmonization on the codon and amino acid level is derived and discussed in the light of previously published results.

4.4.1 Codon usage bias and expression level

The biased usage of codons has been associated with the *expression level* of proteins [Lobry and Gautier, 1994; Karlin et al., 1998], based on the assumption that some codons increase the expression level while others decrease it. The *codon adaptation index* (CAI) [Sharp and Li, 1986] relates the usage of codons in a protein to those in a reference set of highly expressed proteins (ribosomal proteins for examples):

$$CAI = \frac{\sum_a \sum_c n_{ac} \cdot x_{ac}}{\sum_a \sum_c n_{ac}} \quad , \quad x_{ac} = \frac{y_{ac}}{\max(y_a)}. \tag{4.10}$$

If harmonization correlates with the CAI, it is presumably associated to the regulation of expression. Due to time constraints, I have not derived the CAI for the used proteomes but used a different method to study protein codon bias in the following.
The set of optimal codons responsible for a high CAI has been found to be species-specific. The reason for this is that *translational selection* can be caused by different factors such as tRNA gene abundance [Lobry and Gautier, 1994; dos Reis et al., 2004] or the GC content of the silent, third base (GC3) [Sharp and Li, 1986]. In this context I want to note, that an enhanced GC3 content has also been associated with transcription initiation [Karlin et al., 1998] and to be caused by GC-biased mismatch repair systems during recombination processes [Birdsell, 2002; Lesecque et al., 2013]. Given these different constraints shaping the codon usage, it is unclear, which is actually responsible for a harmonization on the codon level.

4.4.2 Directionality of codon bias

The assumption that translational selection is the only cause of harmonization implies that codon usage is directional either towards codons that increase or decrease the expression level. Under the assumption of harmonization along all protein chains, the composition within proteins is the same as that of individual domains. In the following, protein composition are analyzed to investigate the directions of harmonization. The direction, as defined by the CAI, would lead to a correlation between the frequencies of optimal codons in highly expressed and between suboptimal codons in lowly expressed proteins. If other dependencies exist, harmonization may be multi-directional. Frequent co-occurrences of the same codons are indicative for clustering of codons within the same protein and thus hubs of similar compositions. In order to detect compositional hubs, codon composition is correlated within all protein sequences. For this, the relative codon composition of each protein is derived. The Pearson Correlation Coefficient (PCC) between the relative composition of all codons is calculated over all proteins.
In Figure 4.13, the derived PCCs are depicted for the genome of *Haloterrigena turkmenica*. The codons are sorted from top to bottom and left to right with increasing frequency in the whole genome. There is an apparent structure in the co-occurrence of

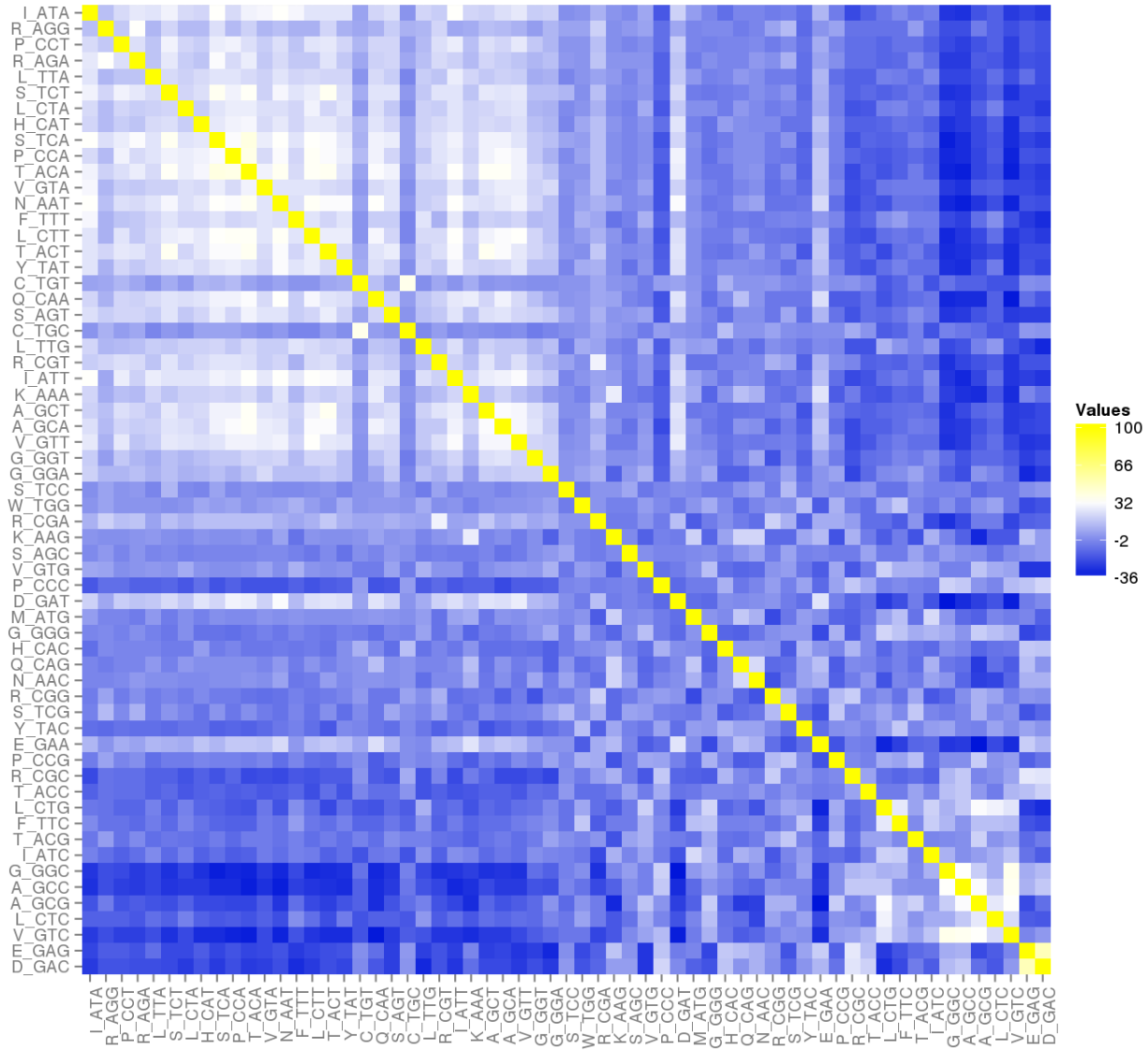

Figure 4.13: Correlation of codon frequencies in *Haloterrigena turkmenica*.

codons. A white square in the upper left corner of the plot indicates that rare codons (which are AT-rich) tend to co-occur, with an exception of the two codons encoding cysteine. These rare codons also anti-correlate with the seven most abundant codons. The other half of codons demonstrates a more complex internal pattern of correlations. However, some codons seem to have similar PCCs for all other codons, implying a clustering of these codons in the same proteins. These results suggest that there are is one large group composed of mostly rare codons and several more complex correlations between smaller subsets of codons. Harmonization in *Haloterrigena turkmenica* may thus be

multi-directional.

To analyze this further, the vectors of PCCs for each codon can be correlated in order to find codons that share certain co-occurrence patterns with other codons. For this, it could be helpful to apply a Principal Component Analysis or other clustering algorithms to this problem. Revealing a multi-directional use of codons can complement the commonly used linear measures of the CAI and the tRNA adaptation index (tAI) [dos Reis et al., 2003], which reflects tRNA-anti-codon affinity and tRNA gene abundance. This can potentially indicate a codon-specific trade-off between the usage of optimal codons and those associated to abundantly encoded tRNAs or even reveal other constraints that have an impact on codon bias.

Binary pattern of codon usage

Moving on to eukaryotic genomes, the pattern in the PCC of codon frequencies changes significantly. It takes the shape of an uneven chess board in the case of the human genome (Figure 4.14). This observation is also shared among all other studied eukaryotic genomes of *Sus scrofa*, *Mus musculus*, *Orcinus orca* (data not shown). The pattern of the PCCs implies an almost binary classification of codons into two groups according to the nucleotide at the third position, which is either GC3 or AT3. Exceptions are the codons TTG (leucine) and ATG (methionine), which seem to co-occur more often with codons of the AT3 group. Exceptional codons differ between organisms: CGA and CGT for example belong to the group of codons with GC3 in *Sus scrofa*. This binary pattern could not be observed in plants (*Solanum tuberosum*, *Arabidopsis thaliana*).
The binary division of codons in the human genome has recently been highlighted in [Hia et al., 2019], further demonstrating a bimodal distribution of GC3-content among proteins. Therein, the authors find strong correlations between mRNA stability and GC3-content, which results in longer half-times of mRNAs and therefore into higher expression levels.
Constraints acting on the *Haloterrigena turkmenica* codon usage and those acting on *Homo sapiens* are essentially different. The abundance of tRNA genes is generally increased in eukaryotes [Goodenbour and Pan, 2006] and translational selection may not play such a great role [Hia et al., 2019; dos Reis et al., 2004]. Instead, the GC3 constraint seems to be the major determinant of codon clustering.
In conclusion, compositional preferences of codon usage within proteins display a complex pattern in archaea and bacteria, which needs to be further investigated to derive concrete assumptions about its relationship to the observed harmonization. In eukaryotic genomes, there seems to be a clear preference for either codons with GC3 or AT3, which was been related to mRNA stability [Hia et al., 2019].

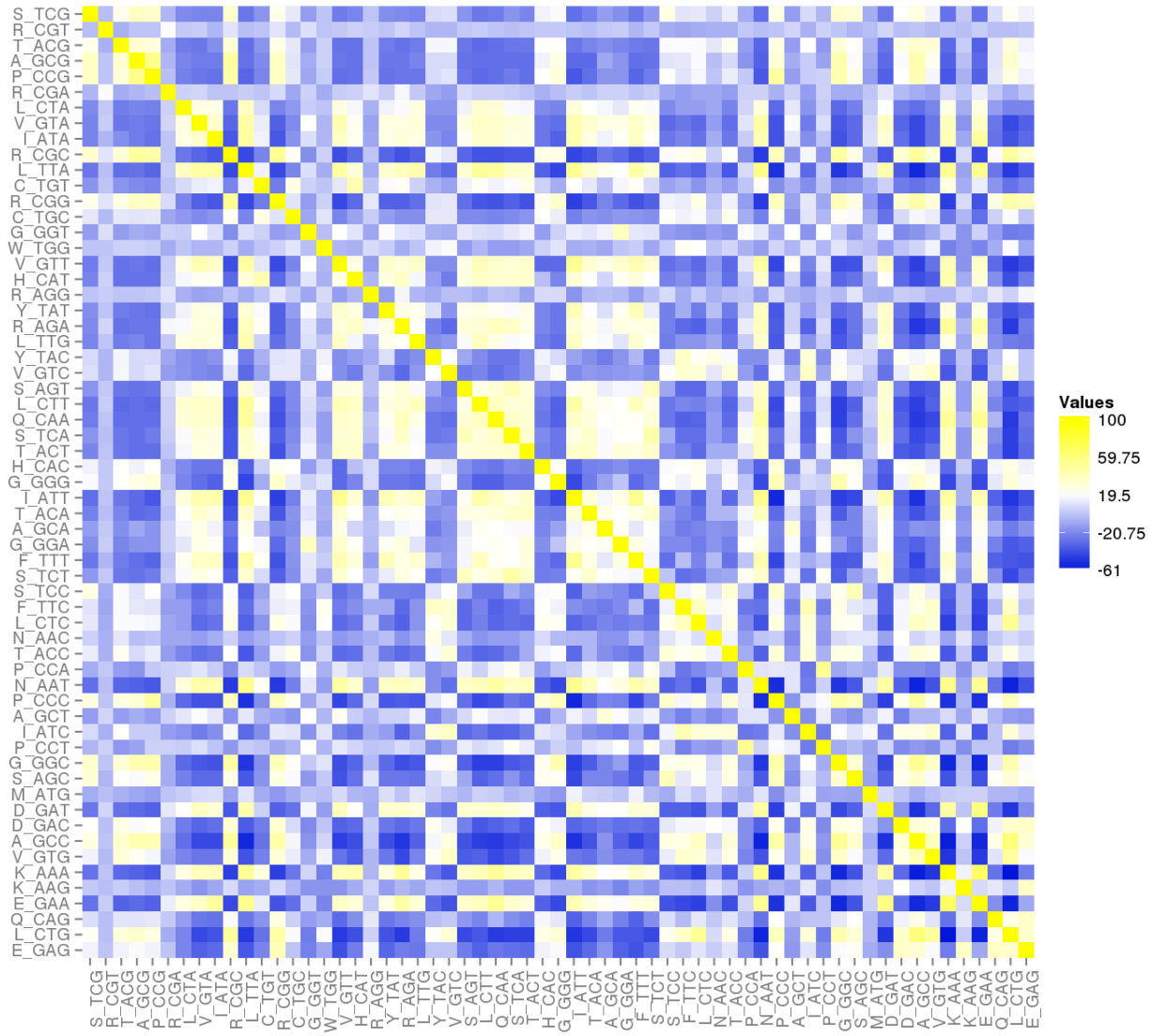

Figure 4.14: Binary co-occurrence of codons in *Homo sapiens*.

4.4.3 Coupling of amino acid with codon harmonization

Having analyzed the nature of codon bias that may lead to the observed harmonization of amino acid compositions, I aimed to characterize the relationships between harmonization on both levels. A certain amino acid composition can be achieved with different codon compositions due to the degeneracy of the genetic code. In order to estimate how strongly coupled the harmonization at the amino acid with that at the codon level is, I derive the common frequency of amino acids that can be explained by the common usage

of codons as

$$CL = 1 - \sum_{a=1}^{20} \min(F_1(a), F_2(a)) - \sum_{c} \min(f_1(c), f_2(c)) \tag{4.11}$$

where $F_i(a)$ is the frequency of amino acid a to occur in sequence i and $f_i(c)$ the frequency of codon c, coding for amino acid a. With this, the usage of identical codons in both compositions opposed to the usage of synonymous codons is captured. As this coupling is more similar to the Manhattan distance, the L1-norm is used in this subsection opposed to the L2-norm as in the other studies. If $CL = 1$, the overlapping amino acid frequencies in both sequences arrives from overlapping codon frequencies and the similarity is coupled between amino acids and codons. This is also the case, if no or little amino acid overlap exists between the two compositions, which is associated to a high percentile rank. With $CL = 0$ the opposite is the case where no overlapping amino acid frequencies arrives from overlapping codon frequencies in the respective domains and the similarity is thus decoupled. This coupling is generally stronger if codon bias constrains the whole genome, for example by higher or lower genomic GC content, as also outlined in [Goncearenco and Berezovsky, 2014]. This effect is here not normalized for. In the case where CL is not compared across genomes such a normalization is not essential.
In Figure 4.15, the distribution of this coupling is depicted for the genomes of *Haloterrigena turkmenica*, *Escherichia coli*, and *Homo sapiens* for both comparisons across and within proteins of domain compositions. For *Haloterrigena turkmenica* and *Escherichia coli*, the coupling is similarly distributed in these comparisons across and within proteins, with a tendency of stronger coupling of compositions within the same protein. This implies that the coupling between amino acid composition and codon bias is close to constant among all domains and amino acid usage is directly coupled to codon usage. This differs form the observation in the human genome, where the comparisons across proteins indicate a smaller coupling between amino acids and codons (Figure 4.15: upper panel, right). In this case, codon choice of a particular amino acid is dependent on the protein. The codon choice is biased in entire proteins but decoupled from a generally defined codon usage. This observation is in line with the differential usage of codons with either GC3 and AT3, as laid out in Subsection 4.4.1, which seems to act as a switch without great impact on the amino acid composition.
With increasing similarity of compositions (decreasing percentile rank), the coupling weakens (Figure 4.15: bottom panel). This effect is due to a numerical issue of Equation 4.11, given that perfect coupling can also be achieved with zero overlap in the amino acid compositions. Further research could account for this effect by incorporating the amount of harmonized amino acids into the coupling, thereby not only including the coupling between amino acid and codon harmonization. Nevertheless, irrespective of the amount of harmonized amino acids, which can be transferred into the significance of compostional similarity by the percentile rank, comparisons within a protein are stronger

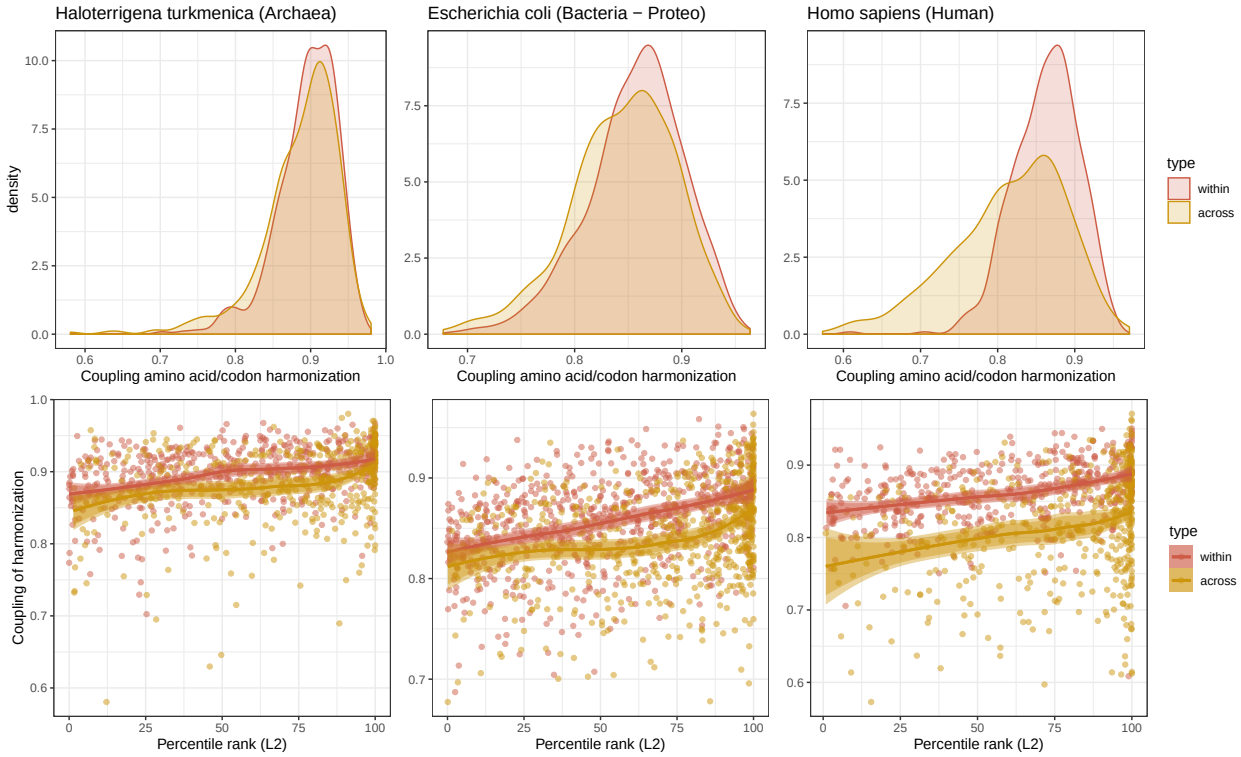

Figure 4.15: Coupling of harmonization between amino acids and codons. The first panel illustrates the distribution of coupling between the harmonization of amino acids and codons according to Equation 4.11. For *Haloterrigena turkmenica* and *Escherichia coli* the coupling in the across and within comparison is similarly distributed, indicating that amino acid and codon composition are generally coupled. In *Homo sapiens*, the coupling is stronger for domains in the same protein. The second panel depicts the coupling along the percentile rank. In all three cases, the coupling decreases with percentile rank.

coupled than comparisons across proteins. In all genomes, the amount of common amino acids in domains within the same protein is thus generally more strongly coupled to common codon usage. Harmonization on the amino acid level is therefore directly coupled to a harmonization on the codon level of the respective amino acids.

4.5 Discussion and Outlook

4.5.1 Summary

There is a tendency of domains in the same protein to have similar amino acid compositions, which is unexpected form random domain recombination in the same genome. This tendency has been captured using a random sequence model that accounts for the

combined composition of domains (D_2-model) and has been measured by comparing percentile rank distributions that indicate the significance of distinct compositions. In proteins of archaea and bacteria, the enhanced similarity of amino acid compositions of domains within the same protein occurs more frequently than in eukaryotes. It was possible to largely detach its origin from fold-specific recombination bias, structural constraints to folds, protein topology and recombination bias in genomic regions. Instead, it was found to correlate with the similarity of codon composition in the respective domains. Using a metric to capture the origins of overlapping compositions, a coupling between amino acid and codon usage could be established. Similar amino acid compositions were more strongly coupled to similar codon usage in domains within the same protein for all analyzed genomes. Eukaryotic genomes additionally demonstrated that similarities of amino acid compositions across different proteins was less coupled to identical codon usage than within the same protein, an effect that is associated to the almost binary usage of codons with either GC3 or AT3 in proteins. The amino acid composition of proteins is thus coupled with codon bias, which here, for the first time, has been studied with respect to domains as independent units in the evolution of proteins.

Reassessment of previously presented results

The results presented in Subsection 2.3.2 have led to the hypothesis of a minor compositional heterogeneity within proteins, indicated by a minor difference between the total residuals between the L- and the P-model. These results are in correspondence with the results presented in this chapter.
As highlighted in Table 4.2, three out of five bacterial genomes possess a compositional homogeneity that is not distinguishable from their combined composition (two out of five in the case of the L1-norm). The remaining two genomes are only slightly below the acceptable risk too assume there is no compositional fluctuation within proteins. This is in line with the observation that the total residual of a random model that considers the composition of local, sub-protein sized fragments (L-model) and a model that considers the natural protein composition (P-model) are very similar.

The study presented here and the previous study use different settings which need to be considered in this comparison. Here, only the compositions of well-structured sequences are compared, while in the previous study the data set consists of 30% residues that could not be assigned to a structured domain and are thus potentially unstructured. The contribution of these unstructured and structured parts to the heterogeneity of composition within proteins was already stated in Subsection 2.3.3. Also, using a data set of 1,307 bacterial genomes in the previous study, some of these may possess a compositional heterogeneity within proteins, some do not. In conclusion, the homogeneity of amino acid composition within proteins depends greatly on the given data set. Here, it is found to be generally increased in archaea and bacteria and less in eukaryotes.

Topology and time-dependency

In order to validate that topology has no impact on harmonization, further analysis is necessary. First of all, a diversity of sequence data needs to be investigated to evaluate if this observation is universal as it may be a bias of the studied genome of *Escherichia coli*. Also, the effect of recombination of homologous domains should be investigated further and its relationship to topology.
Another reason for compositional divergence of higher typologies may be related to the age of a recombination event. Under the assumption that there is no mechanism that selectively recombines domains with similar codon compositions (a counter example for this assumption related to GC-content is given in [Birdsell, 2002]), it is selection and evolutionary pressures that shape the evolution of the DNA sequence after recombination. This suggest that only with time the observed harmonization of domains within the same protein will appear. Recent recombinations may not have undergone enough selection and may thus display weaker compositional similarities.
There is a recognized correlation between time of recombination and the number of domains in a protein [Kummerfeld and Teichmann, 2005], implying that proteins composed of multiple domains map back to evolutionary more recent recombination events. In the general assessment of the percentile rank distribution, such a correlation between the number of domains and harmonization could not be established. This may be due to the used genomes, which may possess mostly proteins where recombination has happened in ancient times or that the harmonization process by itself takes only little time. It is also possible that recombination events occur frequently and those that are harmonized from the beginning are more likely to be selected.
Similar to this selection argument is the possibility of a harmonizing process to act before recombination on the single domains. This would also lead to instantly harmonized proteins with no transition. A plausible analysis to investigate this possibility is to trace recombined domains and non-recombined orthologs in other genomes. If these orthologs are harmonized, it is possible that harmonization occurs even before the recombination event.

Coupling - which way does the pressure go?

In this section I have demonstrated that similar amino acid usage of domains within the same protein is coupled to a similar codon usage. It is not directly clear, which of these two has a stronger impact on the other.
There are evolutionary pressures that constrain proteins to generally possess a homogeneous amino acid composition along the chain. Possible reasons are the use of amino acids of lower biochemical costs, which is more biased in highly expressed proteins [Akashi and Gojobori, 2002]. From a functional perspective, enhanced solubility or hydrophobicity of all domains in the same protein may lead to an advantage as interactions with water or a membrane become more optimal [Trevino et al., 2007]. If these factors

lead primarily to a harmonization of amino composition, codon harmonization may only be following afterwards. This could in principle be reflected in a diversity of synonomous codons. A tendency of the percentile ranks in amino acid and codon composition (see Figure 4.15) to be biased towards similar amino acid but dissimilar codon usage could support the notion of amino acids to harmonized before codons, which could however not be established for the general case. Another fact to be considered in this context is that harmonization at the codon level could go hand-in-hand with amino acid harmonization, leaving no traces of a amino acid harmonization before a codon harmonization.
Apart from pressures at the amino acid level, there are many factors that contribute to an evolutionary pressure to a biased codon usage along a protein chain such as translation efficiency, tRNA abundancies and the regulation of the expression level. Given the great amount of research substantiating these pressures that are manifested in the DNA sequence, the argument that codon bias to impact amino acid composition is more straight forward. The relevance of the resulting codon bias is undisputed and within the boundaries of acceptable, functional protein sequences, evolution optimizes the usage of codons. This optimization may cause mutations at the amino acids level, which can be interpreted as codons to harmonize and by necessity, causing amino acids to harmonize. This scenario may be the most parsimony explanation for the direction of the pressure.
In specific cases of individual proteins, the amount of pressure for harmonization and also for diversification at the amino acid and codon level are likely to differ. In general, it is most certainly the interplay between forces on both sides, from the DNA and the protein level that lead to the observed coupled harmonization.

Keep away from cliches, this world is much more complicated.
- Noam Chomsky

4.5.2 Further studies

Relationship to common measures of codon bias

Codon bias has been studied for the *effective number of codons* (ENC) [Wright, 1990] in individual protein sequences. It is indicative of the constraints arriving from codon bias and accounts for the compositional entropy among codons of the same amino acid. A small ENC results from a biased use of codons, which is achieved when specific codons are more often used than others for the same amino acid. The ENC is coupled to the codon adaptation index (CAI) [Sharp and Li, 1986] and the tRNA adaptation index (tAI) [dos Reis et al., 2003], as these measures indicate synonymous codon preferences.
Relating the ENC to the amino acid composition of the individual domains and the entire protein will put my work in perspective to related work concerning codon bias. Harmonization is likely to be coupled to the ENC and also to the CAI or tAI. It can further relate protein structural and functional constraints to these commonly used measures.

Optimal translation by optimizing amino acid sequence

Optimizing genes for enhanced expression level is a hot topic in molecular design [Saito et al., 2019; Webster et al., 2017]. It is mostly based on the idea of finding an optimal codon sequence that translates into a desired amino acid sequence. Given the many options in protein sequence space for the same structure and the same function, it seems appealing to also optimize the amino acid sequence for this purpose. Certainly, mutations on the amino acid level can effect stability and function of the protein, which complicates the proposed strategy. Enriching a protein sequence to a profile by including closely-related proteins however should allow for a conservative radius of possible sequence space. Such attempts of a hybrid amino acid and codon optimization have not been published before.

Constraints to harmonization

Harmonization is derived by the cumulative contribution of each amino acid or codon. Likewise an increased ENC can be associated to specific amino acids or codons. Their individual contributions to the overall harmonization may be different. For example, if the substitution to a specific amino acid is less often deleterious, codon harmonization could possibly proceed more easily through this amino acid than more deleterious amino acids. Preliminary results (data not shown) indicate that indeed harmonization differs between amino acids. If this kind of constraint from the protein level is universal for all genomes, it is possible to estimate which amino acids obey the rules from the DNA level more or less. This may help to estimate if the amino acid composition of a given protein is affected by harmonization at the codon level and maybe allows to estimate which residues are under greater evolutionary constraints at the protein level.

Chapter 5

Conclusion and Discussion

5.1 Conclusion

5.1.1 It's never just one thing

Proteins are facing a great amount of challenges that all together influence their evolution. First of all, proteins must maintain their function if it is essential to the organism. This function emerges through important interactions with other molecules that mostly need to be conserved in evolution. For this, proteins need to be transported or diffuse to the right location in the cell and escape the degradation process before fulfilling their functions. The structure of proteins is the next most important feature of proteins as it defines interaction sites with other molecules and possible conformational changes of the protein. Stability is key for a successful structure. In order to obtain its structure, a protein needs to fold, which requires to be rapid and to circumvent undesired low-energy states that lead to mis-folding. Only after the translation of an mRNA sequence into a protein sequence, a protein can fold into its functional structure. This translation process is often optimized for expression level, mRNA degradation, mRNA folding, tRNA abundance and codon affinities to tRNAs among other features that impact translation. mRNA itself is transcribed from the DNA, which is the level where most information of the molecular machinery is stored. DNA is exposed to many other factors that directly change the stored information, which are foremost mutations, but also mutational biases, methylation states, histone-interaction, general DNA-organization and DNA repair mechanisms. Although these forces are seemingly limitless, nature has proven that all of them can be embedded into single sequences. The balance and tradeoff between these different forces is not the same for each sequence and may differ between genomes, causing variability among protein sequences. The reason why protein sequences are able to account for all of these features, is the vastness of the possible space. Sequence space comprises many sequences that can perform the same function or possess the same structure and that on top of this also can adapt to constraints on the DNA level. Protein sequences are a balanced compromise of all of these constraints together.

5.1.2 Contributions of this thesis

One of the great challenges in evolutionary biology is to reveal the strongest forces that define constraints of evolution. Given the variability and possible interactions between all constraints, it is not easy to separate the effects of these forces from each other. With this thesis, I have contributed to a better understanding of several constraints and their impact to present-day protein sequences.

Above all, the sequence space occupied by natural proteins is globally of a mostly random structure, an observation that has previously been demonstrated [Weiss et al., 2000; Strait and Dewey, 1996; Lavelle and Pearson, 2009]. Apart from this randomness, the strongest and most abundant feature of natural protein sequences is their diversity in amino acid composition usage, across genomes and proteins. Although the precise reasons for such compositional fluctuations have not been analyzed here in detail, other research efforts were able to associate compositions with specific forces that act on most proteins. This includes, constraints on genomic GC-content due to environmental [Fukuchi and Nishikawa, 2001; Fukuchi et al., 2003], mutational biases [Palacios and Wernegreen, 2002], amino acid biosyntetic costs [Akashi and Gojobori, 2002] or translation efficiency [Quax et al., 2015].
Instead of concluding that this compositional fluctuation is the most relevant feature of natural protein sequences, it can be interpreted as the great capability of proteins to adapt to a diversity of constraints. The same functional protein can in distinct genomes possess very different overall compositions. Although proteins are known to be fragile molecules that can break easily when applying random mutations, they demonstrate to be extremely versatile and to be able to embed many different constraints at once, leading to the different usage of sequence space. This compositional modulation causes long-range sequence correlations that become more pronounced with the length of the considered sequences. Random sequence models that do not account for local amino acid composition have here been demonstrated to reflect similarities among natural sequences less accurately with increasing sequence length, whereas models that account for the local composition reflect similarity with a comparable accuracy irrespective of the used sequence length. This suggest that the most suited random sequence model, should account for the amino acid composition at the level of proteins. The statistics of BLAST are based on a similar model as used in this thesis [Schaffer, 2002]. The remaining similarities that cannot be captured by these models could be associated with short-range sequence correlations among structured, non-related protein sequences. Sequence patterns that are curated from compositional biases could here be derived, which may correspond to truly biochemically preferred sequence patterns related to common secondary structure formation.

Constraints that limit the sequence space exploration of diverging proteins are specific for each protein family and have here been presented to have no global impact on natural

sequence preferences. This observation is promoted by the vastness of sequence space that leads to a given function or structure to be spread across sequence space. Sequences that are positioned in distant locations of this space share a randomly expected similarity but can still be united by common descent [Rost, 1997] or common structure [Tian and Best, 2017]. Thus, although belonging evolutionary or structurally to the same cluster, sequences may be far apart in sequence space. This effect causes evolutionary footprints to dissipate in sequence space, while common constraints among many sequences cause convergence that has a detectable impact in the overall structure of how sequence space is occupied by natural proteins. Divergent evolution nevertheless leads to an increased abundance of similar sequences in local proximity, allowing to infer common descent from a high sequence identity. The transition between local and global sequence space, where similarity between homologs is separated from randomly expected similarity has been coined as the twilight zone [Rost, 1999]. Here, I have revisited the definition of the twilight zone, focusing of the statistical significance of similarity. This work is therefore detached from structural similarity, which was used in previous studies to confirm common descent.

With the here presented approach that contrasts the bias of the amino acid composition of genomes, proteins and domains to that among natural sequences, it was possible to compare the impact of these individual biases. The difference between the composition of proteins and that of domain-sized fragments led to comparable results in a data set of bacterial genomes, which was unexpected at first. The origin of this compositional similarity between domains of the same protein could be dissociated from structural constraints, biases of recombinations in genomic regions and was demonstrated it to be coupled to the usage of identical codons. Although correlations between amino acid and codon compositions have been studied several times [Lobry and Gautier, 1994], for the first time, this approach combines the knowledge of domain structure and constraints on a functional protein level to the biased codon usage on the DNA-level. In most proteins both amino acid and codon compositions have harmonized to a noticeable amount between the comprised domains, an effect that is more pronounced in archaea and bacteria than in eukaryotes. Proteins that possess a heterogeneous amino acid and codon composition, supposedly experience more pressure from the function level than from the DNA-level, suggesting that codon bias may not be essential in individual cases. With this, the presented work has shed more light on the constraints that balance between the DNA and protein world.

5.2 Final discussion

5.2.1 The curse of dimensionality

The vastness of sequence space is not easy to grasp and relating the few observed sequences that nature has explored to it, is a challenging task. This global occupation by natural sequences was assayed in this thesis using an approach that compares distances between sequences. Representing objects by distances in their respective space, becomes however less comprehensive with increasing dimensionality. That is because the number of maximal distant locations grows exponentially, while the amount of direct neighbors grows only linearly. This leads to a rapidly decreasing ratio r between short to long distances with fragment length n:

$$r(n) = \frac{n}{19^{n-1}}. \tag{5.1}$$

The longer a sequence, the smaller is the ratio between the number of sequences in its nearest neighborhood and maximally distant sequences. For long sequences, mostly all positions in sequence space are far away from most other points. This effect is reflected by the fact that the distance distribution of 100mers is centered around a low value of sequence identity (see Figure 2.2). Proximity becomes far less probable and hence also a more significant feature, resulting in the shift of a significant sequence identity with sequence length (see Figure 3.9).
An accumulation or clustering of sequences in such a high-dimensional space needs to overcome this dimensionality in order to be globally detectable. Composition and frequent convergent sequence features could be found to be strong enough to have a global influence on the occupation of sequence space. Divergent evolution, which relates only few of all explored sequences to each other, does however not result into a globally traceable footprint.
Due to this curse of dimensionality, the work in Chapter 2 has received major criticism in the reviewing process. Protein sequences are part of a high-dimensional space and there is no way around this fact. Our approach of studying the abundance of specific distances relative to a random model allowed to overcome the trailed problem of sparsity caused by this high-dimensionality. With this, the presented work succeeded to investigate long sequence fragments and is a continuation of previous approaches [Lavelle and Pearson, 2009; Poznański et al., 2018] that stagnated at a sequence length of six residues.
This distance-based approach could further capture how strong specific biases impact the relative distances of sequences to each other in this space. It does not pin-point these biases to specific regions in sequence space but only indicates that regions of increased occupation exist or sequences that are relatively distant from other sequences. This limitation indicates that a distance distribution is completely detached from the actual locations in sequence space that have been occupied. Therefore, a given distance distribution can be achieved by a tremendous amount of possible sequence space occupations. The

dimension n	0	1	2	3	4	5	point mutations
1	1	19	-	-	-	-	count
2	1	$2 \cdot 19$	19^2	-	-	-	
3	1	$3 \cdot 19$	$3 \cdot 19^2$	19^3	-	-	
4	1	$4 \cdot 19$	$\binom{4}{2} \cdot 19^2$	$\binom{4}{3} 19^3$	19^4	-	
5	1	$5 \cdot 19$	$\binom{5}{2} \cdot 19^2$	$\binom{5}{3} 19^3$	$\binom{5}{4} 19^4$	19^5	

Table 5.1: Size of sequence space point mutation neighborhood. For any sequence with n residues, the number of sequences with a certain number of point mutations is given in each row. The neighborhood of maximal distant sequences is always the largest and grows exponentially with increasing sequence length.

natural data set can thus easily be replaced by an artificial data set that leads to the same distance distribution. Hence, not all sequence space occupations that possesses the characteristic distance distribution of a natural data set comprises foldable protein material; in fact most of them do not.

5.2.2 Islands in sequence space

A recurring concept in my doctoral studies was the idea of nature to populate sequence space like islands a vast sea.

> *"Proteins are not spread uniformly across the full sequence space; instead, they are clustered tightly into families"*
>
> [Huang et al., 2016]

This idea is challenging the observation that the populated sequence space is indeed of a mostly random structure. While the image of natural islands in sequence space is not unsubstantiated, it needs to be carefully considered when drawing assumptions about the occupation of sequence space.

Islands scattered over space

In terms of sequence space occupation, observed sequences are associated with land and unobserved, mostly invalid ones with a huge sea. An island is generally characterized by sharp boundaries between land and sea. In a two dimensional space, boundaries of an island can be defined by a single line, while in a higher dimensional spaces boundaries that mark the transition between two subspaces tend to become complex. Instead of drawing a line in multidimensional space, sequences that are thought to be on an island are, therefore, often enumerated, summarized in a multiple sequence alignment and represented by a consensus profile or sequence, which can be imagined as the center of

the island. However, a consensus profile does not directly draw a line in sequence space, as correlations between positions in a sequence are not reflected. It rather spans over all possible combinations of amino acids that are given in the consensus profile, also over sequences not actually observed in nature, which are part of the unknown sea.
At the core of most profiles, there are often few anchor residues [Rost, 1997] and the rest of the sequence may assume seemingly arbitrary amino acids. It is thus easily possible to derive two sequences that are actually very distant in sequence space while coming from the same profile. In fact, most homologous relationships share an almost randomly expected sequence similarity [Rost, 1999] (see homology convergence). Hence, sequences that are related to each other may span over the whole space, implying that their corresponding island is scattered everywhere, without an obvious global structure. The main reason for this is the dimensionality of sequence space and the fact that the large majority of pairwise distances is maximal (see Subsection 5.2.1). This scattering across the whole space is the main reason why the image of tightly clustered islands is often unsuited to describe how sequence space is occupied by natural protein sequences.

The bonds of evolution

Nevertheless, the way evolution explores the sequence space can be compared with a random walker that can stand on land and that would drown when stepping too deep into water. This concept was first described by [Smith, 1970], where nature is assumed to traverse only over viable sequences through the possible space. Therefore, related sequences can be assumed to be somehow inter-connected. While not all related sequences are in proximity to each other in sequence space, there are or have been intermediate sequences that bridge these sequences together.
These traceable connections are relevant for establishing evolutionary relationships and also to infer knowledge between less similar sequences. They define the local structure of sequence space, they teach us how sequence space got populated and which sequences are more relevant than others.
Although the global occupation of sequence space is almost random, it is evolution that has chosen the ways.

Appendix A

Composition studies

A.1 Composition of genomes

Clustering genomes by composition

The first approximation of a more local composition than the overall composition, as captured by the A-model, was achieved by using the composition of individual genomes in Chapter 2. The resulting G-model is thus based on the background frequency of the 1,307 comprised bacterial data sets and their respective sizes. In the case that each genome has the same composition, i.e., the propensities of each amino acid is roughly the same of different genomes, there would be no difference between the results derived from the A-model and the G-model. However, if amino acids are distributed differently between genomes, A- and G-model are capturing different compositional effects of natural sequences, where the G-model accounts for the fluctuation between genomes. Here, I focus on determining where this fluctuation comes from and which factors are driving the differences the most.
For each of the 1,307 bacterial genomes, its amino acid composition is derived by accounting for all occurring residues, resulting into 1,307 composition vectors. To analyze the fluctuation caused by specific amino acids, the distribution of the <A>-content across all genomes, where A represents one of the 20 possible amino acids, was derived. The distribution are presented in Figure A.1 along with their variance, reflecting the magnitude of fluctuation arriving from this particular amino acid.
The shapes of the amino acid content distribution are manifold. Some display a sharp spike, such as cysteine (C), methionine (M), histidine (H) and tryptophan (W). Others display a wide distribution, such as alanine (A), isoleucine (I), lysine (K) and arginine (R). Hence, the less frequent amino acids (C, H, M and W) contribute less to the overall fluctuation compared to the more frequent amino acids (A, I, K and R). The largest variance comes from the alanine distribution. It is a bimodal distribution, indicating that bacteria demonstrate a tendency for higher or lower alanine-content compared to the mean. This bimodal behavior also occurs for glycine, isoleucine, lysine and less pronounced, or even trimodal also for other amino acids.
In order to further analyze the origin of these bimodal behaviors the relative amino acid content of pairs of amino acids was correlated. The relative <A>-content of genome g

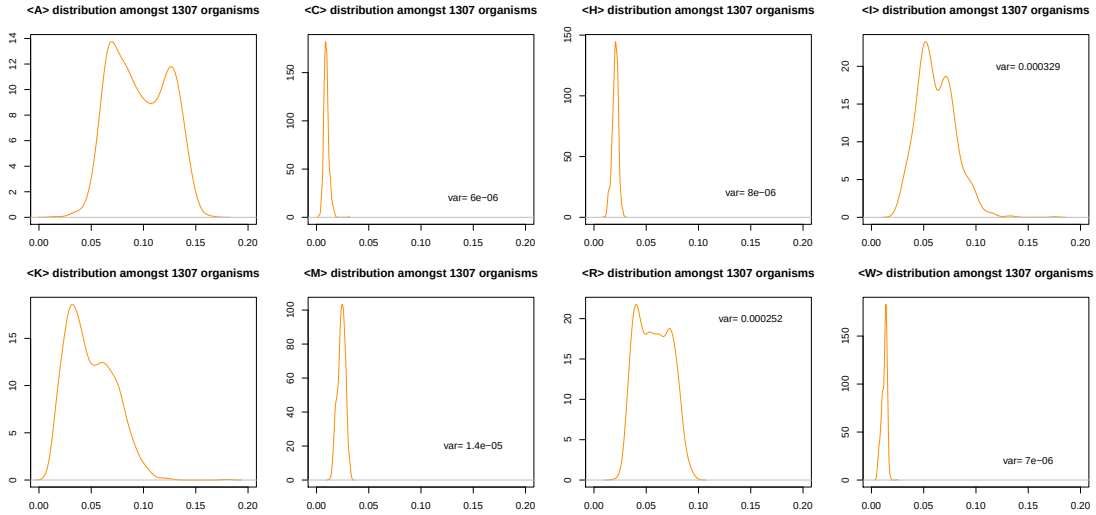

Figure A.1: Amino acid content of genomes.

was determined as the mean <A>-content of all genomes:

$$C_g^{rel}(a) = \frac{C_g(a)}{\sum_h C_h(a)/1,307}. \tag{A.1}$$

In Figure A.2 examples of the correlation between two amino acids for all genomes, along with their Pearson correlation coefficient are presented. The relative amino acid frequencies of amino acids with GC-rich codons (G, A, P and R) or GC-poor codons (I, K and N) are strongly correlating. Amino acids with codons of dislike GC-content are rather negatively correlated. This correlation between GC-content and amino acid composition has previously been recognized [Lightfield et al., 2011].
Having analyzed the fluctuation of amino acid contents and its connection to the GC-content, the composition was used to classify bacterial genomes. For this, the obtained composition vectors of all genomes were used to derived the pairwise compositional difference between genomes, as defined by the Manhattan distance (see Section 4.2). From all distances the 1% smallest were selected and clustered with the tool Cytoscape [Shannon et al., 1971]. The resulting cluster is depicted in Figure A.3 and colored according to the most abundant phyla. The cluster map clearly demonstrates a separation between most phyla. It also possesses an elongated shape, that may reflect a gradient of the GC-content with high content to the left and low content to the right. Actinobacteria have generally a high GC-content while Firmicutes have a low GC-content [Mann and Chen, 2010]. As no DNA-data was available for used genomes, any statement about the real underlying GC-content is only conjectured. Hence, from the amino acid composition by itself, it is often possible to cluster genomes into their corresponding phyla, possibly

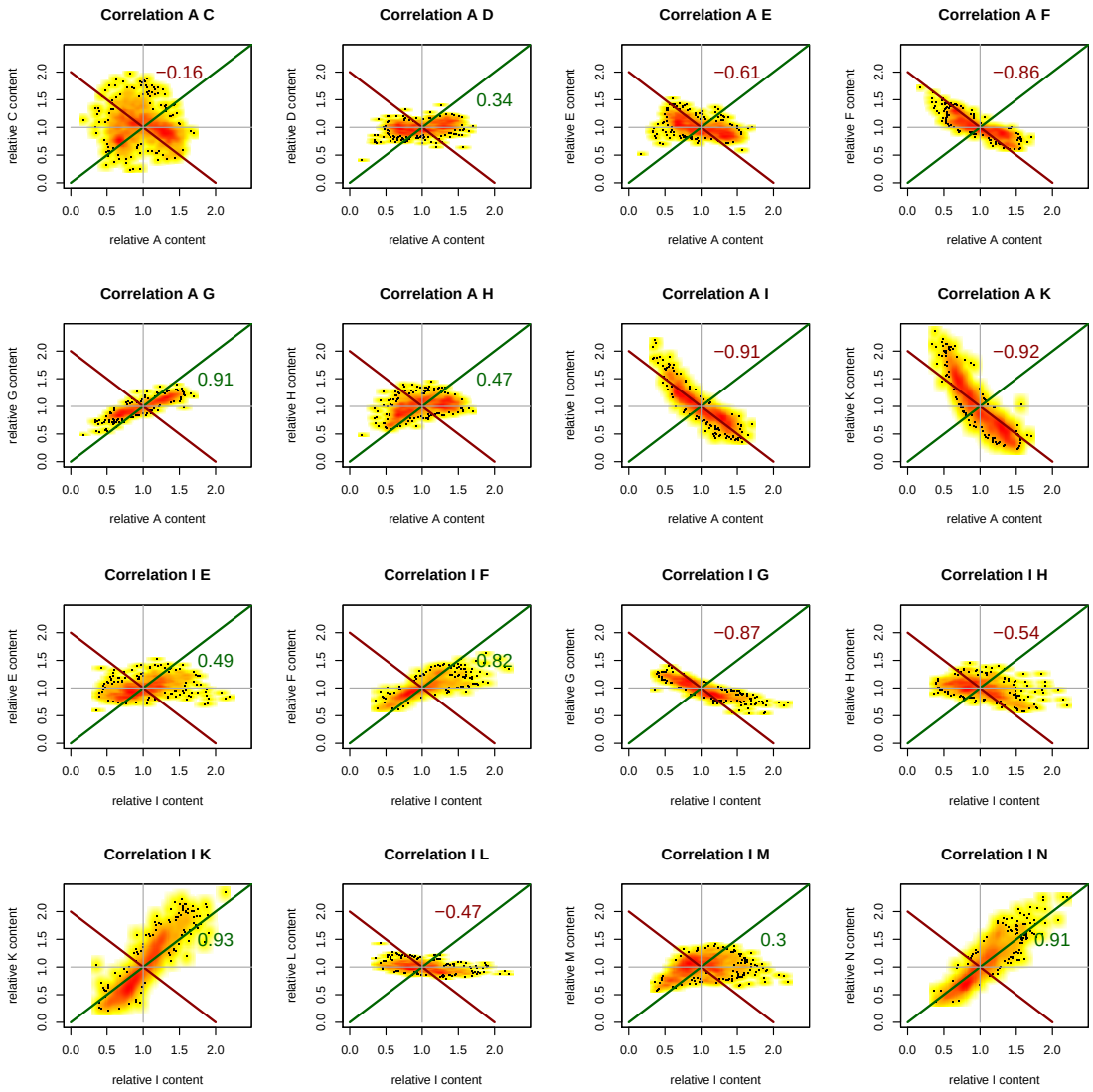

Figure A.2: Correlations of amino acid contents in genomes.

even along their GC-content.
From this study of the composition of genomes it is possible to conclude that there is already a great amount of information stored in the composition of genomes. A normalization by the composition of individual genomes thus approximates the background noise arriving from compositional effects across genomes better than a model based on the overall amino acid frequency. This is especially important when being interested in sequence relationships across distinct genomes.

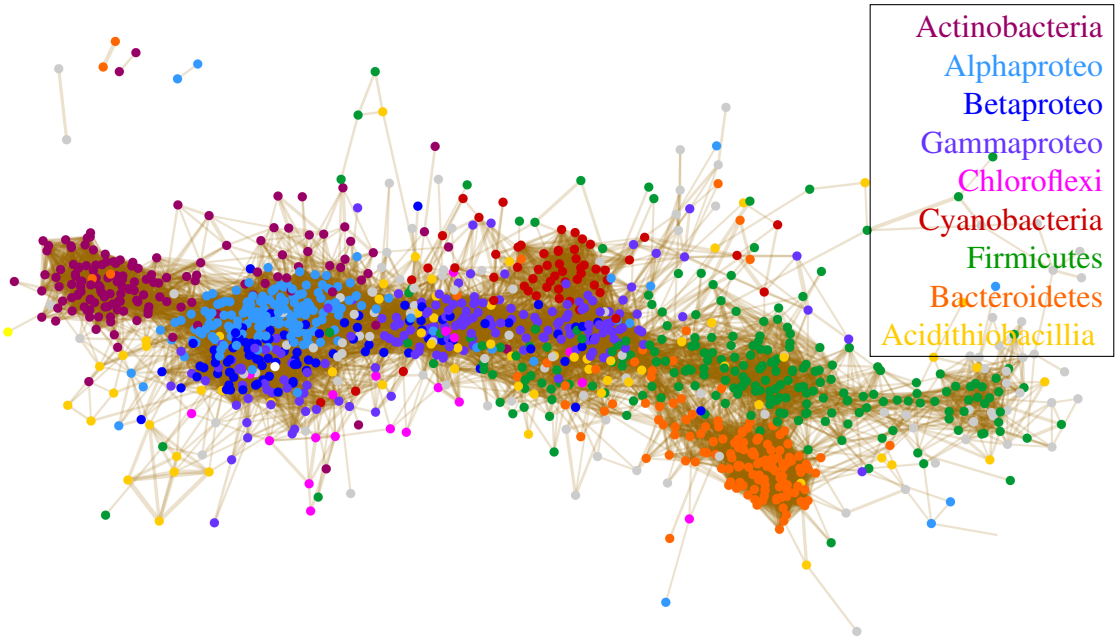

Figure A.3: Cluster of genomes according to genome composition. Bacteria of the same phyla, indicated by distinct colors, tend to cluster. This observation is related to a similar amino acid composition of their genomes. Related phyla such as the Proteobacteria tend to cluster in proximity to each other. An exception of this relatedness are genomes from Acidithiobacillia, which are seemingly randomly distributed over the cluster map. The elongated shape of the cluster may be reflecting a gradient of the GC-content from high GC-content of Bacteria located in the left of the map to Bacteria with lower GC-content in the right of the map.

A.2 Compositions within genomes

Diversity of protein composition

The next more local consideration of composition goes down to the level of proteins. Here, the amino acid content across proteins is analyzed, relative to the respective genome composition. The results are presented in Figure A.4. For each genome, the composition vector of each comprised protein is derived and the amino acid content distribution over the proteins for each genome (pAAC) is computed. Genomes with an amino acid content greater than the mean are plotted in light blue, genomes with a lower content in light green. The average amino acid content of distinct proteins is derived by combining all proteins together, thereby weighing the contribution of genomes by their size and plotted in dark blue. The previously presented amino acid content distribution amongst genomes is colored in orange. The results of alanine (A) and cysteine (C) are depicted. The alanine-content across proteins is greatly genome dependent. According to the alanine-content of genomes, there are some distributions with a high and some with a

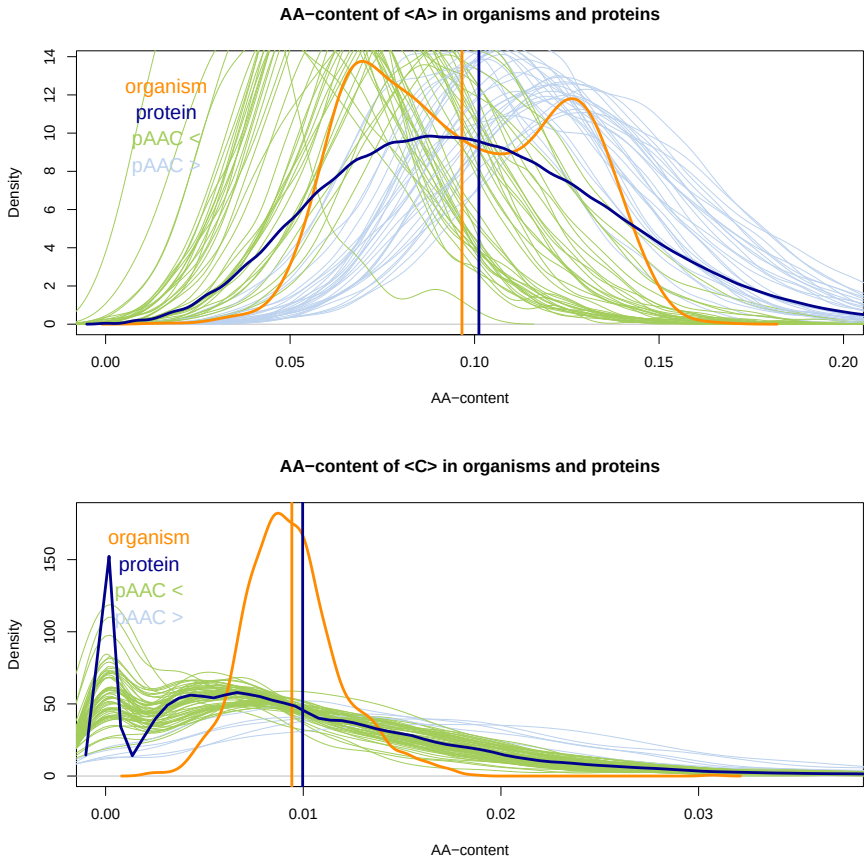

Figure A.4: Amino acid composition of proteins of district bacterial genomes.

low mean. What is more, irrespective of the genome, all content distribution of proteins possess a Gaussian shape. Thus, within a genome, alanine is distributed with a certain variance, but not in a bimodal fashion.
In contrast, the cysteine-content of proteins within the same genome demonstrates a bimodal behavior. There is a large peak with many proteins that have a cyteine-content of 0% . The next largest peak is still below a content of 1% and flattens out towards a higher content. Thus, in all genomes, proteins demonstrate to possess either no cysteines or a certain low amount up to mostly 3%.
In all cases the amino acid content distribution across genomes and proteins differs substantially. This is partially due to the finite sampling problem of short proteins sequences. Other deviations may be related to specific evolutionary pressures that impact the composition of proteins, resulting into the observed compositional fluctuations.

Compositional entropy

One well-known characteristic of natural sequences is the occurrence of low-complexity regions (LCR). These regions, where the same amino acid occurs repeatedly, can be a result of error-prone polymerases. As these sequences are very distinct from random sequences due to their restricted composition, LCRs were eliminated from the used data set. However, there is a sill remaining tendency towards sequences of lower complexity. In Figure A.5 the Shannon-entropy distribution (per residue) over all valid 100mers of our dataset together with the distribution derived from the A-model is presented. There is

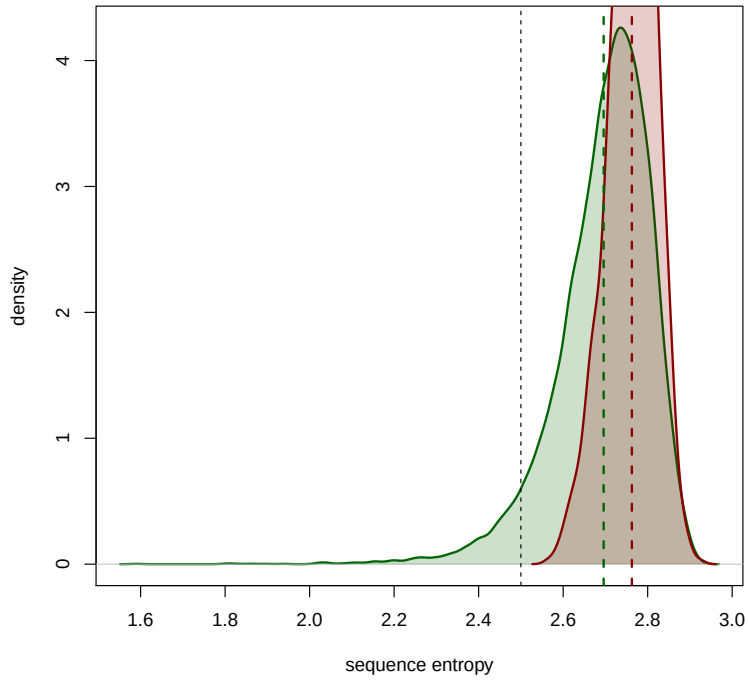

Figure A.5: Entropy of composition for natural and random sequences of length 100. Sequence entropy is determined as Shannon-entropy of composition vectors.

an obvious shift in the mean entropy from 2.76 bits to 2.7 bits, reflecting a decrease in the information content of natural sequences through their composition. Below a threshold of 2.5bits, there are no more random sequences with a lower information content. There are still several natural 100mers with a lower complexity. In some cases this observation was associated with repeat-units.

Waiting distance

A tendency towards lower compositional entropy results into amino acid of the same kind *a* to locally accumulate. By necessity this also leads to regions of lower *a*-content.

Another way of interpreting this effect is that given an amino acid of type a the distance to the next a in the sequences tends to be smaller or larger than randomly expected. In this study, the distribution of this distance is being analyzed, referred to as *waiting distance* for all amino acids. The waiting distance is expected to be

$$W(d) = \lambda \cdot (1-\lambda)^{d-1} \tag{A.2}$$

where λ is the overall propensity of the respective amino acid. This expected waiting distance together with those derived from the natural data set and that from the data set biased by the overall amino acid composition is depicted in Figure A.6. The simulated waiting distance derived from the randomized data set is colored in red, which is perfectly overlapping with the expected waiting distance in black. The waiting distance in the natural data set, plotted in green, deviates from these random distributions. The distributions of those amino acids where the natural deviates the most from the expected distribution as presented here.
For all amino acids, shorter waiting distances are increased. This corresponds to the expectation of locally accumulated identical residues to result into short waiting distances. Long waiting distances are also increased in amino acid specific regions, generally around a distance of 100 residues. Intermediate distances are correspondingly under-represented. The most significant deviation is a frequency of the CxxC pattern

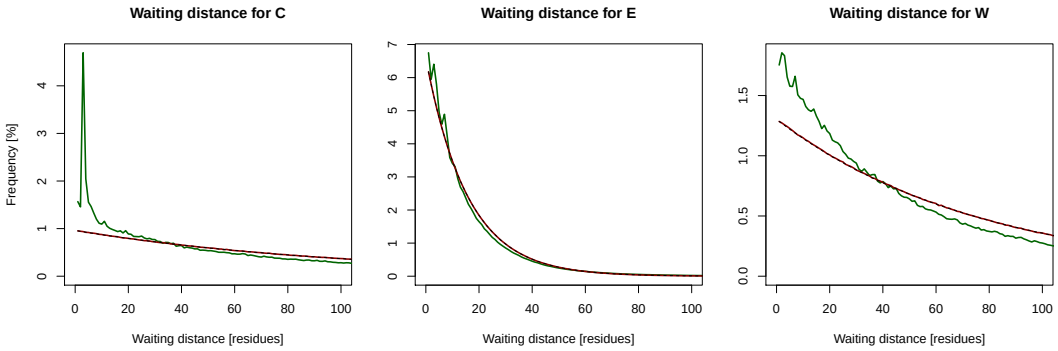

Figure A.6: Sequence bias amino acids with distance 1 of unrelated sequences.

with a frequency of more than 4.5% compared to the expected frequency of less than 1%. Short waiting distances of glutamate and tryptophan are also deviating significantly from the expected waiting distances. The roughed patter of the waiting distance distribution of glutamate also indicates its specific pattern, implying short-range correlations of which some seem to be slightly more preferred than others.

A.3 Sequence bias curated from local composition bias

Sequence bias on top of local composition

The study of the waiting distance already revealed that certain patterns of amino acids occur more often than others. However, the waiting distance does not reveal the frequency of specific patterns, irrespective of the residues in between. In order to overcome this dependency, the frequency of specific patterns was analyzed at different distances in sequence. Such patterns are known to exist and to reflect biophysical constraints of secondary structure elements. Generally, such biases can be derived by contrasting the observed, natural frequency to that expected from the overall amino acid composition. Here, the natural frequencies are contrasted to those of patterns derived from the A- and the P-model, thereby accounting for the compositional bias at different levels. The values of an over- or underrepresentation V of two amino acids a and b at a distance of d amino acids in sequence, is derived as the logarithm of the ratio between their frequency in the natural data set and the frequency in the respective model:

$$V(a,b,d) = \log_2 \left(\frac{f_{\text{natural}}(a,b,d)}{f_{\text{model}}(a,b,d)} \right) \tag{A.3}$$

In Figure A.7 these values are plotted for $d = 1$ and $d = 2$, corresponding to the two patterns XX and X.X. Notably, the deviation from the P-model is decreased compared to that of the A-model, especially in the case for correlations between residues at a distance of two or more amino acids in sequence. The derived relative frequencies after the normalization with the P-model are curated from composition bias on the protein level. Potentially, these frequencies reflect sequence preferences more accurately than those that are normalized by the composition over the entire data set.
A possible extension of this study would be to separate the observed effects by using different data sets. As the composition and sequence patterns in α-helices differ from those in β-sheets, performing the same study on all-α or all-β domains can separate the sequence effect arriving due to common sequence-structure relationships. The here presented biases represent an average over all the considered natural sequences.

Constraints on the DNA-level

A small-scale study to investigate the correlations between GC-content and amino acid composition was performed. For a set of DNA sequences derived from the *Escherichia coli* genome, the expected amino acid composition is calculated from its GC-content GC, assuming an equal distribution over all 61 amino acid-conding codons in the genetic code. In Equation A.5 the derivation of specific condon frequencies is described. It is corrected for the expected frequency of the three stop-codons α:

$$\alpha = 7 \cdot (1 - GC) \cdot 2 \cdot GC. \tag{A.4}$$

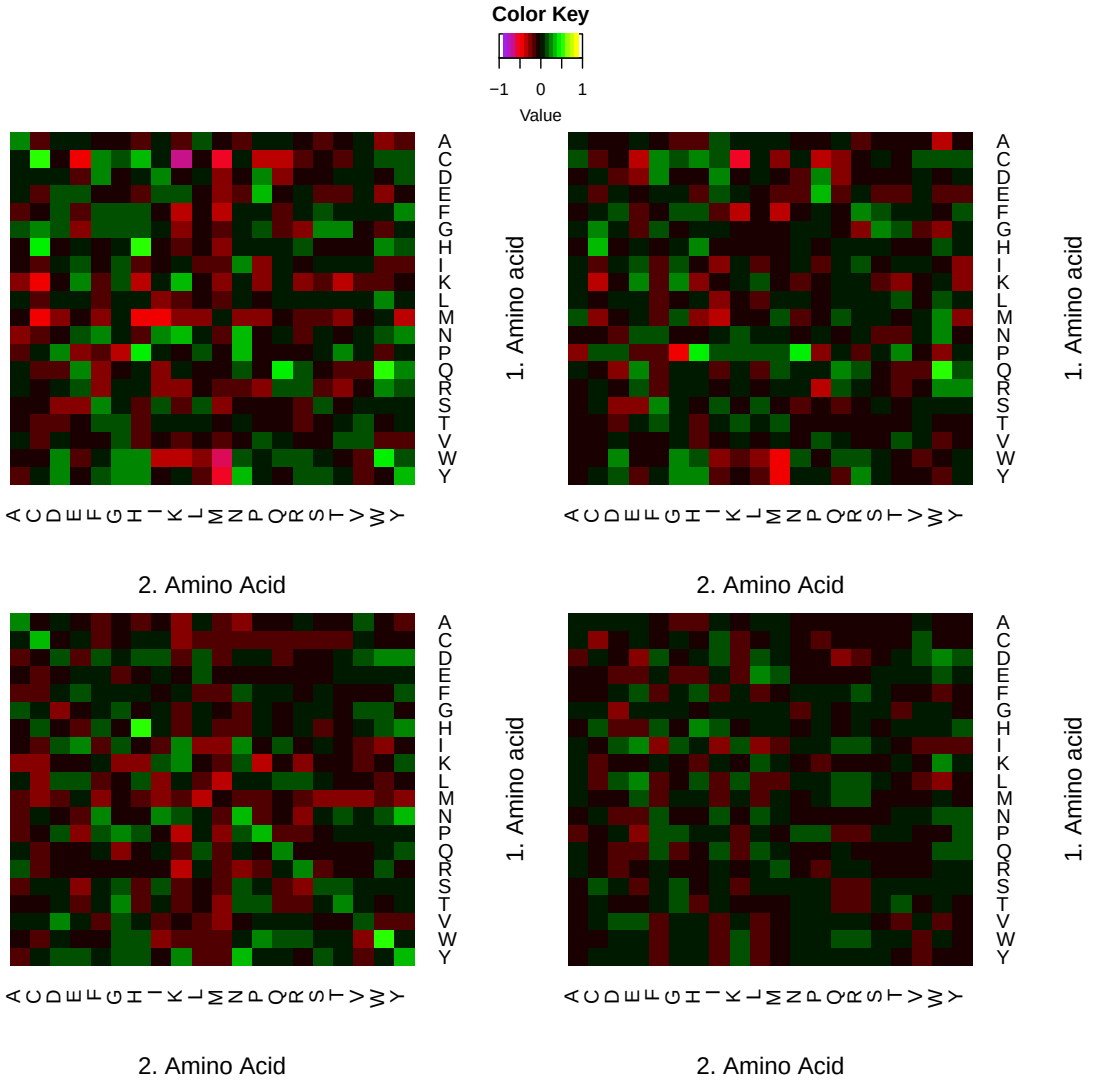

Figure A.7: Sequence bias of residues with distance 1 or 2 in unrelated sequences. The number of XX (upper row) and X.X (bottom row) patterns in our bacterial data set is derived for all possible amino acid combinations. This number is normalized by the respective pattern count derived from the A-model (left side) and the P-model (right side). Colors in red indicate a under-representation of patterns in the natural data, green an over-representation. In the relative frequencies normalized by the A-model, the diagonal displays a tendency towards green, indicating that the same amino acids like to re-occur in close proximity. Overall the deviations are therein more pronounced compared to those normalized by the P-model. At a distance of 2, the deviations converges towards an expected frequency.

$$f_{\text{closed}}(c) = (1+\alpha) \prod_{c_i \in c} \begin{cases} GC & \text{if } c \in \{G,C\} \\ (1-GC) & \text{if } c \in \{A,T\} \end{cases} \tag{A.5}$$

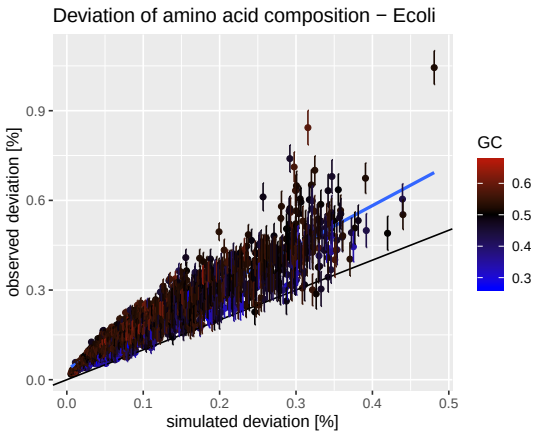

Figure A.8: Constraint amino acid composition by GC-content.

This closed form codon composition is contrasted with the observed natural codon composition and a shuffled version of the natural nucleotides, that did not contain stop codons. All compositions are translated into amino acid composition vectors and their deviation as defined by the L1-norm is derived. On the x-axis the deviation of the mean simulated composition from the closed form is plotted; on the y-axis the deviation between the natural and the closed form along with the variance derived from the simulated composition, indicating the error due to the finite sampling of the natural composition.
The diagonal black line indicates a perfect correlation between the observed and simulated deviation. However, almost all observed deviations are greater than the simulated. This trend indicates, that there are constraints acting on the codon usage, which result into an amino acid composition that is deviating from the expected value. Such effects have previously been associated to codon bias.

Appendix B

Search of a transition

One of the main goals of this thesis was motivated by the hypothesis of an evolution of proteins from a set of small, ancient peptides and the idea that sequence space is being populated accordingly (see Section 1.3.1).
In Section B.3 the progression of sequence space occupation as a function of fragment length is investigated, aiming to detect a transition at a certain peptide length related to the the length of ancient peptides. The results closely relate to those in Section 3.1, as the progression of the transition between expected and unexpected similarities progresses with fragment length. More specifically, it demonstrates how this transition proceeds for different level of sequence similarity.
Another approach to describe local coherences between natural sequences is taken by clustering them according to similarity. A large-scale clustering approach is presented in Section B.1, where the bacterial data set comprising 10^9 amino acids has been successfully clustered into connected components via single point mutation distances. This approach has been applied to fragments of all lengths up to 100 residues, allowing to compared clustering results across fragment length.

B.1 Large-scale cluster analysis

Cluster analysis is a common approach to study local coherences between sequences. A clustering that is unexpected from random sequences indicates more densely populated areas of sequence space, that may be important in the context of natural sequences. With the evolution of proteins from a set of small, ancient peptides in mind, one of the main goals of this study was to identify a transition in the sequence space occupation at a certain sequence length. In order to characterize these regions as well as the clustering behavior, I developed and implemented a large-scale clustering algorithm. Here, I present the strategies and algorithms used for this clustering and the analysis of the main data set of bacterial genomes used in this thesis.

Strategies to reduce complexity

An accurate clustering requires calculating in principle all pairwise similarities between fragments. The main data set of bacterial genomes is composed of $\approx 10^9$ sequence fragments and thus 10^{17} sequence pairs need to be compared. This huge amount of comparisons required to reduce the complexity of any clustering attempt severely.
The first reduction of this complexity was achieved by deriving the set of diverse fragments and to cluster these. After the generation of connected components, each diverse fragment was assigned the number of its occurrences in the data set indicating its prevalence in the data set. The number of diverse fragments grows exponentially with fragment length and is thus mostly beneficial in cases of an exhaustive sampling of sequence space, thus for short fragments.
The next applied strategy to overcome this great complexity was to use a computationally cheap way to compare sequences in the first place. Thereby, the position-wise mismatches of fragments of the same length was computed, which corresponds to a Hamming distance of ungapped sequences. Sequences with one mismatch were group together to obtain connected components among all the observed sequences. The principle of connecting sequences with one point mutation is often used in geno-phenotype cluster maps as it reflects an often accepted mutation of natural sequences. A connected component therefore comprises sequences that are all inter-reachable via intermediate sequences through a path of consecutive point mutations.
Furthermore, the computation was paralleled using 50 CPUs units at once, reducing computation time to couple of days for each calculation of fragments longer than 20 residues. Parallelization was, however, applied differently depending on the expected size of the calculated components. While for short fragment lengths (up to 12 residues) the great majority of fragments all belong to one connected component, long fragments mostly belong to individual components. An optimized implementation for both these cases emerged to be the best strategy to reduce the computing time. The connected component containing the poly-A fragment was the largest component in a reproducible way. Thus, parallelization across the poly-A component was applied, using all cores to calculate this cluster, while for the remaining clusters only one core were utilized.

Algorithm to calculate the large components

Connected components among sequence fragments of length w, are generated by grouping together fragments with a point mutation distance of 1. For a given w, the set of pairwise distinct fragments of the whole data set was extracted and clustered, as the actual number of identical fragments is irrelevant for the construction of connected components.

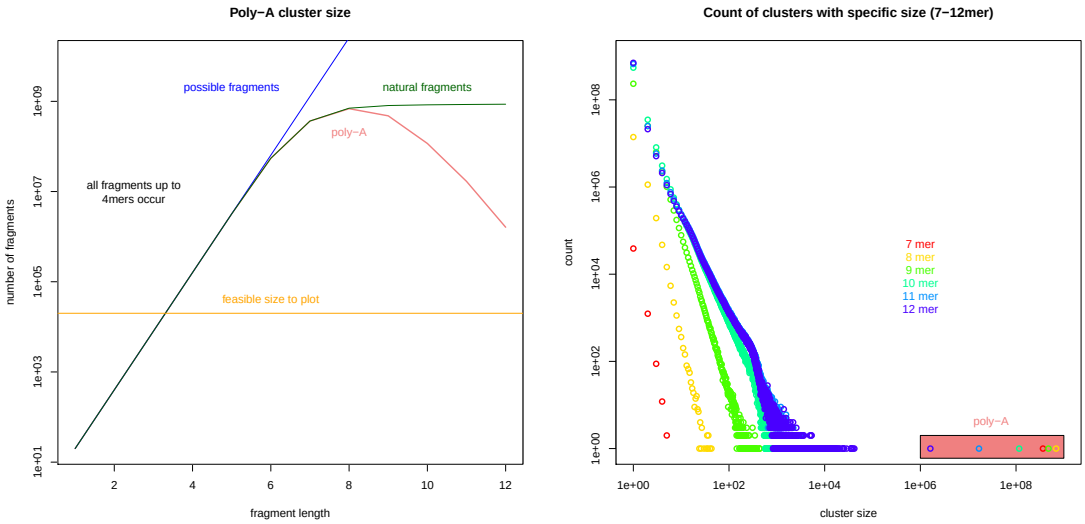

Figure B.1: Size of the poly-A cluster.

Compute neighbors of active fragments

An array was initialized, referred to as C with incrementing values, that has a length defined by the number of distinct fragments such that $C(i) = i$. Each entry $C(i)$ corresponds to the fragment at the index i in a list of fragments with lexicographical order and indicates its associated connected component. In the beginning, every fragment is assigned to its own connected component, assuming it is a singleton.
The computation of the poly-A cluster starts with one fragment, the first entry of the diverse fragments list, which is the poly-A fragment. This fragment is assigned as the active fragment and one thread scans the set of diverse fragments for existing point mutation neighbors. All point mutation neighbors at indices n are collected, $C(n)$ is set to 0 and the set of active fragments is defined by the set of point mutation neighbors.
In the second and all following iterations the active set of fragments is split among the available threads. For each active fragment in a thread, its point mutation neighbors n are derived. If $C(n)$ is already set to 0, the fragment at index n is neglected, as it has been assigned to the cluster already, thereby avoiding an infinite loop. Only those neighboring fragments are assigned to be the next active fragments that have not been assigned to the poly-A cluster yet. After each thread has processed its share of active fragments, the threads synchronize by joining all valid (possibly overlapping) neighbors to the set of active fragments and the next iteration starts. In the case where the active set is empty past synchronization, the computation of the poly-A cluster terminates.
In principle this algorithm can be applied to any starting fragment in place of the poly-A fragment. However, the strategy of applying this algorithm generally to all fragments fails, as starting with fragment length 12, most clusters only comprise few fragments.

In each iteration, only few active fragments exist, leading to many idle threads. For all other connected components, a different strategy was applied.

Efficient parallel computation of connected components

The challenge in constructing an algorithm that calculates all other components efficiently in parallel is that it is essentially unclear which fragments belong together in the beginning. Starting with arbitrary fragments and collecting their neighbors iteratively may sooner or later lead to overlaps of the components. Thus a synchronization between the threads that started with distinct fragments is be necessary. I implemented multiple versions of such frequent synchronization strategies but they all did not perform well. This was simply because the sizes of the currently computed connected component differed largely, leading to many threads waiting for the slowest. Also a mixed strategy that applies the previously described algorithm of multiple threads computing one component were not performant.
Instead, a strategy that did not require synchronization between threads emerged to be sufficiently performant for the purpose of constructing connected components in parallel. This strategy requires a deterministic assignment of all connected components to a specific thread. In principle each possible connected component of the fragment at index i was assigned to the thread with id $i\%t$, where t is the number of available threads and all threads possessed an unique id ranging from 0 to $t-1$. Furthermore, for all fragments at indices i and j with $i < j$ are in the same component, the calculation of their connected component is assigned to thread $i\%t$. This guarantees that for each connected component the thread is assigned according to the lexicographically smallest fragment.

Each thread, starts with the fragment at index of its own thread id and proceeds in steps of size t to the next fragments until it reaches the end of the list of diverse fragments. In the case that the fragment at index i has not been assigned yet ($C(i) = i$), the tread assumes, that it is its job to compute the connected component of this fragment and proceeds with the iterative collection of point mutation neighbors of the corresponding fragments. If any neighbor n occurs with $n < i$, the thread recognizes, that calculating the connected component is the responsibility of a different thread. Instead of synchronizing with the responsible thread, all calculations are simply dismissed and the thread continues with the computation of the connected component of the next fragment. In case of having computed a whole connected component, starting from the lexicographically smallest fragment, the thread assigns the value i to all collected neighbors to the C array. In this moment, other threads are able to recognize, that the corresponding fragments are already assigned to a connected component.
With this approach, the partial computation of connected components is frequently being dismissed, allowing threads to communicate only passively over the shared array C without the need to synchronize. However, the rejection of a partial computation does not occur as often as its acceptance. That is because all threads commonly start with frag-

ments of a low index. In the ideal case that all threads move on to the next fragments, the only possible overlap that can occur is that they compute simultaneously the same connected component. For the thread with the largest id, there is a possible overlap with $t-1$ fragments of in total 10^9 fragments (for fragment length greater than 10 residues). In the case that one thread processes the fragment at a index l, which is significantly higher that of the thread with the next highest index $s < l$, the likelihood of computing a connected component that is not assigned to it increases. That is because the number of unassigned fragments between this leading thread and those behind is at least $l-s$ and possibly large. The more probable rejection of partially computed connected components leads to a deceleration of the leading threads. In contrast, the thread with the most workload (processing the fragment at smallest index amongst all threads) will not reject its calculation, as all previous fragments are already assigned. (The only exception is a case, when after starting its calculation at index i, a different thread of a previous fragments at index $j < i$ assigns this fragment (at index i) to its own connected component ($C(i) = j$) and moves on before thread i recognizes that its fragment has been assigned itself.) This principle leads to a self-regulating effect of threads and they will finish around the same time.

Connected component size distribution

The connected components possess different sizes depending on the chosen length of clustered fragments. Up to a fragment length of 8 residues, most fragments belong to the poly-A cluster (Figure B.1: A) with few smaller connected components. For longer fragments, the poly-A cluster starts to fall apart and the component size distribution of up to 12mers indicates an increase in the number of large cluster (Figure B.1: B). As soon as the poly-A cluster achieves a size comparable with all other clusters (around 15mers), the connected components size distributions resemble each other (Figure B.2). They possess a power-law shape, where the exponent notably increases with fragment length. This steady decay can be captured by comparing the number of clusters of a specific size along an increasing fragment length. According to Figure B.3, the number of small clusters decays slower than the number of large components. In the case of cluster size 10, the number of clusters decays by 5.4% with incrementing fragment length. For a cluster size of 90, this decay in the number of clusters increased to 11%. For larger clusters, the value becomes more noisy. It appears that the decay of the connected component sizes display a continuous behavior, irrespective of the chosen fragment length.
This decay is, however, unexpected in the connected components of A-model sequences. Therein, the number of clusters with three sequences decays by 97% going from 11 to 13mers and by 83% of clusters with two sequences. Thus, natural connected components are much denser and only break down very slowly with increasing fragment length.

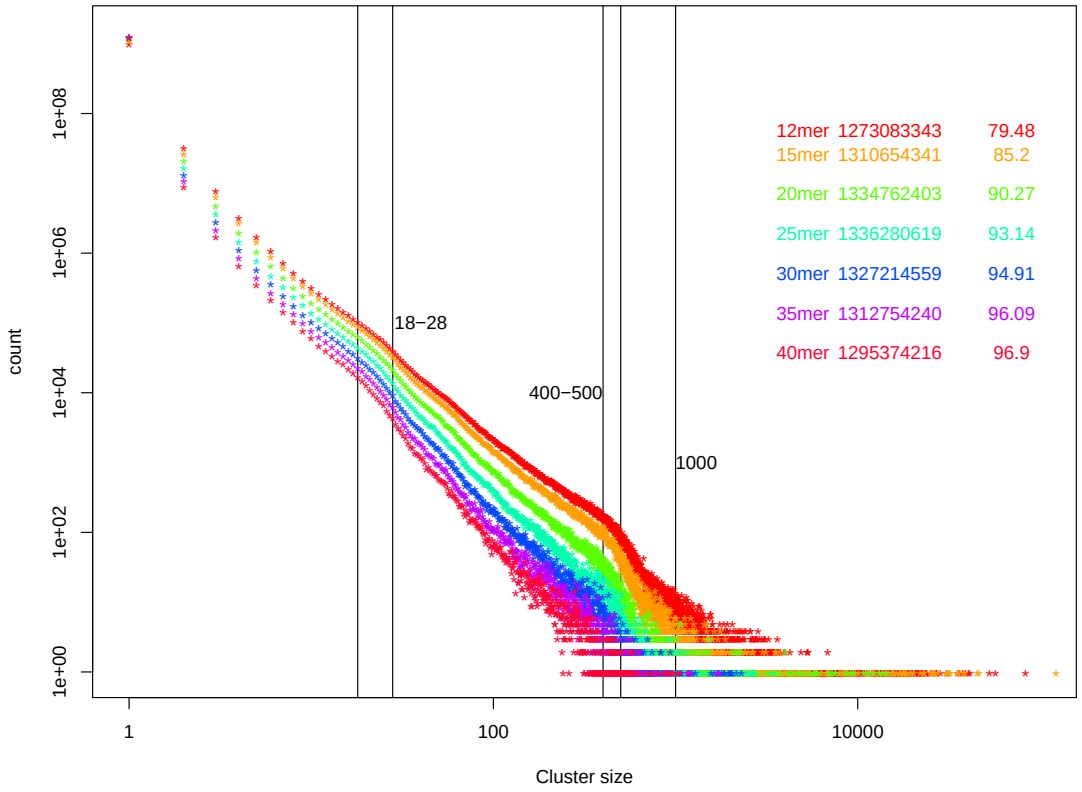

Figure B.2: Connected component size distribution.

B.2 Word count analysis

Similar to other approaches that focus on the frequency of identical sequence, I derived the word count distribution for different sequence lengths and plotted those of 10mers or longer in Figure B.4. Similarly to the cluster size distributions, these also possess a power-law shape. For each frequency of occurrence, the number of sequences decreases slowly with increasing fragment length. This implies that there are many and long sequences in the natural data set that are reoccurring.

Curation of homology

The reuse of identical sequnces is often related to homologous descent. The word count frequencies is likely to be biased by homologous sequences that are over-represented in the data set. In order to estimate how much homology contributes to the observed over-representation of some peptides, I curated the frequencies according to the number of homologous clusters they come from. This approach was performed on 5-8mers, of which the results of 7mers are presented.

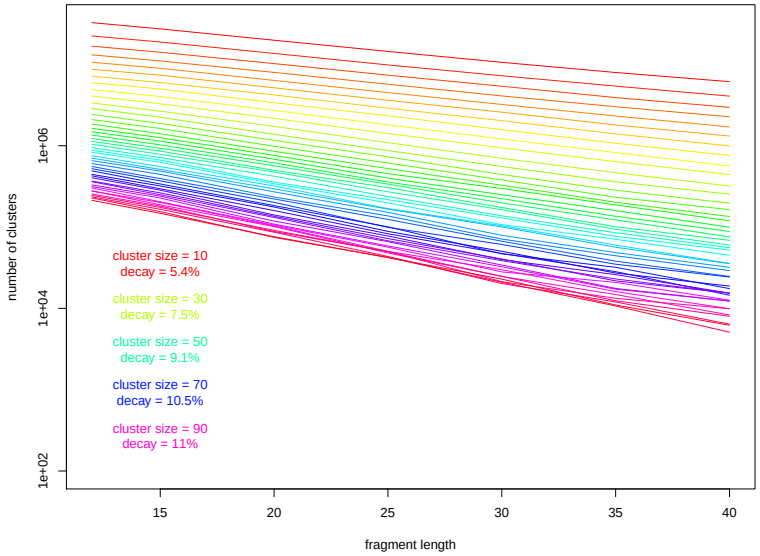

Figure B.3: Logarithmic decay of cluster frequency.

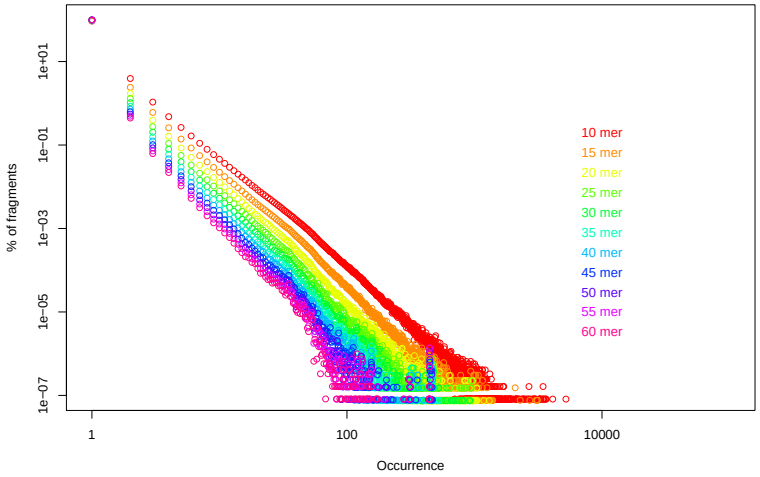

Figure B.4: Word count analysis.

In Figure B.5 the natural abundance of specific 7mers is plotted against the expected abundance derived from the closed form A-model Section 1.2.1 in green circles. The perfect correlation is plotted as a blue line. Obviously, there is a great amount of over-represented fragments above the blue line, which I expected to reduce after removing identical 7mers from homologous regions.
For each 7mer, I collected their original context by extracting additional 20 amino acids at each sides from their original protein sequence, resulting in a set of 47mers. In cases

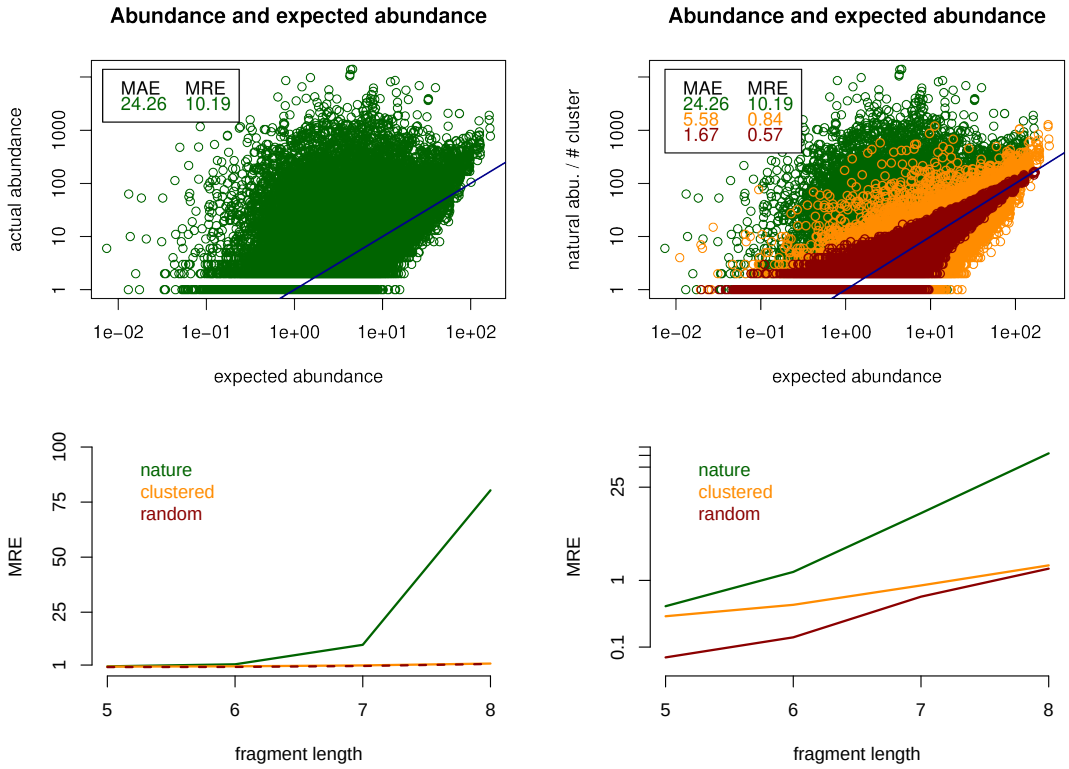

Figure B.5: Homology curated 7mer frequency.

where the 7mer occurred in the beginning or end of a sequence and their context was shorter than 20 amino acids and shorter sequences were retrieved. The final set of all embedded 7mers was clustered using the tool cd-hit [Li et al., 2001] with standard parameters. The number of clusters was then assigned to the 7mer. Instead of plotting the 7mer's occurrence against their expected occurrence, the number of clusters was plotted against the expected occurrence (Figure B.5: B, orange circles). The correlations between cluster count and expected abundance are notably closer to the ideal correlation, indicated by the blue line. Thus, a great amount of over-representation could be removed by accounting only for identical peptides in different contexts.

A certain amount of error is simply expected due to the finite sampling problem. To check its impact, the deviation of sampled (A-model) random sequences was derived an plotted in red. For sequences with an expected low abundance, corresponding to the bottom left area of the plot, the deviation from the ideal correlation was indeed larger than for sequences with an expected high abundance. For an infinite amount of data, this error would converge to zero and the red circles would approach the blue line.

The deviation was captured numerically by two metrics corresponding to the mean ab-

solute error and mean relative error, where the error for each fragment f is defined as:

$$AE(f) = |y - A(f)|$$

$$RE(f) = \left| \frac{y - A(f)}{A(f)} \right|$$

. The absolute error reflects how many more (or fewer) instances of a specific fragment occurred in the natural data set. The relative error additionally normalizes this effect by the expected abundance. The mean errors MAE and MRE are derived over the set of all possible fragments, also those that are not occurring once in the natural data set. For different fragment lengths, the errors obtains different values (Figure B.5: C, D). As the absolute error is deviating significantly more than the relative error, the mean relative error is a more conservative measure to indicate deviations. With fragment length, the MRE of the natural data set increases significantly to a value of more than 75% for 8mers. The MRE of the clustered peptides increases with about the same pace as the error of the sampled random sequences, reaching 1% for 8mers (Figure B.5: D).

B.3 Progression of neighborhood with sequence length

The theory of restricted sequence space occupation

One of the initial targets of this thesis was to determine, if there is a change in the way sequence space is occupied for a certain fragment length. Research focusing around such a transition point was motivated by the theory, that natural proteins have emerged from a small subset of primordial peptides [Alva et al., 2015]. The concept of this theory is explicitly illustrated in Figure 1.9 and will shortly be repeated in this context. This theory suggests that evolution had time to explore sequence space up to a certain peptide length in an exhaustive manner. Of the emerged relatively long peptides, those gained acceptance that could fulfill essential functions such as DNA- and RNA-binding. Due to accretion, the selected fragments become longer and evolve to domains and proteins. As the primordial peptides have a limited size of approximately 20 amino acids, it is standing to reason, that sequence space of peptides longer than 20 has been explored being restricted to these peptides.
This concept is clearly visible on the structural level, where no domain is found to be made of multiple such primordial peptides. Also the limited number of found domain structures, for example 2000 ECOD X-groups (see Subsection 1.2.3), hints to the fact that nature is not exploring sequence space in an exhaustive manner anymore.
In this study, I aim to find a transition in the sequence space occupation by natural sequences the may be suggestive of a limited number of primordial peptides and protein folds.

Approach

For this, I use an approach that describes the occupation of the sequence space by capturing the relative occupation of the local neighborhoods of existing natural sequences to an expected occupation. The significance of similarities between peptides is derived as a function of sequence length. The previously discussed transition point of different sequence space occupation should become visible when interpreting significance of similarities as a function of sequence length.
I use the term n-point mutation neighborhood for all neighbors that are n point mutations away from a specific sequence. This corresponds to the point mutation distance of ungapped sequences, hence the so termed Hamming distance. Different natural data sets are used with their respective random data set of the A-model. The distance distributions using the Hamming distance are generated for the natural and random data for all data sets to derive the frequency of how often an existing fragment possesses neighbor at a certain distance. A transition in the occupation of sequence space, should be imprinted into the neighborhood occupation at a certain sequence length.

Enrichment factor

In order to compare the naturally observed neighborhood with that of a random sequence model, I calculate the ratio between the natural and the random distance distributions at each point mutation distance d for all fragment lengths l. In the following this ratio is termed as the *enrichment factor* EF.

$$EF(d,l) = \frac{DD_{\text{nat}}(d,l)}{DD_{\text{rand}}(d,l)} \tag{B.1}$$

For point mutation distances below 6, the enrichment factor $EF_d(l)$ is plotted as a function of fragment length for a specific point mutation distance d. As depicted in Figure B.6 the enrichment factors $EF_d(l)$ do not increase significantly for small fragment lengths. It starts to increase with a double-exponent behavior and transitions into an exponential behavior after a length of roughly five more residues.
For all possible transition points on the fragment length axis, a double-exponential function was parametrized to fit the first part up to the defined transition point and a exponential function to fit the remaining part optimally to each specific $EF_d(l)$. Those fits were picked that deviate least and plot the double-exponential function with dots and the exponential function as dashes. The used parameters for the fits are indicated in colors corresponding to d where x stands for fragment length.

Interpretation of the change in the enrichment factor

The observed behavior of the enrichment factor with sequence length seems arbitrary and is not-intuitive at first. However, the transition from a horizontal line at zero can be ra-

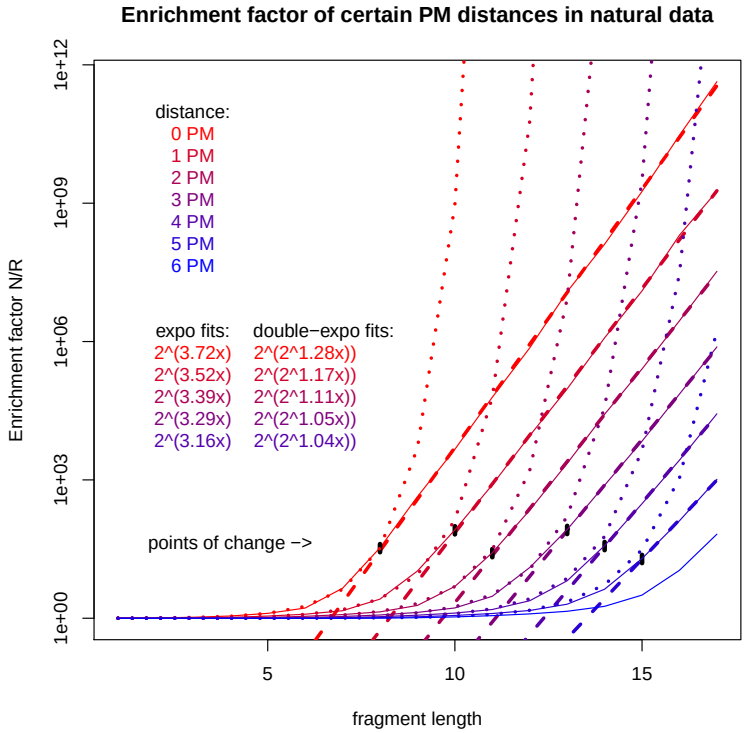

Figure B.6: Enrichment factor of specific point mutation distances. Applied on the bacterial data set.

tionalized with the transition from randomly expected distances to unexpected distances. Such a transition marks the twilight zone in the distance distributions as a function of point mutation distance for a static length, which is studied in Section 3.3.
Identical 8mers (point mutation distance of 0) are almost unexpected to occur in the random data set (compare to Figure 3.9, left-most green circle corresponds to 8mers). 10mers with a 1-point mutation distance are similarly unexpected. These points map to the region before the twilight zone, of the long tail of presumably homologous similarities (see Figure 2.2: B). Hence, such over-represented distances between natural fragments are probably of homologous origin.
Foremost, the steady increase of the enrichment factor can be interpreted as a reflection for a continuous over-representation of natural point mutation neighborhoods across fragment length and increasing point mutation distance. There is no indication of a harsh transition or a change in this increase. Such a behavior has also been reported in the decay of connected component sizes in Section B.1. This may point to the fact that natural sequence radiate from important hubs in sequence space (premordial peptides or protein family motif) through duplication and diversification and that these hub are located rather

arbitrary to each other in sequence space.
However, this conclusion is strictly dependent on the method, which only can detect changes in the relatively close-by neighborhood of sequence space. Sequences that have diverged beyond the level of significant distances (twilight zone) are out of reach as they are indistinguishable from a randomly expected distances.

Data size dependency

In order to test the results for different data sizes, the same approach of deriving the enrichment factor was applied to a 400th of the original bacterial data set. In order to sample equally over all bacterial genomes, every 400th protein was considered. This procedure resulted in to data set of $2.5 \cdot 10^6$ residues, relative to the original data set containing 10^9 residues. The enrichment factors are depicted in B.7, which are almost identical to those of the foll data set in Figure B.6.

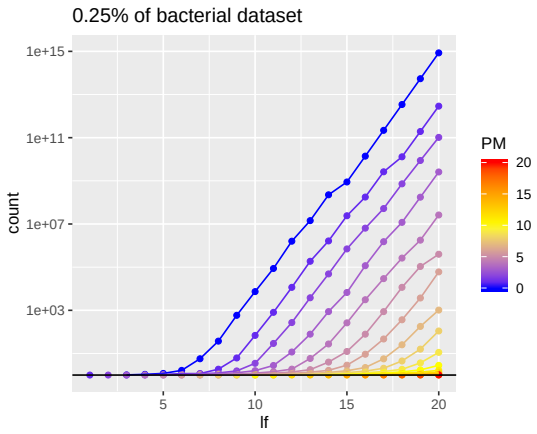

Figure B.7: Data size dependencies effect the enrichment factor.

Appendix C

Curriculum vitae

2010 Abitur at the LGH in Schwäbisch Gmünd

2013 Bachelor of Science at the University of Tübingen
Methods for differential data analysis and their application to HLA ligandome data from tumor and benign tissue

2015 Master of Science at the University of Hamburg
Filtering Method for Protein Structure Similarity

2020 Doctoral studies at the Max Planck Institute for Developmental Biology in Tübingen
The Sequence Space of Natural Proteins

Abbreviations

PDB	Protein Data Bank
PDBID	Protein Data Bank identifier
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
mRNA	Messenger Ribonucleic acid
tRNA	Transfer Ribonucleic acid
SNP	Single-nucleotide polymorphism
InDel	Insertion and deletion
ECOD	Evolutionary classification of protein domains database
Pfam	Protein family database
SCOP	Structural classification of proteins database
HMM	Hidden Markov Model
PCC	Pearson correlation coefficient
SMART	Simple Modular Architecture Research Tool
NCBI	National Center for Biotechnology Information
LCR	Low-complexity region
RMSD	Root-mean-square deviation
BLAST	Basic Local Alignment Search Tool
BLOSUM	BLOcks SUbstitution Matrix
PDF	Probability density function
PR	Percentile rank
PRD	Percentile rank distribution
TKS	Kolmogorov–Smirnov statistic

List of Tables

List of Figures

Bibliography

J. Aguirre, P. Catalán, J. A. Cuesta, and S. Manrubia. On the networked architecture of genotype spaces and its critical effects on molecular evolution. *Open Biology*, 8(7), 2018.

H. Akashi and T. Gojobori. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proceedings of the National Academy of Sciences of the United States of America U. S. A.*, 99(6):3695–3700, 2002.

B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*, chapter 6. Garland Science, 2002.

P. A. Alexander, Y. He, Y. Chen, J. Orban, and P. N. Bryan. The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proceedings of the National Academy of Sciences of the United States of America U. S. A.*, 104(29):11963–11968, 2007.

S. F. Altschul and B. W. Erickson. Optimal sequence alignment using affine gap costs. *Bulletin of Mathematical Biology*, 48(5):603 – 616, 1986.

V. Alva, M. Remmert, A. Biegert, A. N. Lupas, and J. Söding. A galaxy of folds. *Protein Science*, 19(1), 2009.

V. Alva, J. Söding, and A. N. Lupas. A vocabulary of ancient peptides at the origin of folded proteins. *eLife*, 4, 2015.

G. Apic, J. Gough, and S. A. Teichmann. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology*, 310(2):311–325, 2001.

R. Apweiler. The universal protein resource (UniProt) in 2010. *Nucleic Acids Research*, 38:D190–5, 2009.

U. Bastolla, Y. Dehouck, and J. Echave. What evolution tells us about protein physics, and protein physics tells us about evolution, 2017.

S. Bershtein, A. W. Serohijos, and E. I. Shakhnovich. Bridging the physical scales in evolutionary biology: from protein sequence space to fitness of organisms and populations. *Current Opinion in Structural Biology*, 42:31–40, 2017.

T. R. Bhangale, M. J. Rieder, R. J. Livingston, and D. A. Nickerson. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Human Molecular Genetics*, 14(1):59–69, 2005.

J. A. Birdsell. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Molecular Biology and Evolution*, 19

(7):1181–1197, 2002.

T. Bitard-Feildel, M. Heberlein, E. Bornberg-Bauer, and I. Callebaut. Detection of orphan domains in Drosophila using "hydrophobic cluster analysis". *Biochimie*, 119: 244–253, 2015.

E. Bornberg-Bauer. How are model protein structures distributed in sequence space? *Biophysical Journal*, 73(5):2393–2403, 1997.

P. C. Buchholz, C. Zeil, and J. Pleiss. The scale-free nature of protein sequence space. *PLoS ONE*, 13(8):1–14, 2018.

CATH database. *Latest Release Statistics*, accessed 2020-01-19. URL `https://www.cathdb.info/`.

S. Cebrat and M. R. Dudek. The effect of DNA phase structure on DNA walks. *The European Physical Journal B*, 3(2):271–276, 1998.

J. Cedano, P. Aloy, J. A. Perez-Pons, and E. Querol. Relation between amino acid composition and cellular location of proteins. *Journal of Molecular Biology*, 266(3):594–600, 1997.

H. Cheng, R. D. Schaeffer, Y. Liao, L. N. Kinch, J. Pei, S. Shi, B.-H. Kim, and N. V. Grishin. ECOD: An Evolutionary Classification of Protein Domains. *PLoS Computational Biology*, 10(12):e1003926, 2014.

C. Chothia and J. Janin. Principles of protein-protein recognition. *Nature*, 256:705–708, 1975.

C. Chothia and A. Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, 5(4):823–826, 1986.

C. Chothia, J. Gough, C. Vogel, and S. A. Teichmann. Evolution of the Protein Repertoire. *Science*, 300(5626):1701–1703, 2003.

K.-C. Chou. Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins: Structure, Function, and Bioinformatics*, 43(3):246–255, 2001.

K.-C. Chou and C.-T. Zhang. Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology*, 30(4):275–349, 1995.

P. Dasmeh, A. W. Serohijos, K. P. Kepp, and E. I. Shakhnovich. Positively Selected Sites in Cetacean Myoglobins Contribute to Protein Stability. *PloS Computational Biology*, 9(3):1–12, 2013.

D. de Lucrezia, D. Slanzi, I. Poli, F. Polticelli, and G. Minervini. Do natural proteins differ from random sequences polypeptides? natural vs. random proteins classification using an evolutionary neural network. *PLoS ONE*, 7(5):e36634, 2012.

E. J. Deeds, N. V. Dokholyan, and E. I. Shakhnovich. Protein Evolution within a Structural Space. *Biophysical Journal*, 85(5):2962–2972, 2003.

P. J. Diggle. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC press, 2014.

K. A. Dill. Theory for the Folding and Stability of Globular Proteins. *Biochemistry*, 24 (6):1501–1509, 1985.

N. V. Dokholyan, B. Shakhnovich, and E. I. Shakhnovich. Expanding protein universe and its origin from the biological Big Bang. *Proceedings of the National Academy of Sciences of the United States of America U. S. A.*, 99(22):14132–14136, 2002.

M. dos Reis, L. Wernisch, and R. Savva. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. *Nucleic Acids Research*, 31(23):6976–6985, 2003.

M. dos Reis, R. Savva, and L. Wernisch. Solving the riddle of codon usage preferences: A test for translational selection. *Nucleic Acids Research*, 32(17):5036–5044, 2004.

Z. Dosztányi, V. Csizmók, P. Tompa, and I. Simon. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *Journal of Molecular Biology*, 347(4):827–839, 2005.

I. Dubchak, I. Muchnik, S. R. Holbrook, and S.-H. Kim. Prediction of protein folding class using global description of amino acid sequence. *Proceedings of the National Academy of Sciences*, 92(19):8700–8704, 1995.

G. Duranton and H. G. Overman. Testing for Localization Using Micro-Geographic Data. *The Review of Economic Studies*, 72(4):1077–1106, 2005.

M. Elgamacy, M. Coles, P. Ernst, H. Zhu, M. D. Hartmann, A. Plückthun, and A. N. Lupas. An Interface-Driven Design Strategy Yields a Novel, Corrugated Protein Architecture. *ACS Synthetic Biology*, 7(9):2226–2235, 2018.

EMBL-EBI. *Pfam*, accessed 2020-01-19. URL `http://pfam.xfam.org/`.

P. Erdős and A. Rényi. On the evolution of random graphs. *Théorie de l'information*, 5 (1):17–60, 1960.

A. C. Fisher, M. A. Rocco, and M. P. Delisa. *Genetic selection of solubility-enhanced proteins using the twin-arginine translocation system*, volume 705. 2011.

P. J. Fleming and F. M. Richards. Protein packing: Dependence on protein size, secondary structure and amino acid composition. *Journal of Molecular Biology*, 299(2): 487–498, 2000.

P. L. Foster. Stress responses and genetic variation in bacteria, 2005.

M. W. Franklin, S. Nepomnyachiy, R. Feehan, N. Ben-Tal, R. Kolodny, and J. S. Slusky. Efflux pumps represent possible evolutionary convergence onto the β-barrel fold. *Structure*, 26(9):1266 – 1274.e2, 2018.

S. Fukuchi and K. Nishikawa. Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. *Journal of Molecular Biology*, 309(4):835–843, 2001.

S. Fukuchi, K. Yoshimune, M. Wakayama, M. Moriguchi, and K. Nishikawa. Unique amino acid composition of proteins in halophilic bacteria. *Journal of Molecular Biol-*

ogy, 327(2):347–357, 2003.

C. Gaboriaud, V. Bissery, T. Benchetrit, and J. Mornon. Hydrophobic cluster analysis: An efficient new way to compare and analyse amino acid sequences. *FEBS Letters*, 224(1):149–155, 1987.

A. Goncearenco and I. N. Berezovsky. The fundamental tradeoff in genomes and proteomes of prokaryotes established by the genetic code, codon entropy, and physics of nucleic acids and proteins. *Biology Direct*, 9:29, 2014.

J. M. Goodenbour and T. Pan. Diversity of tRNA genes in eukaryotes. *Nucleic Acids Research*, 34(21):6137–6146, 2006.

G. Grigoryan and W. F. Degrado. Probing designability via a generalized model of helical bundle geometry. *Journal of Molecular Biology*, 405(4):1079–1100, 2011.

Grishin lab. *Evolutionary Classification of Protein Domains*, accessed 2020-01-19. URL `http://prodata.swmed.edu/ecod/`.

J. J. Grzymski and A. G. Marsh. Protein languages differ depending on microorganism lifestyle. *PLoS One*, 9(5):1–12, 2014.

M. J. Harms and J. W. Thornton. Evolutionary biochemistry: Revealing the historical and physical causes of protein properties. *Nature Reviews Genetics.*, 14(8):559–571, 2013.

M. J. Harms and J. W. Thornton. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature*, 512(7513):203–207, 2014.

S. Henikoff and J. G. Henikoff. Amino acid substitution matrices. *Advances in Protein Chemistry*, 54(November):73–97, 2000.

F. Hia, S. F. Yang, Y. Shichino, M. Yoshinaga, Y. Murakawa, A. Vandenbon, A. Fukao, T. Fujiwara, M. Landthaler, T. Natsume, S. Adachi, S. Iwasaki, and O. Takeuchi. Codon bias confers stability to human mRNA s . *EMBO reports*, pages 1–19, 2019.

P.-S. Huang, S. E. Boyken, and D. Baker. The coming of age of de novo protein design. *Nature*, 537(7620):320–7, 2016.

F. Jacob. Evolution and Tinkering. *Science*, 196:1161–1166, 1977.

C. Kamp and S. Bornholdt. Coevolution of quasispecies: B-cell mutation rates maximize viral error catastrophes. *Physical Review Letters*, 88:068104, 2002.

S. Karlin, A. M. Campbell, and J. Mrázek. Comparative Dna Analysis Across Diverse Genomes. *Annual Review of Genetics*, 32(1):185–225, 1998.

M. Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1983.

D. A. Kondrashov and F. A. Kondrashov. Topological features of rugged fitness landscapes in sequence space. *Trends in Genetics*, 31(1):24–33, 2015.

E. V. Koonin, R. L. Tatusov, and K. E. Rudd. Sequence similarity analysis of Escherichia coli proteins: Functional and evolutionary implications. *Proceedings of the National*

Academy of Sciences of the United States of America U. S. A., 92(25):11921–11925, 1995.

E. V. Koonin, Y. I. Wolf, and G. P. Karev. The structure of the protein universe and genome evolution. *Nature*, 420:1–6, 2002.

A. Krause. The SYSTERS protein sequence cluster set. *Nucleic Acids Research*, 28(1): 270–272, 2000.

A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, and J. Moult. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics.*, 86(September):7–15, 2018.

S. K. Kummerfeld and S. A. Teichmann. Relative rates of gene fusion and fission in multi-domain proteins, 2005.

D. T. Lavelle and W. R. Pearson. Globally, unrelated protein sequences appear random. *Bioinformatics*, 26(3):310–318, 2009.

S. Lee, B. C. Lee, and D. Kim. Prediction of protein secondary structure content using amino acid composition and evolutionary information. *Proteins: Structure, Function, and Bioinformatics*, 62(4):1107–1114, 2006.

Y. Lesecque, D. Mouchiroud, and L. Duret. GC-biased gene conversion in yeast is specifically associated with crossovers: Molecular mechanisms and evolutionary significance. *Molecular Biology and Evolution*, 30(6):1409–1419, 2013.

I. Letunic and P. Bork. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Research*, 46(D1):D493–D496, 2018.

W. Li, L. Jaroszewski, and A. Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases . *Bioinformatics*, 17(3):282–283, 2001.

J. Lightfield, N. R. Fram, and B. Ely. Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. *PLoS One*, 6(3), 2011.

J. R. Lobry and C. Gautier. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 escherichia coli chromosome-encoded genes. *Nucleic Acids Research*, 22(15):3174–3180, 1994.

P. Luigi Luisi. Contingency and determinism. *Philosophical Transactions of the Royal Society A*, 361(1807):1141–1147, 2003.

A. Lupas and K. Koretke. *Evolution of Protein Folds*, pages 131–151. 2008.

A. N. Lupas, C. P. Ponting, and R. B. Russell. On the Evolution of Protein FOlds: Are Similar Motifs in Different Protein Folds the Result og Convergence, Insertion, or Relics of an Ancient Peptide World? *Journal of Structural Biology*, 134(2-3):191–203, 2001.

S. Mann and Y. P. P. Chen. Bacterial genomic G + C composition-eliciting environmental adaptation, 2010.

R. G. Martin, J. H. Matthaei, O. W. Jones, and M. W. Nirenberg. Ribonucleotide composition of the genetic code. *Biochemical and Biophysical Research Communications*, 6 (6):410–414, 1961.

C. Mignon, N. Mariano, G. Stadthagen, A. Lugari, P. Lagoutte, S. Donnat, S. Chenavas, C. Perot, R. Sodoyer, and B. Werle. Codon harmonization – going beyond the speed limit for protein expression. *FEBS Letters*, 592(9):1554–1564, 2018.

G. Minervini, G. Evangelista, L. Villanova, D. Slanzi, D. De Lucrezia, I. Poli, P. L. Luisi, and F. Polticelli. Massive non-natural proteins structure prediction using grid technologies. *BMC Bioinformatics*, 10, 2009.

M. Mirdita, L. von den Driesch, C. Galiez, M. J. Martin, J. Söding, and M. Steinegger. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research*, 45(D1):D170–D176, 2016.

A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540, 1995.

NCBI. *RefSeq Growth Statistics*, accessed 2019-11-07. URL `https://www.ncbi.nlm.nih.gov/refseq/statistics/`.

NCBI. *The Statistics of Sequence Similarity Scores*, accessed 2019-11-17. URL `https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html`.

S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443 – 453, 1970.

S. Nepomnyachiy, N. Ben-Tal, and R. Kolodny. Global view of the protein universe. *Proceedings of the National Academy of Sciences of the United States of America*, 111(32):11691–11696, 2014.

S. Nepomnyachiy, N. Ben-Tal, and R. Kolodny. Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. *Proceedings of the National Academy of Sciences of the United States of America*, 114(44):11703–11708, 2017.

J. Ngo and J. Marks. Computational complexity of a problem in molecular structure prediction. *Protein Engineering, Design and Selection*, 5(4):313–321, 1992.

E. M. Novoa and L. Ribas de Pouplana. Speeding with control: Codon usage, tRNAs, and ribosomes, 2012.

Y. Ofran and H. Margalit. Proteins of the same fold and unrelated sequences have similar amino acid composition. *Proteins: Structure, Function, and Bioinformatics*, 64(1): 275–279, 2006.

H. Osmanbeyoglu and M. K. Ganapathiraju. N-gram analysis of 970 microbial organisms reveals presence of biological language models. *BMC Bioinformatics*, 12(1):12, 2011.

C. Palacios and J. J. Wernegreen. A Strong Effect of AT Mutational Bias on Amino Acid

Usage in Buchnera is Mitigated at High-Expression Genes. *Molecular Biology and Evolution*, 19(9):1575–1584, 2002.

V. S. Pande, A. Y. Grosberg, and T. Tanaka. Nonrandomness in protein sequences: evidence for a physically driven stage of evolution? *Proceedings of the National Academy of Sciences of the United States of America U. S. A.*, 91(26):12972–12975, 1994.

W. R. Pearson. An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics*, (SUPPL.42):1–8, 2013.

J. B. Plotkin and G. Kudla. Synonymous but not the same: The causes and consequences of codon bias. *Nature Reviews Genetics.*, 12(1):32–42, 2011.

J. Poznański, J. Topiński, A. Muszewska, K. J. Dębski, M. Hoffman-Sommer, K. Pawłowski, and M. Grynberg. Global pentapeptide statistics are far away from expected distributions. *Scientific Reports*, 8(1):15178, 2018.

K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(SUPPL. 1):61–65, 2007.

O. B. Ptitsyn. Random sequences and protein folding. *Journal of Molecular Structure: THEOCHEM*, 24(1-2):45–65, 1985.

M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn. The Pfam protein families database. *Nucleic Acids Research*, 40(D1):D290–D301, 2012.

T. E. Quax, N. J. Claassens, D. Söll, and J. van der Oost. Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular Cell*, 59(2):149–161, 2015.

K. Reinert, T. H. Dadi, M. Ehrhardt, H. Hauswedell, S. Mehringer, R. Rahn, J. Kim, C. Pockrandt, J. Winkler, E. Siragusa, G. Urgese, and D. Weese. The seqan c++ template library for efficient sequence analysis: A resource for programmers. *Journal of Biotechnology*, 261:157 – 168, 2017.

M. Remmert, A. Biegert, A. Hauser, and J. Söding. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2):173–175, 2012.

C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker. Protein structure prediction using rosetta. *Methods in Enzymology*, 383:66 – 93, 2004.

B. Rost. Protein structures sustain evolutionary drift. *Folding and Design*, 2(3):S19–S24, 1997.

B. Rost. Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection*, 12(2):85–94, 1999.

B. Rost. Review: Protein secondary structure prediction continues to rise. *Journal of Structural Biology*, 134(2-3):204–218, 2001.

B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, 232(2):584–599, 1993.

Russell F. Doolittle. Similar Amino Acid Sequences:Chance or Common Ancestry? *Science*, 84(1951):74–84, 1981.

Y. Saito, W. Kitagawa, T. Kumagai, N. Tajima, Y. Nishimiya, K. Tamano, Y. Yasutake, T. Tamura, and T. Kameda. Developing a codon optimization method for improved expression of recombinant proteins in actinobacteria. *Scientific Reports*, 9(1):1–10, 2019.

A. Šali, E. Shakhnovich, and M. Karplus. How does a protein fold? *Nature*, 369(6477): 248–251, 1994.

A. A. Schaffer. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research*, 29(14): 2994–3005, 2002.

R. Schneider, A. de Daruvar, and C. Sander. The HSSP database of protein structure-sequence alignments. *Nucleic Acids Research*, 25(1):226–230, 1997.

P. Shah, D. M. McCandlish, and J. B. Plotkin. Contingency and entrenchment in protein evolution under purifying selection. *Proceedings of the National Academy of Sciences of the United States of America*, 2015, 2015.

E. I. Shakhnovich and A. M. Gutin. Engineering of stable and fast-folding sequences of model proteins. *Proceedings of the National Academy of Sciences of the United States of America U. S. A.*, 90(15):7195–7199, 1993.

P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A Software Environment for Integrated Models. *Genome Research*, 13(22):426, 1971.

P. M. Sharp and W.-H. Li. The codon adaptation index - a measure of directional sysonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, 14(11): 4683–4690, 1986.

M. Y. Shen, F. P. Davis, and A. Sali. The optimal size of a globular protein domain: A simple sphere-packing model. *Chemical Physics Letters*, 405(1-3):224–228, 2005.

S. Shifman, A. Pisanté-Shalom, B. Yakir, and A. Darvasi. Quantitative technologies for allele frequency estimation of SNPs in DNA pools. *Molecular and Cellular Probes*, 16(6):429–434, 2002.

J. M. Smith. Natural selection and the concept of a protein space. *Nature*, 225(5232): 563, 1970.

T. F. Smith, M. S. Waterman, et al. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

J. Söding. Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21 (7):951–960, 2005.

H. Spencer. *The Principles of Biology*. Williams and Norgate, 1864.

T. N. Starr, L. K. Picton, and J. W. Thornton. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature Reviews Genetics*, 549, 2017.

B. J. Strait and T. G. Dewey. The Shannon information entropy of protein sequences. *Biophysical Journal*, 71(1):148–155, 1996.

J. Söding and A. N. Lupas. More than the sum of their parts: On the evolution of proteins from peptides. *BioEssays*, 25(9):837–846, 2003.

R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1):33–36, 2000.

The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 2016.

P. Tian and R. B. Best. How Many Protein Sequences Fold to a Given Structure? A Coevolutionary Analysis. *Biophysical Journal*, 113(8):1719–1730, 2017.

S. R. Trevino, J. M. Scholtz, and C. N. Pace. Amino Acid Contribution to Protein Solubility: Asp, Glu, and Ser Contribute more Favorably than the other Hydrophilic Amino Acids in RNase Sa. *Journal of Molecular Biology*, 366(2):449–460, 2007.

UK MRC. *Statistics of the SCOP2 database*, accessed 2020-01-19. URL `http://scop.mrc-lmb.cam.ac.uk/stats`.

S. Urlinger, U. Baron, M. Thellmann, M. T. Hasan, H. Bujard, and W. Hillen. Exploring the sequence space for tetracycline-dependent transcriptional activators: Novel mutations yield expanded range and sensitivity. *Proceedings of the National Academy of Sciences of the United States of America*, 97(14):7963–7968, 2000.

R. E. Valas, S. Yang, and P. E. Bourne. Nothing about protein structure classification makes sense except in the light of evolution. *Current Opinion in Structural Biology*, 19(3):329–334, 2009.

A. E. Vinogradov. Compactness of human housekeeping genes: Selection for economy or genomic design? *Trends in Genetics*, 20(5):248–253, 2004.

J. Vymětal, J. Vondrášek, and K. Hlouchová. Sequence Versus Composition: What Prescribes IDP Biophysical Properties? *Entropy*, 21(7):654, 2019.

Z. X. Wang and Z. Yuan. How good is prediction of protein structural class by the component- coupled method? *Proteins: Structure, Function, and Bioinformatics*, 38 (2):165–175, 2000.

J. D. Watson and F. H. Crick. Molecular structure of nucleic acids: A Structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

G. R. Webster, A. Y. Teh, and J. K. Ma. Synthetic gene design—The rationale for codon optimization and implications for molecular pharming in plants. *Biotechnology and Bioengineering*, 114(3):492–502, 2017.

L. Weidmann, T. Dijkstra, O. Kohlbacher, and A. Lupas. Where natural protein sequences stand out from randomness. *bioRxiv*, 2019. URL `https://www.biorxiv.org/content/early/2019/07/28/706119`.

O. Weiss, M. A. Jimé Nez-Montaño, and H. Herzel. Information Content of Protein Sequences. *Journal of Theoretical Biology*, 206(3):379–386, 2000.

S. J. Wheelan, A. Marchler-Bauer, and S. H. Bryant. Domain size distributions can predict domain boundaries. *Bioinformatics*, 16(7):613–618, 2000.

S. H. White and R. E. Jacobs. Statistical distribution of hydrophobic residues along the length of protein chains. Implications for protein folding and evolution. *Biophysical Journal*, 57(4):911–921, 1990.

D. N. Woolfson, G. J. Bartlett, A. J. Burton, J. W. Heal, A. Niitsu, A. R. Thomson, and C. W. Wood. De novo protein design: How do we expand into the universe of possible protein structures? *Current Opinion in Structural Biology*, 33:16–26, 2015.

J. C. Wootton and S. Federhen. Analysis of compositionally biased regions in sequence databases. *Methods in Enzymology*, 266:554–571, 1996.

F. Wright. The 'effective number of codons' used in a gene. *Gene*, 87(1):23–29, 1990.

S. Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Preceedings of the 6th International Congress of Genetics*, 1:356–366, 1932.

K. Wüthrich. *NMR of Proteins and Nucleic Acids*. Wiley-Interscience, 1986.

J. F. Yu, Z. Cao, Y. Yang, C. L. Wang, Z. D. Su, Y. W. Zhao, J. H. Wang, and Y. Zhou. Natural protein sequences are more intrinsically disordered than random sequences. *Cellular and Molecular Life Sciences*, 73(15):2949–2957, 2016.

J. Zhang. Evolution by gene duplication: An update. *Trends in Ecology and Evolution*, 18(6):292–298, 2003.

L. Zimmermann, A. Stephens, S.-Z. Nam, D. Rau, J. Kübler, M. Lozajic, F. Gabler, J. Söding, A. N. Lupas, and V. Alva. A completely reimplemented mpi bioinformatics toolkit with a new hhpred server at its core. *Journal of Molecular Biology*, 430(15): 2237 – 2243, 2018. Computation Resources for Molecular Biology.

G. K. Zipf. *The psycho-biology of language: An introduction to dynamic philology*. Routledge, 2013.

Index