Intrinsic disorder and tandem repeats - match made in evolution

Computational studies of molecular evolution

Oxana Lundström

This dissertation explores the role of tandem repeats and intrinsic disorder in protein functions within eukaryotic genomes, and how they contribute to molecular evolution and human disease. Genetic variation in intrinsically disordered regions can lead to protein length expansion, a key factor in rapid evolution. The author examines human short tandem repeat variation and its impact on gene expression patterns in colorectal cancer. This work emphasizes the importance of bioinformatics in the study of potential biomarkers and drug targets based on genetic variations in these proteins.
Intrinsic disorder and tandem repeats - match made in evolution
Computational studies of molecular evolution

Oxana Lundström

Academic dissertation for the Degree of Doctor of Philosophy in Biochemistry towards Bioinformatics at Stockholm University to be publicly defended on Monday 11 December 2023 at 09.00 in G-salen, Arrheniuslaboratorierna hus F, Svante Arrhenius väg 20 C and online via Zoom, public link is available at the department website.

Abstract
Proteins are both the building blocks and workers of the cell, carrying out most of the important functions. For a long time, their structure has been regarded as the primary factor for their function, but intrinsically disordered proteins demonstrate an alternative to this paradigm. Disordered proteins can temporarily assume different forms based on their interactions with other molecules and play critical roles in several biological processes, including cell signaling and regulation of gene expression.

Tandem repeats are repeated patterns in genetic sequence. The role of tandem repeats in many protein structures is well documented today, but their role in disordered proteins is not entirely clear. This thesis aims to shed light on the mechanisms by which protein disorder and tandem repeats are linked.

Only 2.5% of residues in all known protein sequences are characterized by the overlap of tandem repeats and protein disorder as described in Paper III, but many of these proteins have crucial functions and are linked to human diseases. Short tandem repeats emerge in this study as most frequently occurring in disordered regions. Genetic variation in disordered proteins accounts for length differences in eukaryotic genes (Paper I) and many orphan, recently evolved proteins, are disordered due to high GC content (Paper II).

A medical application of this research is illustrated in the thesis with examples of variations in short tandem repeats (STRs) and their role in human diseases. Paper IV presents a comprehensive resource of human STR variation and Paper V illustrates how it can be used to identify specific STRs of interest, such as in the case of colorectal cancer where variations in certain STRs lead to altered gene expression patterns in tumors.

Keywords: Protein evolution, intrinsically disordered proteins (IDPs), tandem repeats, short tandem repeats (STRs), genetic variation, orphan proteins, GC content, human STR variation, colorectal cancer; gene expression.
INTRINSIC DISORDER AND TANDEM REPEATS - MATCH MADE IN EVOLUTION

Oxana Lundström
Intrinsic disorder and tandem repeats - match made in evolution
Computational studies of molecular evolution

Oxana Lundström
“All that is gold does not glitter,
Not all those who wander are lost;
The old that is strong does not wither,
Deep roots are not reached by the frost.”

J.R.R. Tolkien
Abstract

Proteins are both the building blocks and workers of the cell, carrying out most of the important functions. For a long time, their structure has been regarded as the primary factor for their function, but intrinsically disordered proteins demonstrate an alternative to this paradigm. Disordered proteins can temporarily assume different forms based on their interactions with other molecules and play critical roles in several biological processes, including cell signaling and regulation of gene expression.

Tandem repeats are repeated patterns in genetic sequence. The role of tandem repeats in many protein structures is well documented today, but their role in disordered proteins is not entirely clear. This thesis aims to shed light on the mechanisms by which protein disorder and tandem repeats are linked.

Only 2.5% of residues in all known protein sequences are characterized by the overlap of tandem repeats and protein disorder as described in Paper III, but many of these proteins have crucial functions and are linked to human diseases. Short tandem repeats emerge in this study as most frequently occurring in disordered regions. Genetic variation in disordered proteins accounts for length differences in eukaryotic genes (Paper I) and many orphan, recently evolved proteins, are disordered due to high GC content (Paper II).

A medical application of this research is illustrated in the thesis with examples of variations in short tandem repeats (STRs) and their role in human diseases. Paper IV presents a comprehensive resource of human STR variation and Paper V illustrates how it can be used to identify specific STRs of interest, such as in the case of colorectal cancer where variations in certain STRs lead to altered gene expression patterns in tumors.
# Contents

Included publications ................................................................. 1  
Author contributions ................................................................. 3  
Other relevant publications ......................................................... 4  
Populärvetenskaplig sammanfattning .............................................. 5  
Introduction ................................................................................. 7  
Biological background ................................................................. 9  
  DNA, RNA and proteins .............................................................. 9  
  Genes and genome ................................................................. 10  
  Proteins .................................................................................. 10  
  The three domains of life .......................................................... 12  
Molecular evolution ........................................................................ 13  
  Evolutionary theory ................................................................. 13  
  Protein sequence variations ...................................................... 14  
    Mutations ............................................................................. 15  
    Copy Number Variations (CNVs) ............................................ 16  
  Orthologs, paralogs and orphans ............................................... 17  
  Evolutionary constraints on protein-coding genes ....................... 18  
    Structural constraints ......................................................... 18  
    Functional constraints on protein evolution ......................... 18  
    Protein domains .................................................................. 19  
  Protein intrinsic disorder .......................................................... 19  
  Tandem repeats ........................................................................ 21  
    Short Tandem Repeats (STRs) ............................................... 22  
    Intrinsic disorder and STRs ................................................. 24  
  Next-generation sequencing technologies (NGS) ......................... 25  
Computational methods ................................................................ 26  
  Multiple sequence alignments (MSAs) ....................................... 26  
  Homology detection and orphans ............................................. 27  
  Protein structure and disorder prediction .................................. 28  
    TOP-IDP Scale .................................................................... 30
Genotyping tandem repeat variations.................................31
Tandem repeats annotation..................................................32
Biological databases............................................................33
  Genes duo: Ensembl & NCBI/Entrez..................................33
  Protein trio: Uniprot/Swiss-Prot, PFAM & PDB.................33
  Evolutionary databases: Inparanoid and OrthoDB..........34
Summary of results.....................................................................35
  Protein expansion is due to long indels in disordered regions
    (Paper I)........................................................................35
  High GC content causes orphan proteins to be intrinsically
    disordered (Paper II)..........................................................36
  Protein tandem repeats and their relationship with intrinsic
    disorder (Paper III)............................................................36
  Database of short tandem repeat variation in humans. (Paper IV)....37
    Short tandem repeat mutations regulate gene expression in
    colorectal cancer (Paper V)..............................................38
Concluding remarks.....................................................................39
Future outlook............................................................................40
Acknowledgements......................................................................41
References..................................................................................43
Included publications


* = shared first authorship

Reprints were made with permission from the publisher.
Author contributions

Paper I - I took over the initial scripts and analysis done by R.S. and D.E., rewritten and improved them. I produced the figures for publication and contributed to the manuscript text by proofreading.

Paper II - I performed the analysis together with W.B., mainly responsible for intrinsically disordered regions analysis of the dataset. Set up the pipeline for disorder prediction and contributed to the final manuscript version by proofreading.

Paper III - I designed the analysis pipeline together with E.S., extracted disordered regions predictions and performed overlap calculations, analyzed and visualized the results and produced publication graphics of the part of the manuscript about protein disorder. Together with E.S., M.A. and A.E. wrote the first draft of the manuscript. Together with M.D., I worked on the final version of the revised manuscript and provided proofreading and adjustments to match peer review comments.

Paper IV - I came up with the initial idea to make part of our project results into a public database with programmatic access, designed data structures and software architecture, implemented and deployed the software. I set up and run the pipeline to produce the genotyping data. Together with M.G. revised the project plan and merged with another database and service that they had developed earlier. I wrote and revised the manuscript, produced publication figures and worked on adjustments after the peer-review.

Paper V - M.V. and M.A. were responsible for experiment planning. I have assisted with the data collection, interpretation of results and visualization, as well as editing and proofreading of the manuscript.
Other relevant publications


Populärvetenskaplig sammanfattning

Proteiner är både cellens byggnstenar och dess arbetare och utför de flesta viktiga uppgifter i våra celler. Länge har deras struktur betraktats som den primära faktorn för deras funktion, men i slutet av 90-talet, när de experimentella metoderna för att studera proteinstrukturer förbättrades, blev det tydligt att inte alla proteiner följer den traditionella idén att funktionen är en direkt följd av strukturen. Proteinoordning hänvisar till områden inom proteiner eller hela proteiner som inte har fasta strukturer. Istället kan dessa områden eller proteiner tillfälligt anta olika former baserat på sina interaktioner med andra molekyler. Dessa oordnade proteiner har visat sig ha kritiska roller i flera biologiska processer, inklusive cellsignalering och reglering av genuttryck.

Tandemrepetitioner är upprepade mönster i genetisk sekvens som först identifierades i proteiners strukturella delar. Tandemrepetitioners roll i många proteinstrukturer är väl beskriven idag, men vilken roll de spelar i oordnade proteiner är inte helt klart. Till exempel kan tandemrepetitioner fungera som flexibla “länkar” i proteiner och låta dem ändra sin form när de interagerar med andra proteiner eller DNA.

Denna avhandling syftar till att belysa de mekanismer genom vilka oordning i proteiner och tandemrepetitioner är kopplade i friska celler och cancerceller.

Endast 2,5% av proteinbaser i alla kända proteinsekvenser uppvisar överlappning av tandemrepetitioner och proteinoordning som beskrivs i Artikel III, men många av dessa proteiner har avgörande funktioner och är kopplade till mänskliga sjukdomar. Korta tandemrepetitioner framträder i denna studie som mest förekommande i oordnade regioner. Genetisk variation i oordnade proteiner ansvarar för större längder i eukaryota gener (Artikel I) och många föräldralösa, nyligen utvecklade proteiner, är oordnade på grund av högt GC-innehåll (Artikel II). Detta kopplar starkt samman proteiner med oordningen till processen för snabb evolution och deras förmåga att ta upp nya funktioner.

En medicinsk tillämpning av denna forskning visas i avhandlingen med exempel av variationer i korta tandemupprepningar (STR) och deras roll i mänskliga sjukdomar. STR-utvidgningar kan leda till produktionen av proteiner med oordning, som vid neurodegenerativa sjukdomar. I många mänskliga cancer förekommer STR-utvidgningar på många olika platser i genomet. Artikel IV presenterar en omfattande resurs av mänsklig
STR-variation och Artikel V illustrerar hur den kan användas för att identifiera specifika STRs av intresse, som i fallet med kolorektal cancer där variationer i en vissa STRs leder till förändrade genuttrycksmönster i tumörer.
Introduction

Protein structure has for a long time been regarded as the primary determinant of function, with the “lock-and-key” model providing a simplified yet efficient framework for explaining protein-protein interaction and functions. However, as the experimental methods to study protein structure advanced, it became apparent that not all proteins conform to the traditional structural paradigm.

In the early 1990s, scientists started to encounter regions in proteins that lacked a stable three-dimensional structure yet played a crucial role\(^1\). These discoveries paved the way for further investigations of the intrinsically disordered regions and their dynamic nature enabling proteins to engage in flexible interactions.

Tandem repeats, another significant concept in this thesis, are consecutive identical or similar motifs initially identified in structural parts of the proteins\(^2\). Their contribution to the diversity of protein architectures is well understood today. Tandem repeats also play an important role in evolutionary processes, as many arise due to duplication events\(^3\).

However, the association between intrinsic disorder and tandem repeats is a much more recent finding, first observed as an evolutionary mechanism\(^4\). Today, we have only begun to tap into the world of dynamic, rapidly evolving, flexible and multifunctional proteins that employ both of these phenomena. In transcription factors, tandem repeats often serve as flexible linkers and regulatory regions, allowing intrinsically disordered proteins to change conformations depending on the interactions with other proteins and DNA. In signal transduction pathways, tandem repeats serve as binding sites in the middle of a flexible intrinsically disordered region. In scaffold proteins, disorder and tandem repeats co-occur to enable the assembly of multivalent complexes. Finally, in neurodegenerative diseases and cancer, the interplay between tandem repeats and intrinsic disorder contributes to disease pathology and progression, and these regions become potential targets for various therapies.

This work aims to shed some light on the mechanisms through which intrinsic disorder and tandem repeats are connected in healthy and cancerous
cells. I first introduce some essential biological concepts mentioned in the publications, as well as the computational resources and methods used in my research. Summary of the five publications follows: **Paper I** starts with the broad introduction of evolutionary mechanisms in intrinsically disordered regions; **Paper II** continues on the topic of evolutionary innovation and disorder in orphaned proteins; **Paper III** presents a thorough annotation of tandem repeats in different species and introduces the connection between repeats and intrinsic disorder in proteins; **Paper IV** describes the development of a database of short tandem repeat variations that can be used for medical research in conditions where repeat expansion leads to protein misfolding. **Paper V** extends the topic of short tandem repeat variation and demonstrates how it can influence tumor phenotype in a case study of colorectal cancer. Finally, I will discuss the future outlook of the field of evolution of intrinsic disorder by repeat expansions and its role in human disease.
Biological background

DNA, RNA and proteins

A cell is the smallest unit of living matter; it requires energy to maintain its organization and is capable of reproducing itself. These properties are central to the definition of life. All known living organisms consist of one or multiple cells, and in all cases, the whole organism is generated by cell divisions starting from a single cell. The single cell, therefore, contains all the necessary information and machinery that defines the organism. Since the discovery of the structure of DNA by Watson and Crick in 1953 [5] we know that all living cells store this information in the form of \textbf{double-stranded DNA} - long molecular chains formed of the same building blocks, nucleic acids (nucleotides). DNA consists of four different types of nucleic acids: adenine, guanine, cytosine and thymine (denoted by A, G, C and T).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{The central dogma of molecular biology: the flow of genetic information from DNA, through RNA and to protein.}
\end{figure}
The nucleic base pairs are attached in a linear sequence that encodes the hereditary information of the cell. Information stored in DNA is read out and put to use through a two-step process: first, in transcription, segments of DNA sequence are used to guide the synthesis of many RNA molecules. Then, in translation, some RNA molecules are used to synthesize protein molecules (Figure 1), while others help in this process.

Genes and genome
A specific set of rules for translating DNA to protein is shared between all living cells. Different combinations of three base pairs in the DNA, named codons, code for a certain amino acid or serve as a signal for the start or end of a protein molecule. This genetic code rule set is redundant, i.e., some codons code for the same amino acid\(^6\). Each segment of DNA that encodes for a functional molecule, either RNA or protein, is called a gene\(^6\). Regulatory elements, i.e. sequences that influence the gene transcription processes, are today also considered an integral part of a gene\(^7\). The complete set of biological information, often but not always manifested as DNA, that together with other sources of information, produces and maintains the organism, is called a genome\(^8\).

Genetic sequences are characterized by numerous attributes that have significant implications on their function and influence on the phenotype of the organism. The length of the genes varies in different types of genomes, with bacteria and archaea typically having smaller and more compact genes compared to plants and animals\(^9\). However, the length of the gene can significantly vary even within the same organism, depending on the function the gene is performing\(^10\). The GC content or the proportion of Guanine and Cytosine in the DNA sequence affects the molecule’s structure, stability, and gene expression and is also an important property of genetic sequences\(^11\). Codon usage is another characteristic connected to the base composition of genes and is referred to as the “codon usage bias”. Different mutation rates influence codon usage in DNA, GC content of the genome, gene expression and function\(^12\).

Proteins
Proteins are long, unbranched polymer chains formed by different monomeric building units, the amino acids. Each protein molecule, or polypeptide, is created by joining amino acids in a particular sequence, which folds into a three-dimensional form. Proteins constitute more than 50% of the cell's dry mass\(^13\) and perform nearly all functions in a cell: catalyzing metabolic reactions, replicating DNA, responding to stimuli, and transporting molecules from one location to another.

The sequence of amino acids forming a protein is also referred to as the primary structure of the protein. Some of the local arrangements of amino
acid residues have been observed in a variety of different proteins. Combinations of such conformations represent the secondary structure of a given protein. There are two main kinds of secondary structure: spiral conformations, where the amino acids are packed tightly together, called alpha-helices, and long flat sheets, extended in a way that the amino acids are stretched out as far from each other as they can be\textsuperscript{14}. Each extended chain is called a beta-strand, and two or more beta-strands held together are called a beta-sheet.

Once the sequence of the secondary structure is formed, the protein folds further into a unique three-dimensional shape, referred to as its tertiary structure. For some proteins, multiple protein chains, or subunits, come together to form a functional protein complex (quaternary structure). Proteins can have a wide range of structures - they can be globular like enzymes, fibrous like collagen, or even form complex ring structures like ATP synthase\textsuperscript{13}. This structural diversity underlies the wide array of functions proteins perform in the body.

![Protein secondary and tertiary structure](image)

**Figure 2.** Protein secondary and tertiary structure. On the left: cartoon representation of the two regular secondary structure elements, alpha helix and beta sheet, with highlighter hydrogen bonds stabilizing the secondary structure elements’ backbones. On the right: cartoon representation of a protein tertiary structure (PDB:3H4L\textsuperscript{15}) with secondary structure elements and unstructured regions highlighted.
**Protein folding** is a complex process driven by various types of molecular interactions: hydrophobic interactions, where nonpolar amino acids move towards the inside of the protein; electrostatic interactions that occur between positively and negatively charged side chains; disulfide bonds that form between the side chains of cysteine residues and Van der Waals forces that further stabilize the structure. Protein folding is driven by the need to minimize the energy of the molecule. Single protein molecules demonstrably and repeatedly fold spontaneously into their native state. This is known as Anfinsen’s dogma\textsuperscript{16}.

**The three domains of life**

Based on their evolutionary descent and biochemical properties, living organisms are classified into three main branches: Archaea, Eukarya and Bacteria. In Archaea and Bacteria, DNA is not separated from the cytoplasm, while the genetic material in Eukarya, in contrast, is packed in the **nucleus**. Archaea are evolutionarily equally distant from Bacteria and Eukarya and possess unique lipid compositions and translational mechanisms. Bacteria and Archaea are single-celled, while eukaryotes can be unicellular or consist of many cells of different types and functions. **Eukarya** is further subdivided into five kingdoms: Protozoa ( unicellular eukaryotic organisms), Chromista (algae, simple organisms with chloroplasts), Plantae (plants), Fungi (yeast, molds and mushrooms) and Metazoa (animals).

The three domains of life demonstrate fundamental differences in genome organization and mechanisms of evolution. The bacterial genome is limited to 180kb - 13 mb and the archaeal one to 500kb - 5 mb\textsuperscript{9}, while that of eukaryotes can be characterized by a significant expansion in size and complexity\textsuperscript{17,18}. The evolution of bacterial and archaeal genomes involves two major processes: acquisition of exogenous DNA through horizontal gene transfer and genome decay through deletion\textsuperscript{19}. The genomes of Archaea are similar to those of Bacteria in size and gene density, but some translational mechanisms are the same as those in eukaryotes.

Eukaryotes demonstrate a significant increase in genetic information compared to the other two domains of life. An abundance of non-coding DNA, gene duplications and alternative splicing creates more possibilities for complicated evolutionary pathways in eukaryotic genomes. Eukaryotic genes are made up of sequences that contain both coding regions, known as **exons**, and non-coding regions, known as **introns** (Figure 3). Exons are portions of a gene that are expressed, or transcribed, to later produce the final protein. Introns, on the other hand, are intervening sequences within a gene that are not translated into the mature protein. These sequences are transcribed into RNA, but then are removed or “spliced out“ during the RNA
processing, leaving only the exons to be expressed. Introns often contain regulatory elements and influence gene expression or splicing.

**Figure 3.** Eukaryotic gene structure showing a schematic representation of exons, introns and regulatory elements.

**Molecular evolution**

**Evolutionary theory**

Evolutionary theory famously dates back to 1858, when the most influential papers in biology were published. Darwin and Wallace independently presented the hypothesis of “descent with modification” that accounts for the diversity and appearance of life on Earth. Advances in the fields of population genetics, paleontology, developmental and molecular biology formed the integrated view of evolution, primarily developed in the 1930-50s by Fisher, Kutschera, Mayr and others. In the early 1960s, the first sequencing data from different groups of organisms became available. It was only a few small proteins, such as insulin, hemoglobin, and ribonuclease, but already then it became apparent that the evolution on the molecular level cannot only be explained by the adaptive processes and the mechanisms of introducing variety in the genome are not limited to mutation and recombination. A new field, **molecular evolution**, emerged and new theories to explain the underlying processes of phenotypic diversity and genome size evolution were proposed.

According to modern views, evolution is the change in the inherited characteristics in interbreeding populations of living organisms, species. Genetic and phenotypic variety in these populations is introduced through mutations, i.e., changes in the DNA sequence of an organism. In sexually reproducible organisms, another source of variation comes into play - recombination, an exchange of genetic information between two organisms. The amount of variation that is passed from one generation to the next is quite small; almost all variations will be lost or neutral (not affect the fitness
of the organism). However, some of the changes in hereditary material can lead to changes in the phenotype, physical characteristics of an organism, which in turn affect the survival of the organism and its ability to produce offspring. Fossil, phenotypic and genetic evidence can be combined when studying the evolution of organisms and create a more complete picture of historical events that led to the presented evidence. A **phylogeny** is a tree-like diagram that represents the evolutionary connections between different biological species or entities, which are determined by their physical or genetic traits. The branches of the phylogeny illustrate the evolutionary path of species from their shared ancestors over a period of time (Figure 4).

![Evolutionary tree](image)

**Figure 4.** Evolutionary tree illustrating the candidate common ancestor of Diptera genus - *Tipulidae (Crane Fly)* and the phylogeny of existing mosquito and fly species based on the current fossil, phenotypic and genetic evidence. Tipulidae is an artist reconstruction courtesy of The US National Park Service.

**Protein sequence variations**

It is impossible to study the molecular evolution of protein-coding genes separated from other essential regions of the genome. Regulatory elements on the non-coding part of the DNA and non-coding RNA molecules can significantly impact the phenotype of the organism by altering gene transcription and expression levels. Structural elements of the DNA, such as telomeres and the organization of chromatin also play an important role. When I embarked on my research journey, the scientific community had only recently begun tapping into the vast amounts of unknown mechanisms of the
non-coding part of genome evolution. Therefore, my focus has been primarily on the protein-coding genes for which we have a lot of well-structured and reliable data from many different organisms. This section describes important factors in the evolution of protein-coding genes - ways that sequence variations can be introduced, how they can lead to novel functions and what constraints are present in this process.

Mutations
Mutations are sequence changes in DNA and are one of the main sources of genetic diversity. For mutations to affect an organism's offspring, they must occur in cells that produce the next generation (as opposed to somatic mutations) and affect the hereditary material. In the smallest type of mutation event, a single base pair in a DNA sequence is changed into another base pair. This type of mutation can lead to changes in the protein sequence if this segment or be synonymous, i.e., code for the same amino acid. Synonymous mutations exist because many amino acids are encoded by multiple codons, as described earlier.
The frequency by which a certain type of mutation occurs is generally expressed as the number of mutations per biological unit (i.e., per cell division, per gamete, or per round of replication).
Based on the effect of the mutation on the fitness of the organism, it can be beneficial, harmful, or neutral. Beneficial or harmful mutations are what allow us to see natural selection in action. Mutations that increase the fitness of the organism and become heritable tend to increase in frequency in the population; harmful or deleterious mutations, on the other hand, become extinct or remain, albeit at lower prevalence, in the population. Neutral mutations don't have any effect on fitness and can widely spread in a population. According to the neutral theory, the majority of mutations that persist in the population (become fixed) are neutral.
Mutations may also take the form of insertions or deletions, which together are known as indels. The number of nucleotide bases that are inserted or deleted from the DNA sequence can vary a lot. Indels of one or two bases in the coding sequences are the most common and can have a substantial impact on protein translation by causing a frameshift because of the triplet nature of codons.
Short indels that don't cause a frameshift in translation and become fixed in the population often occur in exposed loop regions. Longer indel events might involve an insertion or deletion of an entire protein domain. The selective pressure acting on these events is less well understood, but fixed long indels are often associated with functional change.
Figure 5. Examples of point mutations (orange vs blue) and an insertion event (red) in three fragments of homologous protein sequences of fruit flies and a yellow fever mosquito.

Copy Number Variations (CNVs)
These are variations where entire genes or sections of genes are duplicated or deleted. Several mechanisms of these variations are known in eukaryotes: unequal crossing-over, unequal sister chromatid exchange, replication slippage and retrotransposition. These errors can lead to phenotypic variation, gene expression changes and, in some cases, in regions of lower evolutionary pressure, can serve as functional and structural redundancies and sources of evolutionary innovation.

Gene duplication is the most important process for generating new genes during molecular evolution.

There are several ways in which it can occur:
- By duplication of a single gene or a group of genes
- By whole genome duplication
- By speciation

The first of these mechanisms is the most common: multigene families are widely spread in all known species. By comparing the sequences of the members of these families, it is possible to trace individual duplications that occurred in a common ancestral genome. The initial result of such an event is two identical genes. Selective constraints will ensure that one of these copies is still capable of maintaining the original function; the daughter copy, in the meantime, is subject to relaxed evolutionary pressure.

Most of these excess gene copies acquire deleterious mutations that render them inactive and become pseudogenes\textsuperscript{32}. Occasionally, though, the mutations that accumulate in this gene copy lead to a new function or a slightly altered sub-function of a parent gene. Genes coding for different types of globin proteins that bind oxygen, evolved from a common ancestral gene, which, after successive duplications and speciation events, led to the genes that encode the widespread globin superfamily\textsuperscript{33}. 

16
Whole Genome Duplication (WGD) is the most rapid mechanism of increasing gene copy number and can occur due to an error during meiosis. The evidence from both experimental and computational studies suggests that many gene innovations in yeast and animals are the results of the WGDs during their evolutionary history. Increased complexity in the vertebrate developmental regulatory network has been reported to be the result of the two rounds of WGD at the base of vertebrate evolution.

Speciation causes genes to become orthologs, as described in the next chapter.

Orthologs, paralogs and orphans

Homologs are any genes or proteins from different species that share a common ancestral origin, implying they have similar structures and often a related function.

During protein evolution analysis, it is important to distinguish between proteins that are derived from a single ancestral gene in the last common ancestor of the given two species, i.e., as a result of a speciation event (orthologs), and proteins that evolved through duplication within the same (perhaps ancestral) genome, which are referred to as paralogs (Figure 6).

Figure 6. Gene trees with duplication preceding speciation event a) and evolutionary relationships are shown for the genes. b) de novo orphan gene origination from an ancestral non-coding sequence.
Lastly, orphan genes, also known as taxonomically restricted genes, are those that are found in only one or a few species and lack detectable homologs in other lineages. These often play a role in species-specific adaptations and are believed to have evolved through *de novo* gene creation from a non-coding sequence\(^\text{37}\).

Evolutionary constraints on protein-coding genes

**Structural constraints**

The selective pressures that might increase or reduce the benefits of acquired traits in a population shape the evolution of animal phenotypes. Selective pressures also exist both on the genome level (gene copy numbers, genome size) and on the molecular level. The evidence for selective pressures acting on the protein level includes the fact that three-dimensional structure tends to be more conserved than the corresponding amino acid sequence\(^\text{38}\). Many unrelated proteins exhibit a high degree of structural conservation, even though their primary sequences significantly diverge. This suggests that the maintenance of structural folds strongly constrains protein evolution\(^\text{39}\).

![Protein MSA alignment vs structure](image)

**Figure 7.** Protein MSA alignment vs structure. Residues that form a contact in the structure of a given protein are usually strongly conserved in homologous proteins.

**Functional constraints on protein evolution**

Additional constraints are added by protein-protein interactions\(^\text{40}\) and differences in mutation rate for different proteins. For example, highly expressed proteins are constrained to have fewer mutations to avoid the cost of misfolding. Understanding the constraints on protein sequence variation is essential for comprehending the mechanisms of protein evolution, and it needs to be incorporated into all subsequent analyses.
Protein domains

Most eukaryotic proteins consist of structural domains, each comprising a segment of the polypeptide chain corresponding to one or more elements of secondary structure. Each structural domain folds independently and can even perform an independent function. Protein domains can also be seen from the evolutionary perspective as independent units of protein evolution. When studying duplication and rearrangement events in protein sequences, it is important to keep these structural components in mind. Duplication of a protein domain might result in making a protein product more stable and put it under a more relaxed evolutionary pressure. The domain copy can change over time, leading to a modified structure and protein activity. This evolutionary mechanism is particularly common in Metazoa, and more often occurs at the N and C-terminals with the exception of protein domain repeats\textsuperscript{41}, like in titin\textsuperscript{42} and nebulin proteins.

Protein intrinsic disorder

The notion that a protein requires a rigid 3D structure to function has been prevalent ever since 1931 when it was shown that protein denaturation leads to a complete loss of function\textsuperscript{43}. In the 1970s, the first unstructured, but functional region was identified in fibrinogen; this region plays a key role in blood clotting\textsuperscript{44}. Nevertheless, up to the 1990s, most scientists were convinced that disordered regions they could observe by NMR spectroscopy were nothing more than artifacts. The structure of p21, though, which was discovered to be fully disordered by Kriwacki\textsuperscript{45}, could not be overlooked. Despite the fact that its amino acids only assumed a stable conformation when bound to a kinase, p21 was still able to perform its critical regulatory function.

Several experimental approaches to identify intrinsically disordered regions in proteins have since been developed or adapted from existing ones. In NMR chemical shifts analysis, the pulse sequences have been tailored for resolving chemical shifts in the unfolded state; TROSY and CRINEPT-TROSY NMR spectroscopy now allow for the identification of disorder-order transition during protein binding\textsuperscript{46}. Temperature- and pressure-dependent changes in the NMR results can indicate conformational fluctuation of disordered regions. NMR spin relaxation is also advancing to help characterize protein flexibility. H/D exchange, in combination with mass spectrometry (HXMS), is used for high-throughput identification of disordered regions since disordered regions have higher H/D exchange rates to be able to fold before binding to different partners\textsuperscript{47}.

Experimental identification of protein structure and disorder-to-order transitions requires careful sample preparation, expensive equipment and is
time-consuming. Here is where computational biology comes to the rescue. Based on the current knowledge of the physical properties of protein disorder and amino acid sequences for experimentally identified unstructured regions, several methods to predict disorder from protein sequence have been developed\textsuperscript{48}. On the basis of amino acid composition, hydropathy, capacity of polypeptides to form stabilizing contacts and other differences to known globular proteins, these predictors label each amino acid in a protein sequence as ordered or disordered. While using these methods to study protein disorder and its evolution, it is important to keep in mind that they are limited to recognizing patterns observed in experimentally verified disordered regions, and each predictor is tailored to identify a certain type of features\textsuperscript{49}.

Eukaryotic proteomes are distinguished by their varying levels of intrinsic disorder content, which can range from a few percent to over 40\%\textsuperscript{50}. Among human proteins, roughly 30\% are predicted to contain significant disordered regions\textsuperscript{51}. Many of these proteins perform critical functions in signaling, gene regulation and different protein interactions\textsuperscript{52}. Disordered regions are of particular interest because they lack selective pressures against aggregation and misassembly described earlier, which can make them subject to a higher mutation rate, possibly giving rise to novel protein functions.

The “continuum of intrinsic disorder”\textsuperscript{53} refers to a range of protein structures, from fully folded (globular proteins) to completely unfolded (random coils or completely disordered proteins), with intermediary states featuring proteins with disordered regions (Figure 8). Globular proteins, such as hemoglobin and albumin, are compact, mostly spherical proteins with a structured tertiary formation, which is believed to be crucial for their function. In contrast, proteins with intrinsically disordered regions (IDRs) are in a constant state of fluctuation. These IDRs can either be independent protein molecules or segments of larger proteins with irregular structures. Random coils or fully intrinsically disordered proteins (IDPs) lack any stable, specific structure, allowing them to continuously shift between various conformations. This lack of structure provides these proteins with the flexibility to interact with a multitude of molecules.

This continuum is not a rigid classification; proteins can transition from one state to another through a process called disorder-to-order transition, often triggered by binding with other molecules. For instance, extended IDPs (both intrinsic coils and intrinsic pre-molten globules) and ordered proteins in the pre-molten globule intermediate state both exhibit properties of squeezed coils, since water is a poor solvent for a polypeptide\textsuperscript{54}. 

20
Figure 8. Continuum of intrinsic disorder from a globular protein to protein containing intrinsically disordered regions and finally a random coil, or a fully disordered protein. Adapted from 53.

Tandem repeats

A significant number of protein-coding genes contain repeated motifs. According to the latest estimates, up to 60% of proteins in the human proteome have regions with such motifs55. These regions are called **tandem repeats** (TRs) and are characterized by the unit size, number of the repeated units (length) and the similarity between the units.

At its emergence due to some duplication or insertion event, a TR tends to be perfect - with every unit having exactly the same sequence. Subject to point mutations, indels and other evolutionary events, repeat units diverge from each other. Different studies tend to allow a variable amount of diversity between repeat units, some claiming that more divergence gradually leads to the loss of a TR through the loss of structure56.

TRs in proteins tend to form a specific type of structures57 that can be roughly classified into the following categories (see Figure 9 for examples):

1. Fibrous structures formed by units of 3-7 amino acids, stabilized by inter-chain interactions
2. Solenoids - elongated structures with unit sizes 5-40 amino acids
3. Toroids - closed structures with repeats of 30-60 amino acids
4. Beads on a string - repeats with units over 50, that contain certain stable domains, connected in a consecutive fashion

Proteins with tandem repeats can play a structural role, like the alpha-helical coiled coils, or serve as protein-protein interaction modules. Another well-known example of tandem repeats is the zinc-finger proteins, which bind DNA to regulate gene expression.
Short Tandem Repeats (STRs)

**Short tandem repeats (STRs)** in proteins consist of just one or two amino acids and at the DNA level can be coded by 1-6 repeating nucleotide bases. STRs are also referred to as microsatellites or low-complexity regions. They usually don’t form any predetermined structures, although some amino acids aggregate into crystalline clusters or form structural fibers. STRs are highly abundant in the human genome. When considering non-coding regions, these short repeated sequences constitute 3% of the human genome. They are highly diverse due to lower evolutionary pressure and this property has been used to employ STRs in genetic profiling of individuals for forensic or archaeological purposes.

In the coding regions of the human genome, STRs are often found in promoter, enhancer or intron regions and variation in repeat copy number can impact gene regulation through a variety of mechanisms, including modifying transcription factor binding sites, altering DNA methylation patterns or influencing gene expression. In many proteins, larger expansions of the number of repeated units in STRs, lead to misfolding or toxic aggregation and cause severe diseases, such as Huntington’s Disease (the mechanism described in Figure 10) and Fragile X Syndrome. More subtle changes have been shown to alter complex phenotypic traits or serve as genetic biomarkers for different conditions.
**Figure 10.** STRs at the DNA and protein levels illustrated by the CAG/PolyQ repeat in the Huntington gene. The head (colored in green and orange) of the Huntington (HTT) protein structure, which resembles a tadpole, continuously grows larger as the number of glutamine repeats increases. HTT structure adapted from\textsuperscript{64}.

In some cancers, the mismatched DNA repair machine consisting of a set of genes (MLH1, MLH3, MSH2, MSH3, MSH6, PMS1 and PMS2) can be disrupted due to mutations in these genes or their altered expression. This leads to the accumulation of expanded STRs or other errors in these regions not being corrected. This condition is called **microsatellite instability (MSI)** and, in colorectal and breast cancer, it is used as a predictive marker and patient stratification tool. MSI tumors also have a distinct tissue phenotype and can be diagnosed using tumor imaging (See Figure 11 illustrating this phenomenon in colorectal cancer). The precise mechanism of microsatellite instability is still under active research and its role in patient stratification and treatment selection is still debatable.
Figure 11. Microsatellite instability (MSI) in colorectal cancer and the three aspects of studying it. MSI can be diagnosed by DNA testing either through NGS sequencing or a simpler PCR test accessing a certain known number of loci. Corresponding tumor phenotype can be identified by image analysis with the help of AI and, finally, characteristic tumor microenvironment markers can also signal the MSI status of the tumor. Illustration of tumor imaging and AI classification is adapted from 65.

Intrinsic disorder and STRs

Short tandem repeats (STRs) are strongly associated with unstructured regions. Repeated motifs found in IDR s tend to evolve rapidly, leading to repeat expansion events66, which create elongated unstructured regions that still have low evolutionary pressure. An abundant number of biological processes benefit from this overlap. IDRs and TRs enable protein-protein interactions with a large number of partners in protein hubs67, assist in multiprotein complex scaffolding events68, regulate 3D chromatin formation69, alternative splicing and transcription factor binding. The list just goes on. In many cases, such repeat expansion events can also lead to translation inhibition or aggregation of mutant proteins. These classes of repeats are associated with so-called repeat expansion disorders in humans. For example, SCA1 is caused by the expansion of CAG repeats in the gene-encoding ataxin 1 (ATXN1), which results in polyQ expression.
Next-generation sequencing technologies (NGS)

The development of the Next-generation sequencing (NGS) technologies has largely enabled most of my research, providing a cost-efficient way to obtain large amounts of sequencing data. However, it is important to be aware of its caveats, especially in relation to repeated DNA sequences and how they are handled by bioinformatics software processing NGS data. Repeating DNA sequences can introduce and propagate errors originating during the sequencing itself or later on when trying to assemble the short reads.

In Sanger sequencing, each read comes with a fluorescent peak trace chromatogram, allowing researchers to verify correct base incorporation, especially in challenging areas like repeats. High-throughput sequencing technologies also provide similar information as quality scores, but the sheer volume of data makes manual quality checks impractical.

The most commonly used sequencing technology is Illumina sequencing, known for its low error rate (<0.1%). However, Illumina reads are relatively short (<250 bp), limiting their ability to resolve longer repeat regions effectively.

Genome assembly is the process of reconstructing a complete genome from sequencing reads and other data sources, like linkage maps. Two major approaches are the de Bruijn graph and the overlap/layout/consensus (OLC) methods that differ in how they handle repeat regions during assembly.

De Bruijn graph-based methods require repeat regions to be shorter than a specified k-mer length (usually between 21 and 96, often set at 31). For example, ALLPATHS-LG collapses repeats longer than 96 during initial processing but can expand them later. Newer implementations like SPAdes and SKESA use multiple k-mers to improve assembly but are not designed for large genomes like plants or vertebrates.

OLC methods can resolve repeats shorter than the read length and are not restricted by k-mer size. Historically, Illumina reads were assembled with de Bruijn methods due to the computational demands of OLC. However, with the cost of long reads like PacBio and Nanopore decreasing, most genome projects now use OLC-based assemblers, resulting in more complete genomes with better repeat resolution. In papers IV and V we still had to work with the Illumina platform reads and mitigate challenges with repeat assembly using bioinformatics methods.
Computational methods

Multiple sequence alignments (MSAs)

MSA is generally a method of arranging biological sequences (protein or nucleic acid) to identify regions of similarity. Identical or similar residues in the same columns are aligned together, and gaps are added to compensate for the differences. By combining a global MSA with phylogenetic information, evolutionary events can be studied in detail (Figure 12). There are many tools available for performing an MSA on protein sequences. Clustal Omega is one of the rapid MSA tools\textsuperscript{79} that uses seeded guide trees and HMM profile-profile techniques to generate alignments. It is suitable for large multiple sequence alignments needed for a genome-wide evolutionary study.

![Figure 12](image)

**Figure 12.** Schematic representation of how evolutionary events in protein sequences can be studied with the help of a multiple sequence alignment. In this example from Paper I, we made alignments of homologous proteins of the two closely related species of *Drosophila* and an outgroup species of *Aedes aegypti* with Clustal Omega to identify specific insertions and deletions in the protein of interest.

Another MSA tool, T-Coffee\textsuperscript{80} (Tree-based Consistency Objective Function For alignment Evaluation), uses an approach that is quite different from other alignment algorithms. T-Coffee creates a ‘library’ of pairwise
alignments, which includes both the alignments encountered during the progressive alignment step as well as additional pairwise alignments. This library is then used to create the final multiple alignment using a consistency-based approach. This means that the score of an alignment is calculated based on how well it fits with the pairwise alignments in the library, thus capitalizing on the additional information provided by these pairwise alignments. This approach allows T-Coffee to produce alignments of superior quality\(^1\), both when aligning sequences that share only distant similarity and very closely related sequences. However, its computational intensity makes it less suitable for very large datasets.

Homology detection and orphans

The essential step for studying protein evolution from sequence data is the detection of homologous protein sequences. Sequence similarity searching is the main method of identifying homology. It involves searching the target sequence in the database of other known proteins and statistically assessing how well the sequences match one another. One of the most popular sequence similarity search tools is PSI-BLAST. Another approach to inferring homology is implemented by HHblits\(^2\), a general-purpose tool that represents both query and database sequences by profile Hidden Markov Models (HMMs). Compared to PSI-BLAST, HHblits is faster and has 50-100% higher sensitivity\(^3\).

In the case of orphan gene identification, also known as de novo genes, careful homology search is an essential step of a bioinformatics pipeline. Orphan genes are expected not to return any significant hits from the database in homology search. If the predicted gene has no homologs, it is often necessary to examine the homology closer; for instance, explore potential related structural domains with HMMscan\(^4\) and perform homology searches on both DNA and protein sequences (Figure 13), as well as on the expanded surrounding genomic context.
Protein structure and disorder prediction

In-silico protein structure identification includes many approaches to addressing the protein folding problem from a computational perspective. One of these methods is related to studying protein evolution. It involves restricting the search space of possible 3D structures by identifying the key residues that are in contact with each other in the resulting structure. Information we get from contact predictions can be beneficial in identifying structural constraints on protein evolution and key functional sites in flexible disordered regions.

Earlier methods of protein contact prediction have relied on large sets of multiple aligned sequences and would investigate correlations between columns in these alignments in order to identify indirect interaction between them in the structure. A more modern approach involves deep learning algorithms, like convolutional neural networks (CNNs) and recurrent neural networks (RNNs). These methods extract complex features from protein sequences and integrate co-evolutionary and structural information to predict protein contacts and structure. The most popular deep learning-based method for protein structure prediction is AlphaFold developed by the AI research lab at Google's DeepMind. AlphaFold’s neural network has two distinct modules: the first, Evoformer, predicts contact maps, which describe the distances between residues in the protein sequence, and the second uses a neural network to generate a 3D model of the protein based on these contact maps. AlphaFold generates a probability distribution for every residue, indicating how likely it is to be in a structured region. Residues with lower probabilities of being in structured regions are more likely to be located in IDRs.
Figure 14. Example results of the intrinsic disorder prediction for the Myc-proto oncogenic protein in Human (Uniprot ID P01106) given by the three predictors - IUPred\textsuperscript{87}, using energy function estimates, MobiDB\textsuperscript{88} using a consensus approach and, finally, its predicted structure by AlphaFold\textsuperscript{88}.

While AlphaFold has made significant advances in predicting protein structure and is undoubtedly the leading prediction tool, there are other methods for intrinsic disorder prediction that are still useful today. These include IUPred, a sequence-based predictor that calculates energy estimations of the single amino acids in the sequence to identify if they promote order or disorder\textsuperscript{87}. These energy estimations are derived from known contacts between amino acids in structures of globular proteins. There are machine learning-based disorder predictors that outperform IUPred in accuracy. However, since it does not rely on evolutionary information, it is suitable for use on proteins without known homologs. Another method that is worth mentioning is fDPnn\textsuperscript{89}, a recently developed machine-learning model that integrates several putative disorder properties, including IUPred energy functions, and produces accurate predictions even for de novo proteins.
TOP-IDP Scale

Soon after the discovery of IDPs, researchers noticed that there are significant differences in the abundance of amino acids in unstructured vs structured regions\(^90\). IDPs have been shown to be significantly depleted in bulky hydrophobic (Ile, Leu, and Val) and aromatic amino acid residues (Trp, Tyr, and Phe), which would normally form the hydrophobic core of a folded globular protein, and also possess low content of Cys and Asn residues. These depleted residues (Trp, Tyr, Phe, Ile, Leu, Val, Cys and Asn) were proposed to be called order-promoting amino acids. On the other hand, natively unfolded regions are enriched in polar amino acids (Ala, Arg, Gly, Gln, Ser, Pro, Glu, and Lys), that have been named disorder-promoting\(^91\).

Since then, many attempts have been made to classify amino acids in relation to their tendency to be in a disordered region. One of the classifications that proved accurate in discriminating between order and disorder is the TOP-IDP scale\(^92\). TOP-IDP was constructed by surveying 517 different amino acid scales describing different properties of amino acids. The simulated annealing procedure has been subsequently performed on this data to optimize the scale values to achieve higher power to discriminate between order and disorder. This power is measured as the area ratio value (ARV) formed by two conditional probability curves, one describing the probability of disorder \(P(\text{sd} \mid x)\) and one for order \(P(\text{so} \mid x)\), given the attribute value \(X\). The ranking of amino acids in TOP-IDP is presented in the illustration below (Figure 15). TOP-IDP scale has been used in publications in this thesis to quickly illustrate the disorder propensity of different amino acids in graphs, but it has also become a feature in many disorder predictor methods.
Figure 15. TOP-IDP scores\textsuperscript{92} for each amino acid, sorted by increasing disorder-propensity.

Genotyping tandem repeat variations

As described in the previous chapters, STR regions are characterized by sequencing errors, thus short-read NGS pipelines have been built to ignore these types of repeats due to challenges of assembly of repeats spanning regions longer than read length\textsuperscript{93}. For a long time, we could reliably assess the length of repeats only fully encompassed within the read length. HipSTR\textsuperscript{94} is an example of a method that accurately annotates STRs on each allele, taking into account both the motif and the number of repeated units. HipSTR uses a haplotype-based approach and accurately genotypes repeats encompassed in the read-length, but is not suitable for larger repeat expansions spanning several reads.

However, a variety of computational approaches have been recently developed to profile STRs in sequencing data despite the short read pipelines’ limitations. GangSTR\textsuperscript{95} and ExpansionHunter\textsuperscript{96} are two notable tools that incorporate multiple properties of paired-end reads into machine learning models that allow the genotyping of longer repeats and estimate the accuracy of the estimation. ExpansionHunter developed by Illumina is optimized specifically for finding larger repeat expansions that occur throughout the genome and doesn’t require any previous knowledge on the location of repeats. ExpansionHunter is not suitable for the whole-exome sequence data or detecting subtle STR variation\textsuperscript{97}. 
GangSTR requires a reference panel of known STR locations (Figure 16). This means that, prior to genotyping, one needs to run a repeat annotation pipeline on a reference genome and supply the results in a specific format to the genotyping tool. Repeat annotation in itself is not a trivial task, often producing contradictory results. Therefore, it is important to consider the correct annotation method to use depending on the goals of the project.

![Figure 16. Overview of the STR genotyping pipeline as used in papers IV and V. Variant catalog is generated using TRAL repeat annotation software on a reference genome. Along with sequencing data, it serves as an input for the STR genotyping software GangSTR, which uses a likelihood model to resolve genotypes from paired-end reads. Results are reported as a VCF file for each genotyped individual.]

Tandem repeats annotation

There are two main categories of methods that annotate tandem repeats. The first approach is to identify repeats from the sequence based on homology search or structural domain assignment. These methods often miss the highly diverged protein repeats that still maintain characteristic structures. This problem has been addressed by developing methods that are agnostic to prior knowledge of repeat motifs and can be based on clustering, like XSTREAM or T-REKS or hidden Markov models, like HHRep. To combine outputs of different methods that specialize on the identification of
different types of tandem repeats, researchers can also use a meta-search and scoring method, like TRAL\textsuperscript{101} that also takes evolutionary information into consideration. Alternatively, structure-based annotation methods rely on the modularity of known PDB structures to find repetitive elements and then map them back to the sequence. Such methods like RAPHAEL\textsuperscript{102} and TAPO\textsuperscript{103} are superior for the identification of structural repeats, but completely ignore low-complexity and unstructured repeats.

**Biological databases**

Biological databases serve to enable access to experimental and predicted data concerning genes, proteins, evolutionary relationships and interactions. These databases provide efficient retrieval and integration of different sources of information and we can often combine data obtained from various databases using standardized unique gene or protein identifiers. This section introduces different biological databases that have been used throughout my research projects. In paper IV, I present a new biological database of short tandem repeat variation and describe how it has been developed and integrated with other canonical databases.

**Genes duo: Ensembl & NCBI/Entrez**

Due to historical reasons, there are two databases and reference systems for genomic information. Ensembl\textsuperscript{104} is a joint project between EMBL-EBI (European) and the Sanger Institute that provides sequence information and annotations on a selected set of eukaryotic genomes. Entrez is the name of the US NCBI infrastructure, which provides access to various genomic information for a broad set of organisms\textsuperscript{105}. These databases use different numbering systems for the base pairs in the DNA sequence: Ensembl is 1-based and NCBI is 0-based. Depending on organisms and other goals of the research project, one might prefer one system over the other, but, in most of my projects, the Ensembl identification system is used for human genes.

**Protein trio: Uniprot/Swiss-Prot, PFAM & PDB**

The first database that is essential for computational research on proteins is Uniprot\textsuperscript{106}. Uniprot is a comprehensive resource that provides protein sequences and their high-level annotations, like function and domain structure. Swiss-Prot is the curated part of Uniprot that contains manually verified entries\textsuperscript{107}. For structural information, Uniprot links to PFAM, a protein family annotation resource that provides multiple sequence alignments for related protein structural domains\textsuperscript{108} and PDB, which contains the structures of proteins that are sourced from experimental data\textsuperscript{109}. 

33
Evolutionary databases: InParanoid and OrthoDB

Traditional phylogenetic trees have been built using the morphology of organisms or fossils. In molecular evolution, phylogenetic trees are used to describe the historical relationships and events between groups of genes and proteins. The genetic sequence serves as a footprint of all accumulated mutations and, by studying sequence similarity between genes from different species, the evolutionary history can be revealed from the genes like from a molecular fossil.

There are two main approaches to identifying orthologous genes: (i) from a reconciliation of gene trees with the species phylogeny and (ii) from the classification of all-against-all sequence comparisons of complete genomes.

Two main resources for genes’ evolutionary information are InParanoid and OrthoDB. The InParanoid project gathers proteomes of completely sequenced eukaryotic species plus Escherichia coli and provides phylogenetic relationships between them. It is specifically aimed at discovering orthologous genes and paralogous genes that arose by duplication after some speciation event. Such protein-coding genes are called “inparalogs”. The pairwise orthologous relationships among these proteins are found using BLAST, followed by a clustering step using the InParanoid algorithm. OrthoDB is also a catalog of orthologous protein-coding genes and provides information on evolutionary relationships across vertebrates, arthropods, fungi, plants, and bacteria. OrthoDB constructs orthologous groups at different taxonomic levels, corresponding to the last common ancestors of the species considered. It uses a hierarchical clustering algorithm, resulting in more comprehensive ortholog classifications. Depending on the goals of the project, these two databases can complement each other or be used independently.

Protein disorder resources: Disprot and MobiDB

The DisProt database contains experimentally identified disordered proteins. Current release DisProt 9 includes 2038 proteins and 4,477 pieces of evidence of state transitions, interactions and functions. All proteins in the database are annotated with accession numbers from UniProt, Swiss-Prot, NCBI and other databases. The second resource, MobiDB, extends the content provided by Disprot and other protein structure databases with the goal of providing a centralized source for data on disordered regions in protein structures, featuring full coverage of the SwissProt database. Annotations from other disorder predictors (ESpritz, IUPred and DisEMBL and more recently also AlphaFold) are also included in the database to provide information in cases where experimental annotations are not available. MobiDB includes 26,933 experimentally annotated proteins and, in its latest release, all Swiss-Prot and model organism proteins are processed and stored in the database, for a total of 1,121,068 entries.
Summary of results

Protein expansion is due to long indels in disordered regions (Paper I)

The goal of this project was to explain the length differences between homologous proteins. We could observe that the difference in protein length and the difference in the number of intrinsically disordered residues are strongly linked in all datasets studied (Fungi, Mammals, Nematodes and Insects). The strong coupling between length difference and disorder content is clear regardless of the disorder classifier used to predict unstructured regions and across all ranges of evolutionary distances. Using IUPred and an evolutionary distance of one, i.e., one substitution per residue, a completely ordered protein shows a length difference of about 3%, while the proteins with more than 30% disordered residues have a length difference of about 6%.

We explored three plausible explanations for the correlation between the protein length differences and disorder. In the first scenario, the event of insertion/deletion could lead to a protein becoming increasingly disordered. Another option is that disorder could already be an inherent characteristic of the protein region before the insertion took place. Finally, the disordered content could potentially be explained by the general content in the proteome, rather than being related to specific regions. To investigate different courses of evolutionary events, we developed three statistical models to explain the expected variance in the number of disordered residues, \( \delta D \), given a length difference of \( \delta L \).

The proximity model was shown to explain a large part of the observation for the IUPred predictor. In this model, we assumed that the disorder content of a particular indel is determined by the disorder content of the neighboring residues, i.e., that an unstructured region tends to grow with evolutionary time. We could also show that there are no significant differences between insertions and deletions in terms of disorder content.

Our results indicate that indels located in disordered regions are comparatively less influenced by purifying selection than their ordered counterparts. We demonstrate that regions that are intrinsically disordered have a greater propensity for expansion and contraction compared to other
regions. However, the majority of insertions and deletions in disordered regions typically occur within these already disordered areas.

**High GC content causes orphan proteins to be intrinsically disordered (Paper II).**

The aim of this project was to investigate the potential causes of the differences in disorder content among the very young and orphan genes in *Saccharomyces Cerevisiae* and *Drosophila Melanogaster*. Based on our previous research on intrinsically disordered regions, we followed the intuition to quantify the occurrence of G+C nucleotides in these genomes of interest. This is due to codons that contain G+C coding for amino acids that, on the TOP-IDP scale, are regarded as disorder-promoting. Indeed, the genome of Drosophila Melanogaster has, in general, higher GC content than *Saccharomyces Cerevisiae*, but to dive deeper into specific genes, we ran the ProteinHistorian pipeline on around 200 related eukaryotic species. We divided the protein-coding genes into four age groups: ancient, intermediate, genus-specific orphans and true orphans. We then could show that the relationship between GC content and disorder is more prominent in young genes across all species of interest, and we speculate that there is selective pressure to maintain disorder content in the genome, leading to new proteins acquiring more definite structures with time.

**Protein tandem repeats and their relationship with intrinsic disorder (Paper III).**

This study is an overdue update of the census of protein tandem repeats in SwissProt that was published in 1999. Since then, the count of annotated tandem repeats has grown 7-fold; over 50% of proteins in the curated catalog have been found to contain tandem repeats. This was possible largely due to improved prediction methods and the protein sequence database's growth. In this project, we used a meta-annotation approach to find TR regions in the sequence. We then explored different properties of these regions, including intrinsic disorder content and statistics on the occurrence of different amino acids in the sequence. One of the goals of the project was to examine and quantify the interplay between TRs and intrinsically disordered regions (IDRs). Only a small portion of residues (2.5%) was found to be both in TRs and IDRs, with the
majority of those being TRs fully encompassed within the IDRs. This overlap includes proteins with many essential functions. This paper also emphasizes disorder-promoting homorepeats, which are linked to human diseases and form a significant portion of tandem repeats overlapping with intrinsically disordered regions. Our results show that these homorepeats, particularly those with disorder-promoting amino acids (AAs), tend to be longer within IDRs. Proteins with long polyQ stretches are more prevalent in complex species with rich protein interactions, aligning with the general role of intrinsically disordered TRs in transcriptional regulation and protein-protein interactions.

Database of short tandem repeat variation in humans. (Paper IV).

WebSTR is a database and a web portal initially started as a resource to manage our data for a research project on colorectal cancer but has further been expanded to present broader variations in the human genome. Short tandem repeats (STRs) span around 3% of the human genome, are enriched in gene regulatory regions and can impact gene regulation through various mechanisms. Larger expansions of STR repeat units have been shown to be implicated in disorders like Huntington’s Disease, but they are also accounted for in complex phenotypic traits, like neurodivergence, and play an important role in many cancers. Although STRs have traditionally been excluded by short-read NGS pipelines due to polymorphism and sequencing errors, recent advancements allow accurate genotyping of STRs from sequencing data. WebSTR presents STRs variation data obtained using tools like GangSTR, ExpansionHunter, and HipSTR that can genotype both normal and expanded repeats. WebSTR is a comprehensive resource for genome-wide STR variation, offering a web interface and programmatic access for exploring allele frequencies, mutation rates, and trait associations, contributing to the study of genome-wide STR variation in healthy human samples as well as in colorectal cancer. We also provide a pipeline and means of expanding our catalog in the hope that it will serve a broad community of genomic and medical researchers.
Short tandem repeat mutations regulate gene expression in colorectal cancer (Paper V)

Paper V illustrates the potential of performing a systematic study of how stepwise STR mutations affect human phenotypes using pipelines and data described in Paper IV.
Here, we present evidence of gene expression changes mediated by mutations affecting STR lengths in colorectal cancer (CRC). To perform this study, we first created a new reference panel of STR loci for all human protein-coding genes. Then, we genotyped these reference STRs, calling on the whole-exome sequencing (WES) data of 412 patients diagnosed with CRC from The Cancer Genome Atlas (TCGA)\textsuperscript{116}. Through this analysis, we could confirm many previously reported factors that influence microsatellite instability (MSI) characteristics for this disease.
Additionally, we identified a group of expression STRs (eSTRs) in CRC tumors, where the lengths of alleles are correlated with the expression levels of neighboring genes. We have shown that this collection of eSTRs enables us to predict changes in gene expression resulting from somatic STR mutations in patient-matched samples. Furthermore, we observed a higher rate of mutations in these eSTRs in MSI tumors. This suggests that eSTR mutations may be positively selected for in cancer under certain conditions.
While further data is required to confirm which specific eSTRs play a causal role in regulating gene expression, our discoveries shed light on a relatively unexplored mechanism that influences tumor phenotypes. This set of interesting STR loci can serve as a starting point for further microbiology and pharmacological studies on colorectal cancer.
Concluding remarks

Tandem repeats and intrinsic disorder are the two features of proteins that can co-occur and aid each other in performing important functions in eukaryotic genomes. Only a small fraction of protein sequences exhibit the overlap of repeated sequences and protein disorder, but many of these proteins are essential and have important functions like host-pathogen interactions, as described in Paper III. Genetic variation in intrinsically disordered regions is responsible for protein length expansion in eukaryotic genes (Paper I). Many orphan, newly evolved proteins are intrinsically disordered due to high GC content (Paper II). This links intrinsic disorder closely to the process of rapid evolution and their ability to take up new functions.

A more practical medical science example of this overlap is the role of short tandem repeat variation in human disease. STR expansions can lead to the production of proteins with intrinsic disorder, like in neurodegenerative disorders, where expanded STRs within certain genes can result in abnormally long stretches of repetitive amino acids in the corresponding protein, which promotes misfolding and aggregation. In many human cancers, the mismatch-repair mechanism is impaired, which leads to STR expansions in many different locations of the genome. Paper IV presents a comprehensive resource of human STR variation, and Paper V illustrates how it can be used to identify loci of interest, like in the case of colorectal cancer, where variations in a certain set of STRs lead to altered gene expression patterns in tumors. These findings advance our understanding of protein intrinsic disorder and tandem repeats and pave the way for further studies on potential biomarkers and drug targets based on genetic variations in these proteins.

The role of bioinformatics in this field is extremely important, and I would like to further highlight it. Not only do computational resources and tools enable the processing of large amounts of data that come from NGS experiments or allow running evolutionary algorithms on hundreds of thousands of sequences in a matter of hours, but machine learning models also enable us to mitigate technological limitations and still obtain statistically significant results, as seen in the case of genotyping STRs. All of the projects described in this work relied heavily on high-performance cloud computing and machine learning models.
Future outlook

With the last decade of advancement in both sequencing technologies and the rise of deep learning in bioinformatics, we are stepping into a new era where these powerful tools can together expand our knowledge of biological processes. Deep learning and the release of AlphaFold have significantly influenced the field of intrinsic disorder\textsuperscript{117}, offering a new way to accurately predict disordered regions and interaction sites in proteins. With an expanded set of known intrinsic disorder regions, our understanding of the energy landscape of these proteins and their dynamic behavior will inevitably grow. The new frontier in this field is predicting complex conformational ensembles based on deep learning. It will allow drug discovery projects to target intrinsically disordered regions, which still remains challenging today\textsuperscript{118}.

Long-read sequencing (LRS) technologies such as PacBio and Oxford Nanopore have rapidly advanced in recent years as well and are becoming more accurate and cost-effective\textsuperscript{119}. These methods will allow researchers to seamlessly genotype large STR expansions in the future. This can lead to a more accurate estimate of already known STR variations, but maybe even discovering and quantifying more subtle changes or novel STRs. LRS, along with spatial transcriptomics data and single-cell sequencing, will open up a completely different level of precision when performing GWAS in health and disease\textsuperscript{120}. Not only am I hopeful for the development of personalized cancer treatments derived from STR profiling due to the advancement of our research through sophisticated sequencing, but I am also confident that we will witness an increase in evidence supporting the significant roles of STRs in the human genome, beyond just their pathogenic functions.
Acknowledgements

Writing this thesis more than 10 years after I, for the first time, stepped into the oval buildings of the newly built SciLifeLab (and accidentally calling it SciFiLab at the interview) feels completely surreal. It hasn’t been a straightforward journey presenting the challenges both in my scientific explorations and in personal life. I would like to acknowledge everyone who has been with me on this rollercoaster.

My supervisor Arne Elofsson. Thank you for seeing potential in me during our first meeting, for encouraging me in the many times I was prepared to quit this whole idea, and for accepting me back after my long sabbatical. I honestly don’t know where you find so much patience and wisdom for your students.

Elofsson and Lindahl’s Labs of 2013-2016 were a great bunch of people - Ruan Sagit, Walter Basile, Minttu Virkki, Karolis Uziella, Christoph Peters, Marco Salvatore, Enrichetta Miletta, Mirco Michel, Sara Light, Sikander Hayat, David Rodriguez, Anirudh Ranganathan. I enjoyed our discussions on string theory and deep learning during lunch breaks and truly miss the atmosphere of SciLifeLab back then.

Special thanks go to Kostas Tsirigos for supporting me during the tough times in the hospital and becoming my friend, for encouraging me to push through and for proofreading this text. Ευχαριστώ πάρα πολύ.

This thesis would never have been completed without the guidance of Maria Anisimova, who became my co-supervisor beyond borders. I have had an amazing time at ZHAW, and the beautiful Zurich is forever in my heart. Maria, you are my inspiration and a role model.

Maria’s Lab 2016-2023 - Elke Schaper, Jūlija Pečerska, Max Verbiest, Feifei Xia and Tugce Bilgin Sonay. Thank you for our collaborations and especially Max, for amazing teamwork through remote communication challenges.

Melissa Gymrek is another amazing woman and my collaborator. Thank you for great discussions while working on WebSTR, I am so happy that we managed to do it, and I hope it will serve your research in the future.

I would also like to thank Fredrik Ahlgren for providing me with another academic home at Linnaeus University in Kalmar during the last months of my PhD.

My friends outside of the University, Anna Kazakova, Andrey Ershov and Timofey Mucha. Thank you for being there through all these years, listening
to my rants, and genuinely trying to understand why proteins are not just what you bring with you to the gym. Since then, Anna became a data scientist, Andrew started with bioinformatics and even contributed to WebSTR, so maybe it is a bit contagious? Well, Tim already managed with a PhD way before me, and I still don’t fully understand what his thesis was about.

**Fadi Bitar** and **Pernilla Hallerdal** (and Benji) - your contribution to this thesis is extra valuable as you guys have been my Stockholm base after the move. Thank you for all the warm conversations, laughter, burlesque, photography and delicious food. I think anyone would want to complete their PhD if they had a chance to stay with you while doing it.

**Maria Wool**, thank you for all the inspiration and support and for making an amazing illustration for the cover of this book.

**Erika Lundby** and **Ella at Happy Dance Kalmar**, thank you for supporting me and giving me a chance to put myself out there more through dance.

Finally, behind every successful woman stands her man. **Johan**, I love you and would like to thank you for always being there (don’t know where you find your patience and wisdom either). You should seriously get an honorary doctorate for all the help with this thesis.

This work would have probably been completed a lot sooner if not for my sons **Vidar**, **Viggo** and **Ville**. But maybe they will one day read it and be a little proud of mama and appreciate the time I took to be with them too.

**Ulf** and **Lena**, thank you for accepting me into your family and for watching the three Vs so I can work on my thesis.

My brother **Alex** and my sister **Ioulia**, I am forever the “big sis“ who knows all the answers and can solve all the problems, but believe me, you give me the strength to keep going.

I would also like to thank my parents for giving me a good start despite the difficult circumstances and for setting an example of perseverance. I wouldn’t be here without you.

**Galina Galkina** is another person without whom you wouldn’t be reading this thesis now. Thank you for everything that you do and for becoming my mentor.

And last but not least to my grandma **Lidiya**, who has always been my biggest fan and, unfortunately, didn’t live to see the day this goes to print.

Thank you.
References

15. Functional residues on the surface of the N-terminal domain of yeast
34. Dehal, P. & Boore, J. L. Two rounds of whole genome duplication in


68. Luo, Y. *et al.* mRNA interactions with disordered regions control...


