DISS. ETH NO. 21851

The Evolution of Protein Tandem Repeats

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH Zurich

(Dr. sc. ETH Zurich)

presented by

ELKE SCHAPER

Master of Science, Complex Adaptive Systems, Gothenburg University

born 17 July 1986

citizen of Germany

accepted on the recommendation of

Prof. Dr. Lukas Sebastian Bonhoeffer, examiner
Dr. Maria Anisimova, co-examiner
Prof. Dr. Niko Beerenwinkel, co-examiner
Prof. Dr. Erich Bornberg-Bauer, co-examiner

2014

SUMMARY

Genomes of species from all domains of life are rich in repetitive sequence motifs. Tandem repeats, TRs, are repetitions that occur right next to each other in the sequence. This general definition comprises everything from copy number variations, repetitions that may cover several megabases of nucleic sequence, to microsatellites, repetitions of a single or few base pairs. TRs appear in coding sequence and then are frequently translated to protein TRs. Protein TRs span a subcosmos of sequence diversity in their own right: There are TRs of entire protein domains. There are shorter TRs that lack intrinsic stability, but form stable domains when they appear in tandem on the sequence. There are repetitions of single amino acids – homorepeats – that are often unstructured due to the strong local concentration of a single physiochemical property. For the bulk of TRs however, neither structure nor function are determined.

TRs evolve by duplications and losses of single or multiple TR units. A famous example is the protein Huntingtin, where the wild-type contains a glutamine homorepeat with 6-35 TR units. During autosomal cell proliferation, TR duplications can increase the number of adjacent glutamines, leading to the onset of Huntington's disease. A second example are Zinc fingers. They constitute one of the most abundant protein motifs in metazoan transcription regulating proteins. Duplications and losses of TR units have created Zinc finger TRs literally covering the whole range of TR unit numbers from individual copies within a protein to about 40 TR units. Whereas much research has gone into a few particular TRs, the mode of evolution is currently unknown for the vast majority of TRs.

The aim of this thesis was to gather comprehensive knowledge on the evolution of the whole spectrum of protein TRs. Are protein TRs generally fast-evolving, similar to their nucleic cousins, the tiny microsatellites? If yes, this would make a strong case for a role

of protein TRs in adaptive protein evolution: If the phenotypes connected to different TR confirmations are under selectional pressure, fast mutation rates would allow to quickly adapt to changing outer environments. Are protein TRs generally conserved, such as is often the case for protein sequence? If this were true, the TR region would likely be of structural importance to the protein function. Or are TRs left to evolve freely within the protein?

Chapter by chapter, I address these questions in my thesis. My first mission was to establish trustworthy proteome wide TR annotations for a large range of species. Chapter 2 discusses the hurdles with *de novo* TR detections, combined with a benchmark of current algorithms. Further, I introduce statistical tests to best discern false positive from true positive TR annotations. To accurately annotate TRs, I propose a meta-tool, combining predictions from multiple sources followed by a validation with the presented tests.

My second mission was to compile a framework to research the mode of evolution of TRs. The fundamental idea is to backtrack the history of duplications and losses of TR units by a comparative analysis of TRs from different samples. Chapter 3 presents a model-based approach to annotate homologous TRs for this purpose. Further, I describe a novel phylogenetic method allowing to discern how long TR have been conserved, and when the last TR unit duplications and losses occurred.

My third mission was a proteome-wide analysis of protein TR evolution for two data sets. Since human TRs are of interest owing to the large number of diseases caused by TR expansions/contractions, I first investigated the evolution of human TRs in a comparative study across the eukaryotes. Chapter 3 contains the results of this study, peppered with an analysis of the correlation of the mode of protein TR evolution, and functional and structural aspects of the TR. Second, I studied the evolution of protein TRs for 25 plant species, this time incited by the particularly strong challenges for pathogen and stress resistance for sessile species. Are (some) protein TRs functional in resistance related plant genes? How do protein TRs in metazoans and plants compare with respect to evolution and functionality? In Chapter 4, I look into these question, whilst presenting the results on plant protein TR evolution.

ZUSAMMENFASSUNG

Die Genome von Spezies aller Domänen sind reich an repetitiven Sequenzmotiven. Benachbarte Wiederholungen, BWs, sind Wiederholungen die direkt nebeneinander auf der Sequenz liegen. Diese allgemeine Definition beinhaltet alles von „copy number variations“, Wiederholungen, die sich über mehrere Millionen Basenpaare DNA-Sequenz erstrecken, bis hin zu „microsatellites“, Wiederholungen von einzelnen oder wenigen Basenpaaren. BWs wurden auch in kodierender Sequenz gefunden, in welchem Fall sie zumeist im Zuge der Translation in Protein-BWs übersetzt werden. Protein-BWs selbst umfassen wiederum einen eigenen Subkosmos an Sequenzdiversität: Es gibt BWs von ganzen Proteindomänen. Es gibt kürzere BWs, die zwar nicht intrinsisch stabil sind, aber stabile Domänen formen, wenn sie nebeneinander auf der Sequenz liegen. Es gibt Wiederholungen von einzelnen Aminosäuren – „homorepeats“, die oftmals aufgrund der starken lokalen Konzentration einer einzigen physikochemischen Eigenschaft unstrukturiert sind. Für das Gros der BWs sind allerdings zu dato weder Struktur noch Funktion bekannt.

BWs evolvieren durch Duplikationen und Verluste von einzelnen oder mehreren BW units (Mit „BW unit“ ist eine einzelne Einheit bzw. ein einzelnes Motiv gemeint, das innerhalb eines BWs mehrmals wiederholt ist). Ein geläufiges Beispiel ist das Protein Huntingtin, dessen Wildtyp ein Glutamin-Homorepeat mit 6-35 BW units enthält. Während des autosomalen Zellwachstums können BW-Duplikationen zur Zunahme der Zahl der benachbarten Glutamine führen, was das Einsetzen von Chorea major Huntington, auch Veitstanz oder im Englischen „Huntington's disease“, auslöst. Ein zweites Beispiel sind Zinkfinger. Sie stellen eines der am weitesten verbreiteten Proteinmotive in Transkriptions-regulierenden Proteinen der Metazoa dar. Duplikationen und Verluste vom BW units haben Zinkfingerproteine mit circa 40 bis hin

zu einzelnen wenigen BW units hervorgebracht. Während einzelne BWs in der Forschung tiefgehend behandelt wurden, ist die evolutive Entwicklung des Großteils der BWs unbekannt.

Ziel dieser Dissertation war es, umfassendes Wissen über die Evolution des gesamten Spektrums von Protein-BWs zu sammeln. Evolvieren Protein-BWs in der Regel sehr schnell, ähnlich wie die verwandten DNA-Mikrosatelliten? Falls ja, wäre dies ein starkes Argument für eine Rolle von Protein-BWs in adaptiver Evolution: Wenn unterschiedliche Phänotypen, hervorgerufen durch unterschiedliche BW-Konfirmationen, unter Selektionsdruck stehen, würden hohe Mutationsraten eine schnelle Adaptation an veränderliche äußere Einflüsse ermöglichen. Sind Protein-BWs allgemein eher konserviert, wie oft der Fall für Proteinsequenzen? Falls ja, würden die BW wahrscheinlich bedeutsam für die Proteinstruktur und somit -funktion sein. Oder können BWs frei innerhalb des Proteins evolvieren?

Kapitel für Kapitel behandele ich diese Fragen in meiner Dissertation. Meine erste Mission war es, vertrauenswürdige proteomweite BW-Annotationen für eine breite Spannweite an Arten zu erstellen. In Chapter 2 werden die Schwierigkeiten von *de novo* BW Detektionen diskutiert, sowie ein Benchmark von gängigen Detektionsalgorithmen vorgestellt. Ferner diskutiere ich statistische Tests zur Unterscheidung von falschpositiven und korrekten BW-Annotationen. Um BWs akkurat zu annotieren, schlage ich ein Meta-Tool vor, das Detektionen aus mehreren Quellen bündelt, gefolgt von einer Validierung mit Hilfe der vorgestellten statistischen Tests.

Meine zweite Mission war es, ein Framework zur Erkennung des Modus der BW-Evolution zu erstellen. Die grundlegende Idee war es, die Geschichte der Duplikationen und Verluste von BW-units durch eine vergleichende Analyse der BW units von mehreren Samples zurückzuverfolgen. Hierfür wird in Chapter 3 eine modellbasierte Methode zur Annotation von homologen BWs vorgestellt. Weiterhin beschreibe ich ein neues phylogenetisches Verfahren zur Analyse der Dauer der Konservation von BWs, sowie des Zeitpunktes der letzten BW Duplikationen und Verluste.

Meine dritte Mission war eine Proteom-weite Analyse der Protein-BW Evolution anhand zweier Datensätze. Aufgrund des Interesses an menschlichen BWs wegen der vielen durch BW unit Duplikate/Verluste verursachten Krankheiten, habe ich zunächst die Evolution der menschlichen BWs innerhalb der Eukalypten in einer vergleichenden Studie behandelt. Chapter 3 beschreibt die Ergebnisse dieser Studie mitsamt einer Korrelationsanalyse des Protein-BW Evolutionsmodus und funktioneller und struktureller BW-Charakteristika. Zum zweiten habe ich die Evolution von Protein-BWs in 25 Pflanzen untersucht. Von Interesse sind bei den sessilen Pflanzen die hohen Herausforderungen an Pathogen- und Stressresistenz. Fungieren manche pflanzliche Protein-BWs als Teil von Resistenzgenen? Wie unterscheiden sich die Protein-BWs der Metazoa und der Pflanzen in Bezug auf Funktion und Evolution? In Chapter 4 erörtere ich diese Frage, und stelle die Ergebnisse der Studie zu Protein-BW-Evolution bei Pflanzen vor.