VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA CHEMICKÁ
FACULTY OF CHEMISTRY

ÚSTAV CHEMIE POTRAVIN A BIOTECHNOLOGIÍ
INSTITUTE OF FOOD SCIENCE AND BIOTECHNOLOGY

ANALÝZA LOKALIZACE INVERZNÍCH REPETIC V BAKTERIÁLNÍCH GENOMECH
ANALYSES OF INVERTED REPEATS LOCALIZATION IN BACTERIAL GENOMES

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. Michal Šedý**
AUTHOR

VEDOUCÍ PRÁCE — **doc. Mgr. Václav Brázda, Ph.D.**
SUPERVISOR

BRNO 2021

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA CHEMICKÁ

Zadání diplomové práce

Číslo práce:	FCH-DIP1449/2019	Akademický rok:	2020/21
Ústav:	Ústav chemie potravin a biotechnologií		
Student:	**Bc. Michal Šedý**		
Studijní program:	Chemie a technologie potravin		
Studijní obor:	Potravinářská chemie a biotechnologie		
Vedoucí práce:	**doc. Mgr. Václav Brázda, Ph.D.**		

Název diplomové práce:

Analýza lokalizace inverzních repetic v bakteriálních genomech

Zadání diplomové práce:

Literární rešerše se zaměřením na inverzní repetice, využití nástroje Palindrome finder pro charakterizaci přítomnosti a lokalizaci inverzních repetic v bakteriálních genomech se zaměřením na genomy organismů důležitých v potravinářství. Analýza a zpracování dat.

Termín odevzdání diplomové práce: 30.7.2021:

Diplomová práce se odevzdává v děkanem stanoveném počtu exemplářů na sekretariát ústavu. Toto zadání je součástí diplomové práce.

Bc. Michal Šedý student(ka)	doc. Mgr. Václav Brázda, Ph.D. vedoucí práce	prof. RNDr. Ivana Márová, CSc. vedoucí ústavu

V Brně dne 1.2.2021

prof. Ing. Martin Weiter, Ph.D.
děkan

ABSTRAKT

Inverzní repetice (IR) jsou přirozenou součástí DNA všech známých prokaryotických i eukaryotických organismů. Inverzní repetice mají důležitou roli při regulaci základních buněčných procesů. Jsou zodpovědné za vznik křížových struktur. Inverzní repetice taktéž zapříčiňují genomovou nestabilitu a mohou být zdrojem řady mutací. Křížové struktury mohou být rozeznávány celou řadou DNA vazebných proteinů a také mohou fungovat jako transkripční regulátory. Pomocí nástroje *Palindrom analyser* byla analyzována frekvence výskytu a lokalizace inverzních repetic v bakteriálních genomech. Frekvence výskytu inverzních repetic napříč bakteriálními genomy vykazuje variabilní charakter. Frekvence výskytu krátkých inverzní repetic vykazuje přibližně kvadratickou závislost na % obsahu GC párů v genomu s minimem okolo 50 % obsahu GC. Lokalizace inverzních repetic vzhledem k funkčním oblastem DNA vykazuje nenáhodný charakter rozložení. Frekvence výskytu IR u většiny sledovaných funkčních oblastí je vyšší „vně“ než „uvnitř“.

ABSTRACT

Inverted repeats (IR) are common part of DNA of all living prokaryotic and eukaryotic organisms. Inverted repeats plays an important role in the regulation of basics cells processes. They are responsible for formation of cruciform structures. Inverted repeats also cause genomic instability and can be a source of numerous mutations. Cruciform structures can be recognized by DNA-binding proteins and can also act as a transcriptional regulators. Using the *Palindrome Analyser* tool, the frequency of IR and localization of inverted repeats in bacterial genomes was analyzed. The frequency of IR across the bacterial genome is variable. The frequency of short inverted repeats shows an approximately quadratic dependence on the %GC content in the genome with a minimum of about 50% of GC content. The localization of inverted repeats with respect to “annotated features” show a non-random distribution. The frequency of IR for most features is higher “outside” than “inside”.

KLÍČOVÁ SLOVA

Inverzní repetice, křížové struktury, *Palindrome analyser,* protein p53

KEY WORDS

Inverted repeats, cruciform structures, *Palindrome analyser*, protein p53

ŠEDÝ, Michal. *Analýza lokalizace inverzních repetic v bakteriálních genomech.* Brno, 2020. Dostupné také z: https://www.vutbr.cz/studenti/zav-prace/detail/124203. Diplomová práce. Vysoké učení technické v Brně, Fakulta chemická, Ústav chemie potravin a biotechnologií. Vedoucí bakalářské práce doc. Mgr. Václav Brázda, PhD.

PROHLÁŠENÍ

Prohlašuji, že jsem diplomovou práci vypracoval samostatně a že všechny použité literární zdroje jsem správně a úplně citoval. Bakalářská práce je z hlediska obsahu majetkem Fakulty chemické VUT v Brně a může být využita ke komerčním účelům jen se souhlasem vedoucího bakalářské práce a děkana FCH VUT.

…….

podpis studenta

PODĚKOVÁNÍ

Děkuji panu doc. Mgr. Václavovi Brázdovi, Ph.D. za odborné vedení mé diplomové práce. Dále děkuji Ing. Otílii Porubiakové za čas, který mi věnovala při zpracování této práce.

OBSAH

1 ÚVOD

Inverzní repetice jsou přirozenou součástí DNA prokaryotických a eukaryotických buněk a vyskytují se napříč celým genetickým kódem. Vyskytují se v různých oblastech DNA s různou četností. Jsou rozeznávány celou řadou proteinů jako jsou restrikční enzymy, helikázy a transkripční faktory [1]. Inverzní repetice jsou schopné ovlivňovat strukturu DNA. Také mohou způsobovat genetickou nestabilitu. Například místa translokačních zlomů v genomu lidského karcinomu bývají často obohacena o inverzní repetice [2]. Taktéž mohou být mutagenní a stimulovat tvorbu DNA dvou-řetězcových zlomů, které mohou vést k delecím v savčích i kvasinkových buňkách [2]. Nejenom v lidském genomu, ale také například v genomu viru SARS-COV-2 jsou místa výskytu bodových mutací obohaceny o inverzní repetice [3].

Kromě známé dvoušroubovicové struktury DNA objevené Watson a Crickem v roce 1953 [4] byly postupně objeveny další formy sekundárních struktur DNA, které se účastní celé řady biologických procesů jako například levotočivá Z-DNA, vlásenkové struktury, křížové struktury, triplexy, kvadruplexy a další [5]. Právě inverzní repetice se podílejí na vzniku křížových a vlásenkových struktur. K tvorbě křížové struktury je zapotřebí minimálně 6 nebo více bází dlouhé inverzní repetice [6]. Jejich struktura je stabilizována nadšroubovicovým vinutím DNA [7].

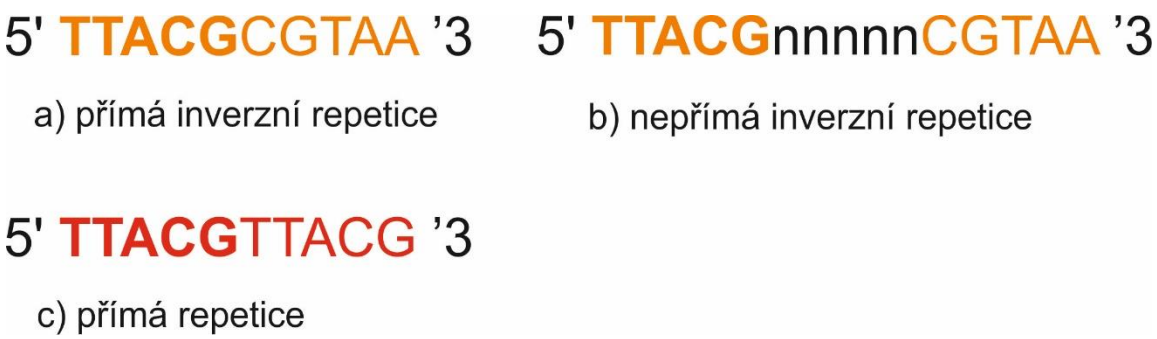
Křížové struktury hrají důležitou roli v regulaci přirozených procesů zahrnujících DNA. Nejsou rozmístěny zcela náhodně. Častěji se vyskytují v blízkosti bodů zlomu, oblasti promotorů a místech počátků replikace DNA [7]. Také mohou ovlivnit stupeň nadšroubovicového vinutí, umístění nukleosomů *in vivo*, a tvorbu dalších sekundárních struktur. Jsou zásadní pro celou řadu biologických procesů zahrnujících replikaci, regulaci genové exprese, strukturu nukleosomů a rekombinaci. Také byly zapojeny do evoluce a vývoje některých onemocnění jako je rakovina nebo Wernerův syndrom (syndrom předčasného stárnutí) a další [8; 9].

Křížové struktury jsou cílem celé řady stavebních a regulačních proteinů. Byla identifikována celá řada proteinů mající afinitu ke křížovým strukturám včetně proteinu 14-3-3, tumor suppresového proteinu p53 s jeho homologu p73 [1; 7; 10]. Také histony H1 a H5, topoisomeráza IIβ, HMG proteiny, HU, p53, proto-onkogenní protein DEK a další interagují s křížovými strukturami [7]. Několik DNA-vazebních proteinů, jako jsou proteiny z rodiny HMGB, Rad54, BRCA1 a PARP-1 polymeráza se i přes svoji nízkou DNA sekvenční specifitu váží přednostně na křížové struktury. Některé z těchto proteinů jsou dokonce schopné vyvolat tvorbu křížových struktury po navázaní na DNA [7].

2 TEORETICKÁ ČÁST

2.1 Inverzní repetice

Inverzní repetice (IR) je úsek DNA, který sestává ze dvou následných navzájem invertovaných komplementárních sekvencí na jednom vláknu DNA. Mezi těmito dvěma sekvencemi se může nacházet různý počet nukleotidových bází včetně nuly. Rozlišujeme přímé a nepřímé inverzní repetice. V případě přímých inverzních repetic také označovaných jako palindromické se mezi invertovanými sekvencemi nenacházejí další nukleotidy (obrázek 1). Nepřímé inverzní repetice obsahují mezi invertovanými sekvencemi variabilní počet dalších nukleotidů (obrázek 1) [11]. Inverzní repetice ještě dále můžeme dělit na dokonalé a nedokonalé na základě úplné nebo částečné komplementárnosti invertovaných sekvencí. Kromě inverzních repetic rozeznáváme také přímé repetice, které sestávají z opakujících se sekvencí nukleotidů (obrázek 1) [7; 12].



Obrázek 1 *Příklady inverzních sekvencí*

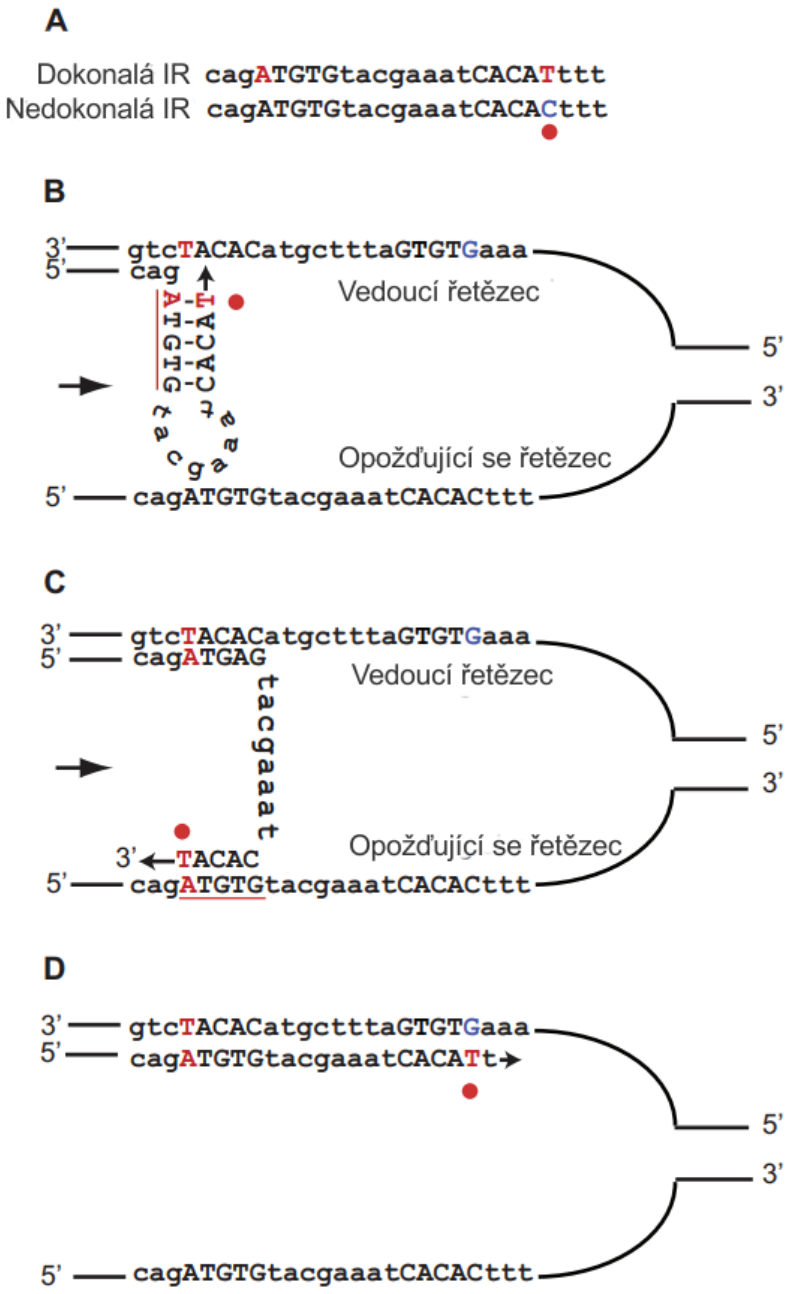
Inverzní repetice plní celou řadu biologicky důležitých funkcí. Hrají důležitou roli při genové expresi, regulaci a replikaci DNA [11]. Inverzní repetice také stimulují vznik delecí během DNA replikace a interchromozomální rekombinaci mezi homologními sekvencemi [13; 14]. Také definují hranice transpozomů a označují oblasti schopné autokomplementárního párování bází [13]. Tyto vlastnosti hrají důležitou roli v případě genové nestability a přispívají nejen ke genetické diverzitě a evoluci, ale také dávají vzniku mutacím a onemocněním. Díky komplementárnosti sekvencí mohou za vhodných podmínek dát vzniku dalším sekundárním strukturám, především křížovým strukturám nebo vlásenkám.

Rozmístění inverzních repetic napříč genetickým kódem vykazuje nenahodilý charakter. Analýza lidského a myšího genomu odhalila, že výskyt inverzních repetic je častější u intronů než u exonů [1]. Dále bylo analýzou promotorů lidského genomu zjištěno, že nejvíce inverzních

repetic s délkou 8+ se nachází ve vzdálenost do 100 bp před počátečním místem transkripce (TSS, z angl.: transcription start site) [1].

2.2 Vliv inverzních repetic na strukturu DNA

Jak již bylo uvedeno výše mohou inverzní repetice způsobovat mutace DNA, tedy měnit její primární strukturu. Nedokonalé inverzní repetice mohou způsobovat bodové mutace. Na obrázku 2 je možné vidět schéma modelu vzniku bodové mutace způsobené nedokonalou IR, poprvé navrženého v roce 1982. Proces mutace zahrnuje dva přeskoky DNA polymerázy. První

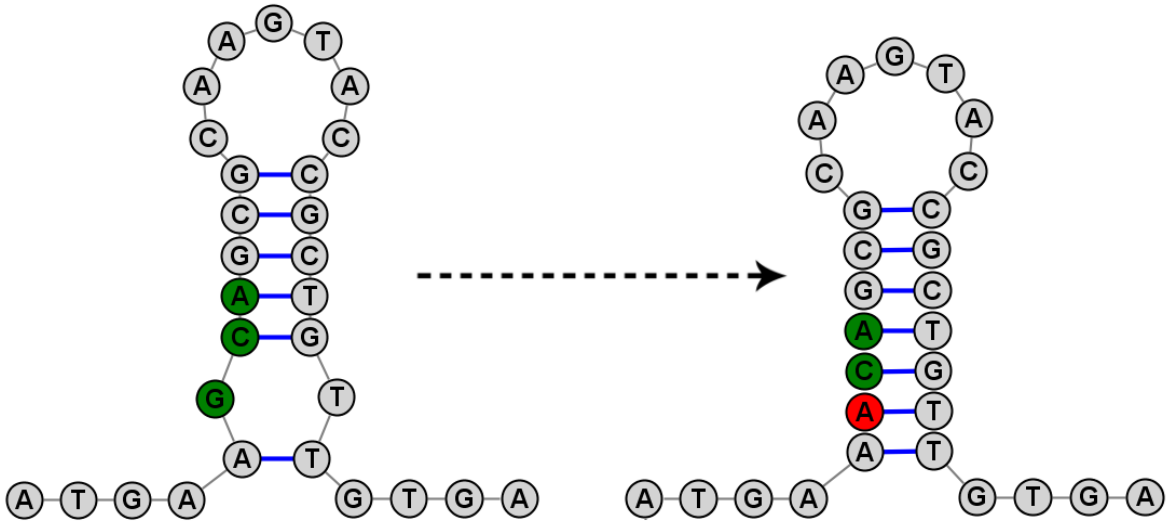


Obrázek 2 *Bodová mutace DNA (převzato a upraveno z [15])*

přeskok, který může být buď intramolekulární, kdy DNA polymeráza přeskakuje z původního

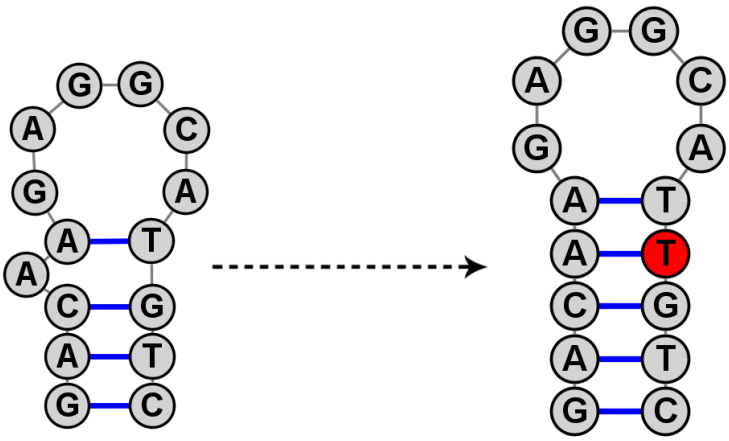
řetězce na nově syntetizovaný (obr. 2 B) nebo intermolekulární, kdy dochází k přeskoku DNA polymerázy na protilehlý řetězec, v tomto případě opožďující se (obr. 2 C). Poté dochází ke druhému přeskoku DNA polymerázy zpátky na replikované vlákno (obr. 2 D). Na konci tohoto procesu se na původním vláknu nachází původní nedokonalá IR a na novém vláknu nová dokonalá IR, která se během další replikace zanese kompletně do genomu dceřiné buňky [15].

Bodové mutace způsobené nedokonalými IR mohou vést k řadě onemocnění. Například oblast genu kódující antitrombin III obsahuje nedokonalou IR (obr. 3), přičemž mutace na dokonalou IR způsobí záměnu guaninu za adenin a konverzi kodonu ACG na ACA (obr. 3), což vede k zařazení aminokys. threoninu namísto alaninu. Následkem je vznik nefunkčního



Obrázek 3 *Bodová mutace genu kódující protein Antithrombin III (kresleno pomocí [82])*

anti-thrombinu III [16].



Obrázek 4 *T-inzerce (kresleno pomocí [82])*

Další onemocnění jako osteogenesis imperfecta, jehož základním projevem je křehkost kostí, které vede ke zlomeninám dlouhých kostí a vzniku dalších deformit je spojováno

s mutacemi genu kolagenu typu I, COL1A1 a COL1A2. Jedná se o delece, inzerce a bodové mutace především v oblasti kódující triple-helikální strukturu kolagenu. Příkladem může být T-inzerce (obr. 4) způsobené konverzí nedokonalé IR na dokonalou IR, která zapříčiní změnu čtecího rámce a následně vznik mutantního propeptidického řetězce proα1(I), který je zkrácený a má jiné složení aminokyselin. Mutantní řetězec proα1(I) znemožňuje tvorbu normální triple-helikální struktury kolagenu [16; 17].

2.3 Vznik a výskyt křížových struktur v genomu

Kromě mutací vedou inverzní repetice k tvorbě sekundárních struktur jako jsou vlásenky nebo křížové struktury. Křížové struktury hrají důležitou roli při regulaci biologických procesů. V normální dvoušroubicovité DNA je jejich výskyt termodynamicky nepravděpodobný. Ovšem s rostoucím stupněm negativního nadšroubovicového vinutí DNA pravděpodobnost jejich výskytu roste z důvodu snižující se energie potřebné pro jejich tvorbu. Tvorba křížových struktur *in vivo* byla prokázána jak u prokaryotických, tak u eukaryotických organismů. Přítomnost křížových struktur byla poprvé popsána u kružnicové plasmidové DNA, kde může být stabilizována negativním nadšroubovicovým vinutím. Například, existence křížové struktury *in vivo* byla prokázána u plazmidu pT181. Delecí sekvence zodpovědné za vznik této struktury vedlo k redukci nebo selhání replikace [7]. Podobně delecí vazebné domény pro křížovou strukturu u proteinů 14-3-3 došlo ke snížení vazby na počátky replikace, což ovlivňuje iniciaci replikace DNA u kvasinek. Ke studiu křížových struktur a jejich izolaci byly také připraveny monoklonální protilátky, které se na tyto struktury specificky váží a tím je stabilizují [18]. Použití monoklonálních protilátek 2D3 a 4B4 se specifitou ke křížovým strukturám vedlo ke dvou až šestinásobnému zvýšení replikace *in vivo* [19]. Podobně u homologních proteinů 14-3-3 kvasinek Saccharomices cerevisiae Bmh1p a Bmh2p, byla pomocí monoklonálních

```
  1 TTCGAAGAAA TGCCAGTGAT GCGGACATCG TTAATATAAA GATTTTACGA AGGAATTCTA
 61 GGTAATGTTG CAATTACTTC TTCTCATGCA CTAACAAGTG AATGATAGAA ATATGTTGAG
121 TTCCTAACTG CCTGATTTTA AATAAGTTTC ATATTATAAT CTTTTAGCAT ATATATATAT
181 ATATTGATCC TCTCTCTTCT TTATTTTCTG CCAGTAACCC ATGTGTGAAG AAGAAAACAT
241 AAATAAAAAA GCAGTAGCAC ATGGACACAT TCACGCCCGA ACACTCCTAA AAAGCAGCCC
301 ACACAAGAAA GTAGATATAG TGTAGGACAC CCAG
```

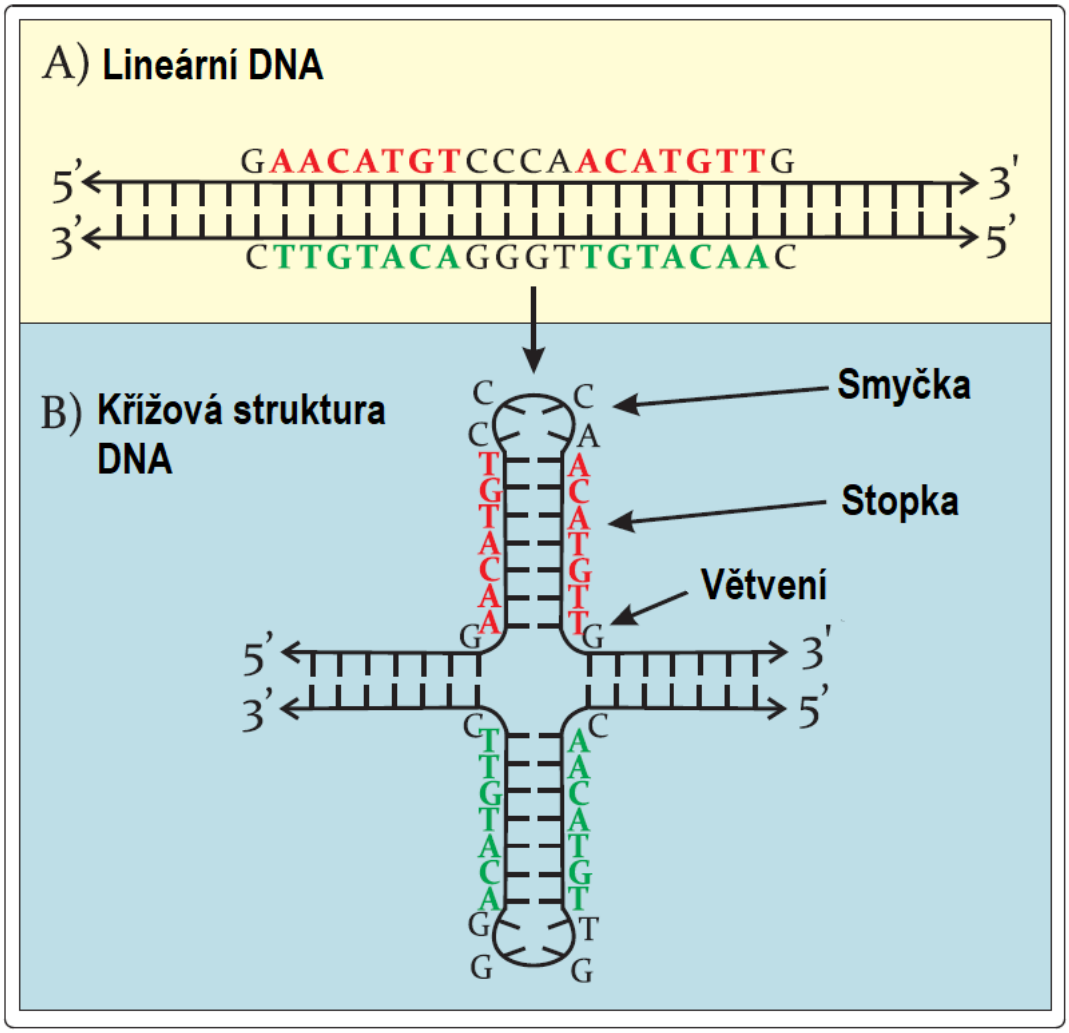
Obrázek 5 *Sekvence ARS307*

protilátek 2D3 prokázána afinita ke křížovým strukturám *in vitro* a *in vivo*. Tyto proteiny se váží v dimerní formě specificky ke křížovým strukturám, přičemž heterodimer Bmh1p-Bmh2p vykazuje silnější afinitu než homodimery výše zmíněných proteinů. Jejich cílová struktura se

nachází na replikačním počátku ARS307, kde je tvořená dokonalou inverzní repeticí o délce 10 bp (obr 5) [20].

Křížové struktury vznikají na základě komplementárnosti bází podobně jako dvoušroubovicová DNA, s tím rozdílem, že k párování bází dochází mezi bázemi jednoho a téhož vlákna DNA. Tímto procesem vzniká útvar, který má po obou stranách výstupky připomínající kříž (obr. 6). Na obr. 6 je možné vidět sekvenci 20 bp promotoru genu p21 obsahující inverzní repetici, která je rozeznávána proteinem p53 a přechod na její křížovou strukturu [7].

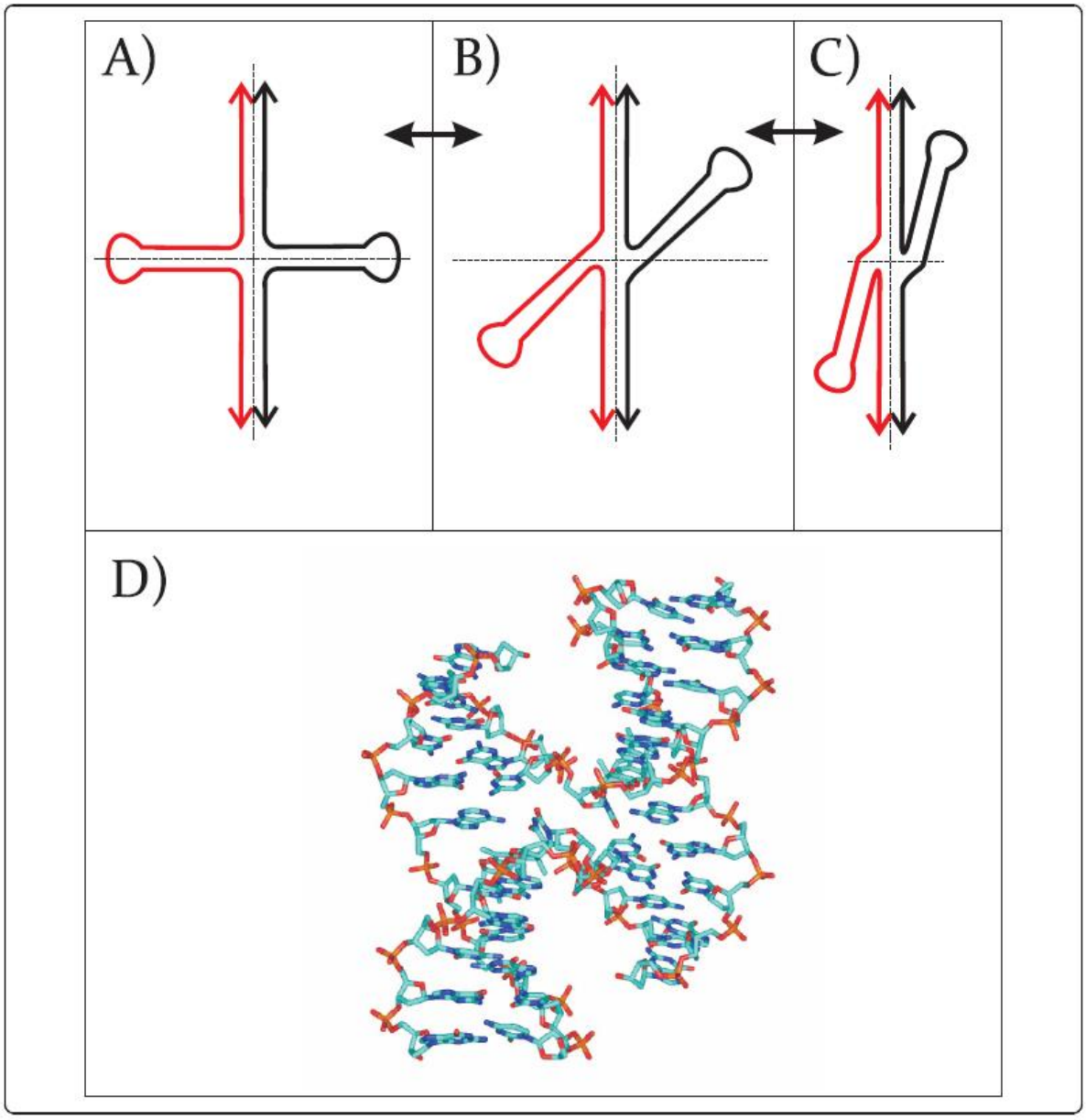
Křížová struktura sestává s větvení, stopky a smyčky (obr. 6 B). Délka smyčky závisí na velikosti mezery mezi invertovanými sekvencemi. Přímé inverzní repetice vedou ke tvorbě smyčky s minimální délkou, která neobsahuje žádnou nukletidovou bázy. Formace křížových struktur z nepřímých inverzních repetic závisí nejen na délce mezery, ale také na její sekvenci. Sekvence mezery bohatá na AT báze zvyšuje pravděpodobnost vzniku křížové struktury.



Obrázek 6 *Přechod lineární DNA na křížovou strukturu DNA (převzato a upraveno z [7])*

Pravděpodobnost formace křížových struktur také podporuje negativní nadšroubovicové vinutí DNA [7; 21].

Kromě křížových struktur vznikajících na jedné dvouvláknové šroubovici DNA, vznikají také křížové struktury při crossing-overu mezi dvěma homologními chromozomy. Při tomto ději dochází k propojení čtyř vláken DNA a vzniká tzv. Hollidayovo spojení, které zaujímá tvar křížové struktury. Toto spojení je následně rozeznáváno a štěpeno speciálními enzymy zvanými resolvázy. Studium křížových struktur pomocí AFM odhalilo několik možných konformací (obr. 7 D). První konformace označovaná jako rozložená má obě ramena kolmá na hlavní řetězec DNA (obr. 7 A), zbylé dvě konformace označované jako skládané svírají ostrý úhel



Obrázek 7 *Konformace křížových struktur (převzato a upraveno z [7])*

s hlavním řetězcem DNA (obr 7 B, C). Právě dvě z těchto konformací můžeme nalézt v Hollidayově struktuře [7].

2.4 Interakce proteinů s křížovými strukturami

Interakce proteinů s molekuly DNA patří mezi základní biologické procesy, bez kterých by nebyl možný život buňky. Hrají důležitou roli při regulaci genové exprese [22], DNA replikaci [23], opravách [24], transkripci [25], rekombinaci [26] a organizaci DNA do chromatinu [27]. Podle způsobu vazby proteinů na nukleové kyseliny rozlišujeme vazbu proteinů na nespecifickou a specifickou mající sekvenční nebo strukturní specifitu. Celá řada proteinů se váže k DNA bez sekvenční či strukturní specifity, například histony, které způsobují organizaci DNA do chromatinu. Velký význam mají i proteiny vážící se specificky k určitým oblastem DNA, například transkripční faktory aj. Velmi často obsahují specifické proteinové motivy, jako zinkový prst, leucinový zip, dva helixy spojené krátkou smyčkou aj. Další významná skupina proteinů jsou proteiny, které rozlišují specificky určité lokální struktury DNA a jsou schopné s nimi interagovat. Mezi takové struktury patří právě výše zmíněné křížové struktury.

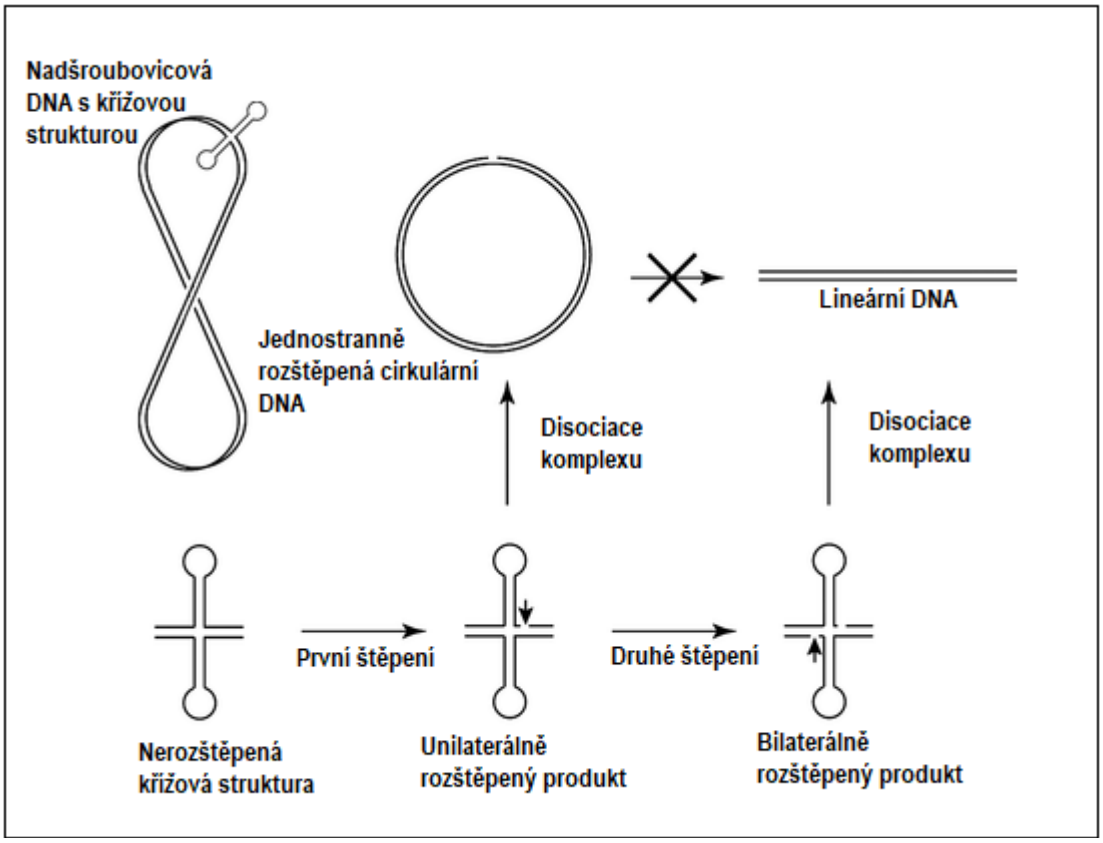
Z hlediska hlavních funkcí proteinů, které se váží na křížové struktury, byly proteiny rozděleny na: a) enzymy rozeznávající Hollidayovo spojení, b) proteiny účastnící se transkripce a opravy DNA, c) chromatin-asociované proteiny, d) proteiny účastnící se replikace a MLL protein [7].

2.4.1 Enzymy rozeznávací Hollidayovo spojení

Enzymy rozeznávají Hollidayovo spojení patří mezi hojně zastoupené endonukleázy, které jsou důležité při opravách DNA a rekombinaci. Jak již bylo zmíněno výše, Hollidayovo spojení je struktura, která sestává s propojení čtyř vláken DNA, vznikající překřížením řetězců DNA při rekombinaci nebo při obrácení replikační vidlice. Proteinů rozeznávající Hollidayovo spojení byla identifikována celá řada, byly identifikovány u mnoha organismů, od bakterií až po kvasinky, archae a savčí buňky [28]. Většina z nich může být rozdělena do dvou základních skupin [29]. Enzymy v první skupině se váží ke všem křížovým strukturám, ale štěpí pouze specifické sekvence. Do této skupiny patří *E. coli* RuvC, kvasinkové integrázy, Cce1, Ydc2 a RnaseH proteiny. Druhá skupiny zahrnuje endonukleázy T7, RecU, resolvázy Hjc a Hje, proteinovou rodinu MutH a příbuzné restrikční endonukleázy [30].

Na obrázku 8 je možné vidět schéma mechanismu bilaterálního (dvoustranného) štěpení křížové struktury lokalizované na cirkulární DNA, která je stabilizována negativním

nadšroubovicovým vinutím. Pokud by docházelo pouze k unilaterálnímu (jednostrannému) štěpení, nevznikala by lineární DNA, ale cirkulární DNA zbavená superhelicity a s reabsorbovanou křížovou strukturou, a tedy s nemožností znovunavázání enzymu. Je tedy důležité, aby proběhlo druhé štěpení před disociací komplexu proteinu a křížové struktury. Toho je dosaženou zrychlením druhé štěpné reakce deseti až stonásobně [28].



Obrázek 8 *Štěpení křížové struktury resolvázou (převzato a upraveno z [28])*

2.4.2 Proteiny účastnící se transkripce a opravy DNA

Oprava DNA patří pravděpodobně mezi nejdůležitější mechanismy k udržování genetické stability. Vazba proteinů na poškozenou DNA a na lokální alternativní DNA struktury tedy hraje klíčovou roli při procesu opravy DNA. Právě tyto alternativní struktury můžeme často najít v promotorech genů, kde se mohou formovat z inverzních repetic, které se zde vyskytují častěji a jsou schopny tvorby křížových struktur *in vivo*. Řada DNA-vazebných proteinů, jako například proteiny rodiny HMGB-box [31], Rad54 [32], BRCA1 [33], MutS [34] a také PARP-1 [24], vykazují nízkou sekvenční, ale vysokou strukturní specifitu ke křížovým strukturám. Mezi proteiny účastnící se opravy DNA, které rozeznávají křížové struktury, patří již výše zmíněné enzymy Ruv, RuvB, helikáza DNA, XPG protein a multifunkční proteiny z rodiny

HMG-box BRCA1, rodina proteinů 14-3-3 a kvasinkové homology Bmh1 a Bmh2, a také rostlinný homolog GF14.

PARP-1 - je nejhojněji se vyskytující enzym z rodiny proteinů PARP. Proteiny z rodiny PARP jsou DNA-závislé polymerázy, vyskytující se hojně v jádrech buněk (cca 1 enzym na 50 nuklezómů), které obsahují strukturu zinkových prstů. Jejich substrátem je molekula NAD+, ze které odštěpují monomery ADP-ribózy, které polymeryzují na poly(ADP-ribózu), která dále slouží jako signální řetězec pro další enzymy podílející se na opravě DNA [35]. PARP-1 má vysokou afinitu k poškozené DNA, taktéž umožňuje rozrušení histonové struktury a tím ji zpřístupňuje dalším regulačním faktorům [36]. Kromě své afinity k poškozené DNA bylo také zjištěno, že se může vázat a stabilizovat křížové struktury v DNA [24]. Pokusy s negativní nadšroubovicovou plasmidovou DNA obsahující křížové struktury ukázaly, že PARP-1 je schopen vazby a stabilizace těchto struktur, což vede ke změně úrovně superhelicit. Modulace úrovně superhelicity byla navržena jako jeden z možných mechanismů regulace genové exprese [24].

Protein p53 – patří mezi jeden z nejvíce studovaných tumor supresorových proteinů. Více než 50 % všech lidských nádorů obsahuje mutace právě tohoto proteinu [37]. Inaktivace genu *TP53* kódující protein p53 hrají kritickou roli při indukci maligní transformace buňky [38]. Protein p53 se váže jako tetramer a jeho DNA-vazebná afinita je evolučně konzervovaná [15]. Protein p53 vykazuje sekvenční specifitu a jeho cílové místo sestává ze dvou kopií sekvence 5'-RRRC(A/T)(T/A)GYYY-3 [15], která často tvoří křížovou strukturu [39]. Taktéž bylo pozorováno, že jeho vazební afinita je závislá na teplotě a délce fragmentu DNA [40; 41]. Dále bylo prokázáno, *in vivo,* že vazba proteinu p53 na jeho cílovou sekvenci je vysoce závislá na přítomnosti právě inverzních repetic v cílovém místě vazby [39]. Inverzní repetice obsažené v cílovém místě zesilují vazebnou afinitu proteinu p53 k DNA [42]. Také byla popsána jeho vyšší afinita k nadšroubicové DNA [43; 44]. Protein p53 se může selektivně vázat na další nekanonické struktury jako duplexy obsahující chybně párované báze, ohnutá DNA, křížové struktury, hemikatenátové smyčky, DNA výdutě, 3 a 4 - cestné křížení nebo telomerní t-smyčky [7].

2.4.3 Chromatin-asociované proteiny

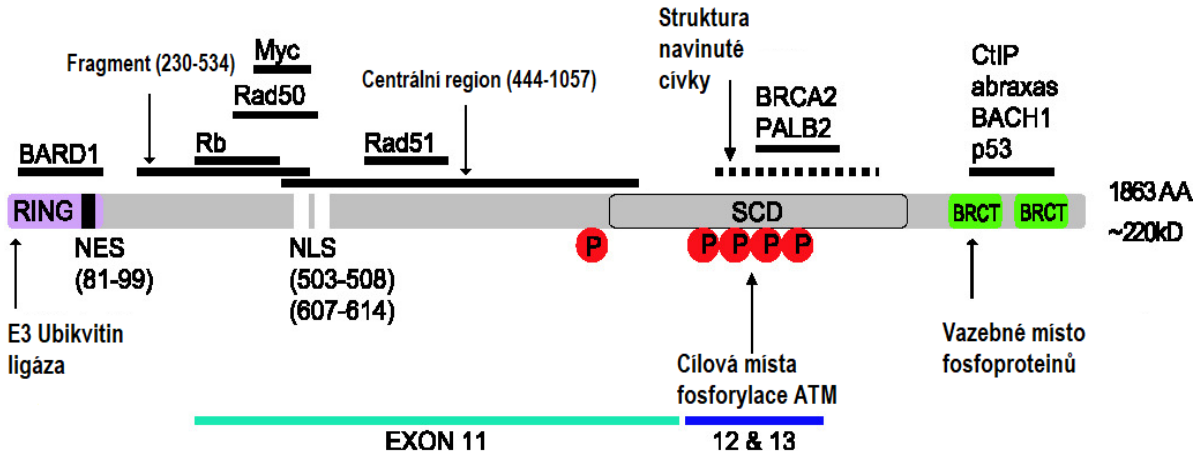
Tvoří širokou skupinou jaderných proteinů. Tyto proteiny, které se částečně podílejí na tvorbě chromatinové struktury a také jsou zapojeny do celé řady procesů souvisejících s funkcí DNA. Do této skupiny patří i proteiny DEK, BRCA1, HMG proteiny, Rad51, Rad54, které se

účastní opravy a replikace DNA. Další skupinou enzymů důležitou v procesech opravy a replikace DNA jsou topoizomerázy. Tyto enzymy se vyskytují ve všech živých organismech, kde mají za úkol řídit topologický stav buněčné DNA [45]. Topoizomeráza I štěpí hollidayova spojení [46] a topoizomeráza II rozeznává a štěpí křížové struktury [47] a taktéž interaguje s HMGB1 proteiny [48]. Tyto procesy jsou důležité pro zachování genomové stability, a to především díky schopnosti snížit torzní napětí, které vzniká v DNA během transkripce a replikace a taktéž schopnosti štěpit dlouhé křížové struktury, které by jinak bránily oddělování vláken DNA. Další protein Rad54 hraje důležitou roli při homologní rekombinaci eukaryot. Kvasinkový a lidský Rad54 se specificky váže na Hollidayova spojení s preferencí k jeho rozložené konformaci, a taktéž podporuje migraci této struktury [32; 49].

DEK protein – je všudypřítomný a hojně se vyskytující protein v mnohobuněčných organismech, kde je obsažen v množství několik miliónů kopií na buněčné jádro. Sestává z 375 amino-kyselin (AA) a má dvě funkční DNA-vazebné domény, kde jedna doména leží v centrální části molekuly a obsahuje sekvence, které jsou shodné s evolučně konzervovanou SAF-box doménou [50]. Druhá DNA-vazebná doména se nachází na C-terminálním konci, který může být posttranslačně modifikován fosforylací, která vede k jeho snížení vazebné afinity k DNA a indukci tvorby DEK multimerů [51]. Protein DEK postrádá sekvenční specifitu, ale vykazuje strukturní specifitu [52]. Studium izolovaného DEK proteinu prokázaly jeho DNA-vazební aktivitu s preferencí ke čtyřcestným spojením a nadšroubovicové DNA proti lineární DNA a jeho schopnost vyvolat pozitvní nadšroubovicové vinutí v relaxované cirkulární DNA [27].

BRCA1 protein – patří mezi lidské tumor supresorové proteiny. Gen BRCA1 byl poprvé identifikován a naklonován v roce 1994 na základě jeho spojitosti s brzkým nástupem rakoviny prsu a vaječníků u žen. BRCA1 protein účinkuje v celé řadě buněčných procesů zahrnující udržování genomové stability, včetně aktivace kontrolního bodu buněčného cyklu vyvolaného poškozením DNA, opravu DNA, ubikvitinaci proteinů, chromatinovou remodelaci, a také transkripční regulaci a apoptózu [7]. Mutace genu BRCA1 jsou spojené se značně zvýšeným rizikem vzniku rakoviny prsu a jsou zodpovědné až za 45 % případů dědičné rakoviny prsu [53].

Protein BRCA1 se skládá z 1 863 aminokyselin a obsahuje několik vazebných domén a řadu fosforylačních míst (obr. 9). Na počátku obsahuje strukturu zinkového prstu (RING), která je zodpovědná za jeho ubikvitin ligázovou aktivitu. Na tuto strukturu se váže protein BARD1, který je důležitý pro jeho aktivitu, a se kterým vytváří heterodimer. Na exonu 11-13 se nacházejí dvě sekvence NLS a vazebná místa pro několik proteinů jako RB, cMyc, Rad50 a RAD51. Dále se zde nachází struktura navinuté cívky, která se účastní interakce s proteinem PALB2 a také část SCD domény, která je cílem fosforylace ATM serin/threonín kinázy. Na konci proteinu se nachází BRCT doména, která zprostředkovává fosfoproteinové interakce mezi BRCA1 proteiny a proteiny fosforylované kinázy ATM a ATR, při zjištění poškození DNA. Také se zde váží proteiny CtIP, abraxas, BACH1 a p53 [54]. Kromě své vazebné afinity k ostatním proteinům vykazuje také afinitu k DNA. Fragment centrálního regionu (444-1 057) BRCA1 se váže silně na negativní nadšroubovicové plasmidové DNA a také vykazuje vysokou afinitu ke křížovým strukturám DNA [55]. Taktéž samotný 1 863 AA dlouhý protein BRCA1 se silně váže k plasmidové scDNA a také na konce DNA. Další fragment BRAC1 (230-534) (obr. 9) se váže silněji na křížové struktury ve srovnání s dvouřetězcovou a jednořetězcovou DNA. Ani 20-násobný přebytek koncentrace lineární DNA nebyl schopný rozrušit vazbu fragmentu



Obrázek 9 *Struktura proteinu BRCA1 (převzato a upraveno z [54])*

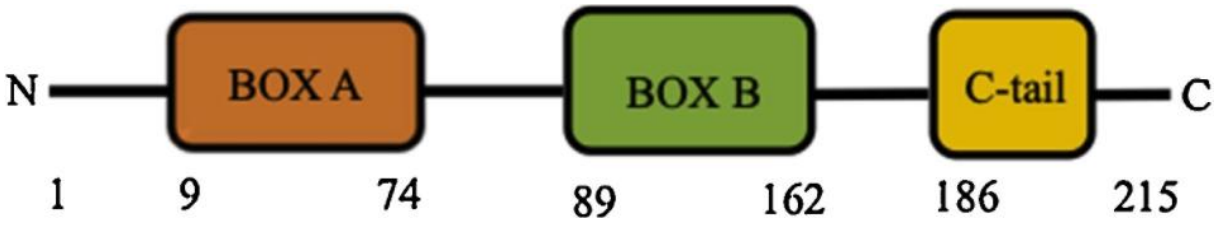
BRCA1 (230-534) vázaného ke křížové struktuře DNA. Oblast aminokyselin 340-534 BRCA1 byla identifikována jako oblast s minimální afinitou k DNA [56]. Funkce proteinu BRCA1 je velmi komplexní a zahrnuje pravděpodobně interakce jak s DNA a jejími sekundárními struktury, tak s celou řadou proteinů.

HMGB rodina – proteiny z rodiny HMG (z angl.: High-Mobility Group) jsou vysoce pohyblivé chromozomální proteiny obsahující HMG-Box doménu. Jsou hojně se vyskytující,

všudypřítomné nehistonové proteiny, které se váží k chromatinu eukaryot. Proteiny HMG jsou rozděleny do 3 rodin, každá obsahující charakteristikou funkční doménu:

a) HMGA – obsahující AT-hook doménu,
b) HMGB – obsahující HMG-box doménu,
c) HMGN – obsahující nukleozóm-vazebnou doménu.

HMGB proteiny nevykazují sekvenční specifitu, ale vykazují strukturní specifitu s vyšší afinitou k některým strukturám DNA (čtyřcestné spojení, DNA minikroužky, cis-platinová DNA aj.) oproti lineární DNA. Proteiny HMG jsou zapojeny do regulace DNA-závislých procesů jako je transkripce, replikace, rekombinace a oprava DNA. HMG-box doména je přibližně 80 aminokyselin dlouhá doména, která se nachází v řadě chromozomálních proteinů a transkripčních faktorů eukaryot. Všechny HMG-box domény vykazují preferenci ke čtyřcestným spojením DNA v rozložené konformaci. Naproti tomu za podmínek, které stabilizují skládanou konformaci čtyřcestného spojení dochází k oslabení afinity HMG-box domény k čtyřcestným spojením [57]. Na obrázku 10 je znázorněna struktura proteinu HMGB1, který se může vázat s vysokou afinitou na DNA smyčky. Struktura sestává ze dvou tandemových HMG-box domén – Box A N-terminální domény, Box B centrální domény a kyselé C-terminální oblasti (obr. 10). Domény A, B jsou zodpovědné za DNA-vazebnou afinitu HMGB1 proteinu, C-terminální oblast hraje roli při regulaci a schopnosti ohybu DNA HMGB proteiny [58]. Bylo zjištěno, že vazba izolované Box A domény ke čtyřcestným spojením DNA je znemožněna, pokud dojde ke dvojité mutaci uvnitř této domény na aminokyselinách Lys2 a Lys11 [59].



Obrázek 10 *Struktura proteinu HMGB (převzato a upraveno z [58])*

2.4.4 Proteiny účastnící se replikace

Existuje řada proteinů účastnící se replikace vykazující afinitu ke křížovým strukturám. Bylo ukázáno, že přechodné změny B-DNA na křížové struktury korelují s replikací a transkripcí DNA [21]. Taktéž mohou křížové struktury sloužit jako rozpoznávací signály uvnitř nebo v blízkosti počátku replikaci DNA eukaryot [23].

Ribozomální protein S16 - kódovaný genem *RPS16* vykazuje strukturní specifitu s preferencí ke křížovým strukturám [60]. Některé proteiny ze skupiny SMC, jako například SMC1, SMC2 vykazují vysokou afinitu ke křížovým strukturám a také fragmentům DNA bohatých na AT sekvence včetně S/MAR regionů [61]. Proteiny ze skupiny SMC patří do velké rodiny vysoce konzervovaných chromozomálních ATPáz, které se podílejí na vyšší organizaci a dynamice chromozomů [62]. Další protein VLF-1, patřící mezi integrázy, se váže k sekundárním strukturám DNA jako jsou Y-vidlice, třícestné spojení a křížové struktury. VLF-1 se podílí na zpracování rozvětvených molekul DNA v pozdních stádiích replikace virového genomu AcMNPV a také se účastní sestavy nukleokapsidů bakuloviru AcMNPV [63].

14-3-3 rodina proteinů - sestává z devíti izoforem (α, β, γ, δ, ε, ζ, η, θ, τ z čehož α a δ jsou fosforylované formy β a γ), kódovaných sedmi různými geny, které se nachází v eukaryotických organismech. Rodina proteinů 14-3-3 byla poprvé identifikována v polovině 60. let jako rodina početných kyselých proteinů vyskytujících se v mozkové tkáni. Byly označeny na základě jejich elučního vzorce při DEAE chromatografii (14. frakce) a následné purifikace pomocí gelové elektroforézy (frakce 3.3) z toho tedy jejich název 14-3-3. Později po 20 letech jim byla připsána jejich první funkce jako aktivátory tyrozin a tryptofan hydroxyláz – enzymů účastnících se biosyntézy neurotransmiterů [64]. Následně byla objevena jejich schopnost vazby k celé řadě rozmanitých signálních proteinů, včetně kináz, fosfatáz a transmembránových receptorů. Tyto rozmanité vlastnosti propůjčují proteinům 14-3-3 schopnost modulovat celou řadu životně důležitých procesů zahrnujících mitogenní signály transdukce, apoptózy a regulaci buněčného cyklu [65].

Proteiny z rodiny 14-3-3 nacházíme především v buněčném jádru, kde se mohou účastnit procesů replikace DNA vazbou ke křížovým strukturám, které přechodně vznikají na replikačním počátku na začátku S fáze buněčného cyklu. Afinita některých izoforem 14-3-3 proteinů ke křížovým strukturám byla potvrzena pomocí imunofluorescenčních metod v HeLa buňkách [66]. Přímé interakce byly potvrzeny u izoforem β, γ, σ, ε, ζ a kvasinkových 14-3-3 analogů Bmh1, Bmh2 a GF14 u rostlin [20]. Taktéž bylo zjištěno, že řada z funkcí proteinů 14-3-3 souvisí s jejich afinitou ke křížovým strukturám [7].

WRN protein – patří mezi evolučně konzervované ATP dependentní 3’→5’ DNA helikázy z rodiny RecQ [67]. Je kódovaný genem WRN a jeho mutace způsobuje jedno z progerických onemocnění tzv. Wernerův syndrom (syndrom předčasného stárnutí). Rodina helikáz RecQ je obsažena jak u eukaryotických, tak u prokaryotických organismů, přičemž u vyšších eukaryot se vyskytuje několik homologů, v případě lidí bylo identifikováno 5 homologů. Všichni

členové sdílejí konzervovanou helikázovou doménu s jednou nebo dvěma C-terminálními doménami, RQC (RecQ karboxy-terminální) doména a HRDC (RNázováD C-terminální) doména. WRN se váže k replikační vidlici a také na Hollidayovo spojení. Tato vazba je vysoce specifická [9]. Při vazbě na DNA dochází k tvorbě velkého komplexu sestávajícího ze 4 monomerů WRN proteinu [9].

2.4.5 MLL protein

Je jaderný protein kódovaný genem *MLL1*, dnes označovaným jako KMT2A nacházející se na chromozomu 11q23 [68]. Chromozomové aberace zahrnující *MLL* gen jsou spojeny s akutní myeloidní i lymfoblastickou leukémií. Produktem genu *MLL1* je 3969 AA dlouhý protein, který patří do rodiny SET1 evolučně konzervovaných methytransferáz H3K4, které jsou zapojeny do regulace řady biologických funkcí. Protein MLL1 funguje jako transkripční koaktivátor, potřebný pro regulaci exprese Hox genů během hematopoézy a vývoje. Obsahuje celou řadu konzervovaných domén jako domény „AT-hook“, CXXC, PHD, BD, TAD, NR box, WDR5 a C-terminální SET doménu, která je zodpovědná za jeho methyltransferázovou aktivitu [69]. Po translaci je proteolyticky rozštěpen enzymem taspázou na dva fragmenty MLL-N a MLL-C. Tyto dva fragmenty tvoří společně s dalšími proteiny multiproteinový komplex, který reguluje modifikaci chromatinu a genovou expresi [69]. Doména „AT-hook“ proteinu MLL se váže ke křížovým strukturám DNA a vykazuje spíše strukturní než sekvenční specifitu a taktéž vykazuje preferenci k AT bohatým oblastem DNA [70].

2.5 Onemocnění a proteiny vážící se na křížové struktury

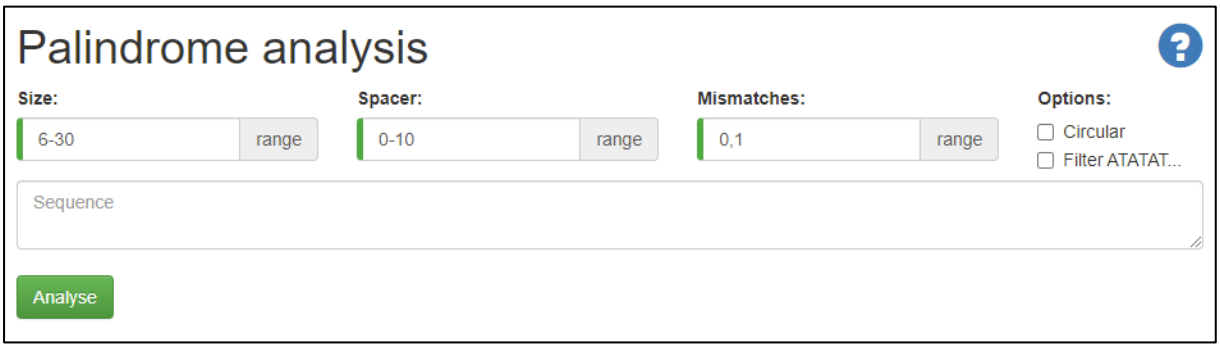
Rozpoznání křížových struktur je důležité pro udržování genomové stability a pro regulaci základních buněčných procesů. Dysregulace mechanismu štěpení Hollidayových spojů a dlouhých křížových struktur může vést k různým chromozomovým aberacím jako translokace, delece, ztráta stability nebo karcinogeneze. Existuje celá řada proteinů rozeznávající a vážící se na tyto struktury s cílem udržet neporušený genom. Mutace a epigenetické modifikace, které mění možnost vzniku křížových struktur mohou mít na buněčné úrovni drastické následky. Tedy dysregulace proteinů s kruciformní specifitou je často spojená s patologickými stavy [7].

Výše zmíněné proteiny jako p53, BRCA1, WRN a protoonkogenní DEK, MLL a HMG jsou spojovány s rozvojem rakoviny. Některý z těchto proteinu zaujímají tak důležitou roli, že jejich mutace nebo inaktivace způsobí vážnou nestabilitu genomu nebo i letalitu. Například kmenové

buňky myší s vyřazeným genem BRCA1 vykazují značnou nestabilitu genomu, přítomnost spontánních zlomů chromozomů a hypersenzitivitu k řadě škodlivých látek poškozující DNA (např.: γ záření), to vše z důvodu snížené schopnosti opravy DNA. Taktéž mutace genu BRCA1 u žen způsobuje zvýšenou pravděpodobnost vzniku rakoviny prsu a vaječníků. Také formace křížových struktur v oblastech promotorů rozeznávaných proteinem p53 může být důležitá pro jeho transkripční aktivitu. Chromozomové proteiny z rodiny HMG (proteiny s vysokou mobilitou) patří mezi transkripční regulátory a hrají důležitou roli při regulaci struktury chromatinu a genové expresi. Jejich nadprodukce je spojená s karcinogenezí, zvýšenou malignitou a potenciálem tumorů metastázovat *in vivo*. Proteiny z rodiny 14-3-3 jsou spojeny s řadou onemocnění jako je rakovina, Alzheimerova choroba, Miller-Diekerův syndrom, Spinocereberální ataxie typu 1 a spongiformní encefalopatie. Lidský protein DEK také může hrát roli v onemocnění způsobující demenci jako Alzheimerova choroba [71]. Také byl nalezen sfúzovaný s nukleoporinem NUP214 u části pacientů s akutní myeloidní leukémií [72]. Také byl identifikován jako autoantigen u celé řady pacientů s autoimunitním onemocněním [73].

2.6 Nástroj *Palindrome analyser*

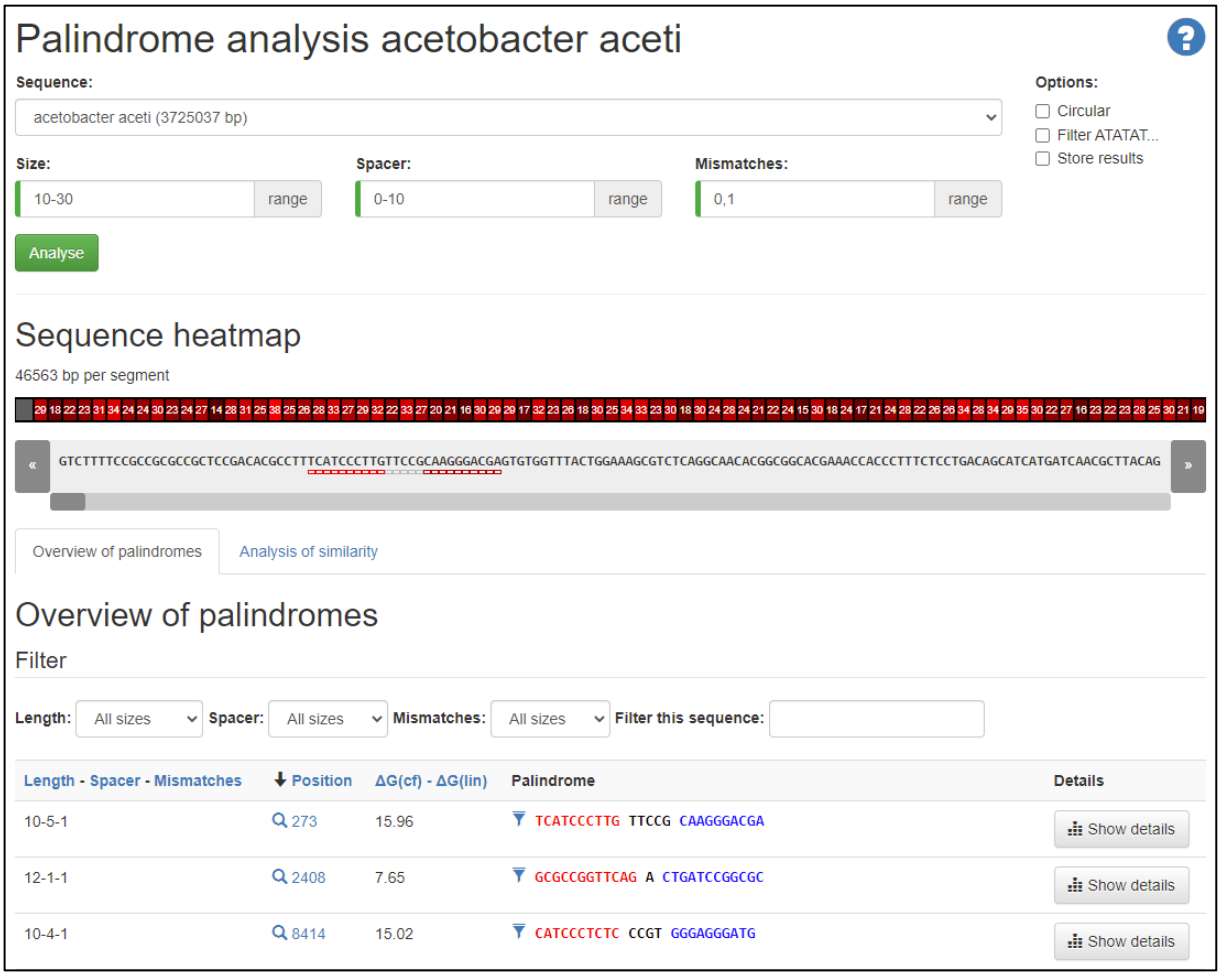
Nástroj *Palindrome analyser* je uživatelsky-přívětivá aplikace vyvinutá Biofyzikálním ústavem akademie věd ČR ve spolupráci Mendelovou univerzitou v Brně pro hledání inverzních repetic v DNA. Aplikace je dostupná z http://palindromes.ibp.cz/ [74]. Po načtení webové stránky se zobrazí úvodní obrazovka a po kliknutí na Palindrome analysis z nabídky



Obrázek 11 *Webová aplikace Palindrome analyser [74]*

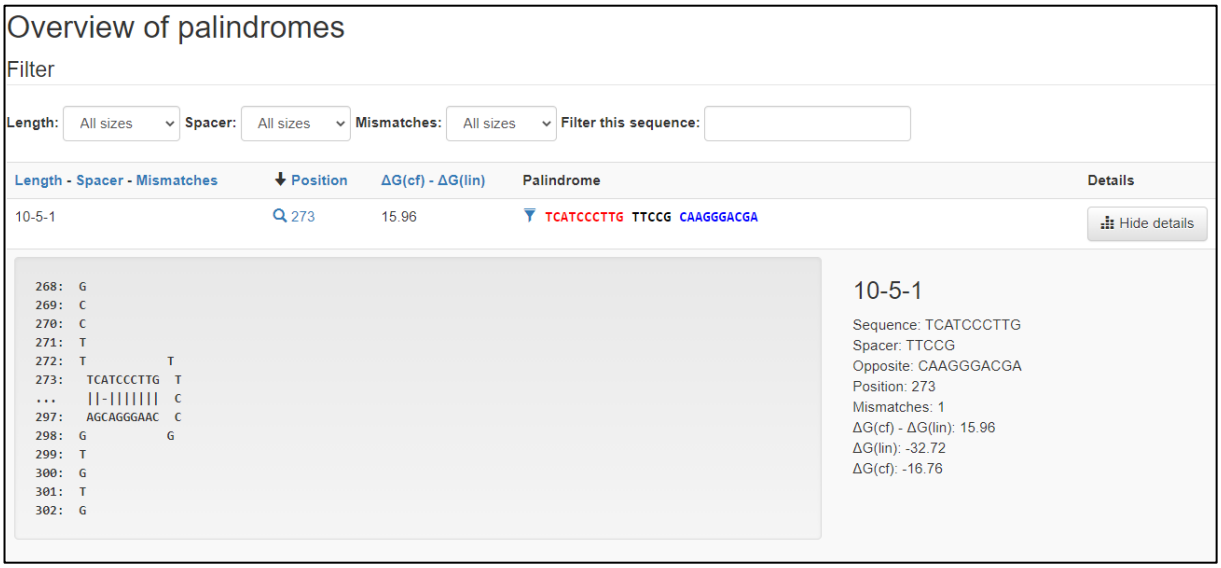
Analysis se zobrazí formulář pro zadání vstupní sekvence a parametrů analýzy (Obr.11). Parametr „size“ – určuje rozsah velikosti vyhledávaných inverzních repetic, „spacer“ – určuje rozsah počtu nukleotidů mezi jednotlivými inverzními sekvencemi, „mismatches“ – určuje maximální počet nekomplementárních bází v hledaných sekvencích. V poli „options“ je možné zatrhnout kružnicovou DNA a filtrování ATATAT sekvencí. Do pole „sequence“ je

zkopírována/zadána ručně sekvence určená k analýze. Aplikace také nabízí možnost bezplatné registrace. Výhodou po přihlášení je možnost importování sekvencí v textovém nebo FASTA formátu a také pomocí NCBI ID s následným uložením analýz s možností stáhnutí výsledků v .csv formátu. Vzhledem k náročnosti zobrazení výsledků je omezen maximální počet vyhledaných inverzních repetic na 5 000. Při překročení tohoto limitu je uživatel vyzván ke změně parametrů analýzy. Na obr. 12 je možné vidět příklad analýzy genomu *Acetobacter aceti*. Pod názvem analyzované sekvence můžeme vidět tzv. heatmap, která rozděluje analyzovanou sekvenci (v tomto případě celý genom) na jednotlivé segmenty délky 46563 bp. Segmenty obsahují číslici vyjadřující počet vyskytujících se IR v daném segmentu a barevné rozlišení (čím jasnější barva, tím vyšší počet IR). Pod nimi lze nalézt graficky vyznačené jednotlivé sekvence IR. O něco níže je celkový přehled nalezených IR s možností filtrace podle uvedených parametrů. Dále jsou zobrazeny jednotlivé výsledky, přičemž na každém řádku je zobrazena jedna IR. První sloupec zobrazuje typ IR ve formátu l-s-m (l – délka sekvence, s – počet nukleotidů mezi komplementárními sekvenci, m – počet nekomplementárních párů bází v IR),

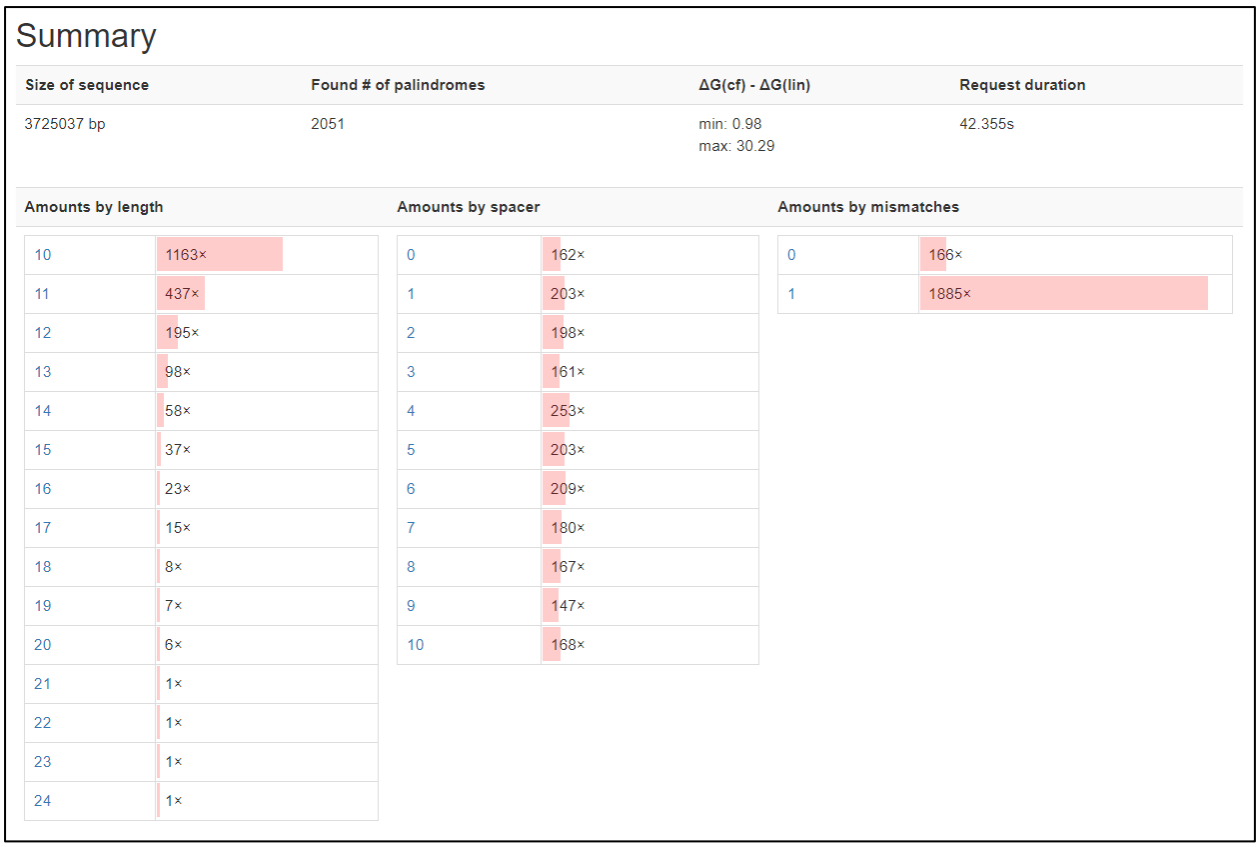


Obrázek 12 *Analýza pomocí Palindrome Analyser [74]*

druhý pozici IR, třetí rozdíl volné energie křížové a lineární struktury, čtvrtý graficky zobrazenou IR, pátý sloupec umožňuje zobrazení detailů IR (obr. 13). Na konci je zobrazeno celkové shrnutí analýzy s celkovým počtem nalezených IR a histogramy výskytu IR podle jednotlivých parametrů viz. obr 14. Navíc k celkovému shrnutí nabízí analýzu podobnosti (obr. 13 odkaz vedle „Overview of palindromes“), která zobrazuje četnost výskytu shodných sekvencí podle zvoleného filtru obr. 15.



Obrázek 13 *Analýza pomocí Palindrome Analyser [74]*

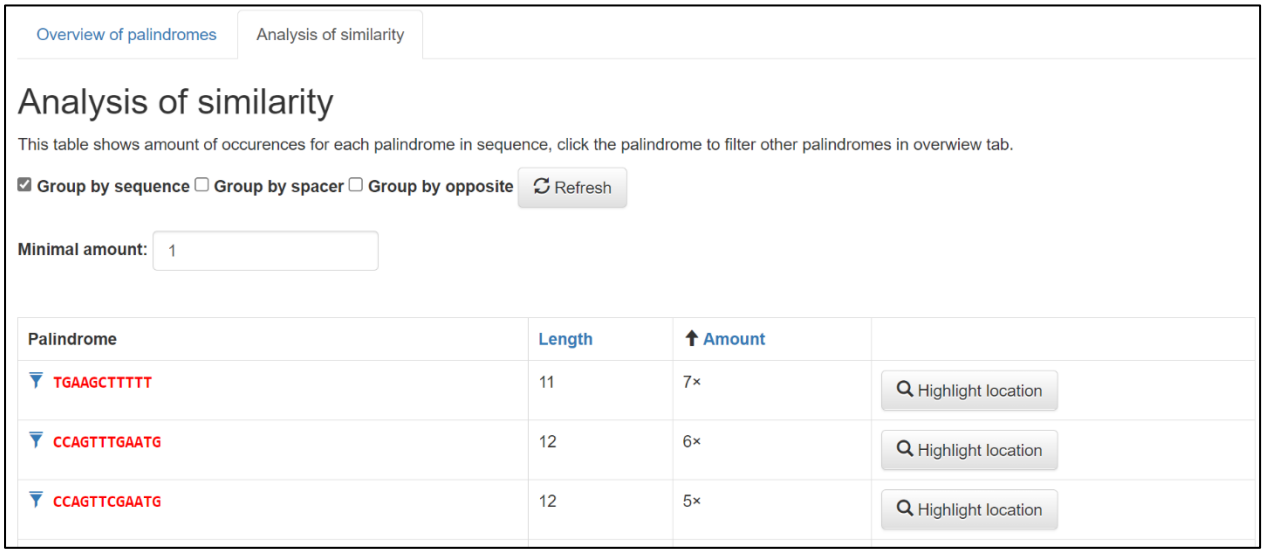


Summary

Size of sequence	Found # of palindromes	ΔG(cf) - ΔG(lin)	Request duration
3725037 bp	2051	min: 0.98 max: 30.29	42.355s

Amounts by length		Amounts by spacer		Amounts by mismatches	
10	1163×	0	162×	0	166×
11	437×	1	203×	1	1885×
12	195×	2	198×		
13	98×	3	161×		
14	58×	4	253×		
15	37×	5	203×		
16	23×	6	209×		
17	15×	7	180×		
18	8×	8	167×		
19	7×	9	147×		
20	6×	10	168×		
21	1×				
22	1×				
23	1×				
24	1×				

Obrázek 14 *Analýza pomocí Palindrome Analyser [74]*



Overview of palindromes | Analysis of similarity

Analysis of similarity

This table shows amount of occurences for each palindrome in sequence, click the palindrome to filter other palindromes in overwiew tab.

Group by sequence Group by spacer Group by opposite Refresh

Minimal amount: 1

Palindrome	Length	Amount	
TGAAGCTTTTT	11	7×	Highlight location
CCAGTTTGAATG	12	6×	Highlight location
CCAGTTCGAATG	12	5×	Highlight location

Obrázek 15 *Analýza pomocí Palindrome Analyser [74]*

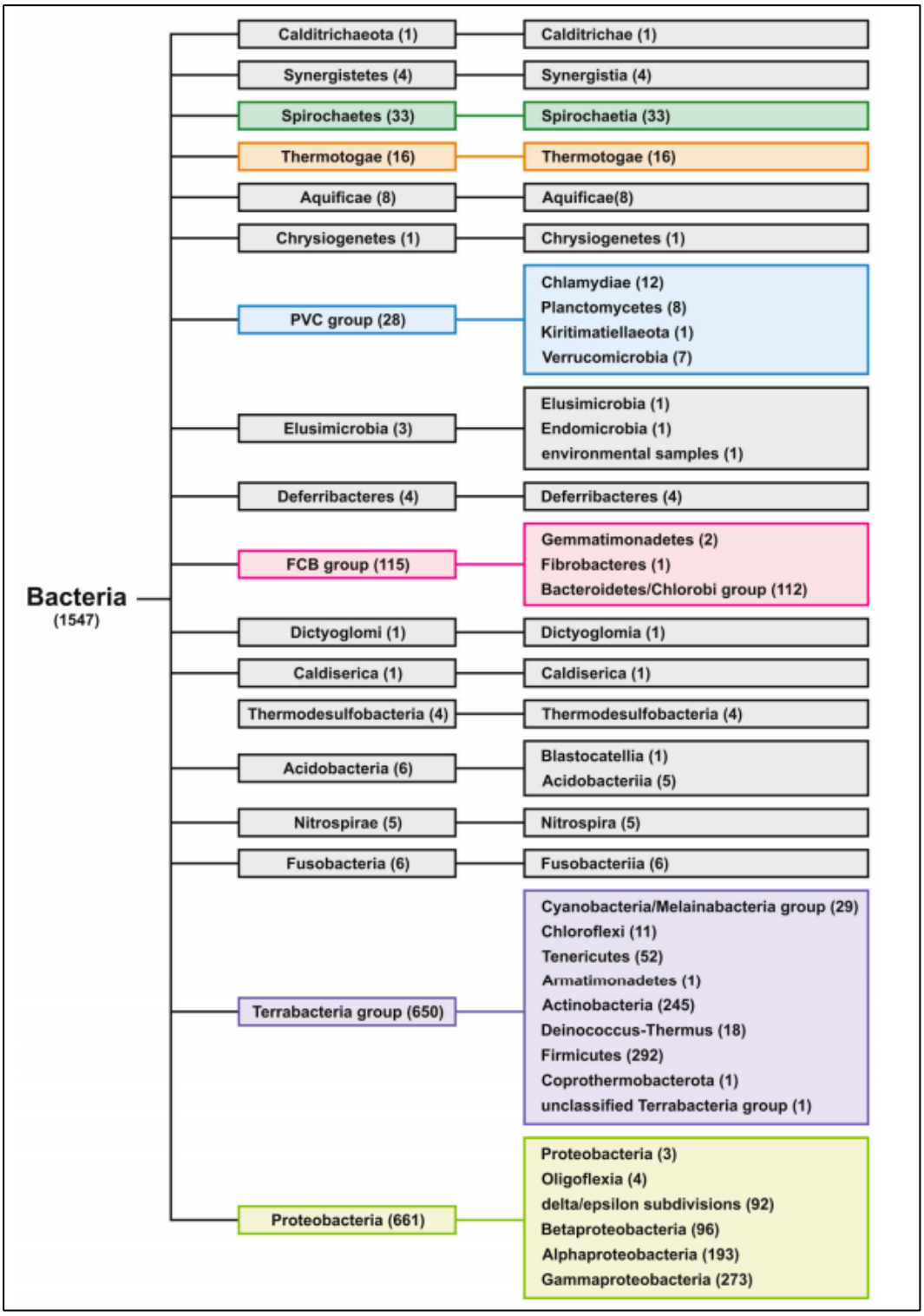
3 EXPERIMENTÁLNÍ ČÁST

3.1 Cíl práce

Cílem bylo vypracovat literární rešerši na téma inverzní repetice. Popsat význam a funkci inverzních repetic v živých organismech a jejich vliv na DNA. Dále zaměření na křížové struktury vznikající z inverzních repetic a jejich charakteristiku. Funkce křížových struktur v organismech, jejich vliv na strukturu DNA a na buněčné funkce. Charakterizace proteinů vážících se na křížové struktury. Využití nástroje *Palindrome analyser* pro charakterizaci a lokalizaci inverzních repetic ve vybraném vzorku bakteriálních genomů.

3.2 Přehled analyzovaných genomů

Pro analýzu inverzních repetic bylo zvoleno 1 547 druhů bakterií roztříděných do jednotlivých skupin a podskupin podle NCBI taxonomie. Rozdělení je možné vidět na fylogenetickým stromu, viz. obr. 16. Celkem bylo analyzováno 1 627 genomů (některé bakterie



Obrázek 16 Fylogenetický strom analyzovaných bakterií (převzato z *[81]*)

mají 2 genomy). Pro statistickou analýzu, byly použity pouze skupiny s 10 a více analyzovanými genomy (barevně zvýrazněny, viz obr. 16).

3.3 Metody

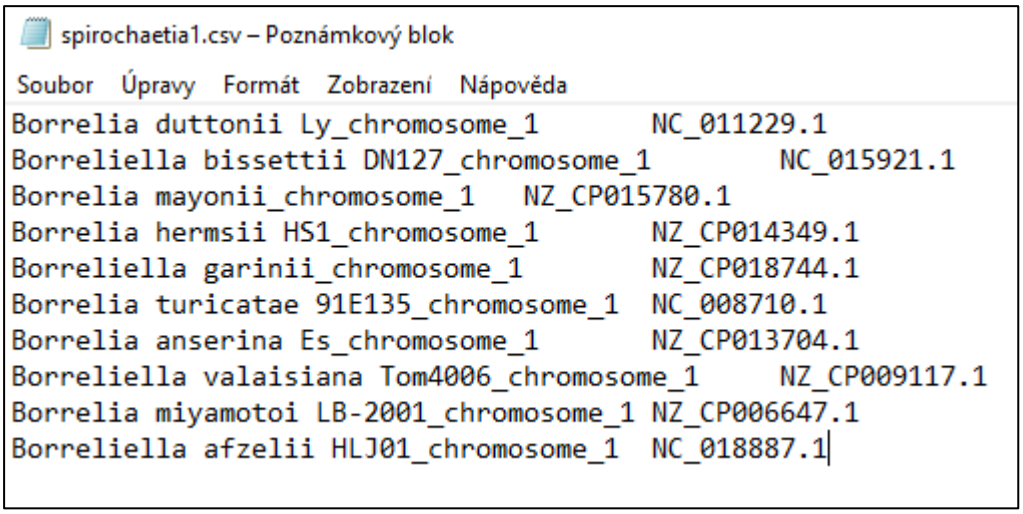
Pro samotnou analýzu byl použit nástroj *Palindrome analyser* (popsaný výše) v jeho offline verzi. Práce s nástrojem *Palindrome analyser* v offline verzi se provádí pomocí příkazů v příkazové řádce. Nejprve je nutné nastavit pracovní adresář tak, aby obsahoval složku s programem *Palindrome analyser* (dostupný z kolomaznik - genetika). Použitelné příkazy si lze zobrazit pomocí spuštění příkazu genetika s parametrem – h, viz. obr. 17. Parametr – acsv stáhne a analyzuje genomy přímo z databáze NCBI, k tomu je zapotřebí vytvořit soubor ve formátu .csv nebo .txt, který obsahuje název analyzovaného genomu (může být libovolný) a

```
C:\genetika>genetika -h
usage: utility-name
 -acsv,--analyse-csv <arg>     Download and analyse genomes and feature
                               tables from NCBI, input is CSV file with
                               name and ID. Remember to set target
                               parameter.
 -adir,--analyse-dir <arg>     Download and analyse genomes and feature
                               tables from NCBI, input is folder with CSV
                               files with name and ID. Remember to set
                               target parameter.
 -c,--circular                 Use this switch to set circular mode.
 -df,--dir-fasta <arg>         Path to FASTA files for bulk analysis.
 -dlcsv,--download-csv <arg>   Download genomes and feature tables from
                               NCBI, input is CSV file with name and ID.
                               Remember to set target parameter.
 -dldir,--download-dir <arg>   Download genomes and feature tables from
                               NCBI, input is folder with CSV files with
                               name and ID. Remember to set target
                               parameter.
 -dr,--dir-raw <arg>           Path to raw files for bulk analysis.
 -e,--extension <arg>          File extension, default is 'fasta' or 'txt'
                               for raw.
 -h,--help                     Print this help message.
 -irmax <arg>                  Max length of IR
 -irmin <arg>                  Min length of IR
 -t,--target <arg>             Set target path for results. Mandatory for
                               download.
C:\genetika>
```

Obrázek 17 *Příkazová řádka programu palindrome analyser [74]*

NCBI identifikátor oddělený tabulátorem. V případě více analyzovaných genomů – je potřeba uvést každý další na nový řádek. Parametr -adir je podobný s tím rozdílem, že vstupem jsou všechny .csv soubory v dané složce. Další parametry -df a, -dldir umožňují stáhnutí jednotlivých genomů z databáze NCBI bez analýzy a parametry -df a, -dr umožňují přímou analýzu ze stažených souborů.

Pro analýzu bylo nutné rozdělit genomy do několika skupin zhruba po 10–20 genomech, protože analýza celého souboru genomů naráz nebyla možná, z důvodu přetečení paměti programu. Na obr. 18 je možné vidět ukázku .csv souboru s genomy ze skupiny spirochaetia.

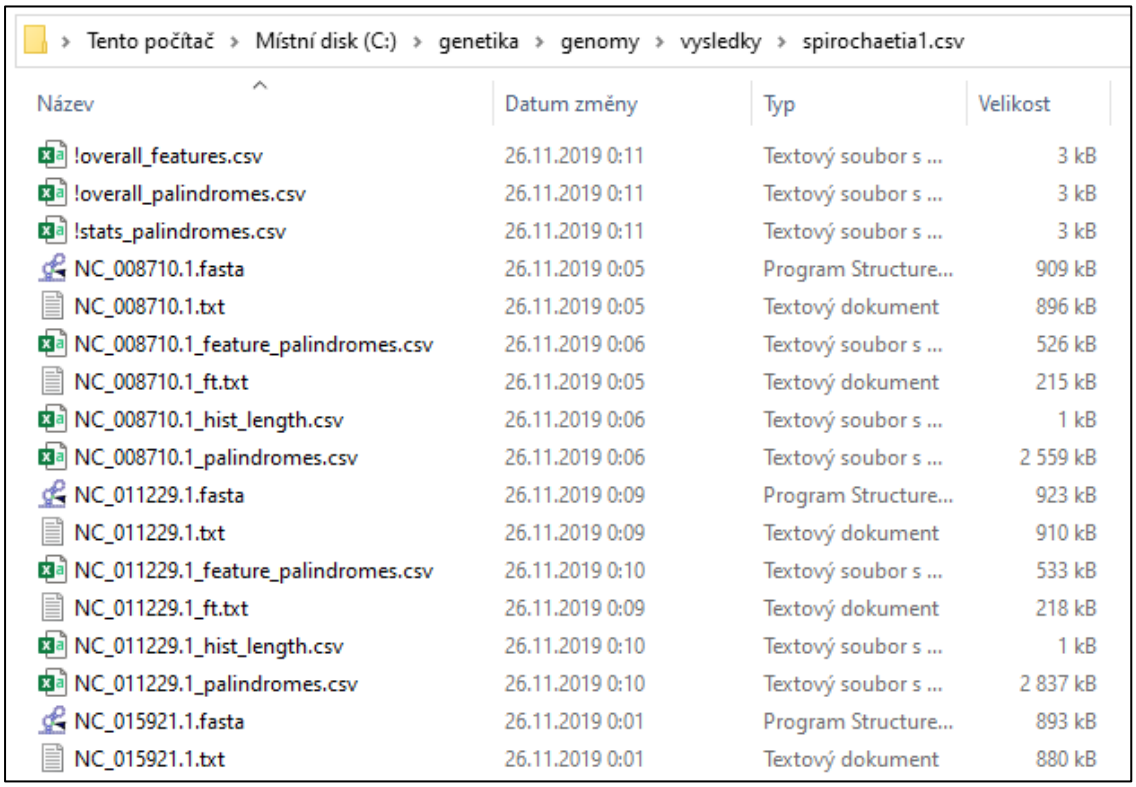


Obrázek 18 *.csv soubor s genomy*

Z takové souboru provedeme analýzu pomocí příkazu:

genetika -acsv C:\genetika\genomy\spirochaetia1.csv -t C:\genetika\genomy\vysledky

V parametru -acsv je nutné uvést celou cestu, kde se nachází .csv soubor s genomy a také další parametr -t, za kterým následuje specifikace cílové složky kam se nahrají stažené soubory a



Obrázek 19 *složka s výsledky analýzy*

výsledky analýzy. Na obr. 19 je možné vidět výslednou složku s výsledky. Ve složce se nachází jednak stažené genomy a jejich funkční oblasti v textovém a FASTA formátu a dále jednotlivé .csv soubory s výsledky analýzy. Specifikaci jednotlivých funkčních oblastí vyskytujících se v souborech s příponou _ft.txt je možné najít na http://www.insdc.org/documents/feature-table#7.2. Na začátku jsou soubory s celkovou analýzou inverzních repetic a analýzou IR vzhledem k funkčním oblastem DNA (overall_palindromes.csv a overall_features.csv). Další .csv soubory s příponou _palindromes.csv obsahují celkový výčet inverzních repetic, jejich lokalizaci, sekvenci, typ a délku. Protože jsou v souborech s příponou .csv hodnoty jednotlivých položek uvedené v uvozovkách a oddělené čárkami, je nutné je pro další zpracování transformovat např. pomocí Excelu. Na obr. 20 je možné vidět soubor overall_palindromes.csv načtený do excelu. Na každém řádku jsou uvedeny výsledky analýzy daného genomu. V jednotlivém sloupci je uveden postupně název genomu, tak jak byl uveden v .csv souboru, potom následují NCBI ID a dále údaje o velikosti genomu (počet bp), obsah GC párů. Dále následují frekvence výskytu všech IR, frekvence výskytu IR o délce 8 bp+, 10 bp+ a 12 bp+. Frekvence výskytu je ve výchozím nastavení vyjádřena jako počet IR na 1 bp analyzovaného genomu. Další sloupce obsahují údaje s počtem IR podle jednotlivých kritérií (all – všechny IR, 8 – IR o délce 8 bp a delší, 10 – IR o délce 10 bp a delší, 12 – IR o délce

Column1	Column2	Column3	Column4	Column5	Column6
Title	NCBI	Size	GC count	fr. all	fr. 8+
Borreliella bissettii DN127_chromosome_1	NC_015921.1	900755	258542	0.05738963	0.00876709
Borreliella garinii_chromosome_1	NZ_CP018744.1	905638	256870	0.05781449	0.00903783
Borrelia miyamotoi LB-2001_chromosome_1	NZ_CP006647.1	907293	260558	0.05178702	0.00716196
Borrelia anserina Es_chromosome_1	NZ_CP013704.1	906833	267413	0.04860763	0.00666054
Borrelia turicatae 91E135_chromosome_1	NC_008710.1	917330	267117	0.05112882	0.00712721
Borreliella afzelii HLJ01_chromosome_1	NC_018887.1	905471	256432	0.05774895	0.00900526
Borreliella valaisiana Tom4006_chromosome_1	NZ_CP009117.1	912160	256525	0.05843054	0.00906091
Borrelia mayonii_chromosome_1	NZ_CP015780.1	904387	256076	0.05804263	0.00893755
Borrelia duttonii Ly_chromosome_1	NC_011229.1	931674	257012	0.05579527	0.00801568
Borrelia hermsii HS1_chromosome_1	NZ_CP014349.1	922500	275143	0.04946016	0.00677073

Column7	Column8	Column9	Column10	Column11	Column12	Column13	Column14	Column
fr. 10+	fr. 12+	all	8	10	12	6	7	
0.00134665	0.00029864	51694	7897	1213	269	30930	12867	
0.00133166	0.00032684	52359	8185	1206	296	31235	12939	
0.00099417	0.00018627	46986	6498	902	169	28958	11530	
0.00084801	0.00017864	44079	6040	769	162	27457	10582	
0.00095277	0.00017442	46902	6538	874	160	28910	11454	
0.00138271	0.0003258	52290	8154	1252	295	31247	12889	
0.0013616	0.00028942	53298	8265	1242	264	31714	13319	
0.00138215	0.00030297	52493	8083	1250	274	31407	13003	
0.0011152	0.00020393	51983	7468	1039	190	31656	12859	
0.00093659	0.00017995	45627	6246	864	166	28310	11071	

Obrázek 20 *soubor overall_palindromes.csv po transformaci do excelu*

12 bp a delší). Poslední sloupce potom obsahují nalezené počty IR jednotlivých délek až do délky 30 (není vyobrazeno na obr. 20). Takto analyzované genomy byly spojeny do jednoho souboru a roztříděny do jednotlivých skupin a dále zpracovány viz. příloha Výsledky.xls.

Feature	Feature coun	Feature total	Avg feature s	fr. all inside	fr. 8+ inside	fr. 10+ inside	fr. 12+ inside
CDS	8155	8485294	1040.5020233	0.05200692	0.00741671	0.00099148	0.00016994
STS	2	1321	660.5	0.04012112	0.00454201	0	0
tRNA	319	24000	75.23510972	0.03904167	0.00345833	0.00025	0
misc_feature	4	2502	625.5	0.0363709	0.00479616	0	0
gene	8566	8611782	1005.34461826	0.05177697	0.00737315	0.00098423	0.00016814
rRNA	41	62580	1526.34146341	0.02937041	0.0035954	0.00036753	0
tmRNA	9	3257	361.88888889	0.03592263	0.00644765	0.00153516	0
ncRNA	19	4369	229.94736842	0.03936828	0.01075761	0.00183108	0.00022889

fr. all around	fr. 8+ around	fr. 10+ around	fr. 12+ around	fr. all before	fr. 8+ before	fr. 10+ before	fr. 12+ before
0.06871919	0.01142612	0.00211833	0.00069099	0.06624647	0.01091355	0.00187983	0.00050399
0.055	0.005	0	0	0.05	0.01	0	0
0.05857367	0.00863636	0.00159875	0.00070533	0.0638558	0.0100627	0.00181818	0.00087774
0.0675	0.0075	0.00125	0	0.0375	0.005	0.0025	0
0.06827107	0.01131158	0.00209258	0.00068877	0.06606584	0.01085104	0.00186318	0.00051483
0.05707317	0.01109756	0.00182927	0.00036585	0.05341463	0.00707317	0.0002439	0
0.07055556	0.01166667	0.00111111	0	0.07	0.00777778	0.00111111	0
0.06131579	0.00789474	0.00105263	0.00026316	0.05263158	0.00736842	0.00052632	0.00052632

fr. all after	fr. 8+ after	fr. 10+ after	fr. 12+ after	all inside	8+ inside	10+ inside	12+ inside
0.07119191	0.01193869	0.00235684	0.00087799	441294	62933	8413	1442
0.06	0	0	0	53	6	0	0
0.05329154	0.00721003	0.00137931	0.00053292	937	83	6	0
0.0975	0.01	0	0	91	12	0	0
0.0704763	0.01177212	0.00232197	0.00086271	445892	63496	8476	1448
0.06073171	0.01512195	0.00341463	0.00073171	1838	225	23	0
0.07111111	0.01555556	0.00111111	0	117	21	5	0
0.07	0.00842105	0.00157895	0	172	47	8	1

Obrázek 21 *overall_features.csv po transformaci do excelu*

Soubor overall_features.csv obsahuje obdobné shrnutí pro výskyt inverzních repetic vzhledem k funkčním oblastem, viz. obr. 21. Na každém řádku je uveden typ funkční oblasti, celkový počet, součet celkové délky funkčních oblastí, průměrná délka a dále následují jednotlivé frekvence výskytu IR vztažené na celkovou délku funkční oblasti a rozdělené podle výskytu IR (nacházející se uvnitř funkční oblasti, do 100 bp před a do 100 bp následující po dané funkční oblasti). Za sloupci s hodnotami frekvencí výskytu následuje obdobně jako u souboru výše sloupce s počty nalezených inverzních repetic roztříděné podle výskytu (uvnitř, před a po) a podle délky (všechny, 8 bp a delší, 10 bp a delší, 12 bp a delší).

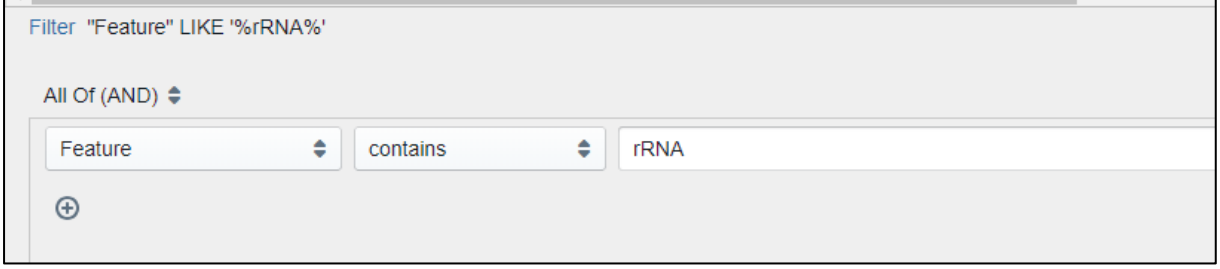
Vzhledem k tomu, že genomy byly analyzovány po skupinách, musela být obdobná tabulka pro celý soubor genomů sestrojena znovu z jednotlivých souborů. Za tímto účelem byly pomocí příkazu

copy *_feature_palindromes.csv souhrn_feature_palindromes.csv

všechny soubory z celé analýzy s příponou feature_palindromes.csv spojeny do jednoho souboru souhrn_feature_palindromes.csv. Vzhledem k jeho velikosti cca 3,11 GB, jej nebylo

možné načíst do tabulkového procesoru excel. Z tohoto důvodu byl zvolen program Tad (dostupný z [75]), využívající SQLite databázi schopný zpracovat tak velký soubor dat. Po načtení souboru do programu Tad, byly vyfiltrovány statistiky k jednotlivým funkčním oblastem a sečteny jednotlivé počty nalezených inverzních repetic. Na obr. 22 je možné vidět příklad filtraci řádků podle rRNA funkční oblasti (oblasti DNA kódující ribozomální RNA) a jednotlivé součty IR. Dále vidíme, že počet všech nalezených inverzních repetic ze všech analyzovaných genomů vyskytujících se uvnitř (all inside) funkční oblasti DNA kódující rRNA je 776 127, v případě IR o délce 8+ je celkový počet 89 156 atd. Z takto vyfiltrovaných výsledků byla zrekonstruována obdobná tabulka jako na obr. 21 pro celý analyzovaný soubor genomů, ze které mohl být dále sestaven graf, viz. kapitola výsledky.

Feature	Info	Feature start	Feature end	Feature size	all inside	8+ inside
rRNA		**33 415 129 009**	**33 442 489 170**	**27 360 161**	**776 127**	**89 156**
rRNA	product (16S ribosomal RNA)	170 007	171 531	1 524	49	8
rRNA	product (23S ribosomal RNA)	171 727	174 646	2 919	91	10
rRNA	product (5S ribosomal RNA), inference (COORDINATES: n...	174 685	174 793	108	1	0
rRNA	product (16S ribosomal RNA)	66 558	68 102	1 544	44	6
rRNA	product (23S ribosomal RNA)	68 561	71 458	2 897	93	8
rRNA	product (5S ribosomal RNA), inference (COORDINATES: n...	71 588	71 701	113	3	0
rRNA	product (16S ribosomal RNA)	172 274	173 818	1 544	44	6
rRNA	product (23S ribosomal RNA)	174 277	177 174	2 897	93	8
rRNA	product (5S ribosomal RNA), inference (COORDINATES: n...	177 304	177 417	113	3	0
rRNA	product (16S ribosomal RNA)	133 836	135 389	1 553	54	9
rRNA	product (23S ribosomal RNA)	135 630	138 576	2 946	70	11
rRNA	product (5S ribosomal RNA), inference (COORDINATES: n...	138 691	138 805	114	4	0
rRNA	product (16S ribosomal RNA)	155 794	157 347	1 553	54	9
rRNA	product (23S ribosomal RNA)	157 588	160 534	2 946	70	11
rRNA	product (5S ribosomal RNA), inference (COORDINATES: n...	160 649	160 763	114	4	0
rRNA	product (5S ribosomal RNA), inference (COORDINATES: n...	260 266	260 369	103	0	0
rRNA	product (5S ribosomal RNA), inference (COORDINATES: n...	260 543	260 646	103	0	0
rRNA	product (23S ribosomal RNA)	810 397	813 298	2 901	87	5



Obrázek 22 Filtrované výsledky v programu Tad

4 VÝSLEDKY

4.1 Frekvence výskytu inverzních repetic v bakteriálních genomech

Analyzovaný dataset obsahoval 1627 genomů z 1547 druhů bakterií. Celková délka genomů se pohybovala od 298 kbp do 20.20 Mbp. Průměrný obsah GC párů ve vzorku byl 50,6 %, s minimem 20,2 % u *Buchnera aphidicola* (Gammaproteobakterie) a maximem 74,70 % u *Corynebacterium sphenisci* (Aktinobakterie). Nejvyšší zastoupení zaujímaly IR o délce 6-7 bp (86,6 % všech IR), méně početné byly IR s délkou 8 bp a delší (13,41 všech IR) a IR s délkou 10 a delší (1,87 % všech IR) a nejméně zastoupení byly IR s délkou 12 a delší (0,42 % všech IR). Tedy čím delší IR, tím menší je její pravděpodobnost výskytu. Celkový souhrn počtu nalezených IR a jejich frekvenci výskytu z celého analyzovaného souboru je znázorněn v tabulce 1.

Tabulka 1 *Celkový počet inverzních repetic jednotlivých délek a frekvence jejich výskytu*

Délka IR	Počet IR	Frekvence výskytu IR na 1000 bp
Vše	241 419 782	41,87
6-7	209 053 900	36,20
8+	32 365 882	5,66
10+	4 504 521	0,80
12+	1 014 643	0,18

Dále byl soubor genomů rozdělen do barevně rozlišených skupin a podskupin, viz. obr 16 fylogenetický strom. Pro statistické vyhodnocení byly brány pouze skupiny a podskupiny obsahující více jak 10 sekvenovaných genomů, zbytek byl zařazen do skupiny ostatní. V tabulce 2 je možné vidět počet analyzovaných sekvencí u jednotlivých individuálních skupin a podskupin a u celého celku. Dále jsou v tabulce 2 uvedené jednotlivé statistické ukazatele, Median, nejkratší a nejdelší genom, průměrný obsah GC párů v genomech, celkový počet nalezených IR a průměrné frekvence nalezených IR včetně minimálních a maximálních hodnot. Průměrná frekvence výskytu inverzních repetic u všech bakteriálních genomů byla 41,87 IR na 1000 bp. V případě skupin byla nejmenší průměrná frekvence výskytu inverzních repetic u skupiny Thermotogae (37,06 IR/kbp) a nejvyšší u skupiny Terrabacteria (44,01 IR/kbp), následované Spirochaetes (43,14 IR/kbp). V případě podskupin byla nejmenší průměrná frekvence výskytu inverzních repetic u podskupiny Chlamydiae (35,66 IR/kbp) a nejvyšší u Tenericutes (61,89 IR/kbp) následovaná skupinou Actinobacteria (49,26 IR/kbp). Nejvyšší frekvence inverzních repetic byla nalezena u bakterie Buchnera aphidicola (BCc) a to 113,38

IR/kbp zároveň s nejnižším obsahem GC párů 20,1 %. Nejnižší frekvence inverzních repetic byla nalezena u bakterie Anaplasma centrale 27,08 IR/kbp.

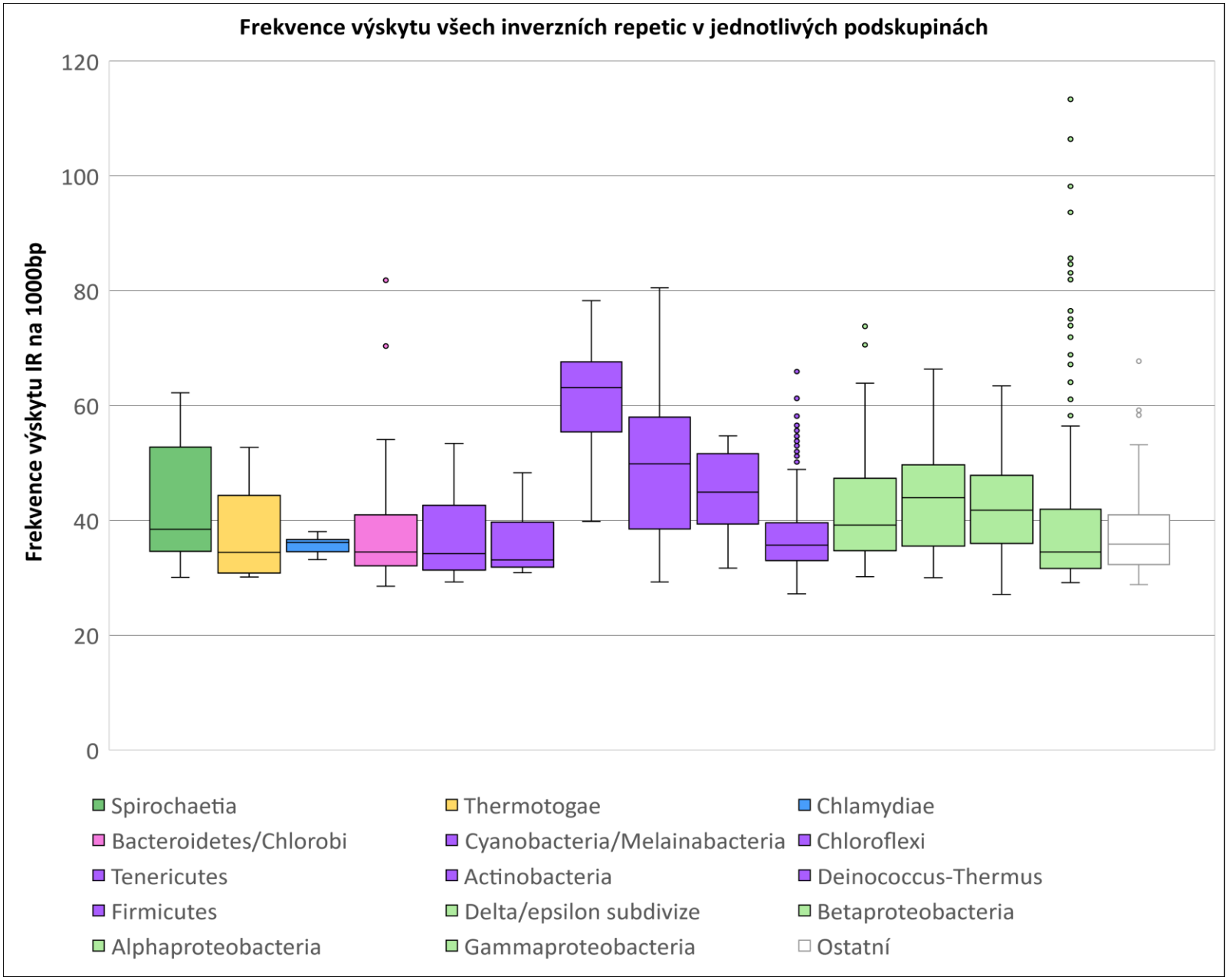
Tabulka 2 *Celkový počet inverzních repetic v jednotlivých skupinách a podskupinách a frekvence jejich výskytu. Sekvence (počet analyzovaných genomů), Median (medián délky genomů), Nejkratší (nejkratší genom), Nejdelší (nejdelší genom), GC (průměrný obsah GC párů v genomech), Počet IR (počet nalezených IR všech délek), Průměr (průměrná frekvence výskytu IR vztažená na 1000bp v dané skupině/podskupině), Min (nejmenší nalezená frekvence výskytu IR vztažená na 1000 bp), Max (největší*

Doména	Sekvence	Median (bp)	Nejkratší (bp)	Nejdelší (bp)	GC (%)	Počet IR	Průměr	Min	Max
Bakterie	1627	3307820	83026	13033779	50,6	241419783	41,87	27,08	113,37
Skupina	**Sekvence**	**Median (bp)**	**Nejkratší (bp)**	**Nejdelší (bp)**	**GC (%)**	**Počet IR**	**Průměr**	**Min**	**Max**
Spirochaetes	38	2646038	277655	4653970	39,7	3490999	43,14	30,06	62,24
Thermotogae	16	2150379	1884562	2974229	39,1	1266797	37,06	30,12	52,72
PVC skupina	28	2917407	1041170	9629675	50,7	3768304	37,17	29,52	67,76
FCB skupina	117	3914632	605745	9127347	42,3	17386216	37,44	28,52	81,83
Terrabacteria	659	3018755	91776	11936683	50,4	107574354	44,01	27,24	80,52
Proteobacteria	724	3551512	83026	13033779	53,4	103582330	41,09	27,08	113,37
Ostatní	45	2157835	1012010	6237577	44,3	4350782	37,97	28,80	59,22
Podskupina	**Sekvence**	**Median (bp)**	**Nejkratší (bp)**	**Nejdelší (bp)**	**GC (%)**	**Počet IR**	**Průměr**	**Min**	**Max**
Spirochaetia	38	2646038	277655	4653970	39,7	3490999	43,14	30,06	62,24
Thermotogae	16	2150379	1884562	2974229	39,1	1266797	37,06	30,12	52,72
Chlamydiae	12	1168953	1041170	3072383	40,3	688281	35,66	33,18	38,08
Bacteroidetes/Chlorob	114	3878527	605745	9127347	41,9	16875955	37,42	28,52	81,83
Cyanobacteria/Melain	29	5315554	1657990	9673108	42,6	5105883	37,57	29,28	53,44
Chloroflexi	12	2333610	1252731	5723298	57,0	1222098	36,09	30,88	48,34
Tenericutes	52	981001	564395	1877792	28,0	3250026	61,89	39,82	78,29
Actinobacteria	246	3960961	775354	11936683	66,2	60232821	49,26	29,30	80,52
Deinococcus-Thermus	18	2895913	2035182	3881839	66,8	2265314	45,21	31,71	54,71
Firmicutes	298	2835823	91776	8739048	40,8	35161375	37,56	27,24	65,93
Delta/epsilon subdivize	92	3136746	1457619	13033779	49,0	14875784	41,93	30,20	73,82
Betaproteobacteria	110	3763620	820037	6987670	60,6	18652683	43,45	30,05	66,39
Alphaproteobacteria	213	3424964	83026	9207384	58,1	30056720	42,48	27,08	63,42
Gammaproteobacteria	302	3777066	200073	7783862	48,8	39242805	39,15	29,16	113,37
Ostatní	75	2664102	1012010	9629675	49,0	9032241	37,51	28,80	67,76

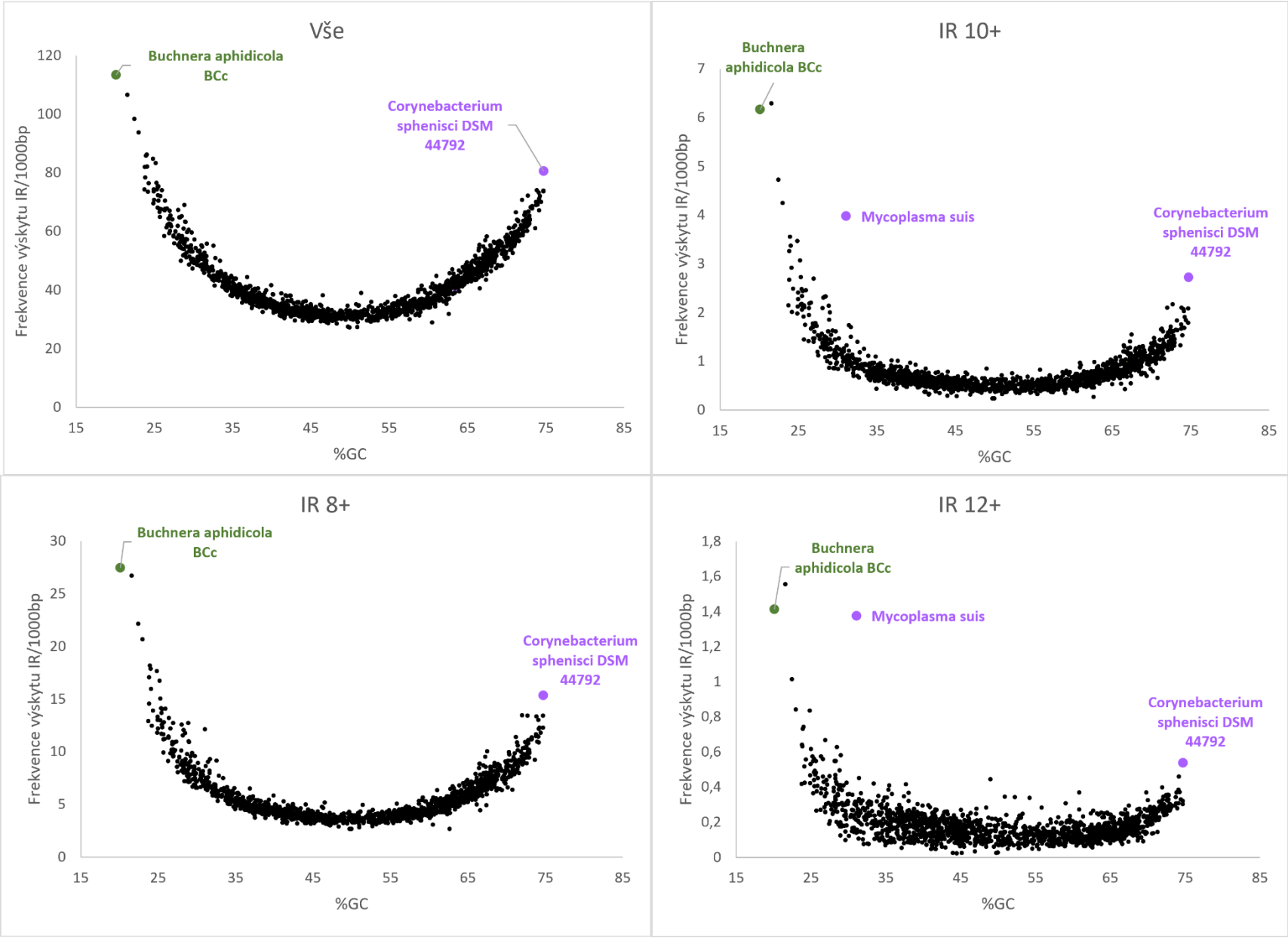
Dále byl pro podskupiny z tab. 2 vytvořen krabicový/boxový graf (obr.23) zachycující detailní charakteristiky výskytu IR v jednotlivých podskupinách. Jednotlivé boxy reprezentují danou podskupinu a jsou ohraničeny shora 3. kvartilem a zespodu 1. kvartilem, mezi nimi se nachází linie vymezující medián. Dále obsahují linie (tzv. vousy), které vycházejí ze střední části boxů kolmo nahoru a dolů, vyjadřující variabilitu dat pod prvním a nad třetím kvartilem s maximální a minimální hodnotou v mezikvartilovém rozpětí. Jednotlivé tečky nad boxy pak znázorňují odlehlé výsledky. Odlehlé výsledky jsou stanoveny na základě jejich hodnot, které leží mimo oblast 1,5násobku mezikvartilového rozpětí. Z grafu je patrné, že genomy z podskupiny Tenericutes mají nejvyšší frekvenci výskytu IR. V případě odlehlých výsledků je nejvíce obsaženo v podskupinách Firmicutes a Gammaproteobacteria. Ve skupině Gammaproteobacteria se jedná především o genomy (celkem 14) bakterie Buchnera aphidicola, izolované z různých hostitelů. Ve skupině Firmicutes se jedná především o bakterie z rodu Clostridium.

Jako další byla sledována závislost frekvence výskytu inverzních repetic na obsahu GC párů. Za tímto účelem byly sestrojeny grafy závislosti frekvenci výskytu IR jednotlivých délek na obsahu GC párů (%) u všech analyzovaných genomů. Na obr. 24 je možné vidět celkem 4 grafy znázorňující závislost frekvence výskytu IR na obsahu GC párů. Jednotlivé popisky u grafů – VŠE, 8+, 10+ a 12+ značí délku zahrnutých IR.

V případě grafu pro frekvenci výskytu IR všech délek je patrné, že zde existuje korelace mezi frekvencí a obsahem GC párů, viz. obr 24. Minimální frekvence výskytu IR se nachází uprostřed, kde obsah GC párů je okolo 50 % a postupně symetricky roste na obě strany až do svého maxima. Tato závislost přestává být s rostoucí délkou inverzních repetic pozorovatelná, jak je možné vidět v případě grafu pro IR 12+. Dále jsou zde zvýrazněny genomy Buchnera aphidicola a Corynebacterium sphenisci DSM 44792, mající minimální a maximální obsah GC párů z celého souboru analyzovaných genomů a jejich korespondující hodnoty frekvence výskytu IR, které jsou taktéž maximální z celého souboru s výjimkou grafu pro IR 10+ a IR 12+. Dále je v grafu pro IR 10+ a IR 12+ zvýrazněn mikroorganismus Mycoplasma suis, jehož frekvence výskytu IR, délky 10+ a 12+ vzhledem k obsahu GC párů výrazně vybočuje z celého souboru.



Obrázek 23 *Boxový graf frekvence výskytu všech IR v jednotlivých podskupinách*



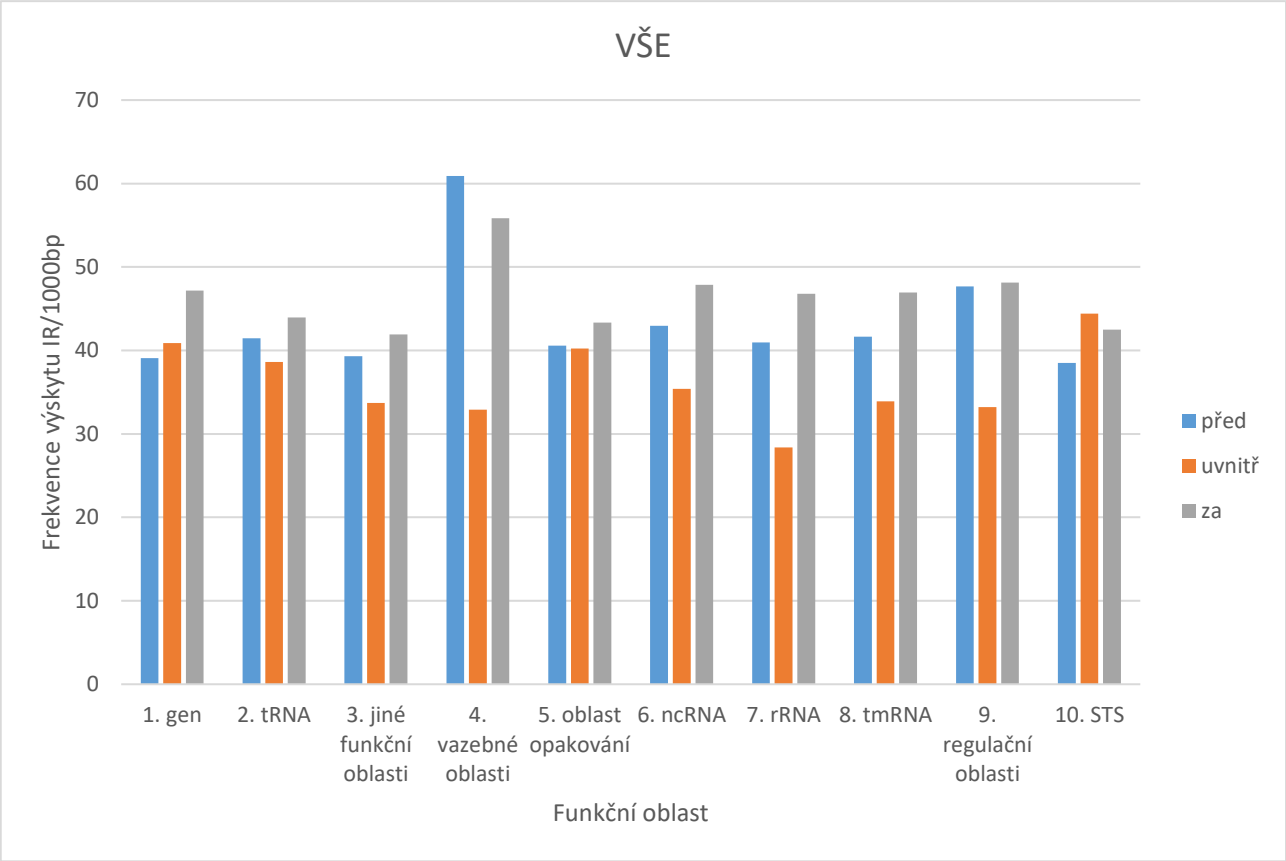
Obrázek 24 *Grafy závislosti frekvence výskytu IR různých délek na procentuálním obsahu GC párů*

4.2 Lokalizace inverzních repetic vzhledem k funkčním oblastem DNA

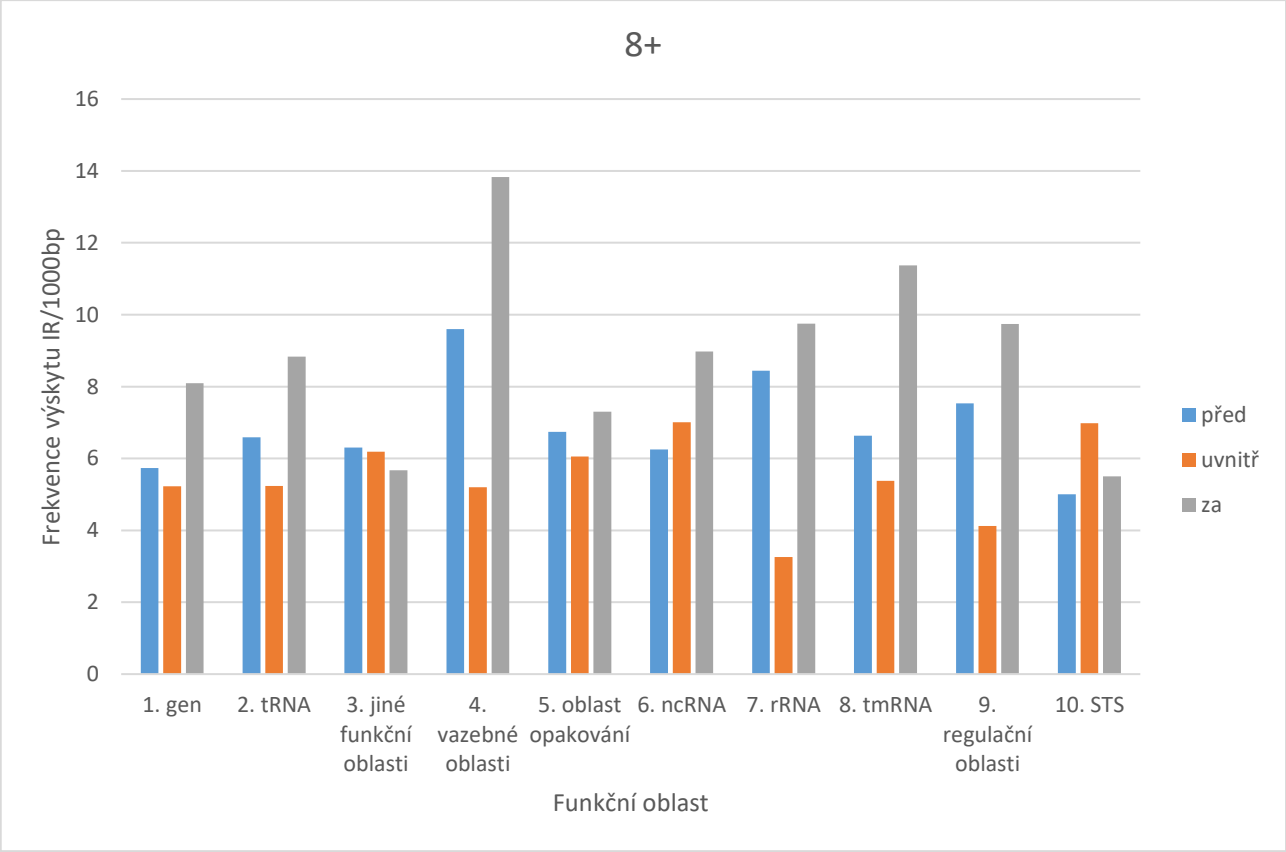
Kromě analýzy frekvence výskytu inverzních repetic v celém genomu byla analýza taktéž zaměřena na sledování frekvence jejich výskytu „uvnitř“ daných funkčních oblastí, do 100 bp „před“ a do 100 bp „za“ nimi. Funkční oblast DNA je úsek DNA, u kterého byla objevena nějaká funkce, například geny pro tvorbu proteinů, RNA nebo různé regulační oblasti. Soubory s polohou výskytu funkčních oblastí jednotlivých genomů byly automaticky staženy programem *Palindrome analyser* z NCBI databáze a rovnou analyzovány, .csv výstupy poté byly zpracovány viz. teorie výše.

Celkem bylo sledováno 10 funkčních oblastí – gen (oblasti s nacházejícími se geny, např. tvorba proteinů, enzymů atd.), tRNA (oblast kódující transferovou RNA), jiné funkční oblasti (funkční oblasti nezařazené do ostatních skupin), různé vazebné oblasti (oblasti ke kterým se váží kovalentně či nekovalentně jiné molekuly, kromě proteinů a primerů), oblasti opakování (oblast obsahující opakující se sekvence nukleotidů), ncRNA (protein-nekódující oblast, která kóduje jinou než ribozomální a transferovou RNA), rRNA (oblast kódující ribozomální RNA), tmRNA (oblast kódující transferovou RNA), regulační oblasti (jakákoliv oblast regulující funkci transkripce, translace, replikace nebo chromatinové struktury), STS (sekvence DNA, která obsahuje orientační bod v genomu, který lze detekovat pomocí PCR).

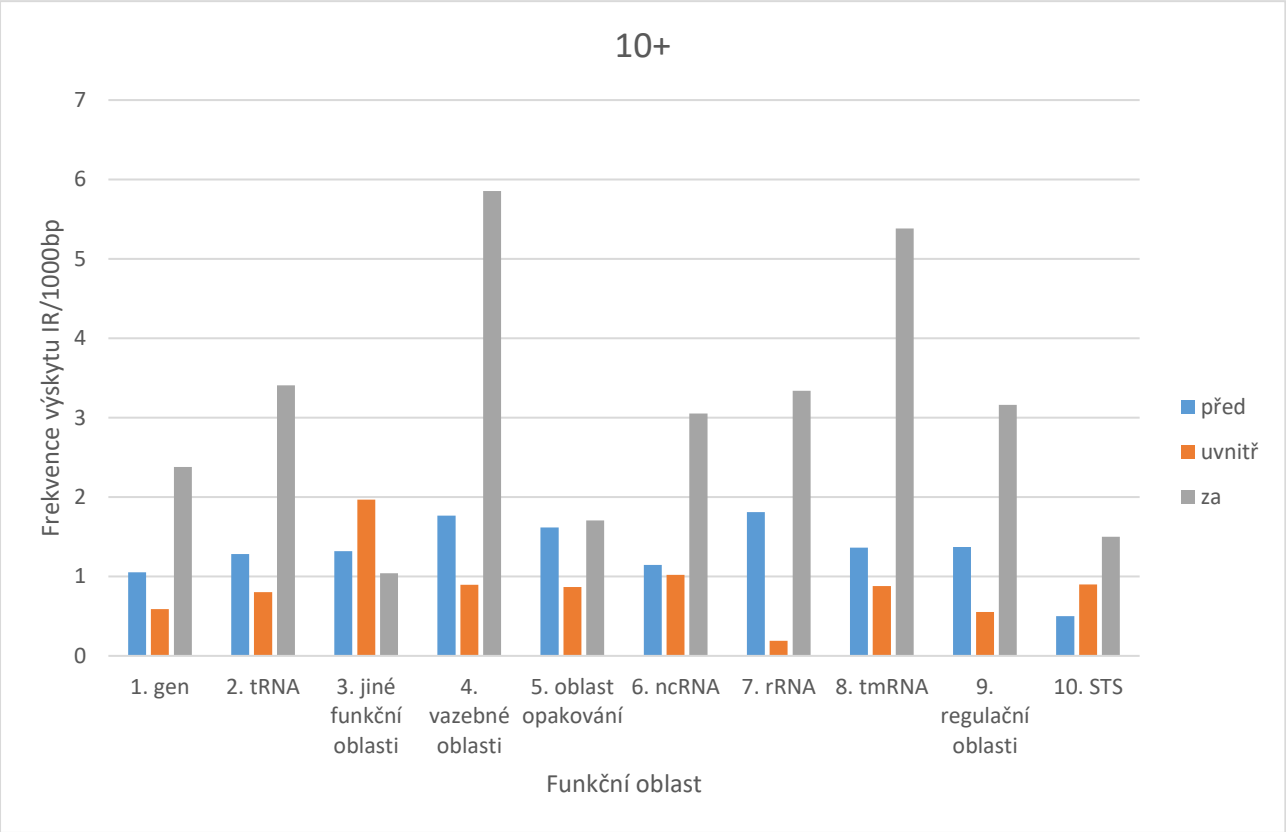
Níže jsou uvedené jednotlivé grafy. Celkem byly sestaveny 4 grafy sledující výskyt inverzních repetic daných funkčních oblastí. V prvním grafu č.1 jsou zahrnuty frekvence výskytu IR všech délek. V dalších grafech č.2, 3 a 4 jsou postupně zahrnuty frekvence výskytu IR délek 8 bp a delší, 10 bp a delší a nakonec 12 bp a delší. V případě grafu č.1 je nejvyšší frekvence IR pozorována „před“ oblastí 4 a „za“ ní, naopak „uvnitř“ je frekvence výskytu IR skoro o polovinu nižší. Podobná je situace u oblastí 6-9 s tím rozdílem, že maximum se nachází až „za“ jednotlivými oblastmi. V případě frekvencí výskytu IR délky 8+ dochází k prohloubení výše uvedené závislosti u oblastí 4, 6-9, v případě oblasti 4 se maximum frekvence IR přesouvá až „za“ oblast. Zajímavá situace nastává u grafů č.3 a 4, kde se maximum frekvence výskytu IR u všech oblastí s výjimkou oblasti č.3 nachází až „za“ nimi a v případě IR délek 12+ dokonce několikanásobně překračuje hodnota frekvence výskytu IR hodnotu výskytu IR „před“ a „uvnitř“ oblastí s výjimkou oblasti 3 a 5. U oblasti 10 nebyly nalezeny žádné IR délky 12+ „před“ ani „po“. Dále byla u všech oblastí pozorována frekvence výskytu IR „uvnitř“ nižší než „vně“ s výjimkou oblastí 10 a 3 (pro IR délky 8+, 10+ a 12+) a také pro oblast 1 v případě grafu č.1.



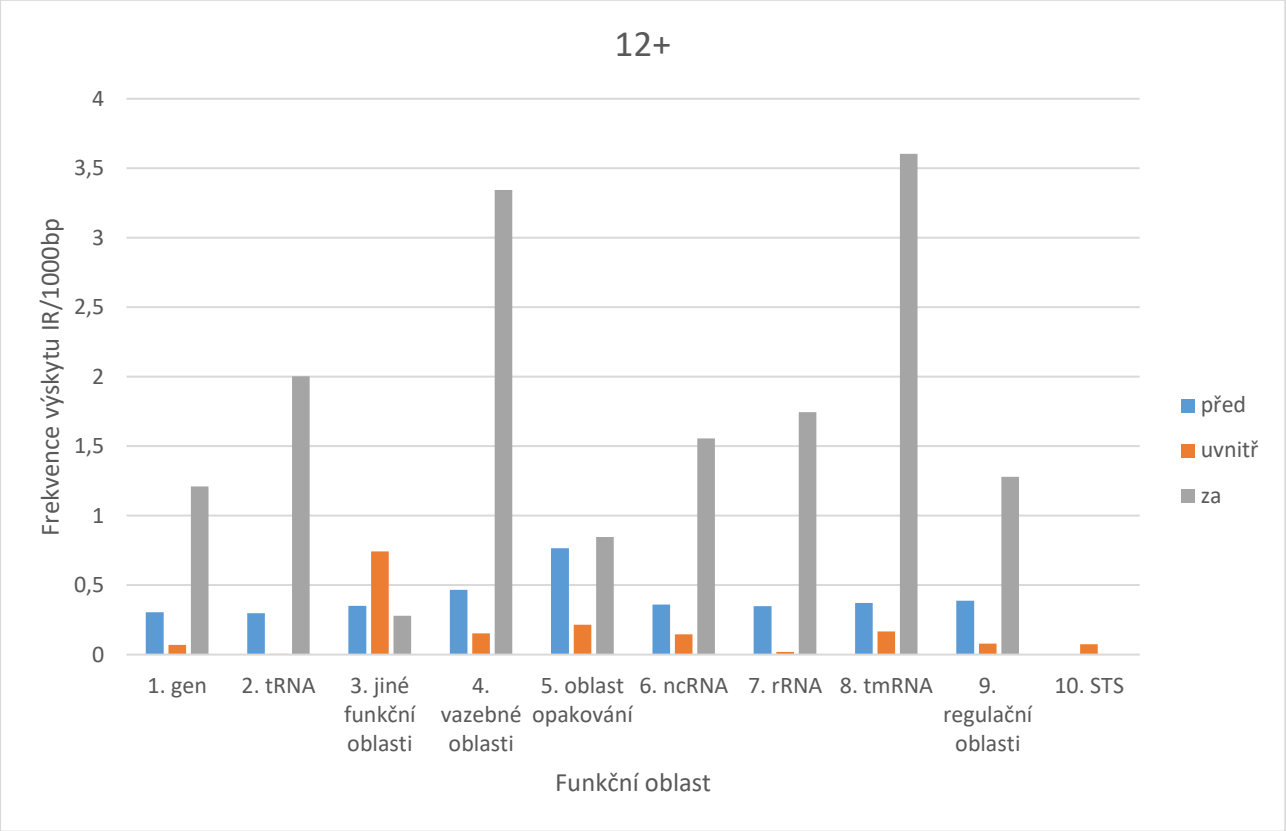
Graf č.1 *Lokalizace frekvence výskytu IR všech délek u jednotlivých funkčních oblastí*



Graf č.2 *Lokalizace frekvence výskytu IR délek 8+ u jednotlivých funkčních oblastí*



Graf č.3 *Lokalizace frekvence výskytu IR délek 10+ u jednotlivých funkčních oblastí*



Graf č.4 *Lokalizace frekvence výskytu IR délek 12+ u jednotlivých funkčních oblastí*

5 DISKUZE

Inverzní repetice se vyskytují nejen v genomu bakterií, ale také v eukaryotické DNA, kde taktéž vykazují nenáhodnou distribuci. Inverzní sekvence bohaté na GC páry, nacházející se v oblasti promotorů, mohou po oxidaci 2'-deoxyguanosinu dát vzniku křížovým strukturám, které následně regulují transkripci [76]. Inverzní repetice se také nacházejí ve virových částicích. Například analýza lidského RNA virusu SARS-CoV-2 zjistila, že bodové mutace se častěji vyskytují uvnitř inverzních repetic [3]. Taktéž dvě konzervované oblasti SARS-CoV-2 a SARS-Co-V zaujímají vlásenkové struktury, které mohou chránit virální RNA před rychlou degradací v lidských buňkách, a tudíž zvyšovat stabilitu virové RNA, zefektivňovat replikaci a virulenci [3].

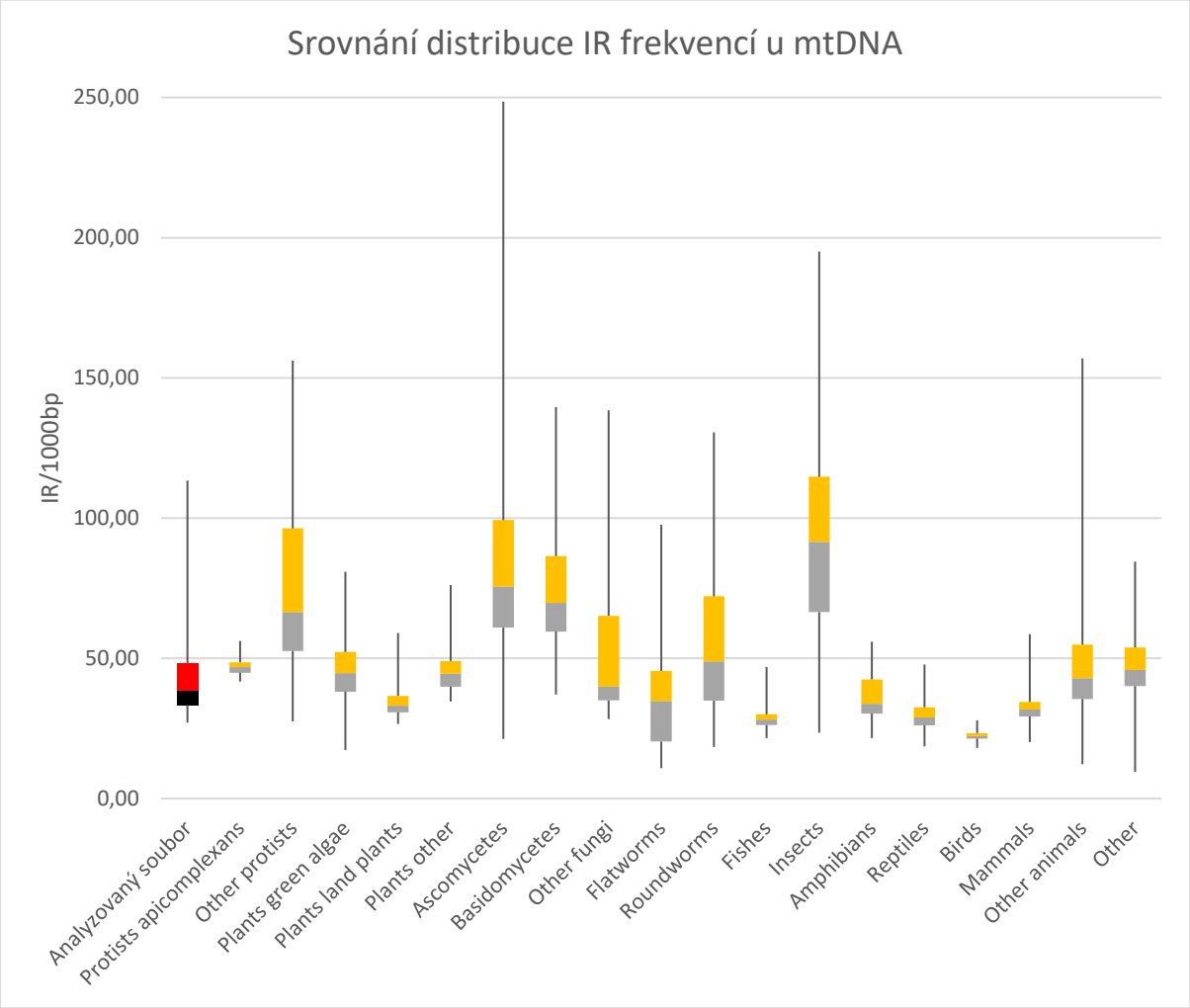
V případě výskytu IR u eukaryotických organismů byl analyzován celý genom *Saccharomyces cerevisiae* (kvasinka pekařská) včetně mitochondriální DNA (mtDNA) [77]. Průměrný výskyt IR u všech chromozomů *S. cerevisiae* byl 0,56 IR/Kbp, což je zhruba 75krát méně, než jaký byl průměrný výskyt IR v analyzovaném souboru bakteriálních genomů [77]. Naproti tomu u mtDNA byl průměrný výskyt IR u *S. cerevisiae* 25,02 IR/Kbp. Tato frekvence výskytu IR se blíží frekvenci výskytu IR u námi analyzovaného souboru bakterií, což podporuje endosymbiotickou teorii vzniku mitochondrií pohlcením bakteriálního genomu dávným předkem eukaryot [78]. Podobně jako v případě analýzy bakteriálních genomů vykazuje analýza výskytu IR vzhledem k funkčním oblastem u S. cerevisiae nenáhodnou distribuci IR s maximem výskytu IR nacházejícím se uvnitř centromerních oblastí [77].

Analýzou mitochondriální DNA různých organismů (celkem 7135 mtDNA sekvencí), byla zjištěna přítomnost inverzních repetic, která vykazovala evoluční konzervaci i jejich nenáhodnou distribuci [79]. Taktéž chloroplastová DNA (cpDNA) vykazuje přítomnost inverzních repetic a jejich nenáhodnou distribuci [80]. Podobně jako v případě frekvence výskytu IR u bakteriálních genomů klesá s rostoucí délkou IR i u mtDNA a cpDNA, s několika výjimkami [79; 80]. Analýza frekvence výskytu IR u cpDNA vzhledem k funkčním oblastem taktéž ukázala jejich nenáhodnou distribuci, tak jako v případě frekvence výskytu IR u analyzovaného souboru bakteriálních genomů [80]. Tabulka 3 ukazuje porovnání statistických ukazatelů analyzovaného souboru bakteriálních genomů s analýzou IR u mitochondriální a chloroplastové DNA.

Tabulka 3 *Srovnání jednotlivých statistických ukazatelů*

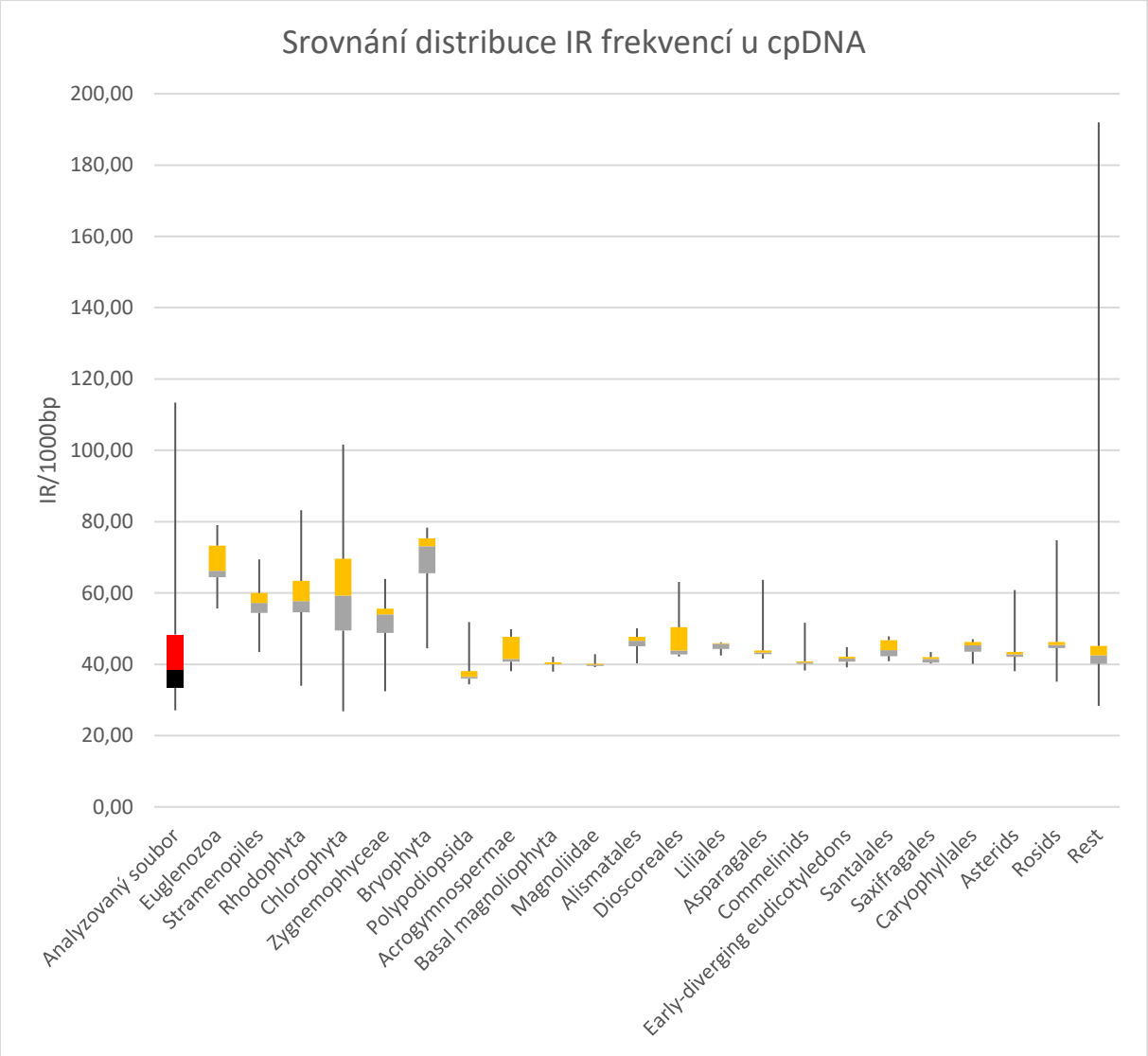
	1. kvartil (IR/Kbp)	3. kvartil (IR/Kbp)	Průměr (IR/Kbp)	Min (IR/Kbp)	Max (IR/Kbp)
Analyzovaný soubor	33,3	48,3	41,9	27,10	113,40
mtDNA	27,0	47,0	41,9	9,47	248,50
cpDNA	40,0	45,0	45,0	26,00	191,98

Polovina analyzované eukaryotické mtDNA vykazovala frekvenci výskytu IR mezi 27 – 47 IR/Kbp [79], u cpDNA bylo rozmezí 40 – 45 IR/Kbp [80] a v případě analyzovaného souboru bakteriálních genomů se 50 % všech hodnot nachází mezi frekvencemi 33 – 48 IR/Kbp. Průměrný výskyt všech inverzních repetic u cpDNA byl 45 IR/Kbp [80] a u mtDNA byl 41,9 IR/Kbp [79], tedy stejný jako v analyzovaném souboru bakteriálních genomů. Ovšem v případě rozpětí hodnot vykazoval náš soubor nejmenší rozpětí, naopak nejvyšší rozpětí hodnot frekvence výskytu IR byl u mtDNA. Dále byly ještě zpracovány dva grafy, které srovnávají frekvence výskytu IR analyzovaného souboru bakteriálních genomů s frekvencí výskytu IR u jednotlivých skupin organismů u mtDNA (graf č. 5) a také u cpDNA (graf č.6). Grafy jsou



Graf č.5 *Srovnání distribuce IR frekvencí u mtDNA (převzato a upraveno z [79])*

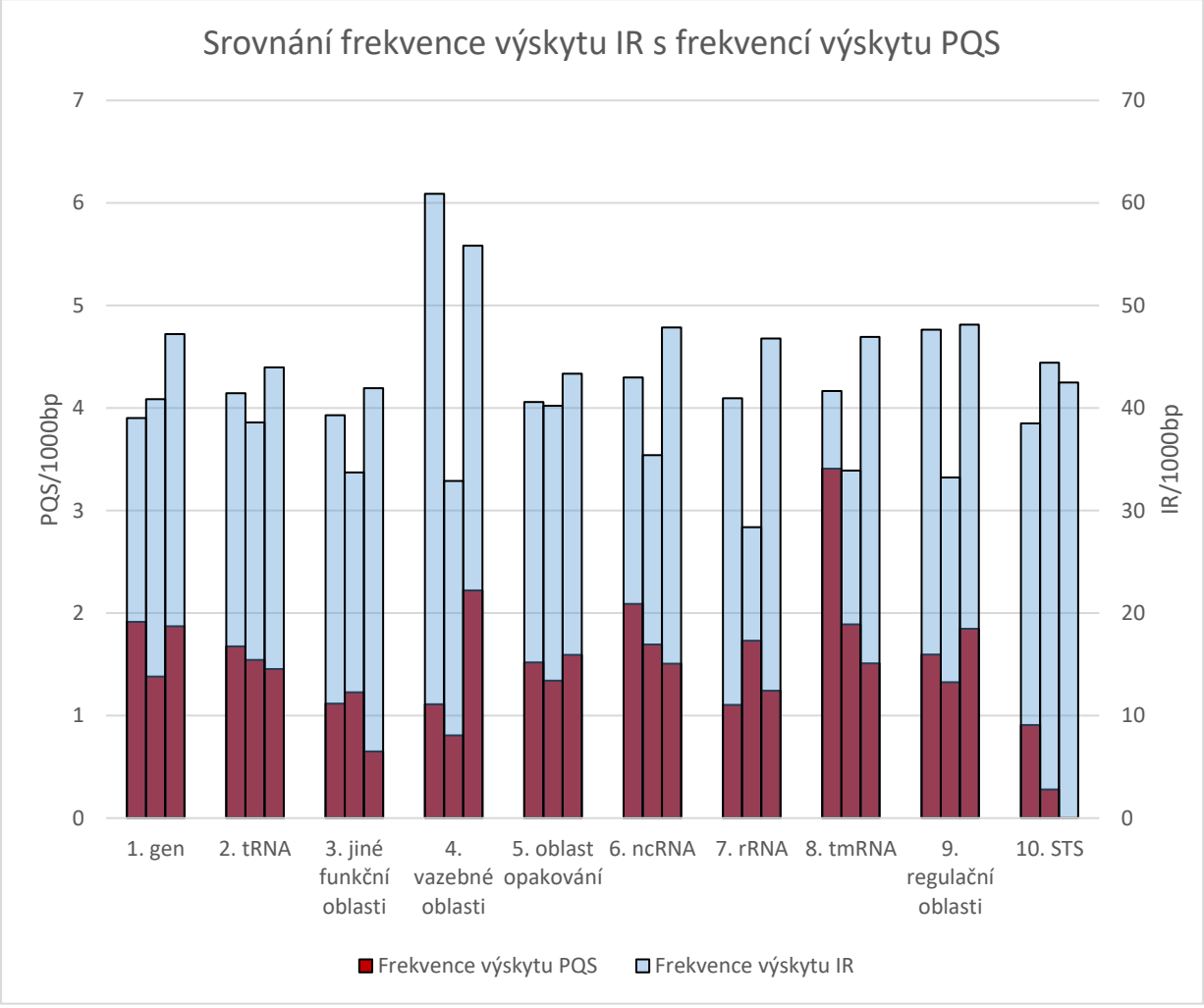
znázorněny jako boxové s vyznačeným mediánem, 3. a 1. kvartilem a s vystupujícími liniemi se střední části udávající maximální a minimální hodnoty.



Graf č.6 *Lokalizace frekvence výskytu IR délek 12+ u jednotlivých funkčních oblastí (převzato a upraveno z [80])*

Dále byla ještě porovnána frekvence výskytu IR analyzovaného souboru bakteriálních genomů s frekvencí výskytu sekvencí pro formaci G-kvadruplexových struktur. Ve srovnání s výskytem sekvencí pro formaci G-quadruplexových struktur (PQS) jsou inverzní repetice zastoupeny poměrně častěji v bakteriálních genomech a jejich závislost na % obsahu GC párů vykazuje odlišný průběh. Obecně platí, že výskyt G-kvadruplexových struktur se zvyšuje s rostoucím obsahem GC % párů, samozřejmě s celou řadou výjimek [81]. Naproti tomu minimum výskytu inverzních repetic nástává okolo 50 % obsahu GC párů a roste symetricky na obě strany a dosahuje své maximální hodnoty v krajních hodnotách % obsahu CG párů. Dále byl zpracován graf č.7 porovnávající frekvenci výskytu PQS sekvencí s IR sekvencemi

vzhledem k daným funkčním oblastem. Graf č. 7 má dvě osy y (osa vlevo znázorňuje frekvenci výskytu PQS, osa vpravo znázorňuje frekvenci výskytu IR) a překrývající se sloupce (hnědá značí frekvenci výskytu PQS a světle modrá frekvence výskytu IR). Průměrná frekvence výskytu IR „před“, „uvnitř“ a „po“ je zhruba 29násobná oproti frekvenci výskytu PQS. V případě rozložení výskytu PQS vzhledem k funkčním oblastem, se maximum a minimum frekvence výskytu PQS nachází „před“ oblastí 8 a „uvnitř“ oblasti 10 a maximum a minimum frekvence výskytu IR „před“ oblastí 4 a „uvnitř“ oblasti 7, viz. graf č. 7. Oblasti 5 a 9 vykazují podobnosti v rozložení frekvence výskytu IR a PQS. Oblasti 3 jsou navzájem invertované. Ostatní oblasti se odlišují v relativním výskytu frekvencí IR a PQS.



Graf č.7 *Srovnání frekvence výskytu IR s frekvencí výskytu PQS*

6 ZÁVĚR

Inverzní sekvence jsou nedílnou součástí DNA, jejich distribuce vykazuje nenahodilý charakter. Mohou zapříčiňovat genomovou nestabilitu, vznik různých inzercí, delecí nebo translokací. Také jsou součástí transpozomů a oblastí rozeznávaných restrikčními endonukleázami. Nedokonalé inverzní repetice mohou být příčinou vzniku mutací a z nich následných onemocnění. Naproti své negativní roli při vzniku genomové instability a mutací, mají taktéž pozitivní roli, protože umožňují vzniku různým sekundárním strukturám, které jsou rozeznávány řadou proteinů. Ze vznikajících sekundárních struktur byla věnována pozornost křížovým strukturám a proteinům, které s těmito struktury interagují. Křížové struktury jsou důležité pro celou řadu biologických procesů včetně transkripce, replikace, rekombinace a regulace genové exprese a organizace genomu. Dále byla představena aplikace *Palindrome analyser* pro hledání inverzních repetic a popsáno využití její online a offline verze při analýze.

Experimentální část byla zaměřena na analýzu frekvence výskytu inverzních repetic v bakteriálních genomech. Za tímto účelem byl analyzován soubor 1547 druhů bakterií, roztříděných do několika skupin a podskupin. Celkem bylo analyzováno 1627 bakteriálních genomů pomocí programu *Palindrome analyser*. Nejvyšší počet inverzních repetic byl nalezen u podskupiny Tenericutes, medián 63,17 IR/1000bp. V případě jednotlivých bakterií byla nejvyšší frekvence výskytu IR zjištěna u bakterie Buchnera aphidicola s frekvencí výskytu 113,37 IR/1000bp a nejnižší frekvence 27,8 IR/1000bp u Anaplasma centrale. Dále byla zkoumána závislost frekvence výskytu IR na obsahu GC párů. Bylo zjištěno, že v případě inverzních repetic všech délek (6-30 bp) vykazuje frekvence výskytu IR přibližně kvadratickou závislost s minimem okolo 50 % obsahu GC párů, která slábne s rostoucí délkou inverzních repetic.

Dále byla zkoumána lokalizace inverzních repetic vzhledem k funkčním oblastem. Celkem bylo zkoumáno 10 různých typu funkčních oblastí u kterých byl sledován výskyt inverzních repetic ve vzdálenosti do 100 bp před, uvnitř a do 100 bp za funkční oblastí, viz. kapitola 4.2. Byla analyzována lokalizace vůči uvedeným funkčním oblastem u celkem 4 druhů délek IR – VŠE (6-30), 8+ (8 a delší), 10+ (10 a delší), 12+ (12 a delší). Zjištěné výsledky byly graficky zpracovány do 4 grafů (č.1 až 4) a zhodnoceny. Ve většině skupin je frekvence výskytu IR „uvnitř“ nižší než „vně“. Tento poměr se zvyšuje s rostoucí délkou inverzních repetic a zároveň dochází ke zvyšování poměru frekvence výskytu IR za funkčními oblastmi oproti výskytu před a uvnitř oblastí. V případě IR délky 12+ je poměr frekvencí výskytu IR „za“ oblastmi dokonce několikanásobně větší než „před“ a „uvnitř“.

V diskuzi byla porovnána frekvence výskytu IR s frekvencí výskytu IR u ostatních organismů včetně mitochondriální a chloroplastové DNA. Výsledky porovnání jsou znázorněny v grafech č.5 až 7. Zajímavé je zjištění, že průměrný výskyt IR u všech chromozomů *S. cerevisiae* (kvasinka pivní) činil 0,56 IR/Kbp, což je zhruba 75krát méně, než jaký byl průměrný výskyt IR v analyzovaném souboru bakteriálních genomů.

7 ZDROJE

[1] BRÁZDA, Václav, Martin BARTAS, Jiří LÝSEK, Jan COUFAL a Miroslav FOJTA. Global analysis of inverted repeat sequences in human gene promoters reveals their non-random distribution and association with specific biological pathways. *Genomics* [online]. 2020, **112**(4), 2772-2777 [cit. 2021-02-26]. ISSN 08887543. Dostupné z: doi:10.1016/j.ygeno.2020.03.014

[2] MANDKE, Pooja, Pallavi KOMPELLA, Steve LU, Guliang WANG a Karen VASQUEZ. *Cruciform DNA structure formed at short inverted repeats: A source of genetic instability in vivo: A source of genetic instability in vivo*. 2019.

[3] GOSWAMI, Pratik, Martin BARTAS, Matej LEXA et al. SARS-CoV-2 hot-spot mutations are significantly enriched within inverted repeats and CpG island loci. *Briefings in Bioinformatics* [online]. 2020 [cit. 2021-03-02]. ISSN 1467-5463. Dostupné z: doi:10.1093/bib/bbaa385

[4] WATSON, J. D. a F. H. C. CRICK. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* [online]. 1953, **171**(4356), 737-738 [cit. 2021-06-23]. ISSN 0028-0836. Dostupné z: doi:10.1038/171737a0

[5] SHARMA, Sudha. Non-B DNA Secondary Structures and Their Resolution by RecQ Helicases. *Journal of Nucleic Acids* [online]. 2011, **2011**, 1-15 [cit. 2021-06-23]. ISSN 2090-021X. Dostupné z: doi:10.4061/2011/724215

[6] ALLERS, Thorsten a David R.F. LEACH. DNA Palindromes Adopt a Methylation-resistant Conformation that is Consistent with DNA Cruciform or Hairpin Formationin Vivo. *Journal of Molecular Biology* [online]. 1995, **252**(1), 70-85 [cit. 2021-06-23]. ISSN 00222836. Dostupné z: doi:10.1006/jmbi.1994.0476

[7] BRÁZDA, Václav, Rob LAISTER, Eva JAGELSKÁ a Cheryl ARROWSMITH. Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Molecular Biology* [online]. 2011, **12**(1) [cit. 2021-02-23]. ISSN 1471-2199. Dostupné z: doi:10.1186/1471-2199-12-33

[8] POGGI, Lucie a Guy-Franck RICHARD. Alternative DNA Structures In Vivo: Molecular Evidence and Remaining Questions. *Microbiology and Molecular Biology Reviews* [online]. 2021, **85**(1) [cit. 2021-06-23]. ISSN 1092-2172. Dostupné z: doi:10.1128/MMBR.00110-20

[9] COMPTON, Sarah A., Gökhan TOLUN, Ashwini S. KAMATH-LOEB, Lawrence A. LOEB a Jack D. GRIFFITH. The Werner Syndrome Protein Binds Replication Fork and Holliday Junction DNAs as an Oligomer. *Journal of Biological Chemistry* [online]. 2008, **283**(36), 24478-24483 [cit. 2021-06-23]. ISSN 00219258. Dostupné z: doi:10.1074/jbc.M803370200

[10] ČECHOVÁ, Jana, Jan COUFAL, Eva JAGELSKÁ, Miroslav FOJTA, Václav BRÁZDA a Sumitra DEB. P73, like its p53 homolog, shows preference for inverted repeats forming cruciforms. *PLOS ONE* [online]. 2018, **13**(4) [cit. 2021-03-02]. ISSN 1932-6203. Dostupné z: doi:10.1371/journal.pone.0195835

[11] GANAPATHIRAJU, Madhavi K., Sandeep SUBRAMANIAN, Srilakshmi CHAPARALA a Kalyani B. KARUNAKARAN. A reference catalog of DNA palindromes in the human genome and their variations in 1000 Genomes. *Human Genome Variation* [online]. 2020, **7**(1) [cit. 2021-03-03]. ISSN 2054-345X. Dostupné z: doi:10.1038/s41439-020-00127-5

[12] USSERY, David W., Trudy M. WASSENAAR a Stefano BORINI. *Computing for comparative microbial genomics: bioinformatics for microbiologists*. Londýn: Springer, 2009. ISBN 978-1-84800-254-8.

[13] ŠEDA, O., F. LIŠKA a L. ŠEDOVÁ. *Aktuální genetika - Multimediální učebnice lékařské biologie, genetiky a genomiky* [online]. Ústav biologie a lékařské genetiky 1. LF. Praha, 2006 [cit. 2021-03-09]. Dostupné z: http://biol.lf1.cuni.cz/ucebnice

[14] LU, Steve, Guliang WANG, Albino BACOLLA, Junhua ZHAO, Scott SPITSER a Karen M. VASQUEZ. Short Inverted Repeats Are Hotspots for Genetic Instability: Relevance to Cancer Genomes. *Cell Reports* [online]. 2015, **10**(10), 1674-1680 [cit. 2021-03-02]. ISSN 22111247. Dostupné z: doi:10.1016/j.celrep.2015.02.039

[15] LAVI, Bar, Eli LEVY KARIN, Tal PUPKO, Einat HAZKANI-COVO a Ruth HERSHBERG. The Prevalence and Evolutionary Conservation of Inverted Repeats in Proteobacteria. *Genome Biology and Evolution* [online]. 2018, **10**(3), 918-927 [cit. 2021-03-14]. ISSN 1759-6653. Dostupné z: doi:10.1093/gbe/evy044

[16] BISSLER, John J. DNA inverted repeats and human disease. *Frontiers in Bioscience* [online]. 1998, **3**(4), 408-418 [cit. 2021-03-18]. ISSN 10939946. Dostupné z: doi:10.2741/A284

[17] BATEMAN, JF, SR LAMANDE, HH DAHL, T MASCARA a WG COLE. A frameshift mutation results in a truncated nonfunctional carboxyl-terminal pro alpha 1(I) propeptide of type I collagen in osteogenesis imperfecta. *J Biol Chem.* 1989, **1989**(264), 10960-10964. PMID: 2500431.

[18] FRAPPIER, L., G.B. PRICE, R.G. MARTIN a M. ZANNIS-HADJOPOULOS. Monoclonal antibodies to cruciform DNA structures. *Journal of Molecular Biology* [online]. 1987, **193**(4), 751-758 [cit. 2021-04-23]. ISSN 00222836. Dostupné z: doi:10.1016/0022-2836(87)90356-1

[19] ZANNIS-HADJOPOULOS, M., L. FRAPPIER, M. KHOURY a G. B. PRICE. Effect of anti-cruciform DNA monoclonal antibodies on DNA replication. *The EMBO Journal* [online]. 1988, **7**(6), 1837-1844 [cit. 2021-04-23]. ISSN 02614189. Dostupné z: doi:10.1002/j.1460-2075.1988.tb03016.x

[20] CALLEJO, Mario, David ALVAREZ, Gerald B. PRICE a Maria ZANNIS-HADJOPOULOS. The 14-3-3 Protein Homologues from Saccharomyces cerevisiae, Bmh1p and Bmh2p, Have Cruciform DNA-binding Activity and Associate in Vivo with ARS307. *Journal of Biological Chemistry* [online]. 2002, **277**(41), 38416-38423 [cit. 2021-04-29]. ISSN 00219258. Dostupné z: doi:10.1074/jbc.M202050200

[21] PEARSON, Christopher E., Haralabos ZORBAS, Gerald B. PRICE a Maria ZANNIS-HADJOPOULOS. Inverted repeats, stem-loops, and cruciforms: Significance for initiation of DNA replication. *Journal of Cellular Biochemistry* [online]. 1996, **63**(1), 1-22 [cit. 2021-04-03]. ISSN 07302312. Dostupné z: doi:10.1002/(SICI)1097-4644(199610)63:11::AID-JCB13.0.CO;2-3

[22] SHLYAKHTENKO, Luda S., Peggy HSIEH, Michael GRIGORIEV, Vladimir N. POTAMAN, Richard R. SINDEN a Yuri L. LYUBCHENKO. A cruciform structural transition provides a molecular switch for chromosome structure and dynamics 1 1Edited by I. Tinoco. *Journal of Molecular Biology* [online]. 2000, **296**(5), 1169-1173 [cit. 2021-07-05]. ISSN 00222836. Dostupné z: doi:10.1006/jmbi.2000.3542

[23] ZANNIS-HADJOPOULOS, Maria, Wafaa YAHYAOUI a Mario CALLEJO. 14-3-3 Cruciform-binding proteins as regulators of eukaryotic DNA replication. *Trends in Biochemical Sciences* [online]. 2008, **33**(1), 44-50 [cit. 2021-07-05]. ISSN 09680004. Dostupné z: doi:10.1016/j.tibs.2007.09.012

[24] CHASOVSKIKH, Sergey, Alexandre DIMTCHEV, Mark SMULSON a Anatoly DRITSCHILO. DNA transitions induced by binding of PARP-1 to cruciform structures in supercoiled plasmids. *Cytometry Part A* [online]. 2005, **68**(1), 21-27 [cit. 2021-05-07]. ISSN 1552-4922. Dostupné z: doi:10.1002/cyto.a.20187

[25] WADKINS, Randy. Targeting DNA Secondary Structures. *Current Medicinal Chemistry* [online]. 2000, **7**(1), 1-15 [cit. 2021-07-05]. ISSN 09298673. Dostupné z: doi:10.2174/0929867003375461

[26] INAGAKI, Hidehito, Tamae OHYE, Hiroshi KOGO, Makiko TSUTSUMI, Takema KATO, Maoqing TONG, Beverly S. EMANUEL a Hiroki KURAHASHI. Two sequential cleavage reactions on cruciform DNA structures cause palindrome-mediated chromosomal translocations. *Nature Communications* [online]. 2013, **4**(1) [cit. 2021-07-05]. ISSN 2041-1723. Dostupné z: doi:10.1038/ncomms2595

[27] WALDMANN, T. Structure-specific binding of the proto-oncogene protein DEK to DNA. *Nucleic Acids Research* [online]. 2003, **31**(23), 7003-7010 [cit. 2021-07-05]. ISSN 1362-4962. Dostupné z: doi:10.1093/nar/gkg864

[28] LILLEY, David M. J. Holliday junction-resolving enzymes-structures and mechanisms. *FEBS Letters* [online]. 2017, **591**(8), 1073-1082 [cit. 2021-05-04]. ISSN 00145793. Dostupné z: doi:10.1002/1873-3468.12529

[29] ARAVIND, L. SURVEY AND SUMMARY: Holliday junction resolvases and related nucleases. *Nucleic Acids Research* [online]. **28**(18), 3417-3432 [cit. 2021-07-05]. ISSN 13624962. Dostupné z: doi:10.1093/nar/28.18.3417

[30] ARAVIND, L. SURVEY AND SUMMARY: Holliday junction resolvases and related nucleases. *Nucleic Acids Research* [online]. 2000, **28**(18), 3417-3432 [cit. 2021-07-05]. ISSN 13624962. Dostupné z: doi:10.1093/nar/28.18.3417

[31] MANDKE, Pooja a Karen M. VASQUEZ. Interactions of high mobility group box protein 1 (HMGB1) with nucleic acids: Implications in DNA repair and immune responses. *DNA Repair* [online]. 2019, **83** [cit. 2021-05-17]. ISSN 15687864. Dostupné z: doi:10.1016/j.dnarep.2019.102701

[32] MAZIN, Alexander V., Olga M. MAZINA, Dmitry V. BUGREEV a Matthew J. ROSSI. Rad54, the motor of homologous recombination. *DNA Repair* [online]. 2010, **9**(3), 286-302 [cit. 2021-07-06]. ISSN 15687864. Dostupné z: doi:10.1016/j.dnarep.2009.12.006

[33] BRÁZDA, Václav, Lucia HÁRONÍKOVÁ, Jack C. C. LIAO, Helena FRIDRICHOVÁ a Eva B. JAGELSKÁ. Strong preference of BRCA1 protein to topologically constrained non-B DNA structures. *BMC Molecular Biology* [online]. 2016, **17**(1) [cit. 2021-07-06]. ISSN 1471-2199. Dostupné z: doi:10.1186/s12867-016-0068-6

[34] LAHIRI, Sudipta, Manju HINGORANI a Ishita MUKERJI. 116 Binding dynamics of yeast MutS homologs Msh4-Msh5 with the Holliday junction. *Journal of Biomolecular Structure and Dynamics* [online]. 2015, **33**(1), 73-73 [cit. 2021-07-06]. ISSN 0739-1102. Dostupné z: doi:10.1080/07391102.2015.1032749

[35] LOIZZI, Vera, Girolamo RANIERI, Mariarita LAFORGIA et al. PARP inhibitors and epithelial ovarian cancer: Molecular mechanisms, clinical development and future prospective (Review). *Oncology Letters* [online]. 2020, **20**(4), 1-1 [cit. 2021-05-07]. ISSN 1792-1074. Dostupné z: doi:10.3892/ol.2020.11951

[36] MARTINEZ-ZAMUDIO, Ricardo a Hyo Chol HA. Histone ADP-Ribosylation Facilitates Gene Transcription by Directly Remodeling Nucleosomes. *Molecular and Cellular Biology* [online]. 2012, **32**(13), 2490-2502 [cit. 2021-07-06]. ISSN 0270-7306. Dostupné z: doi:10.1128/MCB.06667-11

[37] HOLLSTEIN, M, D SIDRANSKY, B VOGELSTEIN a C. HARRIS. P53 mutations in human cancers. *Science* [online]. 1991, **253**(5015), 49-53 [cit. 2021-07-07]. ISSN 0036-8075. Dostupné z: doi:10.1126/science.1905840

[38] ZHU, Yuan, Frantz GUIGNARD, Dawen ZHAO, Li LIU, Dennis K. BURNS, Ralph P. MASON, Albee MESSING a Luis F. PARADA. Early inactivation of p53 tumor suppressor gene cooperating with NF1 loss induces malignant astrocytoma. *Cancer Cell* [online]. 2005, **8**(2), 119-130 [cit. 2021-07-07]. ISSN 15356108. Dostupné z: doi:10.1016/j.ccr.2005.07.004

[39] BRÁZDA, Václav a Miroslav FOJTA. The Rich World of p53 DNA Binding Targets: The Role of DNA Structure. *International Journal of Molecular Sciences* [online]. 2019, **20**(22) [cit. 2021-07-07]. ISSN 1422-0067. Dostupné z: doi:10.3390/ijms20225605

[40] BRAZDA, Vaclav, Petr MULLER, Kristyna BROZKOVA a Borivoj VOJTESEK. Restoring wild-type conformation and DNA-binding activity of mutant p53 is insufficient for restoration of transcriptional activity. *Biochemical and Biophysical Research Communications* [online]. 2006, **351**(2), 499-506 [cit. 2021-07-07]. ISSN 0006291X. Dostupné z: doi:10.1016/j.bbrc.2006.10.065

[41] BRÁZDA, Václav, Eva Brázdová JAGELSKÁ, Miroslav FOJTA a Emil PALEČEK. Searching for target sequences by p53 protein is influenced by DNA length. *Biochemical and Biophysical Research Communications* [online]. 2006, **341**(2), 470-477 [cit. 2021-07-07]. ISSN 0006291X. Dostupné z: doi:10.1016/j.bbrc.2005.12.202

[42] COUFAL, Jan, Eva B. JAGELSKÁ, Jack C.C. LIAO a Václav BRÁZDA. Preferential binding of p53 tumor suppressor to p21 promoter sites that contain inverted repeats capable of forming cruciform structure. *Biochemical and Biophysical Research Communications* [online]. 2013, **441**(1), 83-88 [cit. 2021-07-07]. ISSN 0006291X. Dostupné z: doi:10.1016/j.bbrc.2013.10.015

[43] PALEČEK, Emil, Daniel VLK, Veronika STAŇKOVÁ, Václav BRÁZDA, Bořivoj VOJTĚŠEK, Tedd R HUPP, Achim SCHAPER a Thomas M JOVIN. Tumor suppressor protein p53 binds preferentially to supercoiled DNA. *Oncogene* [online]. 1997, **15**(18), 2201-2209 [cit. 2021-07-07]. ISSN 0950-9232. Dostupné z: doi:10.1038/sj.onc.1201398

[44] BRÁZDA, Václav, Jan PALEĉEK, Šárka POSPÍŠILOVÁ, Borp̂ivoj VOJTêŠEK a Emil PALEĉEK. Specific Modulation of p53 Binding to Consensus Sequence within Supercoiled DNA by Monoclonal Antibodies. *Biochemical and Biophysical Research Communications* [online]. 2000, **267**(3), 934-939 [cit. 2021-07-07]. ISSN 0006291X. Dostupné z: doi:10.1006/bbrc.1999.2056

[45] GHILAROV, D. A. a I. S. SHKUNDINA. DNA topoisomerases and their functions in a cell. *Molecular Biology*. 2012, **46**(1), 47-57. ISSN 0026-8933. Dostupné z: doi:10.1134/S0026893312010074

[46] HEDE, Marianne S., Rikke L. PETERSEN, Rikke F. FRøHLICH, Dinna KRÜGER, Felicie F. ANDERSEN, Anni H. ANDERSEN a Birgitta R. KNUDSEN. Resolution of Holliday Junction Substrates

by Human Topoisomerase I. *Journal of Molecular Biology* [online]. 2007, **365**(4), 1076-1092 [cit. 2021-07-06]. ISSN 00222836. Dostupné z: doi:10.1016/j.jmb.2006.10.050

[47] RENÉ, Brigitte, Serge FERMANDJIAN a Olivier MAUFFRET. Does topoisomerase II specifically recognize and cleave hairpins, cruciforms and crossovers of DNA?. *Biochimie* [online]. 2007, **89**(4), 508-515 [cit. 2021-07-06]. ISSN 03009084. Dostupné z: doi:10.1016/j.biochi.2007.02.011

[48] STROS, M., A. BACIKOVA, E. POLANSKA, J. STOKROVA a F. STRAUSS. HMGB1 interacts with human topoisomerase II and stimulates its catalytic activity. *Nucleic Acids Research* [online]. 2007, **35**(15), 5001-5013 [cit. 2021-07-06]. ISSN 0305-1048. Dostupné z: doi:10.1093/nar/gkm525

[49] MAZINA, Olga M., Matthew J. ROSSI, Nicolas H. THOMAAÖ a Alexander V. MAZIN. Interactions of Human Rad54 Protein with Branched DNA Molecules*. *Journal of Biological Chemistry* [online]. 2007, **282**(29), 21068-21080 [cit. 2021-07-06]. ISSN 00219258. Dostupné z: doi:10.1074/jbc.M701992200

[50] B HM, F. The SAF-box domain of chromatin protein DEK. *Nucleic Acids Research* [online]. 2005, **33**(3), 1101-1110 [cit. 2021-05-08]. ISSN 1362-4962. Dostupné z: doi:10.1093/nar/gki258

[51] KAPPES, Ferdinand, Catalina DAMOC, Rolf KNIPPERS, Michael PRZYBYLSKI, Lorenzo A. PINNA a Claudia GRUSS. Phosphorylation by Protein Kinase CK2 Changes the DNA Binding Properties of the Human Chromatin Protein DEK. *Molecular and Cellular Biology* [online]. 2004, **24**(13), 6011-6020 [cit. 2021-07-06]. ISSN 0270-7306. Dostupné z: doi:10.1128/MCB.24.13.6011-6020.2004

[52] WALDMANN, Tanja, Ingo SCHOLTEN, Ferdinand KAPPES, Hong Gang HU a Rolf KNIPPERS. The DEK protein—an abundant and ubiquitous constituent of mammalian chromatin. *Gene* [online]. 2004, **343**(1), 1-9 [cit. 2021-07-06]. ISSN 03781119. Dostupné z: doi:10.1016/j.gene.2004.08.029

[53] ROSEN, Eliot M., Saijun FAN, Richard G. PESTELL a Itzhak D. GOLDBERG. BRCA1 gene in breast cancer. *Journal of Cellular Physiology* [online]. 2003, **196**(1), 19-41 [cit. 2021-05-09]. ISSN 0021-9541. Dostupné z: doi:10.1002/jcp.10257

[54] CLARK, Serena L., Ana M. RODRIGUEZ, Russell R. SNYDER, Gary D.V. HANKINS a Darren BOEHNING. STRUCTURE-FUNCTION OF THE TUMOR SUPPRESSOR BRCA1. *Computational and Structural Biotechnology Journal* [online]. 2012, **1**(1) [cit. 2021-05-14]. ISSN 20010370. Dostupné z: doi:10.5936/csbj.201204005

[55] BRÁZDA, Václav, Eva B. JAGELSKÁ, Jack C.C. LIAO a Cheryl H. ARROWSMITH. The Central Region of BRCA1 Binds Preferentially to Supercoiled DNA. *Journal of Biomolecular Structure and Dynamics* [online]. 2009, **27**(1), 97-103 [cit. 2021-05-17]. ISSN 0739-1102. Dostupné z: doi:10.1080/07391102.2009.10507299

[56] NASEEM, Riffat, Alice STURDY, David FINCH, Thomas JOWITT a Michelle WEBB. Mapping and conformational characterization of the DNA-binding region of the breast cancer susceptibility protein BRCA1. *Biochemical Journal* [online]. 2006, **395**(3), 529-535 [cit. 2021-07-06]. ISSN 0264-6021. Dostupné z: doi:10.1042/BJ20051646

[57] POHLER, J.R. G. HMG box proteins bind to four-way DNA junctions in their open conformation. *The EMBO Journal* [online]. **17**(3), 817-826 [cit. 2021-07-07]. ISSN 14602075. Dostupné z: doi:10.1093/emboj/17.3.817

[58] MANDKE, Pooja a Karen M. VASQUEZ. Interactions of high mobility group box protein 1 (HMGB1) with nucleic acids: Implications in DNA repair and immune responses. *DNA Repair* [online]. 2019, **83** [cit. 2021-05-17]. ISSN 15687864. Dostupné z: doi:10.1016/j.dnarep.2019.102701

[59] ASSENBERG, René, Michelle WEBB, Edward CONNOLLY, Katherine STOTT, Matthew WATSON, Josie HOBBS a Jean O. THOMAS. A critical role in structure-specific DNA binding for the acetylatable lysine residues in HMGB1. *Biochemical Journal* [online]. 2008, **411**(3), 553-561 [cit. 2021-05-17]. ISSN 0264-6021. Dostupné z: doi:10.1042/BJ20071613

[60] BONNEFOY, Eliette. The Ribosomal S16 Protein of Escherichia Coli Displaying a DNA-Nicking Activity Binds to Cruciform DNA. *European Journal of Biochemistry* [online]. 1997, **247**(3), 852-859 [cit. 2021-07-06]. ISSN 0014-2956. Dostupné z: doi:10.1111/j.1432-1033.1997.t01-1-00852.x

[61] AKHMEDOV, Alexandre T., Christian FREI, Monika TSAI-PFLUGFELDER, Börries KEMPER, Susan M. GASSER a Rolf JESSBERGER. Structural Maintenance of Chromosomes Protein C-terminal Domains Bind Preferentially to DNA with Secondary Structure. *Journal of Biological Chemistry* [online]. 1998, **273**(37), 24088-24094 [cit. 2021-07-06]. ISSN 00219258. Dostupné z: doi:10.1074/jbc.273.37.24088

[62] LOSADA, A. Dynamic molecular linkers of the genome: the first decade of SMC proteins. *Genes & Development* [online]. 2005, **19**(11), 1269-1287 [cit. 2021-05-20]. ISSN 0890-9369. Dostupné z: doi:10.1101/gad.1320505

[63] VANARSDALL, Adam L., Kazuhiro OKANO a George F. ROHRMANN. Characterization of the Role of Very Late Expression Factor 1 in Baculovirus Capsid Structure and DNA Processing. *Journal of Virology* [online]. 2006, **80**(4), 1724-1733 [cit. 2021-05-20]. ISSN 0022-538X. Dostupné z: doi:10.1128/JVI.80.4.1724-1733.2006

[64] HONDERMARCK, Hubert. 14-3-3 Proteins. *Handbook of Cell Signaling* [online]. Elsevier, 2010, s. 1367-1374 [cit. 2021-05-22]. ISBN 9780123741455. Dostupné z: doi:10.1016/B978-0-12-374145-5.00169-8

[65] FU, Haian, Romesh R. SUBRAMANIAN a Shane C. MASTERS. 14-3-3 Proteins: Structure, Function, and Regulation. *Annual Review of Pharmacology and Toxicology* [online]. 2000, **40**(1), 617-647 [cit. 2021-07-06]. ISSN 0362-1642. Dostupné z: doi:10.1146/annurev.pharmtox.40.1.617

[66] TODD, Andrea, Nandini COSSONS, Alastair AITKEN, Gerald B. PRICE a Maria ZANNIS-HADJOPOULOS. Human Cruciform Binding Protein Belongs to the 14-3-3 Family †. *Biochemistry* [online]. 1998, **37**(40), 14317-14325 [cit. 2021-07-07]. ISSN 0006-2960. Dostupné z: doi:10.1021/bi980768k

[67] OZGENC, Ali a Lawrence A. LOEB. Current advances in unraveling the function of the Werner syndrome protein. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* [online]. 2005, **577**(1-2), 237-251 [cit. 2021-07-07]. ISSN 00275107. Dostupné z: doi:10.1016/j.mrfmmm.2005.03.020

[68] WINTERS, Amanda C. a Kathrin M. BERNT. MLL-Rearranged Leukemias—An Update on Science and Clinical Approaches. *Frontiers in Pediatrics* [online]. 2017, **5** [cit. 2021-05-23]. ISSN 2296-2360. Dostupné z: doi:10.3389/fped.2017.00004

[69] COSGROVE, Michael S. a Anamika PATEL. Mixed lineage leukemia: a structure-function perspective of the MLL1 protein. *FEBS Journal* [online]. 2010, **277**(8), 1832-1842 [cit. 2021-05-23]. ISSN 1742464X. Dostupné z: doi:10.1111/j.1742-4658.2010.07609.x

[70] ZELEZNIK-LE, Nancy J., Alanna M. HARDEN a Janet D. ROWLEY. 11q23 Translocations Split the "AT-Hook" Cruciform DNA-Binding Region and the Transcriptional Repression Domain from the Activation Domain of the Mixed-Lineage Leukemia (MLL) Gene. *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 1994, **91**(22), 10610. ISSN 00278424. Dostupné také z: http://www.jstor.org/stable/2366079

[71] GREENE, Allie N., Lois G. PARKS, Matia B. SOLOMON a Lisa M. PRIVETTE VINNEDGE. Loss of DEK Expression Induces Alzheimer's Disease Phenotypes in Differentiated SH-SY5Y Cells. *Frontiers in Molecular Neuroscience* [online]. 2020, **13** [cit. 2021-05-25]. ISSN 1662-5099. Dostupné z: doi:10.3389/fnmol.2020.594319

[72] ZHOU, MIN-HANG a QING-MING YANG. NUP214 fusion genes in acute leukemia (Review). *Oncology Letters* [online]. 2014, **8**(3), 959-962 [cit. 2021-05-25]. ISSN 1792-1074. Dostupné z: doi:10.3892/ol.2014.2263

[73] MOR-VAKNIN, Nirit, Antonello PUNTURIERI, Kajal SITWALA et al. The DEK Nuclear Autoantigen Is a Secreted Chemotactic Factor. *Molecular and Cellular Biology* [online]. 2006, **26**(24), 9484-9496 [cit. 2021-07-09]. ISSN 0270-7306. Dostupné z: doi:10.1128/MCB.01030-06

[74] BRÁZDA, Václav, Jan KOLOMAZNÍK, Jiří LÝSEK, Lucia HÁRONÍKOVÁ, Jan COUFAL a Jiří ŠT'ASTNÝ. Palindrome analyser – A new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. *Biochemical and Biophysical Research Communications* [online]. 2016, **478**(4), 1739-1745 [cit. 2021-02-26]. ISSN 0006291X. Dostupné z: doi:10.1016/j.bbrc.2016.09.015

[75] Tad - A Desktop Viewer App for Tabular Data. *Tad - A Desktop Viewer App for Tabular Data* [online]. [cit. 2021-07-07]. Dostupné z: https://www.tadviewer.com/

[76] FLEMING, Aaron M., Judy ZHU, Manuel JARA-ESPEJO a Cynthia J. BURROWS. Cruciform DNA Sequences in Gene Promoters Can Impact Transcription upon Oxidative Modification of 2′-Deoxyguanosine. *Biochemistry* [online]. 2020, **59**(28), 2616-2626 [cit. 2021-06-30]. ISSN 0006-2960. Dostupné z: doi:10.1021/acs.biochem.0c00387

[77] ČUTOVÁ, Michaela, Jacinta MANTA, Otília PORUBIAKOVÁ et al. Divergent distributions of inverted repeats and G-quadruplex forming sequences in Saccharomyces cerevisiae. *Genomics* [online]. 2020, **112**(2), 1897-1901 [cit. 2021-06-30]. ISSN 08887543. Dostupné z: doi:10.1016/j.ygeno.2019.11.002

[78] ZIMORSKI, Verena, Chuan KU, William F MARTIN a Sven B GOULD. Endosymbiotic theory for organelle origins. *Current Opinion in Microbiology* [online]. 2014, **22**, 38-48 [cit. 2021-06-30]. ISSN 13695274. Dostupné z: doi:10.1016/j.mib.2014.09.008

[79] ČECHOVÁ, Jana, Jiří LÝSEK, Martin BARTAS, Václav BRÁZDA a John HANCOCK. Complex analyses of inverted repeats in mitochondrial genomes revealed their importance and variability. *Bioinformatics* [online]. 2018, **34**(7), 1081-1085 [cit. 2021-02-23]. ISSN 1367-4803. Dostupné z: doi:10.1093/bioinformatics/btx729

[80] BRÁZDA, Václav, Jiří LÝSEK, Martin BARTAS a Miroslav FOJTA. Complex Analyses of Short Inverted Repeats in All Sequenced Chloroplast DNAs. *BioMed Research International* [online]. 2018, **2018**, 1-10 [cit. 2021-06-28]. ISSN 2314-6133. Dostupné z: doi:10.1155/2018/1097018

[81] BARTAS, Martin, Michaela ČUTOVÁ, Václav BRÁZDA et al. The Presence and Localization of G-Quadruplex Forming Sequences in the Domain of Bacteria. *Molecules* [online]. 2019, **24**(9) [cit. 2021-06-13]. ISSN 1420-3049. Dostupné z: doi:10.3390/molecules24091711

[82] REUTER, Jessica S a David H MATHEWS. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* [online]. 2010, **11**(1) [cit. 2021-03-24]. ISSN 1471-2105. Dostupné z: doi:10.1186/1471-2105-11-129

8 POUŽITÉ ZKRATKY

HMGB	proteiny z rodiny HMG (z angl.: High-Mobility Group) vysoce pohyblivých chromozomálních proteinů obsahujících HMG-Box doménu
COL1A1	z angl.: Collagen Type I Alpha 1 Chain
COL1A2	z angl.: Collagen Type II Alpha 1 Chain
AFM	mikroskopie atomárních sil (z angl.: atomic force microscopy)
ARS307	Autonomní replikační sekvence nacházející se na chromozomu III kvasinek Saccharomyces Cerevisiae
SAF-box	z ang. scaffold attachment factor-box
BRCA1	z ang. Breast-associated protein-1
NLS	z ang. nuclear localization signal or sequence (NLS)
BARD1	z ang. BRCA1 Associated ING Domain protein 1
PALB2	z ang. Partner and localizer of BRCA2
SCD	z ang. Serine Containing Domain
ATM	z ang. Ataxia Telangiectasia Mutated protein
RB	z ang. Retinoblastoma protein
BACH1	z ang. BTB and CNC homology 1, basic leucine zipper transcription factor 1
CtIP	z ang. C-terminal binding protein 1 (CtBP1) interacting protein
Abraxas	z ang. BRCA1-A complex subunit, podjednotka komplexu BRCA1-A
scDNA	z ang. supercoiled DNA, superhelikální DNA
SMC	z ang. Structural maintenance of chromosomes proteins
S/MAR	z ang. Scaffold/matrix attachement region
VLF1	z ang. Very late expression factor 1
AcMNPV	z ang. Autographa californica nuclear polyhedrosis virus
DEAE-C	z ang. Diethylaminoethyl cellulose chromatography
HeLa C.	Buněčná linie lidských epitelových buněk odebraných z maligního karcinomu děložního čípku Henrietty Lacksové

MLL	z ang. Mixed Lineage Leukemia
KMT2A	z ang. Lysine [K]-specific MethylTransferase 2A
bp.	z ang.base pair