VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

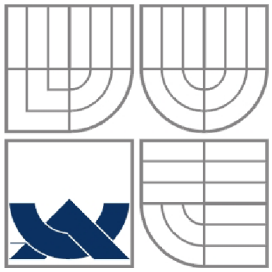
VIZUALIZACE SEKUNDÁRNÍCH STRUKTUR DNA V PROSTŘEDÍ R/BIOCONDUCTOR

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

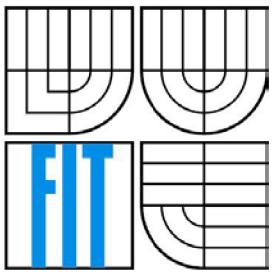
AUTOR PRÁCE
AUTHOR

MARTA ČUDOVÁ

BRNO 2014



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

VIZUALIZACE SEKUNDÁRNÍCH STRUKTUR DNA V PROSTŘEDÍ R/BIOCONDUCTOR

VISUALIZATION OF DNA SECONDARY STRUCTURES IN R/BIOCONDUCTOR

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR
MARTA ČUDOVÁ

VEDOUCÍ PRÁCE
SUPERVISOR
Ing. TOMÁŠ MARTÍNEK, Ph.D.

BRNO 2014

Abstrakt

Tato bakalářská práce se zabývá zobrazovacími technikami sekundárních struktur DNA. Shrnuje a rozebírá některé současné metody vizualizace. Objasňuje problém a jeho význam z hlediska molekulární biologie. Cílem práce je návrh obecného algoritmu pro vizualizaci sekundárních struktur DNA, konkrétně palindromických a triplexových struktur a jeho implementace do prostředí R/Bioconductor. Zaměřuje se zejména na oblast 2D grafického zobrazování zmíněných sekundárních struktur DNA. Výsledkem práce je vytvořená 2D grafická podpora pro softwarové balíčky určené pro prostředí R/Bioconductor, které zabezpečují vyhledávání charakteristických sekvencí palindromů a triplexů v sekvencích DNA.

Abstract

This bachelor‘s thesis deals with the visualization techniques of the secondary structures of DNA. It summarizes and discusses some of the current methods of visualization. It explains the problem and its importance from the standpoint of molecular biology. The main aim of the work is the design of a generic algorithm for visualization of secondary structures of DNA, specifically palindromes and triplexes, and its implementation into the environment R/Bioconductor. It focuses in particular on the area of 2D graphical imaging of these secondary structures of DNA. The result of the work is 2D visualization support for R/Bioconductor‘s software packages which provide searching of the characteristic sequence of palindromes and triplexes in DNA sequences.

Klíčová slova

DNA, sekundární struktury, palindrom, triplex, vizualizace, 2D, diagram, jazyk R, Bioconductor

Keywords

DNA, secondary structures, palindrome, triplex, visualization, 2D, diagram, R language, Bioconductor

Citace

Marta Čudová: Vizualizace sekundárních struktur DNA v prostředí R/Bioconductor, bakalářská práce, Brno, FIT VUT v Brně, 2014

Vizualizace sekundárních struktur DNA v prostředí R/Bioconductor

Prohlášení

Prohlašuji, že jsem tuto práci vypracovala samostatně, pod vedením pana Ing. Tomáše Martínka, Ph. D. Uvedla jsem všechny literární prameny a publikace, ze kterých jsem čerpala.

.....
Marta Čudová
19. května 2014

Poděkování

Na tomto místě bych ráda vyjádřila velké poděkování vedoucímu mé práce, Ing. Tomášovi Martínkovi, Ph. D., za jeho cenné rady, doporučení a odborné směřování v průběhu jejího zpracování.

Obsah

1 Úvod

Sekundární struktury DNA jsou v poslední době středem zájmu mnoha vědeckých týmů, protože se předpokládá, že hrají významnou roli v různých biologických procesech, které probíhají v živých organismech. Biologický význam těchto struktur však doposud není uspokojivě vysvětlen a zmapován. Grafická vizualizace zkoumaných struktur DNA pomocí informačních technologií pomáhá vědeckým pracovníkům lépe pochopit jejich vzhled, prostorovou konfiguraci a vzájemná propojení.

Hlavním cílem mé bakalářské práce navrhnout obecný algoritmus pro vizualizaci sekundárních struktur DNA. Zaměřuji se zejména na oblast 2D grafické reprezentace zájmových sekvencí. Hlavním aplikačním cílem je pak doplnění podpory grafické prezentace výsledků pro softwarové balíčky, určené pro prostředí R/Bioconductor [22], které jsou určeny pro vyhledávání charakteristických sekvencí palindromů či triplexů v DNA řetězcích.

V úvodní části práce stručně uvádím současný stav řešené problematiky s krátkou exkursí do odpovídající oblasti molekulární biologie. Jde zejména o problematiku vybraných sekundárních struktur DNA, zejména palindromů a triplexů tak, aby bylo zřejmé, jakými problémy se ve své práci dále zabývám a co je cílem vizualizace těchto sekundárních struktur DNA.

V následující kapitole zmiňuji některé rozšířené nástroje, které se zobrazováním sekvencí resp. struktur DNA, již zabývají. Stručně zde uvádím také prostředí R/Bioconductor.

Dále uvádím kritické zhodnocení výchozího stavu předmětu řešení, tj. vizualizační podporu pro zkoumané zájmové sekundární struktury DNA, odpovídajících softwarových balíčků prostředí R/Bioconductor. Na základě tohoto zhodnocení následně formuluji konkrétní cíle práce, a to jak hlavní, tak i postupné cíle resp. etapy.

Těžiště práce leží v kapitolách 5 a 6, které obsahují popis řešení konkrétních cílů práce. Jsou zde popsány přístupy k modelování charakteristických entit zájmových sekundárních struktur DNA a jejich využití při sestavování jejich 2D symbolických diagramů. Jsou zde také uvedeny zjednodušené vývojové diagramy popisující algoritmus implementovaných zobrazovacích funkcí a je popsán způsob jejich implementace v prostředí R/Bioconductor. Na závěr kapitoly 6 jsou uvedeny příklady vizualizace reálných sekvencí DNA z lidského genomu.

V závěru práce jsou diskutovány výsledky práce a zhodnoceno plnění cílů bakalářské práce.

2 Základy molekulární biologie

Tato kapitola pojednává o základech molekulární biologie. Je zde popsána primární a sekundární struktura DNA a její prostorové konformace. Text se zejména zaměřuje na B-DNA konformaci, která se za normálních podmínek vyskytuje zcela nejčastěji a mimo jiné je zároveň i předmětem mého grafického zobrazování. Dále se kapitola podrobně zabývá jednotlivými druhy párování, Watson-Crickovým a Hoogsteenovým, a také sekundárními strukturami DNA, konkrétně palindromy a triplexy.

2.1 DNA

Deoxyribonukleová kyselina neboli DNA, je informační makromolekulou nesoucí genetické informace. Nejčastější formou, u níž se polynukleotidová vlákna pravotočivě vinou kolem své osy, je tzv. B-forma DNA neboli B-DNA, což je základní a za normálních podmínek nejčastěji se vyskytující konformací DNA v živých organismech. DNA je obsažena v buňkách všech organizmů a v DNA virových částicích. [1][2]

DNA se tedy může vyskytovat v různých formách a podle zvolených kritérií může být klasifikována do několika základních skupin [29]:

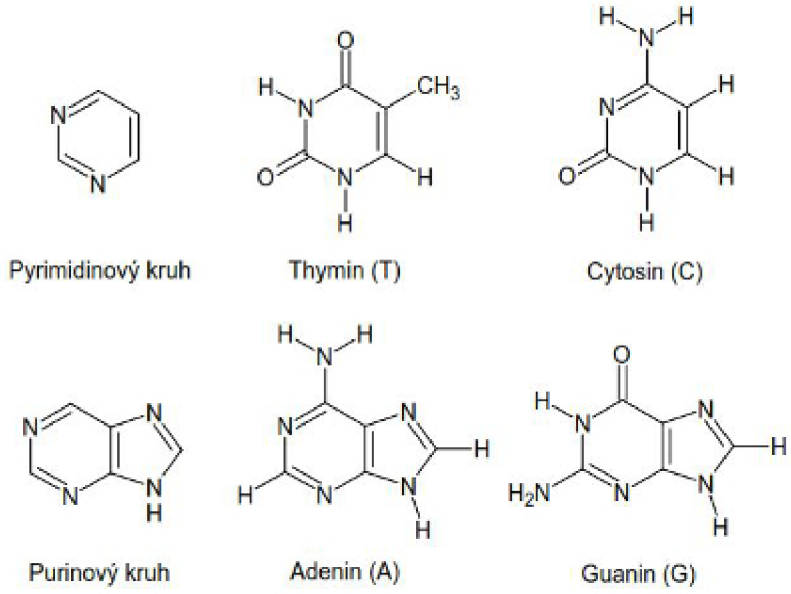
- Podle počtu deoxyribonukleotidových řetězců:
  - *Jednořetězcová DNA*, označovaná též jako ssDNA (z anglického „single stranded“), sestávající se pouze z jednoho deoxyribonukleotidového řetězce.
  - *Dvouřetězcová DNA*, označovaná též jako dsDNA (z anglického „double stranded“), sestávající se ze dvou deoxyribonukleotidových řetězců.
- Podle tvaru:
  - *Kružnicová a lineární DNA*. Obě formy DNA mohou být jednořetězcové i dvouřetězcové.
- Podle lokalizace v buňce:
  - *Jaderná DNA (nDNA)*. DNA se vyskytuje v buněčném jádře eukaryot.
  - *Mimojaderná DNA*. Tato lokalizace je typická pro prokaryoty, DNA se tedy vyskytuje volně v cytoplazmě.

Na strukturální stavbu DNA můžeme pohlížet hned z několika úrovní. Jestliže budeme soustředit náš zájem pouze na jednotlivá vlákna a pořadí nukleotidů, budeme hovořit o *primární struktuře* DNA. Stáčení vláken do útvaru zvaného šroubovice již budeme označovat za *sekundární strukturu* a uspořádání vláken a nukleotidů v prostoru jako *terciální strukturu* DNA [1][2]. Terciální struktura, neboli nadšroubovice (superhelix), vzniká tzv. nadšroubovicovým vinutím sekundární struktury. Tento proces je reverzibilní. Sekundární nebo primární struktura se nazývá relaxovaná, jestliže vznikla relaxací terciální struktury [29].

Nukleotidy

Nukleotidy představují základní stavební kameny molekuly DNA. Mononukleotid obsahuje tři složky – pentózu, což je pětiuhlíkatý cukr, který je v případě molekuly RNA tvořen ribózou a v případě molekuly DNA deoxyribózou. Dále je mononukleotid tvořen fosfátem a dusíkatou, resp. organickou, bází. Podle typu báze rozlišujeme v rámci molekuly DNA čtyři typy nukleotidů, které jsou odvozeny

buď od pyrimidinu, nebo purinu. Mezi deriváty pyrimidinu řadíme báze Cytosin a Thymin. Deriváty purinu jsou báze Adenin a Guanin. Strukturní vzorce nukleotidů DNA jsou uvedeny na obr. 1[1].



Obr. 1: Typy nukleotidů v DNA

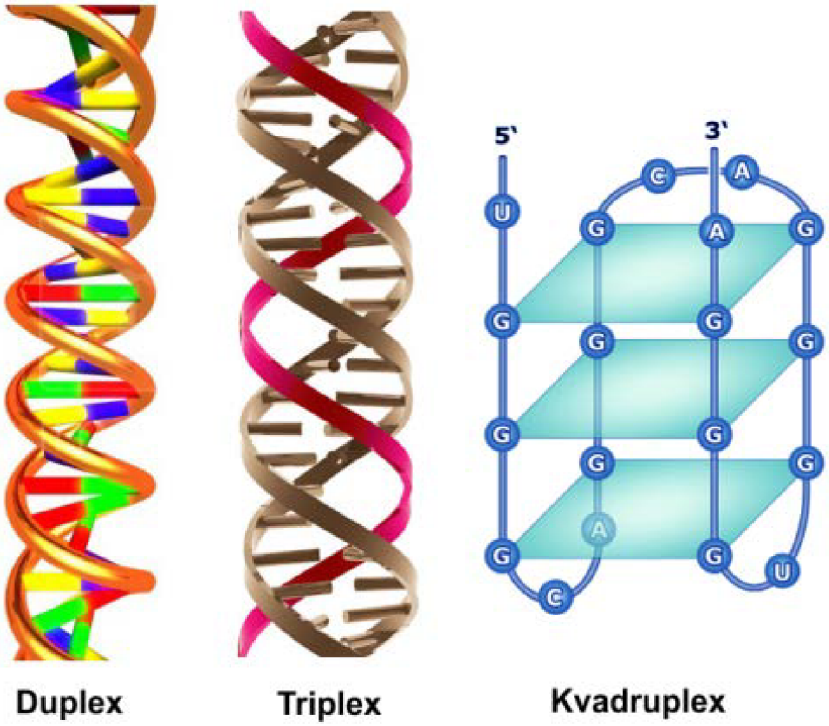
Řetězce nukleotidů jsou vzájemně spojeny fosfodiesterovými vazbami, které vznikají mezi deoxyribózou a fosfátovou skupinou. 3,5 - fosfodiesterová vazba je vazba propojující uhlík C3-deoxyribózy jednoho nukleotidu s uhlíkem C5-deoxyribózy následujícího nukleotidu. Orientace deoxyribózy určuje označení konce vlákna. Pentózafosfátové páteře obou řetězců DNA se otáčejí po zevním povrchu dvoušroubovice, zatímco báze směřují vždy k její ose. [1][2]

Pořadí nukleotidů zapisujeme jako posloupnost nukleotidů ve směru 5'3'. Toto vlákno slouží jako vzor při replikaci. Při syntéze se řetězec prodlužuje na konci označeném jako 3'. V deoxyribonukleotidech je fosfát navázán v poloze 3' nebo 5'. Pokud mají dvě spojená vlákna stejný směr, říkáme, že se jedná o tzv. paralelní vlákna. V opačném případě, tj. pokud vlákna mají opačný směr, nazýváme vlákna antiparalelními. [1][2]

Nukleotid však není chápán jen jako základní složka DNA, ale rovněž i jako délková jednotka. Délku genu nebo jiné sekvence v molekule DNA zpravidla vyjadřujeme počtem nukleotidových párů. Tyto jednotky označujeme jako pár bází (ve smyslu komplementárních bází v nukleotidech), zkratkou „*bp*" (z anglického „*base pair*"). Většina genů dosahuje délky několika tisíc až milionů párů bází. V praxi proto pracujeme s odvozenými jednotkami *kbp* nebo *Mbp*, pro které platí stejný převodní vztah jako u fyzikálních veličin.[6]

Párování bází

Při párování bází dochází ke spojování protilehlých bází vodíkovými můstky podle určitých pravidel. Vodíkové vazby neboli můstky jsou slabé vazebné interakce, které působí mezi molekulami, a díky kterým může vznikat kanonická dvojitá šroubovice DNA.

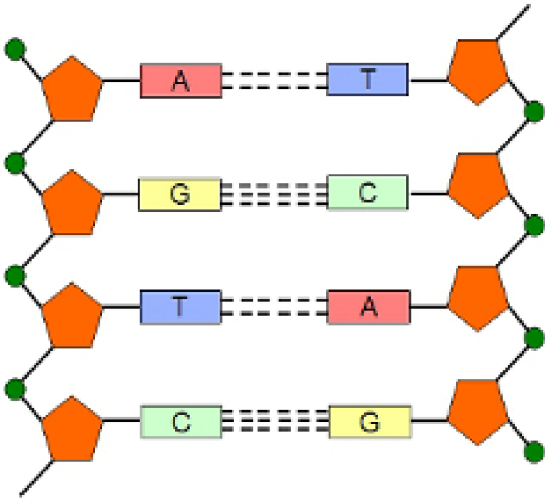


Obr. 2: Různá uskupení podle párování bází mezi více řetězci

K takovému párování může docházet mezi dvěma řetězci. Takové uskupení poté nazýváme jako *duplex* a vzniká tak typické dvouvláknové uspořádání zvané šroubovice. K párování může ovšem docházet i mezi více jak dvěma řetězci. Vznikají tak struktury jako *triplex*, kdy se párují tři vlákna spolu nebo uspořádání zvané *kvadruplex*, kdy dochází k párování čtyř vláken mezi sebou. Viz obr. 2.

Watson - Crickovo párování

Jednotlivá vlákna se mohou vázat pomocí vodíkových můstků mezi nukleovými bázemi. Pomocí vodíkových můstků se váže vždy purinová báze s pyrimidinovou (viz zákon komplementarity), konkrétně Adenin s Thyminem (značeno A:T) dvěma vodíkovými vazbami a Guanin s Cytosinem (G:C) třemi vodíkovými vazbami, viz obr. 3. Tímto párováním je vytvořena kanonická dvojitá šroubovice [2]. Watson-Crickovo párování se znázorňuje pomocí znaku dvojtečky „:“ [20], což používám i nadále v textu.



Obr. 3: Watson - Crickovo párování

Hoogsteenovo (nekanonické) párování

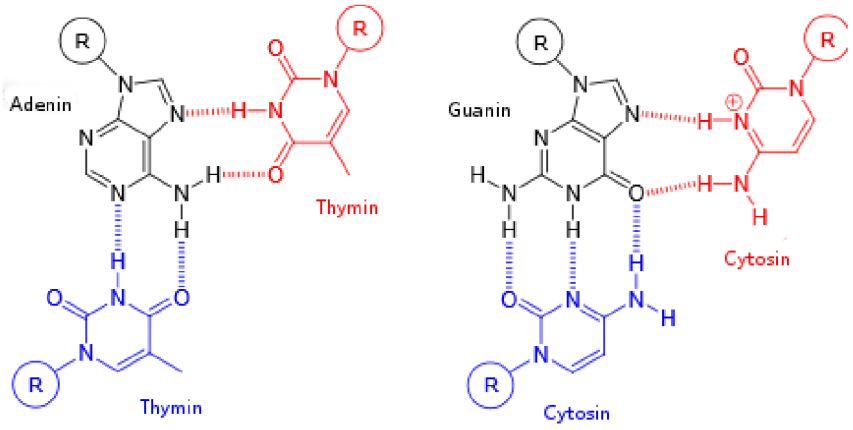
Hoogsteenovo párování, někdy též označované jako obrácené Watson-Crickovo párování, je pojmenováno podle svého objevitele Karsta Hoogsteena. Jedná se o vazby vodíkových můstků mezi bázemi již spárovanými i volnými, které mezi sebou spojují nekomplementární báze nebo báze komplementární v jiných polohách, než jsou vazby obvyklé (kanonické). Na základě jejich existence vznikají na řetězcích nukleových kyselin tzv. nekanonické struktury. Hoogsteenovy vazby jsou sice slabší než Watson-Crickovy, ale v případech tvorby triplexů DNA se oba typy interakcí vyskytují současně [4]. Hoogsteenovo párování se může vyskytnout i samostatně jako upřednostňovaná možnost interakce, a to jak pro nekanonické ale i kanonické páry bází. [16][17]

Hoogsteenovo párování umožňuje vznik triplexových a kvadruplexových struktur. Vznikají takto tzv. triády nebo tetrády. Triády jsou trojice bází spárovaných tak, že stejná purinová báze tvoří jednak Hoosteenův pár s pyrimidinovou bází zároveň a Watson-Crickův pár s pyrimidinovou bází.

Příklad 1: Triáda typu YR*Y a YR*R.
Kde Y značí pyrimidinovou bázi, R značí purinovou bázi, a * obecně znamená odchylku od Watsonova-Crickova párování bází.

Tetrády vznikají při párování mezi čtyřmi bázemi, které se uskuteční mezi molekulami Guaninu a Cytosinu a mezi molekulami Adeninu a Thyminu.

Hoogsteenovo párování vzniká připojením další molekuly Thyminu k páru A:T (T:A.T) nebo další molekuly Cytosinu k páru C:G (C:G.C). Tato struktura však může vznikat pouze tehdy, pokud je dodatečné vlákno homopyridinové povahy. Musí tedy platit, že celé vlákno je tvořeno jen molekulami Cytosinu a Thyminu, neboť každý Hoogsteenův pár vyžaduje v řetězci dvojité šroubovice molekulu Guaninu nebo Adeninu. [16]



Obr. 4: Hoogsteenovo párování.
Hoogsteenovo párování je znázorněno červeně, Watson-Crickovy vazby modře.

Zákon komplementarity

Komplementarita bází vysvětluje důvod, proč každá dvojitá šroubovice DNA obsahuje stejný počet purinových i pyrimidinových nukleotidů. Můžeme tedy říci, že platí Chargaffovo pravidlo o počtu těchto bází: A = T, C = G a A + G= T + C.

Víme, že Adenin se páruje s Thyminem a Cytosin s Guaninem. Z toho plyne, že zastoupení Adeninu, resp. Cytosinu, je stejné jako zastoupení Thyminu, resp. Guaninu.

Stabilita

Stabilita nukleových kyselin závisí na sekvenci a skladbě nukleotidů [9]. Uspořádání bází nad sebou vodorovně umožňuje vzájemné interakce jejich elektronů, označované jako *base - stacking* (stohování bází), což významně přispívá ke konformační stabilitě šroubovice za běžných podmínek v organismech (vysoká koncentrace solí, vysoký stupeň hydratace). [18]

Molekula DNA s převažující počtem párů G:C, které jsou mezi sebou vázány třemi vodíkovými můstky, je odolnější vůči denaturaci. Což znamená vůči rozpadu dvoušroubovice DNA na dvě samostatná vlákna vlivem působení různých vnějších vlivů, které vedou k rozrušení vodíkových vazeb mezi komplementárními bázemi. [18]

Malý a velký žlábek

Malý a velký žlábek jsou útvary, které vznikají na povrchu šroubovice, kde se pravidelně střídají (viz obr. 5). Vznikají zatočením spárovaných řetězců do šroubovice kolem osy, která vede středem vodíkových vazeb mezi bázemi.

Páry bází na těchto spárovaných řetězcích, jejichž glykosidické vazby nejsou vždy úplně stejné, jsou uspořádány asymetricky na cukr-fosfátových kostrách řetězce DNA, což má za příčinu různou hloubku žlábku. Glykosidické vazby tak vždy vychází z obou dvou nukleových bází komplementárního páru na jednu stranu dvoušroubovice. V důsledku toho jsou okraje párů bází ve velkém žlábku širší, než ty ve žlábku malém. Žlábky umožňují interakce řetězce DNA s různými molekulami. Každý žlábek je lemován potenciálními dárci vodíkových vazeb a příjemci atomů, které interagují s DNA vazebnými proteiny. [12][19]

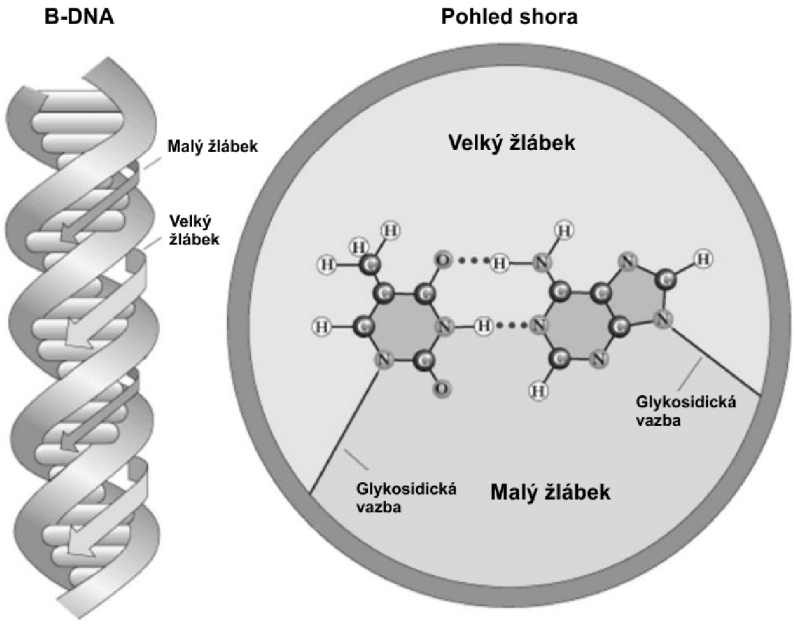
Velký žlábek (GM) je 12 Å široký a 8,5 Å hluboký. Interaguje s α helixy. Helix se orientuje podél osy žlábku, postranní řetězce tvoří vodíkové vazby s bázemi, eventuálně kontakty s fosfátovými skupinami. [12] Úhel C(1') – osa helixu – C(1') větší než 180°. [13]

Malý žlábek (Gm) je 6 Å široký a 7,5 Å hluboký. Interaguje s β skládanými listy. List leží podél osy žlábku, tvoří vodíkové vazby k dusíkům aminoskupin a k fosfátům. [12] Úhel C(1') – osa helixu – C(1') menší než 180°. [13]

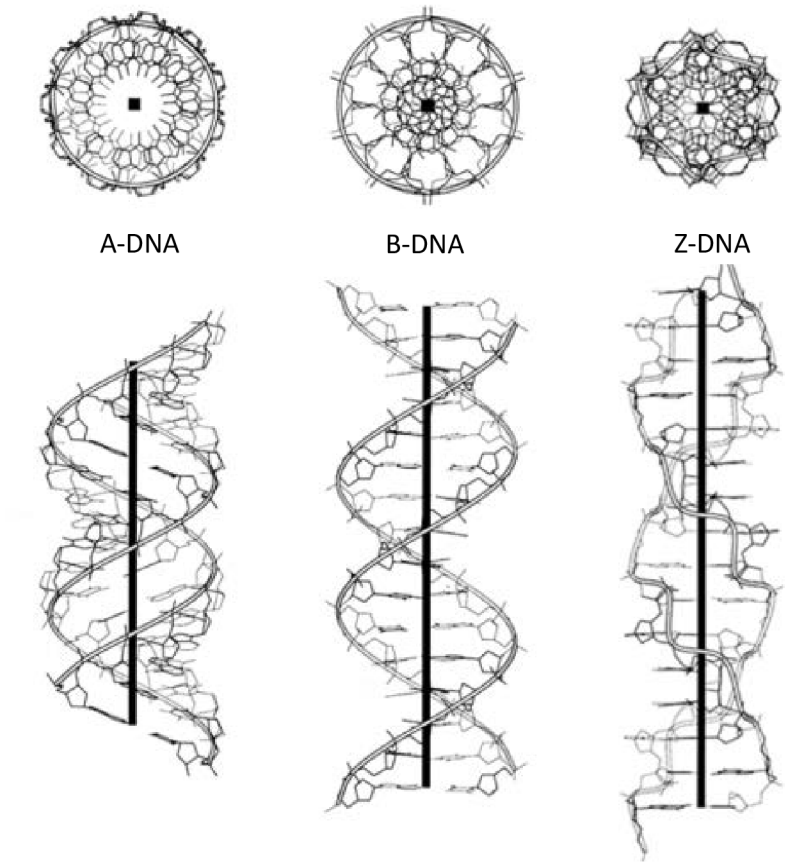
Nukleové kyseliny řadíme je k tzv. heteropolymérům, neboť jsou sestaveny z různých typů základních jednotek. Tato skutečnost je podstatná pro uchovávání a předávání informace.[1]

Konformace DNA

Konformace je prostorové uspořádání biomakromolekuly do struktury, která je za daných podmínek energeticky nejvýhodnější. Konformace jsou ovlivněny svým prostředím, ve kterém se molekula nachází, tj. obsah vody a iontová síla prostředí, ale také samotnou sekvencí nukleotidů. Následující obr. 6 zobrazuje nejznámější konformace DNA, jejich prostorové uspořádání a pohled na strukturu shora.



Obr. 5: Velký a malý žlábek ([32] upraveno)



Obr. 6: Konformace DNA [33].

Jak již bylo zmíněno výše, základní DNA konformací je B-forma, která za normálních podmínek v buňkách převažuje. Blíží se tradičnímu Watson-Crickovskému modelu dvoušroubovice. [8]

Konformace deoxyribóz je 2'– endo. To znamená, že zatímco 4 atomy deoxyfuranózového kruhu jsou v rovině, C2' uhlík vystupuje stejným směrem jako C5'. DNA se v této formě vyskytuje ve vodných roztocích obsahujících nízké koncentrace solí.

Dalšími častými DNA konformacemi jsou A-forma a Z-forma. U A-formy, resp. A-DNA, je deoxyribóza v konformaci 3'– endo. V jednom závitu se vyskytuje 11 bází, které svírají s rovinou dvoušroubovice úhel 20°. Uvnitř šroubovice vzniká poměrně výrazná axiální dutina. Vizuálně se tato forma od B-formy liší tím, že je plošší a širší. V této formě se DNA vyskytuje, pokud je částečně dehydratovaná. [9][15] Z-forma, resp. Z-DNA, je na rozdíl od předchozích dvou forem levotočivá. Její deoxyribóza má konformaci C: C2'-endo, G: C2'-exo, a řetězce se nevinou plynule. Jeden závit obsahuje 12 párů bází. Z-DNA se vyskytuje v molekulách DNA bohatých na páry nukleotidů C:G. [7][9] Tabulka 1 popisuje strukturální rozdíly mezi jednotlivými formami.

Tabulka 1: Konformace DNA a jejich strukturální vlastnosti [9][14][15].

Atribut	A-DNA	B-DNA	Z-DNA
Tvar	nejširší	střední	nejužší
Točivá tendence šroubovice (chiralita)	pravotočivá	pravotočivá	levotočivá
Opakování	po každém páru (1bp)	po každém páru (1bp)	po každých dvou párech (2bp)
Glykosidová vazba	anti	anti	střídavě anti a syn
Úhel otočení po každém opakování	32,7°	36°	-30°
Průměrný počet párů na jedno otočení šroubovice	11	10,5	12
Sklon páru k ose	+13°	0°	-7°
Vzestup vůči ose na jeden pár (vzdálenost vrstev)	2,9 Å (0,29 nm)	3,36 Å (0,336 nm)	3,7 Å (0,37 nm)
Vzestup natočení šroubovice (výška jednoho závitu helixu)	24.6 Å (2.46 nm)	33.2 Å (3.32 nm)	45.6 Å (4.56 nm)
Střední úhel točení	+18°	+16°	0°
Velký žlábek	úzký a velmi hluboký	široký a celkem hluboký	plochý
Malý žlábek	velmi široký a mělký	úzký a celkem hluboký	velmi úzký a hluboký
Konformace cukru	C3' - endo	C2' - endo	C2' - endo, C3' - endo
Průměr	23 Å (2,3 nm)	19 Å (1,9 nm)	18 Å (1,8 nm)

2.2 Sekundární struktura DNA

Jedná se o dvoušroubovici (double helix), kde se vzájemným obtáčením komplementárních řetězců vytváří dvoušroubovicové vinutí, které může být pravotočivé nebo levotočivé. [29] Podmínkou popisovaného uspořádání DNA je opačná polarita fosfodiesterových vazeb obou řetězců. Polynukleotidové řetězce DNA jsou antiparalelní. [1][6]

V závislosti na okolních podmínkách může sekundární struktura přecházet v primární. Tomuto jevu se říká denaturace a dochází při něm k narušení vodíkových vazeb mezi oběma komplementárními řetězci a k jejich vzájemnému rozpojení. [29]

Palindrom

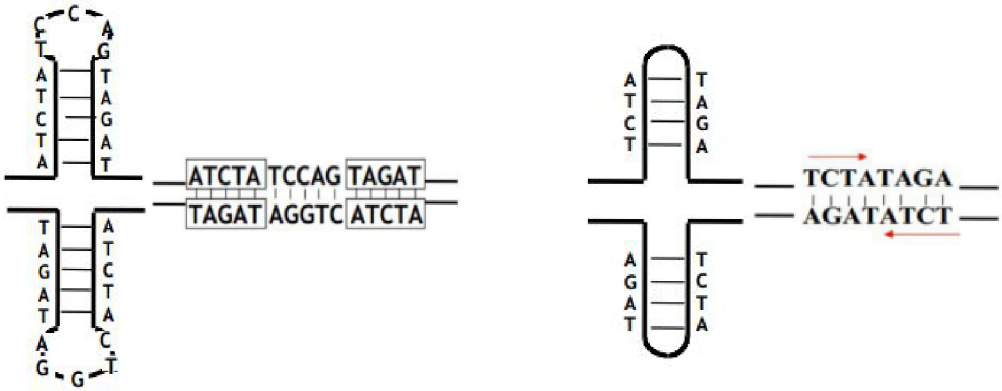
Palindrom je část nukleotidového řetězce DNA nebo RNA. Jedná se o centrálně symetrickou sekvenci nukleotidů. Sekvence nukleotidů se nazývá palindrom, pokud platí, že daná sekvence ve směru 5'3' je komplementární se sekvencí nukleotidů ve směru 3'5'. Palindromy tedy vznikají jako obrácené přilehlé repetice, což znamená, že spolu bezprostředně sousedí. Výskyt obrácených repetic vede k tvorbě vlásenek a křížových struktur.

Pokud se obrácené repetice nacházejí pouze na jednom vlákně, jedná se o vlásenku, což je typické pro molekuly RNA. Jestliže se vyskytují na obou vláknech zároveň, jedná se o křížovou strukturu. Taková struktura je typická pro molekuly DNA. Častý výskyt palindromů je v bílkovinách.

Vlásenka je struktura, která může vznikat v rámci jednoho řetězce DNA, avšak mnohem častěji v rámci řetězce RNA, v místech, kde se vyskytují obrácené repetice. Je to dvoušroubovicová struktura vzniklá párováním komplementárních nukleotidových sekvencí na témže DNA řetězci. Rozlišujeme vlásenku se smyčkou nebo jen samotnou vlásenku (viz obr. 7). V případě vlásenky se smyčkou se mezi těmito dvěma oblasti nachází nespárovaná sekvence nukleotidů tvořící smyčku. [9]

Obr. 7: Vlásenka se smyčkou (vlevo), vlásenka bez smyčky (vpravo)

Křížové struktury vznikají na dsDNA a jsou opět tvořené palindromatickými sekvencemi. Ty vznikají spárováním obrácených repeticí na obou komplementárních řetězcích dsDNA. Křížová struktura může být opět dvojího druhu, se smyčkou nebo bez smyčky (viz obr. 8). [9]



Obr. 8: Křížová struktura se smyčkou (vlevo), bez smyčky (vpravo)

Palindromy plní různé biologické funkce a jsou tak tedy ve středu zájmu biologů. V dnešní době je známo pouze několik biologických významů křížové DNA, avšak komplementární palindromy v lineárním uspořádání hrají důležitou roli jako vazebná místa pro dimetrické proteiny. Křížová DNA

je spojovaná s místy, ve kterých DNA začíná kopírování sebe sama. K vytvoření triplexové struktury DNA (viz podkapitola triplex (H-DNA)) je zapotřebí tří vláken nukleových bází. Frank-Kamenetskii a jeho spolupracovníci dokázali, že DNA musí, mimo jiných podmínek, obsahovat také komplementární palindrom, aby mohlo dojít k vytvoření triplexové struktury. Lidská DNA má potenciál k vytvoření takových struktur. Tyto oblasti jsou často spojované s regulačními oblastmi genů, kde triplex možná může sloužit jako ukončovací značka při kopírování DNA. [2]

Triplex (H-DNA)

Triplex (trojitá šroubovice) je druh šroubovice, v niž jsou základní báze spárované pomocí Watson-Crickových vodíkových můstků a třetího vlákna, které vytváří Hoogsteenovy vodíkové můstky s každým párem šroubovice.

Podle původu vzniku a orientace třetího vlákna a podle původu vzniku samotného triplexu rozlišujeme triplexy následovně:

- Podle původu vzniku triplexu: [9]
    - *Intramolekulární.* Takové triplexy vznikají v rámci jedné molekuly DNA.
    - *Intermolekulární.* Triplexy vznikají spojením dvou molekul DNA. Jeden úsek DNA tvoří šroubovici a třetí komplementární vlákno je tvořeno z rozpojené šroubovice druhého úseku DNA.
- Podle původu vzniku třetího vlákna: [3]
    - *Mezivláknové (z angl. Interstrand) triplexy.*
    - *Vnitrovláknové (z angl. Intrastrand) triplexy.*
- Podle orientace třetího vlákna vzhledem k prostřednímu: [3]
    - *Paralelní triplexy.*
    - *Antiparalelní triplexy.*

Třetí vlákno se vyskytuje ve velkém žlábku dvoušroubovice.

Triplex je tvořen vrstvami zvané triplety. Triplet je tvořen buď dvěma nukleotidy purinu a jedním nukleotidem pyrimidinu (PyPuPu) nebo dvěma nukleotidy pyrimidinu a jedním nukleotidem purinu (PyPuPy). Purin dokáže vytvořit dva vodíkové můstky se třetím vláknem, zatímco pyrimidin může tvořit pouze jeden. [9][10]

Nukleotidy ve třetím vlákně dále nemohou být libovolné, Adenin může vytvářet Hoogsteenovy vazby především s Thyminem nebo Adeninem, zatímco Guanin s Cytosinem nebo Guaninem.

Pro triplet PyPuPy existují klasické trojice C:G.C a T:A.T, a pro triplet PyPuPu trojice C:G.G a T:A.A. Triplety typu PyPuPy jsou závislé na pH, zatímco triplety typu PyPuPu nikoliv. [9]

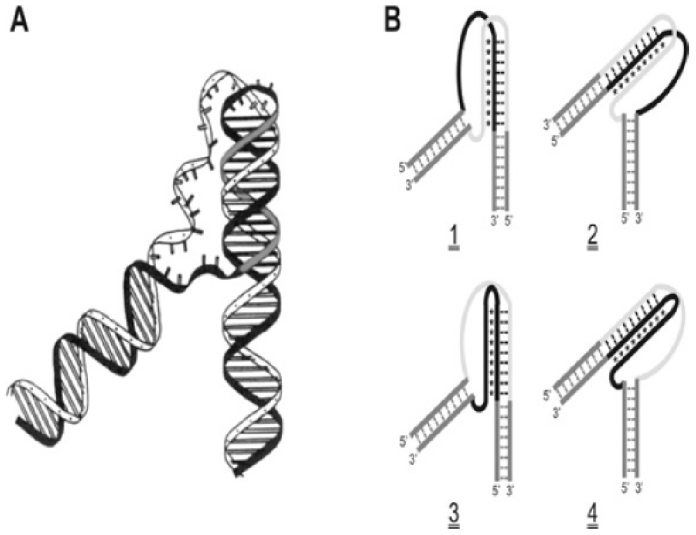
Podobně jako u dvoušroubovice, můžeme i zde strukturu trojité šroubovice popsat podobnými parametry. U trojité struktury je navíc nutné specifikovat vzájemnou polohu tří vláken. Sledujeme tedy průměr tripletu r [Å] a úhel svírající Watson – Crickovy a Hoogsteenovy vazby mezi sebou. Tento úhel značím symbolem α. Hodnoty těchto parametrů závisí na typu tripletu. V jednotlivých vrstvách se však mohou lišit a z toho důvodu není možné určit hodnotu každého parametru pro celý triplex.

Hodnoty úhlů mezi vazbami a průměr triplexu v závislosti na typu tripletu uvádí tabulka 2. Triplet je zapsán ve tvaru H.WC:WC.[3]

Tabulka 2: Parametry triplexové struktury. Převzato a upraveno z [3].

Antiparalelní třetí vlákno						
	A.A:T	A.G:C	C.A:T	G.G:C	T.A:T	T.C:G
α [°]	76,5364	125,6857	72,0865	94,1352	72,2028	93,8591
r [Å]	7,4096	11,9128	72,0865	94,1352	72,2028	93,8591
Paralelní třetí vlákno						
	C.G:C	G.G:C	G.T:A	T.A:T	T.C:G	T.G:C
α [°]	108,8662	74,8281	125,5032	103,7168	70,4748	77,8939
r [Å]	8,8971	7,1615	12,2812	7,8968	6,4697	6,6302

Obr. 10 znázorňuje intramolekulární triplex. V části A je patrno, že šedě znázorněné třetí vlákno se zdržuje ve velkém žlábku dvoušroubovice. Část B reprezentuje čtyři možné formy intramolekulárního triplexu.



Obr. 9: Intramolekulární triplex, převzato z [20]

2.3 Genetické mutace

V této podkapitole čerpám ze zdroje [28]. Genetickými mutacemi rozumíme změny v genotypu organismu. Velká většina mutací vzniká náhodnými mechanizmy, zatímco cílená mutageneze se používá téměř výhradně pro vědecké účely.

Typy mutací můžeme klasifikovat podle následujících kritérií:

- Okolnosti, za jakých došlo k mutaci. Sem řadíme mutace *spontánní* a *indukované*;
- Úroveň, na které je genetická informace ovlivněna. Takové mutace mohou být *genomové*, *genové* nebo *chromozomové*;
- Evoluční pohled. Mutace z daného pohledu může být *výhodná*, *neutrální* nebo *nevýhodná*;
- Z pohledu typu postižených buněk u vyššího organizmu. Do této kategorie řadíme mutace *somatické* a *gametické*.

Dále se již zaměřím pouze na dva typy mutací, inzerci a deleci, které jsou také předmětem zobrazování v mé práci.

Inzerce (adice)

Inzerce je případ, kdy dochází k zařazení jednoho či více nadbytečných nukleotidů do původní sekvence. Pokud je zařazen počet nukleotidů, jehož počet není násobkem čísla tři, dochází k posunu čtecího rámce, tzv. *frameshift mutation*. S tímto souvisí i problém předčasného ukončení proteosyntézy, tj. procesu tvorby bílkovin, z důvodu vzniku terminačního, neboli ukončujícího, kodónu.

Delece

Delece je opačným případem inzerce. Jedná se tedy o případ, kdy dojde ke ztrátě jednoho nebo více nukleotidů z původního řetězce. Delece rovněž může způsobit posun čtecího rámce.

Příklad 2: Inzerce a delece.
Mějme sekvenci nukleotidů AAAGGGCCCTTT.

Příklad inzerce v dané sekvenci může být následující sekvence AAAGG**TT**GCCCTTT, kdy došlo k vložení dvou nukleotidů Thyminu mezi nukleotidy Guaninu. Příklad delece ve stejném vzorovém vlákně může být následující sekvence AAAG-GCCCTTT, kde došlo ke ztrátě nukleotidu Guaninu (znázorněno pomocí pomlčky).

Genetické mutace mají velký význam z hlediska klinické genetiky. Právě mutace jsou příčinou genetických chorob, vývojových vad nebo nádorového bujení. K převážné většině mutací dochází v nekódujících oblastech, tj. oblastech nekódující proteiny. Ani to však neznamená, že nemohou mít negativní vliv. Mnohem závažnější projevy mají však mutace probíhající v kódujících oblastech.

3 Nástroje pro zobrazování sekundárních struktur DNA

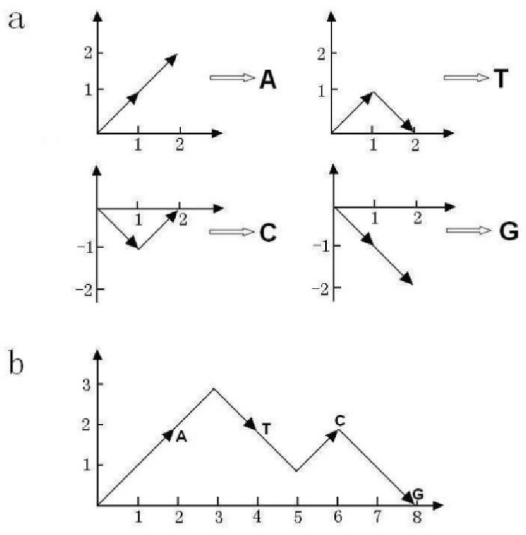
Zobrazování sekvencí DNA je důležitým nástrojem při různých bioinformatických analýzách, porovnávání genových dat apod.

V dostupné literatuře resp. v bioinformatické praxi se používá celá řada zobrazovacích metod, které jsou zaměřeny na zobrazení *(a)* sekvencí primárních struktur DNA, *(b)* sekundárních struktur DNA a *(c)* terciálních struktur DNA.

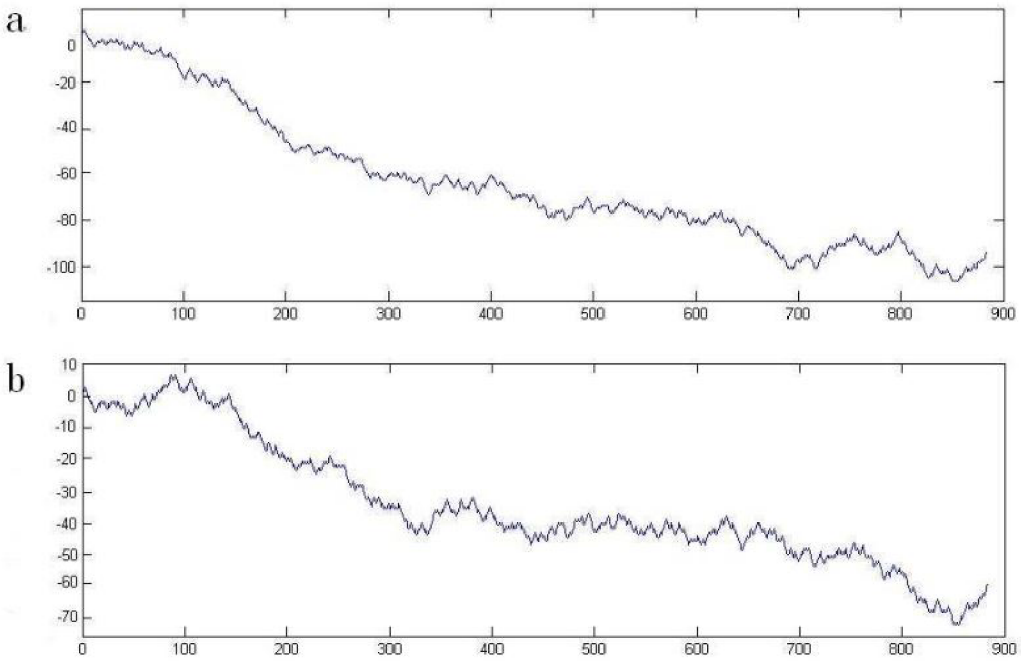
Pro zobrazení sekvencí primárních struktur DNA se využívají zobrazovací techniky s různým rozměrem zobrazované informace označované jako 2D, 3D a 4D (dokonce i 6D) reprezentace [24]. Mezi nejpoužívanější metody patří:

- 2D metoda (původní [25]) zobrazuje jednotlivé nukleotidy A G T C do čtyř ortogonálních směrů. Tento přístup je charakteristický i pro další zmiňované metody zobrazení primárních struktur s tím, že je pochopitelně modifikován. Nevýhodou této původní 2D metody je ztráta vizuální informace při překřížení nebo překrytí křivky;
- 2D IFS (iterovaný funkční systém) [26];
- 2D DV (dual vector) [27];
- 3D PN křivka (PN – páry nukleotidů) [24];

Následující obr. 11 znázorňuje způsob tvorby DV křivky a jednoduchý příklad. Z obr. 12, který znázorňuje reálnou reprezentaci DNA sekvence u člověka a opossuma, je patrno, že jejich DNA sekvence se výrazně liší [27].



Obr. 10: (a) Nukleotidová reprezentace DV křivkou, (b) sekvence ATCG zobrazená pomocí DV křivky [27]



Obr. 11: DV křivka kompletní genové sekvence (a) člověka, (b) opossum [27]

Zobrazení sekundárních struktur DNA přímo souvisí s tématem této bakalářské práce. Samotné vizualizaci sekundární struktury DNA předchází její modelování. Předpokládá se, že aktuální konformace molekuly DNA je, v daných podmínkách, energeticky nejvýhodnější. Metody modelování sekundárních struktur DNA tak využívají různé algoritmy pro určení její minimální volné energie.

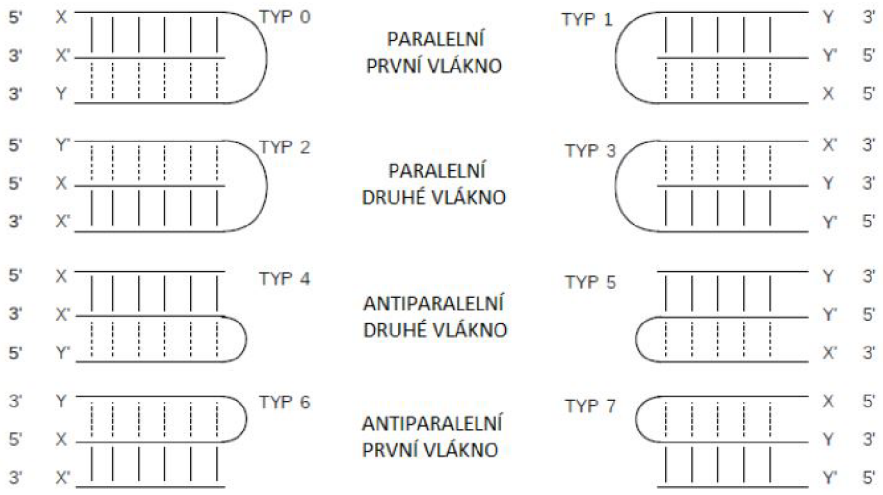
Pro samotné zobrazení sekundárních struktur DNA se využívá *(a)* 2D a *(b)* 3D grafické prezentace.

2D grafická prezentace (např. viz obr. 8) je, v podstatě, symbolické znázornění vláken nukleotidů a jejich vzájemných vazeb. Tato prezentace umožňuje velice názorné zobrazení modelované konformace DNA včetně veškerých vad, nicméně, neposkytuje prakticky žádné informace o její prostorové topologické konfiguraci.

3D grafické zobrazení sekundárních struktur DNA je idealizovaný prostorový obraz jednotlivých vláken DNA, nukleotidů a vazeb mezi nimi, mnohdy i na atomární úrovni.

3.1 Balík Triplex pro R/Bioconductor

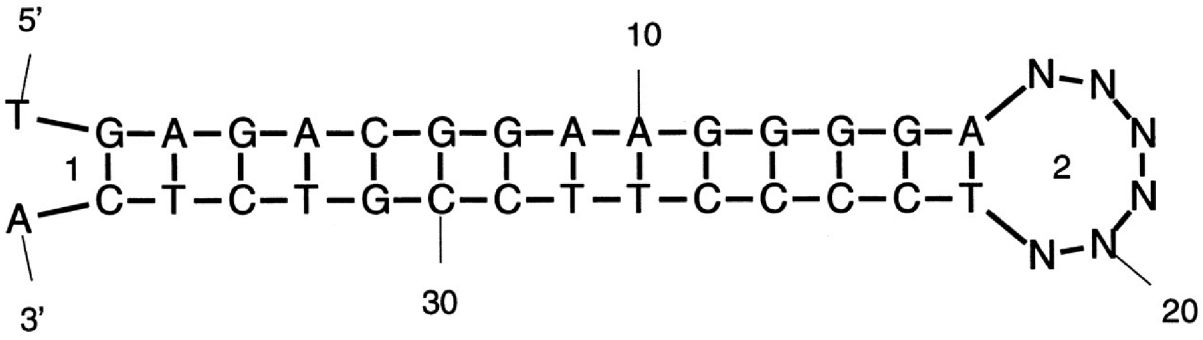
Balík Triplex [21] je vyvinut pro prostředí R/Bioconductor [23], kde je v současnosti jediným nástrojem, který je schopen vyhledávat a zobrazovat triplexové struktury DNA v 1D, 2D i 3D. 1D zobrazení je pouze textové, zobrazující výpis názvů vláken a k nim přiřazenou odpovídající nukleotidovou reprezentaci. 2D i 3D zobrazení je již grafické. Balík používá algoritmus [3], který uvažuje intramolekulární triplexy a rozlišuje u nich osm typů na základě orientace třetího vlákna a pozice triplexu (viz obr. 12)



Obr. 12: Typy triplexů [3]

3.2 Další nástroje

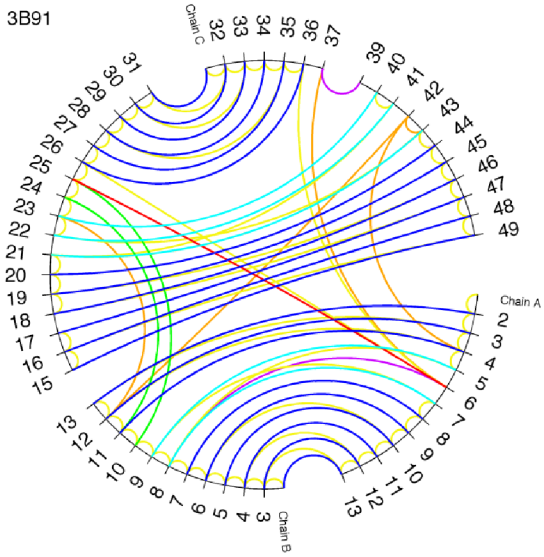
Jedním z nejznámějších programů pro 2D vizualizaci sekundárních struktur je program *Mfold* [11], vytvořený profesorem matematiky Michalem Zukerem. Program je volně šiřitelný a je dostupný z [11]. Program *MFold* používá *Zukerův algoritmus* pro nalezení struktury s minimální energií. Při vývoji své práce jsem se nechala inspirovat, mimo jiné, takovým zobrazením struktur, jaké vytváří program *Mfold* (viz obr. 13).



Obr. 13: DNA struktura vygenerovaná programem Mfold [11]

Rozšířením tohoto programu se stal program *UNAfold* (Unified Nucleic Acid Folding and hybridization package), vytvořený spolupracovníkem profesora Zukera a počítačovým vědcem Nickem Markhamem. *UNAfold* využívá knihovny OpenGL pro 3D vizualizaci struktur. Oba programy jsou založeny na minimalizaci energie molekuly. Příslušný algoritmus obsahuje pravidla pro výpočet celkové energie a hledá se vždy struktura, která má nejnižší energii. [11]

Nussinov diagram (viz obr. 14) zobrazuje sekundární struktury pomocí kružnice. Na této kružnici leží očíslovaná posloupnost bází vláken. Báze, které jsou mezi sebou spojeny Watson-Crickovými nebo Hoogsteenovými vazbami, jsou v kružnici propojeny čárou. Nussinov diagram tedy slouží k jasné identifikaci sekundární struktury RNA. [31]



Obr. 14: Nussinov diagram [34]

4 Kritika současného stavu a definice cílů práce

4.1 Kritika současného stavu v řešené oblasti

Na počátku řešení jsem obdržela aktuální verze softwarových balíčků pro prostředí R/Bioconductor [23], které jsou určeny pro vyhledávání a analýzy palindromů a intramolekulárních triplexů v sekvencích DNA. Konkrétně se jednalo o balíčky *palindrome* (verze 0.99.0) a *triplex* (verze 1.2.1). Poté, co jsem se zevrubně seznámila s implementovanou funkčností obou balíčků, jsem na základě dílčích testů mohla konstatovat následující závěry důležité pro stanovení cílů práce a postupu řešení:

- balíček *palindrome* v současné verzi nemá implementovanou podporu pro vizualizaci zkoumaných struktur DNA;
- balíček *triplex* v dodané verzi poskytuje 2D a 3D zobrazovací metody struktur triplexů, které jsou zahrnuty ve funkcích *triplex.diagram* (2D) a *triplex.3D*. Tento program je v současnosti jediným, který zobrazuje triplexové struktury v prostředí R/Bioconductor. Provedla jsem testy implementovaných zobrazovacích metod a následnou prohlídku odpovídajícího skriptu v jazyce R a konstatovala jsem že:
  - funkce *triplex.diagram* bohužel nezobrazuje korektně složitější struktury triplexů s četným výskytem inzercí/delecí ve vláknech a naprosto selhává při zobrazení některých typů triplexů;
  - skript funkce *triplex.diagram* je koncipován zjevně neefektivně, je nepřehledný a nekomentovaný;
  - funkce *triplex.3D* vykazuje podobné nedostatky, jako funkce *triplex.diagram*.

4.2 Cíle práce

Na základě zhodnocení stavu v řešené oblasti (viz předchozí bod) jsem mohla formulovat cíle této práce podle Zadání.

Hlavním cílem práce je navrhnout obecné algoritmy pro 2D vizualizaci vybraných sekundárních struktur DNA, palindromů a triplexů, implementovat tyto algoritmy v prostředí R/Bioconductor a ověřit funkčnost algoritmů na reálných sekvencích DNA.

Dílčí, postupné cíle práce pak jsou:

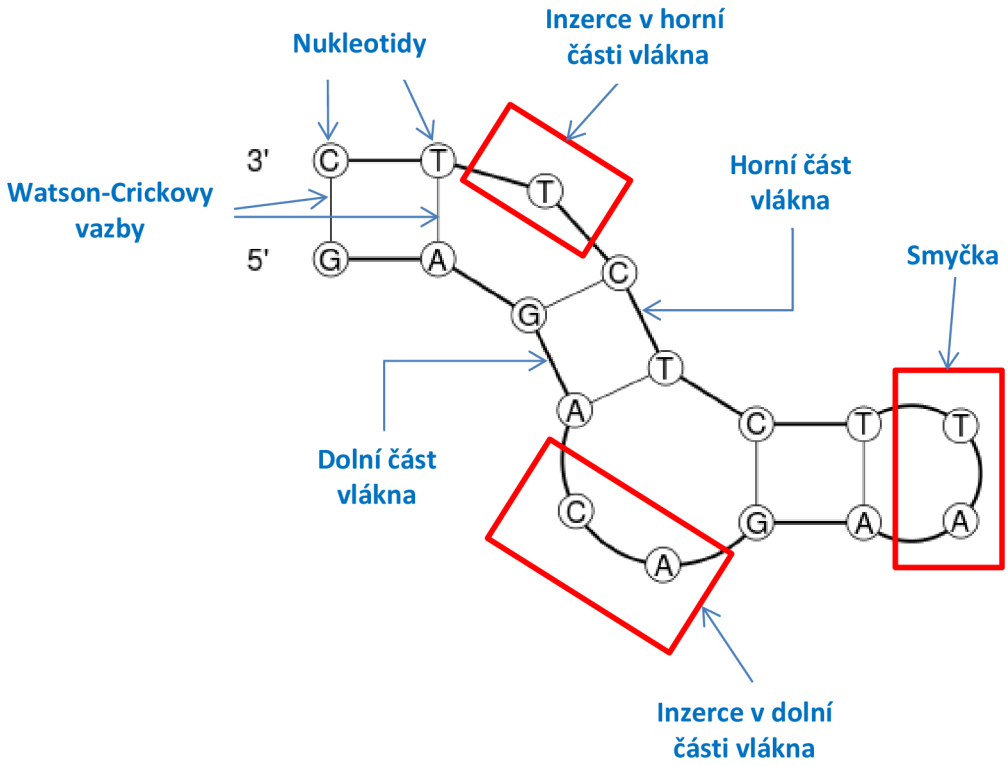
- studium sekundárních struktur DNA s důrazem na zájmové struktury;
- přehled přístupů k vizualizaci struktur DNA;
- studium prostředí R/Bioconductor s důrazem na jeho možnosti vizualizace 2D struktur DNA;
- návrh obecného algoritmu pro sestavení 2D vizualizace zájmových struktur DNA a modelování jejich charakteristických entit;
- implementace navržených algoritmů v prostředí R/Bioconductor, tj. vytvoření funkcí *triplex.diagam* pro balíček *triplex* a *palindrome.diagram* pro balíček *palindrome*;
- ověření implementovaných algoritmů na reálných sekvencích DNA.

Hlavním aplikačním přínosem práce je vytvoření základní zobrazovací podpory balíčků *triplex* a *palindrome*.

5 Návrh algoritmu pro vizualizaci sekundárních struktur DNA

5.1 Palindrom

Cílem algoritmu pro sestavení 2D vizualizace palindromu je určení pozic jednotlivých nukleotidů dané sekvence DNA v 2D diagramu s tím, že se musí zohlednit všechny charakteristické entity dané struktury. Uspořádaná množina těchto pozic, svázaná s dalšími atributy popisu entit, je podkladem pro vytvoření grafické prezentace struktury.

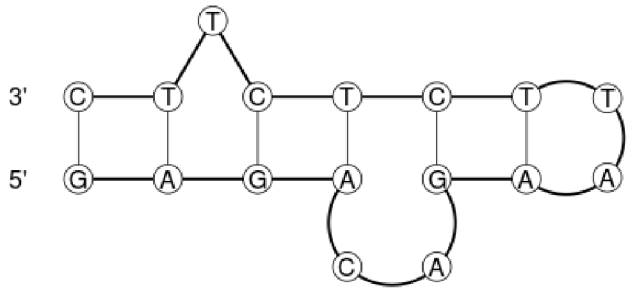


Obr. 15: Entity 2D diagramu palindromu. Modelový diagram vygenerovaný pomocí zpracovaného programu uvažuje deformace tvaru vláken.

Hlavní entity 2D diagramu palindromu, které ve svém algoritmu modeluji, jsou (viz obr. 12):

- **Vlákno DNA** je tvořeno posloupností nukleotidů. V modelu používám termín *horní* a *dolní* část vlákna. Palindrom je tvořen jedním vláknem DNA, jehož nukleotidy se v určitých oblastech spojily pomocí Watson-Crickových vazeb a zformovaly tak ostatní charakteristické entity palindromu;
- **Smyčka** vzniká ohnutím vlákna DNA při formování palindromu. Ve smyčce se nachází jeden nebo více nespárovaných nukleotidů;

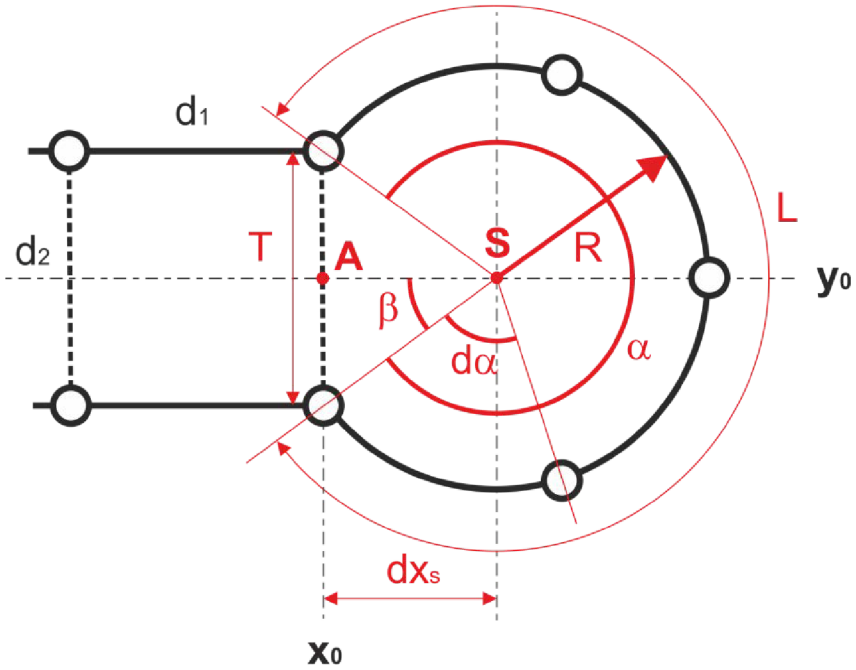
- **Inzerce** jsou oblasti palindromu, kde nedošlo ke spárování nukleotidů z horní a dolní části vlákna pomocí Watson-Crickových vazeb. Část vlákna, která tyto nadbytečné nukleotidy obsahuje, způsobuje napětí ve vláknu palindromu a tím deformace jeho tvaru.



Obr. 16: Modelový diagram palindromu, kde se neuvažují deformace vlákna vlivem inzercí. Jedná se o stejnou strukturu, která je zobrazena na obr. 12.

Model smyčky

Princip modelování smyčky palindromu tak, jak jsem jej pro účely této práce navrhla, je zobrazen na obr. 14. Smyčka samotná je tvořena kruhovým obloukem délky L, na kterém je rozmístěno n nukleotidů v pravidelných intervalech tak, že oblouková vzdálenost mezi nimi je rovna modelové vzdálenosti nukleotidů ve vláknu DNA (d_1), tj. $L = (n-1)\,d_1$. Velikost tětivy oblouku smyčky (T) odpovídá modelové vzdálenosti Watson-Crickových vazeb mezi nukleotidy (d_2), tj. $T = d_2$.



Obr. 17: Popis modelu smyčky palindromu.

Vyjádřením délky oblouku (L) a délky tětivy (T) jako funkce vrcholového úhlu oblouku dostanu, po triviálních úpravách, rovnici (1).

$$f(\alpha): sin(\alpha) - \frac{T}{2\,L}\,\alpha = 0; \tag{1}$$

Vyřešení rovnice (1), tj. určení velikosti vrcholového úhlu oblouku smyčky, je nezbytné pro výpočet jeho ostatních parametrů a následně také pozic nukleotidů, které na něm leží. Rovnice (1) však nemá exaktní řešení a musí se řešit numericky. Pro tento účel jsem implementovala iterační algoritmus, který vypočítává vrcholový úhel kruhového oblouku při známé délce oblouku (L) a délce tětivy (T), založený na Newtonově iterační metodě (metodě tečen) [25]. Podle rekurentní rovnice (2) se provádí opravy argumentu v rekurzivně volané funkci tak dlouho, dokud není dosaženo požadované přesnosti.

$$\alpha_{k+1} = \alpha_k - \frac{f(\alpha)}{f'(\alpha)} = \frac{T\,\alpha - 2\,L\,\sin\left(\frac{\alpha}{2}\right)}{T - L\,\cos\left(\frac{\alpha}{2}\right)}; \tag{2}$$

Konvergence metody je sice citlivá na zvolenou hodnotu počáteční hodnoty, nicméně v tomto případě konverguje velice rychle a spolehlivě při zvolené počáteční hodnotě $\alpha_0 = 6$. Pro výpočet, kdy výsledná hodnota splňuje podmínku $f(\alpha) \leq 10^{-12}$, postačuje maximálně 6 iterací pro široký rozsah[1] počtu nukleotidů ve smyčce palindromu.

Dále již mohu určit zbytek parametrů nutných pro sestavení 2D diagramu smyčky palindromu, tj. poloměr (R) a souřadnice středu kruhového oblouku (x_s, y_s).

$$R = \frac{T}{2\,\sin\left(\frac{\alpha}{2}\right)} = \frac{d_1}{2\,\sin\left(\frac{\alpha}{2}\right)}; \tag{3}$$

$$x_S = x_0 - dx_S = x_0 - R\,\cos\left(\frac{\alpha}{2}\right); \qquad y_S = y_0; \tag{4}$$

Souřadnice n nukleotidů, umístěných na smyčce palindromu pak určím pomocí parametrických rovnic kruhového oblouku.

$$x_i = x_S + R\,\cos(\pi + \beta + i\,d\alpha);\;\; i\epsilon(1,n) \tag{5}$$

$$y_i = y_S + R\,\sin(\pi + \beta + i\,d\alpha);\;\; i\epsilon(1,n) \tag{6}$$

Kde úhly $\beta = \sin^{-1}\left(\frac{T}{2\,R}\right)$ a $d\alpha = \frac{\alpha}{n+1}$ a kde n je počet nukleotidů ve smyčce palindromu.

[1] Testovány byly počty nukleotidů ve smyčce od řádu jednotek až do řádu stovek.

Uvedená konstrukce smyčky palindromu platí pro případ, kdy není uvažována deformace vláken palindromu účinkem inzercí. Pokud je deformace vláken uvažována, je nutné na závěr konstrukce diagramu palindromu smyčku transformovat (posunout a rotovat) tak, aby kotevní bod smyčky $A(x_0, y_0)$ ležel na středu spojnice koncových nukleotidů horního a dolního vlákna a osa oblouku byla shodná s úhlem osy vláken v oblasti, kde je smyčka napojena.

Model inzerce ve vláknu palindromu

Inzerce ve vláknu palindromu vzniká, jsou-li ve vláknu DNA zařazeny nadbytečné nukleotidy, které nejsou spárovány s nukleotidy protější části vlákna. Ve skutečném vláknu tak vzniká napětí, které jej deformuje, tj. mění další směr vlákna v prostoru. Tyto jevy lze při 2D vizualizaci sekundárních struktur DNA zohlednit jen symbolicky. Věrnější modelování těchto jevů je možné pouze v případech 3D modelů sekundárních struktur DNA.

Model inzerce deformující tvar vlákna

V algoritmu konstrukce 2D diagramu palindromu jsem navrhla model symbolického znázornění deformací vlákna vlivem inzercí, které umožňuje názorné vizuální ocenění počtu inzercí ve vláknu a jejich vlivu na prostorové uspořádání palindromu. Modelové uspořádání případu inzerce, kde se uvažuje deformace vláken je uvedeno na obr. 15. Změna směru vlákna v inzerci je odvozena od prodloužení vzdálenosti mezi hraničními body inzerce, které jsou pevně vázány Watson-Crickovými vazbami, v jedné z částí vlákna palindromu (horní nebo dolní). Z hlediska mého modelu jde tedy o problém pouze geometrický a úhel natočení vlákna vždy odpovídá inzerci s jedním nadbytečným nukleotidem (viz např. inzerce v horním vláknu na obr. 12). Pro případy, kdy je v inzerci více nukleotidů, vytvářím další oblouk nad prodlouženou částí vlákna, na kterém jsou tyto nukleotidy umístěny (viz např. inzerce v dolním vláknu na obr. 12).

Hraniční body inzerce (viz obr. 15) tvoří lichoběžník $P_0\,P_1\,L_1\,L_0$, který určuje parametry rotace vlákna. Úhel rotace vlákna α_1 lze určit např. z pravoúhlého trojúhelníku $S_1\,A\,P_0$.

$$\alpha_1 = 2\,\beta = 2\,\sin^{-1}\left(\frac{d_1}{2\,d_2}\right) \tag{7}$$

Souřadnice středu rotace koncových bodů inzerce se pak určí podle rovnic (8) a (9), kde se zohledňuje aktuální úhel natočení vláken (ϑ) a vlákno, kde inzerce vzniká.

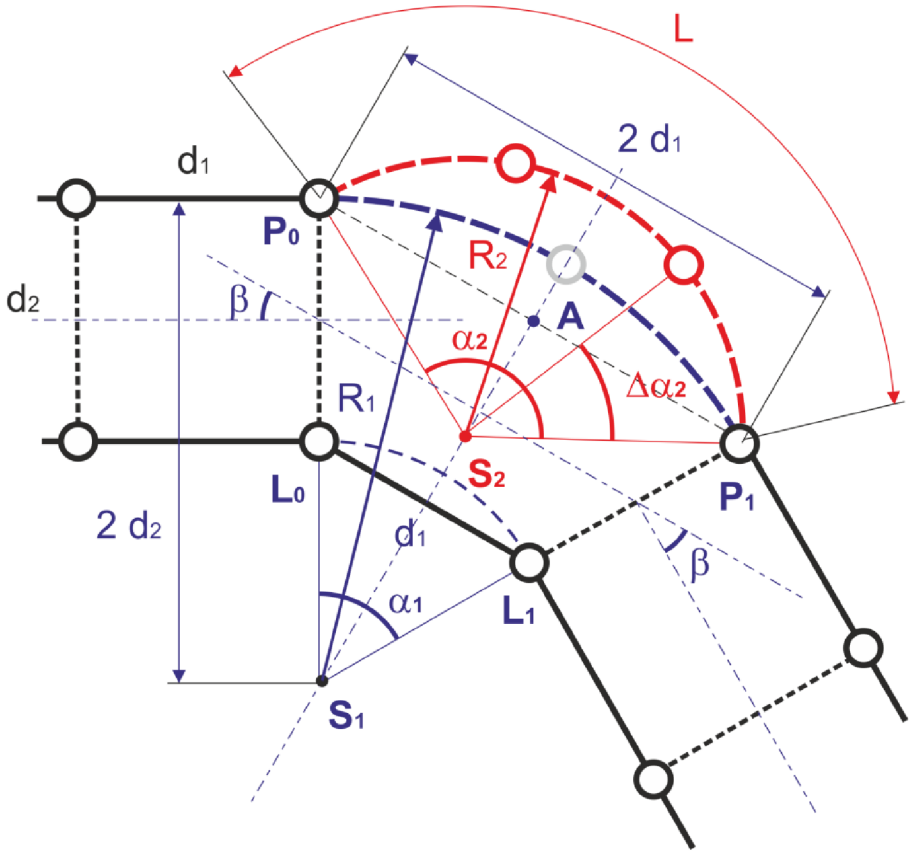
$$x_{S1} = x_0 \pm 2\,d_2\,\sin(-\vartheta) \tag{8}$$

$$y_{S1} = y_0 \pm 2\,d_2\,\cos(-\vartheta) \tag{9}$$

Souřadnice koncových bodů inzerce $P_1\,L_1$ se určí rotací počátečních bodů inzerce $P_0\,L_0$ (viz rovnice (10) a (11)) kolem bodu S_1 o úhel α_1.

$$x_1 = x_{S1} + (x_0 - x_{S1})\,\cos(\alpha_1) - (y_0 - y_{S1})\,\sin(\alpha_1) \tag{10}$$

$$y_1 = y_{S1} + (x_0 - x_{S1})\,\sin(\alpha_1) + (y_0 - y_{S1})\,\cos(\alpha_1) \tag{11}$$



Obr. 18: Model inzerce ve vláknu palindromu. Varianta, kdy inzerce deformují tvar vláken.

Pokud je v inzerci pouze jeden nadbytečný nukleotid, jsou jeho souřadnice určeny obdobnou rotací, avšak pouze o úhel β.

Pokud je v inzerci více nukleotidů, budou umístěny na kruhovém oblouku, který kompenzuje prodloužení vlákna, ve kterém se inzerce nachází. Délka tohoto oblouku je $L = (n + 1)\, d_1$, kde n je počet nukleotidů v inzerci a d_1 je modelová vzdálenost nukleotidů ve vláknu. Délka tětivy oblouku je $T = 2\, d_2$, kde d_2 je modelová vzdálenost mezi nukleotidy ve Watson-Crickově vazbě.

Určení dalších parametrů (vrcholový úhel α_2, souřadnice středu S_2 a poloměr R_2) oblouku známé délky a tětivy přechází na stejnou úlohu, která je již řešena v předchozím bodě (model oblouku smyčky, str. 20), a řeší se tak opět numericky, pomocí Newtonovy iterační metody podle rovnic (1) a (2).

Při sestavení oblouku inzerce, tj. při výpočtu souřadnic jednotlivých nukleotidů inzerce, postupuji tak, že nejdříve sestavím oblouk v pravoúhlé souřadné soustavě, tzn., neuvažuji zde úhel natočení vlákna inzerce. Parametry oblouku určím podle (12) a (13).

$$R_2 = \frac{d_1}{\sin\left(\frac{\alpha_2}{2}\right)}; \tag{12}$$

$$x_{S2} = x_0 + d_1; \quad y_{S2} = y_0 + R_2 \cos\left(\frac{\alpha_2}{2}\right); \tag{13}$$

Poté určím pozice nukleotidů na oblouku inzerce rotací bodu P_0 resp. L_0 podle toho, na které části vlákna palindromu se inzerce nachází, kolem středu rotace S_2 podle rovnic (14) a (15).

$$x_i = x_{S2} + (x_0 - x_{S2})\ \cos(i\ d\alpha_2) - (y_0 - y_{S2})\ \sin(i\ d\alpha_2);\ \ i\epsilon(1,n) \tag{14}$$

$$y_i = y_{S2} + (x_0 - x_{S2})\ \sin(i\ d\alpha_2) + (y_0 - y_{S2})\ \cos(i\ d\alpha_2);\ \ i\epsilon(1,n) \tag{15}$$

Kde úhel $d\alpha_2 = \frac{\alpha_2}{n+1}$ a n je počet nukleotidů v inzerci.

Dále aplikuji na všechny nukleotidy inzerce další rotaci (16) a (17), tentokrát kolem bodu P_0 resp. L_0 o úhel $\vartheta \pm \beta$ podle toho, na které části vlákna palindromu se inzerce nachází. Tím je oblouk (resp. pozice nukleotidů na něm) správně transformován do inzerce příslušné části vlákna.

$$x_i = x_0 + (x_i - x_0)\ \cos(\vartheta \pm \beta) - (y_i - y_0)\ \sin(\vartheta \pm \beta);\ \ i\epsilon(1,n) \tag{16}$$

$$y_i = y_0 + (x_i - x_0)\ \sin(\vartheta \pm \beta) + (y_i - y_0)\ \cos(\vartheta \pm \beta);\ \ i\epsilon(1,n) \tag{17}$$

Zbytek vláken (resp. nukleotidů) za inzercí je rotován do nového směru kolem příslušných koncových bodů rotace, tj. horní vlákno kolem bodu P_1 a dolní vlákno kolem bodu L_1, o úhel $\pm\alpha_1$ (viz rovnice (18) a (19) pro případ horní části vlákna, dolní část vlákna se řeší obdobně, viz (20) a (21)). Současně je upraven nový směr vláken na $\vartheta' = \vartheta \pm \alpha_1$.

$$x'_{Pi} = x_{P1} + (x_{Pi} - x_{P1})\ \cos(-\alpha_1) - (y_{Pi} - y_{P1})\ \sin(-\alpha_1);\tag{18}$$

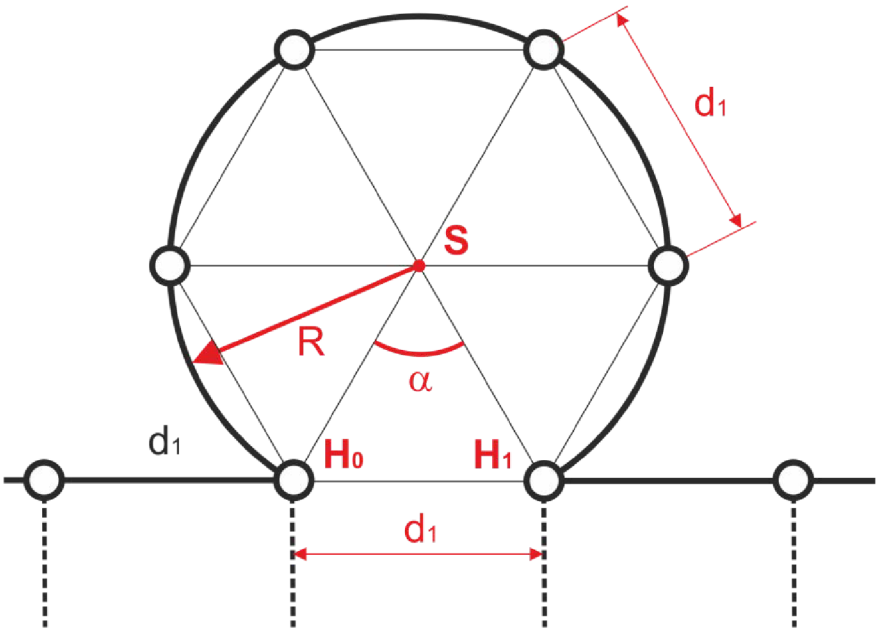
$$y'_{Pi} = y_{P1} + (x_{Pi} - x_{P1})\ \sin(-\alpha_1) + (y_{Pi} - y_{P1})\ \cos(-\alpha_1);\tag{19}$$

$$x'_{Li} = x_{L1} + (x_{Li} - x_{L1})\ \cos(\alpha_1) - (y_{Li} - y_{L1})\ \sin(\alpha_1);\tag{20}$$

$$y'_{Li} = y_{L1} + (x_{Li} - x_{L1})\ \sin(\alpha_1) + (y_{Li} - y_{L1})\ \cos(\alpha_1);\tag{21}$$

Model inzerce nedeformující tvar vlákna

Ve variantě modelu, kdy se nepožaduje modelování deformací vláken inzercemi (viz např. diagram na obr. 13), je použitý model mnohem jednodušší, jak je zřejmé již z jeho náčrtu, který je uveden na obr. 15. Délka vlákna s přebytečnými nukleotidy je zde kompenzována vytvořením kruhové smyčky nad modelovým vláknem. Jednotlivé nukleotidy inzerce, společně s hraničními nukleotidy inzerce (body H_0 a H_1 na obr. 4), zde tvoří pravidelný *(n+2)-úhelník*, kde n je počet nukleotidů v inzerci. Délka strany tohoto pravidelného mnohoúhelníku se rovná modelové vzdálenosti nukleotidů ve vlákně DNA (d_1). Smyčku inzerce pak tvoří kruhový oblouk kružnice opsané tohoto mnohoúhelníku, nebo jednotlivé strany pravidelného mnohoúhelníku (v závislosti na grafické prezentaci modelu).



Obr. 19: Model inzerce palindromu resp. triplexu. Varianta, kdy není modelována deformace tvaru vláken.

Parametry oblouku smyčky inzerce, tj. vrcholový úhel α segmentu, poloměr kruhového oblouku R a souřadnice jeho středu, lze v tomto případě určit pomocí následujících jednoduchých vztahů (22), (23), (24) a (25).

$$\alpha = \frac{2\,\pi}{n+2}; \tag{22}$$

$$R = \frac{d_1}{2\,\sin\left(\frac{\alpha}{2}\right)}; \tag{23}$$

$$x_S = \frac{(x_{H0} + x_{H1})}{2}; \tag{24}$$

$$y_S = y_{H0} + \sqrt{R^2 - \left(\frac{d_1}{2}\right)^2}; \tag{25}$$

Pozice jednotlivých nukleotidů v oblasti inzerce se pak určí pomocí rotace souřadnic pozice hraničního nukleotidu inzerce H_0 kolem středu kruhového oblouku S podle vztahů (26) a (27).

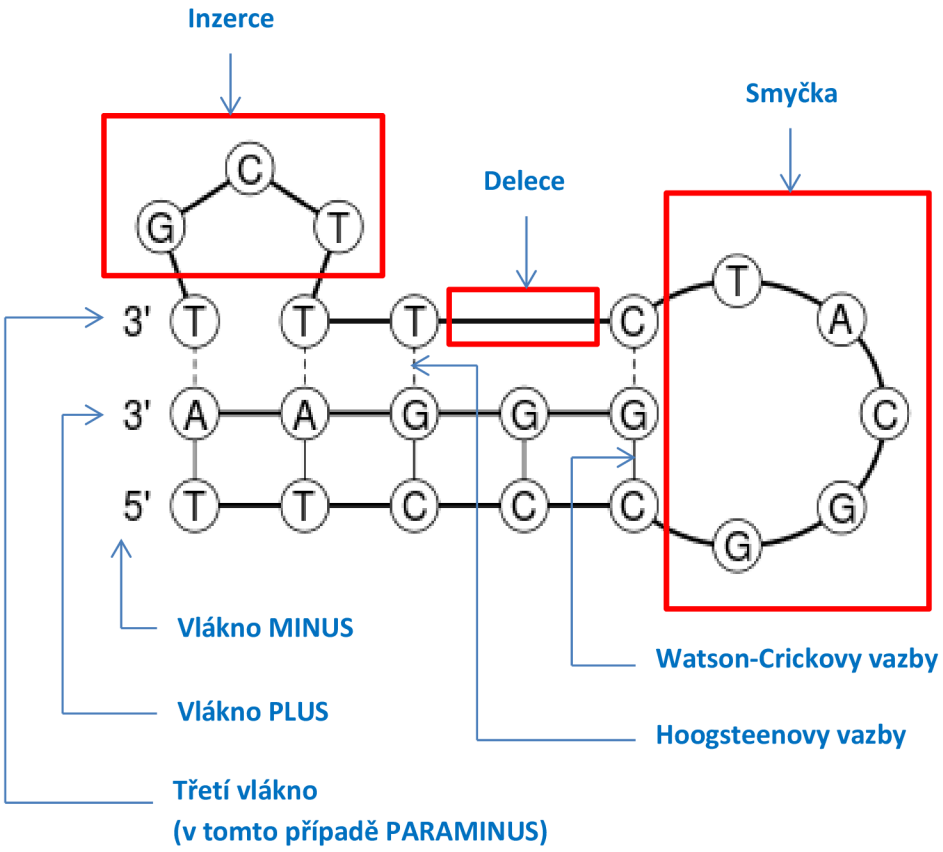
$$x_i = x_S + (x_{H0} - x_S)\,\cos(i\,\alpha) - (y_{H0} - y_S)\,\sin(i\,\alpha);\;\; i\epsilon(1,n) \tag{26}$$

$$y_i = y_S + (x_{H0} - x_S)\,\sin(i\,\alpha) + (y_{H0} - y_S)\,\cos(i\,\alpha);\;\; i\epsilon(1,n) \tag{27}$$

Kde n je počet nukleotidů v inzerci.

5.2 Triplex

Hlavním cílem algoritmu pro sestavení 2D vizualizace triplexu je určení pozic jednotlivých nukleotidů dané sekvence DNA v 2D diagramu s tím, že se zohlední všechny charakteristické entity dané struktury. Uspořádaná množina těchto pozic, svázaná s dalšími atributy popisu jednotlivých charakteristických entit, je podkladem pro následné vytvoření grafické prezentace struktury.



Obr. 20: Entity 2D diagramu triplexu. (Modelový diagram vygenerovaný pomocí zpracovaného programu)

Hlavní entity 2D diagramu triplexu, které ve svém algoritmu modeluji, jsou (viz obr. 17):

- **Vlákna DNA.** Vlákna jsou tvořena posloupností nukleotidů. Konfigurace vláken triplexu závisí na podmínkách při jeho utváření a vzniká tak několik základních typů triplexu (viz bod 3.1). Uvažované triplexy jsou tvořeny dvěma vlákny DNA, které jsou (z části) spojeny Watson-Crickovými vazbami a z nichž jedno tvoří smyčku a jeho část se spojuje s jedním z vláken pomocí Hoogsteenových vazeb;
- **Smyčka** vzniká ohnutím jednoho z vláken DNA při formování triplexu, kdy se část ohnutého vlákna (tzv. třetí vlákno) páruje s jedním z vláken triplexu pomocí Hoogsteenových vazeb. Ve smyčce se nachází jeden nebo více nespárovaných nukleotidů;
- **Inzerce** jsou oblasti ve vláknu triplexu, kde nedošlo ke spárování nukleotidů. Inzerce se kompenzují ve třetím vláknu, tj. v oblasti Hoogsteenových vazeb. Na rozdíl od palindromu (viz 5.1) se neuvažují jiné deformace struktury triplexu vlivem inzercí ve třetím vlákně, než je samotná kompenzace délky vlákna s nepárovanými nukleotidy vytvořením oblouku nad vláknem.

- **Delece** jsou oblasti ve vláknu triplexu, kde chybí nukleotid resp. nukleotidy pro vytvoření vazby s odpovídajícím vláknem. Podobně jako inzerce, se delece kompenzují v třetím vlákně a nedeformují strukturu triplexu.

Model smyčky

Matematický popis a řešení modelu smyčky triplexu je zcela shodný s tím, který jsem vytvořila pro případ palindromu (viz bod Model smyčky, rovnice (1) až (6)) s tím, že podle typu modelovaného triplexu je nutné upravit délku tětivy oblouku. Ta může být, v závislosti na typu triplexu, buď $T = d_2$, nebo $T = 2\ d_2$.

Model inzerce vlákna triplexu

Vzhledem k tomu, že není potřeba modelovat deformace vláken, využívám zde model inzerce, který jsem vytvořila pro model palindromu v případě, kdy se neuvažují deformace vláken, viz str. 24 a odpovídající rovnice (22) až (27).

6 Implementace algoritmu v prostředí R/Bioconductor

6.1 Jazyk R a Bioconductor

R [22][24] je programovací jazyk a vývojové prostředí určené zejména pro statistickou analýzu dat a jejich grafické zobrazení. Jedná se o dialekt jazyka S [22], který původně vznikl v Bellových laboratořích. Jazyk kromě široké škály statistických technik, jako lineární a nelineární modely nebo analýzu časových řad, poskytuje rovněž velmi kvalitní a komplexní podporu pro grafické výstupy, rozhraní pro jiné jazyky resp. systémy a možnosti ladění.

Velkou výhodou jazyka R také je, že se jedná o projekt pod GNU GPL licencí [22][24], a je tak volně šiřitelný. Na jeho využívání se tak orientuje celá řada výzkumných týmů z mnoha oblastí vědy a lze tak předpokládat, že funkce systému odpovídají aktuálnímu stavu poznání v mnoha oblastech vědy. Jazyk R je skriptovací a objektově orientovaný interpret, jehož použití je pro běžnou vědeckou komunitu více akceptovatelné a efektivnější, než použití programovací jazyků na nižší úrovni s nižší mírou abstrakce a uživatelské přívětivosti.

Systém R je jednoduše rozšiřitelný pomocí softwarových knihoven, tzv. R – balíčků [24], které jsou dostupné online prostřednictvím serveru CRAN[2], jehož zrcadlové servery jsou dostupné na celém světě (aktuálně 96 serverů ve 45 zemích).

Silnou stránkou systému R jsou grafické výstupy, což zřejmě souvisí také s jeho původním zaměřením. K dispozici je zde celá řada grafických nástrojů. Většina je zaměřena na tvorbu nejrůznějších grafů, map či různých diagramů, nicméně, je zde i silná podpora tvorby grafických primitiv ve 2D i 3D souřadném systému.

Bioconductor

Rozvoj bioinformatických analýz a modelování si vyžádal vývoj odpovídajících softwarových prostředků. Přibližně v roce 2002 [24] byl zahájen projekt Bioconductor, v rámci kterého se začaly vyvíjet softwarové knihovny umožňující analýzy a modelování genomických dat pro systém R. Na vývoji se podílí mezinárodní tým složený z odborníků předních světových universit a softwarových firem. Bioconductor tedy představuje volně šiřitelný nástroj, který je určen ke zpracování dat z oblasti matematické biologie a bioinformatiky. Distribuován je stejným způsobem, jako ostatní softwarové balíčky systému R, pomocí sítě CRAN.

6.2 Implementace algoritmů

Výše uvedené algoritmy (viz body 5.1 a 5.2) pro zobrazování struktur palindromů a triplexů ve formě 2D diagramů, jsem implementovala pomocí skriptovacího jazyku R jako moduly softwarových balíčků *palindrom* a *triplex*, které rozšiřují možnosti systému R/Bioconductor v oblasti vyhledávání a vizualizace vybraných sekundárních struktur DNA, tj. palindromů a intramolekulárních triplexů.

[2] CRAN – Comprehensive R Archive Network.

Zdrojové soubory modulů *palindrome.diagram.R* resp. *triplex.diagram.R* a *utils.R* jsou součástí programových balíčků *triplex* resp. *palindrome*. Funkčnost 2D vizualizace je dostupná vždy prostřednictvím jedné hlavní funkce, a to *triplex.diagram* resp. *palindrome.diagram*, které jsou exportovány jako veřejné pro uživatele balíčků *triplex* resp. *palindrome* v systému R/Bioconductor (viz Příloha A, soubory *triplex.diagram.help.txt* a *palindrome.diagram.help.txt*).

2D diagram palindromu

Zjednodušený vývojový diagram sestavení 2D vizualizace palindromu je uveden na obr. 18. Vytvořený modul pro 2D vizualizaci palindromu je funkčně integrován se zbytkem balíčku prostřednictvím sdílení dat popisujících nalezené sekundární struktury DNA, které mají být zobrazeny. V tomto případě jde nukleotidovou reprezentaci dolního a horního vlákna palindromu a jeho smyčky. Tyto vstupní informace algoritmu jsou čerpány z modulu *palindrome.alignment.R*.

K uchování těchto vstupních dat, a pro další práci s nimi, slouží třída `CSequence`, kterou jsem definovala v souboru *utils.R*. `CSequence` je tzv. `R5` neboli `Reference Class`. Třída slouží pouze jako kontejner pro data a zahrnuje následující pole:

- `bases` – pole znaků sloužící k uložení původní posloupnosti nukleotidů, která byla získána z funkce *`palindrome.alignment`*;
- `dbases` – pole znaků obsahující upravenou posloupnost `bases` sloužící jako podklad pro vykreslování. Pole `bases` a `dbases` mohou však být úplně identická. Liší se pouze v případech, kdy se ve vlákně palindromu vyskytuje inzerce;
- `n` – původní počet nukleotidů v řetězci `bases`;
- `dn` – počet nukleotidů v poli `dbases`;
- `x` – číselné pole obsahující souřadnice `x` nukleotidů ve vláknu;
- `y` – číselné pole obsahující souřadnice `y` nukleotidů ve vláknu;

V algoritmu vytvářím instanci této třídy vždy pro dolní část vlákna, horní část vlákna a smyčku.

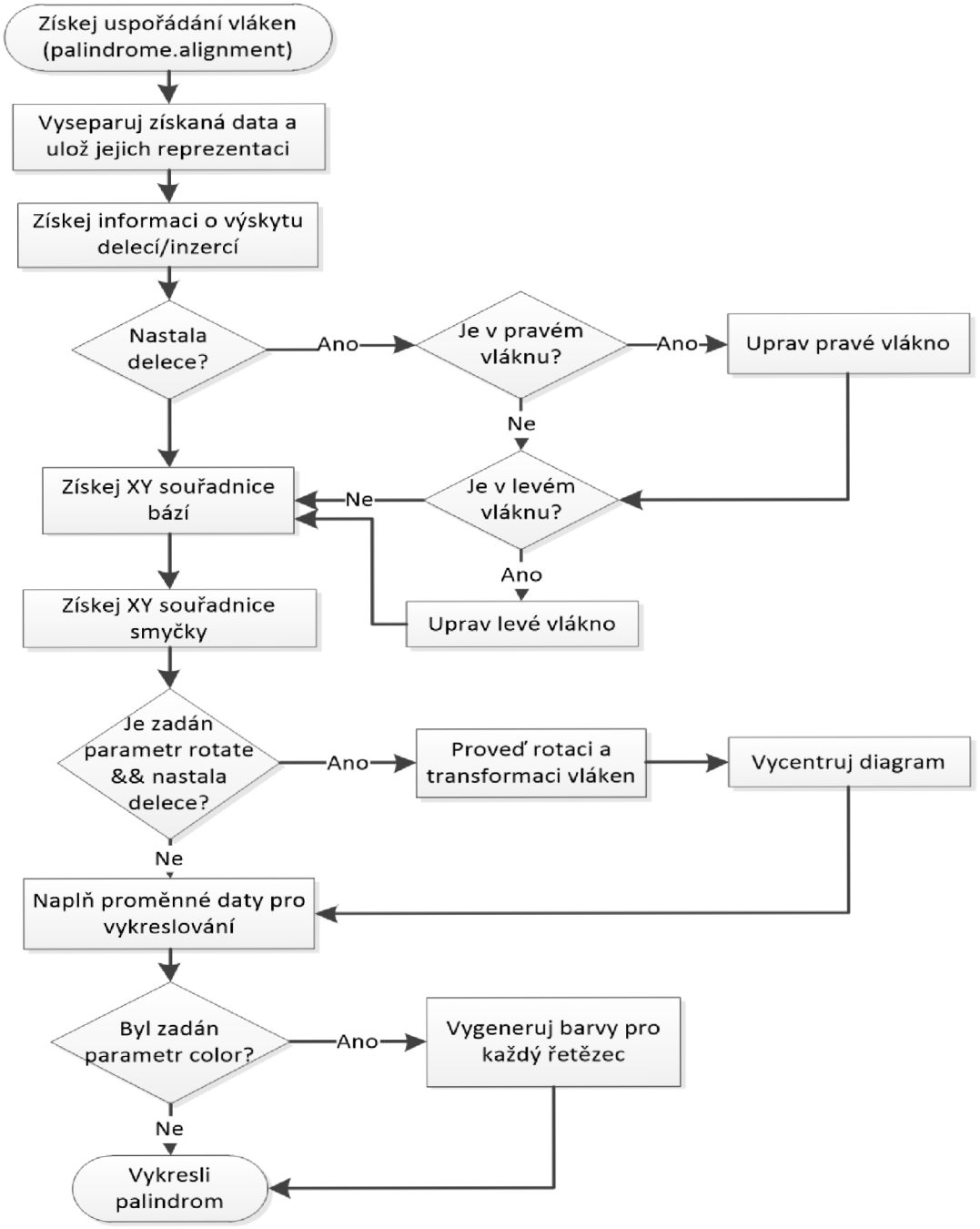
Dále provádím detekci delecí v horním resp. dolním vlákně palindromu. Informace o nalezených delecích a jejich parametrech ukládám v globálních vektorech pro další zpracování. Delece v jedné části vlákna se projeví jako inzerce v části druhé. Musím tedy provést příslušné úpravy v konfiguraci vláken.

V dalším kroku určím, na základě odhadu délky palindromu, modelovou vzdálenost mezi nukleotidy ve vlákně a vygeneruji pozice jednotlivých nukleotidů s uvážením konfigurace všech entit dané struktury. V případě, že je požadována vizualizace znázorňující vliv inzercí na tvar vlákna, provedu navíc konstrukci nezbytných entit a odpovídajících transformací (viz bod 5.1) a vycentruji diagram v kreslícím okně.

Pro 2D grafickou prezentaci dat jsem využila balík `grid`, který umožňuje vytváření grafických objektů (tzv. `grob`), které jsou speciálně určeny pro rychlé a efektivní zobrazování posloupnosti bodů v různých módech, které určují, jaký způsobem mají být jednotlivé body prezentovány. Každý grafický objekt v jednom ze svých slotů obsahuje objekt `gpar`. Nastavením jeho vlastností lze definovat styl, barvu a další vlastnosti, s jakými mají být jednotlivé body vykresleny. V mém případě

používám módy umožňující vykreslení lomených čar (vlákna), vykreslení segmentů (vazby), vykreslení textu (popisy vláken a nukleotidů) a vykreslení kruhů (značky nukleotidů).

V dalším kroku tedy vytvářím potřebné grafické objekty a jejich struktury naplním příslušnými daty popisující způsob jejich vykreslení a daty popisující pozice jednotlivých bodů z příslušných instancí tříd `CSequence`. Na závěr pak grafické objekty vykreslím.



Obr. 21: Vývojový diagram sestavení 2D vizualizace palindromu

2D diagram triplexu

Zjednodušený vývojový diagram sestavení 2D vizualizace triplexu je uveden na obr. 20. Podobně jako v případě palindromu, je modul pro 2D vizualizaci triplexu funkčně integrován se zbytkem balíčku prostřednictvím sdílení dat popisujících nalezené sekundární struktury DNA, které mají být zobrazeny. V tomto případě jde nukleotidovou reprezentaci všech tří vláken triplexu a jeho smyčky a také o typ triplexu (viz bod 3.1). Tyto vstupní informace jsou brány z modulů *triplex.input.R* a *triplex.align.R*.

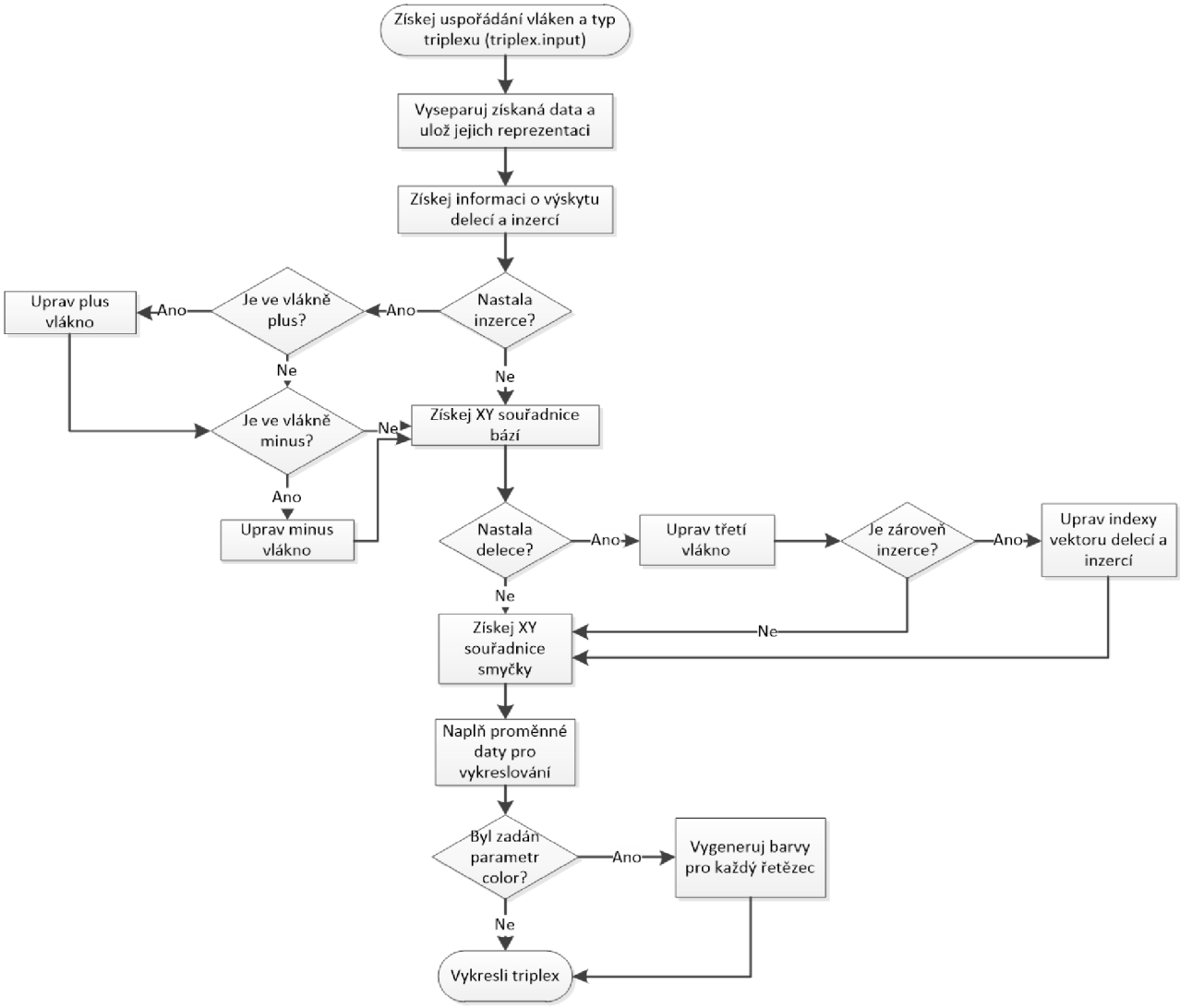
Obdobně, jako v případě palindromu, využívám k uchování těchto vstupních dat, a pro další práci s nimi, instance třídy `CSequence`. Vytvářím instance pro vlákna, plus, minus, loop a třetí vlákno. Na základě informace o typu aktuálního triplexu upravím, podle potřeby, posloupnost nukleotidů ve vláknech.

V dalším kroku provádím detekci inzercí a delecí ve všech vláknech triplexu. Informace o nalezených delecích a inzercích a jejich parametrech ukládám v globálních vektorech pro další zpracování, tj. provedení příslušných úprav v konfiguraci vláken. Výsledné delece resp. inzerce se zobrazují ve třetím vlákně, tj. vlákně, kde jsou nuklidy párovány Hoogsteenovými vazbami.

V dalším kroku určím, na základě odhadu délky triplexu, modelovou vzdálenost mezi nukleotidy ve vlákně a vygeneruji pozice jednotlivých nukleotidů s uvážením konfigurace všech entit dané struktury, které jsou modelovány v souladu s popisem v bodu 5.2.

Pro 2D grafickou prezentaci dat jsem, obdobně jako v případě palindromu, využila grafické objekty (`grob`) z balíku `grid` a využívám objekty pro vykreslení lomených čar (vlákna), vykreslení segmentů (vazby), vykreslení textu (popisy vláken a nukleotidů) a vykreslení kruhů (značky nukleotidů).

Následně tedy vytvořím potřebné grafické objekty a jejich struktury naplním příslušnými daty popisující způsob vykreslení objektů a daty popisující pozice jednotlivých bodů z příslušných instancí tříd `CSequence`. Na závěr pak grafické objekty vykreslím.



Obr. 22: Vývojový diagram sestavení 2D vizualizace triplexu.

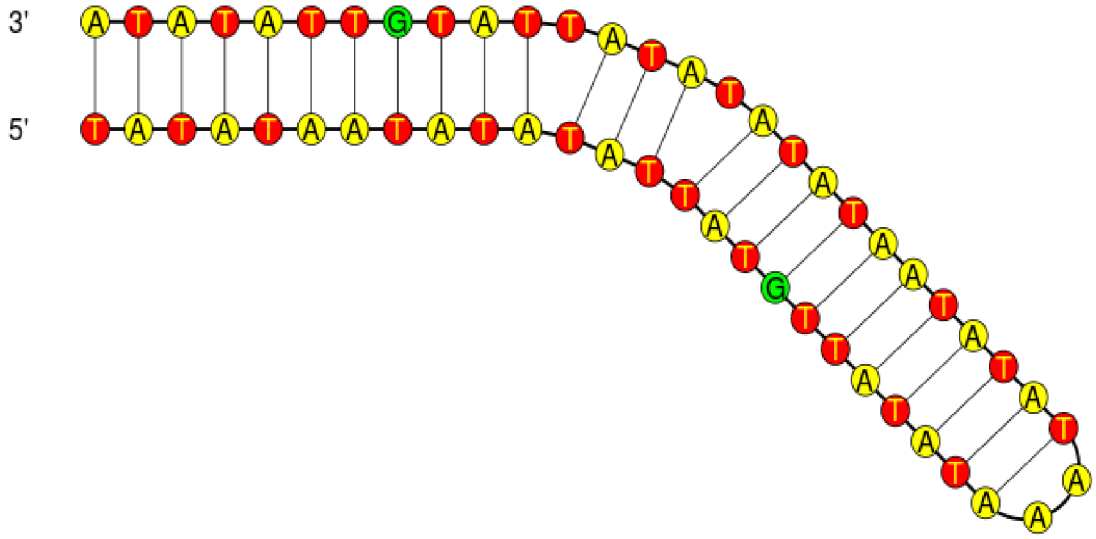
6.3 Příklady vizualizace reálných sekundárních struktur DNA

Palindrom

Příklad vizualizace reálného palindromu je uveden níže, na obr. 19. Struktura byla nalezena ve fragmentu sekvence DNA:

„TTATATATTAATATATAATGTATTATACATATATAATTATATAAATTATGTATGTATTATATAATTATATATAATATAATTATATATAATTGTATATATTAATATATAA“

která pochází z balíku *BSgenome.Hsapiens.UCSC.hg19* [23], který je součástí systému R/Bioconductor.



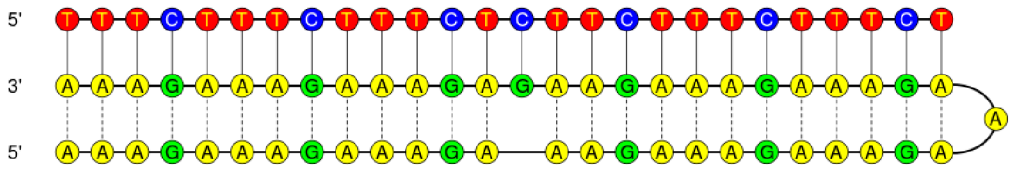
Obr. 23: Vizualizace struktury reálného palindromu, nalezeného ve fragmentu DNA z lidského genomu (*BSgenome.Hsapiens.UCSC.hg19*) [23]. V horní části vlákna se nacházejí dvě inzerce.

Triplex

Příklad vizualizace reálného triplexu je uveden níže, na obr. 21. Struktura byla nalezena pomocí fragmentu sekvence DNA:

„TCTTTCTTCTTTCTTTCTTTCTCTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTCTTTCTTTCTTTTCTCTTTCTCTCTCTCTCTCTTTCTTTTTTTT“

která pochází z balíku *BSgenome.Hsapiens.UCSC.hg19* [23], který je součástí systému R/Bioconductor.



Obr. 24: Vizualizace struktury reálného triplexu, nalezeného ve fragmentu DNA z lidského genomu (*BSgenome.Hsapiens.UCSC.hg19*) [23]. V tzv. třetím vlákně je zřejmá delece.

7 Závěr

Předložená práce se zabývá problémem návrhu algoritmů pro zobrazování vybraných sekundárních struktur DNA, tj. triplexů a palindromů, a jejich implementace v prostředí R/Bioconductor.

V souladu se zadáním práce jsem se zaměřila na oblast 2D grafické reprezentace zájmových struktur ve formě diagramů, které symbolicky zobrazují sekvence nukleotidů, vazby mezi nimi a charakteristické útvary, které mohou v těchto strukturách vznikat. Hlavním aplikačním cílem mé práce bylo doplnění podpory grafické prezentace sekundárních struktur DNA pro ty softwarové balíčky prostředí R/Bioconductor, které se vyhledáváním charakteristických sekvencí palindromů či triplexů v DNA sekvencích zabývají.

Navrhla jsem matematické modely pro konstrukci jednotlivých charakteristických entit zobrazovaných struktur, jako jsou smyčky a inzerce. Pro případ palindromu jsem navrhla způsob modelování symbolické deformace vláken DNA vlivem inzercí. S využitím těchto „stavebních kamenů“ jsem dále navrhla algoritmy pro sestavení 2D diagramů zájmových sekundárních struktur, tj. palindromů a triplexů. Algoritmy jsem implementovala do odpovídajících softwarových balíčků prostředí R/Bioconductor, které jsou pro zkoumání zájmových sekundárních struktur určeny.

Implementované algoritmy jsem ověřovala nejen na modelových, účelově upravených sekvencích DNA, ale také, a to zejména, na reálných sekvencích DNA z lidského genomu. Zde jsem prokázala jejich korektní funkčnost, nicméně předpokládám, že další, velmi důležitá, bude fáze dlouhodobějšího testování po uvolnění odpovídající verze programových balíčků R/Bioconductor.

Cíle práce tak, jak byly stanoveny a uvedeny v úvodní části mé práce (viz bod 4.2), byly splněny. V současné době prostředí R/Bioconductor disponuje nástrojem pro 2D symbolickou vizualizaci nalezených palindromů a triplexů DNA, které poskytují zřetelnou informaci o hledané sekundární struktuře DNA.

Další výzvou, která by měla být při pokračování projektu řešena, je implementace podpory pro 3D zobrazování sekundárních struktur DNA, která umožní analytikům lépe ocenit prostorovou konfiguraci struktury, což je při některých typech bioinformatických analýz důležité.

Literatura

[1] ŠTÍPEK, Stanislav. *Stručná biochemie: Uchování a exprese genetické informace* [online]. 1. vydání. Medprint, 1998. 92 s. ISBN 80-902036-2-0. Dostupné z: http://www.wikiskripta.eu /index.php/Sekund%C3%A1rn%C3%AD_struktura_DNA

[2] SINDEN, R., R.. *DNA structure and function*. San Diego: Academic Press, 1994.

[3] LEXA, M., MARTÍNEK, T., BURGETOVÁ, I., KOPEČEK, D., BRÁZDOVÁ, M.. A dynamic programming algorithm for identification of triplex-forming sequences. *Bioinformatics* [online]. Oxford University Press: 2011. vol. 27, no. 18 2011, p. 2510-2517 [cit. 27.01.2014]. Dostupné z: http://www.fi.muni.cz/~lexa/triplex/

[4] *Genetická informace a její exprese.* [online]. Masarykova Univerzita [cit. 5. 11. 2013]. Dostupné z: http://fybiomuni.wz.cz/Biostranky/download/03_Geneticka_informace_a_jeji _exprese. pdf

[5] *Palindromic sequence*. Wikipedia.org [online]. [cit. 15.10.2013]. Dostupné z: http://en.wikipedia.org/wiki/Palindromic_sequence

[6] KOČÁREK, Eduard. *Molekulární biologie v medicíně*. NCO NZO, 1. vyd., 2007. Str. 33-37. ISBN 978-80-7013-450-4

[7] NOLAN, Jim. *DNA Structure* [online]. Tulane University, 2001. [cit. 13.12.2013]. Dostupné z: http://www.tulane.edu/~biochem/nolan/lectures/rna/DNAstruc2001.htm

[8] POTAMAN, Vladimir, N. and SINDEN, Richard R.. DNA: Alternative Conformations and Biology. In: *Madame Curie Bioscience Database* [online], 2000. [cit. 10.11.2013]. Dostupné z: http://www.ncbi.nlm.nih.gov/books/NBK6545/

[9] SOYFER, Valery N. and POTAMAN, Vladimir, N. *Triple-Helical Nucleic Acids*. New York: Springer, 1996. ISBN 978-0-387-94495-1

[10] FRANK-KAMENETSKII, Maxim, D. and MIRKIN, Sergei, M.. Triplex DNA structures. In: *Annu. Rev. Biochem.* Anuual Reviews, 1995. p. 65-95.

[11] *The mfold Web Server* [online]. New York: The RNA Institute, NY State University at Albany, 2014. [cit. 2.5.2014]. Dostupné z: http://mfold.rna.albany.edu

[12] KING, Robert, C., STANSFIELD, William, D. and MULLIGAN, Pamela, K.. *A Dictionary of Genetics., Seventh Edition*. Oxford University Press, 2006. p.126-185

[13] VOET, Donald a VOETOVÁ, Judith, G.. *Biochemie*. 1. české vydání, Praha: Victoria Publishing, 1995. ISBN 80-85605-44-9

[14] DICKERSON, R., E.. Definitions and nomenclature of nucleic acid structure components. *Nucleic Acids Research* [online]. Volume 17, Number 5, p. 1797–1803. [cit. 6.11.2013]. Dostupné z: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC317523/?page=1

[15] SNUSTAD, D. a SIMMONS, Michael, J.. *Genetika*. 1. české vydání. Brno: Masarykova univerzita, 2009, 871 s. ISBN 978-802-1048-522.

[16] NIELSEN, Peter, E. *Nová molekula života?* [online]. Scientific American, české vydání, 2008, č. 12, str. 34-39. [cit. 12.11.2013]. Dostupné z: http://www.sciam.cz/files/vydani/SA_12_2008/SA_34_41_2008.pdf

[17] SCUDÉ, C. et al. Ligand-induced formation of Hoogsteen-paired parallel DNA. *Chemistry & Biology*. Vol. 3. Cell Press, 1996. Issue 1, p. 57-65

[18] DOSTÁL, J. a kol. *Lékařská chemie II*. Brno: Masarykova Univerzita, 2005. ISBN 80-210-3789-X

[19] LEDVINA, M., STOKLASOVÁ, A. a CERMAN, J.. *Biochemie pro studující medicíny.* Praha: Karolinum, 2005. ISBN 80-246-0851-0

[20] VAN DYKE, M., W. and NELSON, L., D. Triple helix-interacting proteins and cancer. *OA Molecular Oncology*, 2013. Vol. 01, Number 1, p. 5

[21] HON, J., MARTÍNEK, T., RAJDL, K. and LEXA, M. Triplex: an R/Bioconductor package for identification and visualization of potential intramolecular triplex patterns in DNA sequences. *Bioinformatics* [online]. 2013 Aug 1; 29(15):1900-1.
Dostupné z: doi: 10.1093/bioinformatics/btt299.

[22] *The R Project for Statistical Computing* [online]. Dostupné z: http://www.r-project.org

[23] *Bioconductor – open software for bioinformatics* [online]. Dostupné z: http://bioconductor.org

[24] *Comprehensive R Archive Network* [online]. Dostupné z: http://cran.r-project.org/

[25] BARTSCH, Hans, Jochen. *Matematické vzorce*. 4. vydání. Praha: Academia, 2006. 832 s. ISBN 80-200-1448-9

[26] QI, Zhao-Hui and FAN, Tong-Rang. PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization. *Chemical Physics Letters* [online]. Vol. 442. Issue 4-6, pp. 434-440. 2007. Dostupné z: www.sciencedirect.com

[27] NANDY, A.. A new graphical representation and analysis of DNA sequence structure: I. Methodology and Application to Globin Genes. *Current Science.* Vol. 66, pp. 309–314, 1994

[28] BENUSKOVA, L., KROON, R. W., HAVUKKALA, I. *A Compact 2D Representation and Visualization of Large Symbolic Sequences and Applications for Comparative Genome Studies* [online]. [cit. 26. 1. 2014] Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.98.4674&rep=rep1&type=pdf

[29] ZHANG, Zhu-Jin: *DV-Curve: a novel intuitive tool for visualizing and analyzing DNA sequences* [online]. [cit. 26.1.2014] Dostupné z: http://bioinformatics.oxfordjournals.org/content/early/2009/03/10/bioinformatics.btp130.full.pdf

[30] *Genetika-Biologie; Mutace* [online]. [cit. 14.5.2014] Dostupné z: http://www.genetika-biologie.cz/mutace

[31] IPSER, Jan. *Genetika* [online]. Ústí nad Labem: UJEP, 2006. [cit.17.5.2014] Dostupné z: http://biology.ujep.cz/vyuka/file.php/1/opory/Genetika.pdf

[32] NUSSINOV, R. and JACOBSON, A., B. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Biochemistry* [online] Vol. 77, No. 11, pp. 6309-6313. PNAS, 1980. Dostupné z: http://www.ncbi.nln.nih.gov/pmc/articles/PMC350273

[33] *Obrázek: Konformace DNA* [online]. [cit. 30.4.2014] Dostupné z: http://biostudia.blogspot.cz/2013/08/konformacni-rodiny-dna.html

[34] *Obrázek: Malý a velký žlábek DNA* [online].[cit. 30.4.2014] Dostupné z: www.molbio.upol.cz /stranky/vyuka/BPOL/6.%20Struktura%20nukleovych%20kyselin.pdf

[35] *Obrázek: Nussinov diagram* [online]. [cit. 30.4.2014] Dostupné z: http://rna.bgsu.edu/FR3D/AnalyzedStructures/3B91/

Příloha A

Obsah přiloženého CD

Soubor	Složka	Popis
readme.txt	\	Textový soubor s obsahem CD a definicí vlastního podílu práce v dodaných zdrojových kódech
BP_xcudov00.pdf	\Bakalarska_prace\	Písemná zpráva ve formátu pdf
BP_xcudov00.docx	\Bakalarska_prace\	Písemná zpráva ve zdrojovém tvaru MS Word
install.txt	\Dokumentace\	Popis instalace balíčků triplex a palindrome v prostředí R/Bioconductor
triplex.diagram.help.txt	\Dokumentace\	Návod k použití funkce triplex.diagram
palindrome.diagram.help.txt	\Dokumentace\	Návod k použití funkce palindrome.diagram
triplex.tar.gz	\Zdrojove_kody\	Kompletní programový balíček triplex
palindrome.tar.gz	\Zdrojove_kody\	Kompletní programový balíček palindrome