Załącznik 2A: Autoreferat

1. Imię i Nazwisko.

   Maciej Antczak

2. Posiadane dyplomy, stopnie naukowe z podaniem nazwy, miejsca i roku ich uzyskania oraz tytułu rozprawy doktorskiej.

   2.1. Stopień Doktora Nauk Technicznych w dyscyplinie Informatyka uzyskany dnia 21.05.2013 r. na podstawie rozprawy doktorskiej pt. „Algorytmiczne aspekty modelowania i ewaluacji biomolekuł" obronionej na Wydziale Informatyki Politechniki Poznańskiej, promotor: prof. dr hab. inż. Marta Kasprzak, promotor pomocniczy: dr inż. Piotr Łukasiak. **Wysoka wartość naukowa wyników opublikowanych w rozprawie doktorskiej zadecydowała o jej wyróżnieniu przez Radę Wydziału Informatyki w roku 2014.**

   2.2. Tytuł zawodowy: Magister Inżynier Informatyki uzyskany **z wyróżnieniem** dnia 30.06.2005 r. (specjalność: Inżynieria Oprogramowania) na podstawie pracy dyplomowej pt. „Modelowanie i wyznaczanie niepewnych pozycji min morskich" obronionej na Wydziale Informatyki i Zarządzania Politechniki Poznańskiej, promotor: dr inż. Mikołaj Sobczak.

   2.3. Tytuł zawodowy: Inżynier Informatyki uzyskany dnia 30.08.2003 r. na podstawie pracy dyplomowej pt. „Portal edukacyjny dotyczący nowoczesnych sieci optycznych" obronionej na Wydziale Elektrycznym Politechniki Poznańskiej, promotor: dr inż. Maciej Stroiński,

3. Informacje o dotychczasowym zatrudnieniu w jednostkach naukowych.

   3.1. Zakład Bioinformatyki, Instytut Chemii Bioorganicznej Polskiej Akademii Nauk w Poznaniu, adiunkt zatrudniony obecnie w wymiarze ½ etatu od dnia 01.05.2018.

   3.2. Zakład Teorii Algorytmów i Systemów Programowania, Wydział Informatyki, Europejskie Centrum Bioinformatyki i Genomiki, Politechnika Poznańska, adiunkt zatrudniony obecnie w wymiarze 1 etatu od dnia 01.10.2013.

   3.3. Poznańskie Centrum Superkomputerowo-Sieciowe, projektant i programista oprogramowania oraz członek biochemicznego centrum kompetencji powołanego w ramach Programu Obliczeń Wielkich Wyzwań Nauki i Techniki, od 01.04.2012 do 30.11.2012.

   3.4. Zakład Teorii Algorytmów i Systemów Programowania, Wydział Informatyki, Politechnika Poznańska, asystent zatrudniony w wymiarze 1 etatu od 01.10.2011 do 30.09.2013.

3.5. Politechnika Poznańska, kierownik i lider technologiczny specjalizowanego zespołu opracowującego system pozwalający na przeprowadzenie elektronicznej rekrutacji kandydatów na studia na Politechnice Poznańskiej, od 14.12.2006 do 30.09.2009.

3.6. Zakład Teorii Algorytmów i Systemów Programowania, Wydział Informatyki i Zarządzania, Politechnika Poznańska, asystent zatrudniony w wymiarze ½ etatu od 01.10.2005 do 30.09.2011.

4. Wskazanie osiągnięcia wynikającego z art. 16 ust. 2 ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U. 2016 r. poz. 882 ze zm. w Dz. U. z 2016 r. poz. 1311.):

4.1. Tytuł osiągnięcia naukowego:

Metody obliczeniowe wspomagające modelowanie struktur złożonych cząsteczek RNA.

4.2. Wykaz publikacji stanowiących osiągnięcia naukowego:

Podstawą wniosku habilitacyjnego jest osiągnięcie naukowe, które stanowi cykl **ośmiu** powiązanych tematycznie publikacji, wymienionych poniżej. Prace te powstały w latach od 2014 do 2018 i zostały opublikowane w renomowanych czasopismach z listy JCR, specjalizujących się w zakresie bioinformatyki, informatyki i biologii obliczeniowej. W przypadku wnioskowanego osiągnięcia **sumaryczny impact factor** oraz **całkowita liczba punktów MNiSW** stanowią odpowiednio **29,366** oraz **275**.

[A1] **M. Antczak**, T. Zok, M. Osowiecki, M. Popenda, R.W. Adamiak, M. Szachniuk, 22.08.2018, *RNAfitme*: a webserver for modeling nucleobase and nucleoside residue conformation in fixed-backbone RNA structures, *BMC Bioinformatics*, 19(1), 304 (doi:10.1186/s12859-018-2317-9). **[IF (2017): 2,213; MNiSW: 35, Q1]**

[A2] **M. Antczak**, M. Zablocki, T. Zok, A. Rybarczyk, J. Blazewicz, M. Szachniuk, 09.07.2018, *RNAvista*: a webserver to assess RNA secondary structures with non-canonical base pairs, *Bioinformatics*, 35(1), 152-155 (doi:10.1093/bioinformatics/bty609). [**IF (2017): 5,481; MNiSW: 45, Q1**]

[A3] **M. Antczak**, M. Popenda, T. Zok, M. Zurkowski, R.W. Adamiak, M. Szachniuk, 11.12.2017, New algorithms to represent complex pseudoknotted RNA structures in dot-bracket notation, *Bioinformatics*, 34(8), 1304-1312 (doi:10.1093/bioinformatics/btx783). [**IF (2017): 5,481; MNiSW: 45, Q1**]

[A4] **M. Antczak**, M. Popenda, T. Zok, J. Sarzynska, T. Ratajczak, K. Tomczyk, R.W. Adamiak, M. Szachniuk, 14.10.2016, New functionality of *RNAComposer*: application to shape the axis of miR160 precursor structure, *Acta Biochimica Polonica*, 63(4), 737-744 (doi:10.18388/abp.2016_1329). [**IF (2016): 1,159; MNiSW: 15, Q4**]

[A5] **M. Antczak**, M. Kasprzak, P. Lukasiak, J. Blazewicz, 17.09.2016, Structural alignment of protein descriptors – a combinatorial model, *BMC Bioinformatics*, 17, 383 (doi:10.1186/s12859-016-1237-9). [**IF (2016): 2,448; MNiSW: 35, Q1**]

[A6] T. Zok, **M. Antczak**, M. Riedel, D. Nebel, T. Villmann, P. Lukasiak, J. Blazewicz, M. Szachniuk, 30.09.2015, Building the library of RNA 3D nucleotide conformations using clustering approach, *International Journal of Applied Mathematics*

and Computer Science, 25(3), 689-700 (doi:10.1515/amcs-2015-0050). **[IF (2015): 1,037; MNiSW: 25, Q3]**

[A7] A. Rybarczyk, N. Szostak, **M. Antczak**, T. Zok, M. Popenda, R.W. Adamiak, J. Blazewicz, M. Szachniuk, 02.09.2015, New in silico approach to assessing RNA secondary structures with non-canonical base pairs, *BMC Bioinformatics*, 16, 276 (doi:10.1186/s12859-015-0718-6). **[IF (2015): 2,435; MNiSW: 35, Q1]**

[A8] **M. Antczak**, T. Zok, M. Popenda, P. Lukasiak, R.W. Adamiak, J. Blazewicz, M. Szachniuk, 25.04.2014, *RNApdbee* – a webserver to derive secondary structures from pdb files of knotted and unknotted RNAs, *Nucleic Acids Research*, 42(W1), W368-72 (doi:10.1093/nar/gku330). **[IF (2014): 9,112; MNiSW: 40, Q1]**

Wkład habilitanta w powstanie powyższych publikacji znajdują się w załączniku nr 3. Oświadczenia współautorów potwierdzające ich indywidualny wkład w publikacje wchodzące w skład osiągnięcia naukowego znajdują się w załączniku nr 5.

4.3. Omówienie celu naukowego ww. prac i osiągniętych wyników wraz z omówieniem ich ewentualnego wykorzystania.

4.3.1. Wprowadzenie.

Życie, w pewnym uproszczeniu, to wypadkowa wszystkich oddziaływań zachodzących pomiędzy cząsteczkami biologicznymi w komórce. Nawet niewielka ich dezorganizacja często prowadzi do patologii w funkcjonowaniu komórek. Szybki rozwój metod sekwencjonowania zapoczątkował nową erę przewidywania struktur 3D cząsteczek biologicznych, której celem jest zrozumienia mechanizmów leżących u podstaw biofizycznych i biochemicznych procesów ewolucyjnych. Bioinformatyka, jako interdyscyplinarna dziedzina nauki, wyrosła z potrzeby zastosowania komputerów w celu systematycznej analizy i przechowywania ciągle rozrastającego się wolumenu danych sekwencyjnych i strukturalnych cząsteczek biologicznych [E1]. Cząsteczki RNA są liniowymi polimerami zwijającymi sie w konformację natywną podyktowaną fizycznymi i chemicznymi właściwościami łańcucha nukleotydowego. Odgrywają one kluczową rolę w procesach zachodzących w komórkach poprzez np. katalizowanie reakcji chemicznych oraz kontrolowanie ekspresji genów. W opisie struktur cząsteczek RNA rozróżnia się: (1) strukturę pierwszorzędową reprezentowaną przez liniową sekwencje reszt nukleotydowych, (2) strukturę drugorzędową opisującą lokalnie uporządkowane motywy tworzące się w wyniku wiązań wodorowych zachodzących w cząsteczce, (3) strukturę trzeciorzędową reprezentującą specyficzny kształt przestrzenny cząsteczki, (4) strukturę czwartorzędową, gdzie analizowane są oddziaływania pomiędzy kilkoma jednostkami. Za stabilizację kształtu przestrzennego cząsteczek RNA odpowiadają parowania zasad azotowych. Wyróżnia się dwa główne rodzaje par zasad występujące w cząsteczkach RNA, a mianowicie parowania kanoniczne (Watsona-Cricka), wynikające z zasady komplementarności (A:U/G:C), oraz parowania niekanoniczne. W skład tej drugiej grupy wchodzą parowania niekanoniczne typu *wobble* (G:U) oraz pozostałe parowania niekanoniczne, które zwykle nazywane są oddziaływaniami przez przestrzeń lub po prostu przestrzennymi, często dalekiego zasięgu. Przykładowo w przypadku cząsteczek rRNA parowania kanoniczne i niekanoniczne typu *wobble* stanowią zaledwie 71,3%. Resztę stanowią oddziaływania przestrzenne [E2], których występowanie w strukturze 3D RNA jest kluczowe dla zagwarantowania wielu funkcji biologicznych różnych cząsteczek RNA. Kolejnymi charakterystycznymi motywami przestrzennymi, które odgrywają istotną rolę w różnorodnych procesach biologicznych są pseudowęzły.

W ogólności pseudowęzeł [E3] określa motyw, na który składają się przynajmniej dwa parowania (*i, i'*) oraz (*j, j'*) pomiędzy, którymi zachodzi jedna z relacji zaprezentowanych na Rysunku 1.

Rysunek 1. Przykładowy pseudowęzeł zaprezentowany z wykorzystaniem diagramów łukowych.

Struktury RNA, w których występują oddziaływania przestrzenne, tworzące np. rekurencyjne motywy [E4], albo pseudowęzły zwykle określane są jako złożone. Odkrywanie struktur przestrzennych cząsteczek RNA przyczynia się do wyjaśniania pełnionych przez nie funkcji biologicznych, co jest kluczowe, m. in., podczas projektowania nowych leków i rozwiązań terapeutycznych. Określanie struktur 3D RNA stało sie możliwe dzięki rozwojowi metod eksperymentalnych np. krystalografii rentgenowskiej czy spektroskopii jądrowego rezonansu magnetycznego. Niestety w praktyce bardzo często okazuje się, że eksperymentalne określanie struktury 3D RNA jest wyzwaniem wymagającym oraz kosztownym, które nie zawsze prowadzi do upragnionego celu [E5]. Ograniczenia metod eksperymentalnych oraz naturalna dynamika cząsteczek RNA powodują, że ciągle poszerza się przepaść pomiędzy liczbą znanych sekwencji reszt nukleotydowych RNA, a niewielkim ich odsetkiem, dla których eksperymentalnie określona została struktura 3D, zdeponowana w *Protein Data Bank* (*PDB*) [E6]. W związku z powyższym istnieje ogromne zapotrzebowanie na wydajne metody obliczeniowe pozwalające przewidywać *in silico* struktury 3D RNA, bliskie eksperymentalnie określonym strukturom natywnym, których skuteczność została potwierdzona na podstawie wiarygodnych eksperymentów przeprowadzonych przez niezależnych ekspertów. W przypadku modelowania 3D RNA tego typu inicjatywą jest międzynarodowy konkurs o nazwie *RNA-Puzzles* [E7, E8, O3]. W ramach każdego konkursowego wyzwania procedura jest następująca: (1) krystalografowie udostępniają sekwencję cząsteczki organizatorom konkursu, której struktura 3D jest dopiero przygotowywana do publikacji, więc nie jest jeszcze znana szerszej społeczności naukowej, (2) organizatorzy przekazują sekwencję grupom przeprowadzającym modelowanie *in silico* zarówno w kategorii dedykowanej w pełni automatycznym serwerom oraz predykcji manualnej dokonywanej przez ekspertów i w rezultacie otrzymują zbiory modeli 3D, (3) zgłoszone modele 3D są kompleksowo oceniane w kontekście odpowiadających im referencyjnych, eksperymentalnie określonych struktur 3D cząsteczek przez organizatorów, (4) uzyskane wyniki ewaluacji są prezentowane społeczności naukowej w celu motywowania uczestników do dalszych wysiłków na rzecz udoskonalania istniejących rozwiązań. Obecnie istnieje wiele różnych metod obliczeniowych (np. *iFoldRNA* [E9], *MC-Fold/MC-Sym* [E10], *ModeRNA* [E11], *RNAComposer* [O10, O8, A4], *Rosetta RNA* [E12], *SimRNA* [E13], *Vfold* [E14]), które można zastosować do przewidywania struktur 3D RNA. W ostatnich, opublikowanych dwóch rundach konkursu *RNA-Puzzles* [E8, O3] potwierdzono, że: (1) większość metod całkiem dobrze radzi sobie z przewidywaniem kanonicznych par zasad, (2) jakość modeli 3D zgłoszonych dla cząsteczek RNA, dla których z powodzeniem dało się zastosować modelowanie homologiczne i wiedzę ekspercką jest zauważalnie wyższa aniżeli w przypadku pozostałych cząsteczek RNA, (3) wiarygodne modelowanie niekanonicznych par zasad nadal stanowi duże wyzwanie, oraz (4) niektóre modele 3D

RNA występujące wysoko w rankingu, względem miary *RMSD* [E15], zawierają dużą liczbę zderzeń sterycznych. Poruszane cele badawcze są między innymi odpowiedzią na problemy raportowane w artykułach podsumowujących kolejne rundy konkursu *RNA-Puzzles* np. wprowadzenie mechanizmów pozwalających ekspertom wpływać na proces predykcji *in silico* struktur 3D RNA, realizowany w ramach systemu *RNAComposer*, udostępnienie serwera obliczeniowego pozwalającego, między innymi, na przewidywanie niekanonicznych par zasad na podstawie dostarczonej sekwencji reszt nukleotydowych RNA, udostępnienie serwera obliczeniowego pozwalającego neutralizować nieprawidłowe steryczne zderzenia poprzez predykcję konformacji zasad azotowych gwarantując, że struktura 3D szkieletu fosforo-cukrowego pozostanie niezmieniona. Kolejnym problemem jest precyzyjne porównywanie realizowane np. w przypadku analizy skupień przeprowadzonej w obrębie zbioru dostępnych modeli 3D RNA, które często wymaga ekstrakcji wiarygodnej struktury drugorzędowej na podstawie struktury 3D RNA. Niestety tą wiarygodność ciężko jest jednoznacznie zagwarantować, ponieważ topologiczna różnorodność pseudowęzłów powoduje, że ich identyfikacja i klasyfikacja wciąż stanowi wyzwanie, co w dalszej perspektywie utrudnia modelowanie struktur 3D złożonych cząsteczek RNA. Ostatni rezultat skojarzony z efektywnym strukturalnym dopasowaniem motywów 3D o specyficznej topologii, zwanych deskryptorami, zaproponowanymi do opisu struktur 3D białek, może ze względu na swój potencjał wspomóc również analizę struktur 3D RNA.

4.3.2. Cele badawcze.

W ramach prezentowanych badań: (1) określono i opracowano zbiór nowych mechanizmów, udostępnionych w ramach systemu *RNAComposer*, pozwalających ekspertowi skutecznie wpływać na proces modelowania struktur 3D złożonych cząsteczek RNA poprzez aplikowanie wiedzy pochodzącej z literatury lub danych eksperymentalnych, (2) opracowano nowy algorytm wyboru elementów strukturalnych, z repozytorium *RNA FRABABE* [O10, E16], wykorzystywany podczas konstruowania modelu nr 1, którego celem jest maksymalizacja liczby elementów strukturalnych pochodzących z możliwie najmniejszej liczby eksperymentalnie określonych cząsteczek RNA, (3) sformułowano pojęcie rzędu pozwalającego klasyfikować pseudowęzły występujące w cząsteczkach RNA, (4) opracowano nową metodę, udostępnioną w postaci serwera obliczeniowego, pozwalającą na ekstrakcję struktury drugorzędowej złożonych cząsteczek RNA prezentowanej zarówno w postaci tekstowej jak i graficznej, która jako pierwsza pozwalała identyfikować pseudowęzły o rzędach większych od 1, (5) opracowano algorytm określający rzędowość pseudowęzłów w RNA, (6) opracowano rozszerzenie notacji kropkowo-nawiasowej pozwalające na udostępnianie informacji o rzędowości pseudowęzłów w RNA, (7) opracowano rozszerzenia wybranych podejść, wykorzystywanych do wizualizacji struktury drugorzędowej RNA, pozwalające na kolorystyczne prezentowanie rzędowości pseudowęzłów w RNA, (8) opracowano nowe algorytmy dedykowane problemowi kodowania struktury drugorzędowej, w rozszerzonej notacji kropkowo-nawiasowej, złożonych struktur 3D RNA zawierających pseudowęzły wyższego rzędu, (9) zaproponowano nową funkcję kryterialną stosowaną do oceny rozwiązań problemu wspomnianego w poprzednim punkcie, (10) opracowano serwer obliczeniowy dedykowany problemowi przewidywania rozszerzonej struktury drugorzędowej, obejmującej niekanoniczne parowania zasad, na podstawie sekwencji lub kanonicznej struktury drugorzędowej RNA, (11) opracowano nową metodę, udostępnioną w postaci serwera obliczeniowego, pozwalającą na rekonstrukcję pełnoatomowej konformacji cząsteczki RNA charakteryzującej się minimalną wartością

całkowitej energii oddziaływań międzyatomowych na podstawie wprowadzonej sekwencji RNA oraz utrwalonej struktury 3D szkieletu fosforo-cukrowego cząsteczki RNA, (12) opracowano protokół konstrukcji bibliotek reprezentatywnych konformacji reszt nukleotydowych wykorzystujący hierarchiczną analizę skupień w przestrzeni kątów torsyjnych, (13) opracowano algorytm przeszukiwania biblioteki w celu identyfikacji oraz rekonstrukcji rokujących poprawność konformacji reszt nukleotydowych w ramach rozpatrywanej struktury 3D szkieletu fosforo-cukrowego cząsteczki RNA, (14) opracowano wielowarstwową procedurę pozwalającą ewaluować zrekonstruowane konformacje reszt nukleotydowych w celu odfiltrowania tych, które z pewnością nie wejdą w skład rozwiązania charakteryzującego się minimum energetycznym, (15) opracowano algorytm dokładny, którego celem jest przeszukiwanie otoczenia przestrzennego RNA, reprezentowanego w postaci grafu, w celu identyfikacji konformacji o minimalnej energii dla podzbioru reszt zlokalizowanych w eksplorowanym otoczeniu, (16) opracowano nowy model kombinatoryczny bazujący na problemie przydziału o maksymalnym rozmiarze (ang. *maximum-size assignment problem*) oraz szereg nowych algorytmów wielomianowych, opartych na metodzie węgierskiej, oraz algorytm dokładny dla problemu strukturalnego dopasowania pary deskryptorów struktur 3D białek/RNA. W/w badania są nie tylko kontynuacją, ale również twórczym rozwinięciem prac zapoczątkowanych przed uzyskaniem stopnia doktora, w ramach których opracowano metody obliczeniowe i narzędzia dedykowane problemom modelowania i analizy jakości struktur 3D cząsteczek biologicznych, a w szczególności RNA.

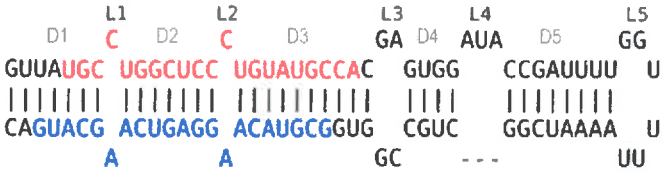
4.3.3. Rezultaty.

4.3.3.1. Udokładnione modelowanie 3D RNA z wykorzystaniem wiedzy eksperckiej.

Podczas opracowywania koncepcji badań dysponowano obliczeniowo efektywnym systemem, o nazwie *RNAComposer* [O10, O8], który pozwalał użytkownikowi, w sposób w pełni automatyczny, przewidywać *in silico* struktury 3D RNA o rozmiarze nie przekraczającym 500nts. Sugestie niektórych badaczy, będących jednocześnie użytkownikami systemu, uzupełnione własnymi przemyśleniami, wynikającymi z czynnego uczestnictwa w procesie modelowania, często homologicznego, struktur 3D RNA w konkursie *RNA-Puzzles* [E8], okazały się motywacją do opracowania nowych mechanizmów, które eksperci mogliby wykorzystywać w celu skutecznego udokładniania procesu modelowania struktur 3D złożonych cząsteczek RNA poprzez wprowadzanie dodatkowych restrykcji strukturalnych wynikających z wiedzy eksperckiej, literatury lub danych eksperymentalnych [A4]. Zbiór narzędzi do predykcji struktur drugorzędowych RNA, udostępnianych przez system, został rozbudowany o *CentroidFold* [E17], *ContextFold* [E18], ponieważ znajdują się one w czołówce rankingów publikowanych na serwerze *CompaRNA* [E19], oraz *IPknot* [E20], który może zostać zastosowany przez użytkownika spodziewającego się pseudowęzłów w wejściowej cząsteczce RNA. Interfejs pozwalający na wprowadzanie dodatkowych restrykcji strukturalnych został unowocześniony, aby zapewnić większą elastyczność, a co za tym idzie łatwość użytkowania. W ramach aktualnie wspieranych restrykcji strukturalnych wyróżnia się: (1) wyłączanie z procesu predykcji wszystkich elementów strukturalnych, przechowywanych w repozytorium systemu *RNAComposer*, pochodzących z cząsteczek RNA, których identyfikatory *PDB* [E6] zostały określone przez użytkownika, (2) przeszukiwanie, podczas procesu predykcji,

elementów strukturalnych pochodzących jedynie z wysokorozdzielczych struktur 3D RNA, przy czym maksymalna dopuszczalna wartość rozdzielczości może być dowolnie konfigurowana przez użytkownika, (3) rozszerzanie wpływu użytkownika na proces minimalizacji energetycznej wstępnego modelu 3D RNA w przestrzeni zarówno kartezjańskiej jak i kątów torsyjnych poprzez udostępnienie mechanizmu pozwalającego użytkownikowi na wprowadzanie dodatkowych restrykcji w zakresie kątów torsyjnych jako uzupełnienie odległości pomiędzy atomami, które mogą się okazać szczególnie użyteczne w przypadku zawiązywania lub utrwalania oddziaływań przestrzennych. Elementy strukturalne, wprowadzane przez użytkownika, w formacie *PDB* mogą się charakteryzować pełną dowolnością w zakresie wykorzystywanych nazw i numerów reszt nukleotydowych oraz identyfikatorów łańcuchów. Jedyne co jest rzeczywiście wymagane i weryfikowane, przez system, to zapewnienie prawidłowego odwzorowania liczby reszt nukleotydowych w rozpatrywanych niciach oraz kolejności ich występowania. Gdy zachodzą niezgodności w zakresie sekwencji RNA pomiędzy elementem, wprowadzonym przez użytkownika, a jego oczekiwanym odpowiednikiem system automatycznie przeprowadza proces podmiany wymaganych zasad. Jeżeli elementy strukturalne, odpowiadające regularnym motywom dwuniciowym i jednoniciowym, automatycznie identyfikowane przez system, pochodzą ze struktur 3D RNA o niskiej rozdzielczości, wtedy warto zastosować kolejną wprowadzoną funkcjonalność pozwalającą na generowanie struktury 3D elementu o oczekiwanej topologii, w oparciu o regularną konformację A-RNA, zamiast przeszukiwania repozytorium. Zapewnienie wysokiej jakości w przypadku pierwszego modelu 3D RNA, rozpatrywanego w ramach rodziny modeli 3D generowanych przez system, stało się przesłanką do opracowania nowego algorytmu wyboru elementów strukturalnych z repozytorium *RNA FRABABE* [O10, E16], który polega na minimalizacji liczności zbioru tzw. *super-elementów*, które pokrywają możliwie najdłuższy fragment przewidywanej cząsteczki RNA. Pod pojęciem *super-elementu* rozumiany jest najdłuższy fragment przestrzenny złożony z sąsiednich (przynajmniej dwóch) elementów strukturalnych pochodzących z jednej cząsteczki RNA, który jako topologicznie zbieżny może zostać w całości zastosowany podczas prowadzonego procesu predykcji struktury 3D RNA. Lista zidentyfikowanych *super-elementów* jest porządkowana nierosnąco względem ich długości reprezentowanej przez liczbę unikalnych reszt nukleotydowych składających się na elementy w nich zawarte. W ostatniej fazie uruchamiany jest specjalizowany algorytm dokładny w celu identyfikacji optymalnego rozwiązania. W celu ewaluacji opracowanego algorytmu przeprowadzono eksperymenty obliczeniowe na reprezentatywnym zbiorze eksperymentalnie określonych cząsteczek RNA zaproponowanym w [O10]. Uśredniona wartość miary *RMSD* [E15] uległa zauważalnej poprawie (obniżając się z 2,682 do 1,445Å). Log opisujący kluczowe elementy procesu predykcji został rozszerzony w celu raportowania szeregu możliwych nieprawidłowości, które mogą zaistnieć w danych wprowadzonych przez użytkownika. System gwarantuje prawidłowość wykorzystywanych elementów strukturalnych weryfikując: (1) zgodność wprowadzanych przez użytkownika struktur 3D RNA z formatem *PDB*, (2) kompletność zbiorów atomów w ramach wszystkich rozpatrywanych reszt nukleotydowych, (3) ciągłość nici w ramach elementów 3D wprowadzanych przez użytkownika, (4) prawidłowość topologii par kanonicznych na końcach nici w każdym rozpatrywanym elemencie strukturalnym. Ponadto, opis rozpatrywanych elementów strukturalnych, prezentowany w logu, został uzupełniony o skojarzone z nimi identyfikatory np. D1, L1 itd. Użytkownik może łatwo pobrać nie tylko

wynikowy(e) model(e) 3D RNA, ale również struktury 3D wszystkich elementów strukturalnych, w jednym pliku, które zostały wykorzystane podczas procesu predykcji danego modelu(i) 3D RNA i mogą posłużyć w celu powtórzenia eksperymentu. Nowe funkcjonalności modułu obliczeniowego spowodowały rozwój interfejsu użytkownika, który koncentrował się również na zapewnieniu większej wygody użytkowania. W celu zaprezentowania nowych aspektów funkcjonalnych systemu *RNAComposer* przeprowadzono eksperymenty udokładnionego modelowania struktur 3D prekursorów miR160 RNA [A4]. Na Rysunku 2 przedstawiona została charakterystyka elementów strukturalnych wybranych automatycznie przez system (po lewej) oraz manualnie przez eksperta (po prawej), które doprowadziły do uzyskania modeli 3D RNA dla prekursora miR160c w trybie interaktywnym (po lewej) oraz wsadowym z wykorzystaniem nowych funkcjonalności (po prawej).



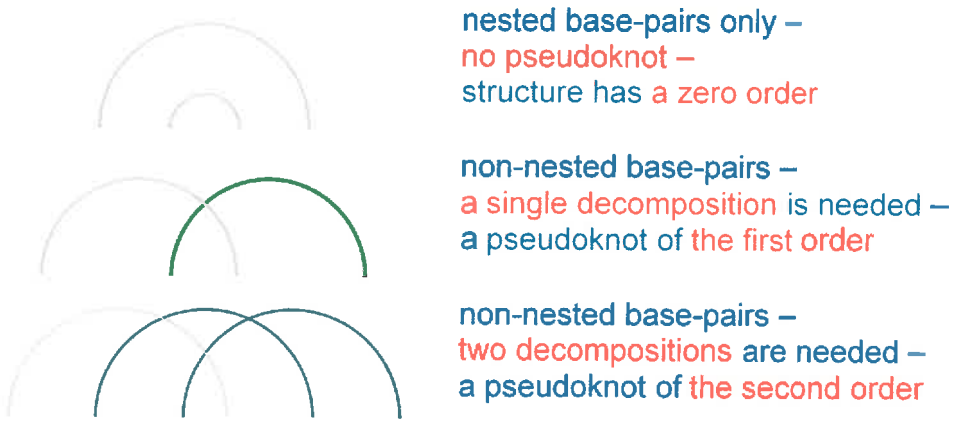
Secondary structure element		RNA FRABASE	3D element selected automatically by RNAComposer				3D element selected manually			
Id	Topology	search pattern	PDB ID Chain	Position	Topology	Res. [Å] PSI [%]	PDB ID chain	Position	Topology	Res. [Å] PSI [%]
L1	CCU (..(AAG).)	NCN (..(NAN).)	3U5H BA	2338-2340, 1895-1897	CCU (..(AAG).)	3.00 100.0	4FE5 B	25-27, 43-45	UCG (..(CAA).)	1.32 33.33
L2	CCU (..(AAG).)	NCN (..(NAN).)	3U5H BA	2338-2340, 1895-1897	CCU (..(AAG).)	3.00 100.0	4FE5 B	25-27, 43-45	UCG (..(CAA).)	1.32 33.33
L3	CGAG (...(CCGG)...)	NRRN (...(NYRN)...)	3J3F AB	490-493, 574-577	CGUG (...(CCGG)...)	5.00 87.50	2ZJR X	1491-1494, 1530-1533	CAAG (...(UCAG)...)	2.91 50.00
L4	GAUAC (...(GC))	(...( ))	1U3K BA	29-33, 6-7	GAUAU (...(AC))	NMR 71.43	397D AB	22-26, 39-40	AUCUG (...(CU))	1.30 0.00

Rysunek 2. Przykładowe wyniki predykcji struktur 3D RNA uzyskane, z wykorzystaniem systemu *RNAComposer*, dla prekursora miR160c w trybie interaktywnym (po lewej) oraz wsadowym (po prawej).

Łatwo zaobserwować, że w przypadku tak krótkich i specyficznych motywów rozdzielczość eksperymentalna okazała się istotniejsza aniżeli identyczność sekwencyjna. System *RNAComposer* jest zasobem, z ugruntowaną pozycją w międzynarodowej społeczności naukowej, często wykorzystywanym do zastosowań badawczych jak i dydaktycznych o czym świadczą: (a) sumaryczna liczba odwiedzin użytkowników w obrębie dwóch udostępnionych instancji przekraczająca ***5 mln***, oraz (b) liczba cytowań pierwszej publikacji na poziomie ***180*** w WoS (na dzień 17.02.2019), która dodatkowo znajduje się **w 1% najlepiej cytowanych artykułów w WoS** od sierpnia 2017 r.

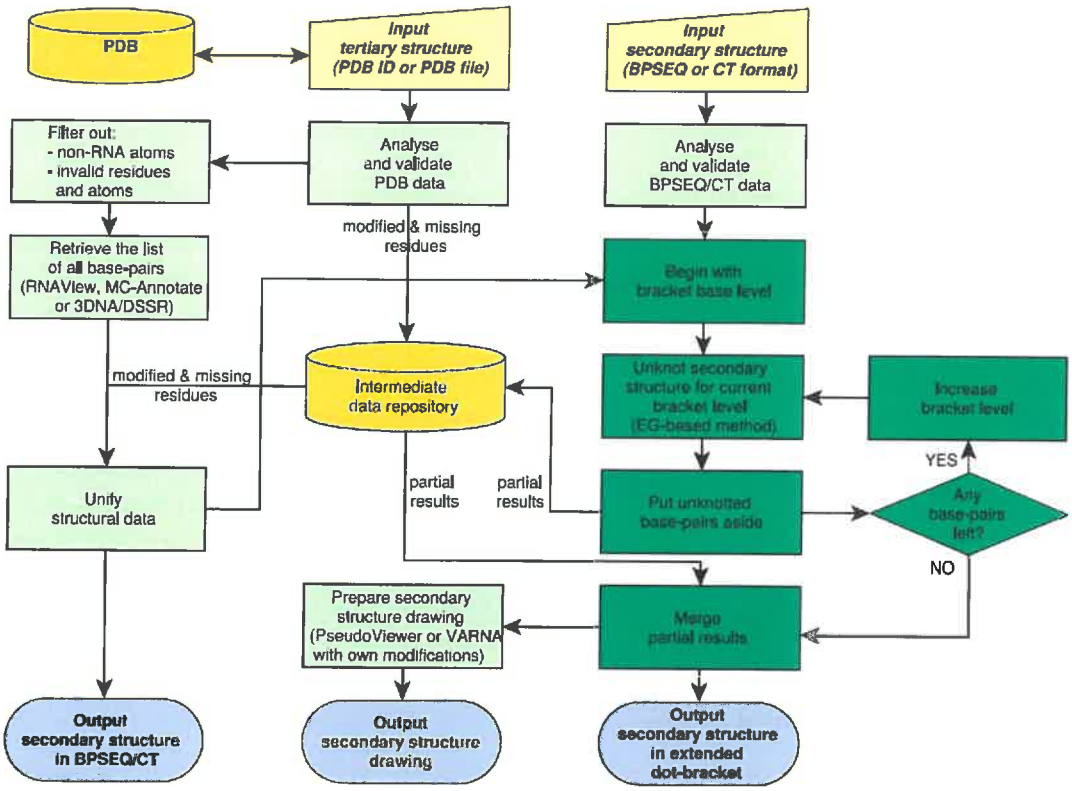
4.3.3.2. Ekstrakcja struktury drugorzędowej w rozszerzonej notacji kropkowo-nawiasowej na podstawie struktury 3D RNA ze szczególnym uwzględnieniem rzędowości pseudowęzłów.

Ekstrakcja struktury drugorzędowej na podstawie struktury 3D RNA jest wykorzystywana w celu: (1) walidacji modeli 3D RNA, uzyskanych z wykorzystaniem metod obliczeniowych, w kontekście np. eksperymentalnie potwierdzonej struktury drugorzędowej RNA, (2) analizy rozkładu struktur drugorzędowych w zbiorze modeli 3D RNA w celu identyfikacji wspólnych elementów topologicznych, (3) przygotowania rankingu modeli 3D RNA na podstawie konsensusu struktur drugorzędowych RNA. Struktura drugorzędowa RNA jest zwykle zapisywana tekstowo w postaci listy par zasad (w formatach odpowiednio *BPSEQ* lub *CT*) oraz notacji kropkowo-nawiasowej, gdzie każda reszta niesparowana jest oznaczana przez znak '.', natomiast reszty sparowane oznaczane są nawiasami '()'. W 2014 roku ogólnodostępne były następujące narzędzia: *RNAView* [E21], *MC-Annotate* [E21], *3DNA/DSSR* [E22] oraz *PseudoViewer* [E23], *VARNA* [E24] służące odpowiednio do identyfikacji i klasyfikacji par zasad RNA oraz wizualizacji struktury drugorzędowej RNA. Niestety ani *MC-Annotate* ani *3DNA/DSSR* nie klasyfikowały precyzyjnie identyfikowanych pseudowęzłów w ramach złożonych cząsteczek RNA [A8]. Podobnie pseudowęzły nie były prawidłowo klasyfikowane przez istniejące narzędzia do wizualizacji. Brak precyzji w tym zakresie zdecydowanie utrudnia modelowanie homologiczne w oparciu o wzorce strukturalne cząsteczek RNA o złożonych topologiach pseudowęzłów. Serwer obliczeniowy, o nazwie *RNApdbee* [A8], był pierwszą aplikacją, udostępniającą nową metodę obliczeniową, która precyzyjnie klasyfikowała złożone topologię pseudowęzłów występujące w cząsteczkach RNA o dużych rozmiarach. Wymagało to wprowadzenia pojęcia *rzędu* pseudowęzła, opartego na koncepcji zagnieżdżonych i niezagnieżdżonych helis RNA [E25], reprezentującego minimalną liczbę dekompozycji zbioru par zasad prowadzącą do uzyskania podzbiorów zawierających jedynie zagnieżdżone parowania zasad azotowych RNA. Ilustracja koncepcji *rzędu* pseudowęzła została zaprezentowana na Rysunku 3.



Rysunek 3. Ilustracja koncepcji *rzędu* pseudowęzła.

W ogólności proces ekstrakcji struktury drugorzędowej na podstawie struktury 3D RNA obejmuje następujące etapy: (1) walidacja struktury 3D RNA wprowadzonej w formacie *PDB* [E6], (2) ekstrakcja listy kanonicznych i niekanonicznych par zasad azotowych RNA z wykorzystaniem jednego z trzech zintegrowanych narzędzi, (3) przeprowadzenie iteracyjnej dekompozycji listy par zasad poprzez identyfikację pseudowęzłów kolejnych *rzędów* z wykorzystaniem, dostosowanej do naszych potrzeb, heurystyki *Elimination Gain* [E25], (4) kodowanie w rozszerzonej notacji kropkowo-nawiasowej i wizualizacja (opcjonalnie) struktury drugorzędowej RNA. Schemat blokowy opracowanej metody został zaprezentowany na Rysunku 4.

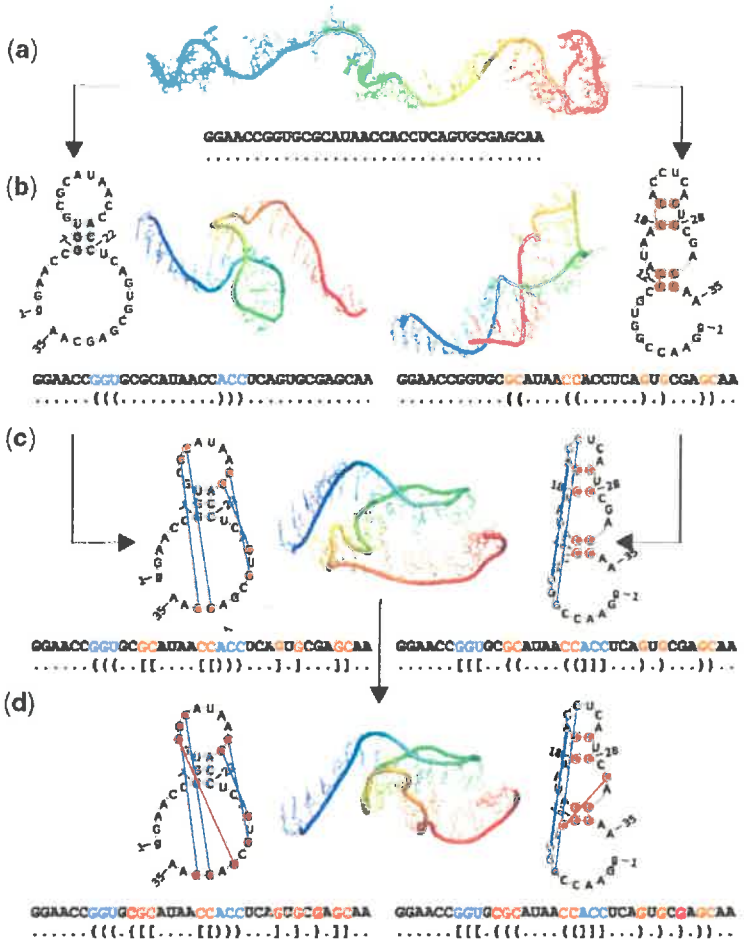


Rysunek 4. Schemat przetwarzania realizowany w systemie *RNApdbee* [A8].

Rozszerzenie notacji kropkowo-nawiasowej polega na wprowadzeniu dodatkowych rodzajów nawiasów, a mianowicie '[,]', '{,}', '<,>', 'A,a', 'B,b', itd., wykorzystywanych w przypadku par zasad sklasyfikowanych w ramach pseudowęzłów o kolejnych *rzędach* ≥ 1. Zintegrowane narzędzia do wizualizacji zostały rozszerzone w celu prezentowania kolejnych *rzędów* pseudowęzłów z wykorzystaniem zaproponowanej palety kolorów [A8]. Ostatecznie przeprowadzono eksperymenty obliczeniowe w celu wyznaczenia rozkładu *rzędów* pseudowęzłów występujących w cząsteczkach RNA, zdeponowanych w *RNA FRABASE* [E26, E27].

Dalsza eksploracja zagadnienia wiarygodnego kodowania struktury drugorzędowej złożonych cząsteczek RNA obejmujących pseudowęzły, w rozszerzonej notacji kropkowo-nawiasowej, była kontynuowana i zaowocowała nową funkcją kryterialną służącą do oceny rozwiązań oraz nowymi algorytmami opublikowanymi w pracy [A3]. Udostępniona wersja serwera opierała się na metodzie przybliżonej, co powodowało, że niekiedy współpracujący eksperymentatorzy sygnalizowali, że niektóre struktury drugorzędowe RNA, zwracane przez system, nie zawsze są wystarczająco realistyczne. Zwijanie struktur 3D RNA to proces hierarchiczny w którym poszczególne domeny zwijają się sekwencyjnie. Następnie wewnątrzcząsteczkowe oddziaływania przestrzenne pozycjonują elementy o określonej topologii względem siebie inicjując tworzenie bardziej złożonych motywów strukturalnych zwanych pseudowęzłami [E28]. Na Rysunku 5 przedstawiono hipotetyczne ścieżki zwijania dość krótkiej cząsteczki RNA (1DDY:A). Łatwo zaobserwować jak zmienia się zarówno struktura drugorzędowa jak i struktura 3D (przewidziana z wykorzystaniem systemu *RNAComposer*),

analizowanego RNA, w zależności od sposobu klasyfikacji rozpatrywanego zbioru par zasad na parowania regularne i pseudowęzły kolejnych *rzędów*.



Rysunek 5. Ścieżka zwijania się niewielkiej cząsteczki RNA (1DDY:A) od stanu (a) jednoniciowego, poprzez tworzenie (b) pętli typu *hairpin* i (c) pseudowęzła pierwszego *rzędu*, aż (d) po uzyskanie finalnej struktury z pseudowęzłem *rzędu* drugiego [A3].

Początkowo założono, że wszystkie algorytmy rozpatrywane w pracy [A3] wymagają zbioru par zasad azotowych RNA (*inputBps*) na wejściu oraz będą generowały strukturę drugorzędową RNA w rozszerzonej notacji kropkowo-nawiasowej na wyjściu. Analizę przeprowadzono w zasadzie od samego początku formułując pierwszy algorytm o nazwie *First-Come-First-Served* (*FCFS*), który został opracowany podczas tworzenia bazy danych *RNA FRABASE* [E26]. Następnie opracowano nowy algorytm dokładny (*DP*), wykorzystujący strategię programowania dynamicznego [E29], gdzie podproblemem jest klasyfikacja pojedynczej pary zasad azotowych RNA. W ramach opracowanego algorytmu iteracyjnie uruchamiane są następujące operacje: (1) identyfikacja maksymalnego podzbioru zagnieżdżonych par zasad (*currentNestedBps*) pochodzących z wejściowego zbioru par zasad (*inputBps*), (2) przypisanie wszystkim parom zasad, należącym do zbioru *currentNestedBps*, bieżącej wartości *rzędu* pseudowęzła (*currentOrder* początkowo równego *0*), (3) inkrementacja bieżącej wartości *rzędu* o 1 (*currentOrder++*), (4) usunięcie wszystkich par zasad należących do zbioru *currentNestedBps* ze zbioru wejściowego *inputBps* (*inputBps* = *inputBps* \ *currentNestedBps*). Powyższy zestaw operacji uruchamiany jest tak długo jak długo wejściowy zbiór par zasad nie jest

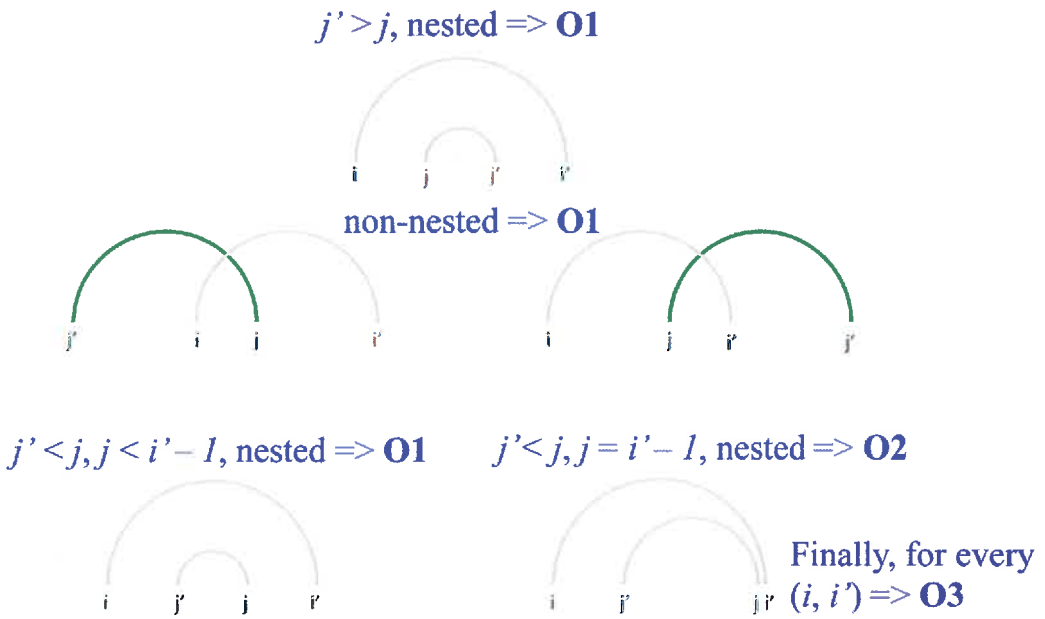
pusty (*inputBps* $\neq \emptyset$). Pierwsza operacja jest kluczową procedurą algorytmu. Rozpoczyna się ona od wstępnego przygotowania danych. W skład tego procesu wchodzą następujące operacje: (1) przenumerowanie (od 1) reszt w zbiorze par zasad biorąc pod uwagę jedynie podzbiór reszt sparowanych, (2) wprowadzenie pary reprezentującej wirtualny łuk, w obrębie którego wszystkie rozpatrywane parowania zasad są zagnieżdżone, (3) porządkowanie rozpatrywanego zbioru par zasad (zapewniając, że indeks otwierającej reszty jest zawsze mniejszy od indeksu zamykającej reszty w każdej rozpatrywanej parze zasad) w porządku niemalejącym najpierw względem różnicy między indeksami odpowiednio zamykającej i otwierającej reszty w parze, a następnie indeksu otwierającej reszty. Następnie dwie macierze DP są alokowane i wypełniane rekursywnie wagami liczbowymi, a mianowicie: (1) macierz przechowująca oceny kolejnych optymalnych rozwiązań tzn. zbiorów zawierających zagnieżdżone pary zasad (*scoreMtx*), (2) macierz przechowująca numer zamykającej reszty najbardziej zewnętrznej pary zasad (tzn. różnica między indeksem reszty zamykającej i otwierającej w przypadku tej pary jest największa) w obecnym zbiorze reszt zagnieżdżonych (*indexMtx*). Zmiana dokonana w macierzy *scoreMtx* powoduje odpowiednią modyfikację w macierzy *indexMtx*. Podczas wypełniania macierzy *scoreMtx* wykorzystywane są następujące operacje:

O1: $scoreMtx[i+1][j]=max\{scoreMtx[i+1][j], scoreMtx[i+1][j-1]\}$

O2: $scoreMtx[i+1][j]=max\{scoreMtx[i+1][j], scoreMtx[i+1][j'-1]+scoreMtx[j'][j]\}$

O3: $scoreMtx[i][i']=1+scoreMtx[i+1][i'-1]$

Dla każdej dwójki par zasad $\{(i, i'), (j, j')\} \in inputBps$, gdzie $i+1 < j < i' - 1$ operacje uruchamiane są zgodnie z zależnościami zaprezentowanymi na Rysunku 6.



Rysunek 6. Zależności pomiędzy rozpatrywanymi dwójkami par zasad azotowych RNA rozróżniane przez algorytm programowania dynamicznego [A3].

Po wypełnieniu macierzy optymalne rozwiązanie (tzn. maksymalny zbiór zagnieżdżonych par zasad azotowych RNA) jest odczytywane z macierzy *indexMtx* w następujący sposób:

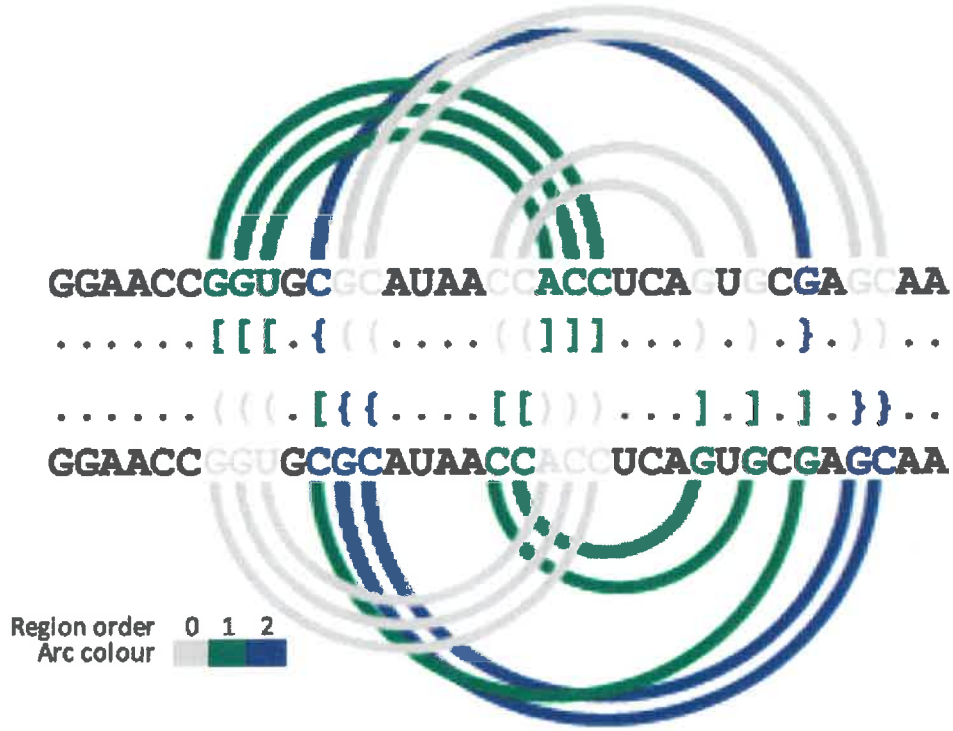
- zaczynając od $i = 1$, $i' = n$, jeżeli $indexMtx[i][i'] = k'$ i k' jest liczbą, wówczas k' to numer zamykającej reszty pary zasad (k, k') w rozwiązaniu,

- numer reszty otwierającej *k* pobierany jest z *inputBps*,
- procedura jest rekurencyjnie kontynuowana odwiedzając kolejne komórki, a mianowicie *indexMtx[k+1][k'-1]* i *indexMtx[i][k-1]* aż do momentu natknięcia się na niewypełnioną komórkę.

Pomimo dysponowania optymalnym algorytmem w przypadku jednokryterialnej funkcji, która dla każdej kolejnej wartości *rzędu* pseudowęzła identyfikowała maksymalną liczbę zagnieżdżonych par zasad nadal pojawiały się rozwiązania, których realistyczność była dyskutowana przez eksperymentatorów. W wyniku przeprowadzonych konsultacji opracowano nową dwukryterialną funkcję oraz relację dominacji [A3]. Ta druga preferuje rozwiązania w których liczba par zasad niezakwalifikowanych jako pseudowęzły (ich *rząd* wynosi 0) jest największa, a w przypadku dwóch rozwiązań, które nie różnią się względem pierwszego kryterium preferowane jest to rozwiązanie, które charakteryzuje się niższą karą reprezentowaną przez średnią ważoną iloczynów kolejnych *rzędów* (od 1) i liczby odpowiadających im par zasad azotowych RNA (zakładając, że liczba par zasad jest dodatnia) ostatecznie ze znakiem przeciwnym (aby maksymalizować oba kryteria). W ramach dalszych prac opracowano nowy algorytm nazwany algorytmem *hybrydowym* (*HYB*). Operuje on na *regionach*, a nie pojedynczych parach zasad azotowych RNA, przy czym *region* reprezentowany jest przez zbiór sąsiednich par zasad azotowych RNA (np. region składający się z trzech par zasad mógłby zostać opisany następującym zbiorem {(2,7),(3,6),(4,5)}). W pierwszej kolejności identyfikowane są wszystkie zagnieżdżone regiony, których *rząd* wynosi 0. W ramach pozostałych regionów identyfikowane są ich rozłączne podzbiory. Dla każdego regionu, rozpatrywanego w ramach zidentyfikowanego podzbioru, musi istnieć przynajmniej jeden inny region w tym podzbiorze pomiędzy którymi nie zachodzi relacja zagnieżdżania. Algorytm podczas rozwiązywania kolejnych rozłącznych podzbiorów regionów wykorzystuje przeszukiwanie wyczerpujące albo losowe (10 000 permutacji), przy czym pierwsza strategia jest wykorzystywana jeśli liczba regionów zawartych w rozpatrywanym podzbiorze nie przekracza ośmiu. Celem algorytmu jest identyfikacja rozkładu *rzędów* dla rozpatrywanych regionów, dla każdego zidentyfikowanego podzbioru, optymalizując zaproponowaną dwukryterialną funkcję z wykorzystaniem zdefiniowanej relacji dominacji. W celu wiarygodnej ewaluacji proponowanych algorytmów w kontekście innych, opublikowanych już, podejść przeprowadzono eksperymenty obliczeniowe oraz interpretację wyników na podstawie obu aktualnie dostępnych funkcji oceny rozwiązań. Na wstępie przygotowano dwa zbiory ewaluacyjne zawierające listy par zasad, w formacie *BPSEQ*, dla reprezentatywnego zestawu nieredundantnych struktur 3D RNA [E30] z wykorzystaniem narzędzia *3DNA/DSSR* [E22] zakładając, że pary niekanoniczne nie są rozpatrywane w ogóle lub są rozpatrywane, ale jedynie w przypadku ich występowania w regionach dwuniciowych. Dla obu rozpatrywanych zbiorów danych przeprowadzono analizę statystyczną liczb rozłącznych podzbiorów regionów oraz oceniono następujące algorytmy: *FCFS* [E26], *EG* i *EC* [E25] oraz *DP* i *HYB* [A3]. Następnie przeanalizowano uzyskane wyniki w celu: (1) wyznaczenia rozkładu liczb cząsteczek RNA dla których rozpatrywane algorytmy zidentyfikowały pseudowęzły o *rzędzie* nie przekraczającym określonej wartości z zakresu 1 do 8, (2) identyfikacji podzbioru instancji dla których zaobserwowano różnorodność w zakresie uzyskiwanych rozwiązań, które ostatecznie posłużyły do przeprowadzenia analizy porównawczej rozpatrywanych algorytmów, oraz (3) wyznaczenia procentowego współczynnika instancji dla których rozpatrywane algorytmy zwróciły niezdominowane rozwiązania (należące do frontu Pareto)

w oparciu o zaproponowaną dwukryterialną funkcję oceny. Analiza porównawcza algorytmów skupiła się na wyznaczeniu jak często poszczególne algorytmy dominowały/były dominowane przez inne algorytmy albo okazywały się bezkonkurencyjne (tzn. lepsze od wszystkich pozostałych). Uzyskane rezultaty potwierdziły, że nowo opracowane algorytmy, a więc *HYB* i *DP* okazały się najlepsze w stawce dla obu rozpatrywanych zbiorów ewaluacyjnych [A3]. Porównanie rozwiązań uzyskanych z wykorzystaniem algorytmów (a) *FCFS* oraz (b) *HYB* dla przykładowej cząsteczki RNA (1DDY:A) obejmujące odpowiadające im wartości zaproponowanej funkcji oceny zostało zaprezentowane na Rysunku 7.

(a) $$fscoreII(R1)=\begin{bmatrix} 4 \\ -1\cdot(1\cdot 3+2\cdot 1)\end{bmatrix}=\begin{bmatrix} 4 \\ -5\end{bmatrix}$$



(b) $$fscoreII(R2)=\begin{bmatrix} 3 \\ -1\cdot(1\cdot 3+2\cdot 2)\end{bmatrix}=\begin{bmatrix} 3 \\ -7\end{bmatrix}$$

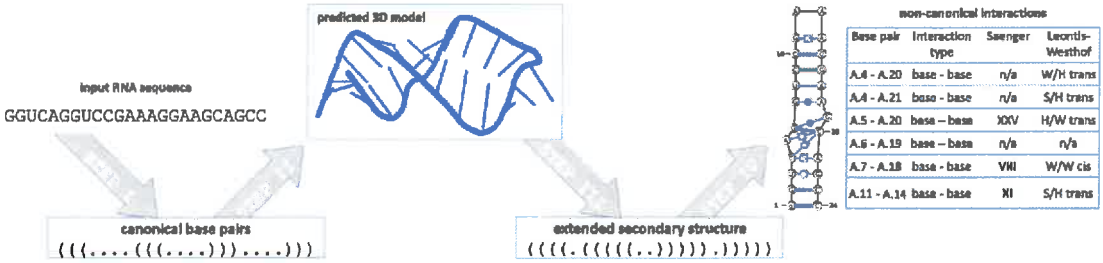
Rysunek 7. Porównanie rozwiązań uzyskanych z wykorzystaniem algorytmów: (a) *FCFS* oraz (b) *HYB* dla przykładowej cząsteczki RNA (1DDY:A) obejmujące odpowiadające im wartości nowej funkcji oceny [A3].

System *RNApdbee* jest również popularnym zasobem chętnie wykorzystywanym przez międzynarodową społeczność naukową o czym świadczą: (a) sumaryczna liczba odwiedzin użytkowników przekraczająca ***12,5 tys.***, oraz (b) liczba cytowań pierwszej publikacji na poziomie ***25*** w WoS (na dzień 17.02.2019).

4.3.3.3. Przewidywanie rozszerzonej struktury drugorzędowej RNA, obejmującej pary niekanoniczne, na podstawie sekwencji lub kanonicznej struktury drugorzędowej RNA.

Uwzględnienie niekanonicznych par zasad RNA jest kluczem do wiarygodnego i precyzyjnego przewidywania struktury 3D RNA [E7, E8, O3]. Spośród przeszło

pięćdziesięciu metod służących do przewidywania struktury drugorzędowej na podstawie sekwencji reszt nukleotydowych RNA zaledwie kilka udostępnia rozszerzoną strukturę drugorzędową RNA zawierającą zarówno kanoniczne jak i niekanoniczne parowania zasad [E31, E32, E10, E33, A7, E34, E35]. Dysponując efektywnymi narzędziami w zakresie zarówno przewidywania wysokorozdzielczych struktur 3D RNA (*RNAComposer* [O10, O8, A4]) jak i adnotacji rozszerzonej struktury drugorzędowej RNA obejmującej, między innymi, niekanoniczne parowania zasad (*RNApdbee* [A8, A2, O2]) zaproponowano nową metodologię w [A7] oraz publicznie dostępny, w pełni automatyczny serwer obliczeniowy [A2] służący do przewidywania rozszerzonej struktury drugorzędowej RNA, obejmującej niekanoniczne parowania zasad, na podstawie sekwencji lub kanonicznej struktury drugorzędowej RNA. Proponowane podejście, zaprezentowane na Rysunku 8, obejmuje następujące etapy: (1) przewidywanie kanonicznej struktury drugorzędowej na podstawie sekwencji reszt nukleotydowych RNA (opcjonalnie), (2) przewidywanie struktury 3D na podstawie kanonicznej struktury drugorzędowej RNA, (3) adnotacja rozszerzonej struktury drugorzędowej na podstawie struktury 3D RNA, (4) prezentacja w postaci tekstowej jak i graficznej niekanonicznych par zasad z wykorzystaniem klasyfikacji Leontisa-Westhofa [E36].



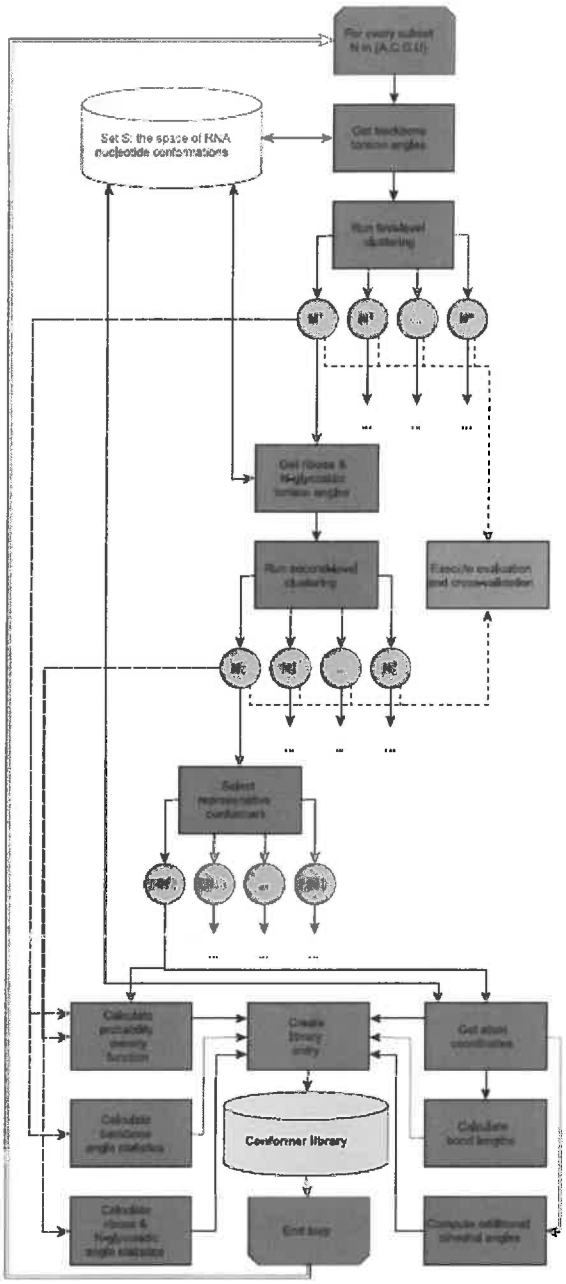
Rysunek 8. Schemat przetwarzania realizowany w systemie *RNAvista* [A2].

Udostępnienie serwera wymagało opracowania i integracji dedykowanych modułów obliczeniowych systemów zarówno *RNAComposer* jak i *RNApdbee* dostosowanych do wymagań wynikających z prowadzonych badań. Warto podkreślić, że użytkownik może dowolnie konfigurować zestaw procedur, spośród tych udostępnionych w ramach zintegrowanych modułów obliczeniowych (np. narzędzia do przewidywania kanonicznej struktury drugorzędowej czy też do identyfikacji listy par zasad azotowych RNA), które powinny zostać zastosowane w bieżącym procesie przewidywania. W pracy [A7] przeprowadzono eksperymenty obliczeniowe w dużej skali w celu ewaluacji precyzji pilotażowego wdrożenia infrastruktury integrującej systemy *RNAComposer* i *RNApdbee*. Na wstępie przygotowano zbiór referencyjny, w którym znalazły się cząsteczki RNA pochodzące z repozytorium *RNA STRAND* [E37] o rozmiarze nie przekraczającym 500nts. Następnie przeprowadzono eksperymenty przewidywania rozszerzonej struktury drugorzędowej RNA na podstawie jedynie sekwencji lub kanonicznej struktury drugorzędowej RNA. Ewaluacja uzyskanych wyników wymagała opracowania nowego narzędzia pozwalającego na automatyczne wyznaczanie wartości miar, a mianowicie liczba oraz procent prawidłowo przewidzianych niekanonicznych par zasad RNA, dokładność (*PPV*), czułość (*TPR*), oraz współczynnik korelacji Mathewsa (*MCC*) [E38], weryfikujących precyzję procesu predykcji par zasad RNA, ze szczególnym uwzględnieniem par niekanonicznych, w kontekście znanej struktury referencyjnej. Wyniki wykazały, że w zależności od długości wejściowej sekwencji RNA, procent prawidłowo przewidzianych niekanonicznych par zasad na podstawie jedynie

sekwencji reszt nukleotydowych albo kanonicznej struktury drugorzędowej RNA wahał się w zakresie odpowiednio 30,64–57,57 oraz 49,91–70,51. Ponadto wysoka jakość predykcji została również potwierdzona na zbiorze struktur 3D RNA pochodzących z *CompaRNA PDB benchmark* [E19]. W kolejnej pracy [A2] przeprowadzono nowe eksperymenty obliczeniowe mające na celu ewaluację stopnia precyzji zaproponowanego serwera w zakresie identyfikacji niekanonicznych par zasad w ramach reprezentatywnego zbioru rekurencyjnych motywów RNA [E4] obejmującego: *K-turn, T-loop, C-loop, Sarcin, GNRA, Double sheared and Triple sheared.* Dla każdego motywu określono identyfikatory *PDB* [E6] referencyjnych struktur 3D RNA o wysokiej rozdzielczości oraz skojarzone z nimi dane, a mianowicie sekwencję w formacie *FASTA*, kanoniczną strukturę drugorzędową w notacji kropkowo-nawiasowej, zakres reszt wchodzących w skład motywu oraz referencyjny zbiór par zasad w formacie *BPSEQ*. Uzyskane rezultaty potwierdziły, że dysponowanie wiarygodną kanoniczną strukturą drugorzędową RNA na wejściu znacznie zwiększa precyzję rozszerzonej struktury drugorzędowej RNA [A2].

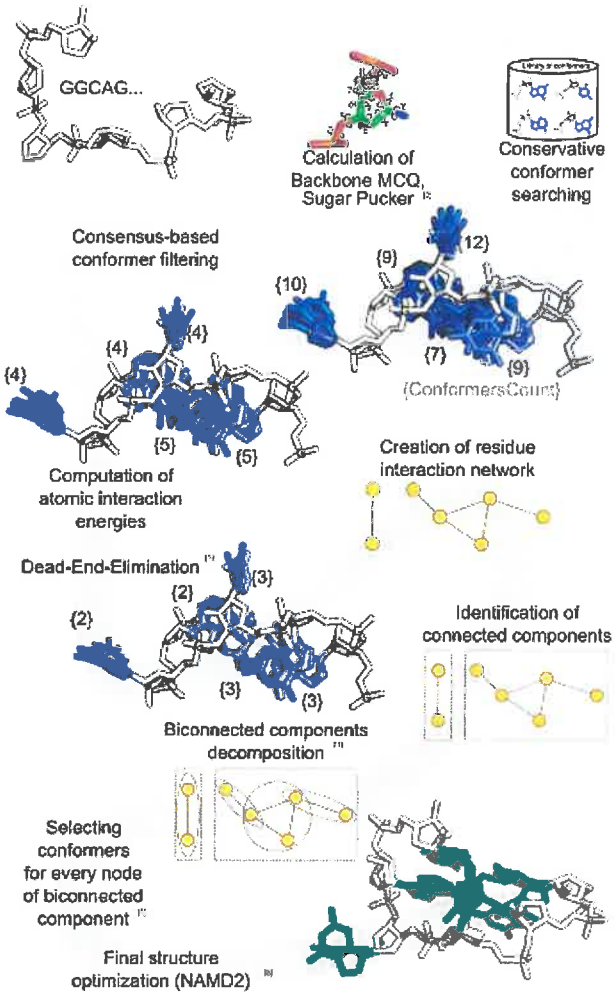
4.3.3.4. Rekonstrukcja pełnoatomowej struktury 3D RNA na podstawie wprowadzonej sekwencji RNA oraz struktury 3D utrwalonego szkieletu fosforo-cukrowego.

W przypadku przewidywania struktur 3D RNA w ramach niektórych wyzwań w konkursie *RNA-Puzzles* [O3] zaistniała potrzeba precyzyjnego, udokładnionego modelowania wybranego podzbioru reszt zasad azotowych/nukleozydowych na strukturze 3D utrwalonego szkieletu fosforo-cukrowego w celu odbudowania brakujących oddziaływań przestrzennych lub identyfikacji pełno-atomowej konformacji RNA w ramach określonego, utrwalonego otoczenia przestrzennego. Ponadto jakość modeli 3D RNA zgłaszanych w konkursie *RNA-Puzzles* [O3] zwykle odbiega od oczekiwań w zakresie wartości miary *Clash score* [E39], co powoduje potrzebę neutralizacji poważnych zderzeń sterycznych bez wpływu na ogólny kształt. Niestety istniejące narzędzia wspierające modelowanie homologiczne np. *ModeRNA* [E11] nie gwarantowały niezmienności struktury 3D szkieletu RNA. W związku z powyższym opracowano nową metodę [A1], udostępnioną w postaci w pełni automatycznego serwera obliczeniowego *RNAfitme*, pozwalającą na rekonstrukcję pełnoatomowej konformacji cząsteczki RNA charakteryzującej się minimalną wartością całkowitej energii oddziaływań międzyatomowych na podstawie wprowadzonej sekwencji RNA oraz utrwalonej struktury 3D szkieletu fosforo-cukrowego cząsteczki RNA. Inspiracją ogólnej koncepcji projektowanej metody był sukces narzędzia *SCWRL* [E40] pozwalającego na identyfikację globalnej konformacji łańcuchów bocznych dla utrwalonej struktury 3D szkieletu białka. *SCWRL* wykorzystuje on bibliotekę reprezentatywnych konformacji rotamerycznych łańcuchów bocznych, zależnych od szkieletu, wyznaczoną na podstawie zbioru eksperymentalnie określonych struktur 3D białek [E41]. Badania rozpoczęto od opracowania nowego protokołu konstrukcji biblioteki reprezentatywnych konformacji nukleotydowych RNA [A6], zaprezentowanego na Rysunku 9. Na wstępnie przygotowano zbiór konformacji reszt nukleotydowych obserwowanych w strukturach 3D cząsteczek RNA zdeponowanych w *RNA FRABASE* [E26, E27]. Każda kompletna i niemodyfikowana reszta nukleotydowa opisana została przez wektor kątów torsyjnych odpowiednio szkieletu (α, β, γ, δ, ε, ζ), rybozy ($\nu 0$, $\nu 1$, $\nu 2$, $\nu 3$, $\nu 4$) oraz zasady (χ) oraz rozdzielczość eksperymentalną. Podczas badań rozważano trzy zbiory konformacji nukleotydów pochodzące odpowiednio



Rysunek 9. Schemat protokołu konstrukcji biblioteki [A6].

z (a) pojedynczej, dużej jednostki 23S rRNA (PDBId: 3CC2), (b) zbioru struktur 3D RNA o rozdzielczości poniżej 2,4Å, oraz (c) wszystkich aktualnie dostępnych struktur 3D zdeponowanych w *RNA FRABASE*. Każdy rozpatrywany zbiór został podzielony na cztery podzbiory odpowiadające typom reszt w RNA. Każdy uzyskany podzbiór został poddany hierarchicznej analizie skupień. Na pierwszym poziomie analiza dotyczyła identyfikacji różnych konformacji szkieletu reszt nukleotydowych. Następnie określono różne konformacje względem kątów rybozy i χ. Reprezentatywne konformacje reszt nukleotydowych zostały wybrane na podstawie macierzy odległości wyznaczonej w trybie każdy z każdym w obrębie każdego uzyskanego skupienia. Analiza skupień została przeprowadzona z wykorzystaniem dwóch podejść, a mianowicie przez partycjonowanie (*k-medoids* [E42]) oraz kwantowanie wektorowe (*median neural gas technique* [E43]). Podczas eksperymentów rozważane były następujące metryki odległości, a mianowicie metryka euklidesowa oraz średnia różnica kątowa (*mean of circular quantities – MCQ* [E44]). Każda konformacja przechowywana w bibliotece opisywana jest przez zbiór indywidualnych danych pozwalających na wydajną jej identyfikację na podstawie charakterystyki szkieletu i/lub rybozy, współrzędne składających się na nią atomów oraz statystyczne właściwości skupienia, w którym została zidentyfikowana. Przeprowadzono eksperymenty weryfikujące jakość uzyskanych bibliotek [A6]. W drugim etapie opracowano algorytm optymalizacyjny, zainspirowany podejściem [E40], pozwalający na modelowanie reszt zasad azotowych/nukleozydowych na podstawie sekwencji i struktury 3D utrwalonego szkieletu fosforo-cukrowego RNA, zaprezentowany na Rysunku 10. Algorytm buduje wstępny, pełno atomowy model 3D RNA, wykorzystując wybraną przez użytkownika bibliotekę reprezentatywnych reszt nukleotydowych RNA, minimalizując zbyt bliskie oddziaływania międzyatomowe. Opracowane podejście realizuje następujące etapy: (1) wyszukiwanie konformacji kandydatów w wybranej bibliotece, z wykorzystaniem



Rysunek 10. Schemat przetwarzania realizowany w systemie *RNAfitme*.

miary *MCQ* wyznaczonej dla odpowiadających kątów torsyjnych szkieletu, dla każdego rozpatrywanego nukleotydu utrwalonej struktury 3D RNA, (2) rekonstrukcja współrzędnych reszty zasady azotowej/nukleozydowej, dla każdego rozpatrywanego nukleotydu utrwalonej struktury 3D RNA, na podstawie odpowiadających kątów torsyjnych rozpatrywanych w konformacjach kandydatów przechowywanych w bibliotece, (3) identyfikacja (maksymalnie dziesięciu) rokujących poprawność konformacji reszt zasad azotowych/nukleozydowych, dla każdego rozpatrywanego nukleotydu utrwalonej struktury 3D RNA, uszeregowanych zgodnie z niemalejącym rankingiem wartości miary *RMSD* [E15] wyznaczonej pomiędzy zrekonstruowaną konformacją reszty nukleotydowej, a jej odpowiadającą konformacją nukleotydu RNA zawartą w bibliotece, (4) dla każdej zrekonstruowanej konformacji reszty zasady azotowej/nukleozydowej $c_m(r_i)$, dla każdego rozpatrywanego nukleotydu r_i utrwalonej struktury 3D RNA, wyznaczane są następujące rodzaje energii wzajemnych oddziaływań międzyatomowych pomiędzy daną konformacją, a: (a) utrwaloną częścią struktury 3D ($E_{fixed}(c_m(r_i))$), (b) każdą konformacją reszty zasady azotowej/nukleozydowej $c_n(r_j)$ rozpatrywanej w ramach każdego innego, dotąd nierozwiązanego nukleotydu r_j ($E_{pair}(c_m(r_i), c_n(r_j))$), (5) eliminowanie wszystkich konformacji reszt zasad azotowych/nukleozydowych, dla każdego rozpatrywanego nukleotydu utrwalonej struktury 3D RNA, które nie wejdą w skład rozwiązania o minimalnej energii z wykorzystaniem procedury *dead-end elimination* [E40], (6) w obrębie rozpatrywanych nukleotydów utrwalonej struktury 3D RNA, dla których nadal występuje więcej niż jedna osiągalna konformacja reszty zasady azotowej/nukleozydowej, tworzona jest sieć oddziaływań w postaci nieskierowanego grafu G; w tej sieci wierzchołkami są rozpatrywane nukleotydy r_i, a krawędź pomiędzy parą reszt (r_i, r_j) tworzona jest wtedy, gdy istnieje przynajmniej jedna para konformacji reszt nukleotydowych ($c_m(r_i)$, $c_n(r_j)$), dla których energia ich wzajemnych oddziaływań ($E_{pair}(c_m(r_i), c_n(r_j))$) jest dodatnia, (7) identyfikacja wszystkich spójnych składowych (ang. *connected components*) w sieci, (8) dekompozycja zidentyfikowanych spójnych składowych do postaci drzewa dwuspójnych składowych (ang. *block-cut tree*) z wykorzystaniem przeszukiwania w głąb oraz identyfikacja wierzchołków należących do przynajmniej

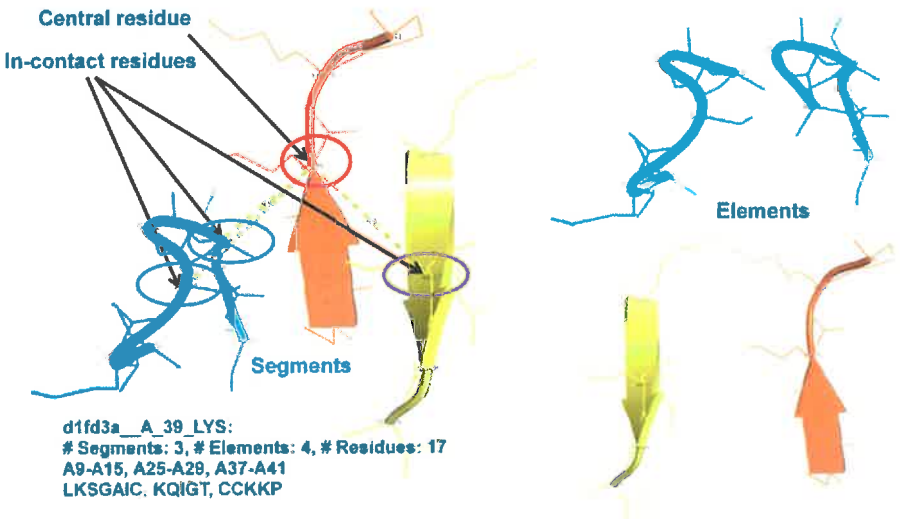
dwóch dwuspójnych składowych (ang. *articulation points*), (9) zidentyfikowane dwuspójne składowe (ang. *biconnected components*), w których występuje dokładnie jeden punkt artykulacji, są kolejno rozwiązywane z wykorzystaniem metody podziału i ograniczeń (ang. *branch-and-bound*); uzyskane rozwiązanie, obejmujące wybrane konformacje reszt zasad azotowych/nukleozydowych i odpowiadające im współczynniki energetyczne, jest zapamiętywane w skojarzonym punkcie artykulacji, (10) minimalizacja energetyczna w przestrzeni kartezjańskiej (opcjonalnie). W celu ewaluacji stopnia precyzji opracowanej metody [A1], w zależności od zastosowanych bibliotek konformacji reszt nukleotydowych, przeprowadzono eksperymenty obliczeniowe na reprezentatywnym zbiorze dziewiętnastu eksperymentalnie określonych struktur 3D RNA. Średnio *Clash score* zmniejszył się o ½ oraz ¾ w trybie modelowania reszt odpowiednio zasad azotowych oraz nukleozydowych. W większości przypadków wynik miary *RMSD*, wyznaczony w oparciu o podzbiór zrekonstruowanych atomów pomiędzy uzyskanym modelem 3D, a wejściową strukturą 3D RNA, nie przekracza 1.0Å. Rezultaty miary *INF* [E38] pozwalają stwierdzić, że struktura drugorzędowa jest prawidłowo rekonstruowana z przeszło 90% skutecznością we wszystkich rozpatrywanych przypadkach niezależnie od trybu predykcji. Warto podkreślić, że metoda najlepiej radziła sobie w przypadku wykorzystywania biblioteki skonstruowanej w oparciu o zbiór wysokorozdzielczych struktur 3D RNA, a więc kluczowym aspektem decydującym o sukcesie modelowania z wykorzystaniem *RNAfitme* jest wysoka jakość wprowadzonej struktury 3D szkieletu fosforo-cukrowego RNA. Uzyskane rezultaty [A1] potwierdziły, że *RNAfitme* jest niezawodnym narzędziem pozwalającym efektywnie rekonstruować struktury 3D dużych cząsteczek RNA oraz ich niewielkich, nieciągłych fragmentów prowadząc jednocześnie do zmniejszenia liczby zbyt poważnych zderzeń sterycznych.

4.3.3.5. Identyfikacja oraz strukturalne dopasowanie lokalnych motywów 3D białek/RNA.

W ogólności deskryptor struktur 3D białek/RNA opisuje lokalne, przestrzenne otoczenie określonej reszty aminokwasowej/nukleotydowej, zwanej resztą centralną, obejmując tylko z te segmenty łańcucha głównego zlokalizowane w jej przestrzennym sąsiedztwie [A5]. Koncepcja lokalnych deskryptorów [E45] została opracowana dla struktur 3D białek. Aczkolwiek traktując deskryptor jako motyw 3D spełniający pewien charakterystyczny wzorzec topologiczny oraz zakładając pewną elastyczność podczas opracowywania rozwiązań można z powodzeniem uzyskać dopasowywanie deskryptorów struktur 3D RNA, co zaprezentowano na Rysunku 11.

Rysunek 11. Przykład strukturalnego dopasowania deskryptorów struktur 3D RNA.

Deskryptory struktur 3D białek zostały z powodzeniem zastosowane do: (1) rozpoznawania kształtów 3D z wykorzystaniem cech wynikających z sekwencji [E45], (2) przewidywania kontaktów zachodzących pomiędzy resztami aminokwasowymi [E46], oraz (3) wyznaczania dopasowania struktur 3D białek, niezależnego od sekwencji [E47]. W przypadku cząsteczek RNA koncepcja deskryptorów wydaje się użyteczna do rozwiązania następujących problemów: (1) wyznaczania strukturalnego dopasowania, niezależnego od sekwencji, struktur 3D RNA, (2) odkrywania i klasyfikacji powtarzających się lokalnych podstruktur 3D w ramach kompleksów RNA-białko, miejsc wiązania RNA-ligand czy pseudowęzłów RNA, (3) odkrywania zależności pomiędzy sekwencją reszt nukleotydowych, a strukturą 3D RNA. Proces konstrukcji struktury 3D deskryptora obejmuje następujące etapy: (1) określenie reszty centralnej, (2) identyfikacja reszt, w analizowanej cząsteczce, które spełniają relację *w kontakcie* z określoną resztą centralną, (3) dla reszty centralnej i wszystkich innych reszt, w analizowanej cząsteczce, będących *w kontakcie* z tą pierwszą tworzone są krótkie, ciągłe fragmenty łańcucha głównego zwane elementami (domyślnie obejmujące pięć reszt; ten parametr można konfigurować zgodnie z oczekiwaniami użytkownika), (4) wszystkie nakładające się elementy tworzą dłuższe, ciągłe fragmenty łańcucha głównego zwane segmentami. Przykładowa struktura 3D deskryptora oraz podstawowe pojęcia wykorzystywane w procesie jego konstrukcji zostały zaprezentowane na Rysunku 12.



Rysunek 12. Ilustracja struktury 3D przykładowego deskryptora oraz podstawowych pojęć wykorzystywanych w procesie jego konstrukcji.

Relacja *w kontakcie* pomiędzy parą reszt cząsteczek biologicznych jest określana bazując na odległościach euklidesowych wyznaczonych pomiędzy wybranymi atomami analizowanych reszt. Eksperymentowanie z tą relacją jest kluczowe w zakresie dostosowania topologii i rozmiaru struktury 3D deskryptora do określonych zastosowań. Eksperymentowanie tego typu zwykle wymaga modyfikacji kodu źródłowego programu, co niestety może być kłopotliwe. Dlatego opracowano mechanizm pozwalający na definiowanie tej relacji w postaci wyrażenia (np. OR(DISTANCE:SCGC ≤ 6.5, AND(DISTANCE:SCGC ≤ DISTANCE:CA – 0.75, DISTANCE:SCGC ≤ 8.0))), wspierającego operatory relacyjne, logiczne i inne operatory specjalizowane np. *DISTANCE:P*, które można łatwo konfigurować bez potrzeby rekompilacji kodu programu. Następnie opracowano sposób przechowywania kluczowych właściwości struktur 3D deskryptorów, a mianowicie,

która reszta pełni rolę tej centralnej oraz jaki jest wykorzystywany rozmiar elementu, w formacie *PDB* [E6] oraz *PDBx/mmCIF*, który pozwala na odtwarzanie topologii deskryptora bezpośrednio na podstawie struktury 3D. W celu zapewnienia efektywności przetwarzania procesu konstruowania struktur 3D deskryptorów przeprowadzono optymalizację z wykorzystaniem mechanizmów współbieżności. Korzystając z opracowanego protokołu konstrukcji struktur 3D deskryptorów przygotowano zbiór struktur 3D deskryptorów na podstawie reprezentatywnego zestawu struktur 3D domen białkowych pochodzących z repozytorium *ASTRAL* [E48]. W ramach dalszych prac [A5] zajmowano się problemem identyfikacji największego strukturalnego dopasowania dla pary struktur 3D deskryptorów *A* i *B*, które spełnia poniższe założenia:

- porównywane deskryptory opisywane są przez zbiory składających się na nie elementów $A=\{a_*,a_1,\ldots,a_n\}$ i $B=\{b_*,b_1,\ldots,b_m\}$, gdzie a_* i b_* reprezentują centralne elementy rozpatrywanych deskryptorów.
- $RMSD(u,v)$ jest funkcją zwracającą wartość miary *RMSD* [E15] wyznaczoną dla motywów 3D *u* i *v* lub stałą wartość $M >> 3{,}5\text{Å}$, jeżeli motywów 3D nie da się porównać, ponieważ składają się z różnej liczby atomów.
- w każdym osiągalnym dopasowaniu strukturalnym zachodzi zależność $RMSD(a_*,b_*)\leq 1.2$.
- zbiory dupleksów dla porównywanych deskryptorów są definiowane jako: $D_A=\{d_{Ai}=(a_*,a_i): i=1,\ldots,n\}$, $D_B=\{d_{Bj}=(b_*,b_j): j=1,\ldots,m\}$.
- każde osiągalne dopasowanie strukturalne spełnia następujące założenia: (1) zawiera elementy centralne, (2) całkowita liczba elementów *N* wchodzących w skład dopasowania spełnia następujące nierówności: $N\geq\frac{4}{5}|A|$, $N\geq\frac{4}{5}|B|$, oraz (3) zawiera przynajmniej $\frac{2}{3}$ reszt obu rozpatrywanych deskryptorów.
- wartość miary *RMSD* wyznaczona zarówno dla wszystkich rozpatrywanych kombinacji par dupleksów (d_{Ai}, d_{Bj}) w ramach osiągalnego dopasowania strukturalnego jak i jego samego nie powinna przekraczać wartości $3{,}5\text{Å}$.

Przykładowa ilustracja instancji wejściowej problemu, zorientowanej na RNA, wraz z uzyskanym rozwiązaniem została zaprezentowana na Rysunku 11. Największe strukturalne dopasowanie pary struktur 3D deskryptorów cząsteczek biologicznych zostało zamodelowane jako problem optymalizacyjny oparty na problemie przydziału o maksymalnym rozmiarze (ang. *maximum-size assignment problem*) spełniającym poniższe ograniczenia programowania liniowego:

$$\text{maximize} \quad \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij},$$

$$\text{subject to} \quad \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}x_{ij} \leq L,$$

$$\sum_{i=1}^{n} x_{ij} \leq 1, \forall_{j=1,\ldots,m}$$

$$\sum_{j=1}^{m} x_{ij} \leq 1, \forall_{i=1,\ldots,n}$$

$$x_{ij} \in \{0,1\}, \forall_{i=1,\ldots,n,\ j=1,\ldots,m}$$

gdzie c_{ij} reprezentuje koszt dopasowania obiektów i oraz j, L reprezentuje ograniczenie skojarzone z całkowitym kosztem rozwiązania, oraz x_{ij} jest zmienną decyzyjną. W przypadku rozważanego problemu, $c_{ij} = \mathrm{RMSD}(d_{Ai}, d_{Bj})$, $i = 1,\ldots,n, j = 1,\ldots,m$, a wartość progowa L jest określana na podstawie oczekiwań użytkownika. Ponadto L zależy od wartości *min(n, m)*. Całkowita liczba elementów N wchodzących w skład rozwiązania jest wartością maksymalizowanej funkcji podwyższonej o jeden. Nie wszystkie założenia określone w sformułowaniu problemu są obecne w zaproponowanym modelu, ponieważ z powodzeniem można je weryfikować przed lub po znalezieniu dopuszczalnego rozwiązania. Następnie dowiedziono, że zaproponowany model może zostać rozwiązany w czasie wielomianowym [A5]. W ogólności problem przydziału może zostać rozwiązany w $O(n^3)$ z wykorzystaniem metody węgierskiej [E49]. W związku z powyższym największe strukturalne dopasowanie pary struktur 3D deskryptorów cząsteczek biologicznych może zostać znalezione w $O(n^4)$. Ostatecznie w pracy [A5] zaproponowano algorytm dokładny z powracaniem oraz trzy warianty algorytmu wielomianowego, bazującego na metodzie węgierskiej, (gdzie rozwiązaniem jest: pierwsze osiągalne rozwiązanie, najdłuższe dopasowanie w obrębie zbioru wszystkich osiągalnych rozwiązań, który w ostatnim wariancie został uzupełniony o osiągalne rozwiązania częściowe). Ogólna koncepcja wydajnych algorytmów została zaprezentowana na Rysunku 13.

```
Alignment hungarian-method-driven-search(int n, int m, float[][] CMA) {
    Alignment result;
    List<List<AlignedDuplexesPair>> assignments;
    for(int K = Math.min(n,m); K >= Math.floor((0.8 * (Math.max(n,m) + 1)) - 1); K--) {
        float[][] CA = transform(CMA, n, m, K);
        List<AlignedDuplexesPair> assignment = hungarianMethod.execute(CA);
        if (isAppropriateAssignment(assignment)) {
            assignments.add(assignment);
            if (firstAlignmentOnly) {
                break;
            }
        }
    }
    for(List<AlignedDuplexesPair> assignment: assignments) {
        Alignment alignment = constructAlignmentBasedOnAssignment(assignment);
        if (isAppropriateAlignment(alignment)) {
            result = alignment;
            break;
        }
    }
    return result;
}
```

Rysunek 13. Koncepcja algorytmów opartych na metodzie węgierskiej.

W celu ewaluacji precyzji oraz efektywności opracowanych algorytmów wielomianowych w kontekście algorytmu dokładnego przeprowadzono eksperymenty obliczeniowe dla reprezentatywnego zbioru strukturalnie podobnych par struktur 3D deskryptorów. Przeprowadzone eksperymenty jednoznacznie potwierdziły, że trzeci wariant algorytmu wielomianowego (z najwyższą wartością $L = 2{,}33 * K$) uzyskał wyniki zbliżone do algorytmu dokładnego pod względem jakościowym w zdecydowanie krótszym czasie [A5]. Opracowane rozwiązania udostępniono w postaci otwartego oprogramowania w repozytorium *GitHub* wraz z dokumentacją użytkową i scenariuszami użycia. Ponadto udostępniono również stronę internetową projektu. Progi liczbowe wykorzystane zarówno w sformułowaniu problemu jak

i modelu kombinatorycznym są traktowane jako parametry konfiguracyjne, które można łatwo dostosować do potrzeb użytkownika.

4.3.4. Podsumowanie.

Wszystkie początkowo zakładane cele badawcze zostały zrealizowane. Opracowano szereg nowych algorytmów, narzędzi i metod obliczeniowych poświęconych tematyce struktur 3D cząsteczek biologicznych, a w szczególności RNA udostępnionych w ramach publicznie dostępnych serwerów oraz w repozytoriach dedykowanych danych biologicznych. Możliwość wstępnego prognozowania przeprowadzanego przez ekspertów z wykorzystaniem precyzyjnych i efektywnych narzędzi obliczeniowych takich jak np. *RNAComposer* czy *RNApdbee* wspiera proces przygotowywania eksperymentów biochemicznych, co znacząco obniża ich koszty, a z drugiej strony zwiększa szansę na zrozumienie procesów biochemicznych zachodzących w komórce, a w końcu jak funkcjonują organizmy istot żywych. Badania w tym obszarze prowadzące do odkrycia algorytmu fałdowania się cząsteczek biologicznych są kluczowe dla zrozumienia wielu procesów w tym również procesów powodujących choroby, a co za tym idzie przyczyniają się do projektowanie nowych leków i rozwiązań terapeutycznych.

5. Omówienie pozostałych osiągnięć naukowo - badawczych.

5.1. Wybrane publikacje lub inne prace niewchodzące w skład osiągnięcia naukowego:

W skład pozostałego dorobku naukowego wchodzi **dziesięć** publikacji, które powstały w latach od 2012 do 2018. Dziewięć z nich zostało opublikowanych w renomowanych czasopismach z listy JCR z czego dwie przed uzyskaniem stopnia doktora. W przypadku poniższego zestawu publikacji **sumaryczny impact factor** oraz **całkowita liczba punktów MNiSW** stanowią odpowiednio **50,527** oraz **295** z czego po uzyskaniu stopnia doktora odpowiednio **33,441** oraz **215**.

[O1] S. Wasik, **M. Antczak**, J. Badura, A. Laskowski, T. Sternal, 14.04.2018, A Survey on Online Judge Systems and Their Applications, *ACM Computing Surveys*, 51(1), 3:1-3:34 (doi:10.1145/3143560). **[IF (2017): 5,550; MNiSW: 50, Q1]**

[O2] T. Zok, **M. Antczak**, M. Zurkowski, M. Popenda, J. Blazewicz, R.W. Adamiak, M. Szachniuk, 02.07.2018, *RNApdbee 2.0*: multifunctional tool for RNA structure annotation, *Nucleic Acids Research*, 46(W1), W30-W35 (doi:10.1093/nar/gky314). **[IF (2017): 11,561; MNiSW: 40, Q1]**

[O3] Z. Miao, R.W. Adamiak, **M. Antczak**, R.T. Batey, A. Becka, M. Biesiada, M.J. Boniecki, J.M. Bujnicki, S.-J. Chen, C.Y. Cheng, F.-C. Chou, A.R. Ferre-D'Amare, R. Das, W.K. Dawson, F. Ding, N.V. Dokholyan, S. Dunin-Horkawicz, C. Geniesse, K. Kappel, W. Kladwang, A. Krokhotin, G.E. Lach, F. Major, T.H. Mann, M. Magnus, K. Pachulska-Wieczorek, D.J. Patel, J.A. Piccirilli, M. Popenda, K.J. Purzycka, A. Ren, G.M. Rice, J.Jr. Santalucia, J. Sarzynska, M. Szachniuk, A. Tandon, J.J. Trausch, S. Tian, J. Wang, K.M. Weeks, B. Williams II,Y. Xiao, X. Xu, D. Zhang, T. Zok, E. Westhof, 30.01.2017, *RNA-Puzzles* Round III: 3D RNA structure prediction of five riboswitches and one ribozyme, *RNA*, 23(5), 655-672 (doi:10.1261/rna.060368.116). **[IF (2017): 4,490; MNiSW: 35, Q1, w 1% najlepiej cytowanych artykułów w WoS]**

[O4] M. Chlopkowski, **M. Antczak**, M. Slusarczyk, A. Wdowinski, M. Zajaczkowski, M. Kasprzak, 24.03.2016, High-order statistical compressor for long-term storage of DNA sequencing data, *RAIRO – Operations Research*, 50(2), 351-361 (doi:10.1051/ro/2015039). [**IF (2016): 0,550; MNiSW: 15, Q4**]

[O5] S. Wasik, **M. Antczak**, J. Badura, A. Laskowski, T. Sternal, 27.02.2016, *Optil.io*: Cloud Based Platform For Solving Optimization Problems Using Crowdsourcing Approach, in *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, San Francisco, USA, 433-436 (doi:10.1145/2818052.2869098).

[O6] P. Lukasiak, **M. Antczak**, T. Ratajczak, J. Blazewicz, 09-12.10.2015, *SphereGrinder* – reference structure-based tool for quality assessment of protein structural models, *Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine*, 665-668 (doi:10.1109/BIBM.2015.7359765). [**indeksowane w WoS bez IF; MNiSW: 15**]

[O7] P. Lukasiak, **M. Antczak**, T. Ratajczak, M. Szachniuk, M. Popenda, R.W. Adamiak, J. Blazewicz, 11.06.2015, *RNAssess* – a web server for quality assessment of RNA 3D structures, *Nucleic Acids Research*, 43(W1), W502-506 (doi:10.1093/nar/gkv557). [**IF (2015): 9,202; MNiSW: 40, Q1**]

[O8] K.J. Purzycka, M. Popenda, M. Szachniuk, **M. Antczak**, P. Lukasiak, J. Blazewicz, R.W. Adamiak, 04.02.2015, Automated 3D RNA structure prediction using the *RNAComposer* method for riboswitches, in Shi-Jie Chen, Donald H. Burke-Aguero (eds.) *Methods in Enzymology: Computational Methods for Understanding Riboswitches*, 553, 3-34 (doi: 10.1016/bs.mie.2014.10.050). [**IF (2014): 2,088; MNiSW: 20, Q3, rozdział na zaproszenie**]

Po uzyskaniu stopnia doktora ⬆

Przed uzyskaniu stopnia doktora ⬇

[O9] P. Lukasiak, **M. Antczak**, T. Ratajczak, J. M. Bujnicki, M. Szachniuk, M. Popenda, R.W. Adamiak, J. Blazewicz, 25.04.2013, *RNAlyzer* - novel approach for quality analysis of RNA structural models, *Nucleic Acids Research*, 41(12), 5978-5990 (doi:10.1093/nar/gkt318). [**IF (2013): 8,808; MNiSW: 40, Q1**]

[O10] M. Popenda, M. Szachniuk, **M. Antczak**, K.J. Purzycka, P. Lukasiak, N. Bartol, J. Blazewicz, R.W. Adamiak, 26.04.2012, Automated 3D structure composition for large RNAs, *Nucleic Acids Research*, 40(14):e112 (doi:10.1093/nar/gks339). [**IF (2012): 8,278; MNiSW: 40, Q1, w 1% najlepiej cytowanych artykułów w WoS**]

Wkład habilitanta w powstanie powyższych publikacji znajdują się w załączniku nr 3.

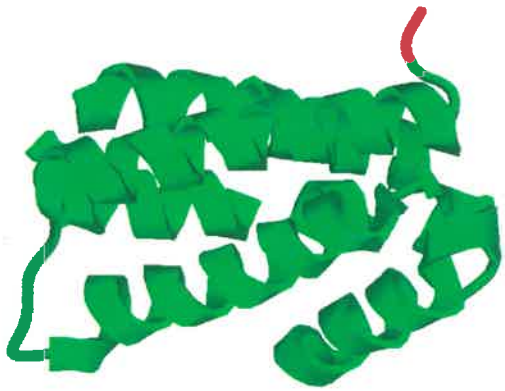
Publikacja o *RNApdbee 2.0* nie została włączona do cyklu reprezentującego osiągnięcie naukowe, ponieważ wkład habilitanta szczególnie podczas przygotowywania wstępnej wersji manuskryptu nie był wystarczający ze względu na opiekę nad żoną po urodzeniu dziecka.

5.2. Opis pozostałych osiągnięć naukowo - badawczych.

5.2.1. Opis badań realizowanych przed uzyskaniem stopnia doktora.

5.2.1.1. Przewidywanie granic domen w białkach na podstawie sekwencji reszt aminokwasowych.

Znajomość struktury przestrzennej białka jest kluczowa w celu identyfikacji pełnionej przez nie funkcji. Domeny zwykle są postrzegane jak kompaktowe moduły przestrzenne występujące w białkach, które mogą fałdować się niezależnie oraz mogą stanowić jednostki ewolucji [E50]. Większość białek zawiera wiele domen, które są zwykle skojarzone z pełnioną przez nie funkcją. Umiejętność wiarygodnego przewidywania domen w białku często decyduje o sukcesie jego klasyfikacji funkcjonalnej czy modelowania homologicznego. W ramach przeprowadzonych badań opracowana została efektywna metoda obliczeniowa [E51], która w wejściowej sekwencji reszt aminokwasowych, pozwala odkrywać obiecujące fragmenty sekwencyjne, a następnie je stopniowo rozszerzać, aż do uzyskania odpowiednio wysokiego współczynnika identyczności sekwencyjnej z wybranym, potwierdzonym fragmentem sekwencyjnym występującym na granicy pomiędzy sąsiadującymi domenami w dowolnym spośród rozpatrywanych, eksperymentalnie określonych, białek. Połączenie automatycznego algorytmu przeszukiwania wzorców sekwencji ze specjalizowaną relacyjną bazą danych, przechowującą uogólnione wzorce fragmentów sekwencyjnych łączących sąsiadujące domeny białkowe, wyznaczone na podstawie potwierdzonych repozytoriów *CATH* [E52], *DALI* [E53], *PFAM* [E54], *SCOP* [E55], pozwala na odkrywanie regionów reprezentujących domeny (zarówno ciągłe jak i nieciągłe) w nieokreślonej sekwencji białkowej bez potrzeby przewidywania jej struktury 3D. Jakość opracowanej metody została zweryfikowana podczas konkursu międzynarodowego *CASP7* (*The Critical Assessment of protein Structure Prediction*) [E56], którego celem jest określanie możliwości i ograniczeń aktualnie dostępnych metod służących do modelowania i analizy struktur białek. Dla dwudziestu sześciu białek, rozpatrywanych w konkursie, zaprezentowana metoda uzyskała najlepszy rezultat. Przykładowa predykcja uzyskana dla białka T0367 (PDB:2HSB) została zaprezentowana na Rysunku 14. Szczegółowy opis podejścia i uzyskanych wyników znajduje się w pracy [E51].



Rysunek 14. Wizualizacja predykcji uzyskanej dla białka T0367 (PDB:2HSB). Kolor zielony oraz czerwony opisuje odpowiednio prawidłowo przewidzianą domenę oraz błąd predykcji [E51].

5.2.1.2. Obliczeniowe przewidywanie struktur 3D białek.

Białka pełnią wiele różnych funkcji w komórce. Liczba eksperymentalnie określonych struktur 3D białek, zdeponowanych w *PDB* (*Protein Data Bank* [E6]), jest tylko niewielkim ułamkiem znanych sekwencji białkowych. Szeroko wykorzystywane techniki eksperymentalne, takie jak np. krystalografia rentgenowska czy spektroskopia magnetycznego rezonansu jądrowego, nie zawsze udaje się z powodzeniem zastosować dla niektórych rodzajów białek. Ponadto są to zwykle złożone, czasochłonne i kosztowne eksperymenty. W związku z powyższym opracowywanie wiarygodnych i efektywnych metod obliczeniowych dedykowanych problemowi przewidywania struktur 3D białek na podstawie sekwencji reszt aminokwasowych jest kluczowe, aby minimalizować rozdźwięk pomiędzy liczbami znanych sekwencji oraz eksperymentalnie określonych struktur 3D białek. W ramach

przeprowadzonych badań opracowane zostało nowe podejście obliczeniowe dedykowane konstrukcji modeli 3D białek bazujące na koncepcji deskryptorów struktur białkowych. Lokalna podstruktura 3D, nazywana deskryptorem, jest reprezentowana przez zbiór kilku stosunkowo krótkich nieciągłych fragmentów łańcucha polipeptydowego. Każdy deskryptor opisuje lokalne otoczenie przestrzenne określonej reszty aminokwasowej, zwanej resztą centralną, zawierając tylko te fragmenty łańcucha głównego białka, które są zlokalizowane w sąsiedztwie reszty centralnej [E45, A5]. Opracowana metodologia [E57] obejmuje następujące fazy: (1) sekwencyjne dopasowanie deskryptorów do wejściowej sekwencji reszt aminokwasowych tworzące zbiór obiecujących przypisań, (2) filtrowanie mało wiarygodnych przypisań deskryptorów, oraz (3) konstrukcja wstępnego modelu 3D białka (jego rdzenia) poprzez asemblację potwierdzonych przypisań deskryptorów. Podstawowym komponentem, na którym opiera się metoda, jest biblioteka grup struktur 3D deskryptorów. Grupa deskryptorów reprezentuje zbiór deskryptorów strukturalnie podobnych do tzw. deskryptora założycielskiego grupy. Przykładowa grupa deskryptorów została przedstawiona na Rysunku 15.

Descriptor name	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Segment 6
d1p1da2_A_206_LEU	FHVKLPK	LGITI	DPLVISD	SVAHRTGTLEL	DKLLAIDN	QILQQCEDLVKLKIRK
d1q5aa__A_679_VAL	KTVLLQK	FGFVL	..QYLES	GVAWR.AGLRM	DFLIEVNG	NMIRQ..NTLMVRVVM
d1y7na1_A_84_MET	TTVLIRR	LGPSV	..SIICS	GIAER.GGVRV	HRIIEING	HILGN..GEIRMKTMP
d1x6da1_A_98_ILE	HVTILHK	AGLGF	..ITVHR	GLASQ.GTIQK	NEVLSING	RQARE..RQAVIVTRK
d1v62a__A_96_LEU	..VEIVK	LGISL	..ITIDR	SVVDR.GALMP	DHILSIDG	KLLASISEKVRLEILP
d1w9ea1_A_188_MET	REVILCK	LRLKS	..IFVQL	SFASL.VGLRF	DQMLQING	KVLKQ..EKITWIIRD
d2cssa1_A_110_ILE	GRVILNK	LKVVG	..AFITK	GLADVVGKLRA	DEVLEWNG	NIILE..PQVEIIVER
d1ufla__A_98_LEU	KKVNLVL	LTIRG	..IYITG	SEAEG.SGLKV	DQILEVNG	RLLKS..RHLILTVKD

Character '*' means that there is no structural mapping for a particular residue between the founder and the group member

Rysunek 15. Przykładowa grupa deskryptorów reprezentowana przez dopasowanie sekwencyjne (z lewej) oraz wizualizację struktur 3D deskryptorów w niej zawartych nałożonych na strukturę 3D deskryptora założycielskiego (z prawej) [A5].

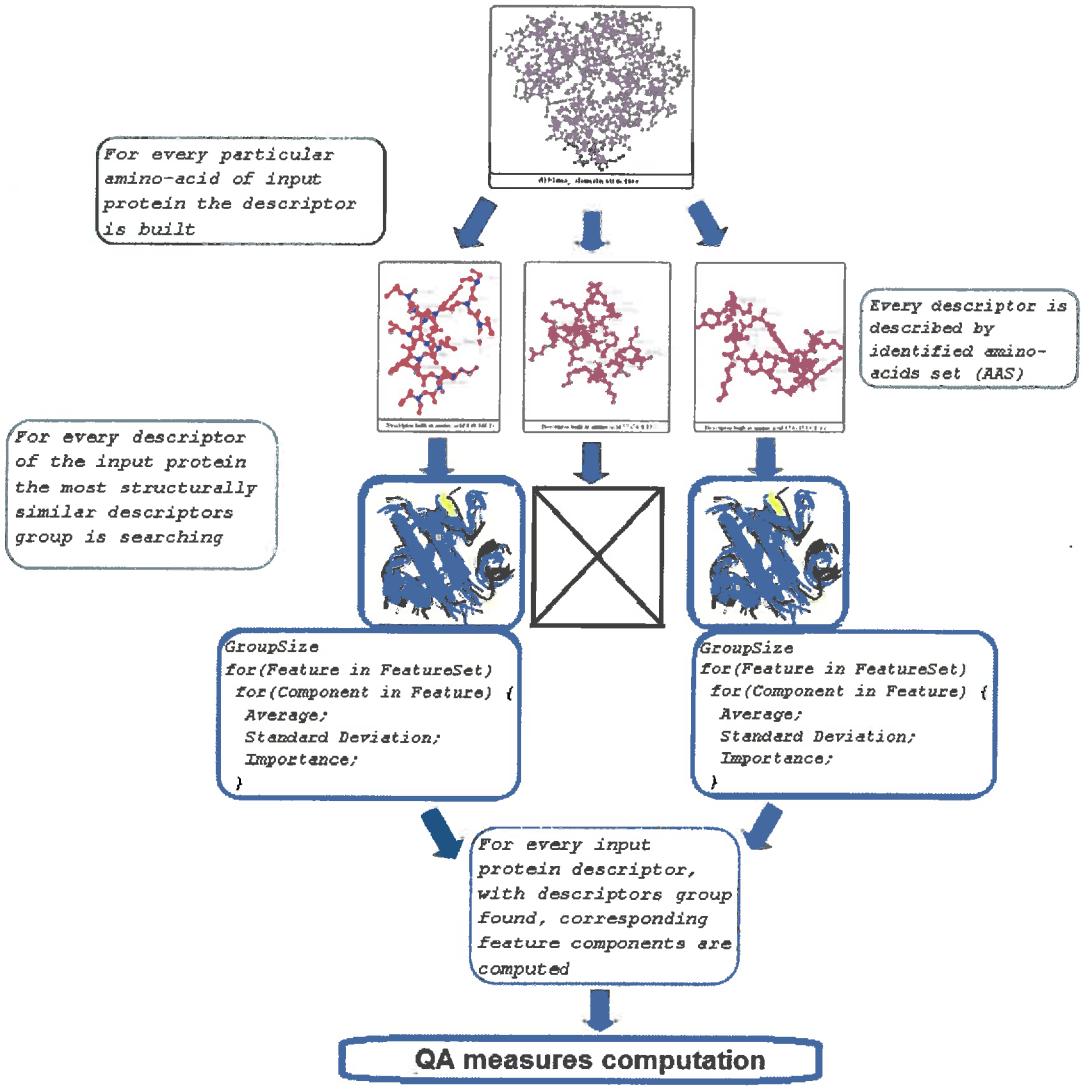
W literaturze opublikowana została wielokryterialna funkcja [E45] pozwalająca odpowiedzieć na pytanie czy dana para deskryptorów jest w relacji strukturalnego podobieństwa czy też nie. W ramach prowadzonych prac zostały opracowane wstępne podejścia algorytmiczne dedykowane problemowi strukturalnego dopasowywania struktur 3D deskryptorów, który został dogłębnie zanalizowany i efektywnie rozwiązany w [A5]. Dla każdej reszty aminokwasowej występującej w strukturach 3D domen, pochodzących z eksperymentalnie określonych białek przechowywanych w bazie *ASTRAL* [E48], skonstruowana została struktura 3D deskryptora wchodząca w skład zbioru. Następnie deskryptory zostały pogrupowane względem relacji strukturalnego podobieństwa tworząc grupy wchodzące w skład biblioteki. Finalną postać biblioteki uzyskano w wyniku odfiltrowania grup, których liczność była niższa od założonego progu. W pierwszej fazie wykorzystany został *FragHMMent* [E46]. W ramach drugiej fazy opracowany został klasyfikator bazujący na maszynie wektorów wspierających (ang. *Support Vector Machine*). W ogólności przygotowane zostały dwa zbiory przykładów zarówno pozytywnych jak i negatywnych. Każdy przykład składał się ze struktury 3D deskryptora oraz przypisanej dla niego sekwencji. W skład przykładów pozytywnych wchodzą: (1) natywne deskryptory, (2) deskryptory, których struktura 3D jest rekonstruowana z wykorzystaniem narzędzia *SCWRL* [E40] (a) przy jednoczesnym zachowaniu ich rzeczywistej sekwencji, albo (b) zastosowaniu sekwencji innego strukturalnie podobnego deskryptora w grupie. W skład przykładów negatywnych wchodzą: deskryptory po rekonstrukcji narzędziem *SCWRL* zakładając, że sekwencja: (a) pochodzi z deskryptora wchodzącego w skład innej strukturalnie niepodobnej

grupy, (b) została wylosowana zachowując rzeczywisty rozkład aminokwasów w znanych sekwencjach białek, oraz (c) została całkowicie losowo wygenerowana. W celu ewaluacji jakościowej struktur 3D deskryptorów, w rozpatrywanych zbiorach, zastosowano potencjały wywodzące się z mechaniki molekularnej (potencjały *Lennarda-Jonesa* [E58] i elektrostatyczny) oraz statystyczny (*DFIRE* [E59]). Ponadto opracowano algorytm udokładniania współrzędnych atomów łańcuchów bocznych zrekonstruowanych konformacji deskryptorów w przestrzeni kątów torsyjnych z wykorzystaniem algorytmu *Metropolis* [E60]. W ostatniej fazie opracowano jedynie wymagania w zakresie hierarchicznej metody asemblacji przypisanych struktur 3D deskryptorów, charakteryzujących się najwyższym współczynnikiem wiarygodności, na podstawie nakładających się fragmentów przestrzennych. Proponowane podejście powinno współbieżnie eksplorować wiele ścieżek budując różne składowe struktury 3D oceniane, na podstawie konsensusu, na każdym etapie w celu identyfikacji jednej (lub wielu) obiecującej(ych) składowej(ych) struktury(r) 3D, która(e) powinna(y) być dalej rozbudowywana(e). Przeprowadzono wstępne eksperymenty obliczeniowe w celu weryfikacji precyzji zarówno klasyfikatora jak i algorytmu wygładzania współrzędnych łańcuchów bocznych konformacji deskryptorów, które zakończyły się sukcesem i wykazały, że proponowana koncepcja wydaje się obiecująca, ale badania nad jej rozwojem powinny być dalej kontynuowane [E57].

5.2.1.3. Ogólna ewaluacja jakości modelu(i) 3D białek.

Ciągły rozwój w dziedzinie obliczeniowego przewidywania struktur 3D cząsteczek biologicznych, a w szczególności białek powoduje, że dysponuje się coraz większą różnorodnością dostępnych modeli przestrzennych. Niestety często zdarza się, że modele 3D uzyskane metodami obliczeniowymi są niedokładne wykazując znaczne odchylenia od referencyjnej struktury 3D, o ile takowa została eksperymentalnie określona. W związku z powyższym pojawia się problem w jaki sposób wybrać model 3D, który jest strukturalnie najbliższy eksperymentalnie określonej strukturze 3D, którą przecież się nie dysponuje w ogólnym przypadku, spośród zbioru dostępnych modeli 3D uzyskanych metodami predykcji *in silico*? To jest kluczowe pytanie adresowane w problemie ogólnej ewaluacji jakości modelu(i) 3D białek (bez posiadania referencyjnej struktury 3D). Ponadto ewaluacja zwykle dokonywana jest na dwóch poziomach, a mianowicie globalnym oraz lokalnym, gdzie wyznaczana jest jedna mierzalna wartość szacująca prawidłowość strukturalną odpowiednio całego modelu 3D oraz otoczenia przestrzennego poszczególnych reszt aminokwasowych. W ramach przeprowadzonych badań opracowano nową metodę obliczeniową pozwalającą szacować jakość strukturalną pojedynczego modelu 3D białka bez posiadania referencyjnej struktury 3D, zarówno z globalnej jak i lokalnej perspektywy, z wykorzystaniem koncepcji lokalnych deskryptorów struktur białkowych [E61, E16]. Podczas prowadzonych prac zaadresowano trzy kluczowe pytania: (1) w jaki sposób wyznaczyć kontekst przestrzenny (substytut struktury referencyjnej), który posłuży do przeprowadzenia oceny?, (2) jakie cechy strukturalne lub znane potencjały molekularne/statystyczne zastosować w celu wiarygodnego rozróżniania geometrii prawidłowych od nieprawidłowych?, (3) jakiego rodzaju oddziaływania międzycząsteczkowe (poziom szczegółowości) powinny być rozpatrywane? Jako kontekst ponownie wykorzystana została biblioteka grup struktur 3D deskryptorów, a w szczególności jej warianty różniące się zastosowaną wartością progu minimalnej liczby deskryptorów wchodzących skład grupy. Budowanie biblioteki jest złożonym obliczeniowo zadaniem, ponieważ wymaga przeprowadzenia

wielu strukturalnych dopasowań pomiędzy parami deskryptorów, dlatego zastosowano mechanizmy współbieżności w celu optymalizacji przetwarzania w zakresie dostępnych zasobów, a mianowicie rozmiaru wolnej pamięci operacyjnej oraz liczby dostępnych jednostek obliczeniowych. Porównanie różnych wariantów bibliotek grup deskryptorów zostało przeprowadzone z wykorzystaniem współczynnika pokrycia wyznaczanego poprzez liczbę aminokwasów pokrytych przez deskryptory (tylko założycielskie, lub wszystkie występujące w bibliotece) w stosunku do liczby aminokwasów wszystkich domen białkowych rozpatrywanych podczas konstruowaniu biblioteki [E16]. W zakresie oceny geometrii zdecydowano się zastosować podstawowe własności strukturalne, a mianowicie gęstość upakowania, całkowity ładunek, hydrofobowość reszt aminokwasowych oraz podobieństwo sekwencyjne i struktury drugorzędowej przy wykorzystaniu macierzy substytucji. W zakresie rodzajów oddziaływań analizowanych w ramach grupy wybrano: (a) odpowiadające sobie aminokwasy w dopasowaniu sekwencyjnym, (b) pary aminokwasów będące w kontakcie przestrzennym w ramach deskryptora założycielskiego, (c) aminokwas centralny oraz wszystkie inne aminokwasy w deskryptorze założycielskim będące z nim w kontakcie przestrzennym, (d) wszystkie unikalne reszty aminokwasowe zidentyfikowane w parach z podpunktu (b), (e) cała struktura 3D deskryptora założycielskiego. Dla każdej grupy deskryptorów, rozpatrywanej w bibliotece, wyznaczana jest dystrybuanta rozkładu normalnego dla kolejnych atrybutów reprezentowanych przez każdą kombinację par (cecha/funkcja oceny, rodzaj oddziaływania), które następnie będą wykorzystywane do rozpoznawania odchyleń, na podstawie struktur 3D deskryptorów występujących w grupie [E16]. Opracowana metoda [E61, E16], zaprezentowana na Rysunku 16, składa się z następujących etapów: (a) konstrukcja struktur 3D deskryptorów dla każdej reszty aminokwasowej w danym wejściowym modelu 3D białka, (b) poszukiwanie strukturalnie podobnej grupy w bibliotece dla każdego deskryptora skonstruowanego w poprzednim etapie, (c) dla każdego deskryptora, pochodzącego z modelu 3D, dokonuje się oceny jego wiarygodności/jakości w kontekście strukturalnie podobnej grupy o ile takową udało się zidentyfikować, (d) wiarygodność danego deskryptora, pochodzącego z modelu 3D, jest wyznaczana jako średnia ważona zbioru rozpatrywanych atrybutów, (e) lokalna wiarygodność danej reszty jest wyznaczana jako średnia wiarygodności uzyskanych dla tej reszty podczas ewaluacji przynajmniej dwóch deskryptorów, w których ta reszta została zidentyfikowana, (f) wiarygodność globalna modelu jest wyznaczana jako średnia lokalnych wiarygodności wszystkich ocenionych reszt . Procesy konstrukcji struktur 3D deskryptorów, wyszukiwania dla nich strukturalnie podobnych grup w bibliotece oraz wyznaczanie dla nich współczynnika wiarygodności zostały ponownie zoptymalizowane z wykorzystaniem mechanizmów współbieżności [E16]. W celu zweryfikowania efektywności i użyteczności scharakteryzowanej metody w zakresie szacowania globalnej wiarygodności strukturalnej modelu 3D przeprowadzono eksperymenty obliczeniowe, w których wyznaczono korelacje Pearsona i Spearmana pomiędzy rezultatem nowo opracowanej metody oraz szeroko wykorzystywana miarą *GDT_TS* (*Global Distance Test – Total Score*) [E62], która ocenia jak bardzo oddalony jest dany model 3D od odpowiadającej referencyjnej struktury 3D białka. Eksperymenty przeprowadzono na reprezentatywnych zbiorach eksperymentalnie określonych białek i nadesłanych dla nich modeli 3D podczas konkursu *CASP9* w kategorii *Template Based Modeling* oraz *Template Free Modeling* [E63].

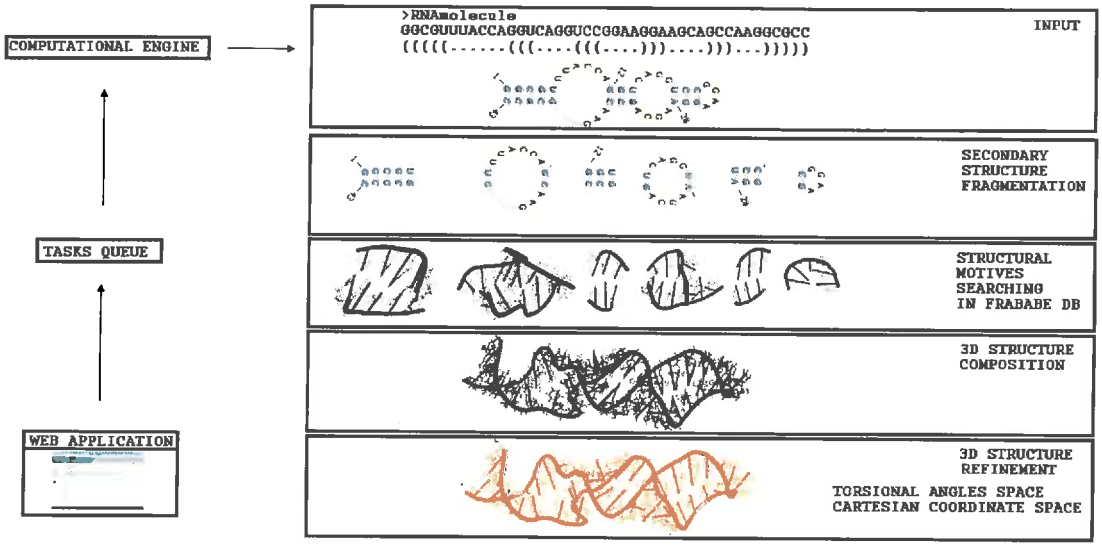


Rysunek 16. Schemat blokowy metody dedykowanej problemowi ogólnej ewaluacji jakości struktur 3D białek [E16].

Uzyskane wyniki pozwoliły zaobserwować, że opracowana metoda sprawdza się dość dobrze jeżeli modele 3D w analizowanym zbiorze zauważalnie różnią sie od siebie. Ponadto najlepiej sprawdzają się pojedyncze cechy takie jak podobieństwo sekwencyjne oraz struktury drugorzędowej, uśredniony współczynnik hydrofobowości, oraz średnia ważona spośród wszystkich rozpatrywanych cech/funkcji strukturalnych. W przypadku struktur 3D deskryptorów skonstruowanych w ramach białek należących do drugiej kategorii zdecydowanie trudniej było zidentyfikować strukturalnie podobne grupy, w ramach dysponowanych bibliotek, co spowodowało, że wrażliwość metody pogorszyła się [E16]. Obiecujące wyniki spowodowały, że opracowano prototyp nowego narzędzia wspomagające proces modelowania homologicznego oraz udoskonalania modeli 3D białek, które pozwalało, między innymi, prezentować wyniki lokalnej ewaluacji reszt aminokwasowych w postaci wykresów słupkowych oraz wizualizować oddziaływania strukturalne zachodzące w analizowanym modelu 3D z wykorzystaniem grafu oraz profili gęstości. Szczegółowy opis i przykładowe wizualizacje można znaleźć w [E61, E16].

5.2.1.4. Automatyczne przewidywanie struktur 3D na podstawie sekwencji i struktury drugorzędowej RNA.

Zrozumienie mechanizmów funkcjonowania cząsteczek RNA wymaga znajomości ich struktury 3D. W przypadku cząsteczek RNA stosuje się dokładnie te same metody eksperymentalne co w przypadku białek. Niestety przepaść pomiędzy liczbami odpowiednio znanych sekwencji reszt nukleotydowych RNA, a odpowiadających im eksperymentalnie określonych struktur 3D jest jeszcze większa niż w przypadku białek, ponieważ strukturalna różnorodność oraz dynamika cząsteczek RNA w kontaktach z jonami, otoczeniem i innymi cząsteczkami biologicznymi powoduje, ze przewidywanie struktury 3D RNA z rozdzielczością atomową jest problemem z pewnością nie łatwiejszym niż w przypadku białek [E64]. W 2012 r. istniejące algorytmy przewidywania struktury 3D RNA często wymagały manualnej manipulacji dokonywanej przez eksperta lub zwracały wiarygodne modele 3D dla niezbyt skomplikowanych cząsteczek RNA [E65]. Różniły się one głównie pod względem danych wejściowych (sekwencja reszt nukleotydowych, które może zostać uzupełniona strukturą drugorzędową lub mapą kontaktów przestrzennych uzyskanych w wyniku przeprowadzenia eksperymentów biochemicznych), reprezentacji nukleotydu/rozdzielczości predykcji (modele gruboziarniste lub pełnoatomowe), oraz stosowanej metodologii (modelowanie homologiczne, *ab initio*, *de novo*). Żadna z nich nie osiągnęła stadium pełnej automatyzacji umożliwiając przewidywanie struktur 3D dużych cząsteczek RNA (500 nts) charakteryzujących się wysoką rozdzielczością, prawidłową geometrią i stereochemią kwasów nukleinowych przy jednoczesnej optymalizacji czasu przetwarzania (rzędu minut dla największych wspieranych rozmiarów instancji) [O10]. W ramach prowadzonych badań opracowano nową metodę obliczeniową, spełniającą powyższe wymagania, pozwalającą na w pełni automatyczne przewidywanie struktury 3D cząsteczki RNA opisanej przez sekwencję i strukturę drugorzędową udostępnioną w ramach efektywnego serwera obliczeniowego, o nazwie *RNAComposer*, otwartego dla społeczności naukowej [O10]. Warto podkreślić, że optymalizacja przetwarzania jest kluczowym wymaganiem pozafunkcjonalnym, ponieważ system pozwala na zgłaszanie zadań w trybie wsadowym multiplikujących liczbę (do 10): (1) sekwencji reszt nukleotydowych, (2) struktur drugorzędowych i (3) wynikowych modeli 3D RNA co w rezultacie pozwala uzyskać aż do 1000 modeli 3D RNA w ramach jednego zgłoszonego zadania. Podczas projektowania rozwiązania zdecydowano się również udostępnić tryb predykcji struktury 3D RNA na podstawie jedynie sekwencji reszt nukleotydowych. W związku z powyższym zostały przeprowadzone eksperymenty obliczeniowe na zbiorze reprezentatywnych, eksperymentalnie określonych struktur 3D RNA, zdeponowanych w *PDB* [E6], w celu określenia trzech metod *in silico* dedykowanych problemowi przewidywania struktur drugorzędowych RNA, które następnie zostały zintegrowane w module obliczeniowym systemu *RNAComposer* [O10]. Koncepcja opracowanej metody obliczeniowej została zaprezentowana na Rysunku 17. Składa się z następujących etapów: (1) fragmentacja wejściowej struktury drugorzędowej na elementy topologiczne, (2) wyszukiwanie najbardziej odpowiednich elementów strukturalnych (szereg kryteriów jest optymalizowanych np. zgodność sekwencyjna, rozdzielczość eksperymentalna), w repozytorium relacyjnym *RNA FRABABE* [O10, E16], dla każdego elementu topologicznego zidentyfikowanego w poprzednim etapie, (3) jeżeli występuje niezgodność sekwencyjna w wybranym elemencie strukturalnym wtedy wszystkie nieprawidłowe sekwencyjnie zasady są podmieniane, (4) jeżeli w repozytorium nie



Rysunek 17. Koncepcja metody obliczeniowej udostępnionej w ramach systemu *RNAComposer* [O10, E16].

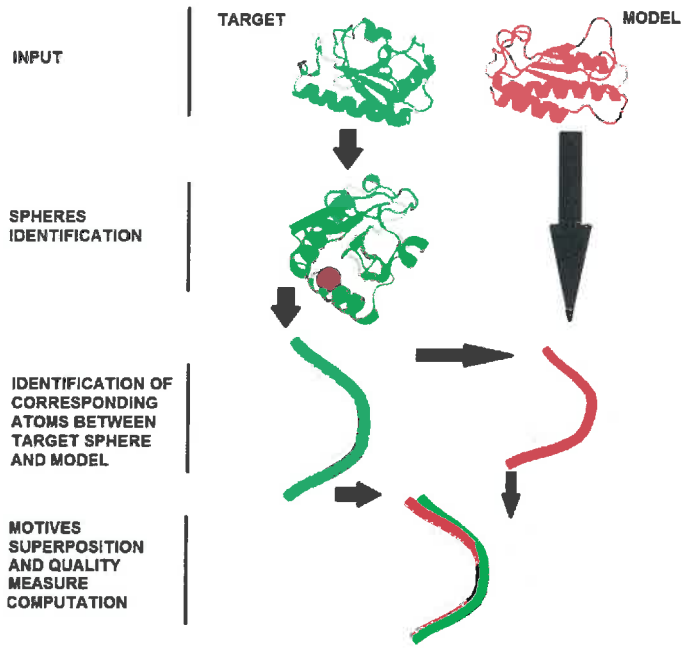
występują elementy strukturalne odpowiadające określonemu elementowi topologicznemu wtedy uruchamiane są dodatkowe mechanizmy pozwalające na wygenerowanie brakujących elementów strukturalnych o określonej sekwencji i topologii struktury drugorzędowej RNA, (5) asemblacja wstępnego modelu na podstawie wybranych elementów strukturalnych, (6) minimalizacja energetyczna w przestrzeniach zarówno kątów torsyjnych jak i euklidesowej prowadząca do uzyskania finalnego modelu 3D. W ramach projektowania i implementacji modułu obliczeniowego systemu *RNAComposer* opracowano szereg rozwiązań [E16], w skład których wchodzą: (1) obiektowo-relacyjne repozytorium elementów strukturalnych, o nazwie *RNA FRABABE* [O10, E16], pozwalające na efektywne wyszukiwanie elementów 3D o określonej topologii struktury drugorzędowej RNA spełniających dodatkowe kryteria kluczowe w procesie przewidywania struktury 3D RNA, (2) narzędzie automatyzujące proces migracji danych pochodzących z macierzystej bazy fragmentów strukturalnych *RNA FRABASE* [E26, E27], (3) algorytmy pozwalające na: (a) fragmentację wejściowej struktury drugorzędowej RNA zapisanej w notacji kropkowo-nawiasowej z wykorzystaniem wyrażeń regularnych, (b) efektywne wielokryterialne wyszukiwanie elementów strukturalnych w repozytorium *RNA FRABABE*, (c) systematyczną asemblację składowych elementów strukturalnych w większe fragmenty 3D, na podstawie grafowej reprezentacji wejściowej struktury drugorzędowej RNA, przy jednoczesnej minimalizacji uszkodzeń strukturalnych, wynikających ze zmiennoprzecinkowej reprezentacji współrzędnych atomowych, (4) warstwa zarządcza odpowiedzialna za integrację, adaptowanie przepływających danych, oraz jednoczesne uruchamianie wielu, często złożonych, narzędzi zewnętrznych w ramach heterogenicznej architektury obliczeniowej optymalizując zużycie aktualnie dostępnych zasobów w zakresie pamięci operacyjnej i liczby jednostek obliczeniowych. Moduł obliczeniowy został poddany szerokiej ewaluacji wydajnościowej w celu identyfikacji wąskich gardeł i przeprowadzenia optymalizacji w zakresie: (a) wykorzystania pamięci operacyjnej, (b) konstruowania i wykonywania zapytań do relacyjnego repozytorium, (c) przetwarzania zadań w trybie wsadowym, z wykorzystaniem mechanizmów współbieżności, [E16]. Użyteczność i efektywność proponowanej metody została oceniona na podstawie szeregu przeprowadzonych

eksperymentów obliczeniowych [O10, E16]: (1) ewaluacja jakości przewidywania struktur 3D RNA na podstawie reprezentatywnego zbioru czterdziestu eksperymentalnie określonych cząsteczek RNA, zdeponowanych w *PDB* [E6], (2) ewaluacja jakości i czasu przetwarzania w kontekście innych, publicznie dostępnych, metod przewidywania struktur 3D RNA, a mianowicie *MC-Fold/MC-Sym* [E10] i *iFoldRNA* [E9], (3) przedstawienie możliwości systemu na podstawie zbioru dziewięćdziesięciu pięciu losowo wygenerowanych sekwencji RNA o rozmiarze do 500nts, (4) ewaluacja wydajnościowa współbieżnej wersji systemu. Przewidywane modele 3D RNA wykazują wysoką jakość odzwierciedlaną zarówno stosunkowo niskimi wartościami globalnego *RMSD* [E15] jak i wiernością odwzorowania oddziaływań przestrzennych zasad (INF_{all}, INF_{cbp} [E38]) w kontekście referencyjnej struktury 3D RNA [O10, E16].

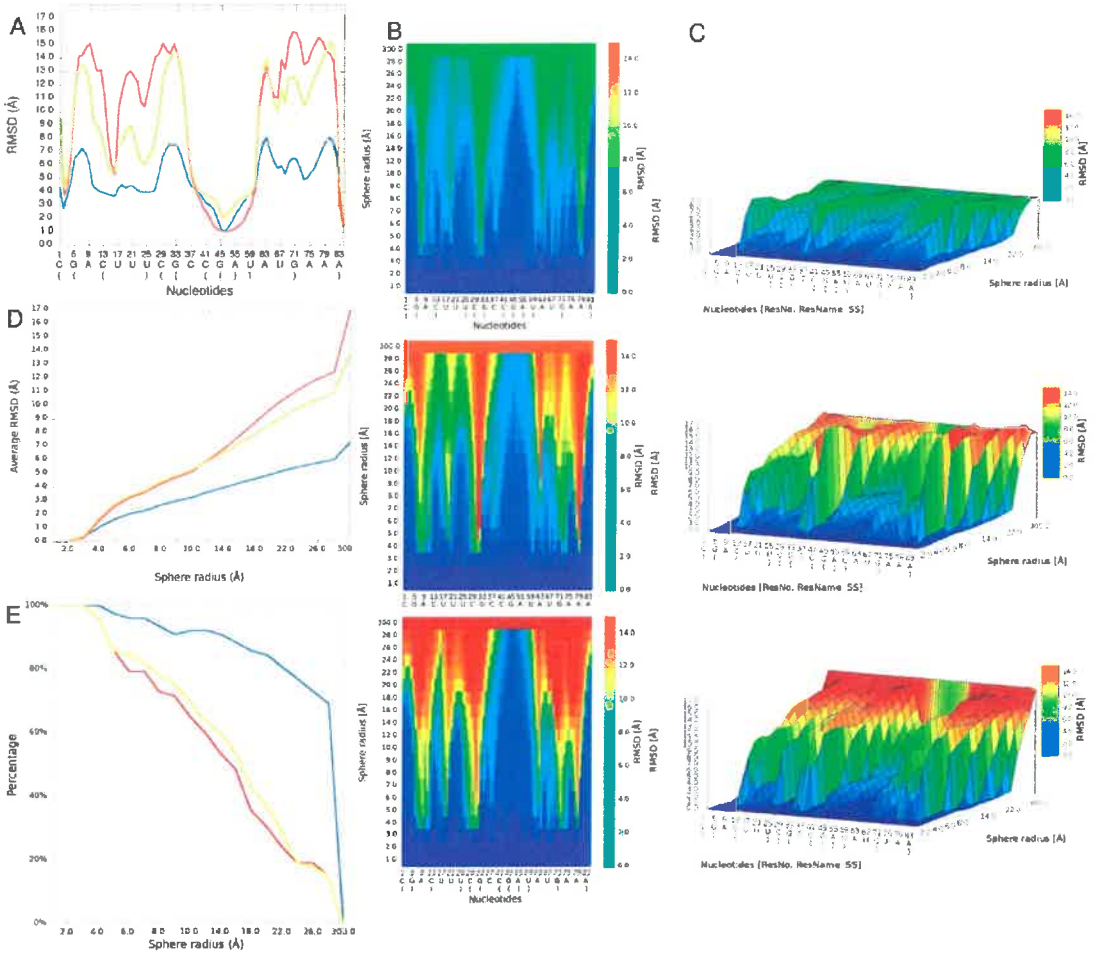
5.2.1.5. Wizualna ewaluacja precyzji modelu(i) 3D w kontekście referencyjnej struktury 3D białek/RNA.

Stale rosnąca liczba metod obliczeniowych, dedykowanych problemowi przewidywania struktur 3D cząsteczek biologicznych, stwarza potrzebę ich obiektywnej oceny prowadzącej do identyfikacji ich atutów i ograniczeń, które są kluczowe w zakresie ich potencjalnego zastosowania. W tym celu zostały powołane dwa międzynarodowe konkursy, a mianowicie *Critical Assessment of protein Structure Prediction* [E66] oraz *RNA-Puzzles* [E7, E8, O3], których celem jest kompleksowa ewaluacja aktualnie dostępnych metod przewidywania struktur 3D odpowiednio białek i RNA. Istnieje wiele różnych podejść pozwalających na globalną ewaluację modelu 3D w oparciu o referencyjną strukturę 3D zarówno dla białek jak i dla RNA [E16]. Niestety istniejące miary zazwyczaj koncentrują się na analizie pewnych, wybranych aspektów co powoduje, że wiarygodna ewaluacja wymaga manualnej agregacji wyników szeregu dostępnych miar. Jedną z najczęściej wykorzystywanych globalnych miar jest *Root-Mean-Square-Deviation (RMSD)* [E15], która reprezentuje uśrednioną odległość pomiędzy dwoma zbiorami odpowiadających sobie atomów, które wcześniej zostały poddane optymalnej superpozycji. Niestety podczas interpretacji należy mieć świadomość, że np. jej wynik zależy od długości łańcucha (liczby rozpatrywanych atomów), czy też globalna superpozycja powoduje, że błędy rozkładają się wzdłuż całej struktury co znacznie utrudnia ich lokalizację i interpretacje. W przypadku cząsteczek RNA miara *RMSD* pozwala na uchwycenie ogólnego kształtu przestrzennego, natomiast niewiele może powiedzieć na temat wierności odwzorowania kluczowych oddziaływań w modelu 3D w zakresie par zasad (ang. *base-pairing*) lub wzajemnej ich orientacji (ang. *base-stacking*). Z drugiej strony miary oparte na analizie oddziaływań wewnątrzcząsteczkowych nie są na tyle wrażliwe, aby zidentyfikować nieprawidłowości występujące w niewielkim fragmencie całej struktury 3D. Ostatnią miarą, która była inspiracją podczas opracowywania koncepcji badań to *Global Distance Test – Total Score* (*GDT_TS*) [E62], która została opracowana dla struktur 3D białek oraz zastosowana w konkursie *CASP*. Miara ta wykorzystuje tzw. *Global Distance Test* (*GDT*) reprezentujący największy możliwy odsetek procentowy liczby atomów C_α odpowiadających sobie pomiędzy modelem 3D, a referencyjną strukturą 3D, dla których *RMSD* jest poniżej określonego progu. W ramach *GDT_TS* wyznaczana jest średnia wyników miary *GDT* uzyskanej dla czterech progów, a mianowicie 1, 2, 4, 8Å. W praktyce okazuje się jednak, że liczba i określone wartości stosowanych progów powinny być elastycznie konfigurowane przez

eksperta, ponieważ zależą od długości referencyjnej struktury 3D, jakości analizowanych modeli 3D, typów motywów 3D, które w rzeczywistości decydują o określonym kształcie referencyjnej struktury 3D cząsteczki. W ramach prowadzonych badań opracowano nowatorską metodę wizualnej ewaluacji modelu(i) 3D w kontekście referencyjnej struktury 3D białek/RNA, która jako pierwsza pozwala na przeprowadzenie analizy w szerszej perspektywie umożliwiając ekspertowi wprowadzenie dowolnego wachlarza wartości progowych (tzw. poziomów szczegółowości). Proponowana koncepcja składa się z następujących etapów: (1) ekspert definiuje parametry konfiguracyjne: (a) struktury 3D w formacie *PDB* [E6] zarówno dla struktury referencyjnej jak i zbioru rozpatrywanych modeli 3D, (b) nazwę atomu pełniącego rolę środka sfery, (c) tryb przetwarzania (jedynie atomy centralne lub wszystkie atomy mogą być rozważane), (d) dyskretyzowany wektor promieni sfer, (2) dla każdej pary rozpatrującej resztę res_i ($i = 1..n$), pochodzącą z referencyjnej struktury 3D, oraz rozpatrywany promień sfery r_j ($j = 1..m$) identyfikowany jest zbiór atomów zlokalizowanych w sferze $sph_{(ref,\ resi,\ rj)}$, skonstruowanej w referencyjnej strukturze 3D wokół reszty res_i o określonym promieniu r_j, (3) w ramach każdego rozpatrywanego modelu 3D mod_k ($k = 1..o$) lokalizowane są współrzędne atomów $sph_{(modk,\ resi,\ rj)}$ odpowiadające zbiorowi atomów rozpatrywanym w każdej sferze (n x m) skonstruowanej w poprzednim etapie, (4) dla każdej pary sfer ($sph_{(ref,\ resi,\ rj)}$, $sph_{(modk,\ resi,\ rj)}$), pochodzących odpowiednio z referencyjnej struktury 3D i jednego z rozpatrywanych modeli 3D mod_k, wyznaczana jest wartość miary *RMSD*. Koncepcja porównywania struktur 3D białek została zaprezentowana na Rysunku 18. Wyniki są prezentowane w postaci następujących rodzajów wizualizacji: (a) wykres liniowy integrujący wyniki uzyskane dla wielu modeli 3D dla określonego promienia sfery, (b) kolorowa mapa 2D oraz (c) krajobraz 3D prezentujące wyniki uzyskane w ramach jednego modelu 3D dla wszystkich rozważanych promieni sfer, zdefiniowanych przez eksperta, (d) i (e) wykresy liniowe integrujące wiele modeli 3D prezentujące odpowiednio uśredniony współczynnik jakości oraz odsetek procentowy reszt, dla których wartość *RMSD* znalazła się poniżej *progu odcięcia*, określonego przez eksperta, w ramach rozpatrywanych poziomów szczegółowości. Przykładowe wizualizacje zostały zaprezentowane na Rysunku 19.



Rysunek 18. Koncepcja porównywania struktur 3D białek [E16].



Rysunek 19. Przykładowe wizualizacje przygotowane z wykorzystaniem pakietu *gnuplot* [O9].

Pierwsze narzędzie implementujące zaproponowaną koncepcję, o nazwie *SphereGrinder* [O6], zostało, po raz pierwszy, wykorzystane podczas konkursu *CASP9*. W oparciu o doświadczenia zdobyte podczas analizy białek zrealizowano również system, o nazwie *RNAlyzer*, dedykowany cząsteczkom RNA [O9]. Dalszy rozwój metody nastąpił w wyniku prowadzonych prac nad rozwojem tego drugiego narzędzia. W celu ujednolicania reprezentacji struktur 3D RNA przygotowano i wdrożono słownik nazw atomów i reszt nukleotydowych, wykorzystywanych przez różne narzędzia. Ponadto opracowano i zastosowano mechanizm pozwalający na oznaczanie struktury drugorzędowej RNA na udostępnianych wykresach, w notacji kropkowo-nawiasowej, na podstawie referencyjnej struktury 3D, a w szczególności zbioru par kanonicznych i *wobble* zidentyfikowanych z wykorzystaniem narzędzia *RNAView* [E21]. Podstawowym niedogodnością użytkowną był brak możliwości wykorzystania grafik, generowanych przez opracowane rozwiązania, ze względu na zbyt niską jakość. W ramach *RNAlyzera* zaproponowano szereg skryptów, wykorzystujących pakiet *gnu plot*, w celu zapewnienia oczekiwanej rozdzielczości udostępnianych wizualizacji.

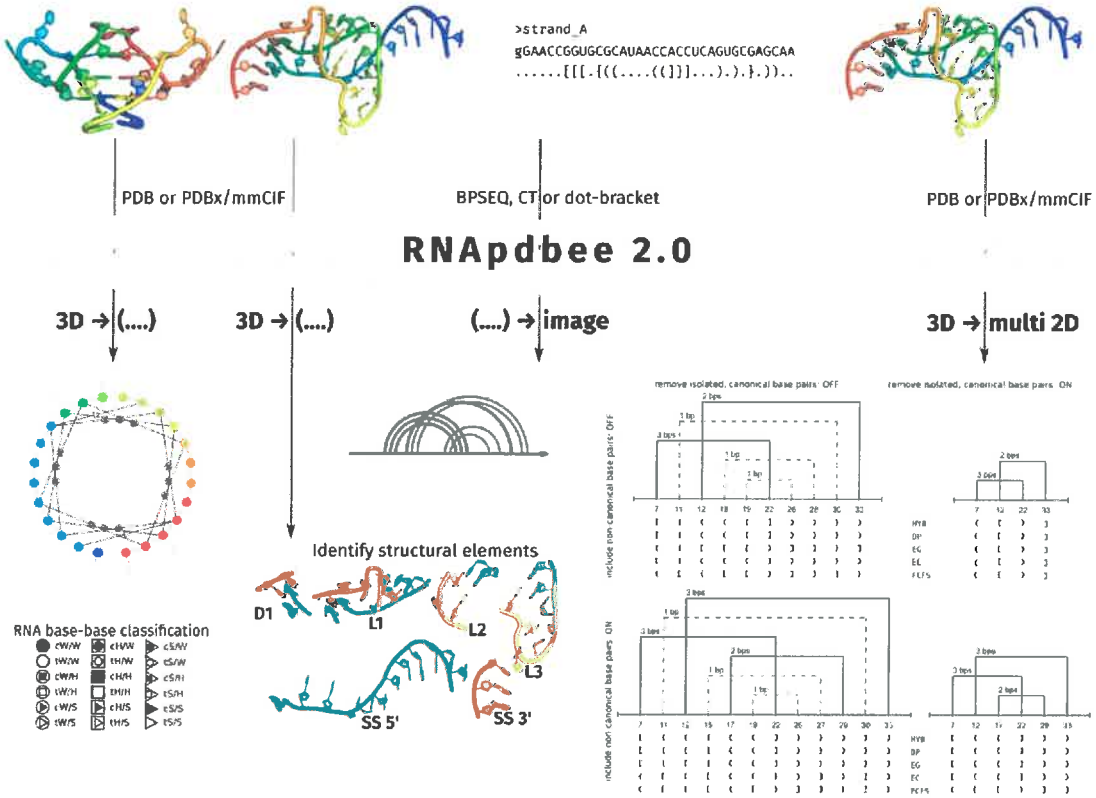
5.2.2. Opis badań realizowanych po uzyskaniu stopnia doktora.

5.2.2.1. Obliczeniowe modelowanie struktur 3D RNA ryboprzełączników.

W pracy [O8] precyzyjnie scharakteryzowano możliwości jak i ograniczenia systemu *RNAComposer* [O10] w ramach automatycznej predykcji struktury 3D RNA na przykładzie wybranych struktur ryboprzełączników RNA. Dodatkowo zaprezentowano również w jaki sposób inne narzędzia takie jak *RNApdbee* [A8] czy *RNAlyzer* [O9] mogą wspomóc proces modelowania, analizy i udokładniania modeli 3D RNA. Ryboprzełączniki, rozpatrywane w ramach niekodujących regionów mRNA, są zdolne do wiązania metabolitów komórkowych w celu modulowania ekspresji genów. W strukturach tych występują albo pseudowęzły albo wielorozgałęzione pętle, które wciąż stanowią duże wyzwanie dla metod przewidywania *in silico* struktur 3D RNA [O8]. W ramach procesu modelowania: (1) określono reprezentatywny zbiór struktur ryboprzełączników pokrywających możliwie największy podzbiór rodzin cząsteczek RNA wykazujących funkcje biologiczną (proces przeszukiwania dostępnych zasobów bazy *Rfam* [E67] został zautomatyzowany), (2) półautomatyczna identyfikacja zbioru identyfikatorów *PDB* [E6] cząsteczek RNA strukturalnie podobnych do tych znajdujących się w zbiorze reprezentatywnym, określonym w poprzednim etapie, których elementy strukturalne powinny zostać odfiltrowane z repozytorium systemu *RNAComposer* w celu zapewnienia wiarygodności procesu modelowania, (3) przeprowadzono modelowanie struktur 3D RNA z wykorzystaniem systemu *RNAComposer*, oraz (4) globalną ewaluację jakości uzyskanych modeli 3D dla rozpatrywanych rodzin w kontekście referencyjnej struktury 3D z wykorzystaniem miar *Root-Mean-Square-Deviation* [E15] i *Interaction Network Fidelity* [E38]. Większość otrzymanych modeli 3D RNA charakteryzowała się wysoką jakością zarówno w zakresie własności stereochemicznych jak i oddziaływań zachodzących pomiędzy zasadami. Ponadto przeprowadzono szczegółową analizę wybranych struktur ryboprzełączników RNA interesujących z punktu widzenia złożoności procesu modelowania [O8].

5.2.2.2. Kompleksowa adnotacja struktur RNA.

System *RNApdbee* [A8] w ciągu kilku lat od zaistnienia w społeczności naukowej przeobraził się w platformę pozwalającą na kompleksową analizę zależności zachodzących pomiędzy strukturą przestrzenną, a drugorzędową cząsteczek RNA [O2]. Wizualizacja najnowszych aspektów funkcjonalnych systemu została zaprezentowana na Rysunku 20. Najnowsza odsłona systemu, jako jedna z nielicznych, aktualnie dostępnych rozwiązań, pozwala efektywnie przetwarzać bardzo złożone struktury 3D RNA przechowywane jedynie w formacie *PDBx/mmCIF*. Wachlarz narzędzi służących do identyfikacji par zasad na podstawie struktury 3D RNA został rozbudowany o dodatkowe narzędzie, o nazwie *FR3D* [E68]. Udostępniono użytkownikowi możliwość decydowania o tym czy należy czy też nie rozważać niekanoniczne i izolowane pary zasad, co w połączeniu z wyborem spośród szeregu nowych algorytmów kodowania rozszerzonej struktury drugorzędowej w notacji kropkowo-nawiasowej, zaproponowanych w [A3], pozwala zaobserwować różnorodność topologiczną szczególnie w przypadku złożonych cząsteczek RNA obejmujących pseudowęzły wysokiego *rzędu* [A8]. Rozbudowano mechanizmy wizualizacji struktury drugorzędowej RNA zarówno w zakresie rozszerzenia istniejących podejść o możliwość prezentowania niekanonicznych par zasad jak i integracji nowych np. diagramy łukowe generowane przy użyciu *R-CHIE* [E69]. Ponadto procedury wizualizacji wspierają aktualnie wieloniciowe struktury RNA co umożliwia np. prezentowanie struktur drugorzędowych kwadrupleksów.



Rysunek 20. Prezentacja nowych funkcjonalności systemu *RNApdbee 2.0* [O2].

W ramach zupełnie nowych funkcjonalności systemu należy wyróżnić: (1) wizualizowanie struktury drugorzędowej RNA na podstawie jej tekstowej reprezentacji wprowadzonej w rozszerzonej notacji kropkowo-nawiasowej, (2) identyfikację elementów strukturalnych, a mianowicie pętli oraz fragmentów jedno lub dwuniciowych, które następnie mogą być wykorzystane, jako elementy 3D wprowadzone przez użytkownika, podczas procesu udokładnionego modelowania struktur 3D RNA z wykorzystaniem systemu *RNAComposer* [O10, O8, A4], (3) jednoczesne uruchomienie wszystkich dostępnych algorytmów oraz narzędzi do identyfikacji par zasad, w ramach jednego efektywnego obliczeniowo procesu, w celu przedstawienia na jednym diagramie różnorodności w zakresie topologii wejściowej złożonej cząsteczki RNA (zawierającej pseudowęzły) z wykorzystaniem nowo opracowanego algorytmu.

5.2.2.3. Kompleksowa ewaluacja modelu(i) 3D w kontekście referencyjnej struktury 3D RNA.

Z narzędziem *RNAlyzer* [O9] skojarzone są następujące niedogodności: (1) obliczenia są przeprowadzane u użytkownika, a więc lokalna infrastruktura obliczeniowa wpływa znacząco na czas przetwarzania, a zarazem na wygodę użytkowania systemu jak i innych aplikacji uruchomionych jednocześnie, (2) narzędzie wspiera tylko jedną miarę, a mianowicie *Root-Mean-Square-Deviation* (*RMSD*) [E15], (3) w przypadku zaobserwowania nieprawidłowości na wykresie liniowym nie ma możliwości, aby jednym kliknięciem przygotować wizualizację nałożonych na siebie struktur 3D, a mianowicie jednej pochodzącej ze referencyjnej struktury 3D RNA odpowiadającej sferze o określonym promieniu obserwowanej wokół wybranej reszty nukleotydowej oraz, w ogólności, wielu odpowiadającym jej struktur 3D pochodzących

z analizowanych modeli 3D RNA, (4) nie do końca czytelne są osie np. w przypadku krajobrazów trójwymiarowych oraz wysokiej jakości grafiki dla udostępnianych wizualizacji nie mogą być łatwo pobrane, (5) brak możliwości, aby w ramach jednej instancji narzędzia zaobserwować więcej niż jeden wykres w danej chwili co zauważalnie obniża ergonomię użytkowania. W celu rozwiązania wszystkich wspomnianych wyżej niedogodności opracowano nową platformę do ewaluacji jakości modeli 3D w kontekście referencyjnej struktury 3D RNA, o nazwie *RNAssess* [O7]. Główną zaletą powstałej aplikacji jest odciążenie infrastruktury użytkownika, ponieważ obliczenia wykonywane są w infrastrukturze skojarzonej z serwerem, a użytkownik zleca zadania, śledzi ich postęp wykonania oraz analizuje uzyskane wyniki z wykorzystaniem przeglądarki. W ich skład wchodzi globalna ewaluacja przeprowadzona z wykorzystaniem szeregu miar, zdefiniowanych poniżej, która wcześniej nie była dostępna. Moduł obliczeniowy systemu jest reprezentowany przez szeroki wachlarz usług sieciowych pozwalających na: (1) weryfikację zgodności struktur 3D RNA z formatem *PDB* [E6], (2) ekstrakcję listy par zasad na podstawie struktury 3D RNA z wykorzystaniem *RNAView* [E21] lub *MC-Annotate* [E21], (3) ewaluację jakości modelu(i) 3D RNA z wykorzystaniem następujących miar: *Root-Mean-Square-Deviation* [E15], *Interaction Network Fidelity (wc, nwc, stacking, all)* [E38], *Deformation Index* [E38], *Clash score* [E39], *P-value* [E70], (4) automatyczne generowanie udostępnianych typów wizualizacji z wykorzystaniem pakietu *gnuplot*, które można łatwo pobrać jednym kliknięciem w przeglądarce, (5) przygotowanie strukturalnej superpozycji umożliwiającej jej dalszą wizualizację w przestrzeni kartezjańskiej pomiędzy zbiorem atomów zlokalizowanych w sferze o określonym promieniu (w Å) obserwowanej wokół wybranej reszty nukleotydowej (np. AG30) w referencyjnej strukturze 3D RNA, a odpowiadającymi zbiorami atomów ocenianych modeli 3D RNA, (6) udostępnienie profilu deformacji (ang. *Deformation Profile*) dla każdego rozpatrywanego modelu 3D RNA skonstruowanego z wykorzystaniem zintegrowanego pakietu *Deformation Profile Scripts* [E38]. Ponadto zastosowano mechanizmy współbieżności oraz opracowano nową koncepcję przetwarzania w celu zapewnienia wysokiej wydajności w wyniku obserwowanego wzrostu liczby zgłaszanych modeli 3D RNA w kolejnych wyzwaniach konkursu *RNA-Puzzles* [E7, E8, O3]. W ramach prowadzonych prac opracowano słownik uśrednionych wartości promieni dla każdej kombinacji par (typ nukleotydu RNA, rozważana nazwa atomu centralnego) uwzględniając jednocześnie czy w zbiorze atomów rozpatrywane są atomy wodoru czy też nie. Podczas analizy referencyjnej struktury 3D RNA zbiory atomów dla sfer o określonych promieniach są wyznaczane dokładnie raz. Dla każdej reszty nukleotydowej wchodzącej w skład referencyjnej struktury 3D RNA tworzona jest lista pojemników uporządkowanych rosnąco względem odpowiadających im wartości promieni sfer. W każdym pojemniku przechowywane są identyfikatory atomów, zakładając, że odległość w przestrzeni kartezjańskiej wyznaczona pomiędzy atomem zawartym w danym pojemniku, a atomem centralnym jest mniejsza lub równa promieniowi skojarzonemu z aktualnym pojemnikiem, a jednocześnie większa od promienia skojarzonego z poprzednim pojemnikiem. Zastosowane podejście powoduje, że część wspólna dowolnej pary pojemników rozpatrywanej dla danego nukleotydu jest zbiorem pustym, a co za tym idzie zbiór atomów skojarzony z określonym promieniem sfery jest sumą mnogościową wszystkich atomów skojarzonego z nim pojemnika oraz wszystkich pojemników go poprzedzających. Ponadto, identyfikatory atomów przechowywanych w koszach są liczbami całkowitymi, a nie jak to się zwykle realizuje w postaci łańcuchów znakowych zbudowanych poprzez

konkatenację nazw odpowiednio reszty nukleotydowej RNA i atomu. Przeprowadzono eksperymenty wydajnościowe, które potwierdziły, że zastosowane rozwiązania znacząco zwiększyły efektywność modułu obliczeniowego [O7].

5.2.2.4. Modelowanie i ewaluacja modeli 3D RNA w konkursie *RNA-Puzzles*.

RNA-Puzzles [E7, E8, O3] to międzynarodowy eksperyment dokonujący wiarygodnej ewaluacji aktualnego stanu wiedzy w zakresie przewidywania *in silico* struktur 3D RNA w celu identyfikacji zarówno mocnych jak i słabych stron oraz ograniczeń wiodących metod w zakresie złożoności strukturalnej jak i długości sekwencji reszt nukleotydowych RNA. Prezentowane wyniki pozwalają potencjalnym użytkownikom ocenić przydatność dostępnych metod w zakresie różnorodnych zastosowań, a zarazem inspirują twórców do dalszych wysiłków w zakresie udoskonalania istniejących rozwiązań. W ramach działań zespołu naukowego „Adamiak group", podsumowanych w pracy [O3], opracowano specjalizowane mechanizmy wspomagające proces modelowania realizowany w ramach automatycznego systemu *RNAComposer* [O10, O8, A4] oraz przeprowadzono szereg eksperymentów obliczeniowych. Ponadto udostępniono narzędzie z myślą o automatyzacji procedury ewaluacji jakości zgłaszanych modeli 3D w kontekście referencyjnej struktury 3D RNA. Wprowadzenie kategorii predykcji skierowanej do w pełni zautomatyzowanych serwerów w ramach inicjatywy *RNA-Puzzles* stało się główną motywacją dla opracowania i implementacji dedykowanej, rozszerzonej funkcjonalnie usługi realizującej w pełni automatyczną predykcję struktury 3D RNA na podstawie sekwencji oraz struktury drugorzędowej oraz samej sekwencji reszt nukleotydowych RNA. W pełni automatyczna usługa została zaprojektowana w postaci dodatkowej warstwy, nasłuchującej w trybie ciągłym, odpowiedzialnej za: (1) przyjmowanie zadań w postaci elektronicznych wiadomości pasujących do określonego wzorca, (2) integrację szeregu modułów obliczeniowych systemu *RNAComposer* charakteryzujących się różnorodną konfiguracją uruchomieniową, (3) przeprowadzenie wstępnego przetwarzania (np. filtrowanie niekanonicznych par zasad zidentyfikowanych w nadesłanej strukturze drugorzędowej RNA), (4) przygotowanie planu dystrybucji zadań oraz optymalizację jego współbieżnego wykonania, (5) ewaluację uzyskanych modeli 3D RNA, z wykorzystaniem *RNApdbee* [A8] i *RNAssess* [O7], w oparciu o szeroki wachlarz miar (obejmujący całkowitą energia uzyskana z wykorzystaniem pakietu *XPLOR* [E71], promień żyroskopowy, INF_{cbp} [E38], *Clash score* [E39]) oraz agregację uzyskanych wyników w postaci końcowego rankingu, oraz (6) przygotowanie pliku w formacie *PDB* [E6], zawierającego do pięciu wybranych modeli 3D RNA, wprowadzonego w postaci załącznika w wiadomości elektronicznej wysłanej do nadawców zleconego zadania. Ponadto usługa wspiera wieloliniową reprezentację struktury drugorzędowej, zapisanej w notacji kropkowo-nawiasowej, zakładając, że kolejne linie (od drugiej włącznie) reprezentują pseudowęzły o rosnących rzędach [A8] (od pierwszego rzędu włącznie). Wzorce wiadomości mogą być elastycznie konfigurowane. W przypadku trybu predykcji opartego jedynie na sekwencji reszt nukleotydowych RNA realizowany jest następujący algorytm: (1) uruchomienie zintegrowanych narzędzi (*RNAstructure, RNAfold, Contrafold, ContextFold, CentroidFold, IPknot, RNAshapes, HotKnots, CentroidAlifold, RNAalifold* [E19]) do predykcji *in silico* struktury drugorzędowej RNA w celu identyfikacji szerokiego, unikalnego zbioru struktur drugorzędowych reprezentowanych w rozszerzonej notacji kropkowo-nawiasowej, (2) uruchomienie procesu predykcji dziesięciu modeli 3D

RNA z wykorzystaniem każdej rozpatrywanej konfiguracji modułu obliczeniowego systemu *RNAComposer* dla każdej unikalnej struktury drugorzędowej RNA uzyskanej w poprzednim etapie, (3) filtrowanie wszystkich modeli 3D RNA dla których proces predykcji zakończył się niepowodzeniem lub całkowita energia uzyskana z wykorzystaniem pakietu *XPLOR* jest dodatnia, (4) ewaluacja uzyskanych modeli 3D RNA oraz utworzenie rankingu względem niemalejącego porządku współczynnika energetycznego. Podczas, gdy *apriori* znana jest zarówno sekwencja jak i struktura drugorzędowa RNA proces predykcji znacząco się upraszcza prowadząc do uzyskania zbioru dwudziestu modeli 3D RNA, które najpierw są porządkowane względem nierosnącego porządku INF_{cbp}, a następnie względem współczynnika energetycznego. W przypadku pojawienia się na wejściu cząsteczki wieloniciowej każde wystąpienie znaku '-' zamieniane jest na krótką czteronukleotydową sekwencję 'AAAA' w celu uzyskania cząsteczki jednoniciowej. Wszystkie możliwe kombinacje łańcuchów są rozpatrywane w ramach procesu predykcji. Następnie wszystkie wynikowe modele 3D RNA są poddawane procesowi odfiltrowywania atomów występujących w sztucznie wprowadzonym fragmencie łączącym poszczególne nici. Dodatkowo istnieje możliwość dołączenia do wynikowej wiadomości elektronicznej: (1) pliku logu pozwalającego prześledzić przeprowadzony proces predykcji dla każdego ostatecznie wybranego modelu 3D RNA oraz (2) plik *CSV* obejmujący wartości rozpatrywanych miar oraz strukturę drugorzędową, w rozszerzonej notacji kropkowo-nawiasowej, wyekstrahowaną z wykorzystaniem *RNApdbee*, wyznaczone dla zestawu ostatecznie wybranych modeli 3D RNA. Usługa została kompleksowo przetestowana pod względem wydajnościowym jak i akceptacyjnym. Jako przykład obiecującego wyniku, uzyskanego z wykorzystaniem opracowanej usługi, można wskazać predykcję serwerową nr 3 zgłoszoną w ramach wyzwania nr 19, która została sklasyfikowana na 4 miejscu, względem miary *RMSD*, spośród 65 zgłoszonych modeli ustępując jedynie trzem predykcjom ludzkim. Ponadto predykcja serwerowa nr 5, w ramach tego samego wyzwania, uzyskała najwyższy wynik względem miary INF_{nwc} *ex aequo* z dwoma innymi predykcjami nadesłanymi w konkursie. W zakresie ewaluacji opracowano narzędzie *RNAQUA* wspomagające manipulowanie strukturami 3D RNA oraz ich analizę jakości, udostępnione w ramach *RNA-Puzzles toolkit*, pozwalające na: (a) wprowadzenie, przez eksperta, jednoznacznego dopasowania pomiędzy każdym rozpatrywanym modelem 3D, a referencyjną strukturą 3D RNA, (b) ocenę modelu(i) 3D RNA z wykorzystaniem następujących miar: *Root-Mean-Square-Deviation* (*RMSD*) [E15], *Interaction Network Fidelity* (*INF*) [E38] *(wc, nwc, stacking, all), Deformation Index* [E38], *Clash score* [E39], *P-value* [E70], (c) ekstrakcję sekwencji reszt nukleotydowych dla modelu(i) i referencyjnej struktury 3D RNA lub składających się na nią fragmentów oraz raportowanie różnic obserwowanych w powstałym dopasowaniu, (d) ekstrakcję zbioru nieciągłych fragmentów 3D struktury referencyjnej/modelu(i) na podstawie dopasowania zdefiniowanego przez eksperta w postaci oryginalnej lub przenumerowanej (można dodatkowo zrealizować proces nakładania wyekstrahowanych struktur 3D z modelu(i) na odpowiadającą im referencyjną strukturę 3D RNA).

5.2.2.5. Ciągła ewaluacja algorytmów rozwiązujących problemy optymalizacyjne.

W ogólności problemy kombinatoryczne polegają na odkrywaniu wartości zmiennych dyskretnych, które spełniają określone ograniczenia. Problemy

przeszukiwania reprezentują jedną z klas problemów kombinatorycznych. Rozwiązaniem problemu przeszukiwania jest konkretny obiekt (lub wartość) spełniający postawione warunki, lub odpowiedź „nie" jeśli takowy nie istnieje. Szczególnym przypadkiem problemu przeszukiwania jest problem optymalizacyjny, którego rozwiązaniem jest obiekt (lub wartość) optymalny(a) pod względem sformułowanej funkcji celu [E72]. Złożone problemy optymalizacyjne są rozwiązywane nie tylko przez naukowców zajmujących się badaniami operacyjnymi czy ekonomią, ale również nie wprost przez przedsiębiorców zainteresowanych np. minimalizacją kosztów transportu czy produkcji. *Online judge* [E73] określa platformę internetową wspierającą w pełni automatyczną, przeprowadzaną w czasie rzeczywistym ewaluację rezultatów uruchomiania programów zgłaszanych przez uczestników biorących udział w konkursie o charakterze programistycznym. Wstępny przegląd literatury wykazał, że nie istnieją rozwiązania tego typu dedykowane rozwiązywaniu problemów optymalizacyjnych, pomimo wydawałoby się potencjalnego szerokiego zapotrzebowania na tego typu platformę. W związku z powyższym przeprowadzono kompleksowy przegląd aktualnie istniejących systemów typu *online judge* z punktu widzenia możliwych zastosowań [O1]. Niestety wstępne obserwacje zostały potwierdzone co stało się inspiracją dla prowadzonych badań. Ponadto ustalono, że istniejące systemy zwykle ukierunkowane są na określoną dziedzinę problemów. Poza tym różnią się znacznie w zakresie udostępnianej funkcjonalności, co powoduje, że ich praktyczne zastosowanie jest zwykle dość trudne ze względu na różne oczekiwania potencjalnych użytkowników [O1]. W rezultacie zdecydowano się wypełnić lukę proponując platformę *Optil.io* [O1, O5] bazującą na koncepcjach *Evaluation-as-a-Service* [E74] oraz *crowdsourcing* [E75], która umożliwia użytkownikom z całego świata ciągłą i wiarygodną ewaluację rozwiązań algorytmicznych złożonych problemów optymalizacyjnych oraz organizację specjalizowanych wyzwań/konkursów inspirowanych zastosowaniami naukowymi lub przemysłowymi. Użytkownicy zainteresowani organizacją wyzwania wprowadzają w systemie: (1) tekstową definicję problemu uzupełnioną, opcjonalnie, o inne treści multimedialne, (2) zbiór prostych przypadków testowych dostarczanych użytkownikom, (3) zbiór złożonych przypadków testowych wykorzystywanych w trakcie ewaluacji oraz (4) ograniczenia zasobowe związane z maksymalnym dopuszczalnym czasem przetwarzania i wykorzystaniem pamięci RAM. Wiarygodna ewaluacja dokonywana jest na podstawie przypadków testowych, które nie są udostępniane uczestnikom konkursu w bezpiecznym, jednorodnym środowisku uruchomieniowym. Uczestnicy mogą zgłaszać rozwiązania w postaci kodów źródłowych w wielu językach programowania oraz obserwować na bieżąco uzyskiwane rezultaty przez cały okres trwania konkursu. Bezpieczeństwo i jednorodność środowiska obliczeniowego zapewniane są poprzez automatyczne tworzenie kontenerów, na platformie *Docker*, dla każdego procesu ewaluacji w celu zapewnienia izolacji wykonania oraz precyzyjnego egzekwowania ograniczeń zasobowych. Warto podkreślić, że platforma obsługuje również wysokowydajne obliczenia pozwalając użytkownikom na wykorzystanie technik opartych na GPU. W pracy [O1] oprócz przeglądu systemów *online judge*: (1) sformułowano formalną definicję systemu *online judge*, (2) scharakteryzowano wspólne aspekty wykorzystywanej przez te systemy metodologii ewaluacji, (3) przedstawiono platformę *Optil.io* oraz przedyskutowano wyniki wstępnych eksperymentów obliczeniowych, przeprowadzonych z jej wykorzystaniem, które potwierdziły, że integracja koncepcji *crowdsourcing* w ramach systemu *online judge* pozwala na efektywne i wiarygodne rozwiązywanie złożonych problemów

optymalizacyjnych. W 2017 r., na platformie *Optil.io*, przeprowadzone zostały następujące konkursy, a mianowicie *Brilliant Challenges* dedykowany ciekawym, praktycznym problemom optymalizacyjnym oraz *PACE: Parameterized Algorithms and Computational Experiments Challenge*.

5.2.2.6. Kompresja rezultatów sekwencjonowania DNA.

Znaczący rozwój metod sekwencjonowania DNA powoduje, że koszty sekwencjonowania pełnego genomu są coraz niższe, a zarazem wolumen surowych danych produkowanych w wyniku rosnącej liczby przeprowadzanych eksperymentów sekwencjonowania stale rośnie. W związku z powyższym centra zajmujące się badaniem genomu aktualnie przechowują i przetwarzają dane liczone w petabajtach, a dysponowana przez nie infrastruktura sprzętowa kurczy się tak szybko, że istnieje ogromna potrzeba rozwoju wysokiej jakości algorytmów kompresji, które pozwolą znacząco zredukować zarówno dynamikę tego procesu jak i rzeczywisty koszt przechowywania danych. W ramach przeprowadzonych badań opracowany został specjalizowany, bezstratny kompresor [O4] przeznaczony do długoterminowego przechowywania danych (w formacie *FASTQ* [E76]) wytwarzanych przez wysokoprzepustowe sekwencery DNA. Podczas projektowania metody postawiono sobie za cel optymalizację jakości kompresji kosztem czasu przetwarzania, który nie powinien znacząco odbiegać od czasów przetwarzania kompresorów ogólnego przeznaczenia, które są najczęściej wykorzystywane w tego typu zastosowaniach. Opracowanie rozwiązanie wykorzystuje modele statystyczne wyższego rzędu (ang. *high-order statistical models*) w połączeniu z kodowaniem przedziałów (ang. *range encoding*). Ponadto wykorzystywane są również podejście słownikowe oraz algorytm bazujący na sortowaniu w celu identyfikacji powtarzających się informacji wraz z metodą kodowania LZ77. Szczegółowy opis algorytmu znajduje się w pracy [O4]. W celu ewaluacji jakości kompresji oraz czasu przetwarzania procesu kompresji/dekompresji opracowanego rozwiązania w kontekście innych dostępnych narzędzi, a mianowicie *DSRC* [E77], *Quip* [E78], *SCALCE* [E79] oraz *7-zip, bzip2, gzip*, przeprowadzono eksperymenty obliczeniowe w oparciu o publicznie dostępne dane pochodzące z sześciu eksperymentów sekwencjonowania [O4]. Opracowane rozwiązanie uzyskało najwyższą jakość kompresji dla sześciu największych plików wykorzystywanych podczas eksperymentów.

5.2.3. Plany naukowe.

Kontynuacja badań w zakresie opracowywania nowych algorytmów, metod obliczeniowych i narzędzi w dziedzinie bioinformatyki strukturalnej cząsteczek biologicznych, a w szczególności RNA. W zakresie automatycznego przewidywania struktur 3D RNA planuje się rozwój metody obliczeniowej, udostępnionej w ramach systemu *RNAComposer*, obejmującej: (1) rozpoznawanie motywów rekurencyjnych przewidywanych na podstawie sekwencji reszt nukleotydowych RNA w celu udokładnienia procesu predykcji 3D RNA w oparciu o zbiór, rozpoznanych w taki sposób, niekanonicznych par zasad, (2) udostępnienie ekspertowi mechanizmu wprowadzania ograniczeń na planarność, które zostaną zastosowane podczas procesu minimalizacji energetycznej wstępnego modelu 3D RNA, (3) opracowanie protokołu minimalizacji pozwalającego na zawiązywanie oddziaływań wewnątrzcząsteczkowych dalekiego zasięgu, (4) rozszerzenie modelu fragmentacji struktury drugorzędowej RNA oraz repozytorium elementów strukturalnych, aby umożliwić modelowanie

specyficznych motywów takich jak np. pseudowęzły i kwadrupleksy, (5) opracowanie nowych mechanizmów wyszukiwania, przekształcania oraz przewidywania *in silico* struktur 3D złożonych wieloramiennych pętli, które nie zostały dotąd zidentyfikowane w ramach żadnej spośród znanych, eksperymentalnie określonych struktur 3D RNA, (6) opracowanie nowego, specjalizowanego algorytmu asemblacji elementów strukturalnych w celu konstrukcji wstępnego modelu 3D RNA, (7) opracowanie dodatkowej warstwy, która będzie zarządzała wieloetapowym, hierarchicznym procesem modelowania struktur 3D RNA o rozmiarze sięgającym 1000nts. Dalszy rozwój platformy *RNApdbee* w zakresie: (1) rozpoznawania i klasyfikacji motywów 3D RNA obejmujących oddziaływania bliskiego i dalekiego zasięgu, (2) analizy nowych typów interakcji np. RNA-ligand czy RNA-białko, (3) rozpoznawania i klasyfikacji motywów 3D RNA występujących w otoczeniu jonów, (4) wizualizacji szeregu struktur w oparciu o jeden wspólny wzorzec struktury drugorzędowej RNA, (5) udoskonalenia sposobu prezentacji struktury drugorzędowej RNA w taki sposób, aby zapewnić współosiowość elementów topologicznych na diagramie, które rzeczywiście są współosiowe w odpowiadającej jej strukturze 3D RNA. Ponadto planowane są prace w zakresie opracowania: (1) nowych algorytmów inspirowanych teorią grafów i badaniami operacyjnymi do problemu kodowania złożonych struktur 3D RNA zawierających pseudowęzły w rozszerzonej notacji kropkowo-nawiasowej, (2) nowej metody obliczeniowej umożliwiającej rekonstrukcję pełnoatomowej struktury 3D RNA na podstawie wprowadzonej sekwencji RNA oraz zbioru współrzędnych atomów fosforu, (3) nowego algorytmu pozwalającego na identyfikację przeplotów w strukturach 3D RNA z wykorzystaniem metod grafiki komputerowej, (4) interaktywnego diagramu prezentującego topologię struktury RNA, (5) nowej metody dopasowywania, niezależnej od sekwencji reszt nukleotydowych RNA, par struktur 3D RNA z wykorzystaniem koncepcji deskryptorów, (6) nowego algorytmu rankingowania modeli 3D RNA opartego na konsensusie zbudowanym na podstawie sieci oddziaływań zachodzących pomiędzy zasadami, (7) nowych podejść (lub miar) rozwiązujących problem ogólnej ewaluacji jakości modeli 3D RNA (zakładając nieznaną referencyjną strukturę 3D RNA) z wykorzystaniem metod uczenia maszynowego. Ponadto planuje się aplikowanie o projekty naukowe (np. SONATA) finansowane przez Narodowe Centrum Nauki.

5.3. Współautorskie narzędzia i serwery obliczeniowe.

W poniższej Tabeli zaprezentowano szeroki wachlarz ogólnodostępnych narzędzi i serwerów obliczeniowych, wykorzystywanych przez społeczność naukową, w których tworzeniu brał udział habilitant w roli projektanta, programisty, konsultanta naukowego lub kierownika projektu.

Nazwa	Opis
RNAfitme	Serwer obliczeniowy pozwalający na rekonstrukcję pełnoatomowej struktury 3D RNA na podstawie wprowadzonej sekwencji RNA oraz struktury 3D utrwalonego szkieletu fosforo-cukrowego. Liczba odwiedzin: 565 (28.01.2019). Publikacje: *AMCS* 2015 [A6], *BMC Bioinformatics* 2018 [A1]. URL: http://rnafitme.cs.put.poznan.pl
RNAvista	Serwer obliczeniowy pozwalający na przewidywanie rozszerzonej struktury drugorzędowej RNA (obejmującej pary niekanoniczne) na podstawie sekwencji lub kanonicznej struktury drugorzędowej RNA.

	Liczba odwiedzin: 397 (28.01.2019). Publikacje: *BMC Bioinformatics* 2015 [A8], *Bioinformatics* 2018 [A2]. URL: http://rnavista.cs.put.poznan.pl
RNAQUA	Narzędzie wspomagające manipulowanie strukturami 3D RNA oraz analizę ich jakości. Liczba odwiedzin: brak danych. Publikacje: w przygotowaniu. URL: https://github.com/mantczak/rnaqua
RNA-Puzzles toolkit	Zasób udostępniający: (1) znormalizowane zestawy modeli 3D zgłoszonych w konkursie *RNA-Puzzles*, (2) skrypty pozwalające wyznaczać wartości miar oceny jakości modeli 3D RNA w kontekście referencyjnej struktury 3D RNA, oraz (3) narzędzia wspomagające manipulowanie strukturami 3D RNA. Liczba odwiedzin: brak danych. Publikacje: w przygotowaniu. URL: https://chichaumiau.github.io/rnapuzzlestoolkit/
RNAComposer	Serwer obliczeniowy pozwalający na w pełni automatyczne oraz udokładnione modelowanie struktur 3D dużych cząsteczek RNA. Liczba odwiedzin: 5 013 083 (28.01.2019). Publikacje: *NAR* 2012 [O10], *MIE* 2015 [O8], *Acta Biochim Pol* (2016) [A4]. URLs: http://rnacomposer.cs.put.poznan.pl, http://rnacomposer.ibch.poznan.pl
descs-standalone	Narzędzie pozwalające na identyfikację oraz strukturalne dopasowanie lokalnych motywów 3D biomolekuł takich jak białka i cząsteczki RNA. Liczba odwiedzin: brak danych. Publikacje: *BMC Bioinformatics* 2016 [A5]. URL: https://github.com/mantczak/descs-standalone
Optil.io	Platforma do ciągłej ewaluacji rozwiązań algorytmicznych dla problemów optymalizacyjnych inspirowanych zastosowaniami naukowymi lub przemysłowymi. Liczba odwiedzin: 140 503 (28.01.2019). Publikacje: *ACM CSCW* 2016 [O5], *CSUR* 2018 [O1]. URL: https://www.optil.io
RNAssess	Serwer obliczeniowy pozwalający na kompleksową ewaluację modeli 3D RNA w kontekście referencyjnej struktury 3D RNA. Liczba odwiedzin: 2 102 (28.01.2019). Publikacje: *NAR* 2015 [O7]. URL: http://rnassess.cs.put.poznan.pl
RNApdbee	Serwer obliczeniowy pozwalający na ekstrakcję struktury drugorzędowej w rozszerzonej notacji kropkowo-nawiasowej na podstawie struktury 3D złożonych cząsteczek RNA, w szczególności zawierających pseudowęzły. Liczba odwiedzin: 12 494 (28.01.2019). Publikacje: *NAR* 2014 [A8], *Bioinformatics* 2018 [A2], *NAR* 2018 [O2]. URL: http://rnapdbee.cs.put.poznan.pl
RNAlyzer	Narzędzie pozwalające na wizualną ewaluację modeli 3D RNA w kontekście referencyjnej struktury 3D RNA. Liczba odwiedzin: brak danych.

	Publikacje: *NAR* 2013 [O9]. URL: http://rnalyzer.cs.put.poznan.pl
SphereGrinder	Narzędzie pozwalające na wizualną ewaluację modeli 3D białek w kontekście referencyjnej struktury 3D białka. Liczba odwiedzin: brak danych. Publikacje: *IEEE BIBM* 2015 [O6]. URL: http://spheregrinder.cs.put.poznan.pl

5.4. Wskaźniki bibliometryczne według bazy Web of Science.

Po uzyskaniu stopnia doktora habilitant opublikował **piętnaście** prac w latach od 2014 do 2018, których **sumaryczny impact factor** oraz **całkowita liczba punktów MNiSW** stanowią odpowiednio **62,837** oraz **490.**

	Przed uzyskaniem stopnia naukowego doktora (maj 2013r.)	Obecnie (dane z dnia 17.02.2019)
Indeks Hirscha	**2**	**8**
Liczba publikacji	**2**	**17**
Liczba cytowań**	**14**	**322**
Średnia liczba cytowań na pracę**	**7,00**	**18,94**
Sumaryczny Impact Factor*	**17,056**	**79,893**
Suma punktów MNiSW*	**80**	**570**

*Dane przygotowane na podstawie roku publikacji. ** Okres *przed doktoratem* zakończył się wraz z końcem roku 2013.

6. Literatura uzupełniająca.

[E1] A. Lesk, 2008, Introduction to Bioinformatics, *Oxford University Press*, New York.

[E2] J.C. Lee, R.R. Gutell, 2004, Diversity of Base-pair Conformations and their Occurrence in rRNA Structure and RNA Structural Motifs, *Journal of Molecular Biology*, 344(5), 1225–1249.

[E3] Z. Miao, E. Westhof, 2017, RNA Structure: Advances and Assessment of 3D Structure Prediction, *Annual Review of Biophysics*, 46, 483–503.

[E4] A.I. Petrov, C.L. Zirbel, N.B. Leontis.2013, Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas, *RNA*, 19, 1327–1340.

[E5] E. Westhof, P. Auffinger, 2000, RNA Tertiary Structure, *Encyclopedia of Analytical Chemistry*, 522–532.

[E6] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, 2000, The Protein Data Bank, *Nucleic Acids Research*, 28(1), 235–242.

[E7] J.A. Cruz, M.F. Blanchet, M. Boniecki, J.M. Bujnicki, S.J. Chen, S. Cao, R. Das, F. Ding, N.V. Dokholyan, S.C. Flores, L. Huang, C.A. Lavender, V. Lisi, F. Major, K. Mikolajczak, D.J. Patel, A. Philips, T. Puton, J. Santalucia, F. Sijenyi, T. Hermann, K. Rother, M. Rother, A. Serganov, M. Skorupski, T. Soltysinski, P. Sripakdeevong, I. Tuszynska, K.M. Weeks,

C. Waldsich, M. Wildauer, N.B. Leontis, E. Westhof. 2012, *RNA-Puzzles*: a CASP-like evaluation of RNA three-dimensional structure prediction, *RNA*, 18, 610–625.
[E8] Z. Miao, R.W. Adamiak, M.F. Blanchet, M. Boniecki, J.M. Bujnicki, S.J. Chen, C. Cheng, G. Chojnowski, F.C. Chou, P. Cordero, J.A. Cruz, A.R. Ferré-D'amaré, R. Das, F. Ding, N.V. Dokholyan, S. Dunin-Horkawicz, W. Kladwang, A. Krokhotin, G. Lach, M. Magnus, F. Major, T.H. Mann, B. Masquida, D. Matelska, M. Meyer, A. Peselis, M. Popenda, K.J. Purzycka, A. Serganov, J. Stasiewicz, M. Szachniuk, A. Tandon, S. Tian, J. Wang, Y. Xiao, X. Xu, J. Zhang, P. Zhao, T. Zok, E. Westhof, 2015, RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures, *RNA*, 21, 1066–1084.
[E9] S. Sharma, F. Ding, N.V. Dokholyan, 2008, *iFoldRNA*: threedimensional RNA structure prediction and folding, *Bioinformatics*, 24, 1951–1952.
[E10] M. Parisien, F. Major, 2008, The *MC-Fold* and *MC-Sym* pipeline infers RNA structure from sequence data, *Nature*, 452, 51–55.
[E11] M. Rother, K. Rother, T. Puton, J.M. Bujnicki, 2011, *ModeRNA*: a tool for comparative modeling of RNA 3D structure, *Nucleic Acids Research*, 39(10), 4007–4022.
[E12] C.Y. Cheng, F.C. Chou, R. Das, 2015, Modeling complex RNA tertiary folds with *Rosetta*, *Methods in Enzymology*, 553, 35–64.
[E13] M.J. Boniecki, G. Lach, W.K. Dawson, K. Tomala, P. Lukasz, T. Soltysinski, K.M. Rother, J.M. Bujnicki, 2016, *SimRNA*: a coarse-grained method for RNA folding simulations and 3D structure prediction, *Nucleic Acids Research*, 44, e63.
[E14] X. Xu, S.J. Chen, 2015, A Method to Predict the 3D Structure of an RNA Scaffold, *Methods in Molecular Biology*, 1316, 1–11.
[E15] W. Kabsch, 1976, A solution for the best rotation to relate two sets of vectors, *Acta Crystallographica*, A32, 922–923.
[E16] M. Antczak, 2012, Algorytmiczne aspekty modelowania i ewaluacji biomolekuł, rozprawa doktorska, Wydział Informatyki, Politechnika Poznańska.
[E17] K. Sato, M. Hamada, K. Asai, T. Mituyama, 2009, *CENTROIDFOLD*: a web server for RNA secondary structure prediction, *Nucleic Acids Research*, 37, W277–80.
[E18] S. Zakov, Y. Goldberg, M. Elhadad, M. Ziv-Ukelson, 2011, Rich parameterization improves RNA structure prediction, *Journal of Computational Biology*, 18, 1525–1542.
[E19] T. Puton, L.P. Kozlowski, K.M. Rother, J.M. Bujnicki, 2013, *CompaRNA*: a server for continuous benchmarking of automated methods for RNA secondary structure prediction, *Nucleic Acids Research*, 41(7), 4307–4323.
[E20] K. Sato K, Y. Kato, M. Hamada, T. Akutsu, K. Asai, 2011, *IPknot*: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming, *Bioinformatics*, 27, i85–i93.
[E21] H. Yang, F. Jossinet, N.B. Leontis, L. Chen, J. Westbrook, H. Berman, E. Westhof, 2003, Tools for the automatic identification and classification of RNA base pairs, *Nucleic Acids Resaerch*, 31, 3450–3460.
[E22] X.J. Lu, W.K. Olson, 2003, *3DNA*: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures, *Nucleic Acids Research*, 31, 5108–5121.
[E23] Y. Byun, K. Han, 2006, *PseudoViewer*: web application and webservice for visualizing RNA pseudoknots and secondary structures. *Nucleic Acids Resarch*, 34, W416–W422.

[E24] K. Darty, A. Denise, Y. Ponty, 2009, *VARNA*: Interactivedrawing and editing of the RNA secondary structure, *Bioinformatics*, 25, 1974–1975.
[E25] S. Smit, K. Rother, J. Heringa, R. Knight, 2008, From knotted to nested RNA structures: a variety of computational methods for pseudoknot removal, *RNA*,14, 410–416.
[E26] M. Popenda, M. Blazewicz, M. Szachniuk, R.W. Adamiak, 2008, *RNA FRABASE* version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures, *Nucleic Acids Research*, 36(1), D386–D391.
[E27] M. Popenda, M. Szachniuk, M. Blazewicz, S. Wasik, E.K. Burke, J. Blazewicz, R.W. Adamiak, 2010, *RNA FRABASE 2.0*: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures, *BMC Bioinformatics*, 11, 231.
[E28] P. Brion, E. Westhof, 1997, Hierarchy and dynamics of RNA folding, *Annual Review of Biophysics and Biomolecular structure*, 26, 113–137.
[E29] R. Bellman, 1952, On the theory of dynamic programming, *Proceedings of the National Academy of Sciences of the United States of America*, 38, 717–719.
[E30] N.B. Leontis, C.L. Zirbel, 2012, Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking. In *Nucleic Acids and Molecular Biology*, 281–298. Springer Nature.
[E31] P. Dallaire, F. Major, 2016, Exploring Alternative RNA Structure Sets Using *MC-Flashfold* and *db2cm*. In *RNA Structure Determination*, 237-251, Humana Press, New York, NY.
[E32] C.H. zu Siederdissen, S.H. Bernhart, P.F. Stadler, I.L. Hofacker, 2011, A folding algorithm for extended RNA secondary structures, *Bioinformatics*, 27(13), i129-i136.
[E33] M. Pietrosanto, E. Mattei, M. Helmer-Citterich, F. Ferrè, 2016, A novel method for the identification of conserved structural patterns in RNA: From small scale to high-throughput applications, *Nucleic Acids Research*, 44(18), 8600-8609.
[E34] M.F. Sloma, D.H. Mathews, 2017, Base pair probability estimates improve the prediction accuracy of RNA non-canonical base pairs, *PLoS Computational Biology*, 13(11), e1005827.
[E35] C. Weinreb, A.J. Riesselman, J.B. Ingraham, T. Gross, C. Sander, D.S. Marks, 2016, 3D RNA and functional interactions from evolutionary couplings, *Cell*, 165(4), 963-975.
[E36] N.B. Leontis, E. Westhof, 2001, Geometric nomenclature and classification of RNA base pairs, RNA, 7(4), 499-512.
[E37] M. Andronescu, V. Bereg, H.H. Hoos, A. Condon, 2008, *RNA STRAND*: the RNA secondary structure and statistical analysis database, *BMC Bioinformatics*, 9(1), 340.
[E38] M. Parisien, J. A. Cruz, E. Westhof, F. Major, 2009, New metrics for comparing and assessing discrepancies between RNA 3D structures and models, *RNA*, 15, 1875–1885.
[E39] I.M. Davis, A. Leaver-Fay, V.B. Chen, J.N. Block, G.J. Kapral, X. Wang, L.W. Murray, W.B. III Arendall, J. Snoerink, J.S. Richardson, 2007, *MolProbity*: all-atom contacts and structure validation for protein and nucleic acids, *Nucleic Acids Research*, 35, W375–W383.
[E40] Q. Wang, A.A. Canutescu, R.L. Dunbrack, SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling, *Nature Protocols*, 2008, 3, 1832–47.
[E41] R.L. Dunbrack, M. Karplus, 1993, Backbone-dependent rotamer library for proteins application to side-chain prediction. *Journal of Molecular Biology*, 230, 543–74.

[E42] L. Kaufman, P. Rousseeuw, 1990, Finding Groups in Data: An Introduction to Cluster Analysis, 1st Edn., *Wiley-Interscience*, New York, NY.
[E43] M. Cottrell, B. Hammer, A. Hasenfuss, T. Villmann, 2006, Batch and median neural gas, *Neural Networks*, 19(6), 762–771.
[E44] T. Zok, M. Popenda, M. Szachniuk, 2014, *MCQ4Structures* to compute similarity of molecule structures, *Central European Journal of Operations Research*, 22(3), 457–473.
[E45] T.R. Hvidsten, A. Kryshtafovych, J. Komorowski, K. Fidelis, 2003, A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins, *Bioinformatics*, 19, ii81–ii91.
[E46] P. Björkholm, P. Daniluk, A. Kryshtafovych, K. Fidelis, R. Andersson, T.R. Hvidsten, 2009, Using multi-data hidden markov models trained on local neighborhoods of protein structure to predict residue-residue contacts, *Boinformatics*, 25, 1264–1270.
[E47] P. Daniluk, B. Lesyng, 2011, A novel method to compare protein structures using local descriptors, *BMC Bioinformatics*, 12, 344.
[E48] S. E. Brenner, P. Koehl, M. Levitt, 2000, The *ASTRAL* compendium for protein structure and sequence analysis, *Nucleic Acids Research*, 28, 254–256.
[E49] H.W. Kuhn, 1955, The Hungarian method for the assignment problem, *Naval Research Logistics Quarterly*, 2, 83–97.
[E50] J. Liu, B. Rost, 2004, Sequence-based prediction of protein domains, *Nucleic Acids Research*, 32(12), 3522–3530.
[E51] M. Antczak, J. Blazewicz, P. Lukasiak, M. Milostan, N. Krasnogor, G. Palik, 2011, *DomAns*—Pattern based method for protein domain boundaries prediction and analysis, *Foundations of Computing and Decision Sciences*, 36(2), 99–119.
[E52] F. Pearl, C. Bennett, J. Bray, A. Harrison, N. Martin, A. Shepherd, I. Sillitoe, J. Thornton, C. Orengo, 2003, The *CATH* database: an extended protein family resource for structural and functional genomics, *Nucleic Acids Research*, 31(1), 452–455.
[E53] L. Holm, C. Sander, 1998, Touring protein fold space with *DALI/FSSP*, *Nucleic Acids Research*, 26, 316–319.
[E54] R. Finn, M. Marshall, A. Bateman, 2005, *iPfam*: visualization of protein-protein interactions in *PDB* at domain and amino acid resolutions, *Bioinformatics*, 21, 410–412.
[E55] A. Andreeva, D. Howorth, S. Brenner, T. Hubbard, C. Chothia, A. Murzin, 2004, *SCOP* database in 2004: refinements integrate structure and sequence family data, *Nucleic Acids Research*, 32, D226–D229.
[E56] M. Tress, J. Cheng, P. Baldi, K. Joo, J. Lee, J.H. Seo, J. Lee, D. Baker, D. Chivian, D. Kim, Ezkurdia, I., 2007, Assessment of predictions submitted for the CASP7 domain prediction category, *Proteins: Structure, Function, and Bioinformatics*, 69(S8), 137–151.
[E57] P. Lukasiak, K. Fidelis, M. Antczak, A. Kryshtafovych, J. Blazewicz, 2011, Automated prediction of 3D structure of proteins based on descriptors approach, in *Machine Learning Reports* 01/2011, 19–25.
[E58] B. Brooks, R. Bruccoleri, B. Olafson, D. States, S. Swaminathan, M. Karplus, 1983, *CHARMM*: a program for macromolecular energy, minimization, and dynamics calculations, Journal of Computational Chemistry, 4, 187–217.

[E59] Z. Hongyi, Z. Yaoqi, 2002, Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction, *Protein Science*, 11, 2714–2726.
[E60] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, 1953, Equation of State Calculations by Fast Computing Machines, *Journal of Chemical Physics*, 21, 1087.
[E61] M. Antczak, P. Lukasiak, K. Fidelis, A. Kryshtafovych, J. Blazewicz, 2012, Protein structure quality assessment and modeling support techniques, in *Machine Learning Reports* 01/2012, 21-26.
[E62] A. Zemla, C. Venclovas, J. Moult, K. Fidelis, 1999, Processing and analysis of *CASP3* protein structure predictions, *Proteins*, 37, 22–29.
[E63] A. Kryshtafovych, J. Moult, S.G. Bartual, H. Bazan, D.E. Casteel, E. Christodoulou, J.K. Everett, J. Hausmann, T. Heidebrecht, T. Hills, R. Hui, J. F. Hunt, J. Seetharaman, A. Joachimiak, M.A. Kennedy, C. Kim, A. Lingel, K. Michalska, G.T. Montelione, J.M. Otero, A. Perrakis, J.C. Pizarro, M.J. van Raaij, T.A. Ramelot, F. Rousseau, L. Tong, A.K. Wernimont, J. Young, T. Schwede, 2011, Target highlights in CASP9: Experimental target structures for the critical assessment of techniques for protein structure prediction, *Proteins*, 79, 6–20.
[E64] C. Laing, T. Schlick, 2011, Computational approaches to RNA structure prediction, analysis, and design, *Current Opinion in Structural Biology*, 21, 306–318.
[E65] N. Leontis, E. Westhof, 2012, RNA 3d Structure Analysis and Prediction. *Springer Science & Business Media.*
[E66] J. Moult, J.T. Pedersen, R. Judson, K. Fidelis, 1995, A large-scale experiment to assess protein structure prediction methods, *Proteins*, 23, ii–iv.
[E67] S.W. Burge, J. Daub, R. Eberhardt, J. Tate, L. Barquist, E.P. Nawrocki, S.R. Eddy, P.P. Gardner, A. Bateman, 2013, Rfam 11.0: 10 years of RNA families, *Nucleic Acids Research*, 41, D226–D232.
[E68] M. Sarver, C.L. Zirbel, J. Stombaugh, A. Mokdad, N.B. Leontis 2007, *FR3D*: finding local and composite recurrent structural motifs in RNA 3D structures, *Journal of Mathematical Biology*, 56(1-2), 215–252.
[E69] D. Lai, J.R. Proctor, J.Y.A. Zhu, I.M. Meyer, 2012, *R-CHIE*: a web server and R package for visualizing RNA secondary structures, *Nucleic Acids Research*, 40(12), e95–e95.
[E70] C.E. Hajdin, F. Ding, N.V. Dokholyan, K.M. Weeks, 2010, On the significance of an RNA tertiary structure prediction. RNA, 16, 1340–1349.
[E71] C.D. Schwieters, J.J. Kuszewski, N. Tjandra, G.M. Clore, 2003, The *Xplor-NIH* NMR molecular structure determination package, *Journal of Magnetic Resonance*, 160, 65–73.
[E72] M.R. Garey, D.S. Johnson, 1979, Computers and intractability: A guide to the theory of NP-completeness. *W.H. Freeman*, San Francisco.
[E73] A. Kurnia, A. Lim, B. Cheang, 2001, Online judge, *Computers & Education*, 36, 4, 299–315.
[E74] H. Müller, T. Norman, D. Kennedy, G. Srinivasa, A. Mamonov, N. Preuss, J. Kalpathy-Cramer, A. Hanbury, K. Farahani, R. Sergeev, J.H. Paik, A. Klein, A. Criminisi, A.Trister, 2016, Report on the cloud-based evaluation approaches workshop 2015. *ACM SIGIR Forum*, 50, 1 38–41.
[E75] J. Howe, 2006. The rise of crowdsourcing, Wired Magazine 14, 6, 1–4.

[E76] P.J.A. Cock, C.J. Fields, N. Goto, M.L. Heuer, P.M. Rice, 2010, The Sanger *FASTQ* file format for sequences with quality scores, and the Solexa/Illumina *FASTQ* variants, *Nucleic Acids Research*, 38, 1767–1771.
[E77] L. Roguski, S. Deorowicz, 2014, *DSRC 2* – industry-oriented compression of *FASTQ* files, *Bioinformatics,* 30, 2213–2215.
[E78] D.C. Jones, W.L. Ruzzo, X. Peng, M.G. Katze, 2012, Compression of next-generation sequencing reads aided by highly efficient de novo assembly, *Nucleic Acids Research*, 40, e171.
[E79] F. Hach, I. Numanagic, C. Alkan, S.C. Sahinalp, 2012, *SCALCE*: boosting sequence compression algorithms using locally consistent encoding, *Bioinformatics*, 28, 3051–3057.

17.02.2019 Maciej Antczak