

Tomasz Woźniak

**AUTOMATYZACJA ANALIZY WIDM NMR
W PROCESIE USTALANIA STRUKTURY
PRZESTRZENNEJ RNA.**

Pracę wykonano w Instytucie Chemii Bioorganicznej
Polskiej Akademii Nauk w Poznaniu.

Promotor pracy: Prof. dr hab. Ryszard W. Adamiak

Praca została przedstawiona
Radzie Instytutu Chemii Bioorganicznej
Polskiej Akademii Nauk w Poznaniu
celem uzyskania stopnia
doktora nauk chemicznych.

Pracę dedykuję moim Rodzicom

Podziękowania

Serdecznie dziękuję mojemu promotorowi prof. dr hab. Ryszardowi W. Adamiakowi za opiekę naukową, przekazaną wiedzę i wartościowe uwagi.

Dziękuję prof. dr hab. Zofii Gdaniec, za pomoc, wsparcie i dobrą radę.

Dziękuję dr Dorocie Gudanis za opinie i uwagi merytoryczne oraz redakcyjne na temat niniejszej pracy.

Dziękuję dr Magdalenie Małgowskiej, mgr Annie Teubert, dr Danielowi Baranowskiemu, oraz wszystkim członkom Zespołu Struktury i Funkcji RNA za dobrą atmosferę w pracy, ciekawe dyskusje a także wszelką pomoc.

Dziękuję Pauli Michalak za wsparcie w ciężkich chwilach.

Dziękuję rodzicom, bratu i babci oraz przyjaciołom za motywację do badań.

SPIS TREŚCI

WYKAZ SKRÓTÓW I NAZW WŁASNYCH	3
I. CEL PRACY	9
II. CZĘŚĆ LITERATUROWA	11
II.1. Wprowadzenie	11
II.2. Budowa RNA	14
II.3. Spektroskopia NMR kwasów nukleinowych	25
II.3.1 Zjawisko rezonansu	27
II.3.2 Przesunięcie chemiczne	29
II.3.3 Objętości i intensywności sygnałów	30
II.3.4 Sprzężenie spinowo-spinowe	30
II.3.5 Efekty relaksacyjne	32
II.3.6 Budowa spektrometru NMR	32
II.3.7 Rejestracja widm NMR	33
II.3.8 Wielowymiarowe widma NMR	37
II.3.9 Izotopowe wzbogacanie próbek	38
II.4. Rozwiązywanie struktur przestrzennych RNA w oparciu o metody spektroskopii NMR	40
II.4.1 Charakterystyczne regiony na widmach NMR dla kwasów nukleinowych	40
II.4.2 Rozwiązywanie struktury w oparciu o analizę sekwencyjnych ścieżek NOE	41
II.4.3 Wyznaczanie więzów strukturalnych	42
II.4.4 Tworzenie modelu w oparciu o metody dynamiki molekularnej	43
II.5. Algorytmy służące do przypisywania sygnałów na widmach NMR RNA	45
III. MATERIAŁY I METODY	47
III.1. Język programowania Python	47
III.2. Biblioteki programistyczne i skrypty	48
III.3. Wzorzec projektowy Django	49
III.4. Biblioteki wykorzystane w interfejsie webowym	49
III.5. Baza danych PostgreSQL	49
III.6. System kolejkowy	50
III.7. Oprogramowanie dla danych NMR	50
III.8. Inne narzędzia	51
III.9. Dane wykorzystane w programie	51

IV.	WYNIKI BADAŃ WŁASNYCH	53
IV.1.	Schemat logiczny działania programu RNA NOE <i>pathfinder</i>	53
IV.2.	Algorytmy przetwarzania wielokierunkowych widm NMR dlaRNA	57
IV.2.1	Rozróżnianie sygnałów od szumów	57
IV.2.2	Wyszukanie regionów na widmach	61
IV.2.3	Korelacja sygnałów między widmami	65
IV.2.4	Wybór optymalnego rodzaju ścieżki NOE do obliczeń	68
IV.2.5	Generowanie ścieżek w rejonie <i>aromatyczno-anomerycznym</i> widma NOESY	71
IV.2.6	Przypisanie sygnałów nie należących do ścieżek NOE	74
IV.2.7	Generowanie ścieżek w innych regionach widma NOESY	74
IV.2.8	Analiza widm 3D	75
IV.3.	Integracja oprogramowania RNA NOE <i>pathfinder</i> i RNAComposer	75
IV.4.	Testy oprogramowania	76
IV.5.	Webserwer RNA NOE <i>pathfinder</i>	79
IV.5.1	Schemat działania webserwisu	79
IV.5.2	Bezpieczeństwo serwisu	82
IV.6.	Programy towarzyszące	85
IV.6.1	SPEC: narzędzie do pobierania sygnałów z widm NMR z użyciem programu Felix	86
IV.6.2	PDBinout: narzędzie do konwersji i personalizowania plików PDB	86
IV.6.3	TotalRestrains: narzędzie do przetwarzania i przygotowania więzów eksperymentalnych	87
IV.6.4	Program do analizy niespełnionych więzów odległościowych „BrokenRestrainsAnalyser”	87
IV.	DYSKUSJA	88
V.	PODSUMOWANIE	90
VI.	LITERATURA	92

WYKAZ SKRÓTÓW I NAZW WŁASNYCH

AFM	mikroskopia sił atomowych (ang. <i>Atomic Force Microscopy</i>)
AJAX	Technika tworzenia stron internetowych, w której elementy strony komunikują się z serwerem asynchronicznie – bez przeładowywania strony (ang. <i>Asynchronous JavaScript and XML</i>)
Blueprint	Szablon do tworzenia stylów stron internetowych
biblioteka (programistyczna)	zbiór klas, metod, wartości, który może być wykorzystywany do tworzenia oprogramowania w oparciu o znajdujące się w nim elementy
Biological Magnetic Resonance Data Bank	Baza danych eksperymentalnych dla spektroskopii biomolekularnego NMR
Bootstrap	szablon HTML, CSS i JavaScript do tworzenia stron internetowych
BPMN	graficzna notacja pozwalająca na zapis modelu problemu biznesowego (ang. <i>Business Process Modeling Notation</i>)
CAPTCHA	test weryfikujący czy użytkownik danej strony jest człowiekiem czy skryptem lub programem (ang. <i>Completely Automated Public Turing test to tell Computers and Humans Apart</i>)
CD	dichroizm kołowy (ang. <i>Circular Dichroism</i>)
Celery	oprogramowanie do kolejkowania zadań wykonywanych asynchronicznie
CNS Solve	Program służący do rozwiązywania struktur przestrzennych z wykorzystaniem danych krystalograficznych lub uzyskanych dzięki spektroskopii NMR (ang. <i>Crystallography and NMR System</i>)
Cron	program (demon) służący do uruchamiania zadań w określonym czasie
cryoEM	mikroskopia elektronowa
CSRF	Atak polegający na wykonywaniu czynności w serwisie bez wiedzy i zgody osoby atakowanej przez kod zamieszczony na innej stronie (ang. <i>Cross-Site Request Forgery</i>)
CSS	kaskadowe arkusze stylów – język pozwalający na zdefiniowanie wyglądu stron internetowych (ang. <i>Cascading Style Sheets</i>)
DNA	kwasy deoksyrybonukleinowe
DNS	system nazw domenowych, umożliwiający tłumaczenie nazw domen na adresy IP i odwrotnie (ang. <i>Domain Name System</i>)
Felix	program do analizy wielowymiarowych widm NMR
FID	sygnał precesji swobodnej (ang. <i>Free Induction Decay</i>)
FLYA	jeden z algorytmów do przypisywania sygnałów na wimach NMR
FN	falszywe wartości negatywne (ang. <i>false negatives</i>)
FP	falszywe wartości pozytywne (ang. <i>false positives</i>)
GIT	rozproszony system kontroli wersji

GPL	licencja wolnego oprogramowania stworzona dla celów projektu GNU (ang. <i>GNU General Public License</i>)
HADDOCK	program służący do dokowania cząsteczek
honeypot	pułapka pozwalająca na wykrywanie prób nieautoryzowanego dostępu np. poprzez udostępnienie fałszywego panelu logowania
HTTPS	Bezpieczny protokół sieciowy, dzięki któremu komunikacja klient – serwer jest szyfrowana (ang. <i>Hypertext Transfer Protocol Secure</i>).
impuls RF	impuls częstotliwości radiowej
ISPA	metoda wyznaczania więzów odległościowych poprzez przybliżenie dla pary izolowanych spinów (ang. <i>isolated spin pair approximation</i>)
Inkscape	program do tworzenia grafiki wektorowej
JavaScript	Język programowania wykorzystywany przy tworzeniu stron internetowych; kod napisany za pomocą JavaScript jest uruchamiany po stronie klienta
JIT	metoda która polega na kompilowaniu kodu tuż przed jego uruchomieniem (ang. <i>just-in-time compilation</i>)
klasa	definicja obiektów danego typu
klastrowanie	przypisywanie obiektów do grup (klastrow) na podstawie podobieństwa określonych cech i ich różnic pomiędzy poszczególnymi grupami
LGPL	licencja wolnego oprogramowania z większymi uprawnieniami niż licencja GPL (ang. <i>GNU Lesser General Public License</i>)
licencja BSD	licencja wolnego oprogramowania opracowana na Uniwersytecie Kalifornijskim w Berkeley (ang. <i>Berkeley Software Distribution License</i>)
lista	struktura danych, w której elementy ustawione są w określonej kolejności
NMR	Magnetyczny Rezonans Jądrowy (ang. <i>Nuclear Magnetic Resonance</i>)
man in the middle	Dosłownie „człowiek pomiędzy” - atak polegający na przejęciu połączenia pomiędzy użytkownikiem a serwerem, co umożliwia posłuchanie i modyfikację przesyłanych wiadomości
matching, (NMR)	dopasowanie efektywności filtra pasmowego
czas mieszania, τ_m	czas pomiędzy dwoma impulsami w sekwencji eksperymentu NOESY podczas którego następuje relaksacja, dzięki czemu możliwe jest obserwowanie efektu NOE (ang. <i>mixing time</i>)
MMTSB	zestaw narzędzi przydatnych m.in. do modelowania molekularnego (ang. <i>Multiscale Modeling Tools for Structural Biology</i>)
Model-Szablon-Widok	wzorzec w którym warstwa modelu (tabele w bazie danych), szablonu opisującego widok wyświetlany użytkownikowi i kontrolera (widoku) są rozdzielone (ang. <i>Model-Template-View</i>)
MySQL	system zarządzania bazami danych SQL rozwijany przez firmę Oracle Corporation
obiekt	struktura która składa się zarówno z danych jak i z metod które operują

	na tych danych
PDB	baza danych struktur przestrzennych cząsteczek (ang. <i>Protein Data Bank</i>)
peak picking	proces zaznaczania sygnałów na widmie
PIP	program służący do pobierania i instalowania bibliotek języka Python
PostgreSQL	system zarządzania bazami danych SQL rozwijany przez PostgreSQL Global Development Group
ppm	części na milion – jednostka przesunięcia chemicznego (ang. <i>parts per milion</i>)
programowanie obiektowe	Wzorzec w którym tworzy się oprogramowanie poprzez łączenie informacji o stanie i zachowaniu w oddziałujące ze sobą obiekty
PyCharm	edytor tekstu wspierający pisanie kodu w języku Python
PyPy	alternatywna implementacja języka Python
Python	język programowania wysokiego poziomu
RabbitMQ	program będący pośrednikiem przy przekazywaniu wiadomości pomiędzy różnymi programami lub systemami
region <i>aromatyczno-anomeryczny</i>	Region na widmach NMR dla kwasów nukleinowych, na widmie 2D NOESY w tym regionie można wyznaczyć ścieżkę NOE H1'-H6/H8
rekurencja	proces polegający na wywoływaniu funkcji lub metody przez samą siebie
RNA	kwasy rybonukleinowe
RNAComposer	webserwer służący do modelowania struktur 3D dużych cząsteczek RNA
serwer www	komputer z oprogramowaniem obsługującym żądania protokołu HTTP, dzięki czemu użytkownicy łączący się z tym serwerem mogą wyświetlać strony umieszczone na tym serwerze
SHAPE	metoda badania struktury drugorzędowej cząsteczek RNA z wykorzystaniem selektywnej acylacji grupy 2'-hydroksylowej rybozy i wydłużania starterów (ang. <i>Selective 2'-Hydroxyl Acylation analyzed by Primer Extension</i>)
shimming	optymalizacja jednorodności pola mająca na celu poprawę rozdzielczości sygnałów na widmie
słownik	struktura danych, w której wartości są przypisane do odpowiednich kluczy
SQLite	system zarządzania bazami danych SQL, w którym dane przechowywane są w jednym pliku
SQL injection	atak polegający na wykonaniu zapytania SQL poprzez lukę w zabezpieczeniach serwisu internetowego
sygnał lock'u	sygnał pochodzący od deuterowanego rozpuszczalnika, wykorzystywany do kompensacji dryftu pola
ścieżka NOE	ścieżka transferu magnetyzacji pomiędzy atomami należącymi do kolejnych reszt nukleotydowych, którą można wyznaczyć na widmach 2D NOESY
tabu-search	algorytm wyszukiwania, który prócz znajdowania lokalnego optimum

	pozwała na kroki nieoptymalne jeśli w kolejnych krokach może to poprawić wartość oceny
test jednostkowy	test automatyczny sprawdzający poprawność niewielkiej ilości kodu (ang. <i>unittest</i>)
TMS	tetrametylosilan
TN	prawdziwe wartości negatywne (ang. <i>true negatives</i>)
TopSpin	oprogramowanie służące do rejestracji i obróbki widm NMR
TP	prawdziwe wartości pozytywne (ang. <i>true positives</i>)
tuning (NMR)	dostrojenie częstości rezonansowej dla danej próbki
UPGMA	metoda klastrowania hierarchicznego, w której wielokrotnie łączy się najbardziej podobne elementy lub grupy elementów (ang. <i>Unweighted Pair Group Method with Arithmetic Mean</i>)
VEGA	program do konwertowania, manipulacji i wizualizacji plików PDB
widmo COSY	widmo wykonane w technice spektroskopii korelacyjnej (ang. <i>Correlation Spectroscopy spectrum</i>)
widmo DQF-COSY	widmo wykonane w technice spektroskopii korelacyjnej z wykorzystaniem filtru dwukwantowego (ang. <i>Double Quantum Filtered Correlation Spectroscopy spectrum</i>)
widmo HSQC	widmo wykonane w technice heterojądrowej, jednokwantowej spektroskopii korelacyjnej (ang. <i>Heteronuclear Single-Quantum Correlation Spectroscopy spectrum</i>)
widmo HMBC	widmo wykonane w technice heterojądrowej spektroskopii korelacyjnej przez wiele wiązań (ang. <i>Heteronuclear Multiple Bond Correlation Spectroscopy spectrum</i>)
widmo NOESY	widmo pozwalające na obserwację jądrowego efektu Overhausera (ang. <i>Nuclear Overhauser Enhancement Spectroscopy spectrum</i>)
widmo TOCSY	widmo wykonane w technice spektroskopii korelacyjnej pozwalające na obserwację całych układów spinowych (ang. <i>Total Correlation Spectroscopy spectrum</i>)
wielowątkowość	możliwość wykonywania różnych obliczeń w tym samym czasie
więzy na kąty torsyjne	więzy nakładane na podstawie danych eksperymentalnych na modelowaną cząsteczkę w celu ograniczenia zakresu dla określonych kątów torsyjnych
więzy odległościowe	więzy nakładane na podstawie danych eksperymentalnych na modelowaną cząsteczkę w celu ograniczenia odległości pomiędzy określonymi atomami
Virtualenv	narzędzie pozwalające na instalowanie wirtualnych środowisk stworzonych dla języka Python
wzorzec projektowy	sprawdzone rozwiązanie lub reguła, która może być wykorzystywana przy budowie różnorodnych programów i aplikacji
XPLOR-NIH	Program służący do rozwiązywania struktur 3D różnych cząsteczek z wykorzystaniem danych eksperymentalnych

XSS	Atak polegający na umieszczeniu niepożądanego kodu na stronie wyświetlanej użytkownikowi (ang. <i>Cross-site scripting</i>)
YAML	format służący do zapisywania strukturalizowanych danych w sposób czytelny zarówno dla człowieka jak i programów komputerowych (ang. <i>YAML Ain't Markup Language</i>)

ABSTRACT

Process of NMR spectra analysis for nucleic acids is very complex. Important phase required for cross-peaks assignment is process of finding NOE pathways – magnetization transfer pathways between protons of consecutive residues engaged in base-pair creation, and assignment on the basis of found pathways. Because of the small differences in chemical shift values for RNA molecules and existence of only four types of residues, unambiguous assignment of cross-peaks is time-consuming process. Research done in our Laboratory and prof. J. Błazewicz Bioinformatics Laboratory resulted in system for analysis NOE pathways on the basis of *tabu-search* algorithm (1). Unfortunately, despite high computational speed this solution requires additional expert knowledge. Moreover because it doesn't use various important information, needed for unambiguous and automatic finding of NOE pathways, it isn't efficient enough. In 2013 program using chemical shifts statistics and FLYA algorithm was created. This algorithm can assign most of cross-peaks in *aromatic-anomeric* region (2).

The purpose of my research was to design and implement of informatics system for multidimensional NMR spectra analysis for [^{13}C , ^{15}N] labeled RNA. This system called RNA NOE*pathfinder* will considerably automatize process of NOE pathways finding, cross-peaks assignment and create structural restraints.

I created program, that performs automatic NMR spectra analysis, reduces noise, splits spectra into regions, and searches for correlations between various spectra. These correlations are used to partially assign some of the cross-peaks. In the next steps, using given RNA sequence and its secondary structure (*dot-bracket* format), short pathway fragments are generated in *aromatic-anomeric* region. These fragments are assigned to sequence and combined with each other to create pairs, that are combined with each other again. Subsequently gaps between fragments and missing pathway fragments at the ends are filled. During each step part of the worst solutions are removed, generated pathways are clustered and scored. Finding NOE pathways allows to assign remaining cross-peaks, and generate distance restraints, that may be used in modeling procedure with molecular dynamics. Program is located on computational server and access is given through web interface located at <http://rnanoepathfinder.ibch.poznan.pl>. In RNA NOE*pathfinder* program automatic peak picking procedure implemented in Felix 2007 (Felix NMR Inc.) program and Topspin 3 (Bruker) are utilized. Peak list generated that way are used as input data. Moreover RNA NOE*pathfinder* allows to introduce modified nucleotide residues using its internal files system.

I. CEL PRACY

Proces analizy widm NMR dla kwasów nukleinowych jest bardzo złożony. Istotnym jego etapem, koniecznym dla przypisania sygnałów, jest wyznaczenie tzw. „ścieżek NOE”- ścieżek transferu magnetyzacji pomiędzy protonami kolejnych reszt nukleotydowych zaangażowanych w tworzenie par zasad i przypisaniu sygnałów na podstawie wyznaczonych w ten sposób ścieżek. Ze względu na niewielkie różnice wartości przesunięć chemicznych dla RNA i obecność jedynie czterech rodzajów reszt nukleotydowych, jednoznaczne przypisanie sygnałów jest niezwykle czasochłonne. Dotychczas prowadzone badania w naszej Pracowni i Pracowni Bioinformatyki prof. J. Błażewicza zaowocowały systemem analizy ścieżek NOE w oparciu o algorytm *tabu-search*. (1). Niestety system ten, mimo dużej szybkości obliczeń, wymaga dodatkowej wiedzy eksperckiej (pracy eksperymentatora) i nie uwzględnia wielu istotnych informacji, bez których jednoznaczne i automatyczne wyznaczenie ścieżek NOE nie jest wystarczająco efektywne. W 2013 roku ukazał się program, który na podstawie opracowanych statystyk dotyczących wartości przesunięć chemicznych oraz wykorzystaniu algorytmu FLYA, może dokonywać przypisywania większości sygnałów w regionie *aromatyczno-anomerycznym* (2).

Celem moich badań było zaprojektowanie i zaimplementowanie informatycznego systemu analizy widm wielokierunkowych NMR dla znakowanych [^{13}C , ^{15}N] RNA, który w dużym stopniu zautomatyzuje proces generacji ścieżek NOE, przypisywania sygnałów i wyznaczania więzów strukturalnych o nazwie RNA NOEpathfinder.

Stworzyłem program, który automatycznie analizuje widma NMR, redukuje szумы, dzieli widma na regiony, a następnie wyszukuje korelacje pomiędzy sygnałami na różnych widmach. Na podstawie tych korelacji przypisywana jest część sygnałów. W dalszych krokach korzystając z podanej sekwencji RNA i struktury drugorzędowej (format *dot-bracket*) program ten generuje krótkie fragmenty ścieżek NOE w regionie *aromatyczno-anomerycznym*, które następnie dopasowuje do sekwencji i kombinuje ze sobą. Tak powstałe pary następnie kombinatorycznie łączy ze sobą a następnie uzupełnia luki pomiędzy fragmentami ścieżek oraz brakujące fragmenty na końcach ścieżek. Na każdym etapie odrzucana jest część najgorszych rozwiązań, a uzyskane ścieżki są następnie klastrowane na podstawie podobieństwa i oceniane. Wyznaczenie ścieżek NOE pozwala na przypisanie pozostałych sygnałów oraz stworzenie więzów odległościowych, które mogą być wykorzystywane do modelowania cząsteczki z wykorzystaniem metod dynamiki molekularnej. Program ten jest umieszczony na maszynie obliczeniowej, a dostęp do niego odbywa się poprzez webserwis zlokalizowany pod adresem <http://rnanoepathfinder.ibch.poznan.pl>. Program RNA NOEpathfinder wykorzystuje możliwość automatycznego zaznaczania sygnałów, jaka jest zaimplementowana w programie Felix 2007 (Felix NMR Inc.) oraz Topspin 3 (Bruker). Tak wyznaczone sygnały są przekazywane jako dane wejściowe. Ponadto program RNA NOEpathfinder zawiera system modyfikowalnych plików

wewnętrznych, dzięki którym istnieje możliwość m.in. wprowadzania modyfikacji reszt nukleotydowych.

II. CZĘŚĆ LITERATUROWA

II.1 Wprowadzenie

Cząsteczki kwasów nukleinowych, do których zalicza się DNA (kwas deoksyrybonukleinowy) i RNA (kwas rybonukleinowy) są długimi liniowymi polimerami. Odpowiadają one za przechowywanie i przekazywanie informacji genetycznej w komórce. Cząsteczki RNA pełnią ponadto wiele kluczowych funkcji w komórce. Zalicza się do nich przede wszystkim udział w ekspresji genów. W komórce występują cztery główne typy cząsteczek RNA:

- **mRNA** – informacyjny (matrycowy) RNA
- **tRNA** – transportujący (transferowy) RNA
- **rRNA** – rybosomowy RNA
- **ncRNA** – niekodujący RNA

mRNA

Cząsteczki informacyjnego RNA tworzą różnorodną grupę cząsteczek, której zadaniem jest przenoszenie informacji genetycznej. Średnia długość mRNA dla ssaków wynosi 2,2 tysiąca zasad (3). Cząsteczki te są rozpoznawane i wiązane przez rybosomy, gdzie służą jako matryca dla syntezy białek. Na końcach cząsteczek mRNA znajdują się regiony niekodujące, tzw. UTR (ang. *untranslated regions*). W regionach tych często tworzą się złożone struktury drugo- i trzeciorzędowe mające wpływ na ekspresję genów.

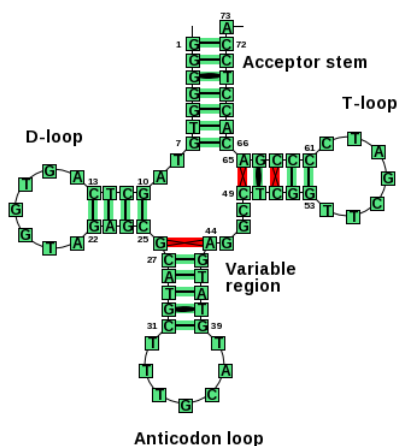
tRNA

Cząsteczki tRNA stanowią około 15 procent RNA obecnego w komórce (4). Struktura drugorzędowa cząsteczek tRNA ma kształt „liścia koniczyny” przedstawiony na rysunku 1 (5). Cząsteczki tRNA zawierają liczne modyfikowane lub hipermodyfikowane reszty nukleotydowe. Struktura trzeciorzędowa przypomina literę „L” na końcach której znajduje się ramię antykodonowe i akceptorowe (rysunek 2). Ramię akceptorowe posiada charakterystyczną sekwencję CCA, do której wiąże się aminokwas.

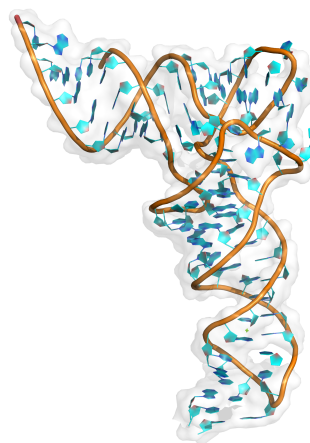
Cząsteczki tRNA pełnią istotną rolę w translacji. Wiążą się one specyficznie do syntetaz aminoacylo-tRNA, które do ich 3' końca dołączają aminokwas odpowiadający antykodonowi danego tRNA. Aminoacylowane tRNA są wykorzystywane w syntezie łańcucha polipeptydowego na rybosomie. Cząsteczki tRNA o sekwencji antykodonu komplementarnej do sekwencji mRNA są przyłączane w miejscu A, po czym następuje dołączenie aminokwasu do rosnącego łańcucha polipeptydowego, a cząsteczka tRNA zajmuje miejsce P.

rRNA

Cząsteczki rRNA stanowią około 80% puli RNA w komórkach (4) i wraz z białkami wchodzi w skład rybosomów (rys. 3). Rybosomy składają się z dużej i małej podjednostki, które u Eucaryota mają stałe Svedberga 60S i 40S a u Procaryota 50S i 30S odpowiednio. Rybosomy

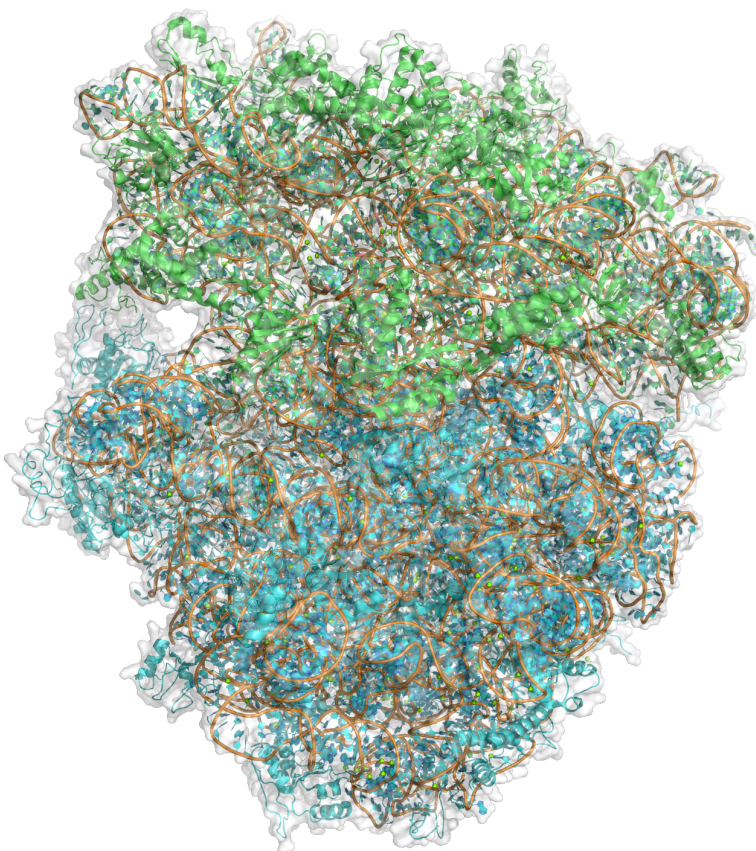


Rys. 1. Struktura drugorzędowa tRNA człowieka kodującego alaninę, rysunek pobrany z bazy tRNAdb (<http://trna.bioinf.uni-leipzig.de>)



Rys. 2. Struktura przestrzenna tRNA z drożdży, PDB ID: 2TRA

odpowiadają za katalizowanie reakcji dołączania aminokwasów do syntetyzowanego białka. Rolę katalityczną pełnią w tym przypadku cząsteczki rRNA, które znajdują się we wszystkich kluczowych miejscach dla tego procesu (6). Rybosom prokariotyczny ze względu na różnice w budowie względem rybosomu eukariotycznego stanowi jeden z bardziej istotnych celów terapeutycznych (7).



Rys. 3. Struktura przestrzenna rybosomu z *Thermus Thermophilus*, PDB ID: 2WDK, 2WDL

NcRNA

Wykazano, że aż 90% eukariotycznego, genomowego DNA może ulegać transkrypcji (8, 9). Niewielki procent z tej puli (1-2%) koduje białka, podczas gdy większość stanowią cząsteczki niekodującego RNA. Do cząsteczek ncRNA zaliczamy takie cząsteczki jak: miRNA, piRNA, siRNA, lncRNA, eRNA czy RNA związane z promotorem. Oto ich krótka charakterystyka. **MiRNA** są to krótkie (20-24 reszty nukleotydowe) cząsteczki odpowiedzialne za regulację około 50% genów (8, 10). Powstają w wyniku cięć katalitycznych struktur spinkowych przez enzymy Drosha i Dicer (8, 11, 12). Oddziałują z białkiem Argonaute wchodzącym w skład kompleksu RISC wyciszającego geny. **PiRNA** są krótkimi (24-31 reszt nukleotydowych) cząsteczkami wchodzącymi w oddziaływanie z białkami Pivi (8, 13). Ich główną rolą jest blokowanie aktywności transpozonojowej podczas rozwoju linii zarodkowych (8, 14). **SiRNA** są to krótkie (20-24 nukleotydowe), dwuniciowe cząsteczki powstające w wyniku cięć katalitycznych przez białko Dicer i biorące udział w wyciszaniu genów poprzez oddziaływanie z kompleksem RISC. SiRNA mogą również wpływać na ekspresję genów poprzez zwiększanie heterochromatyzacji DNA (8, 15). **LncRNA** są klasyfikowane jako cząsteczki o długości powyżej 200 reszt nukleotydowych, które charakteryzują się lokalizacją jądrową, niskim stopniem zachowawczości i ekspresją (8, 16). Ich rola w komórce jest bardzo zróżnicowana. Cząsteczki **eRNA** ulegają ekspresji w rejonach enhancerowych w związku z czym posiadają specyficzne dla nich wzorce metylacji. Ich długość waha się w granicach 0.1 do 9 tysięcy par zasad (8, 17). Sugeruje się ich rolę jako aktywatorów transkrypcji. Cząsteczki **RNA związane z promotorem** to szereg cząsteczek wiążących się w rejonach promotorowych genów. Ich długość może wynosić od 16 do ponad 200 reszt nukleotydowych. Wiadomo, że wpływają na proces transkrypcji, jednak ich dokładna funkcja nie jest w pełni poznana (8, 18).

Poza wymienionymi wyżej rodzajami cząsteczek, istnieją również cząsteczki **snoRNA** (ang. *small nucleolar RNA*), które biorą udział w modyfikowaniu cząsteczek rRNA (19). Cząsteczki RNA są również wykorzystywane jako startery (tzw. fragmenty Okazaki), podczas syntezy nowego DNA na nici opóźnionej. Istotnym odkryciem była obserwacja, że cząsteczki RNA są też odpowiedzialne za autokatalityczny splicing, co było pierwszym dowodem na aktywność katalityczną RNA (20). Istnieją również tzw. ryboprzełączniki, które znajdują się w mRNA i poprzez wiązanie się z określoną cząsteczką zmieniają swoją strukturę przestrzenną, czym zmieniają poziom ekspresji genu kodowanego przez dane mRNA.

Dla wielu wirusów RNA stanowi jedyny materiał genetyczny przenoszący informacje genetyczne. Takie rodziny wirusów jak *Paramyxoviridae*, *Rhabdoviridae*, *Arenaviridae*, *Bunyaviridae* czy *Orthomyxoviridae* mają genom w postaci pojedynczej nici RNA (-), *Reoviridae* zawierają dwuniciowe cząsteczki RNA, natomiast wirusy z rodzin *Retroviridae*, *Hepadnaviridae* i *Caulimoviridae* wykorzystują odwrotną transkryptazę do przepisania cząsteczek RNA na DNA (21).

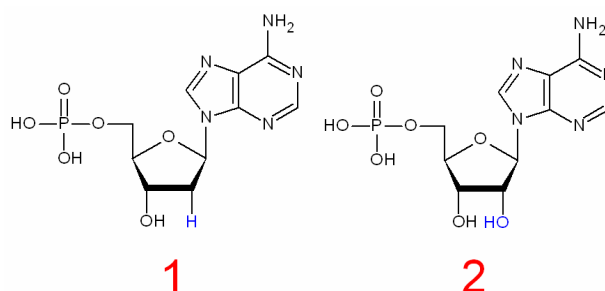
II.2 Budowa RNA

Cząsteczki kwasów nukleinowych są to biopolimery zbudowane z rybonukleotydów połączonych wiązaniami 3'-5' fosfordiestrowymi.

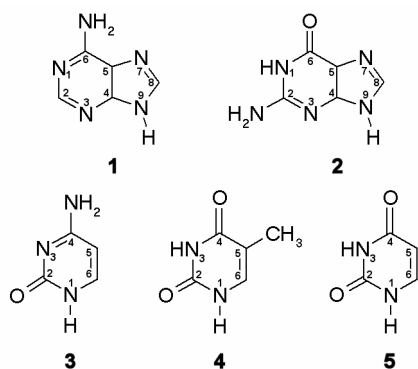
Nukleotydy składają się z:

- zasady azotowej
- części cukrowej
- grupy fosforanowej

Budowę nukleotydów przedstawiłem na rysunku 4.



Rys. 4. Wzory strukturalne nukleotydów – 1) deoksynukleotyd adeninowy, 2- rybonukleotyd adeninowy, na niebiesko zaznaczono grupę wodorotlenową oraz atom wodoru różniące te cząsteczki



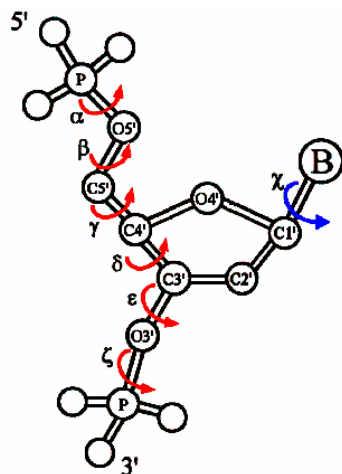
Rys. 5. Wzory zasad azotowych: 1 - adenina, 2 - guanina, 3 - cytozyna, 4 – tymina, 5 - uracyl

Wyróżniamy cztery podstawowe rodzaje zasad azotowych występujące w cząsteczkach RNA: adeninę, guaninę, cytozynę i uracyl. W cząsteczkach DNA zamiast uracylu występuje tymina (rysunek 5). Nukleozydy w łańcuchu połączone są poprzez wiązania 3'-5' fosfordiestrowe. Kolejność ułożenia poszczególnych reszt w cząsteczce RNA nazywamy jej sekwencją nukleotydową.

Konformacja

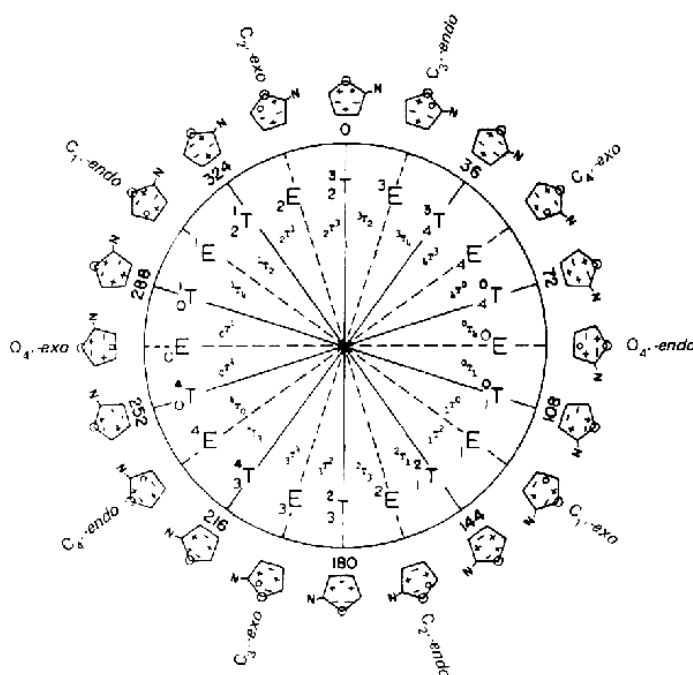
Strukturę przestrzenną oligonukleotydów najczęściej przedstawia się w postaci współrzędnych w trójwymiarowym układzie współrzędnych kartezjańskich. Można również opisać ją z wykorzystaniem kątów torsyjnych pomiędzy określonymi atomami. Dla cząsteczek RNA są to kąty: α (O3'-P-O5'-C5'), β (P-O5'-C5'-C4'), γ (O5'-C5'-C4'-C3'), δ (C5'-C4'-C3'-O3'), ϵ (C4'-C3'-O3'-P), ζ (C3'-O3'-P-O5') oraz kąt χ (O4'-C1'-N9-C4 – dla puryn lub O4'-C1'-N1-C2 – dla pirymidyn), które zaznaczyłem na rysunku 6. Określenie kątów dla zasad azotowych nie jest konieczne, gdyż ich pierścienie są praktycznie płaskie. Część cukrowa może

przyjmować różne konformacje. Z jednej strony można je opisać za pomocą pięciu kątów torsyjnych ($\nu_0, \nu_1, \nu_2, \nu_3, \nu_4$), z drugiej strony pierścień cukrowy dążąc do minimalizowania naprężeń może przyjmować jedynie pewne konformacje, które można opisać za pomocą



Rys. 6. Kąty torsyjne zaznaczone na pojedynczym nukleotydzie – na czerwono kąty $\alpha, \beta, \gamma, \delta, \epsilon, \zeta$, na niebiesko kąt N-glikozydowy χ .

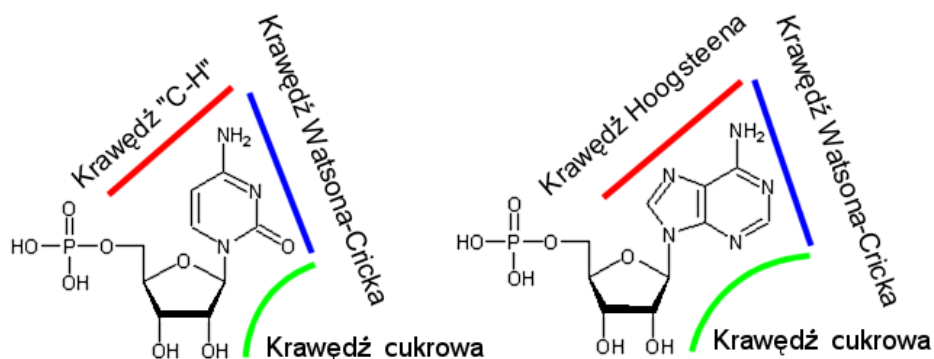
dwóch parametrów: kąta pseudorotacji P oraz amplitudy pofałdowania (22, 23). Zależność kąta pseudorotacji od konformacji przedstawiłem na rysunku 7. Dla cząsteczek RNA najczęściej spotykana jest konformacja C3'-endo, natomiast dla DNA konformacja C2'-endo. Kąt χ definiuje orientację zasady azotowej względem części cukrowej. Jeśli atom O2 pirymidyny lub N3 puryny znajduje się „nad” pierścieniem części cukrowej można mówić o konformacji *syn*, w przeciwnym wypadku jest to konformacja *anti*.



Rys. 7. Koło pseudorotacji opisujące konformację reszty cukrowej. E (ang. *envelope*) – jeden atom wysunięty poza płaszczyznę pierścienia; T (ang. *twist*) – dwa atomy wysunięte poza płaszczyznę pierścienia. Znaki „+” i „-” oznaczają wartość dodatnią lub ujemną kąta torsyjnego (24).

Pary zasad

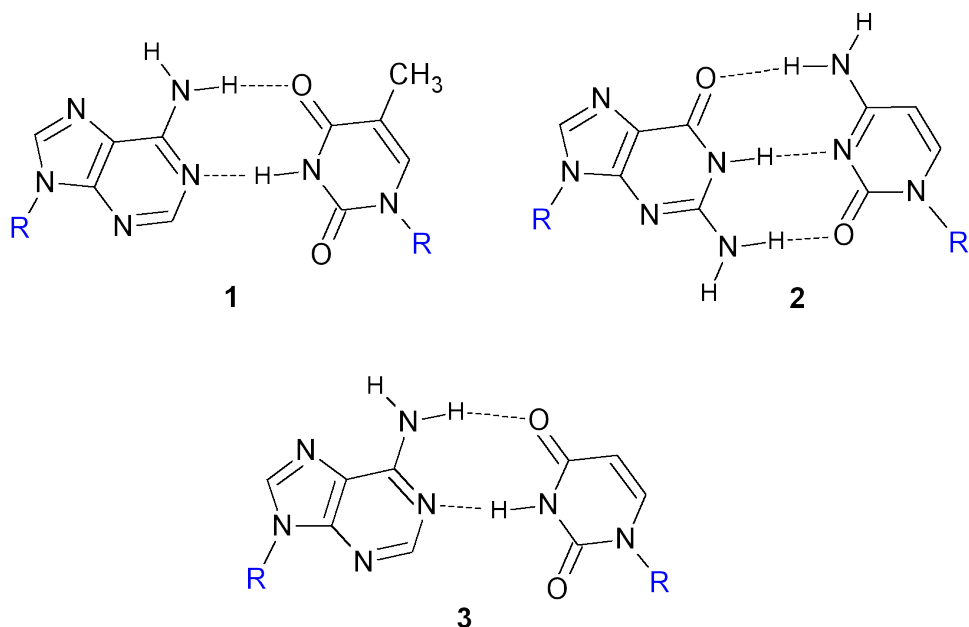
Poszczególne reszty nukleotydowe mogą być zaangażowane w tworzenie par zasad z innymi resztami nukleotydowymi. Nukleotydy posiadają trzy krawędzie zlokalizowane na zasadzie azotowej, które mogą być wykorzystane do tworzenia wiązań wodorowych z innymi zasadami. Zostało to przedstawione na rysunku 8.



Rys. 8. Krawędzie, które mogą zostać wykorzystane przez nukleotydy do tworzenia par zasad.

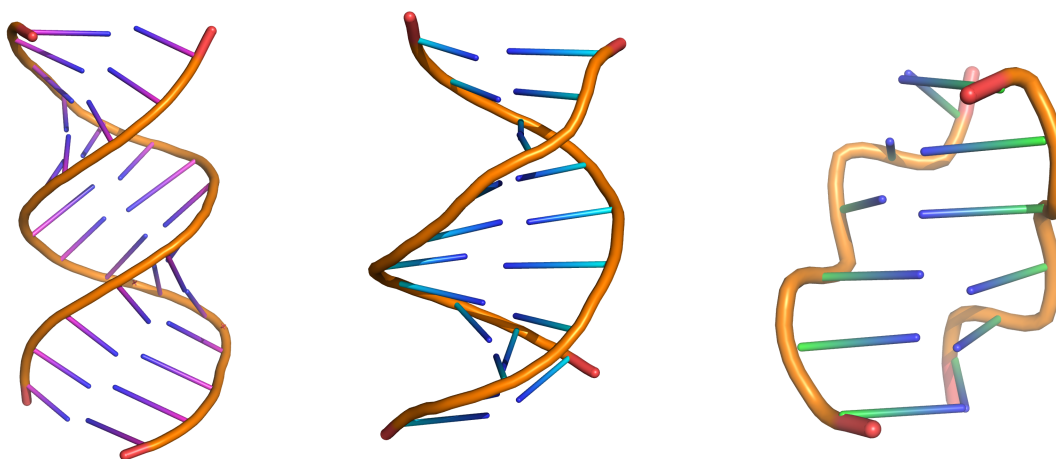
Przykłady tworzonych par zasad

Najczęściej spotykane jest tworzenie par zasad typu Watson-Crick/Watson-Crick pomiędzy resztami guanozyny i cytozyny oraz adenozyiny i urydyny (rys. 9). Niemniej w znanych strukturach RNA stanowią one około 60% wszystkich par zasad. Poznane niekanoniczne pary zasad zostały zebrane w bazie danych Database of non-canonical base pairs (http://prion.bchs.uh.edu/bp_type/) (25).

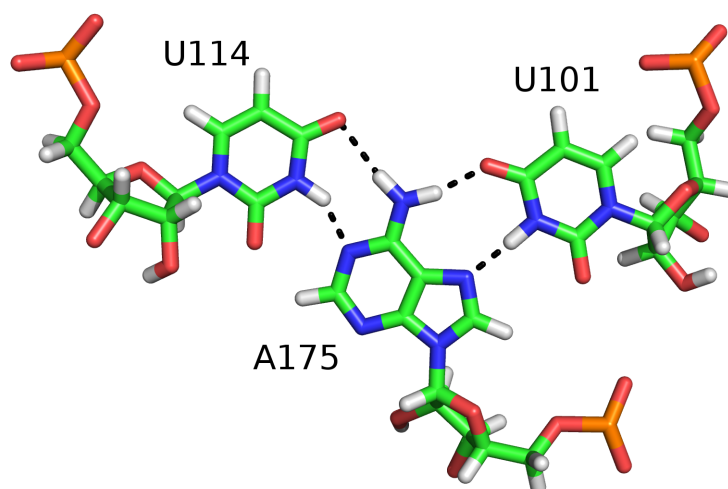


Rys. 9. Kanoniczne pary zasad 1- para A-T; 2- para G-C; 3 – para A-U

Tworzenie par zasad umożliwia powstanie trwałych struktur zbudowanych z dwóch antyrównoległych nici RNA. Układ taki, nazywany helisą RNA jest dodatkowo stabilizowany poprzez oddziaływania warstwowe równoległe ułożonych par zasad. Znane są trzy główne rodzaje helis: A, B oraz Z (rys. 10). Helisy A i B są prawoskrętne, natomiast Z – lewoskrętne (26). Dla cząsteczek RNA typowa jest forma A (27).



Rys. 10. Porównanie struktur przestrzennych helis: A, B, Z DNA (PDB ID: 3V9D, 3IXN, 2HTT odpowiednio)



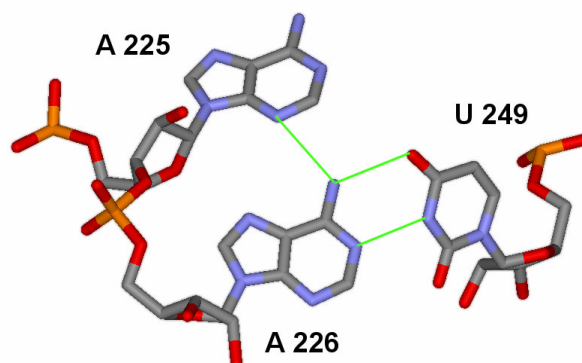
Rys. 11. Trójka zasad U:A-U znajdująca się w strukturze RNA telomerazy ludzkiej (PDBID: 2K95).

Trójki zasad

W cząsteczkach RNA można znaleźć trójki zasad, które tworzą ze sobą wiązania wodorowe. Powstają one w miejscach, gdzie jednoniciowy fragment cząsteczki RNA wchodzi w oddziaływanie z rejonem dwuniciowym (28). Fragment jednoniciowy może oddziaływać z dwuniciowym zarówno w dużym jak i małym rowku helisy. Na rysunku 11 przedstawiłem trójkę zasad znajdującą się w strukturze pseudowęzła RNA telomerazy ludzkiej.

Platforma adenzynowa

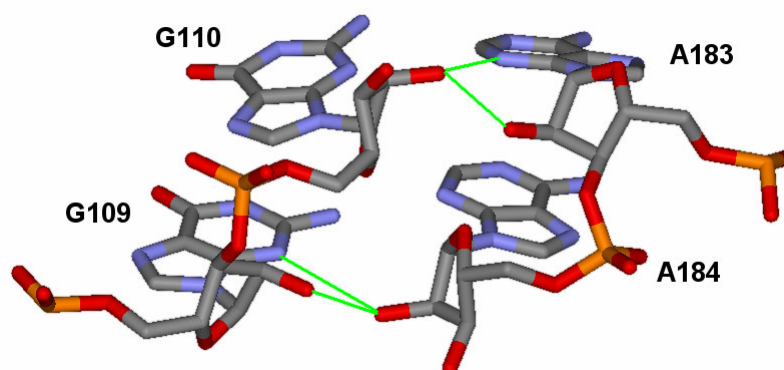
Jest to motyw strukturalny, w którym dwie kolejno ułożone adenyzy układają się w jednej płaszczyźnie i tworzą wiązania wodorowe niezależnie od tworzenia oddziaływania typu Watsona-Cricka z resztą urydyny przez drugą z adenyzy. Poniżej platformy adenzynowej znajduje się niekanoniczna para zasad typu *Hoogsteena* (A-U lub G-U), która stabilizuje taką strukturę. Pierwszy motyw platformy adenzynowej odkryto w domenie P4-P6 grupy I intronu z *Tetrahymena* (rys. 12).



Rys. 12. Motyw platformy adenzynowej, PDB ID: 1HR2

Suwaki rybozowe

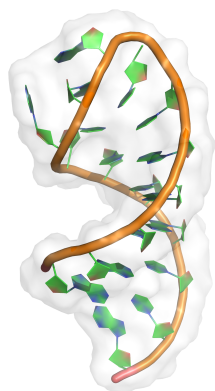
Motyw suwaków rybozowych (ang. *ribose zippers*) występuje dość powszechnie w strukturach RNA. Występuje on, gdy obserwujemy wiązania wodorowe pomiędzy atomem N3 puryny lub O2 pirymidyny a grupami 2'-OH drugiej nici, co widać na rysunku 13.



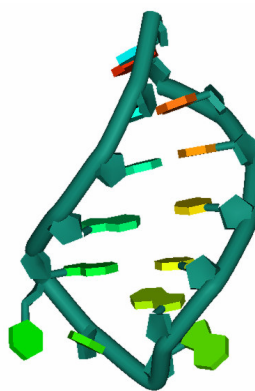
Rys. 13. Suwaki rybozowe w domenie P4-P6 zmutowanego intronu grupy I z *Tetrahymena*, PDB ID: 1L8V. Kolorem zielonym zaznaczyłem wiązania wodorowe.

Motyw spinki do włosów

Motyw spinki do włosów jest najczęściej występującym motywem strukturalnym. Powstaje on na skutek tworzenia par zasad w pojedynczej nici RNA. Część reszt nukleotydowych tworzy pętlę terminalną (ang. *terminal loop*), podczas gdy pozostałe tworzą trzon pętli (ang. *stem*) (rys. 14).



Rys. 14. Struktura przestrzenna motywu spinki do włosów, PDB ID: 2KVN

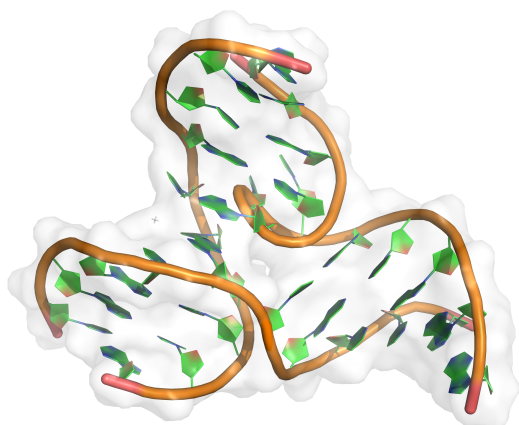


Rys. 15. Struktura pętli czteronukleotydowej AGUU, PDB ID: 1K4B.

Pętla czteronukleotydowa.

Pętla czteronukleotydowa jest to specjalny przypadek motywu spinki do włosów, który warto wyróżnić ze względu na jego stabilność oraz częstość występowania wśród naturalnie występujących motywów strukturalnych (rys. 15). Stanowi on ponad połowę wszystkich struktur spinkowych RNA. Pętle te najczęściej mają sekwencję UNCG, GNRA lub CUYG, gdzie N to dowolny nukleotyd, R - puryna a Y – pirymidyna (29). Najbardziej stabilne są pętle GNRA, gdyż w obrębie pętli G tworzy parę zasad typu *sheared* z A. Pętle czteronukleotydowe typu GNRA specyficznie wiążą się do wielu motywów strukturalnych RNA.

Styki



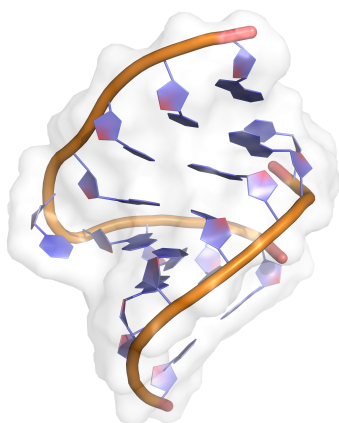
Rys. 16. Struktura przestrzenna trójczłonowego styku, PDB ID: 1EKW

Styk (ang. *junction*) jest tworzony przez trzy lub więcej helis łączących się w jednym miejscu. Styki trójczłonowe w RNA mogą pełnić rolę ryboprzełączników lub rybozymów (30). Styki trójczłonowe najczęściej przyjmują kształt zbliżony do litery Y (rys. 16), który jest stabilizowany przez współosiowe oddziaływania warstwowe (omówione dalej) oraz inne oddziaływania pomiędzy helisami. Przykładem styku trójczłonowego jest

rybozym *hammerhead*. Pełną listę znanych

styków można znaleźć w bazie RNAJunction Database (<http://rnajunction.abcc.ncifcrf.gov/>) (31).

Pętle wewnętrzne i wybrzuszenia

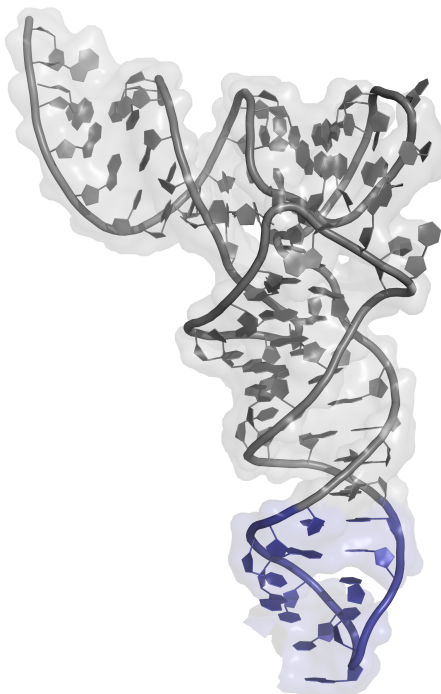


Rys. 17. Struktura duplexu RNA zawierającego jednonukleotydowe wybrzuszenie, PDB ID: 2K41

Pętle wewnętrzne to struktury powstające na skutek braku parowania jednej lub kilku reszt nukleotydowych w strukturze helikalnej. Reszty, które nie tworzą par zasad ulegają wypętlaniu. Liczba niesparowanych reszt może być identyczna lub różna w obu niciach, przez co powstają odpowiednio pętle symetryczne i niesymetryczne. Jeśli jedna lub kilka reszt ulega wypętlaniu tylko w jednej nici, to możemy mówić o wybrzuszeniach (ang. *bulge*). Takie wybrzuszenie przedstawiłem na rysunku 17 (32).

Zwrot U

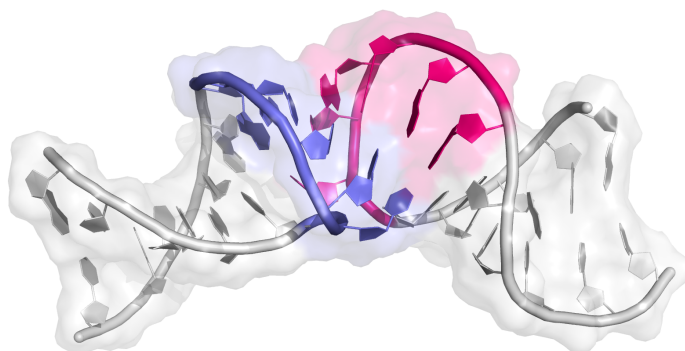
Jest to motyw pętli terminalnej, w której następuje zwrot o 180 stopni. Zwrot U charakteryzuje się specyficznym układem oddziaływań stabilizujących, w które zaangażowane są pierwsza i trzecia reszta (iminowy atom wodoru pierwszej reszty oddziałuje z tlenem grupy fosforanowej trzeciej reszty, a grupa 2'-OH pierwszej reszty z atomem N7 trzeciej). Najbardziej znanym przykładem zwrotu U jest ramię antykodonowe cząsteczek tRNA (rys. 18) (33).



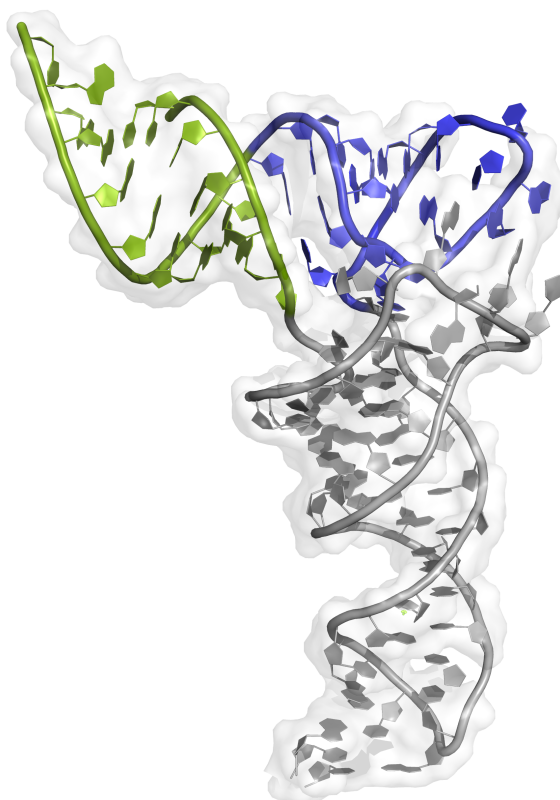
Rys. 18. Cząsteczka tRNA z zaznaczoną pętlą zawierającą zwrot U (kolor niebieski), PDB ID: 2TRA

Motyw oddziałujących pętli

Motyw oddziałujących pętli (ang. *kissing loops*) powstaje, gdy reszty nukleotydowe należące do różnych pętli terminalnych tworzą ze sobą pary zasad (rys. 19). Pierwszy raz struktura ta została opisana dla pętli TAR wirusa HIV (34). Struktura ta stabilizowana jest przez jony magnezu. Listę motywów oddziałujących pętli, podobnie jak listę znanych styków można odnaleźć w bazie RNAJunction.



Rys. 19. Motyw oddziałujących pętli, PDB ID: 1KIS. Reszty nukleotydowe zaangażowane w oddziaływanie zaznaczyłem kolorem niebieskim i różowym.



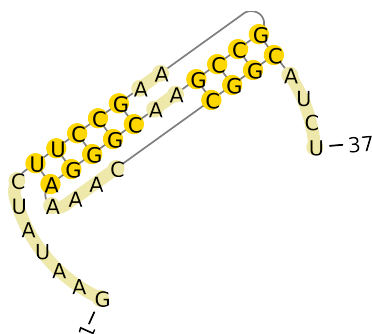
Rys. 20. Częsteczka tRNA z zaznaczonymi ramieniem akceptorowym (kolor zielony) oraz ramieniem T (kolor niebieski) biorącymi udział we współosiowym oddziaływaniu warstwowym, PDB ID: 2TRA

Współosiowe oddziaływania warstwowe

Struktury helikalne mogą się wzajemnie stabilizować poprzez współosiowe oddziaływania warstwowe. Do sytuacji takiej dochodzi, gdy dwie helisy są ułożone jedna nad drugą tak, że terminalne pary zasad obu helis mogą ze sobą oddziaływać warstwowo. Oddziaływanie takie można zaobserwować na przykład w strukturze rybozemu typu hammerhead oraz cząsteczkach tRNA (rys. 20).

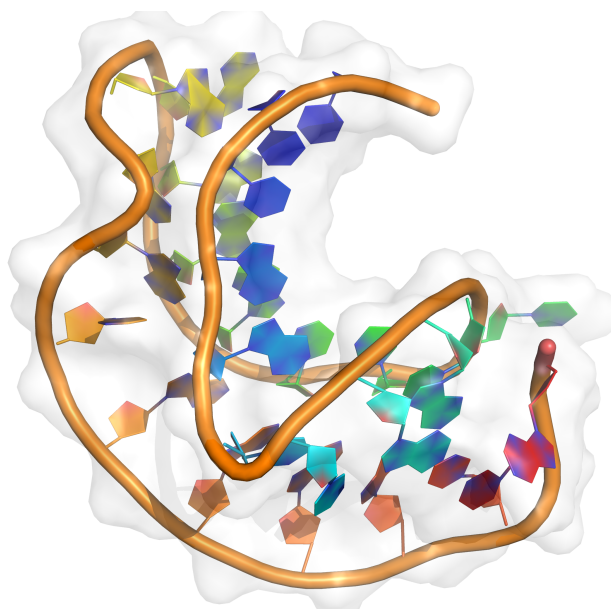
Pseudowęzły

Prócz prostych pętli, cząsteczki RNA mogą tworzyć bardziej skomplikowane struktury, w których dwie nici oddziałują w ten sposób, że tworzona jest pętla, która oddziałuje z fragmentem jednoniciowym nie należącym do tej pętli (rys. 21, 22). Strukturę taką można



Rys. 21. Struktura drugorzędowa pseudowęzła - aptamer przeciw odwrotnej transkryptazie wirusa HIV-1 (36)

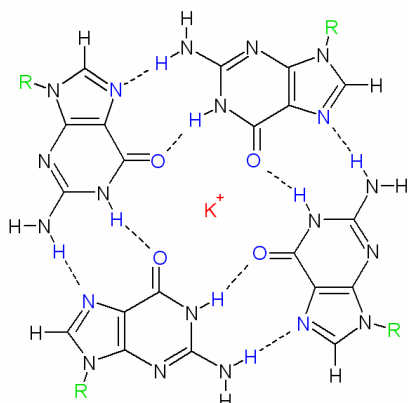
„rozplątać” bez rozrywania wiązań kowalencyjnych. Ponieważ obecnie w strukturach RNA znajdowanych jest coraz więcej pseudowęzłów, są one katalogowane w bazie PseudoBase++ (35) (<http://pseudobaseplusplus.utep.edu/>). Do wizualizacji struktury drugorzędowej można wykorzystać oprogramowanie Pseudoviewer (<http://wilab.inha.ac.kr/pseudoviewer2/>).



Rys. 22. Struktura przestrzenna pseudowęzła odpowiadającego za regulację przesunięcia ramki odczytu z wirusa *Pea Enation Mosaic Virus Type I*; PDB ID: 1KPZ

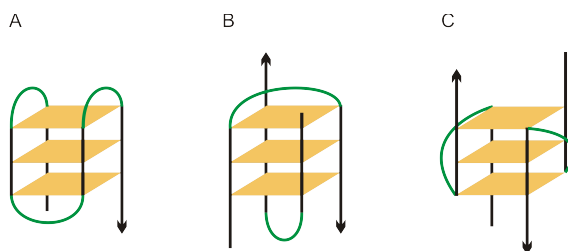
G-kwadrupleksy i trypleksy

G-kwadrupleksy są cząsteczkami o dużej zawartości reszt guanozyny, które przyjmują charakterystyczną formę zawierającą tetrady guanozynowe (rys. 23). G-tetrady te stabilizowane



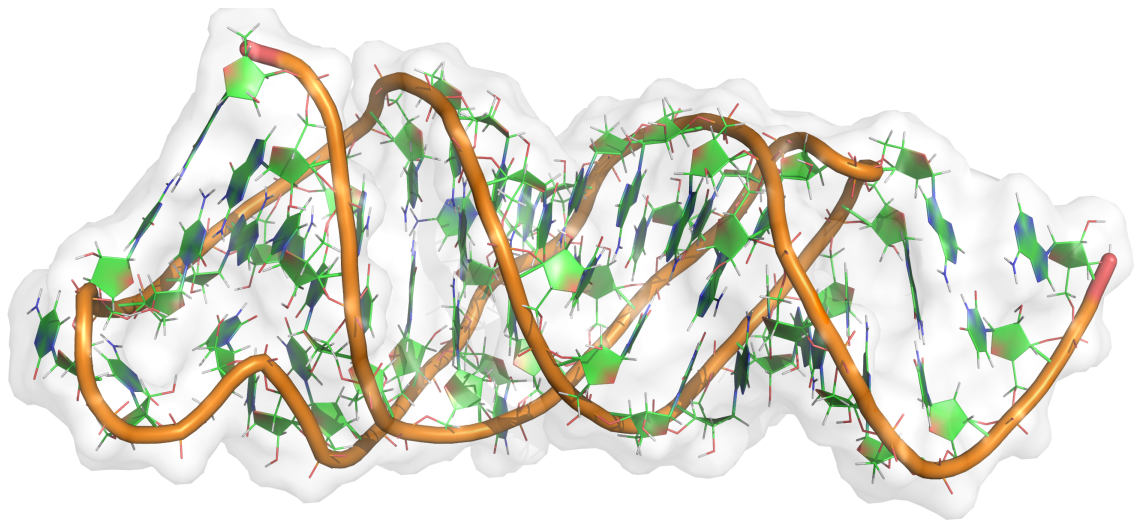
Rys. 23. Struktura tetrady guanozynowej, na czerwono zaznaczony jon potasu, na niebiesko zaznaczone atomy tworzące biorące udział w tworzeniu wiązań wodorowych: tlen grupy karbonylowej O6 i atom wodoru H1 oraz atom azotu N7 i atom wodoru H2.

są przez kationy jednowartościowe zlokalizowane w kanale utworzonym w centrum G-kwadrupleksu oraz przez oddziaływania warstwowe z sąsiadującymi G-tetradami (37). Istotną rolę w stabilizacji kwadrupleksów pełni też kierunkowość wiązań wodorowych w G-tetradach, biorących udział w oddziaływaniach warstwowych. G-kwadrupleksy mogą mieć różną topologię, wynikającą z odmiennego ułożenia łańcuchów względem siebie. Wyróżniamy kwadrupleksy równoległe, antyrównoległe oraz hybrydowe. Reszty nukleotydowe niezaangażowane w tworzenie G-tetrad formują łączące je pętle (rys. 24) (37).



Rys. 24. Występujące w strukturach G-kwadrupleksów pętle: A- boczne, B – przekątne, C – zewnętrzne (37).

Odkryto również struktury, które zbudowane są z trzech nici kwasu nukleinowego. Przykładem może być struktura pseudowęzła występującego w RNA wchodzącym w skład telomerazy (38) przedstawiona na rysunku 25. Trzecia nić asocjuje do dużej bruzdy duplexu tworzonego przez pozostałe nici.



Rys. 25. Struktura przestrzenna potrójnej helisy znalezionej w RNA telomerazy z drożdża *Kluyveromyces lactis*, PDB ID: 2M8K

II.3 Spektroskopia NMR kwasów nukleinowych

Badanie struktury przestrzennej biomolekuł pozwala na lepsze zrozumienie mechanizmów reakcji oraz interakcji cząsteczek ze sobą. Badania te pozwalają z jednej strony poszerzyć wiedzę dotyczącą funkcjonowania organizmów, a z drugiej umożliwiają projektowanie leków nowych generacji wykorzystujących poznane właściwości makrocząsteczek.

Do badań struktury biomolekuł służy szereg metod badawczych.

Metody te można podzielić na:

- metody wykorzystujące różnice mas cząsteczkowych
- metody określania kształtu makrocząsteczek ze średnią rozdzielczością
- metody określania struktury z rozdzielczością atomową
- metody obliczeniowe

Do metod wykorzystujących różnicę mas cząsteczkowych zalicza się przede wszystkim spektrometrię masową oraz sekwencjonowanie, dzięki którym można poznać masę i strukturę pierwszorzędową makromolekuły. Rozwijanych jest również wiele metod wykorzystujących modyfikowanie lub rozcinanie poszczególnych reszt, dzięki czemu uzyskuje się informacje o strukturze drugo- lub trzeciorzędowej. Dla RNA do takich metod należą metody wykorzystujące cięcia enzymatyczne (39), z wykorzystaniem jonów ołowiu (40) czy metoda SHAPE (ang. *selective 2'-hydroxyl acylation analyzed by primer extension*) (41).

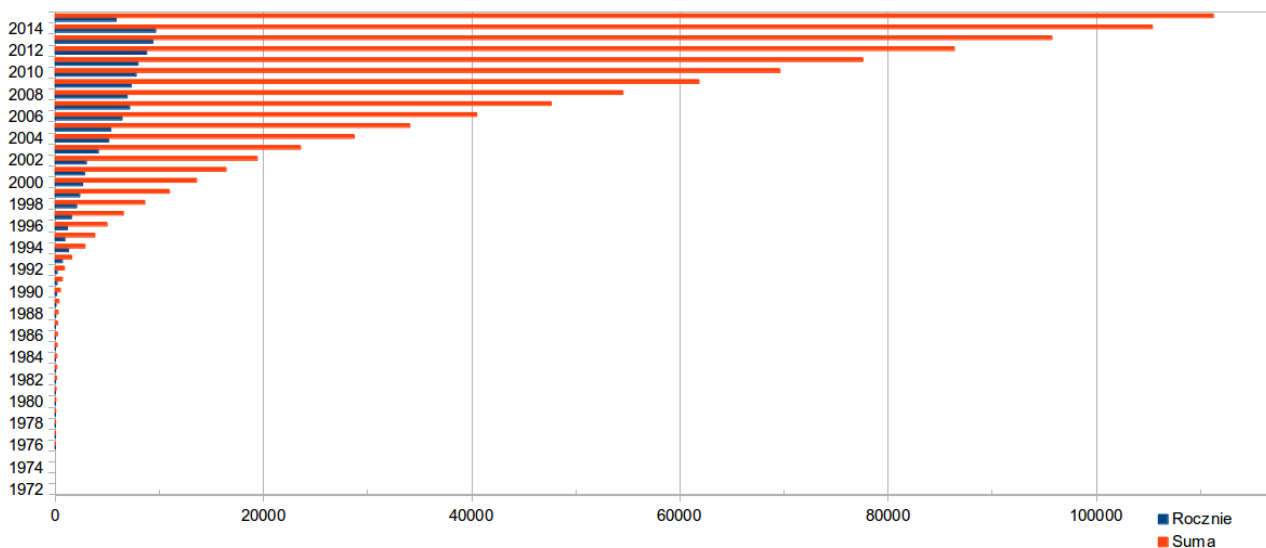
Do metod wyznaczania struktur przestrzennych ze średnią rozdzielczością służą mikroskopia sił atomowych (AFM) oraz mikroskopia elektronowa (cryoEM). Pierwsza z tych metod wykorzystuje oddziaływania międzyatomowe między mikroskopem a badaną powierzchnią (42), natomiast druga polega na obserwacji elektronów przechodzących przez próbkę w warunkach kriogenicznych (43). Pewne informacje o konfiguracji cząsteczek, np. dotyczące skrętności helisy można uzyskać dzięki dichroizmowi kołowemu - CD (ang. *Circular Dichroism*) (44).

Metody pozwalające na rozwiązywanie struktur z dokładnością atomową to krystalografia oraz spektroskopia NMR. Pierwsza z tych metod wykorzystuje ugięcie promieni rentgenowskich na strukturze krystalicznej zbudowanej z cząsteczek badanego związku. Obecnie metodą tą można uzyskiwać struktury o rozdzielczości poniżej 1 Å (45) oraz struktury dużych makromolekuł oraz kompleksów (46). Główną wadą tej metody jest konieczność uzyskania kryształu, niemożliwość uzyskania informacji o regionach mało ustrukturalizowanych oraz możliwość wystąpienia lokalnych zmian struktury powstałych w wyniku tworzenia kryształu. Spektroskopia NMR pozwala na badanie struktur przestrzennych w roztworze, czyli warunkach bardziej zbliżonych do fizjologicznych (47). Pozwala również na badania dynamiki cząsteczek oraz regionów nieuporządkowanych. Główną wadą tej metody są problemy w rozwiązywaniu

struktur związków o dużych masach cząsteczkowych. Największą cząsteczką RNA rozwiązana za pomocą spektroskopii NMR jest fragment RNA HIV-1 o długości 155 nukleotydów, zawierający sygnał pakowania wirusa (48).

Ze względu na trudności związane z uzyskaniem danych eksperymentalnych, często stosuje się metody modelowania molekularnego. Modelowanie molekularne wykorzystuje dwa przeciwne podejścia – modelowanie homologiczne, w którym do uzyskania modelu struktury przestrzennej wykorzystuje się znane struktury albo modelowanie „ab initio” - opierające się o dynamikę molekularną. Metody te pozwalają na badania zarówno na poziomie atomowym (49) jak i „gruboziarnistym” (ang. *Coarse-grained*), gdzie pojedyncza reszta jest reprezentowana tylko przez jeden lub kilka węzłów (50). W naszym zespole (Zespół Struktury i Funkcji RNA) rozwijana jest nowa metoda przewidywania struktury przestrzennej cząsteczek RNA zaimplementowana w oprogramowaniu RNAComposer (<http://rnacomposer.cs.put.poznan.pl/>) (51, 52). W metodzie tej wykorzystywane są krótkie fragmenty znanych struktur przestrzennych cząsteczek RNA, które następnie są łączone w większą strukturę.

Struktury przestrzenne poznanych cząsteczek są zdeponowane w bazie PDB (ang. *Protein Data Bank*) (<http://www.rcsb.org>) (53). Na dzień 17 sierpnia 2015 roku, liczba struktur zdeponowanych w bazie PDB wyniosła 111241 (rys. 26). Większość z nich stanowią struktury białek. Znacznie mniejszą grupę stanowią cząsteczki kwasów nukleinowych, w tym 1581 struktur cząsteczek DNA, 1103 struktur cząsteczek RNA, oraz 5050 kompleksów białko - kwas nukleinowy. Jednocześnie około 42% struktur cząsteczek RNA (nie uwzględniając kompleksów) jest rozwiązanych za pomocą spektroskopii NMR. Dla porównania, w przypadku białek jest to około 9%.



Rys. 26. Liczba struktur RNA zdeponowanych w bazie PDB na przestrzeni kolejnych lat

II.3.1 Zjawisko rezonansu

Jądro atomowe obdarzone jest spinem jądrowym. Jest to wektor momentu pędu wynikający z ruchu obrotowego jądra. Wartość spinu zależy od spinowej liczby kwantowej i przyjmuje wartości będące wielokrotnością $\frac{1}{2}$. Każde jądro posiadające niezerowy spin jądrowy posiada moment magnetyczny $\mu = \gamma \hbar I$, gdzie $\hbar$ to kreślona stała Plancka, I – spin jądrowy a γ to współczynnik giromagnetyczny. Wartość współczynnika giromagnetycznego zależy od rodzaju jądra. Zjawisko rezonansu można obserwować dla jąder posiadających moment magnetyczny (czyli jąder posiadających niezerowy spin jądrowy). Ponieważ jądra posiadające spin większy niż $\frac{1}{2}$ posiadają również moment kwadrupolowy, sygnały dla nich są relatywnie szerokie. Z tego powodu do badań nad cząsteczkami RNA wykorzystuje się między innymi jądra ^1H , ^{13}C , ^{15}N , ^{31}P i ^{19}F posiadające spin $\frac{1}{2}$.

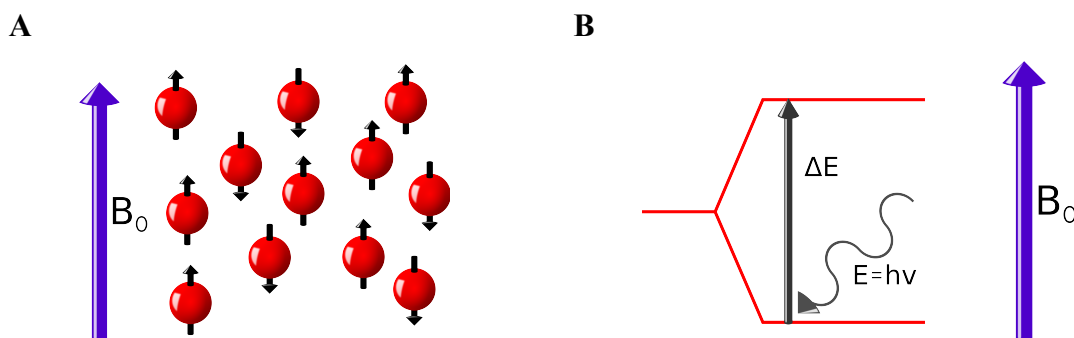
Tabela 1. Właściwości magnetyczne wybranych jąder atomowych (54)

Atom	Izotop	Spinowa liczba kwantowa	Moment magnetyczny	Współczynnik giromagnetyczny	Częstotliwość rezonansowa dla pola 1T	Czułość detekcji	Naturalna zawartość
H	^1H	0.5	2.79277	2.675	42.577	1.0	99.98
	^2H	1	0.85735	0.411	6.536	0.009	0.0156
B	^{10}B	3	1.8007	0.288	4.575	0.02	18.83
	^{11}B	1.5	2.6880	0.858	13.660	0.165	81.17
C	^{13}C	0.5	0.70216	0.673	10.705	0.016	1.108
N	^{14}N	1	0.40369	0.193	3.076	0.001	99.635
	^{15}N	0.5	-0.28298	-0.271	4.315	0.001	0.365
O	^{17}O	2.5	-1.8930	-0.363	5.772	0.029	0.037
F	^{19}F	0.5	2.6273	2.517	40.055	0.834	100.0
Si	^{29}Si	0.5	-0.55492	-0.531	8.460	0.079	4.70
P	^{31}P	0.5	1.1316	1.083	17.235	0.066	100.0

Umieszczenie cząsteczki w silnym polu magnetycznym B_0 spowoduje powstanie dyskretnych stanów energetycznych dla atomów wchodzących w jej skład. Wektor precesji spinu jądra dla atomu o połówkowym spinie jądrowym jest wtedy zorientowany równolegle lub antyrównolegle do kierunku pola magnetycznego (rys. 27 A). Różnica energii pomiędzy tymi stanami zależy od wartości indukcji zewnętrznego pola magnetycznego (B_0) oraz rodzaju jądra. Różnica ta wpływa na obsadzenie obu stanów energetycznych, co z kolei przekłada się na czułość eksperymentów. Ta zależność wywołała potrzebę budowania coraz silniejszych magnesów, by poprawić stosunek sygnału do szumu.

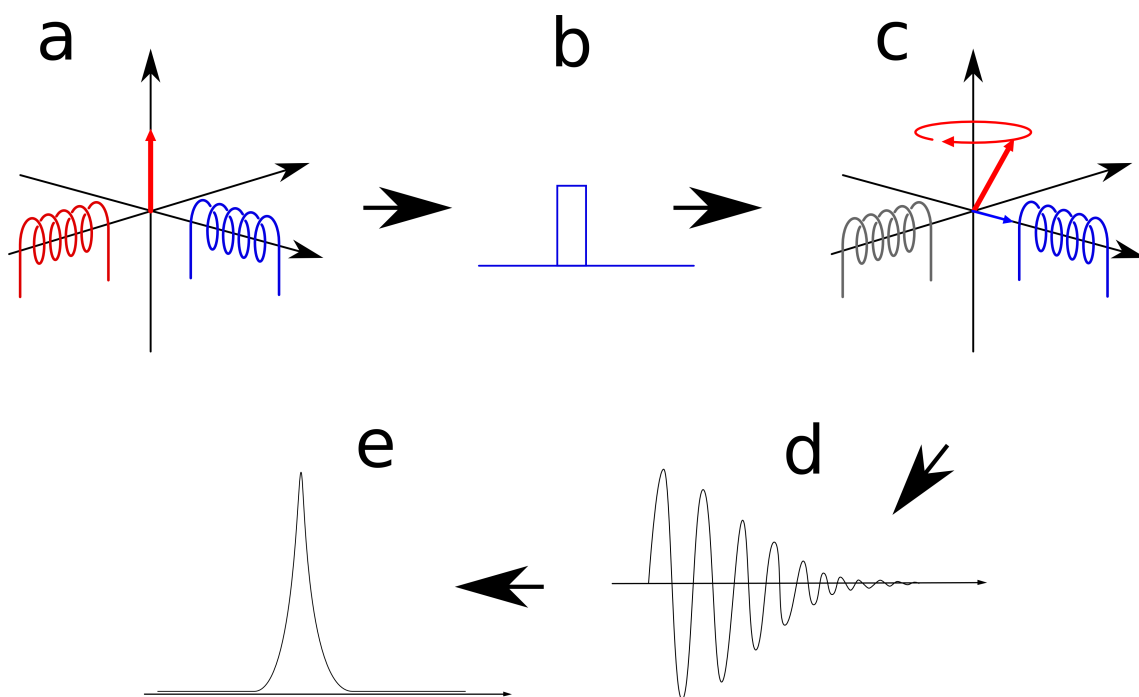
Działając na próbkę polem elektromagnetycznym o odpowiednio dobranej częstotliwości można stymulować przejście z niższego stanu energetycznego na wyższy. Spektroskopia NMR polega na obserwacji absorpcji energii podczas przejścia pomiędzy stanami energetycznymi

(rys. 27 B). Zmieniając częstotliwość można w ten sposób otrzymać widmo NMR. Metoda ta nosi nazwę metody fali ciągłej (ang. *continuous wave*) – CW NMR. Można stosować również zmianę pola magnetycznego przy stałej częstotliwości naświetlania (ang. *field sweep*), jednak już w latach siedemdziesiątych ubiegłego wieku zaniechano tej metody.



Rys. 27. A) Przedstawienie orientacji wektora precesji spinów jądrowych dla atomów o połówkowym spinie jądrowym w silnym polu magnetycznym, B) Przejście pomiędzy stanami o różnej energii po spełnieniu warunku rezonansu. B_0 – wektor indukcji pola magnetycznego.

Aby uzyskać widmo wyżej opisaną metodą potrzebna jest znaczna ilość czasu, gdyż należy dokonywać pomiaru dla wielu częstotliwości po kolei. Obecnie stosuje się metodę wykorzystującą transformację Fouriera (FT NMR), w której za pomocą nawet pojedynczego pomiaru można uzyskać widmo. Metoda ta polega na naświetlaniu próbki pojedynczym silnym sygnałem o częstości zbliżonej do częstości Larmora, czyli częstości precesji spinu jądrowego (54). Sygnał ten trwa najczęściej kilkadziesiąt mikrosekund i wychyla magnetyzację o określony kąt – najczęściej 90° ($\pi/2$) od położenia równowagi w kierunku osi y. Wektor magnetyzacji zaczyna precesję wokół osi z i z czasem powraca do położenia równowagi (rys. 28). Rzut precesji na oś y jest rejestrowany przez odbiornik i zapisany w postaci sygnału precesji swobodnej tzw. FID (ang. *Free Induction Decay*). Wykonanie operacji matematycznej, transformacji Fouriera na tych danych powoduje otrzymanie takiego samego widma jak w przypadku spektroskopii NMR fali ciągłej. Przez lata opracowano wiele różnych rodzajów widm, które uzyskuje się poprzez dodawanie dodatkowych impulsów o określonej sile, częstotliwości, kształcie oraz odstępie czasowym od pozostałych impulsów. Przykładem prostej techniki może być technika echa spinowego (54). Technikę tę stosuje się w celu wyeliminowania efektu niejednorodności pola, które występuje w różnych częściach próbki, co przekłada się na gorszy kształt sygnałów. Jeśli rozłożymy wektor magnetyzacji na sumę elementarnych momentów magnetycznych, to po przyłożeniu impulsu $\pi/2$ możemy zaobserwować rozwijanie się magnetyzacji w wachlarz składowych. Jeśli po czasie τ przyłożymy impuls π to poszczególne składowe będą dążyły do zbieżności, którą osiągną po czasie 2τ . W tym momencie w cewce odbiorczej (opisanej w rozdziale „Budowa spektrometru NMR”) pojawi się zaindukowane napięcie – echo spinowe.



Rys. 28. Uzyskiwanie sygnału NMR w przypadku spektroskopii wykorzystującej transformację Fouriera; a – układ w stanie równowagi; b – impuls; c – wychylenie i precesja magnetyzacji; d – sygnał indukcji swobodnej (FID); e – widmo NMR.

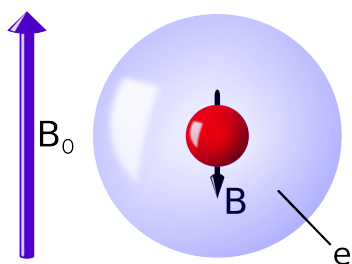
II.3.2 Przesunięcie chemiczne

Absorpcję energii podczas zjawiska rezonansu rejestruje się w postaci widma. Na widmie tym widoczne są sygnały absorpcyjne dla częstotliwości dla których zaszło zjawisko rezonansu. Ze względu na to, że częstotliwość zależy od przyłożonego wektora indukcji pola magnetycznego do określania położenia sygnału na widmie stosuje się wielkość bezwymiarową - przesunięcie chemiczne (δ). Jednostką przesunięcia chemicznego jest 10^{-6} lub inaczej ppm (ang. *parts per milion*). Wartość przesunięcia chemicznego wynosi:

$$\delta = (v_{\text{substance}} - v_{\text{reference}})/v_0$$

gdzie $v_{\text{substance}}$ to częstotliwość rezonansowa dla jądra badanej próbki, $v_{\text{reference}}$ to częstotliwość rezonansowa wzorca, a v_0 to częstotliwość podstawowa spektrometru.

Przesunięcie chemiczne danego sygnału mierzy się względem położenia sygnału wzorca. W tym celu najczęściej wykorzystywany jest TMS (tetrametylosilan). Wzorzec może być zewnętrzny (wprowadzany do próbki w węższej próbce) oraz wewnętrzny – dodawany do badanego preparatu. Wartość przesunięcia chemicznego dla danego typu jądra może się różnić.



Rys. 29. Przesłanianie jądra przez chmurę elektronową (e) B_0 – zewnętrzne pole magnetyczne, B – pole indukowane

Wynika to z tego, że ciągle poruszające się elektrony oddziałują z zewnętrznym polem magnetycznym B_0 , przez co następuje zmiana ich ruchu, co z kolei przekłada się na stworzenie indukowanego pola magnetycznego o zwrocie przeciwnym do B_0 (rys. 29). Pole to powoduje ekranowanie lub odsłanianie jąder atomowych. Największy wpływ na ekranowanie danego

jądra mają elektrony znajdujące się w jego bezpośrednim sąsiedztwie, w związku z czym głównym czynnikiem wpływającym na wartość przesunięcia chemicznego dla danego jądra jest jego otoczenie chemiczne (55).

Ze względu na podobieństwo w budowie chemicznej – linie rezonansowe dla jąder znajdujących się w określonej grupie chemicznej zazwyczaj mają podobne wartości przesunięć chemicznych. Linie rezonansowe obserwowane na widmach NMR powstają w wyniku uśrednienia sygnałów pochodzących od jąder cząsteczek znajdujących się zarówno w różnej orientacji przestrzennej względem pola magnetycznego jak i posiadających różną konfigurację przestrzenną wynikającą z obrotu grup atomów względem wiązań chemicznych. Różnice w wartościach lokalnego pola magnetycznego wynikające z efektów ekranowania stanowią najczęściej niewielki procent pola B_0 . W związku z tym rejestracja dobrej jakości widma wymaga niezwykle czulej aparatury pomiarowej.

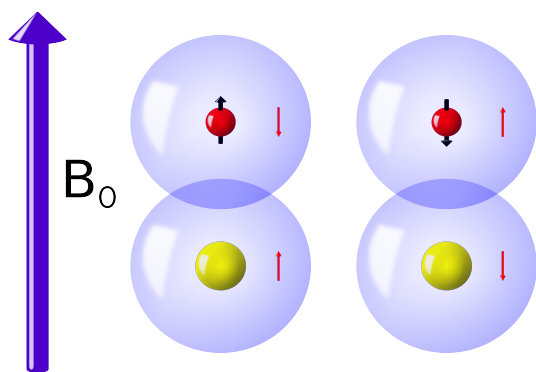
II.3.3 Objętości i intensywności sygnałów

Zarówno intensywność i objętość sygnału są proporcjonalne do ilości jąder, od których pochodzi ten sygnał. Amplituda sygnału może się różnić ze względu na różną szerokość i multipletowość sygnałów. W związku z tym bardziej użyteczna wydaje się objętość danego sygnału. Porównując objętości sygnałów można między innymi rozpoznać które sygnały należą do poszczególnych związków w mieszaninie lub znaleźć grupy chemiczne posiadające kilka identycznych magnetycznie jąder.

II.3.4 Sprzężenie spinowo–spinowe

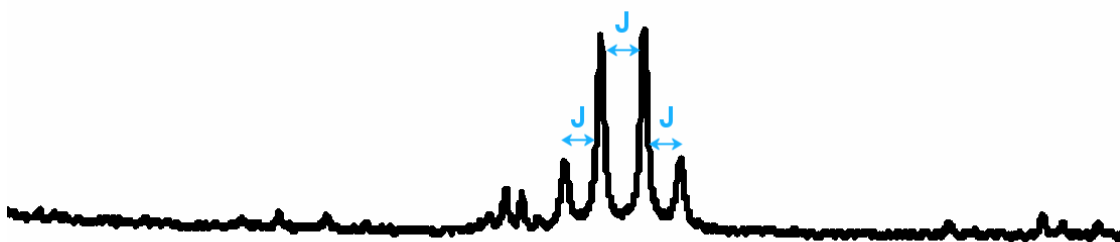
Spin jądrowy powoduje słabą polaryzację magnetyczną elektronów otaczających dane jądro. Polaryzacja ta jest przekazywana dalej poprzez wiązania chemiczne (rys. 30). W ten sposób pole magnetyczne wokół drugiego jądra jest wzmacniane lub osłabiane. Zjawisko to nazywamy sprzężeniem spinowo–spinowym. Jeśli rozpatrujemy przypadek jąder o spinie połówkowym, to dla pól magnetycznych wykorzystywanych w spektroskopii magnetycznego rezonansu jądrowego populacja spinów zorientowanych równolegle lub antyrównolegle jest prawie identyczna. W związku z tym osłabienie lub wzmocnienie pola magnetycznego jest tak

samo prawdopodobne. Prowadzi to do powstania dwóch linii rezonansowych dla drugiego jądra (rys. 31). Ponieważ energia tych oddziaływań zależy od orientacji molekuł a nie od

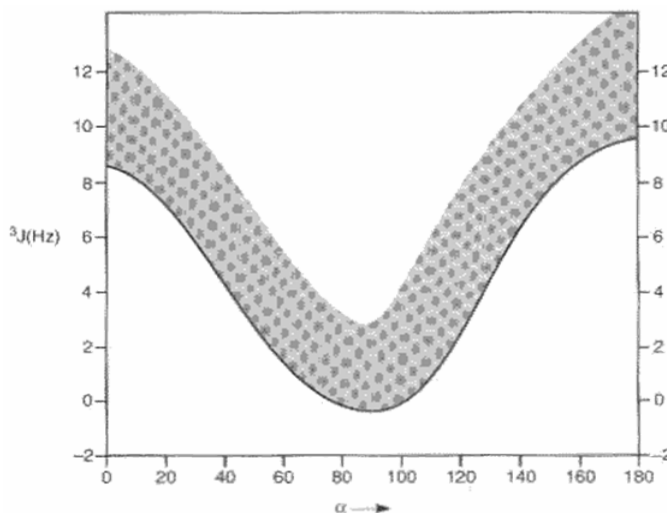


Rys. 30. Transfer polaryzacji powłoki elektronowej (czerwona strzałka) wywołanej przez spin jądrowy (czarna strzałka) przez wiązanie chemiczne.

indukcji pola, linie te są oddzielone od siebie o stałą, niezależną od natężenia pola zewnętrznego wartość. Z tego powodu wartość określającą to oddziaływanie – stałą sprzężenia skalarnego podaje się w jednostkach częstotliwości (Hz). W przypadku sprzężenia przez jedno wiązanie (1J) mówimy o sprzężeniu jąder bezpośrednio ze sobą związanych, przez dwa wiązania (2J) – sprzężeniu geminalnym a przez trzy (3J) – sprzężeniu wicynalnym.



Rys. 31. Fragment widma jednowymiarowego ^1H NMR, na którym zaznaczono stałą sprzężenia skalarnego.



Rys. 32. Zależność wartości stałej sprzężenia w zależności od wartości kąta torsyjnego (24).

Dla sprzężeń wicynalnych istnieje zależność pomiędzy kątem torsyjnym ϕ tworzoną przez te wiązania a wartością stałej sprzężenia. Zależność ta jest opisana równaniem Karplusa: $^3J=A+B*\cos\phi+C*\cos2\phi$. Wartości A, B i C są różne dla różnych grup związków i wyznacza się je doświadczalnie. Dla kątów torsyjnych bliskich 0° lub 180° wartości stałej sprzężenia są największe, natomiast dla kątów bliskich 90° stała sprzężenia jest bliska 0 (rys. 32).

II.3.5 Efekty relaksacyjne

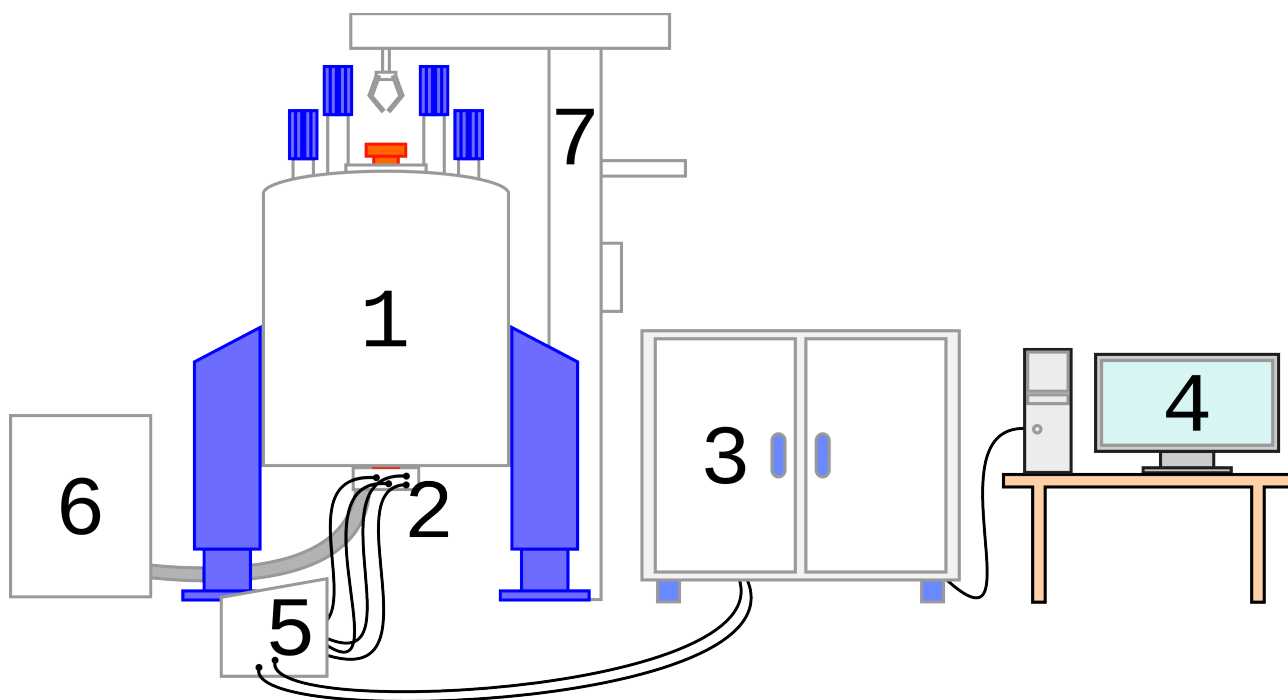
Po wzbudzeniu podczas pomiaru, jądro powraca do stanu równowagi. Dzieje się to dzięki efektom relaksacyjnym. Relaksacja zachodzi ze względu na lokalne zaburzenia, które wpływają na orientację spinu. Wyróżniamy dwa rodzaje relaksacji: relaksacja spin–sieć (podłużna) oraz relaksacja spin–spin (poprzeczna).

Relaksacja spin-sieć dotyczy magnetyzacji zgodnej z kierunkiem pola magnetycznego. Relaksacja podłużna polega na oddaniu energii zaabsorbowanej podczas wzbudzenia, do otoczenia. W dużym stopniu zależy ona od poruszających się lokalnych dipoli magnetycznych. Czasem T_1 określa się czas, który licząc od zakończenia impulsu RF potrzebny jest do ustalenia równowagowego namagnesowania. Wpływa on w znaczącym stopniu na kształt i intensywność sygnału rejestrowanego na widmie.

Relaksacja spin–spin polega na utracie spójności magnetyzacji dla wzbudzonego układu jąder. Dzieje się to poprzez niejednorodności i fluktuacje pola magnetycznego, które umożliwiają oddziaływanie ze sobą poszczególnych spinów. W połączeniu z ruchami cząsteczkowymi prowadzi to do powstawania rozbieżności w fazie.

II.3.6 Budowa spektrometru NMR

Spektrometr magnetycznego rezonansu jądrowego jest zbudowany z magnesu nadprzewodzącego, w środku którego umieszczona jest sonda. Sonda zawiera przede wszystkim cewki nadawczą i odbiorczą, oraz zestaw cewek poprawiających jednorodność pola (ang. *shimming*). Dostępnych jest wiele rodzajów sond, które umożliwiają pomiary na kilku kanałach jednocześnie. W magnecie można zainstalować również tzw. kriosondę, która dzięki schłodzeniu elementów elektronicznych w bezpośrednim otoczeniu pomiaru pozwala na znaczną poprawę stosunku sygnału do szumu na widmach. Sonda połączona jest z konsolą poprzez układ wzmacniaczy oraz opcjonalnych filtrów. Konsola jest nadzorowana przez użytkownika poprzez stację roboczą. Dodatkowo spektrometr może być wyposażony również w stację kontroli temperatury lub robotę do automatycznej zmiany próbek.

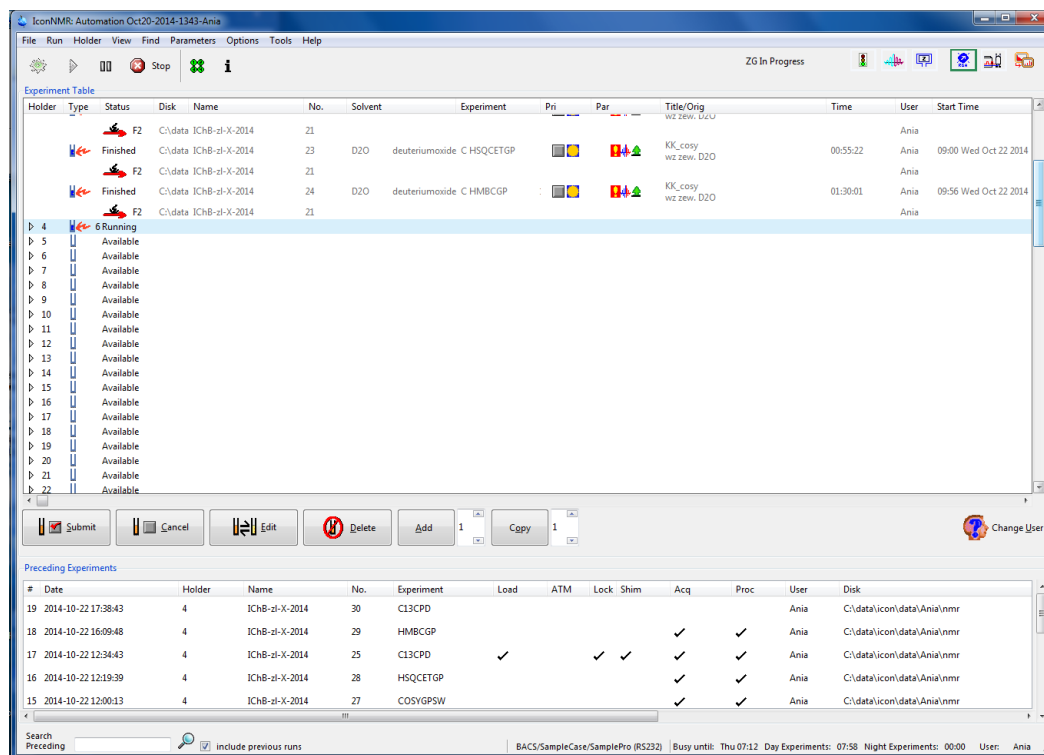


Rys. 33. Schemat nowoczesnego spektrometru NMR z wyposażeniem. 1) Magnes, 2) Sonda, 3) Cyfrowa konsola wysokiej rozdzielczości, 4) Stacja kontrolna, 5) Wzmacniacz sygnału, 6) Jednostka kontroli temperatury, 7) Robot do zmiany próbek.

II.3.7 Rejestracja widm NMR

Aby otrzymać widmo NMR najpierw należy przygotować próbkę do badań. Najlepiej jeśli użyty rozpuszczalnik nie posiada atomów wodoru, dzięki czemu sygnał od rozpuszczalnika nie będzie utrudniał odczytywania innych sygnałów. W tym celu stosuje się rozpuszczalniki deuterowane. Dla cząsteczek RNA wykorzystuje się wodę deuterowaną. W celu maksymalnej eliminacji zawartości H_2O oraz HDO , próbkę można dwukrotnie odparować z wodą deuterowaną. Aby uzyskać informacje o atomach wodoru wymienialnych z wodą stosuje się mieszaninę 90% H_2O i 10% D_2O , natomiast sygnał wody wygasza się za pomocą jednej z kilku technik wygaszania wody, np. z wykorzystaniem techniki rzeźbienia wzbudzenia z gradientami (ang. *water suppression using excitation sculpting with gradients*) (56). W celu uzyskania jak największego stężenia w jak najmniejszej objętości stosuje się probówki 3mm, których budowa pozwala zredukować objętość do około 200 μl .

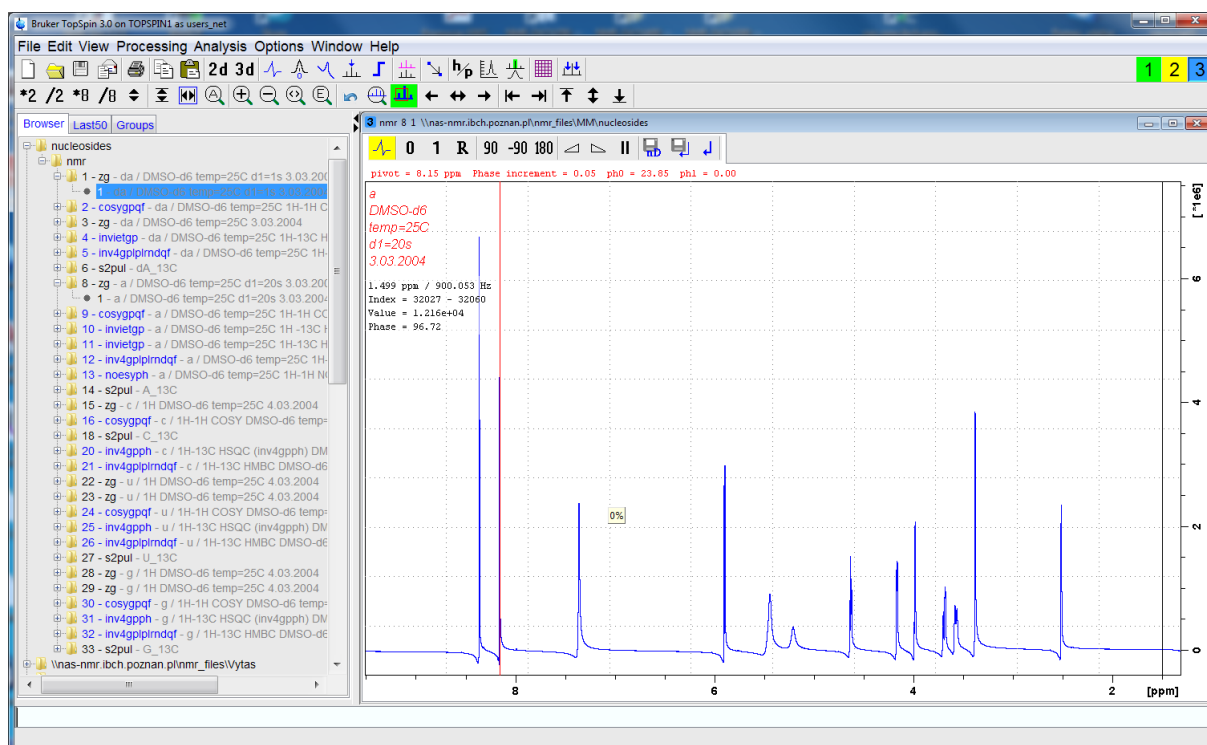
Do wykonywania pomiarów wykorzystuje się specjalne oprogramowanie, które pozwala na zmianę parametrów eksperymentów, zebranie danych i wstępną analizę wyników. Przykładowe oprogramowanie służące do uruchamiania eksperymentów przedstawiłem na rysunku 34.



Rys. 34. Icon NMR – przykładowe oprogramowanie służące do wykonywania eksperymentów

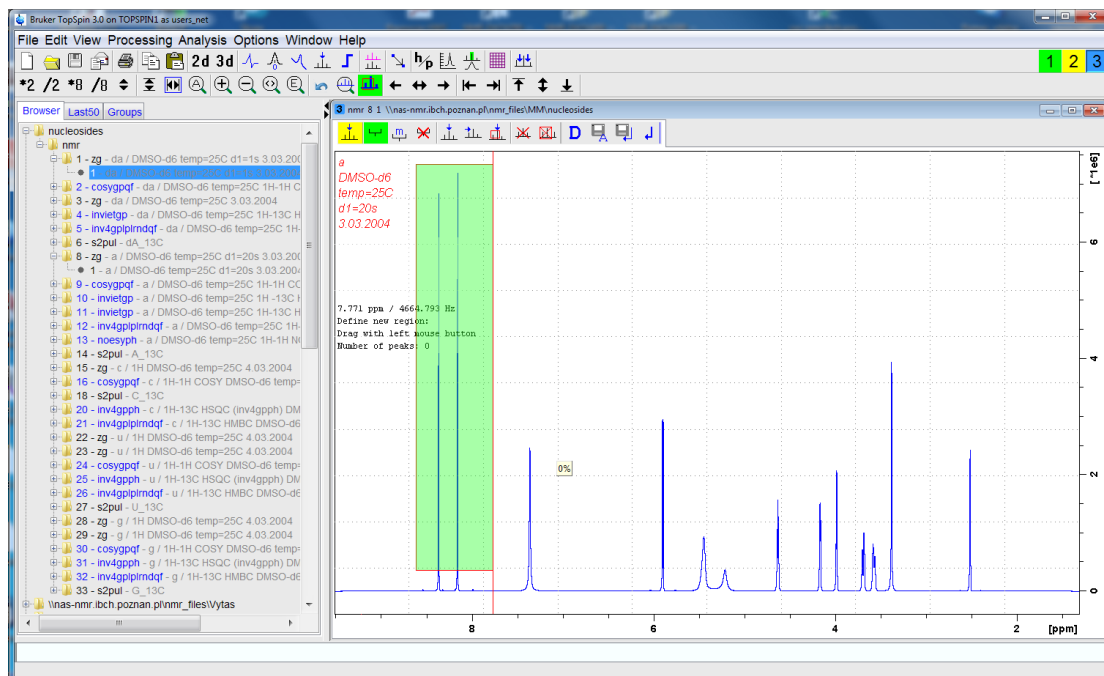
Próbkę podczas pomiarów można rotować, co pozwala na uzyskanie większej jednorodności pola magnetycznego, jednak ze względu na coraz bardziej doskonałe magnesy, wykonywane z dużą dokładnością próbówki oraz bardzo czyste rozpuszczalniki nie zawsze to jest konieczne.

Podczas wykonywania eksperymentu ważne jest, by natężenie pola miało cały czas taką samą wartość. W tym celu wykorzystuje się kompensację zaburzeń zewnętrznych zależnych od czasu, tzw. *lock*, który pozwala na wykrycie i korektę dryftu pola magnetycznego. Jako sygnał *lock'u* wykorzystuje się sygnał od deuterowanego rozpuszczalnika. Następnie wykonuje się optymalizację jednorodności pola (ang. *shimming*) dzięki której można uzyskać lepszą rozdzielczość sygnałów. Kolejnym krokiem jest *tuning* i *matching*, które odpowiadają za dostrojenie cewek do częstotliwości badanego jądra oraz dopasowanie do standardowej impedancji układu. Następnie wyznacza się optymalne wzmocnienie sygnału i rejestruje widmo NMR. Uzyskany tą drogą sygnał FID przede wszystkim poddaje się transformacji Fouriera, a następnie dopasowuje fazę, zaznacza sygnały (ang. *peak picking*), które można następnie zintegrować.

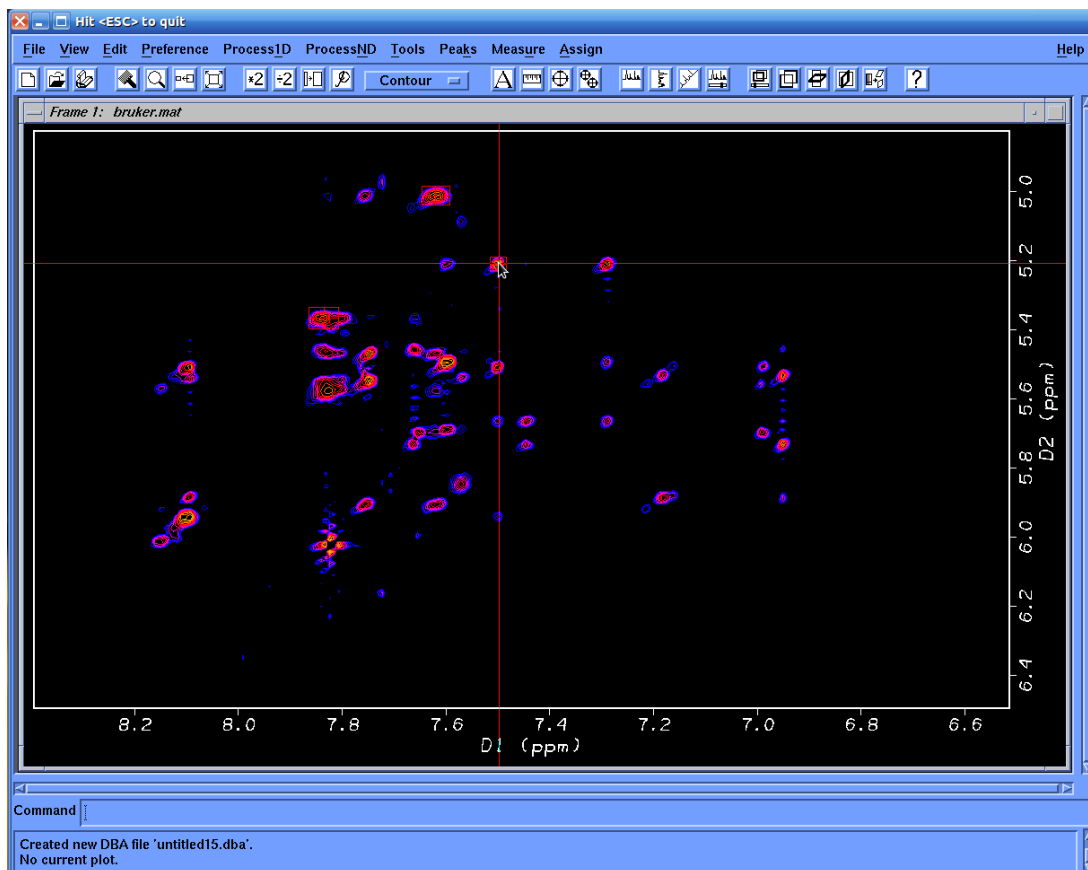


Rys. 35. Ustawianie fazy sygnałów na widmie 1D dla cząsteczki adenozyiny

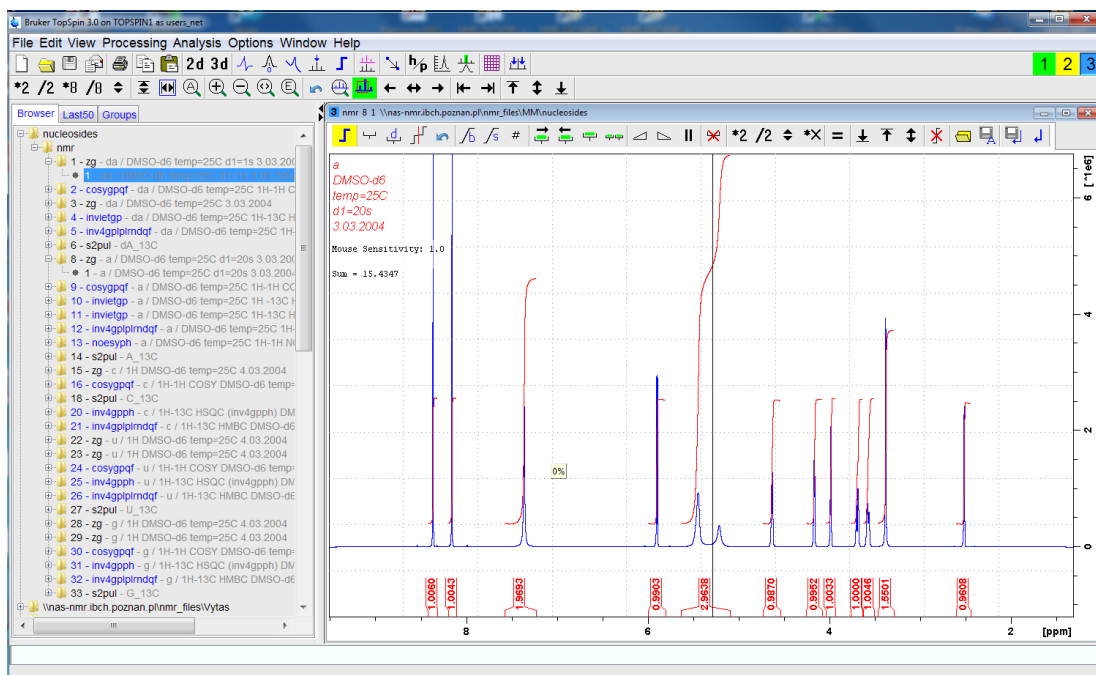
A



B



Rys. 36. Zaznaczanie sygnałów na widmie 1D w programie Topspin (A) oraz 2D w programie Felix (Accelrys) (B)



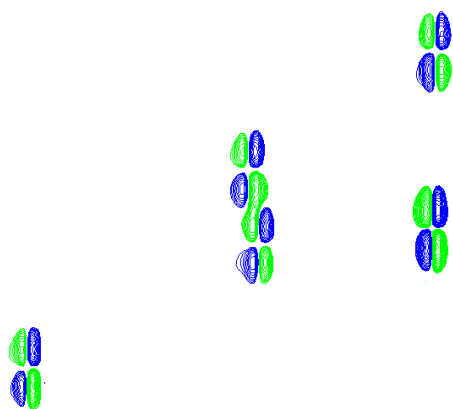
Rys. 37. Integracja sygnałów na widmie 1H dla adenozyiny

II.3.8 Wielowymiarowe widma NMR

Jedną z technik, które miały znaczący wpływ na rozwój i wykorzystanie spektroskopii NMR do badań strukturalnych jest opracowanie techniki widm wielowymiarowych. Pozwalają one na rozdzielnie i korelowanie nakładających się na siebie sygnałów. Widma te wykonuje się stosując sekwencję impulsów rozdzielonych w czasie, z których ostatni jest impulsem $\Pi/2$. Zmieniając odstęp pomiędzy sygnałami można zarejestrować serię widm, które następnie poddaje się transformacji Fouriera, dzięki czemu uzyskuje się widmo 2D lub o większej liczbie wymiarów (57).

Na przestrzeni lat opracowano wiele widm wielowymiarowych, wykorzystujących różne sekwencje w celu uzyskania różnych informacji. Poniżej przedstawiłem kilka wybranych rodzajów eksperymentów, które znajdują szerokie zastosowanie w badaniach struktury przestrzennej RNA.

Widma korelacyjne COSY (ang. *Correlation Spectroscopy*) pozwalają na uzyskanie informacji o sprzężonych ze sobą spinach jądrowych (62, 63). Dla sprzężonych ze sobą jąder na widmie



Rys. 38. Złożona multipletowa struktura sygnałów widocznych na widmie 2D DQF-COSY

widoczny jest sygnał korelacyjny, którego przesunięcie chemiczne w obu wymiarach odpowiada wartościom przesunięcia chemicznego dla oddziałujących jąder. Istnieje wiele odmian techniki COSY, przykładowo znaczną poprawę jakości widma COSY można uzyskać poprzez zastosowanie filtra dwukwantowego, który filtruje przejścia jednokwantowe. Uzyskuje się w ten sposób widmo DQF-COSY (ang. *Double Quantum Filtered*), które znacznie łatwiej zinterpretować (64, 65, 66). Sygnały widoczne na widmach DQF-COSY posiadają złożoną multipletową strukturę (rys. 38).

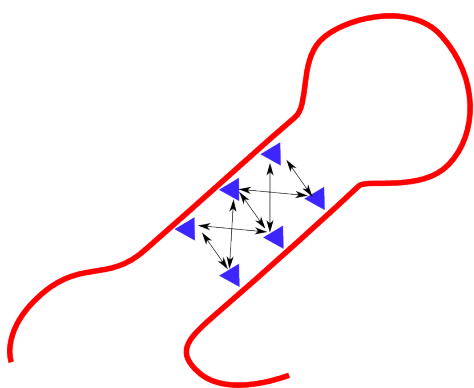
Widma HSQC (ang. *Heteronuclear Single-Quantum Correlation Spectroscopy*) pozwalają na obserwację sprzężeń pomiędzy jądrami różnych pierwiastków chemicznych (67, 68). Odbywa się to przez transfer polaryzacji pomiędzy sprzężonymi jądrami, przy czym wykorzystuje się transfer od jądra ^1H do heterojądra X a następnie z powrotem do jądra ^1H . W efekcie tego procesu, na widmie 2D HSQC widoczne są sygnały korelacyjne pomiędzy jądrami oddzielonymi od siebie pojedynczym wiązaniem atomowym. W przypadku kwasów nukleinowych wykorzystuje się głównie widma ^1H - ^{13}C HSQC oraz ^1H - ^{15}N HSQC.

Widma HMBC (ang. *Heteronuclear Multiple Bond Correlation Spectroscopy*) podobnie jak widma HSQC pozwalają na uzyskanie informacji o sprzężeniach pomiędzy jądrami różnych pierwiastków chemicznych, przy czym na tych widmach widoczne są sygnały korelacyjne

między jądrami oddzielnymi o dwa, trzy a czasem też cztery wiązania atomowe (69).

Widma TOCSY (ang. *Total Correlation SpectroscopY*) pozwalają na identyfikację całych systemów sprzężonych ze sobą spinów (70, 71, 72). Ze względu na zależność odkrytą przez Karplusa, są wykorzystywane głównie w przypadku cząsteczek DNA, w których konformacja 2'-endo (w przeciwieństwie do 3'-endo) przekłada się na dużą wartość sprzężenia spinowo-spinowego dla całego pierścienia cukrowego.

Widma NOESY (ang. *Nuclear Overhauser Enhancement Spectroscopy*) pozwalają na obserwację sprzężeń dipolowych pomiędzy jądrami atomów wodoru (73). W tym celu wykorzystuje się sekwencję impulsów, w której na sygnał detekcji składają się dwa impulsy oddzielone od siebie o tzw. czas mieszania τ_m (ang. *mixing time*). Podczas tego czasu część magnetyzacji jest przekazywana do jąder znajdujących się w bezpośrednim sąsiedztwie w przestrzeni. Efekt ten jest proporcjonalny do odwrotności szóstej potęgi



Rys. 39. Rysunek przedstawiający w schematyczny sposób możliwe sprzężenia dipolowe dla jąder atomów należących do oddalonych od siebie reszt nukleotydowych.

odległości, w związku z czym można go obserwować dla jąder atomów wodoru oddalonych od siebie o mniej niż około 5Å (rys. 39). Poprzez zwiększenie czasu mieszania można uzyskać silniejsze sygnały korelacyjne, jednak w takiej sytuacji wielkość efektu NOE przestaje być wprost odwrotnie proporcjonalna do szóstej potęgi odległości. W związku z czym najczęściej wykonuje się serię widm NOESY z różnymi czasami mieszania, aby z jednej strony ułatwić przypisanie sygnałów a z drugiej uzyskać bardziej dokładne więzy (opisane dalej).

II.3.9 Izotopowe wzbogacanie próbek

Naturalna zawartość izotopów węgla i azotu o pożądanymi właściwościami magnetycznymi (^{13}C , ^{15}N) w cząsteczkach jest niewielka. Ponadto sygnały ^1H , ^{13}C i ^{15}N dla cząsteczek RNA charakteryzują się słabą separacją na widmie. Skutkuje to problemami z przypisaniem sygnałów, zwłaszcza dla większych cząsteczek. Wykorzystanie próbek wzbogaconych izotopowo (^1H , ^{13}C , ^{15}N) pozwala między innymi na zmniejszenie nakładania się sygnałów na siebie oraz wykonywanie wielowymiarowych eksperymentów heterojądrowych.

W celu uzyskania próbek wzbogaconych izotopami o pożądanymi właściwościami magnetycznymi wykorzystuje się dwa główne podejścia (57, 58, 59, 60, 61):

- cząsteczka jest znakowana podczas syntezy chemicznej
- cząsteczka jest znakowana podczas syntezy enzymatycznej

Jeśli chodzi o syntezę chemiczną, to wykorzystuje ona sekwencyjne dołączanie jednostek nukleotydowych rozpoczynając najczęściej od 3' końca. Wszystkie centra nie biorące udziału

w reakcji są blokowane dzięki odpowiednim grupom ochronnym, które są usuwane odpowiednio pomiędzy cyklami i po zakończeniu syntezy. Metoda ta pozwala na uzyskanie w pełni selektywnego znakowania. Ponadto uzyskane cząsteczki bardzo dobrze nadają się do połączenia z innymi cząsteczkami RNA poprzez ligację. Głównym ograniczeniem tej metody jest wielkość cząsteczki, do około 55 nukleotydów, oraz wysoka cena znakowanych izotopowo nukleotydów.

Synteza RNA *in vitro* odbywa się poprzez transkrypcję z wykorzystaniem polimerazy T7 oraz znakowanych nukleotydów. Metoda ta choć pozwala na uzyskanie dłuższych cząsteczek ma pewne wady. Najważniejszą z nich jest niehomogenność 3' i 5' końców oraz pewne wymagania dotyczące sekwencji znajdującej się na 5' końcu. Z tego powodu wymagane dodatkowe oczyszczanie na żelu poliakrylamidowym lub rozdział poprzez jedną z metod chromatografii (74). Można również zastosować sekwencję rybozymów umieszczoną w bezpośrednim sąsiedztwie syntezy cząsteczki, które wycinają z transkrybowanego fragmentu pożądany łańcuch oligonukleotydowy (75). Znakowane nukleotydy najłatwiej uzyskać dzięki hodowli komórek na medium zawierającym $^{15}\text{NH}_4\text{Cl}$ i $^{13}\text{C}_6\text{-Glc}$ lub ^{13}C -metanol. Z tak uzyskanej hodowli izoluje się RNA i trawi do poziomu pojedynczych nukleotydów, które później można wykorzystać do syntezy.

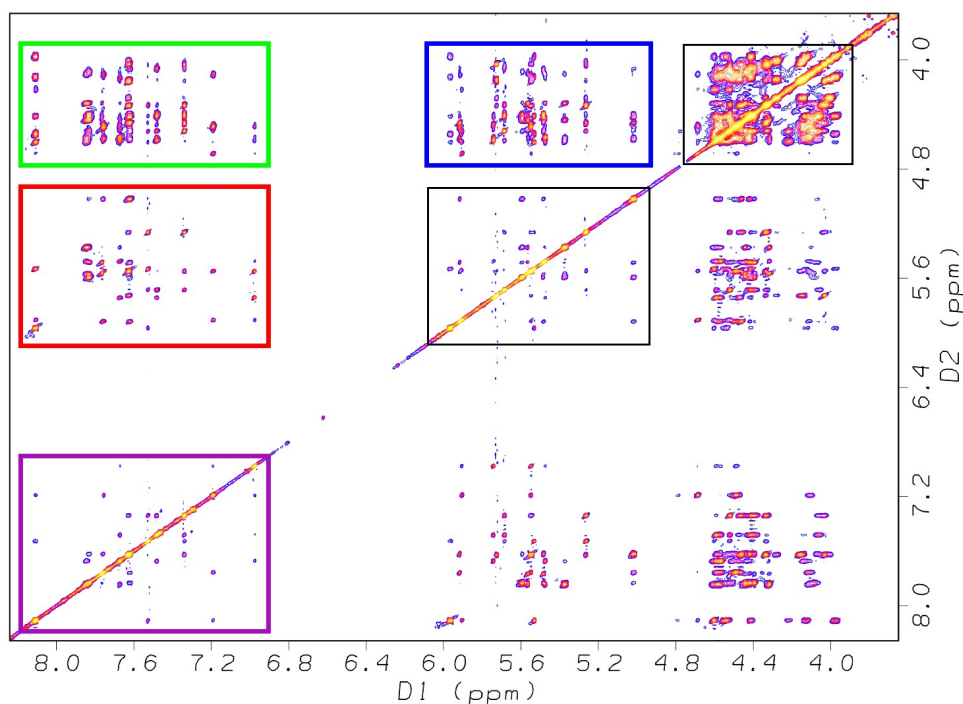
II.4 Rozwiązywanie struktur przestrzennych w oparciu o metody spektroskopii NMR

II.4.1 Charakterystyczne regiony na widmach NMR dla kwasów nukleinowych

Kwasy nukleinowe zbudowane są z reszt nukleotydowych o podobnej budowie. Powoduje to grupowanie się linii rezonansowych w określonych zakresach przesunięć chemicznych, przedstawionych w tabeli numer 2. Dzięki temu na widmach wielowymiarowych możemy obserwować charakterystyczne regiony (rys. 40). Regiony te można analizować w dużym stopniu oddzielnie, co ułatwia proces przypisywania sygnałów.

Tabela 2. Zakresy spektralne przesunięć chemicznych charakterystyczne dla poszczególnych atomów w resztach rybonukleotydowych.

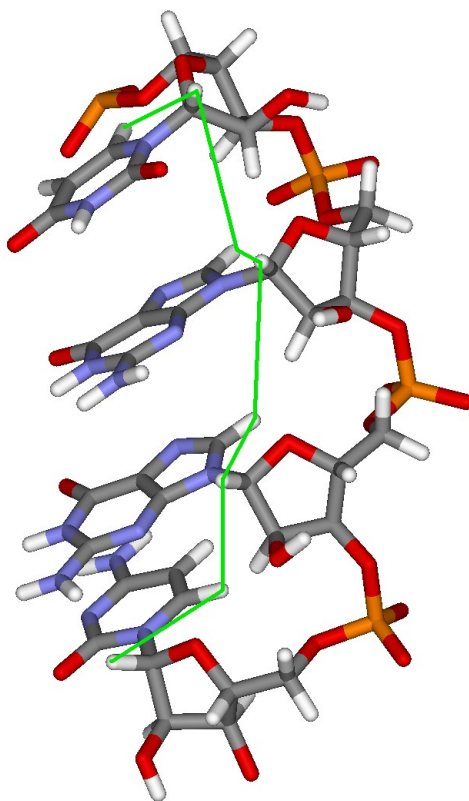
Zakres	Protony
6.5 ppm – 8.5 ppm	H2 (adenozyny), H6 (pirymidyny), H8 (puryny)
5.25 ppm – 6.25 ppm	H1', H5 (pirymidyny)
3.8 ppm – 5.2 ppm	H2', H3', H4', H5', H5''



Rys. 40. Widmo 2D NOESY z zaznaczonymi charakterystycznymi regionami. Czerwonym kolorem zaznaczyłem region *aromatyczno-anomeryczny* H6/H8-(H1'H5), niebieskim rejon H1'-(H2',H3',H4',H5'), fioletowym region *aromatyczno-aromatyczny* H6/H8- H6/H8, zielonym – rejon H6/H8-(H2',H3',H4',H5'), a czarnym pozostałe rejony.

II.4.2 Rozwiązywanie struktury w oparciu o analizę sekwencyjnych ścieżek NOE

Atomy należące do kolejnych reszt nukleotydowych są ze sobą sprzężone dipolowo i jeśli reszty te należą do regionu, który jest regularną helisą to można dla nich zaobserwować charakterystyczną ścieżkę transferu magnetyzacji (rys. 41). Na widmach 2D NOESY jest to widoczne w postaci sygnałów korelacyjnych pomiędzy dwoma atomami tej samej reszty oraz atomami tej samej i kolejnej reszty nukleotydowej. Sygnały te można połączyć prostymi odcinkami. Przykładem takiej ścieżki jest ścieżka NOE $H6/H8_{(i)} - H1'_{(i)}-H6/H8_{(i+1)}$ (rys. 42).

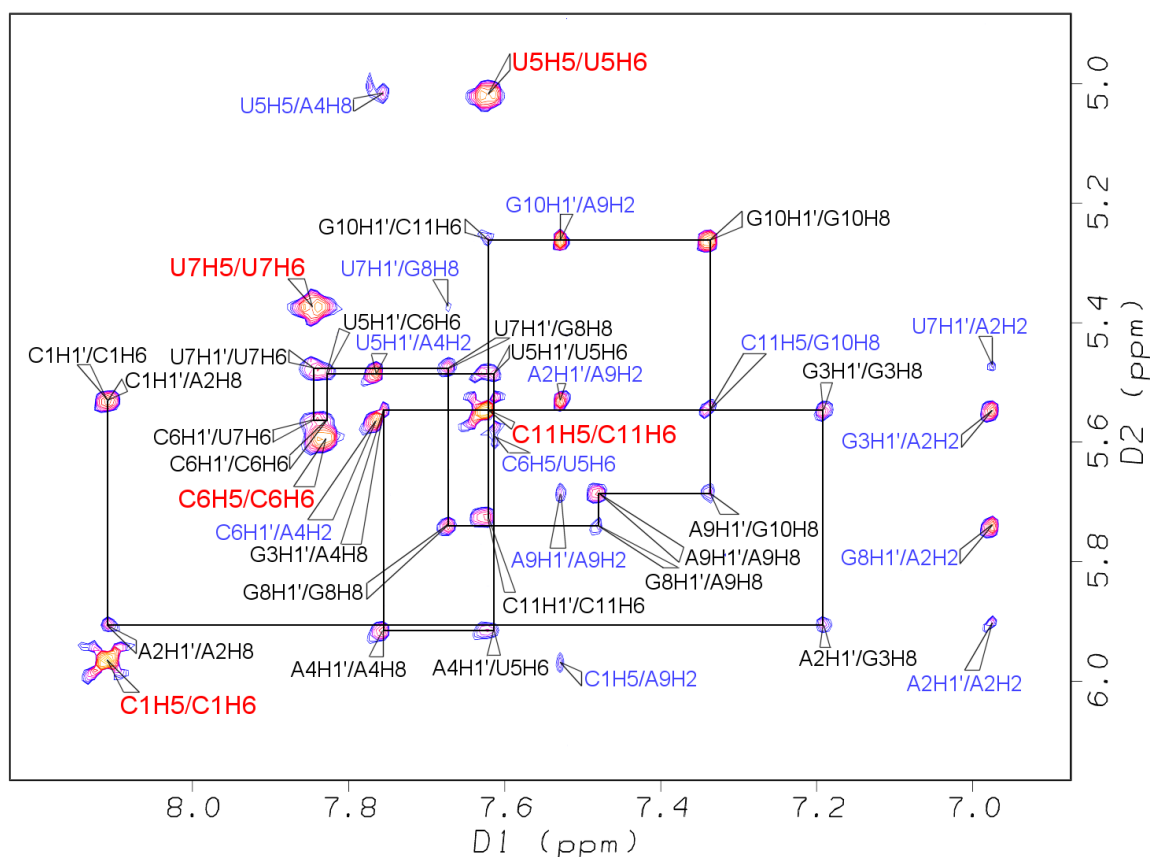


Rys. 41. Transfer magnetyzacji pomiędzy jądrami wodoru należącymi do kolejnych reszt nukleotydowych

Analizę widma przy pomocy techniki sekwencyjnych ścieżek NOE ułatwia analiza innych widm. Przykładowo z widma DQF-COSY można uzyskać informację o wartościach przesunięć chemicznych dla jąder H5 i H6 należących do pirymidyn a z widma $1H\ 13C$ HSQC informację o przesunięciu chemicznym dla H2 adenozyń, oraz informację o możliwych nałożeniach sygnałów korelacyjnych na widmie NOESY. W przypadku nałożonych sygnałów przydatna może się okazać analiza symetrycznego regionu, gdyż ze względu na różną rozdzielczość widma w poszczególnych wymiarach w symetrycznym regionie sygnały te mogą okazać się lepiej rozdzielone.

Analizę najlepiej rozpocząć od regionu *aromatyczno-anomerycznego*, gdyż znajduje się tam tylko jeden rodzaj ścieżek NOE. Uzyskane przypisania można wykorzystać do wyznaczenia

ścieżki H1'-H2', w skład której wchodzi stosunkowo silne sygnały, a następnie ścieżek H6/H8 – H2' oraz H6/H8 – H3'. W zależności od obecności i intensywności sygnałów na widmie, można też wyznaczyć ścieżki H6/H8 – H5', H6/H8 – H5'', H6/H8 – H4' lub ich fragmenty.



Rys. 42. Sekwencyjna ścieżka NOE w regionie *aromatyczno-anomerycznym* dla wzajemnie komplementarnego dupletu o sekwencji 5'-CAGAUCUGAGC-3'. Kolor czerwony – sygnały pochodzące od oddziaływań H5-H6 pirymidyn, kolor czarny – sygnały wchodzące w skład ścieżki NOE, kolor niebieski – pozostałe sygnały.

II.4.3 Wyznaczanie więzów strukturalnych

Na podstawie analizy widm NMR można wyznaczyć szereg więzów, które są niezbędne dla rozwiązania struktury przestrzennej badanej cząsteczki. Zaliczają się do nich więzy odległościowe oraz więzy na kąty torsyjne (76).

W celu uzyskania więzów odległościowych ^1H - ^1H należy poddać analizie widmo na którym zarejestrowany jest efekt NOE, przykładowo 2D NOESY. Objętość sygnału korelacyjnego NOE jest odwrotnie proporcjonalna do odległości pomiędzy protonami od których on pochodzi. Więzy można otrzymać jedną z trzech metod:

- poprzez bezpośrednie porównanie sygnałów korelacyjnych i diagonalnych (ang. *direct comparison of cross and diagonal peaks*)
- poprzez metody iteracyjne (ang. *reference structure based iterative methods*)
- poprzez przybliżenie dla pary izolowanych spinów (ang. *isolated spin pair*)

W Pracowni Biomolekularnego NMR w naszym instytucie stosuje się trzecią z wymienionych metod. Wykorzystując sygnały korelacyjne, dla których odległość jest znana i stała można obliczyć średnią odległość pomiędzy protonami, dla których obserwuje się efekt NOE według wzoru:

$$r_{ij}^6 = (V_{wz}/V_{ij})r_{wz}^6$$

gdzie r_{ij} to odległość pomiędzy protonami, od których pochodzi dany sygnał, r_{wz} – odległość dla wybranego wzorca, a V_{ij} i V_{wz} to odpowiednio objętość wzorcowa i objętość wybranego sygnału. Jako sygnały wzorcowe w przypadku kwasów nukleinowych wykorzystuje się sygnały korelacyjne H5-H6 pirymidyn, które są zarówno intensywne, jak i łatwe do przypisania na podstawie widma COSY. W niektórych przypadkach można również wykorzystać sygnały H5'-H5'' a w przypadku DNA również H2' – H2''. Ponieważ widma 2D NOESY są symetryczne, dla każdej odległości otrzymujemy dwa sygnały, których wartości objętości należy uśrednić. Aby wykorzystać więzy do obliczeń struktury należy określić dla nich zakres tolerancji. Wykorzystuje się w tym celu określony procent wartości danego sygnału zarówno in plus jak i in minus lub klasyfikuje do różnych grup sygnałów o odgórnie określonych zakresach.

Więzy na kąty torsyjne wyznaczane są na podstawie dokładnej analizy różnych widm NMR. Dla kątów torsyjnych w cząsteczce, dla których można zaobserwować sprzężenie spinowo-spinowe można obliczyć wartość kąta na podstawie równania Karplusa. W cząsteczkach RNA są to między innymi sprzężenia H1'-H2' których wielkość, lub brak obserwowalnych sygnałów na widmie DQF-COSY, pozwala na nałożenie więzów na konformację pierścienia cukrowego (77, 78). Wielkość kąta torsyjnego χ (kąt tworzony przez atomy O4'-C1'-N9-C4 puryn lub O4'-C1'-N1-C2 pirymidyn) można oszacować na podstawie analizy odległości pomiędzy protonem H1' a H6 lub H8. Dla nietypowej konformacji *syn*, wartość odległości wynosi pomiędzy 2 a 2,5 Å, co można zaobserwować w postaci znacznie większych wartości objętości sygnałów na widmie. Więzy na kąty α i ζ można nałożyć na podstawie analizy przesunięć chemicznych atomów jąder fosforu ^{31}P (76, 79). Więzy na kąty β i ϵ nakłada się na podstawie analizy sprzężeń skalarnych pomiędzy atomem $\text{P}_{(i)}$ a $\text{H}3'_{(i-1)}$ i $\text{H}5'_{(i)}$, a na kąt γ na podstawie wartości sprzężeń pomiędzy protonami H4' a H5' i H5''.

II.4.4 Tworzenie modelu w oparciu o metody dynamiki molekularnej

W celu stworzenia modelu cząsteczki RNA w oparciu o dane eksperymentalne uzyskane dzięki spektroskopii magnetycznego rezonansu jądrowego najczęściej wykorzystuje się metody dynamiki molekularnej. Obliczenia te polegają na rozwiązywaniu równania ruchu Newtona w określonym polu siłowym, dla niewielkich okresów czasu. Zmiany konformacji cząsteczki wpływają na jej energię potencjalną zgodnie ze zdefiniowanym polem siłowym. Następnie

stosuje się minimalizację energii w której wykorzystuje się semiempiryczne pole siłowe. Są one opisane potencjałem V:

$$V=V_{\text{BOND}}+V_{\text{ANGLE}}+V_{\text{TORS.}}+V_{\text{VDV.}}+V_{\text{EL.}}+V_{\text{EXP.}}$$

gdzie $V_{\text{BOND}}+V_{\text{ANGLE}}+V_{\text{TORS.}}+V_{\text{VDV.}}+V_{\text{EL.}}$ to potencjały odpowiadające za poprawne wartości długości wiązań, kątów płaskich, kątów torsyjnych, oddziaływania Van der Waalsa, oddziaływania elektrostatyczne odpowiednio. $V_{\text{EXP.}}$ odpowiada za dane uzyskane eksperymentalnie, w tym więzy odległościowe i więzy katowe. Przykładowe wartości parametrów wykorzystywanych podczas obliczeń przedstawiłem w tabeli 3 (80).

Obliczenia można prowadzić w przestrzeni kątów torsyjnych (TAMD – *Torsion Angle Molecular Dynamics*), co znacząco zmniejsza liczbę stopni swobody cząsteczki i ułatwia tworzenie modelu (81, 82). Rozwiązanie to zostało zaimplementowane między innymi w programach XPLOR (83, 84) oraz CYANA (85).

Tabela 3. Przykładowe parametry poszczególnych etapów dynamiki molekularnej w przestrzeni kątów torsyjnych

	Etap 1	Etap 2	Etap 3	Etap 4
	Dynamika wysokotemperaturowa TAMD	Powolne schładzanie TAMD	Powolne schładzanie (kartezjański układ współrzędnych)	Minimalizacja metodą gradientów sprzężonych
temperatura	20000 K	20000 K => 3000 K	3000 K => 300 K	-
liczba kroków	8000	10000	6000	1200
długość kroku	0,008 ps	0,007 ps	0,003 ps	-
Δt	64 ps	70 ps	18 ps	-
W_{NOE}	2 => 75	100	215	200
W_{dihedral}	100	100	200	200
W_{vdw}	5 => 0,3	0,3 => 1,0	1,0	1,0

II.5 Algorytmy służące do przypisywania sygnałów na widmach NMR

Proces przypisywania sygnałów stanowi jeden z najtrudniejszych i najbardziej czasochłonnych etapów rozwiązywania cząsteczek metodami spektroskopii NMR. Dla białek zaproponowano szereg rozwiązań wykorzystujących różne algorytmy, jednak ze względu na liczne różnice rozwiązania te nie mogą zostać jednak z sukcesem wykorzystane do automatycznej analizy widm NMR dla cząsteczek RNA. Pierwszy algorytm dedykowany dla RNA zaproponowano w 1990 roku. Bazował on na zredukowanej macierzy odległości (ang. *reduced adjacency matrix*) oraz procedurze wycofywania (ang. *backtracking*) (86).

W 2004 roku w naszym zespole (Zespół Struktury i Funkcji RNA) wspólnie z zespołem prof. J. Błazewicza zaproponowano rozwiązanie bazujące na teorii grafów (1). W rozwiązaniu tym sygnały wchodzące w skład ścieżki to węzły grafu. W założeniu graf jest prosty, czyli każdy węzeł i każda krawędź występują tylko raz w danej ścieżce. Algorytm próbuje dołączyć nową krawędź do istniejącej ścieżki tak długo jak to jest możliwe. Następnie sprawdzana jest zgodność z danymi wprowadzonymi przez eksperymentatora. Ostatnim krokiem jest usunięcie części krawędzi i próba zastąpienia ich nowymi zgodnie z wyżej opisaną procedurą. Do danych eksperckich zaliczają się w tym przypadku długość ścieżki, objętości sygnałów, rozdzielczość widma, nakładające się sygnały, wielkość określająca jaką odległość należy interpretować jako dublety, lista wykluczonych sygnałów, sekwencja RNA, początkowe sygnały, przypisane sygnały, sygnały korelacyjne H5-H6. W tym samym roku zaproponowano również inne podejście (87), w którym wykorzystano algorytm ewolucyjny.

W 2005 roku dla celów generowania ścieżek NOE wykorzystano algorytm tabu search (88). Algorytm ten opiera się na wyszukiwaniu lokalnym, podczas którego dodawany jest pojedynczy węzeł (sygnał) ścieżki i każdorazowej ocenie otrzymanego rozwiązania. W zaproponowanym podejściu prócz najlepszego rozwiązania, do dalszych obliczeń wykorzystywane są także inne, najbardziej obiecujące rozwiązania. Podejście to okazało się szybsze i bardziej precyzyjne od algorytmu ewolucyjnego. W tym przypadku również skuteczność algorytmu jest bardzo zależna od dodatkowych informacji eksperckich dotyczących ścieżki i poszczególnych sygnałów.

Do rozwiązywania nieregulanych struktur RNA, zawierających wybrzuszenia również wykorzystano metody heurystyczne – zarówno algorytm tabu search jak i beam search (89). Obie metody dają dobre rezultaty, ale aby uzyskać jeszcze lepsze wyniki powinny zostać bardziej dostosowane do analizy widm dla nieregularnych struktur RNA.

Innym podejściem jest metoda oparta o statystyki przesunięć chemicznych dla poszczególnych reszt nukleotydowych w określonym otoczeniu chemicznym (2). Na podstawie dostępnych w bazie BMRB - Biological Magnetic Resonance Data Bank dostępnej na stronie <http://www.bmrb.wisc.edu/> opracowali dokładne statystyki przesunięć chemicznych dla jąder atomów węgla i wodoru każdej z reszt zaangażowanych w tworzenie par zasad

z komplementarną nicią RNA. Na podstawie tych danych, dzięki wcześniej zaimplementowanemu algorytmowi FLYA (90) stworzyli możliwość przypisywania sygnałów dla cząsteczek RNA, których reszty są zaangażowane w tworzenie par zasad z komplementarną nicią. Algorytm ten dokonuje przypisań na podstawie mapowania rzeczywistych i przewidzianych na podstawie zadanych reguł sygnałów oraz wykorzystuje algorytm ewolucyjny i znajdowanie minimum lokalnego w celu maksymalizacji wartości oceny.

Problem znajdowania ścieżki NOE na widmach 3D można przedstawić jako problem znalezienia najdłuższej ścieżki o uporządkowanych kolorach (ang. *Orderly Colored Longest simple Path problem*) (91, 92). Dla tak zdefiniowanego problemu zaproponowano dwa matematyczne modele :

- model wykorzystujący cykliczny, skierowany graf c-dzielny (ang. *cyclic c-partite directed graph*)
- model wykorzystujący cykliczny, skierowany graf c-spójny (ang. *cyclic c-connected directed graph*)

Oba modele wykorzystano do obliczeń metodą Branch & Cut. Oba rozwiązania pozwalają na rozwiązanie założonego problemu dla dużej liczby sygnałów korelacyjnych, co przetestowano na sztucznie wygenerowanych danych testowych.

Wedle mojej najlepszej wiedzy, przedstawione wyżej rozwiązania nadal nie pozwalają na w pełni automatyczną analizę wielowymiarowych widm NMR dla długich, znakowanych izotopowo cząsteczek RNA. Ponadto każde z powyższych rozwiązań wymaga wstępnej i często czasochłonnej obróbki widm NMR przez eksperymentatora.

III. MATERIAŁY I METODY

III.1 Język programowania Python

Program RNA NOE*pathfinder* jest napisany w języku Python (<http://www.python.org>). O wyborze języka zadecydowało kilka czynników. Przede wszystkim prosta składnia, podczas tworzenia której postawiono na czytelność kodu. Drugą istotną cechą tego języka jest to, że należy on do języków orientowanych obiektowo (ang. *object-oriented programming languages*). Programowanie obiektowe pozwala na wyodrębnianie fragmentów kodu w postaci klas pełniących określone role w systemie. Każda z klas może posiadać przypisane do niej dane oraz metody – czyli funkcje, za pomocą których można operować na tych danych. Kolejnymi istotnymi kwestiami są powszechne wykorzystanie języka Python w bioinformatyce, dobre wsparcie techniczne, udostępnienie przez różnych programistów modułów pozwalających na łatwiejszą obróbkę danych, możliwość dołączania modułów napisanych w innych językach (C++, Java - <http://www.jython.org>, R - <http://rpy.sourceforge.net>) oraz wsparcie dla wielowątkowości. Ponadto istnieją projekty znacznie wpływające na szybkość działania programu (np. Cython - <http://cython.org>, PyPy - <http://pypy.org>). Połączenie prostej i przejrzystej składni oraz języka orientowanego obiektowo pozwoliło mi na stosunkowo szybkie testowanie różnych algorytmów i rozwiązań poszczególnych problemów, integrację danych i dołączanie nowych elementów w miarę pojawiających się potrzeb.

W celu zapewnienia dużej szybkości obliczeń oraz minimalizacji czasu oczekiwania, program RNA NOE*pathfinder* wykorzystuje wielowątkowość, która umożliwia prowadzenie obliczeń równolegle na kilku procesorach. Poszczególne obliczenia są dzielone na podobnej wielkości zadania, np. wyszukanie krótkich fragmentów ścieżek NOE, rozpoczynające się od różnych punktów startowych. Po zakończeniu danej partii obliczeń, zbiór rozwiązań danego problemu jest synchronizowany.

W początkowej fazie wdrażania program był uruchamiany za pomocą PyPy – alternatywnej implementacji języka Python, która wykorzystuje ideę kompilacji fragmentów kodu przed jego uruchomieniem - JIT (ang. *just-in-time compilation*). Dzięki temu zwiększyła się szybkość obliczeń oraz zmniejszyła ilość wymaganej pamięci obliczeniowej. Niestety jedna z bibliotek programistycznych (Matplotlib) okazała się być niekompatybilna z PyPy. Ponadto pojawiły się liczne problemy z wielowątkowością oraz połączeniem z interfejsem użytkownika. W związku z tym na chwilę obecną program jest uruchamiany za pomocą standardowej implementacji języka Python, jednak w przyszłości planuję migrację na PyPy.

Obliczenia prowadzące do przypisania sygnałów są kosztowne obliczeniowo. Ponieważ trudno osiągnąć satysfakcjonujące moce obliczeniowe na komputerach w laboratoriach NMR, najlepszym rozwiązaniem jest udostępnienie zasobów do obliczeń poprzez stronę internetową. Rozwiązanie takie pozwala również na uniknięcie problemów związanych z instalacją na różnego rodzaju maszynach posiadających odmienne systemy operacyjne oraz aktualizacją

oprogramowania. Interfejs webowy, który szerzej opisałem w rozdziale „Webserwer RNA NOEpathfinder”, stworzyłem za pomocą wzorca projektowego (ang. *framework*) Django opisanego w rozdziale „Wzorzec projektowy Django”.

III.2 Biblioteki programistyczne i skrypty

Python posiada wbudowane standardowe biblioteki, które pozwalają na rozwiązywanie różnych, często spotykanych zadań.

Tabela 4. Biblioteki standardowe języka Python wykorzystane do stworzenia programu RNA NOEpathfinder.

Biblioteka	Opis
re	Biblioteka pozwalająca na wyszukiwanie określonych wzorców w tekście oraz manipulację nimi.
os, sys	Biblioteki umożliwiające współdziałanie z systemem operacyjnym oraz wykorzystanie obiektów i funkcji interpretera języka Python.
argparse	Biblioteka pozwalająca na łatwe przetwarzanie parametrów podanych w linii komend za poleceniem uruchomienia programu
datetime	Moduł pozwalający na wykonywanie różnych operacji na danych dotyczących czasu i daty.
PyYAML	Biblioteka do obsługi formatu YAML, który pozwala na przechowywanie złożonych danych w pliku i jest czytelny dla użytkownika
logging	Biblioteka pozwalająca na zapisywanie etapów działania programu wraz z ostrzeżeniami i błędami w pliku tekstowym.
unittest	Biblioteka ułatwiająca wykonywanie testów jednostkowych, czyli testowania każdej z metod i funkcji w każdej klasie.

Oprócz bibliotek włączonych do programu Python, istnieje bardzo duża liczba bibliotek udostępnianych i rozwijanych przez różnych użytkowników. Część z nich wykorzystałem w programie RNA NOEpathfinder, zarówno w części obliczeniowej jak i w obrębie webserwera.

Aby umożliwić wygodną edycję i konwersję różnych danych wykorzystywanych w programie RNA NOEpathfinder wykorzystałem bibliotekę PyYAML (<http://pyyaml.org/>). YAML (<http://yaml.org/>) to prosty format tekstowy, który jest zarówno czytelny dla użytkownika jak i pozwala na łatwe wczytywanie danych i tworzenie z nich odpowiednich struktur w programie, które można potem łatwo wykorzystać. Biblioteka ta posłużyła przede wszystkim do importowania parametrów oraz plików wewnętrznych programu.

Kolejną biblioteką jest Matplotlib (<http://matplotlib.org/>), która posłużyła mi do automatycznego tworzenia wykresów przedstawiających poszczególne fazy obliczeń oraz ich wyniki. Plikami wynikowymi są pliki graficzne w formacie PNG (ang. *Portable Network Graphics*), które pozwalają na bezstratny zapis danych przy zachowaniu niewielkiej wielkości plików, co jest szczególnie istotne w przypadku systemów webserwerowych.

Ponadto podczas tworzenia oprogramowania wykorzystałem bibliotekę logbook, która była bardzo przydatna do testów jednostkowych.

III.3 Wzorzec projektowy Django

Do stworzenia interfejsu webowego wykorzystałem wzorzec projektowy Django w wersji 1.5 (<https://www.djangoproject.com/>) (93). Jest to oprogramowanie, które pozwala na tworzenie dynamicznych stron www. System ten jest napisany w języku Python i udostępniony na licencji BSD (ang. *Berkeley Software Distribution License*) umożliwiającej jego szerokie wykorzystanie. Na jego wybór zdecydowałem się ze względu na jego uniwersalność oraz łatwość tworzenia spersonalizowanych rozwiązań internetowych z wykorzystaniem języka Python, dużą liczbę funkcjonalnych dodatków oraz szerokie wsparcie na forach internetowych.

Jedną z najistotniejszych cech Django jest implementacja architektury Model-Szablon-Widok (ang. *Model-Template-View*), w której następuje oddzielenie logiki biznesowej (poziom modelu) od logiki aplikacji (poziom widoku) oraz wyglądu (poziom szablonu). Dzięki temu możliwa jest łatwa zmiana zarówno działania webserwisu, architektury bazy danych oraz wyglądu strony bez zmiany pozostałych elementów strony.

III.4 Biblioteki wykorzystane w interfejsie webowym

Kluczową funkcjonalnością w systemie webserwerowym jest rejestracja oraz logowanie użytkowników do serwisu. Podczas rejestracji wykorzystywane są kody CAPTCHA, które są generowane dzięki aplikacji `django-simple-captcha` 0.4.0 pobranej ze strony <https://django-simple-captcha.readthedocs.org/en/latest/>. Wszystkie przesyłane pliki są zabezpieczone przed nieuprawnionym dostępem dzięki mechanizmowi zaimplementowanemu w aplikacji `django-private-files` (<https://bitbucket.org/vvangelovski/django-private-files/src>), co szerzej opisałem w rozdziale „Bezpieczeństwo serwisu”. Z kolei statystyki na stronie wyświetlane są dzięki wykorzystanej przeze mnie aplikacji `django-tracking` (<https://bitbucket.org/codekoala/django-tracking>). Dostęp do panelu administracyjnego dodatkowo zabezpieczyłem z pomocą `django-admin-honeypot` (<https://github.com/dmpayton/django-admin-honeypot>).

III.5 Baza danych PostgreSQL

Dane wysyłane przez użytkowników do systemu webserwerowego muszą zostać zapisane, by mogły być później wykorzystane. W tym celu najlepiej jest wykorzystać bazę danych. Django posiada wbudowane mechanizmy ułatwiające wykorzystanie różnych baz danych. Początkowo wykorzystywałem bazę SQLite (<http://www.sqlite.org/>), w której dane zapisywane są w pojedynczym pliku na dysku, jednak moje testy wykazały, że niektóre skrypty aktualizujące i pobierające informacje z bazy danych mogą powodować czasowe blokowanie webserwisu. Problem ten rozwiązałem zmieniając bazę danych na PostgreSQL (<http://www.postgresql.org/>), ze względu na doświadczenie uzyskane podczas pracy nad innymi projektami.

Baza danych PostgreSQL jest systemem zarządzania bazami danych, rozwijanym przez PostgreSQL Global Development Group. PostgreSQL jest dostępny na otwartej licencji The

PostgreSQL Licence, która umożliwia szerokie wykorzystanie oprogramowania bazodanowego w różnych systemach informatycznych, w szczególności w webserwisach i stronach internetowych.

III.6 System kolejkowy

Aby umożliwić płynne korzystanie z webserwisu podczas trwania obliczeń, oraz by zabezpieczyć serwer przed nadmiernym obciążeniem wynikającym z jednoczesnego uruchomienia kilku zadań obliczeniowych, niezbędne okazało się wprowadzenie systemu kolejkowego. Pierwotnie było to moje autorskie rozwiązanie oparte o plik tekstowy z listą eksperymentów oraz skrypt powłoki uruchamiany przez program Cron. Niestety rozwiązanie to okazało się mało efektywne i trudno rozwijalne a w niektórych okolicznościach mogło powodować problemy z uruchomieniem zadania. Jednocześnie okazało się, że obliczenia będą uruchamiane na tym samym serwerze na którym będzie się znajdować strona internetowa. W związku z tym do kolejkovania zdecydowałem się na użycie oprogramowania Celery (<http://docs.celeryq.org/en/latest/>), RabbitMQ (<http://www.rabbitmq.com/>) oraz biblioteki dedykowanej dla wzorca projektowego django – Django-celery (<https://pypi.python.org/pypi/django-celery>). Oprogramowanie to pozwala na dużą dowolność w konfiguracji uruchamiania zadań przy zachowaniu możliwości wykorzystania wielowątkowości. Wdrożenie tego rozwiązania wiązało się z refaktoryzacją części obliczeniowej oraz włączeniu jej w strukturę katalogów webserwisu, ale jednocześnie uzyskałem możliwość łatwiejszego zapisu danych do bazy danych.

III.7 Oprogramowanie dla danych NMR

Aby z widma NMR uzyskać dane niezbędne do obliczeń z wykorzystaniem RNA *NOEpathfinder*, należy zaimportować dane widmo do jednego z programów do przetwarzania danych eksperymentalnych a następnie ręcznie lub za pomocą automatycznych procedur należy zaznaczyć sygnały widoczne na widmie.

Do zbierania danych eksperymentalnych niezbędnych do testów, wykorzystałem oprogramowanie Felix oraz TopSpin 3.1 (Bruker). Program Felix jest wykorzystywany w naszym zespole do przypisywania sygnałów oraz obliczania objętości sygnałów, natomiast program TopSpin stanowi oprogramowanie dostarczane przez firmę Bruker do spektroskopów NMR wykorzystywanych w naszym Instytucie. Oba programy pozwalają na automatyczne lub półautomatyczne generowanie listy sygnałów obecnych na widmach. W przypadku programu Felix pobierane są również objętości, które można wykorzystać do stworzenia więzów strukturalnych, w przeciwieństwie do intensywności które generuje program TopSpin. Niestety w przypadku tego programu uzyskuje się współrzędne na widmie, które następnie trzeba zamienić na wartości przesunięć chemicznych. Z tego powodu użytkownik jest proszony

o podanie wartości przesunięć chemicznych dla dwóch sygnałów mających możliwie skrajne wartości przesunięć chemicznych w każdym z wymiarów.

III.8 Inne narzędzia

Oprócz wyżej wymienionych programów i bibliotek, wykorzystałem również kilka innych programów i metodologii, które warto wymienić. Pierwszy z programów - Apache w wersji 2.2.22 odpowiada za udostępnianie strony w Internecie zainstalowanego na systemie Debian (SMP Debian 3.2.51-1 x86_64).

Grafikę dla strony startowej stworzyłem z wykorzystaniem programu Inkscape 0.48.3.1 (<http://inkscape.org/>) oraz Gimp w wersji 2.6.12 (<http://www.gimp.org/>). Szablony CSS stworzyłem z wykorzystaniem szablonu Blueprint (<http://www.blueprintcss.org/>). Wygląd panelu użytkownika oparłem o gotowy szablon wykonany w technologii bootstrap (<https://wrapbootstrap.com/theme/detail-admin-responsive-theme-WB07061TJ>)

Schematy wykorzystane w tej dysertacji, ukazujące sposób wykorzystania programu oraz interfejsu webowego wykonałem zgodnie z regułami BPMN (ang. *Business Process Modeling Notation*). Diagramy klas i kolaboracji wykonałem zgodnie z konwencją UML (ang. *Unified Modeling Language*). Wykresy te opracowałem za pomocą programu DIA (<https://live.gnome.org/Dia>).

Zarządzanie plikami programu realizowałem najpierw poprzez tworzenie kopii zapasowych oprogramowania a następnie poprzez rozproszony system kontroli wersji (oprogramowania) GIT (<http://git-scm.com/>). Oprogramowanie instalowałem przy użyciu wirtualnego środowiska języka Python – Virtualenv (<http://virtualenv.readthedocs.org/en/latest/>). Pakiety instalowałem przy użyciu oprogramowania PIP (<https://pip.readthedocs.org/en/latest/>), które bardzo dobrze współpracuje ze środowiskiem virtualenv.

Kod programu pisałem wykorzystując zarówno dostępny w wielu dystrybucjach linuxa edytor tekstowy Gedit jak i dedykowany dla programistów program PyCharm (<https://www.jetbrains.com/pycharm/>), który ułatwia zarówno tworzenie oprogramowania w języku Python oraz stron wykorzystujących wzorzec projektowy Django, ale również ułatwia pracę z oprogramowaniem GIT i Virtualenv.

III.9 Dane wykorzystane w programie

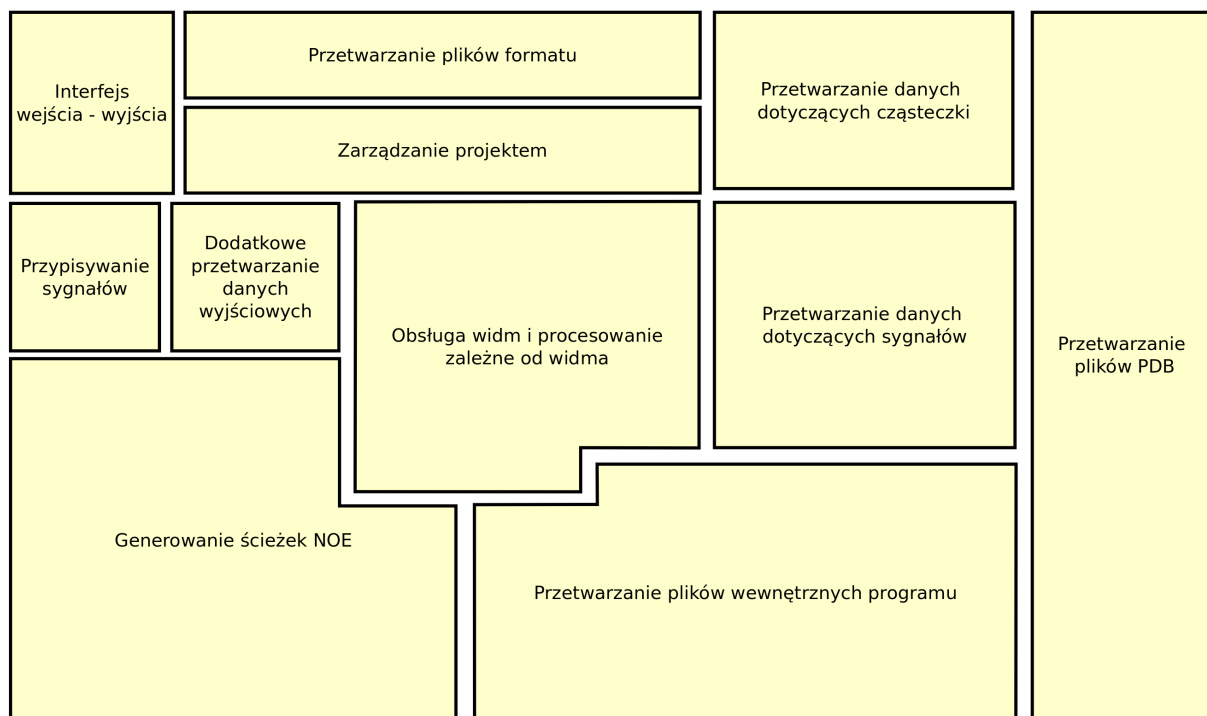
Dane, które wykorzystałem podczas tworzenia programu RNA NOEpathfinder zostały zebrane z kilku różnych źródeł. Wartości przesunięć chemicznych dla poszczególnych atomów w każdej z reszt nukleotydowych wprowadziłem korzystając z bazy Biological Magnetic Resonance Data Bank (<http://www.bmrb.wisc.edu/>). Typowe wartości odległości pomiędzy poszczególnymi atomami wprowadziłem dzięki uśrednieniu wartości obliczonych dla cząsteczek RNA zdeponowanych w bazie Protein Data Bank (<http://www.rcsb.org/pdb/home/home.do>).

Dokładne wartości przesunięć chemicznych dla każdej z reszt nukleotydowych znajdującej się w określonym otoczeniu chemicznym wprowadziłem na podstawie danych uzupełniających z publikacji dotyczącej metody przypisywania sygnałów w oparciu o statystyki przesunięć chemicznych (2).

IV. WYNIKI BADAŃ WŁASNYCH

IV.1 Schemat logiczny działania programu RNA NOEpathfinder

Program RNA NOEpathfinder składa się z wielu klas i metod, które można przypisać do określonej funkcji i etapu analizy. Schematyczny podział programu na funkcjonalne grupy przedstawiłem na rysunku 43.



Rys. 43. Schematyczne przedstawienie funkcjonalnych części programu RNA NOEpathfinder odpowiedzialnych za przetwarzanie poszczególnych partii danych na odpowiednich etapach analizy.

Reprezentacja danych

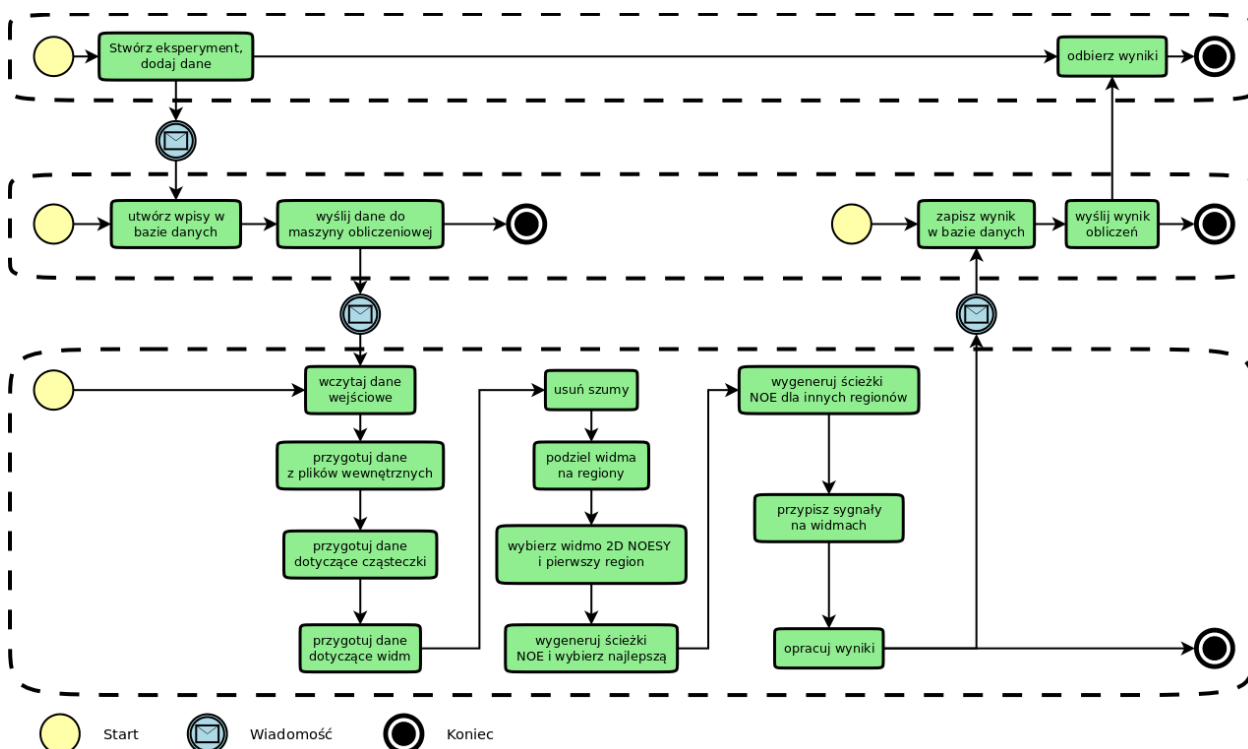
Program RNA NOEpathfinder wykorzystuje szereg różnych danych wejściowych. Do najistotniejszych należą informacje dotyczące położenia sygnałów na widmach, ich szerokości w każdym z wymiarów oraz objętości lub intensywności. Dane dotyczące pojedynczego sygnału są przechowywane w słowniku, gdzie kluczami są nazwy poszczególnych parametrów obejmujące również nazwy i numeracje reszt i atomów uzyskane podczas przypisywania sygnałów. Struktury te są zebrane w innym słowniku, którego kluczami mogą być albo identyfikatory z programu używanego do zaznaczenia sygnałów (ang. *peak picking*), albo kolejne liczby naturalne. Opisywana wyżej struktura jest przechowywana w obiektach klasy **Peaks**, która pomaga zarządzać sygnałami, oraz weryfikuje ich poprawność, albo w obiektach jednej z dziedziczących z niej klas: **FelixPeaks** lub **TopspinPeaks**.

Oprócz danych dotyczących sygnałów, każde widmo jest reprezentowane również przez szereg parametrów, określających między innymi rodzaj atomów w każdym wymiarze czy

warunki wykonania eksperymentu (klasa **ExperimentalParameters**). Dane te wraz z sygnałami przechowywane są w obiektach klasy **Spectrum**, lub jednej z dziedziczących z niej klas. Widma są zorganizowane w większą strukturę w klasie **Spectra** pozwalającą na wyszukanie odpowiednich widm dla danego problemu. Regiony na widmie są reprezentowane jako n-poziomowa struktura słownikowa, gdzie n odpowiada liczbie wymiarów widma, a identyfikatorem jest numer regionu w określonym wymiarze. Na ostatnim poziomie tej struktury znajdują się obiekty klasy **Peaks**. Struktura ta przechowywana jest w obiektach klasy **Regions** zawartych w **Spectrum**.

Parametry niezbędne dla uruchomienia programu zgromadzone są w klasie **InternalParameters**. Struktura drugorzędowa znajduje się z kolei w obiektach klasy **Molecule**, a struktura przestrzenna w obiektach **PDBMolecule** dziedziczących z **MoleculeData**. Dane z plików wewnętrznych programu przetwarzane i gromadzone są w obiektach odpowiednich klas dziedziczących z klasy **TemplateElement**. Wymienione wcześniej dane zebrane są w obiekcie klasy **InternalFormat**, do której referencja przekazywana jest pomiędzy obiektami na każdym etapie analizy widm NMR. Wszystkie klasy wraz z łączącymi je relacjami zostały przedstawione na schemacie załączonym do rozprawy doktorskiej.

Przebieg programu (ang. *program flow*)



Rys. 44. Uproszczony diagram choreografii przedstawiony zgodnie z BPMN ukazujący przebieg programu RNA NOEpathfinder. Licząc od góry linią przerywaną zaznaczyłem odpowiednio użytkownika, webserwer, superkomputer. W chwili obecnej rolę zarówno webserwera jak i superkomputera pełni ta sama maszyna.

Ogólny schemat współdziałania poszczególnych elementów systemu z użytkownikiem przedstawiłem na rysunku 44. Użytkownik posiadający konto wprowadza dane dotyczące eksperymentu, poszczególnych widm, struktury drugo- lub trzeciorzędowej oraz parametry a następnie zatwierdza je do obliczeń. Tak przygotowane dane są zapisywane w bazie danych. Następnie następuje uruchomienie zadania obliczeniowego z wykorzystaniem systemu kolejkowego. Status obliczeń jest na bieżąco aktualizowany w bazie danych i dostępny poprzez stronę internetową. Po zakończeniu obliczeń dane są zapisywane w bazie danych i udostępniane użytkownikowi. Obliczenia prowadzone na superkomputerze obejmują:

- wczytanie parametrów
- wczytanie danych z plików wewnętrznych programu lub zastępujących je plików od użytkownika oraz przetworzenie w celu uzyskania odpowiednio wyliczonych danych i reguł
- wczytanie i przetworzenie danych dotyczących struktury drugo- lub trzeciorzędowej badanej cząsteczki
- wczytanie widm NMR wraz z niezbędnymi parametrami
- eliminację szumów na widmach
- podział widm na regiony
- korelację poszczególnych sygnałów między widmami
- przypisanie na podstawie zebranych danych części sygnałów
- wybór widma NMR o potencjalnie najlepszych właściwościach dla wyznaczenia ścieżki NOE oraz regionu na którym należy wyznaczyć ścieżkę
- wygenerowanie słownika możliwych połączeń pomiędzy poszczególnymi sygnałami
- przygotowanie funkcji oceniających
- wygenerowanie krótkich fragmentów ścieżek NOE w rejonie *aromatyczno-anomerycznym* oraz ich dopasowanie do sekwencji
- kombinowanie ze sobą krótkich elementów ścieżek NOE, uzupełnienie luk pomiędzy nimi i na końcach ścieżek oraz wybór optymalnej ścieżki lub ścieżek
- przypisanie sygnałów znajdujących się w wybranej ścieżce NOE
- opracowanie wyników, w tym graficzną reprezentację ścieżek NOE

Obsługa błędów

Podczas działania programu może dojść do problemów z przetwarzaniem danych dostarczonych przez użytkownika, np. z powodu błędnego formatu lub braku danych. Powoduje to sytuacje, które program powinien odpowiednio obsłużyć i poinformować użytkownika o zaistniałym problemie. Do obsługi błędów i ostrzeżeń wykorzystany został moduł logging. W programie RNA NOEpathfinder wykorzystuję trzy poziomy błędów:

- FATAL – błąd, którego nie da się ominąć i który powoduje poważne problemy w działaniu programu
- ERROR – błąd, który powoduje, że część danych nie będzie uwzględniona i może prowadzić do błędnych wyników.
- WARNING – Ostrzeżenie o działaniu, które choć dopuszczalne, najprawdopodobniej nie jest zamierzeniem użytkownika

Dodatkowo wykorzystany jest poziom INFO, który pozwala na poinformowanie o prawidłowo wykonanych operacjach. Niezależnie od wymienionego wyżej systemu stworzona została klasa **StatusUpdater**, która służy do aktualizacji danych na stronie internetowej i szybkiego poinformowania użytkownika o statusie obliczeń oraz potencjalnych błędach.

IV.2 Algorytmy przetwarzania wielokierunkowych widm NMR dla RNA

IV.2.1 Rozróżnianie sygnałów od szumów

Jednym z kluczowych elementów analizy widm NMR jest zaznaczenie sygnałów na widmach. Można to wykonać ręcznie, bazując na doświadczeniu eksperymentatora lub z pomocą odpowiednich programów. Automatyczne zaznaczanie sygnałów na widmach jest skomplikowaną procedurą. W większości przypadków polega ono na wykryciu lokalnego maksimum które może być sygnałem a następnie wykonanie na nim odpowiednich operacji matematycznych. Do najprostszych metod należy wykorzystanie symetrii widm homojądrowych, w celu potwierdzenia istnienia danego sygnału oraz eliminacja wszystkich sygnałów, których objętość lub intensywność jest poniżej określonej wartości. Często stosuje się również przewidywanie kształtu sygnałów, dzięki czemu można z dużym prawdopodobieństwem określić ile i jakiej wielkości sygnały znajdują się w danym miejscu (94). Procedury te są jednak zależne od jakości widma i poziomu szumów. W związku z tym, w wyniku wykorzystania automatycznego znajdowania sygnałów uzyskuje się albo niekompletną listę rzeczywistych sygnałów – TP (ang. *true positives*), albo listę zawierającą również błędnie zaklasyfikowane szумы - FP (ang. *false positives*). Problem ten jest szczególnie istotny w przypadku widm 2D-NOESY, na których obserwuje się dużą liczbę mało intensywnych sygnałów.

Jednym z założeń oprogramowania RNA NOEpathfinder jest stworzenie systemu, który będzie minimalizował ilość pracy eksperymentatora podczas przygotowywania danych wejściowych. W związku z tym stworzyłem uproszczony system (który opisałem w rozdziale „Programy towarzyszące”), który na podstawie listy widm automatycznie zaznacza na nich sygnały oraz wyznacza ich objętości w programie Felix. Program ten korzysta z udostępnionego w tym programie interfejsu, oraz procedury zaznaczania sygnałów powyżej określonego poziomu szumów. Niestety system ten nie potrafi sam oszacować poziomu szumów na widmie. W związku z tym często dochodzi do sytuacji, że prócz rzeczywistych sygnałów, na liście sygnałów znajdują się również błędnie zakwalifikowane szумы. Brak sygnału na liście, może spowodować trudności lub niemożliwość ustalenia ścieżki NOE, natomiast znajdujące się tam szумы w bardzo dużym stopniu zwiększają czas obliczeń. Z tego powodu podczas tworzenia programu RNA NOEpathfinder, olbrzymią wagę przyłożyłem do procedur eliminacji szumów. Proces ten podzieliłem na kilka etapów wykorzystujących odmienne algorytmy:

- eliminacja szumów na podstawie przewidywanej liczby sygnałów
- eliminacja szumów dla widm homojądrowych
- eliminacja szumów dla widm heterojądrowych

Redukcja szumów na podstawie przewidywanej liczby sygnałów.

W trakcie generowania listy sygnałów tracona jest informacja dotycząca ich kształtu. W związku z tym najprostszą metodą eliminacji szumów jest wykorzystanie określonej wartości objętości lub intensywności. Problem jaki się w tym wypadku pojawia to wyznaczenie wartości

granicznej. W tym celu można jednak wykorzystać przewidywaną liczbę sygnałów. Dla cząsteczki o określonej strukturze drugorzędowej można przewidzieć ile sygnałów będzie widoczne na danym widmie. W tym celu w moim programie wykorzystuję informacje o typowych sygnałach jakie powinny być widoczne dla danej reszty, zależnie od jej położenia względem innych reszt nukleotydowych. Uzyskana wartość N_{ref} nie jest dokładna i wykorzystanie jej mogłoby prowadzić do utraty informacji o części rzeczywistych sygnałów. W związku z tym jako wartość odcięcia N_a w moim programie wykorzystuję jej wielokrotność. Zadowalające wyniki udało mi się uzyskać dla wartości między $1.6 * N_{ref}$ a $2.5 * N_{ref}$.

Zaobserwowałem, że dla niektórych widm widoczna jest istotna różnica pomiędzy sygnałami z listy wśród których znajdują się sygnały TP, a sygnałami z listy wśród których nie znajdują się TP. W moim programie próbuję wykorzystać tą obserwację poprzez wybór maksymalnej wartości uzyskiwanej ze wzoru:

$$score = diff - (((maxdifinrange / N_{ref}) * ((N_{curr} - N_{ref}) * (N_{curr} - N_{ref})) ** 0.5)))$$

w przedziale wartości $N_{curr} \in <A * N_{ref}, B * N_{ref}>$

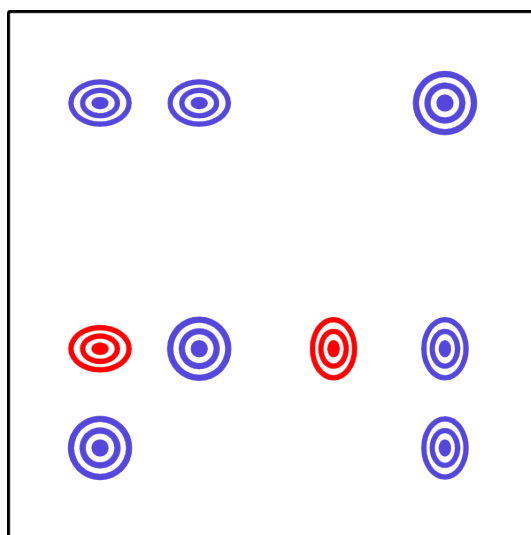
gdzie $maxdifinrange$ to maksymalna różnica objętości pomiędzy sygnałami z zadanego przedziału, wartości A i B wyznaczyłem jako 1.6 i 2.5 odpowiednio.

Procedurę tą można wykorzystać na dwóch etapach – pierwszy z nich jest to wstępna selekcja, natomiast drugi to eliminacja szumów po etapie podziału widm na regiony. W moim programie wykorzystuję ją tylko na pierwszym z tych dwóch etapów ze względu na ryzyko utraty informacji o TP. Pozostałe rozwiązania, opisane niżej pozwalają na usunięcie wystarczającej ilości FP.

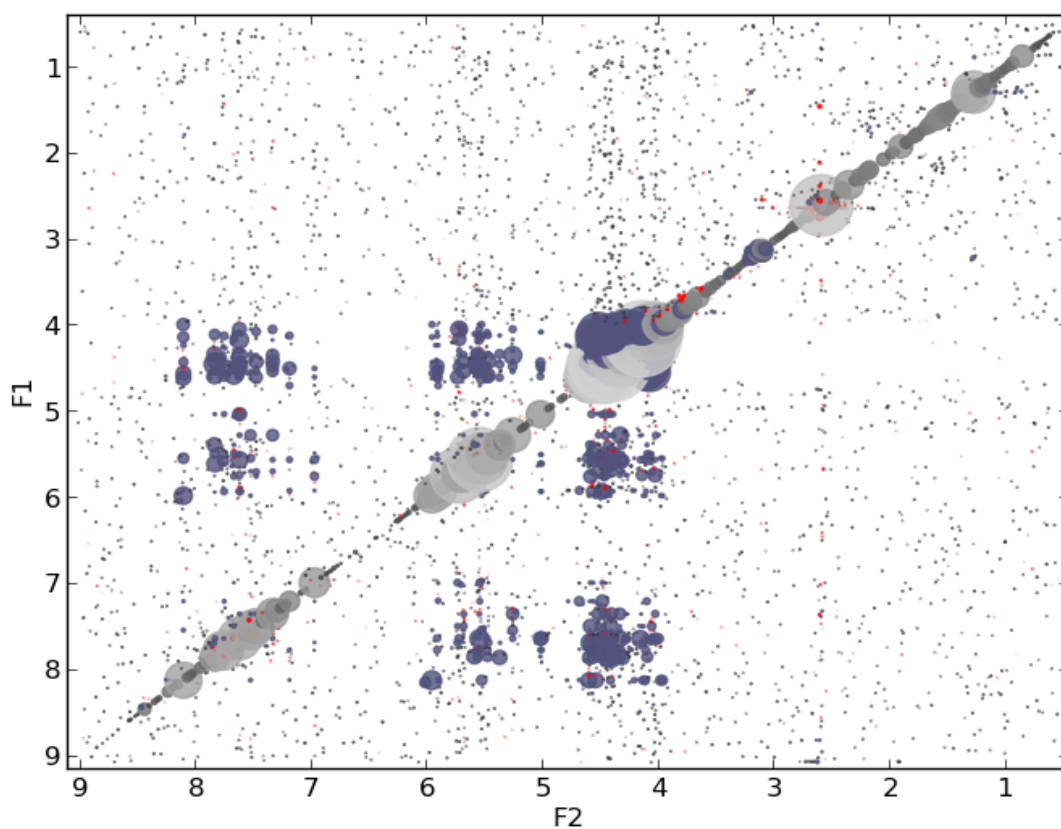
Redukcja szumów dla widm homojądrowych.

Widma homojądrowe wykazują symetrię osiową. Oś symetrii stanowi diagonalna, na której można zaobserwować sygnały odpowiadające sygnałom widma 1D. Symetrycznie po obu stronach diagonalnej zlokalizowane są sygnały odpowiadające za korelację pomiędzy sygnałami pochodzącymi od poszczególnych jąder. W związku z tym jeśli po jednej stronie widma zaznaczony jest sygnał, który nie ma odpowiednika po drugiej stronie, jest on najprawdopodobniej FP (rys. 45). Wykorzystując tę właściwość widm homojądrowych można dla nich dokonać dalszej eliminacji szumów.

Pierwszym krokiem jest wyszukanie i zaznaczenie sygnałów znajdujących się na diagonalnej. Procedura ta opiera się o wyszukanie sygnałów o takich samych wartościach przesunięć chemicznych w obu rozpatrywanych wymiarach w zakresie błędu. Niestety efektem ubocznym tej procedury jest eliminacja sygnałów korelacyjnych dla jąder, których przesunięcia chemiczne różnią się o niewielką wartość.



Rys. 45. Symboliczne przedstawienie sygnałów na homojądrowym widmie 2D. Kolorem niebieskim zaznaczone są symetryczne sygnały (TP), natomiast kolorem czerwonym – potencjalne szumy (FP)



Rys. 46. Widmo 2D NOESY dla wzajemnie komplementarnego oligonukleotydu. Kolorem niebieskim zaznaczone są sygnały posiadające symetrycznie położonego partnera, kolorem szarym – nie posiadające partnera a kolorem czerwonym sygnały o ujemnej wartości objętości

Kolejnym krokiem jest podzielenie sygnałów na dwie grupy – sygnały znajdujące się „nad” diagonalną oraz „pod” diagonalną. Następnie tworzona jest struktura słownikowa w której zawarta jest informacja o możliwych dopasowaniach sygnałów z obu stron widma wraz z wartościami oceny tego dopasowania. Wartość score obliczana jest według wzoru:

$$A*(ave/2 - diff) + B*(2 - (volume_{stronger}/volume_{weaker})) \text{ dla } volume_{stronger}/volume_{weaker} < 2$$

$$A*(ave/2 - diff) \text{ dla } volume_{stronger}/volume_{weaker} \geq 2$$

gdzie A i B to współczynniki ustalane przez użytkownika (ich wartości domyślne są równe 3), ave to średnia wielkość szerokości sygnału w wybranych wymiarach, diff to odległość od idealnie symetrycznej pozycji sygnałów a $volume_{stronger}$ i $volume_{weaker}$ to objętości silniejszego i słabszego sygnału odpowiednio. Obliczona w ten sposób wartość score uwzględnia zarówno różnicę położenia jak i różnicę objętości sygnałów, co ułatwia dopasowanie sygnałów.

Kolejnym krokiem jest stworzenie klastrów skupiających rozwiązania zawierające te same sygnały. Tak stworzone klastry są następnie przeszukiwane za pomocą algorytmu brute-force w celu znalezienia najlepszego dopasowania każdego z elementów danego klastra. Sygnały nie posiadające swojego odpowiednika są odrzucane jako szumy. Potencjalne pomyłki powstać mogą jedynie dla grupy bardzo silnie nałożonych sygnałów o podobnych wartościach objętości. Etap ten jest wykonywany przed procesem dopasowywania sygnałów pochodzących od różnych widm NMR.

Redukcja szumów dla widm heterojądrowych

Widma heterojądrowe, a w szczególności widmo 1H - ^{13}C HSQC są wykorzystywane do wykrywania sygnałów pochodzących od określonych rodzajów jąder, na przykład sygnałów od jąder H2 adenzyn, lub H5/H6 pirymidyn oraz do wykrywania wartości przesunięć chemicznych dla których może występować nakładanie sygnałów na innych widmach. Z tego powodu redukcja szumów do możliwie niskiego poziomu przy zachowaniu wszystkich widocznych sygnałów pełni dużą rolę w programie RNA NOEpathfinder.

Metodę redukcji szumów omówię na przykładzie widma 1H - ^{13}C HSQC, dla którego optymalizowałem tą metodę. Widmo 1H - ^{13}C HSQC można podzielić na regiony co dokładnie omówiłem w rozdziale „Wyszukiwanie regionów na widmach”, w których występują sygnały pochodzące od określonych typów jąder. Jednocześnie wyznaczane są regiony „puste”, na których nie powinny być widoczne żadne sygnały. W związku z tym wszelkie sygnały które znajdują się w tych regionach mogą zostać potraktowane jako FP i na ich podstawie oszacowanie poziomu szumów. Ze względu na możliwość pojawiania się stosunkowo silnych FP, zdecydowałem się dodać dodatkowe ograniczenie, które do wyznaczania poziomu szumów nie

wykorzystuje ułamka najsilniejszych FP na widmach, przez co nie jest tracona informacja o TP o niewielkiej objętości lub intensywności.

IV.2.2 Wyszukanie regionów na widmach

Dla regularnej prawoskrętnej helisy ścieżki NOE są widoczne w określonych regionach na widmach NOESY. Z tego powodu aby wyznaczyć ścieżki NOE dla danej cząsteczki najlepiej jest osobno analizować poszczególne regiony. W tym celu należy poprawnie wykryć położenie tych regionów na widmie. Zadanie to realizowane jest przez obiekt klasy **RegionsFinder**, która w wyniku zwraca wartości graniczne poszczególnych regionów. Najprostszym rozwiązaniem byłoby wykorzystanie stałych wartości przy których następowałby podział na poszczególne regiony, ale w niektórych przypadkach może się okazać że kilka sygnałów zostanie błędnie zaliczonych do innego regionu niż tego do którego należą. Ponadto dla zmodyfikowanych par zasad mogą wystąpić niewielkie różnice jeśli chodzi o granice regionów na widmie NOESY. Aby umożliwić prawidłowe wyszukiwanie regionów i zapewnić jak największą uniwersalność, zastosowałem połączenie kilku algorytmów wyszukiwania, które następnie są używane do wspólnego wyznaczenia najbardziej prawdopodobnej wartości rozdzielającej poszczególne regiony.

Wyszukanie według wartości określonych w szablonie

Jak wspominałem wyżej, najłatwiejszą metodą jest wyszukiwanie regionów w oparciu o ustalone wartości, uzyskane przez obliczenia prowadzone w klasie **TemplateRegions**. Dodatkowo jeśli w miejscu podziału znajduje się sygnał lub sygnały, to wyszukiwane jest najbliższe miejsce, gdzie można dokonać podziału. Niemniej testy wykazały, że w niektórych przypadkach tak przygotowane wartości podziału widma mogą znacząco odbiegać od wartości oczekiwanych. W związku z tym wycofałem to rozwiązanie z programu RNA NOEpathfinder. Niemniej wartości odcinające oraz zakresy regionów są wykorzystywane przez pozostałe algorytmy.

Wyszukanie według największych obszarów bez sygnałów

Pomiędzy poszczególnymi regionami zawierającymi określone sygnały najczęściej obserwuje się obszary całkowicie pozbawione sygnałów. Można to wykorzystać w celu wyznaczenia wartości przesunięć chemicznych oddzielających regiony. W tym celu algorytm sortuje sygnały według przesunięć chemicznych w wybranym wymiarze. Następnie wyszukuje wybraną liczbę największych różnic pomiędzy przesunięciami chemicznymi. Procedura jest powtarzana dla każdego z wymiarów. Dodatkowo, aby poprawnie wyznaczyć regiony nawet w przypadku istnienia dużej ilości szumów (FP), lub pojedynczych słabych sygnałów pomiędzy regionami, procedura jest wykonywana dla 40, 50, 60, 70 i 80 procent najsilniejszych sygnałów na widmie.

Metoda ta jest stosunkowo mało czuła na szумы. Jest jednak bardzo wrażliwa na obecność pojedynczych FP o stosunkowo dużej objętości mających przesunięcia chemiczne oddalone od typowych zakresów rzeczywistych regionów. Z tego powodu przed wykonywaniem wyszukania regionów z użyciem tej metody – eliminowane są najpierw regiony, które nie zawierają ani jednego sygnału a następnie eliminowane są te sygnały, które nie należą do żadnego z pozostałych regionów i dopiero na tak przygotowanych danych prowadzone są obliczenia. Może się również zdarzyć, że metoda ta może powodować błędne zakwalifikowanie słabych sygnałów na widmie do danego regionu.

Wyszukiwanie według klastrów

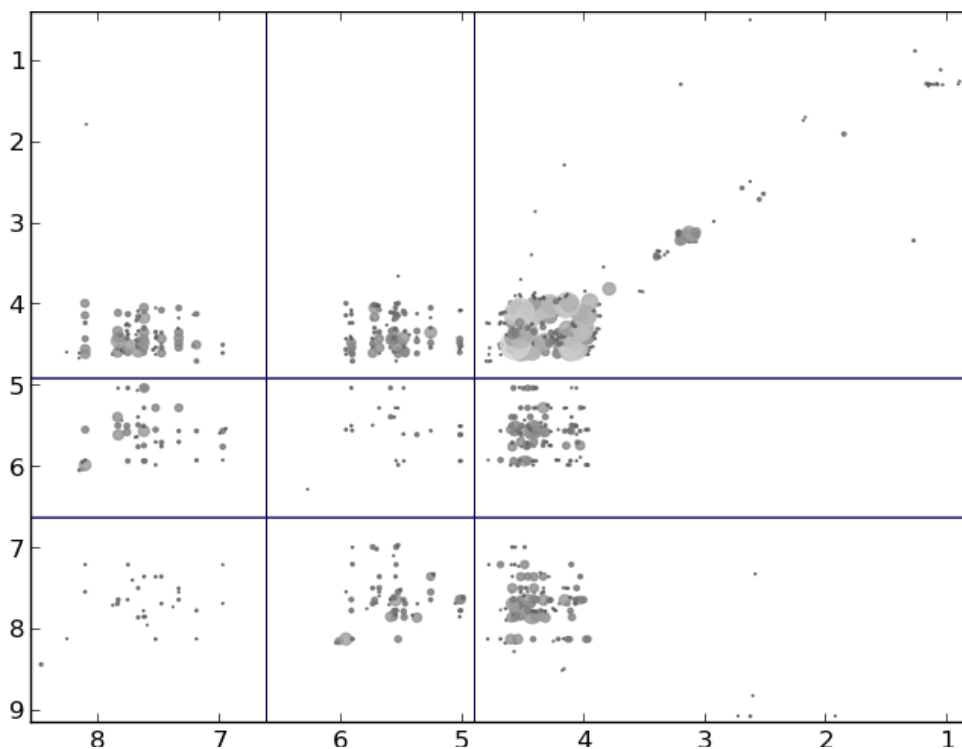
W celu wyszukania regionów, można wykorzystać klastrowanie danych. W tym celu wykorzystałem zmodyfikowany algorytm UPGMA (ang. *Unweighted Pair Group Method with Arithmetic Mean*). Algorytm ten polega na tym, że z całego zbioru sygnałów wybierane są dwa sygnały dla których różnica wartości przesunięć chemicznych w wybranym wymiarze jest najmniejsza. Następnie dla tych sygnałów tworzony jest zmodyfikowany wpis zawierający średnią, maksymalną i minimalną wartość przesunięcia chemicznego w tym wymiarze oraz informację o całkowitej liczbie sygnałów z jakich został utworzony. Wpis ten zastępuje w zbiorze sygnały, z których został utworzony. Wartość średnia jest używana podczas porównywania wartości przesunięcia chemicznego wraz z innymi sygnałami. Następnie krok ten jest powtarzany tak długo, aż pozostanie liczba wymiarów zdefiniowana przez dane wejściowe. Kolejnym krokiem jest wybranie określonej (przez obiekt klasy **TemplateRegions**) ilości przesunięć chemicznych, które rozdzielają poszczególne podzbiory, co odbywa się poprzez obliczenie wartości średnich pomiędzy maksymalną wartością przesunięcia chemicznego dla jednego regionu i minimalną dla kolejnego. Klastrowanie to wykonywane jest dla każdego z wymiarów widma osobno.

Metoda ta jest bardzo czuła na szумы lub sygnały od innych związków nie wchodzące w skład regionów, gdyż jeśli pośród sygnałów w programie, pojawiają się sygnały lub szумы o wartościach oddalonych od wartości typowych dla rzeczywistych regionów na widmie, mogą powstać regiony złożone wyłącznie z szumów, a rzeczywiste regiony mogą zostać połączone w pojedynczy region. Problem ten rozwiązałem poprzez dodatkową modyfikację algorytmu UPGMA. Nowa wersja algorytmu najpierw wyszukuje sygnały znajdujące się najbliżej środka każdego z idealnych regionów wyznaczonych przez obiekt klasy **TemplateRegions**. Następnie dołącza kolejne sygnały tylko do klastra zawierającego jeden z wyszukanych sygnałów. W ten sposób unika się wyszukiwania regionów w miejscach gdzie one nie występują oraz zapewnia wyszukanie każdego ze zdefiniowanych regionów.

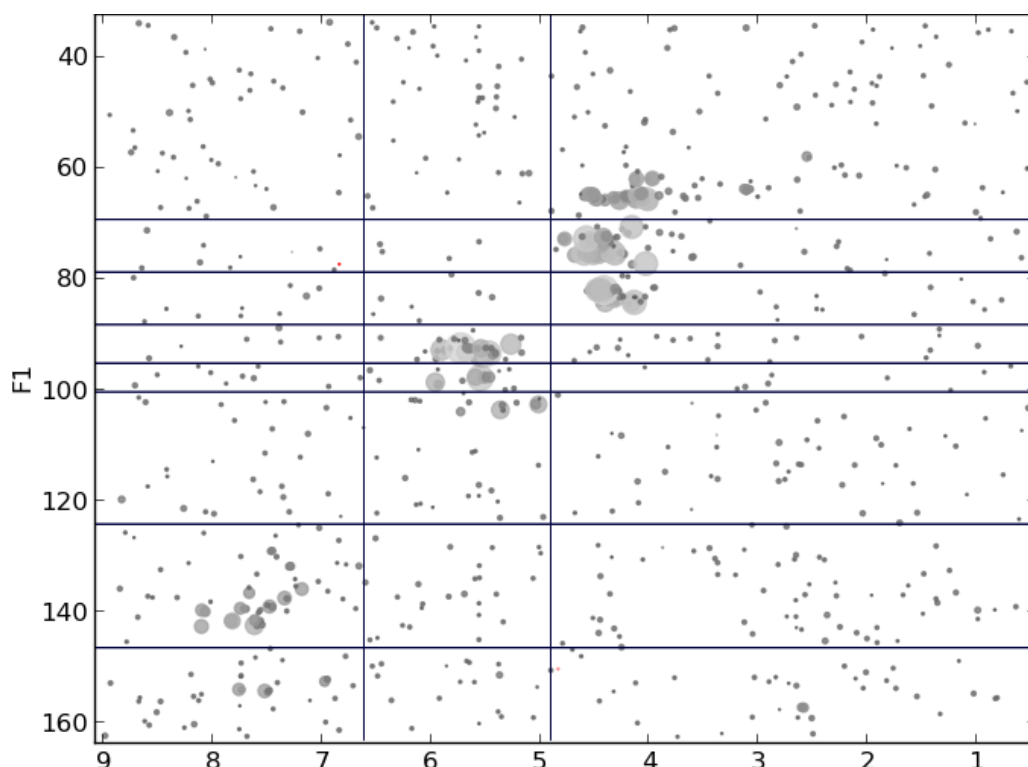
Połączenie wyszukanych wartości i wybór przesunięcia

Po wyszukaniu klastrów oraz największych różnic w wartościach przesunięć chemicznych konieczne jest połączenie tych danych w celu uzyskania jak najbardziej wiarygodnego wyniku. W tym celu na podstawie wyznaczonych klastrów wybiera się graniczne wartości przesunięć pomiędzy tymi klastrami i tworzy z nich przedziały, którym przypisuje się wartość współczynnika istotności.

Kolejnym krokiem jest dodanie informacji uzyskanych dzięki obliczeniom największych różnic przesunięć chemicznych pomiędzy posortowanymi sygnałami. Jest to osiągane poprzez wybieranie części wspólnej z dotychczasowymi przedziałami. Operacja ta jest powtarzana dla wszystkich zbiorów zawierających określony procent najsilniejszych sygnałów. Część wspólna otrzymuje dodatkowy punkt oceny. Nienakładające się fragmenty przedziałów są usuwane, gdyż testy wykazały, że pozostawienie tych fragmentów może prowadzić do błędnych wyników. Z tak uzyskanego zbioru wybierane są środki tych przedziałów, które mają najlepsze wartości punktacji (rys. 47, 48).



Rys. 47. Przykładowe widmo 2D NOESY. Granatowymi liniami zaznaczone są regiony wyszukane przez program RNA NOEpathfinder



Rys. 48. Przykładowe widmo ^1H - ^{13}C HSQC . Granatowe linie wyznaczają regiony wyszukane przez program RNA NOEpathfinder

Wybieranie regionów na widmach

Obiekt klasy **RegionsFinder** zwraca wartości oddzielające poszczególne regiony na widmie. W celu uzyskania struktury słownikowej zawierającej sygnały podzielone na regiony, w której kluczami są numery poszczególnych regionów, konieczna jest procedura podziału na regiony. Zadanie to realizowane jest przez obiekt klasy **Spectrum** zawierający sygnały należące do analizowanego widma. W związku z tym, że RNA NOEpathfinder zaprojektowałem do pracy z praktycznie dowolną liczbą wymiarów, idealnym rozwiązaniem okazała się rekurencja. Dla każdego wymiaru na widmie, zbiór sygnałów jest dzielony na regiony w tym wymiarze. Każdy z tak uzyskanych zbiorów jest przekazywany dalej do podziału w kolejnym z wymiarów. Uzyskane zbiory sygnałów w postaci obiektów klasy **Peaks** zapisywane są pod odpowiednimi indeksami w strukturze słownikowej.

Żeby ułatwić podział na regiony, dla każdego zbioru widm wykonanych w tych samych warunkach, wartości oddzielające poszczególne regiony dla wymiarów protonowych są obliczane na podstawie widma 2D NOESY a następnie wykorzystywane we wszystkich pozostałych widmach wykonanych w tych samych warunkach.

Sygnały błędnie zakwalifikowane do danego regionu

Dla niektórych cząsteczek RNA może zdarzyć się sytuacja, że dla jednego lub kilku

sygnałów wartość przesunięcia chemicznego znajdzie się w rejonie charakterystycznym dla sygnałów pochodzących od innych jąder. W tym celu stworzyłem dodatkowy algorytm zaimplementowany w klasie **ImpositionsManager**, który na podstawie widm $^1\text{H}^{13}\text{C}$ HSQC w wielu przypadkach potrafi znaleźć sygnały od takich jąder i przenieść je do właściwych regionów na widmie. Metoda ta polega najpierw na wyszukaniu regionów na widmie $^1\text{H}^{13}\text{C}$ HSQC, w których nie powinno być sygnałów. Następnie program sprawdza, do którego sąsiadującego regionu jest temu sygnałowi najbliżej (procentowo) oraz czy znajduje się on blisko, czy daleko od sąsiadującego regionu. Wartość odcinającą wyznaczyłem na 10% szerokości danego regionu. Następnie korzystając z danych uzyskanych z korelowania ze sobą sygnałów z różnych regionów (który opisałem w kolejnym rozdziale) znajduwane są sygnały powiązane z danym sygnałem na różnych widmach. Sygnały te są następnie przenoszone do właściwych regionów na widmach. Metoda ta jest stosunkowo czuła na błędy i szумы, w związku z czym uwzględniane są tylko te sygnały na widmie HSQC, których intensywność jest większa niż 25% najsłabszych sygnałów na tym widmie.

IV.2.3 Korelacja sygnałów między widmami

Zbiór widm przekazanych do obliczeń nie musi pochodzić z pojedynczej serii eksperymentów. Eksperymenty mogą zostać wykonane w różnej temperaturze, z użyciem wody deuterowanej lub zawierającej mieszaninę H_2O i D_2O , z buforami zawierającymi różne sole lub w różnych stężeniach. Może to powodować, że wartości przesunięć chemicznych dla sygnałów na widmach wykonanych w różnych warunkach nie będą się na siebie idealnie nakładać. Możliwe są dwie główne sytuacje – przesunięcie położenia sygnałów dla pojedynczego jądra oraz przesunięcie wszystkich sygnałów na widmie. Pierwsza z wymienionych sytuacji może być pożądana, gdyż umożliwia rozdzielenie nałożonych na siebie sygnałów należących do odmiennych układów spinowych. Rozdzielenie sygnałów może umożliwić łatwiejsze przypisanie sygnałów oraz pozwolić na wyznaczenie większej ilości więzów strukturalnych. Niemniej stanowi ona wyzwanie z programistycznego punktu widzenia, gdyż uniemożliwia korelowanie sygnałów poprzez proste dopasowanie wartości na dwóch widmach.

W celu wyeliminowania tego problemu, stworzyłem algorytmy zaimplementowane w klasie **SpectralImposer**, pozwalające na korelowanie sygnałów z uwzględnieniem wyżej wymienionych problemów. Algorytm bazuje na założeniu, że wszystkie sygnały mogą zostać przesunięte o tą samą wartość w każdym z wymiarów, a ponadto każdy z sygnałów może być dodatkowo przesunięty o inną niezależną wartość w każdym z wymiarów na widmie.

Pierwszym etapem jest znalezienie wymiarów, w których wykorzystane są te same rodzaje jąder dla dwóch różnych widm n-wymiarowych. Dla tych widm następnie są sprawdzane informacje z pliku wewnętrznego, który odpowiada za rodzaje dostępnych widm. Znalezione regiony będą tymi które będą na siebie nakładane. W związku z tym kolejnym krokiem jest „spłaszczenie” widma poprzez połączenie odpowiednich regionów we wszystkich pozostałych

wymiarach dla obydwóch widm. Jeśli widmo nie ma wyznaczonych regionów, korelowane są wszystkie sygnały obecne na widmach.

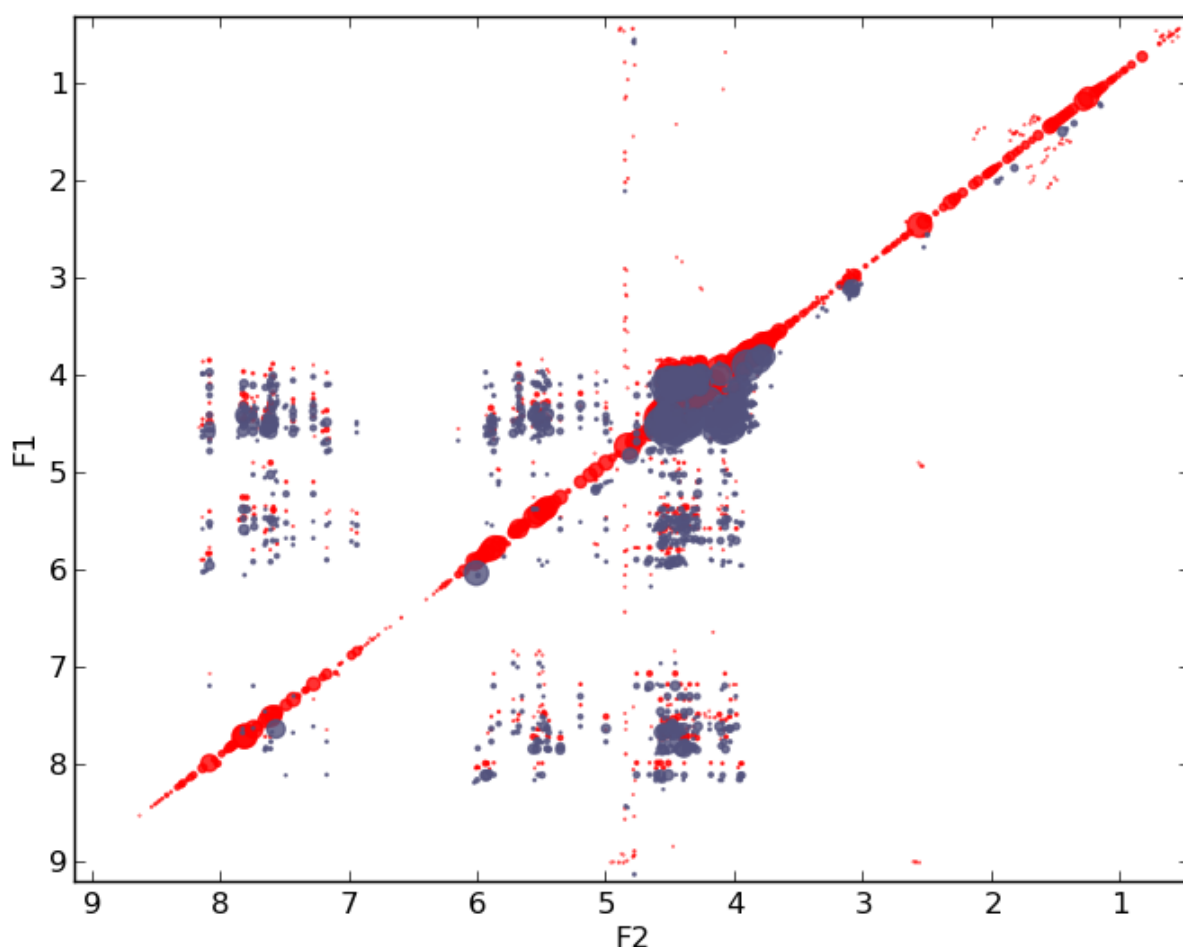
Jeśli widma są tego samego typu to wykonywany jest uproszczony proces normalizacji objętości. W tym celu obliczana jest całkowita objętość sygnałów na obydwu porównywanych regionach i obliczany jest mnożnik, jakiego należy użyć na wartościach objętości sygnałów jednego z widm, aby wartości były w miarę porównywalne.

Dla każdej pary sygnałów w obydwu analizowanych regionach (lub całych widmach), dla każdego z wymiarów, które są porównywane obliczana jest odległość pomiędzy nimi. Następnie obliczana jest „odległość” sygnałów na widmie dla każdej pary sygnałów z obu widm, liczona tak jakby korelowane sygnały znajdowały się na jednym widmie. Jeśli uzyskana wartość jest mniejsza niż współczynnik odcięcia zdefiniowany w parametrach eksperymentu, to tworzony jest wektor, o który będą przesunięte wszystkie pozycje sygnałów na jednym z widm podczas dalszych etapów nakładania sygnałów. Współczynnik odcięcia powinien być stosunkowo niewielki, żeby zmniejszyć możliwość kombinacji i jednocześnie odpowiednio duży, by móc uwzględnić rzeczywiste przesunięcie sygnałów na widmach.

Kolejnym krokiem jest ponowne obliczenie odległości uwzględniającej wektor przesunięcia, dla każdego wektora i każdej pary sygnałów widocznych na widmach. Jeśli widma są tego samego typu to następuje także porównanie względnych objętości sygnałów, gdyż odpowiadające sobie sygnały na widmach tego samego typu powinny mieć porównywalne wartości, co jest bardzo przydatne dla znalezienia odpowiadających sobie sygnałów. Jest to szczególnie przydatne w sytuacji jeśli proces eliminacji szumów z widma nie przebiegł prawidłowo i obok właściwego sygnału znajduje się FP. Wartość względnej objętości jest mnożona przez zadany współczynnik i dodawana do odległości. Wartości te są zapisywane w odpowiedniej strukturze słownikowej, gdzie kluczami są indeksy dla obu sygnałów. Dla każdego z sygnałów wybierany jest następnie ten, który ma najmniejszą wartość. Obliczana jest całkowita suma wartości i wybierane jest rozwiązanie posiadające najmniejszą sumę obliczonych wartości odległości sygnałów.

Podczas tworzenia algorytmu oraz w celu weryfikacji poprawności jego działania wykonałem serię testów. Obejmowały one korelowanie ze sobą widm 2D NOESY o różnych parametrach oraz widm ^1H - ^{13}C HSQC i DQF-COSY z widmami 2D NOESY. Testy te pokazały, że praktycznie niemożliwe jest korelowanie ze sobą widm 2D NOESY bez wykorzystania objętości lub intensywności sygnałów, a wartość współczynnika dla objętości może przełożyć się na błędne korelacje. W przypadku korelowania widm ^1H - ^{13}C HSQC z widmami 2D NOESY, kwestią kluczową okazały się szумы na pierwszym z tych widm.

W przyszłości można również rozwinąć wspomniany algorytm o korelację przesunięć



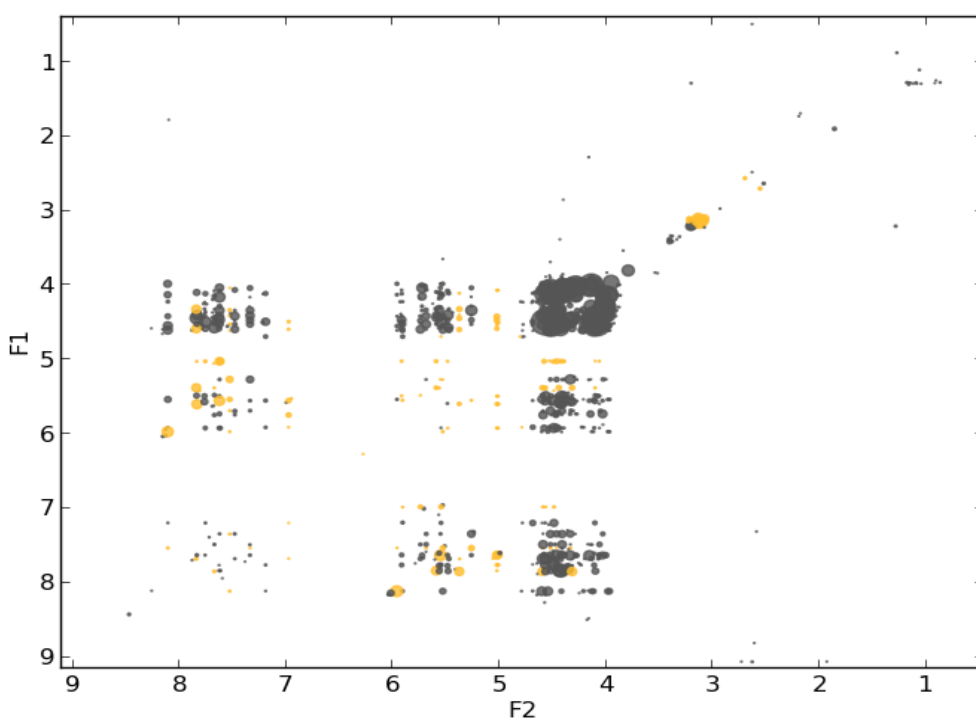
sygnałów, co może ulepszyć wyszukiwanie przesunięć całych systemów spinowych o określoną wartość w jednym z wymiarów.

Rys. 49. Nałożone na siebie dwa widma 2D NOESY dla wzajemnie komplementarnego oligonukleotydu. Kolorem czerwonym i niebieskim zaznaczone są sygnały na obu widmach, wyraźnie widać przesunięcie położenia sygnałów o niewielką wartość przesunięcia chemicznego.

Częściowe przypisanie sygnałów

Uzyskane korelacje sygnałów są wykorzystane do częściowego przypisania sygnałów,

gdyż w niektórych regionach widma ^1H - ^{13}C HSQC i DQF-COSY znajdują się tylko sygnały od jąder określonych grup atomów. Pozwala to przykładowo na znalezienie systemów spinowych tworzonych dla jąder H2 reszt adenozyń lub zidentyfikowanie sygnałów korelacyjnych H5-H6 na widmie NOESY. Wymaga to wstępnych przypisań grup sygnałów co opisałem dokładniej w rozdziale „Przypisanie sygnałów nie należących do ścieżek NOE”. Wstępne przypisanie sygnałów na widmie NOESY, realizowane w klasie **Assignmenter**, dzięki danym uzyskanym z klasy **ImpositionsManager**, jest praktycznie niezbędne dla poprawnej oceny ścieżek NOE, gdyż pozwala dopasować wygenerowaną ścieżkę do sekwencji nukleotydowej. Najpierw znajduwane są sygnały z widm HSQC i COSY, które zostały już przypisane i następnie sygnał na widmie NOESY mający najbardziej zbliżoną wartość przesunięcia chemicznego (wybrany dzięki korelowaniu widm ze sobą) zostaje przypisany na tej podstawie (rys. 50). Aby przypisania te były użyteczne do tworzenia ścieżek NOE, informacje z przypisanych sygnałów na widmach NOESY są przekazywane na całe układy spinowe, oraz nanoszone w obu wymiarach protonowych widma. Przypisania na całe układy spinowe nie są wykonywane, gdy daną wartość przesunięcia chemicznego posiadają sygnały pochodzące od różnych atomów.



Rys. 50. Widmo 2D NOESY dla wzajemnie komplementarnego oligonukleotydu. Kolorem żółtym zaznaczone są sygnały posiadające częściowe przypisanie

IV.2.4 Wybór optymalnego rodzaju ścieżki NOE do obliczeń

Na widmie NOESY można obserwować wiele sekwencyjnych ścieżek transferu magnetyzacji. Należy do nich między innymi ścieżka H1' – H6/H8 w rejonie *aromatyczno-*

anomerycznym. W przypadku wprowadzenia do badanej cząsteczki modyfikowanych reszt nukleotydowych, ścieżki takie muszą uwzględniać sygnały korelacyjne pochodzące od nowych atomów. Aby zapewnić maksymalną swobodę wyboru sekwencji nukleotydowej podczas korzystania z programu RNA NOE*pathfinder*, stworzyłem serię plików wewnętrznych oraz system klas, które procesują dane z tych plików. Klasy wraz z ich funkcją przedstawiłem w tabeli 5.

Jednym z wyzwań jakie napotkałem było stworzenie algorytmu, który byłby w stanie wyznaczyć optymalny rodzaj ścieżki NOE do rozpoczęcia obliczeń oraz zdefiniować reguły według których ta ścieżka ma być tworzona. Ostatecznie pierwszy z tych problemów rozwiązałem w trzech etapach:

- stworzenie możliwych rodzajów ścieżek NOE
- sprawdzenie, czy każdy rodzaj ścieżki jest poprawny
- ocena każdego rodzaju ścieżki NOE i wybranie optymalnego

Tabela 5. Klasy odpowiadające za przetwarzanie plików wewnętrznych programu oraz ich główne zadania.

Nazwa klasy	Funkcja
TemplateAtomType	Reprezentacja poszczególnych pierwiastków chemicznych
TemplateResidue	Reprezentacja poszczególnych reszt nukleotydowych
TemplateAtomInResidue	Reprezentacja atomów zajmujących określone pozycje w danych resztach nukleotydowych wraz z typowymi dla nich wartościami przesunięć chemicznych
TemplateSpectrum	Reprezentacja widm oraz ich właściwości i typu atomów obserwowanych na nich w każdym z wymiarów
TemplateRegions	Wyznaczenie możliwych podziałów widm na regiony, co szerzej opisałem w rozdziale „Wyszukanie regionów na widmach”
TemplateVisibleCrosspeaksFromResidues	Przetwarzanie informacji o sygnałach, które można obserwować na danym typie widm NMR
TemplateTypicalConnections	Przetwarzanie informacji o odległościach międzyatomowych dla poszczególnych par atomów
TemplatePathways	Wyszukanie możliwych ścieżek NOE oraz wybór najbardziej optymalnej do rozpoczęcia obliczeń
TemplatePathwayRules	Stworzenie reguł wyboru i oceny ścieżek NOE oraz ich fragmentów
TemplateAnalyser	Organizacja sekwencyjnego przetwarzania danych przez wyżej wymienione klasy

Etap tworzenia możliwych rodzajów ścieżek NOE to wyszukanie zestawów atomów dla każdej reszty nukleotydowej na podstawie informacji uzyskanych dzięki instancji klasy **TemplateTypicalConnections**. Stworzenie pełnej listy wszystkich możliwych typów ścieżek okazało się bardzo złożone obliczeniowo, w związku z czym wprowadziłem pewne ograniczenie, a mianowicie, że jeden z typów atomów obecnych w ścieżce musi należeć do zbioru (H1', H2', H3', H4', H5'). Drugi etap polega na weryfikacji, czy w każdej z potencjalnych

ścieżek znajduje się element łączący dwie dowolne reszty. Następnie każdy rodzaj ścieżki NOE jest oceniany na podstawie tego, czy znajduje się w regionie w którym jest mało sygnałów, oraz czy wszystkie sygnały korelacyjne w zaproponowanej ścieżce należą do jednego regionu. Przeprowadzone przeze mnie testy pokazały, że spełnione są założenia i dla niemodyfikowanych reszt nukleotydowych wybierana jest ścieżka H1'-H6/H8 w regionie *aromatyczno-anomerycznym*.

Znacznie trudniejszym zadaniem okazało się stworzenie spójnych i prostych reguł tworzenia i oceny ścieżek NOE dla wybranego rodzaju ścieżki. Reguły te kilkakrotnie zmieniałem dopasowując do wymagań programu w danej chwili oraz upraszczając ich formę. Ostatecznie przyjęły formę słownika, w którym kluczami są nazwy poszczególnych reguł, a wartości określają nazwy reszt, atomów oraz pozostałe parametry danej reguły. Istnieją dwa główne rodzaje reguł: dodaj sygnał posiadający wskazane przypisania oraz nie dodawaj sygnału posiadającego dane przypisania. Przykładową regułą jest nie włączanie do ścieżek NOE w regionie *aromatyczno-anomerycznym* sygnałów, które w jednym z wymiarów są przypisane jako sygnały pochodzące od protonów H2 adenozy.

Kolejnym parametrem, który może być wykorzystywany do oceny sekwencyjnych ścieżek NOE jest intensywność sygnałów wchodzących w skład ścieżki, a raczej rozkład intensywności w obrębie ścieżki. Sygnały korelacyjne pomiędzy jądrami atomów należących do tej samej reszty są zazwyczaj znacznie silniejsze od sygnałów korelacyjnych pomiędzy jądrami atomów należących do kolejnych reszt nukleotydowych. W związku z tym program sprawdza, czy różnica pomiędzy średnią wartością wzorcowych odległości dla obu typów sygnałów jest większa niż 10% średnich wartości wzorcowych odległości dla wszystkich sygnałów wchodzących w skład ścieżki NOE. Wynik ten jest przekazywany w programie jako reguła „strongerweaker” i wykorzystywany podczas oceny zarówno fragmentów jak i całych ścieżek NOE.

Do oceny są również wykorzystywane sygnały z innych regionów. Jeśli dla kolejnych dwóch sygnałów wchodzących w skład ocenianej ścieżki NOE istnieje sygnał korelacyjny w innym regionie (przede wszystkim w regionie *aromatyczno-aromatycznym*), to te dwa sygnały z dużo większym prawdopodobieństwem stanowią kolejne elementy rzeczywistej ścieżki NOE. W programie RNA NOEpathfinder w takiej sytuacji dodawane są dodatkowe punkty do całkowitej oceny. Wykorzystanie sygnałów informacji pochodzącej z innych regionów wiąże się jednak z dodatkowymi problemami. W przypadku jeśli sygnał korelacyjny nie został zaznaczony, lub kolejne elementy ścieżki NOE mają bardzo podobną wartość przesunięcia chemicznego, program może gorzej ocenić niektóre fragmenty ścieżek NOE tylko ze względu na to, że dla niektórych ocenianych fragmentów ścieżek istnieje więcej sygnałów korelacyjnych. Z tego powodu w sytuacji gdy na widmie widocznych jest wiele nakładających się na siebie sygnałów – lepiej jest zmniejszyć liczbę punktów przyznawanych za znalezienie sygnału korelacyjnego w innym regionie widma.

Do oceny ścieżek NOE mogą również zostać wykorzystane sygnały korelacyjne z tego

samego regionu, które już zostały przypisane do danego typu jądra atomowego. Przykładowo sygnały, które zostały przypisane do jąder atomów H2 w regularnych dupleksach RNA mogą tworzyć korelację do atomów H1' tej samej (i), następnej (i+1), naprzeciwległej (j) oraz kolejnej po naprzeciwległej (j+1) reszty. Można tą informację wykorzystać do oceny danej ścieżki NOE.

IV.2.5 Generowanie ścieżek w rejonie *aromatyczno-anomerycznym* widma NOESY

Pierwszym etapem jest częściowe przypisanie sygnałów na widmie NOESY, które opisałem w rozdziale „Korelacja sygnałów między widmami”. W rejonie *aromatyczno-anomerycznym* takimi sygnałami są sygnały korelacyjne H5-H6 pirymidyn oraz sygnały pochodzące od protonów H2 adenozyń. Przypisanie to odbywa się w dwóch etapach. W pierwszym z nich przypisywane są te sygnały, dla których istnieje korelacja z innym widmem, natomiast w drugim etapie przypisywane są te sygnały, które mają takie same wartości przesunięć chemicznych w wymiarze, dla którego są przypisane. W przypadku sygnałów których w jednym z wymiarów wartości przesunięć chemicznych są bardzo zbliżone, sygnały te nie są przypisywane ze względu na możliwość pomyłki. Wyjątkiem jest sytuacja, gdy analizowane są korelacje pomiędzy sygnałami na widmach DQF-COSY i NOESY. W tym przypadku bezpośrednio skorelowane sygnały są przypisywane.

Kolejnym etapem jest wybranie możliwych połączeń pomiędzy poszczególnymi atomami, które będą mogły wejść w skład nowo generowanych ścieżek NOE. Tworzony jest słownik, gdzie dla każdego sygnału wybierane są takie sygnały, które w danym wymiarze nie przekraczają 1.3 szerokości danego sygnału. Wartość ta zapewnia stworzenie połączenia w przypadku nałożonych sygnałów o niecharakterystycznym kształcie. Dopasowanie w tym wymiarze jest zapisywane w słowniku, co jest wykorzystywane w późniejszych etapach tworzenia widma. Co więcej na tym etapie na podstawie danych dotyczących wartości przesunięć chemicznych dla których może występować nakładanie sygnałów oraz wcześniejszych przypisań, eliminowane są sygnały, które nie wchodzą w skład ścieżki NOE. W przypadku wartości przesunięć chemicznych dla których może występować nakładanie się sygnałów, tworzone jest dane połączenie, by uniknąć sytuacji w której prawidłowa ścieżka NOE nie zostanie wygenerowana. Z tworzonych połączeń wykluczane są te, dla których istnieją już częściowe przypisania, które eliminują je ze ścieżki NOE (H2 adenozyń, H5 pirymidyn). Wyjątek stanowią połączenia, dla których może występować nakładanie się sygnałów. Rozwiązanie takie znacznie ułatwia dalszą analizę oraz zmniejsza liczbę obliczeń, gdyż nie trzeba każdorazowo porównywać wartości przesunięć chemicznych dla dwóch sygnałów by sprawdzać, czy mogą one tworzyć kolejne elementy ścieżki NOE. Na tym etapie niezwykle istotne jest aby zaznaczone sygnały dla widma były pozbawione wszystkich lub prawie wszystkich szumów, gdyż fałszywe sygnały pochodzące od szumów mogą spowodować znaczne zwiększenie czasu obliczeń dla dalszych etapów analizy.

Kolejnym etapem jest wygenerowanie krótkich fragmentów ścieżek NOE, domyślnie łączących 7 sygnałów należących do ścieżki. W tym celu program wybiera najsilniejsze sygnały,

które nie zostały wcześniej wykluczone ze ścieżki NOE, ani nie są przypisane do jąder atomów H5 i H6 pirymidyn. Dodatkowo dla każdego z takich fragmentów tworzone są specjalne słowniki ukazujące możliwości dalszych połączeń po obu stronach danego fragmentu. Ze względu na dużą możliwość powstałych kombinacji, część najgorzej ocenionych rozwiązań jest odrzucana. Dla etapu tworzenia krótkich fragmentów przyjąłem założenie, że dany sygnał nie może się pojawiać w danym fragmencie dwukrotnie, nawet w przypadku możliwego nakładania sygnałów. W pewnych rzadkich sytuacjach może to powodować problemy ze znalezieniem ścieżki, jednak użytkownik oprogramowania może je rozwiązać poprzez ręczną korektę i uzupełnienie sygnałów obecnych na widmie lub zmianę parametrów. Do oceny tych fragmentów wykorzystywana jest funkcja oceniająca, która uwzględnia różnice w wartościach przesunięć chemicznych dla sygnałów (artefakty automatycznego wyszukiwania sygnałów), możliwą ilość sygnałów dla danej wartości przesunięcia chemicznego oraz to czy kolejne sygnały wchodzące w skład fragmentu są na przemian słabsze i silniejsze. Ponadto sprawdzane jest, czy dla danej pary sygnałów w ścieżce w innym regionie istnieje sygnał posiadający takie same wartości przesunięć chemicznych, jak wartości przesunięć chemicznych, którymi różnią się kolejne elementy ścieżki NOE.

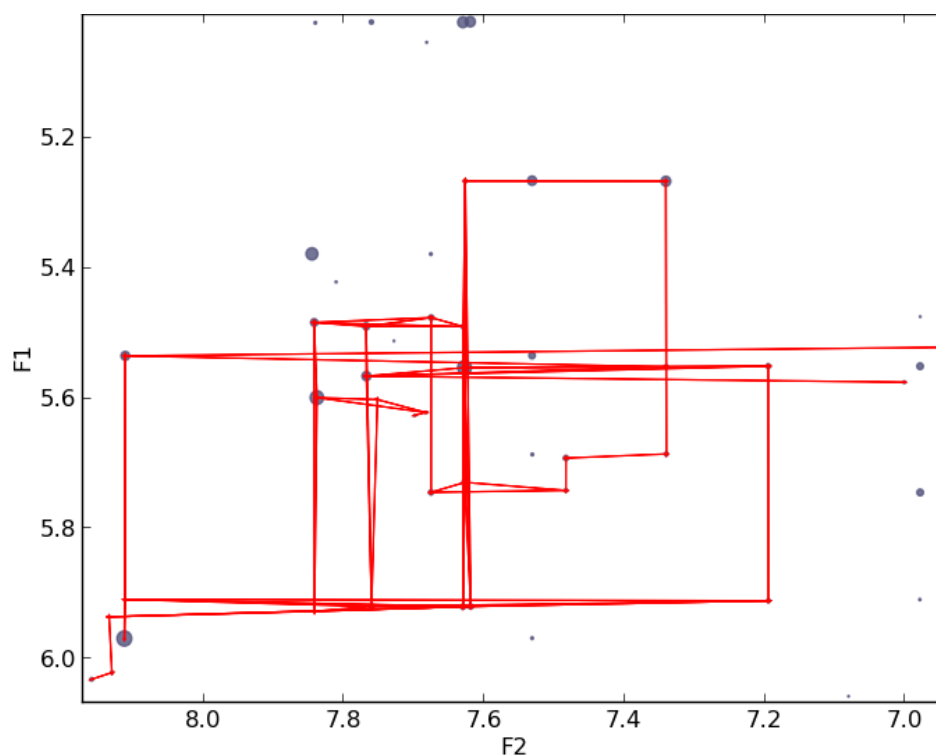
Tak wygenerowane krótkie fragmenty są następnie dopasowywane do sekwencji nukleotydowej. W tym przypadku wykorzystywana jest inna funkcja oceniająca, która oprócz wyżej wymienionych elementów uwzględnia również to czy dla danego przesunięcia chemicznego istnieje sygnał H5/H6 pirymidyn, co pozwala na wyeliminowanie tych dopasowań, gdzie tej wartości przesunięcia chemicznego odpowiadała by puryna. Ponadto sprawdzane są inne sygnały w regionie *aromatyczno-anomerycznym* dla danej wartości przesunięć chemicznych oraz to czy na podstawie danych z plików wewnętrznych programu, są one typowymi sygnałami. Ponieważ możliwości przypisać jest kilkakrotnie więcej niż samych fragmentów ścieżek, na tym etapie ponownie odrzucana jest część najgorszych rozwiązań.

Następnym etapem jest tworzenie bloków złożonych z dwóch dopasowanych fragmentów ścieżki NOE, pomiędzy którymi brakuje pewnej liczby sygnałów. Jest to uzyskane poprzez kombinatoryczne dopasowywanie wszystkich fragmentów, a następnie dzięki wcześniej wygenerowanym słownikom możliwych połączeń znajduwane są możliwości uzupełnienia luki pomiędzy fragmentami. Nowo powstały większy fragment ścieżki jest również oceniany, aby wyeliminować jak najwięcej błędnych rozwiązań na tym etapie.

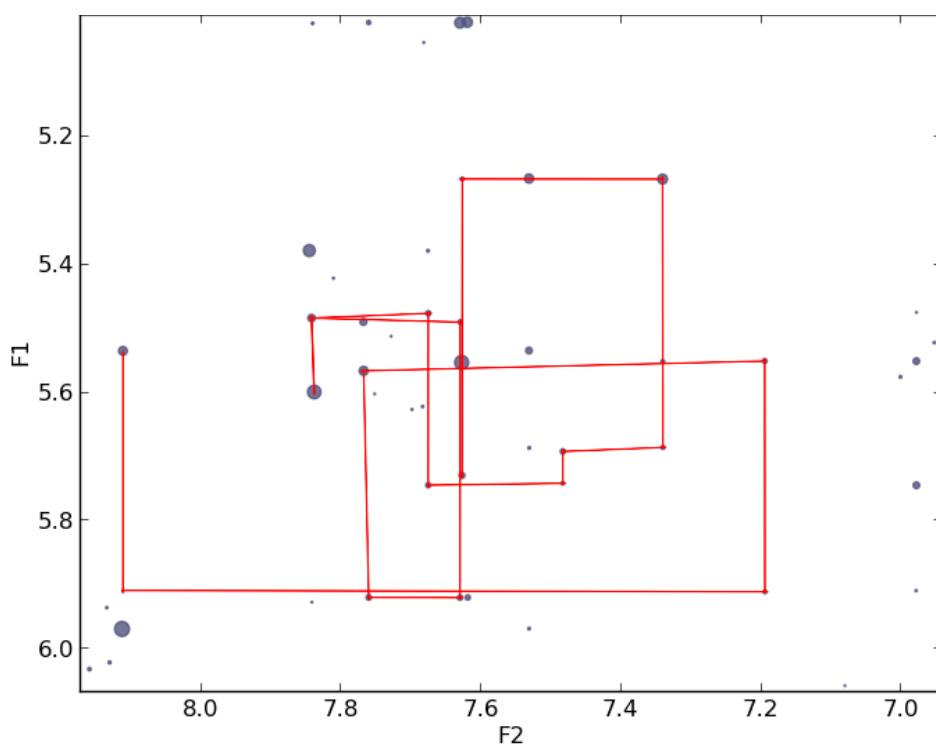
W kolejnym kroku tak utworzone bloki są łączone ze sobą – na zasadzie pierwszy element następnego bloku jest drugim elementem poprzedniego bloku. Procedura ta jest powtarzana tak długo, aż nie dojdzie do sytuacji, gdzie na końcach nie pozostanie do uzupełnienia kilka sygnałów. W tym momencie bloki są ze sobą łączone i ponownie oceniane, a następnie uzupełniane są brakujące elementy na końcach, po czym znów następuje ocena ścieżek. Tak powstałe potencjalne ścieżki NOE są do siebie dopasowywane, by ostateczne rozwiązanie obejmowało wszystkie możliwe do wyznaczenia ścieżki na widmie. Na tym etapie

ocenie poddawana jest tylko liczba elementów wszystkich ścieżek mająca takie same wartości przesunięć chemicznych.

A



B



Rys. 51. Region *aromatyczno-anomeryczny* widma 2D NOESY dla wzajemnie komplementarnego oligonukleotydu. A – kolorem czerwonym zaznaczono wszystkie możliwe połączenia pomiędzy sygnałami, B – kolorem czerwonym zaznaczono połączenia wchodzące w skład rzeczywistej ścieżki NOE

IV.2.6 Przypisanie sygnałów nie należących do ścieżek NOE

Po przypisaniu sygnałów należących do ścieżek NOE możliwe jest również przypisanie innych sygnałów- nie należących do ścieżek NOE. W tym celu program RNA NOE*pathfinder* wykonuje wyszukiwanie w dwóch krokach. W pierwszym wyszukuje wartości przesunięć chemicznych, dla których nie ma pełnego przypisania sygnałów, następnie tworzy możliwe przypisania osobno dla każdego sygnału, następnie sprawdza najlepsze możliwości dla całego przesunięcia chemicznego. Tak uzyskane zestawy przypisań są następnie kombinowane ze sobą i wybierane jest najlepsze rozwiązanie. Dzięki temu można na przykład przypisać sygnały pochodzące od H2 adenozyń. W drugim kroku przypisywane są pozostałe sygnały, przy czym w przypadku możliwości przypisania danego sygnału do różnych atomów, wybierane jest rozwiązanie, dla którego istnieje najbardziej zbliżony wpis w plikach wewnętrznych programu.

Na chwilę obecną funkcjonalność ta jest przetestowana jedynie dla regionu *aromatyczno-anomerycznego*, gdzie przypisywane są w szczególności sygnały pochodzące od jąder atomów H2-H1', sygnały H5-H6/H8 pochodzące od różnych reszt, a sygnały H5-H6 pirymidyn są przypisywane do danej reszty. W pewnych szczególnych przypadkach przypisanie wszystkich sygnałów może okazać się niemożliwe. Dotyczy to przede wszystkim sytuacji, w których nieprzypisane sygnały pochodzące od różnych jąder atomowych mają bardzo podobne wartości przesunięć chemicznych.

Na podstawie przypisań oraz znanych objętości sygnałów, można stworzyć więzy odległościowe. Podstawą do utworzenia więzów odległościowych jest porównanie objętości danego sygnału oraz sygnałów wzorcowych, dla których odległość pomiędzy atomami jest względnie stała. Tworzenie oraz przekształcanie więzów można wykonać za pomocą wcześniej stworzonego przeze mnie programu TotalRestrains, opisanego w rozdziale „Programy towarzyszące”.

IV.2.7 Generowanie ścieżek w innych regionach widma NOESY

Na podstawie przypisań sygnałów znajdujących się w regionie aromatyczno-anomerycznym możliwe jest wyznaczenie ścieżek w innych regionach widma NOESY. Najprostszym przykładem jest ścieżka H1'-H2', która w większości przypadków zawiera najsilniejsze sygnały w regionie H1'-H2'/H3'/H4'/H5'. Program RNA NOE*pathfinder* wyszukuje sygnały, które na podstawie korelacji z widmem ¹H-¹³C HSQC pochodzą od sygnałów H2' oraz pochodzą od jąder sygnałów, których odległość w badanej cząsteczce jest porównywalna do wartości z plików wewnętrznych programu. Na tej podstawie tworzona jest ścieżka NOE w tym regionie. W podobny sposób mogą być wykonane próby wygenerowania innych ścieżek NOE. Na chwilę obecną metoda ta nie jest przetestowana ani udostępniona w wersji online programu.

IV.2.8 Analiza widm 3D

Do analizy większych cząsteczek RNA, ze względu na silne nakładanie się sygnałów na widmach 2D konieczne jest wykorzystanie cząsteczek znakowanych izotopem węgla ^{13}C lub ^{15}N , które umożliwiają wykonywanie widm wielowymiarowych. Dzięki temu uzyskujemy między innymi rozseparowanie sygnałów na widmach, co znacznie ułatwia analizę.

Oprogramowanie RNA *NOEpathfinder* zaprojektowałem tak, by umożliwić wykorzystanie danych pochodzących z widm wielowymiarowych a w szczególności z widm 3D NOESY-HSQC. W tym celu wszystkie klasy i metody stworzyłem w ten sposób aby umożliwić analizę danych wielowymiarowych. Dopasowanie to polega między innymi na każdorazowym sprawdzaniu w których wymiarach zebrane zostały dane dotyczące którego rodzaju jądra, oraz wykorzystaniu metod rekurencyjnych dla celów dzielenia widma na regiony oraz wybierania danych z danego regionu. Większość informacji specyficznych dla widm wielowymiarowych jest analizowane we wstępnych etapach przetwarzania danych z widm. Na etapie tworzenia ścieżek NOE oraz przypisywania sygnałów, algorytmy są praktycznie identyczne z tymi dla widm dwuwymiarowych. Niemniej na chwilę obecną program wymaga wielu testów na rzeczywistych danych pochodzących z widm N-wymiarowych oraz poprawek i udoskonaleń które okażą się niezbędne podczas tych testów.

IV.3 Integracja oprogramowania RNA *NOEpathfinder* i RNAComposer

Stworzony przeze mnie program ma na celu ułatwienie rozwiązywania problemów związanych z rozwiązywaniem struktur przestrzennych cząsteczek RNA w oparciu o dane uzyskane dzięki spektroskopii NMR. Aby móc lepiej realizować ten cel, można połączyć zalety modelowania cząsteczek RNA w oparciu o znane struktury zaimplementowane w programie RNAComposer (51, 52) stworzonym w Zespole Struktury i Funkcji RNA oraz analizy danych eksperymentalnych z wykorzystaniem programu RNA *NOEpathfinder*. Integracja jest możliwa na dwa zupełnie różne sposoby, opisane poniżej.

Wykorzystanie modeli przestrzennych dla celów wyznaczania ścieżek NOE

RNA *NOEpathfinder* potrzebuje informacji o sekwencji oraz strukturze drugorzędowej cząsteczki. Dane te mogą zostać wprowadzone w formacie Dot-Bracket. Niemniej program RNA *NOEpathfinder* pozwala na wykorzystanie w tym celu modelu cząsteczki zbudowanego w programie RNAComposer lub za pomocą innych metod modelowania molekularnego. Koordynaty przestrzenne muszą być dostarczone w formacie PDB. Dzięki wykorzystaniu klas oraz danych z programu *PDBinout* opisanego w rozdziale „Programy towarzyszące”, program RNA *NOEpathfinder* jest w stanie przeanalizować plik dostarczony w dowolnej wersji formatu PDB. Plik PDB jest wczytywany przez obiekt klasy **PDBReader**, który generuje obiekt klasy **PDB**. Wygenerowany obiekt posiada złożoną strukturę słownikową w której przechowywane są

obiekty klasy **Residue**, które z kolei zawierają obiekty klasy **Atom**. Ta złożona struktura umożliwia uwzględnienie poszczególnych modeli, łańcuchów lub fragmentów łańcuchów podczas przetwarzania plików PDB. Następnie dzięki metodom zaimplementowanym w klasie **FormatDetector**, na podstawie danych wczytanych przez klasę **FormatReader** wyszukiwany jest schemat nazewnictwa poszczególnych atomów, najbardziej zbliżony do formatu wejściowego pliku PDB. Następnym etapem jest wyszukanie par zasad. Odbywa się to na prostej zasadzie zmierzenia odległości między kluczowymi atomami zbudowanego modelu PDB, a następnie przeszukaniu listy odległości w celu znalezienia określonych wartości które użyłem do wyszukania określonych par zasad. W ten sposób można uzyskać rzeczywisty obraz tworzonych par zasad oraz odległości międzyatomowych w cząsteczce kwasu nukleinowego. Wartości odległości obliczone na podstawie wstępnego modelu mogą w przyszłości posłużyć do automatycznego wyznaczenia ścieżek NOE dla struktur zawierających wybrzuszenia lub rejony jednoniciowe. Z kolei wartości odległości międzyatomowych mogą zostać wykorzystane do ulepszenia funkcji oceniających. Podobne rozwiązanie zostało zaproponowane dla algorytmu FLYA, gdzie struktury uzyskane z programu RNAComposer zostały wykorzystane jako dane wejściowe do obliczeń (100).

Wykorzystanie więzów z danych eksperymentalnych w oprogramowaniu RNAComposer

Oprogramowanie RNAComposer pozwala na wprowadzanie danych eksperymentalnych w postaci więzów strukturalnych. Dane uzyskane z programu RNA NOEpathfinder można wykorzystać do stworzenia więzów odległościowych, które następnie można wykorzystać do wygenerowania struktury, co przełoży się na jakość modelowania, w szczególności na etapie udokładniania struktury przestrzennej.

IV.4 Testy oprogramowania

Program RNA NOEpathfinder wymagał bardzo starannego testowania. Odbywało się ono na czterech poziomach:

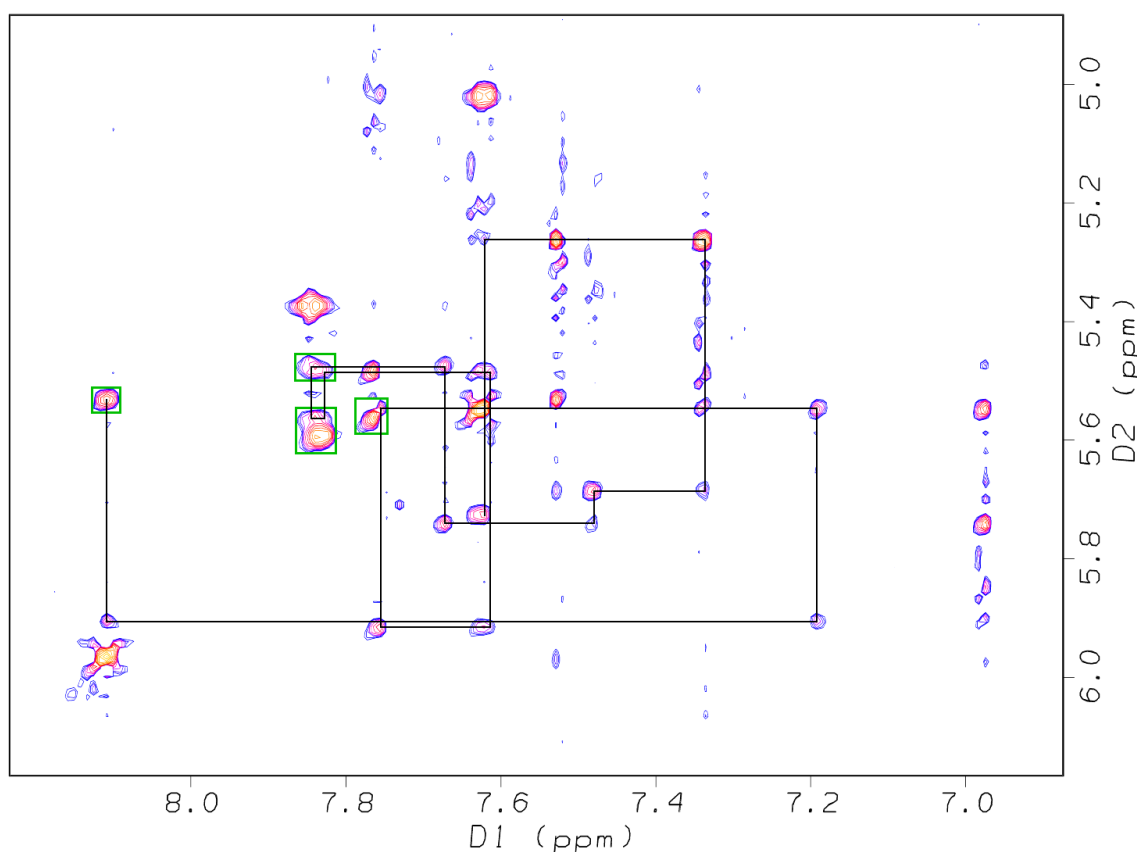
- testowanie klas i metod
- testowanie zintegrowanych mechanizmów
- testowanie całego programu dla danych testowych
- testowanie interfejsu użytkownika

W celu przetestowania poszczególnych klas i metod, stworzyłem szereg testów jednostkowych (ang. *unittest*). Za pomocą tych testów sprawdziłem możliwie dużo kombinacji danych wejściowych i wyników obejmujące losowe przypadki oraz wartości skrajne. Testowanie zintegrowanych mechanizmów wykonałem częściowo w oparciu o mechanizmy testów jednostkowych a częściowo ręcznie badając wpływ poszczególnych elementów na działanie oprogramowania. Testy interfejsu webowego wykonałem poprzez szczegółową ręczną analizę

poszczególnych stron systemu.

Testy całego oprogramowania dla danych testowych zawierających widma 2D wykonałem wykorzystując widma i przypisania dla małych cząsteczek RNA, uzyskane w przez Dorotę Gudanis oraz przeze mnie w Pracowni Biomolekularnego NMR w naszym Instytucie. Testy wybranych algorytmów testowałem również na cząsteczkach zbadanych przez Łukasza Popendę oraz Magdalenę Małgowską. Całościowe testy oprogramowania przeprowadziłem dla dwóch cząsteczek:

1. Cząsteczka 1 - wzajemnie komplementarny dupleks z wystającymi 3' końcami o sekwencji 5' – CAGAUCUGAGC – 3', testy na danych surowych
2. Cząsteczka 2 - wzajemnie komplementarny dupleks o sekwencji 5' – GCAGCUGC – 3', testy na danych poddanych wstępnemu przygotowaniu.



W przypadku testów dla cząsteczki 1 testowałem przede wszystkim możliwość znalezienia prawidłowej ścieżki w przypadku kiedy nie są zaznaczone wszystkie sygnały na widmie 2D – NOESY (ze względu na silne nałożenie sygnałów) oraz występuje wiele szumów błędnie zakwalifikowanych jako sygnały. Specjalnie przygotowane testy wykazały, że napisany przeze mnie program pozwala na wygenerowanie rzeczywistej ścieżki NOE. Niestety testy te ukazały również problemy i ograniczenia. Ponieważ dane zawierały szumy błędnie zakwalifikowane jako sygnały, a nie zawierały wszystkich sygnałów widocznych na widmach,

konieczne było umożliwienie wielokrotnego wykorzystania tego samego sygnału jako elementu ścieżki. W ten sposób znacząco zwiększyła się ilość możliwych kombinacji na każdym z etapów generowania ścieżki. Ponadto funkcje oceniające nie mogły nadawać dużych kar za wielokrotne wykorzystanie tego samego sygnału, ani w dużym stopniu uwzględniać różnicy w wartościach przesunięcia chemicznego pomiędzy kolejnymi elementami ścieżki, gdyż powodowałoby to znacznie niższą ocenę rzeczywistej ścieżki (rys. 52). Brak tych elementów oceny wymusił obniżenie kryteriów wyboru najlepszych krótkich fragmentów ścieżki NOE, co przełożyło się na dalsze zwiększenie liczby analizowanych kombinacji. Ze względu na ograniczone zasoby sprzętowe przeznaczone do obliczeń (w szczególności pamięć RAM) musiałem zdefiniować dodatkowe ograniczenia, które eliminowały część nieprawidłowych fragmentów ścieżki NOE. W ten sposób udało się uzyskać listę możliwych ścieżek NOE, wśród których znalazła się ta właściwa. Wartość nadana prawidłowej ścieżce NOE była niższa niż znaczącej liczby innych, nieprawidłowych ścieżek NOE. Liczne i czasochłonne próby optymalizacji funkcji oceniającej nie dały w tej kwestii znaczącej poprawy.

Rys. 52. Rejon *aromatyczno-anomeryczny* widma NOESY dla cząsteczki 1. Kolorem czarnym zaznaczona jest ścieżka NOE, a zielonymi prostokątami nałożone na siebie grupy sygnałów w tym sygnały wchodzące w skład ścieżki NOE.

Doświadczenia i problemy z testami dla cząsteczki 1 skłoniły mnie do zmiany części założeń oraz uproszczenia testów. Z tego powodu w przypadku testów dla cząsteczki 2 testowałem możliwość uzyskania szybkich wyników w przypadku gdy dane zostały wstępnie przygotowane. Główną zmianą której dokonałem była edycja regionu *aromatyczno-anomerycznego* na widmie 2D NOESY, w obrębie którego dokonałem ręcznego zaznaczenia brakujących sygnałów oraz usunięcia nadmiarowych. Ze względu na niewielką długość cząsteczki zdecydowałem się również na zwiększenie długości generowanych fragmentów ścieżki oraz wykluczenie w obliczeniach możliwości tworzenia ścieżek w których jeden sygnał pojawia się więcej niż jeden raz. Dla cząsteczki 2 udało się uzyskać prawidłową ścieżkę NOE, a wartość nadana przez funkcję oceniającą była wyższa niż dla pozostałych wygenerowanych ścieżek NOE.

IV.5 Webserwer RNA NOEpathfinder

IV.5.1 Schemat działania webserwisu

Jak wspomniałem wcześniej, dostęp do oprogramowania RNA NOEpathfinder odbywa się poprzez stronę internetową pod adresem <http://rnanoepathfinder.ibch.poznan.pl> (rys 53).

RNA NOEpathfinder

Institute of Bioorganic Chemistry, Polish Academy of Sciences
Department of Structural Chemistry and Biology of Nucleic Acids
Group of RNA Structure and Function

Home
Tools
Help
About
References
Links
Contact us

Welcome to RNA NOEpathfinder, automated NMR spectra analysis tool for RNA molecules.

RNA NOEpathfinder is a tool for automatic analysis of multidimensional NMR data for RNA and other nucleic acids. Tool allow analysis of various spectra including **NOESY**, **DQF-COSY**, **^1H - ^{13}C HSQC** and **NOESY-HSQC**.

Analysis includes: automatic noise detection and elimination, detection of regions, imposition of different spectra one into another, generation of NOE pathways, assignment of cross-peaks and finally preparation of results for structure calculation.

Necessary data: peaklists and volumes/intensities for NOESY 400 ms mixing time, HSQC (or NOESY-HSQC spectrum), DQF-COSY and secondary structure in dot-bracket format (or preliminary 3D model of the molecule).

Additional data: NOESY spectra performed in other temperatures (to allow distinguishment of imposed cross-peaks), NOESY 150 ms (or similar) mixing time for restraints generation, user defined "internal files" and parameters.

All spectra should contain high signal to noise ratio, and if possible should contain all existing peaks. Accepted peak formats: **Felix** (Accelrys), **Topspin** (Bruker).

To use RNA NOEpathfinder registration and login is required.

Your account

User ID:
Password:

[Forgot your password?](#)
[Create an account](#)

Number of visitors 0
Visitors online 1
Number of users 2
Total number of experiments 4
Number of currently submitted experiments 4

Current jobs

User name	Status	Creation date
anonymous	waiting	Oct. 14, 2013, 11:26 a.m.
anonymous	waiting	Oct. 2, 2013, 9:09 a.m.
anonymous	waiting	Sept. 23, 2013, 12:55 p.m.
anonymous	waiting	Sept. 23, 2013, 8:47 a.m.

[Terms of use](#) [Privacy policy](#) [Contact us](#) created by **Tomasz Woźniak**

Rys. 53. Schemat graficzny webserwisu RNA NOEpathfinder na przykładzie strony startowej.

Aby użytkownik mógł uruchomić własne obliczenia, niezbędna jest jego rejestracja poprzez stronę internetową. Podczas rejestracji należy podać hasło oraz adres e-mail, na który zostanie wysłany link aktywacyjny. Ponadto należy podać swoje imię, nazwisko oraz nazwę instytucji oraz wybrać opcję czy chce się pozostać anonimowym w systemie. Formularz rejestracyjny jest zabezpieczony kodem CAPTCHA (ang. *Completely Automated Public Turing test to tell Computers and Humans Apart*), który zapobiega masowemu tworzeniu kont przez roboty, co utrudniałoby funkcjonowanie serwisu. (rys. 54).

Wszystkie dane, włącznie z hasłem użytkownik może w każdej chwili zmienić

i uaktualnić. W przypadku zapomnienia hasła użytkownik może je zmienić korzystając

RNA NOEpathfinder webserver - registration

e-mail:

password:

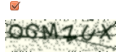
repeat password:

name:

surname:

university/company/private:

anonymous: ☐

security code: 

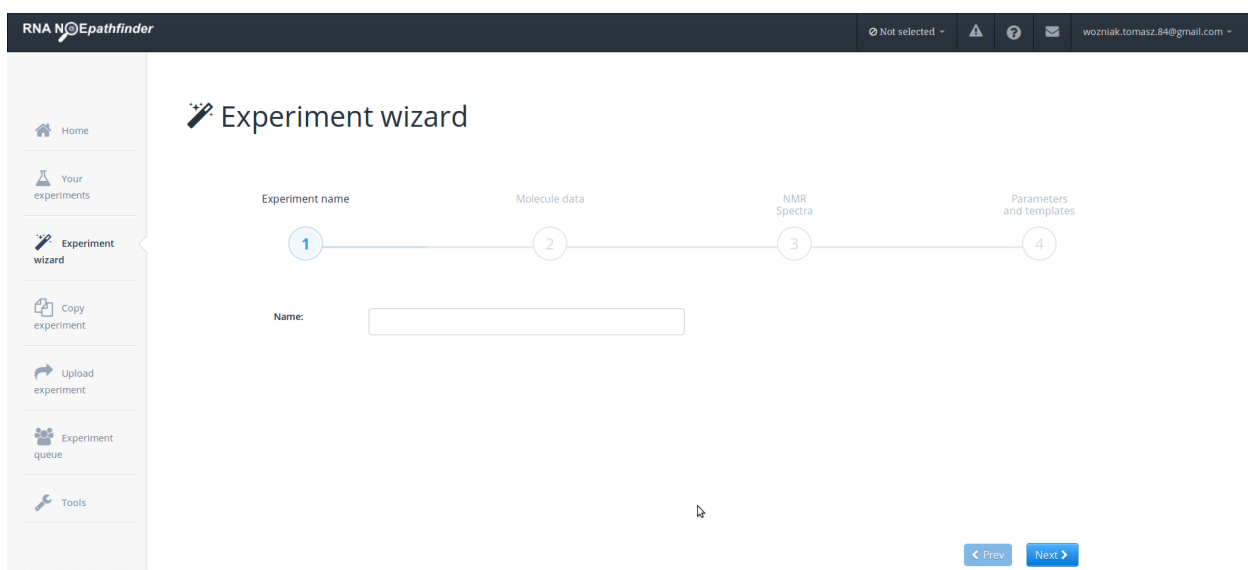
☐ By registering I accept: [terms of use](#) and [privacy policy](#)

Rys. 54. Formularz rejestracyjny w serwisie RNA NOEpathfinder.

z generowanego linku przesyłanego na adres e-mail podany podczas rejestracji.

Po zalogowaniu, użytkownik kierowany jest na własną stronę startową, gdzie może zobaczyć główne parametry swojego konta. Aby rozpocząć obliczenia trzeba stworzyć nowy eksperyment. Eksperyment można stworzyć na trzy różne sposoby: Poprzez kreatora eksperymentu, mechanizm wczytywania danych z plików archiwum oraz poprzez skopiowanie istniejącego eksperymentu użytkownika.

Kreator eksperymentu pozwala na ręczne dodanie informacji o strukturze drugorzędowej lub przestrzennej cząsteczki, listy sygnałów obecnych na widmach oraz ustalenia parametrów startowych (rys. 55).

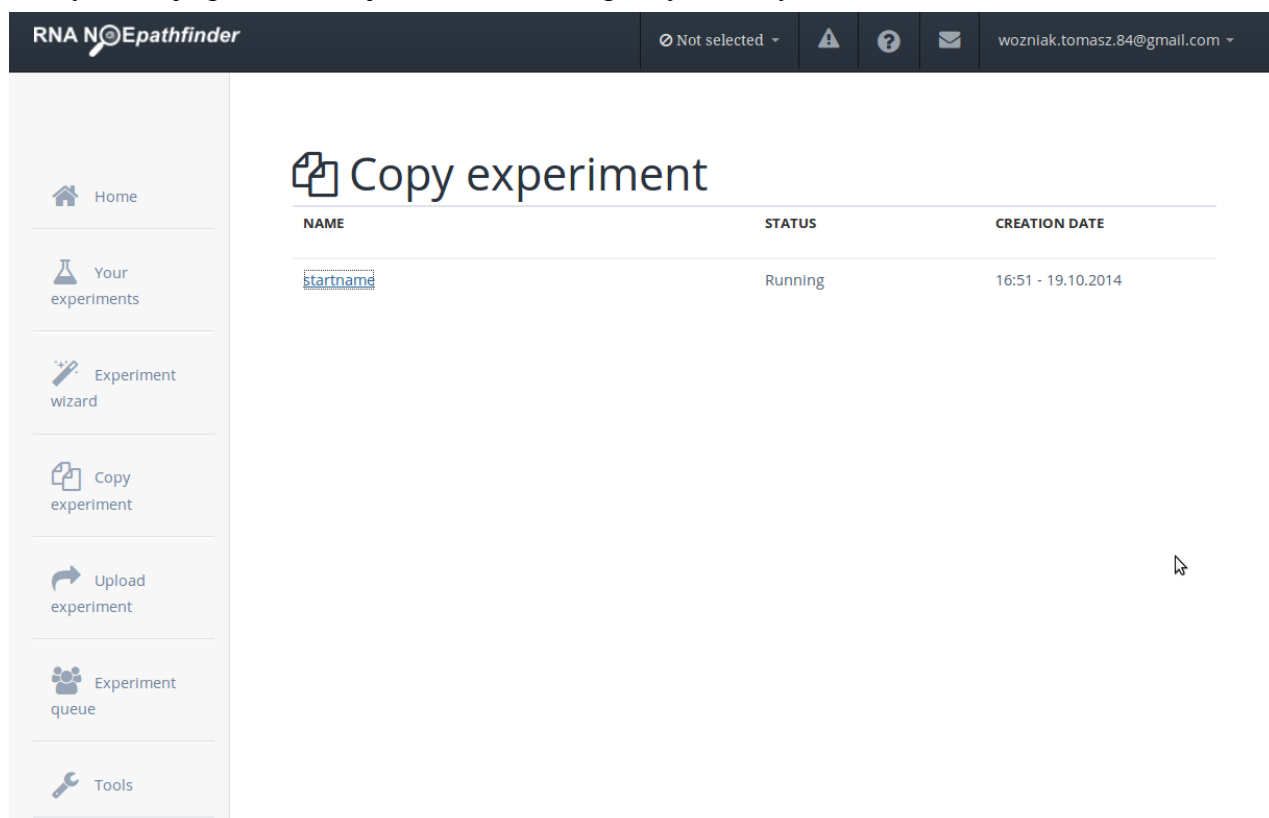


Rys. 55. Kreator eksperymentu w panelu użytkownika.

Kopiowanie eksperymentu jest stworzone w celu umożliwienia użytkownikowi szybkiego przetestowania innego zestawu parametrów do obliczeń. Jest to szczególnie przydatne, gdy z różnych powodów nie można uzyskać prawidłowej ścieżki NOE (rys. 56).

Wczytywanie danych z archiwum jest prostym mechanizmem, który pozwala na ponowne wczytanie wcześniej użytych danych, ręczne przygotowanie plików, które będą potem użyte lub półautomatyczne wygenerowanie tych plików. W celu ułatwienia pobierania

i przetwarzania tych danych z programu Felix stworzyłem skrypt napisany w języku Python, który szerzej opisałem dalej w rozdziale „Programy towarzyszące”.



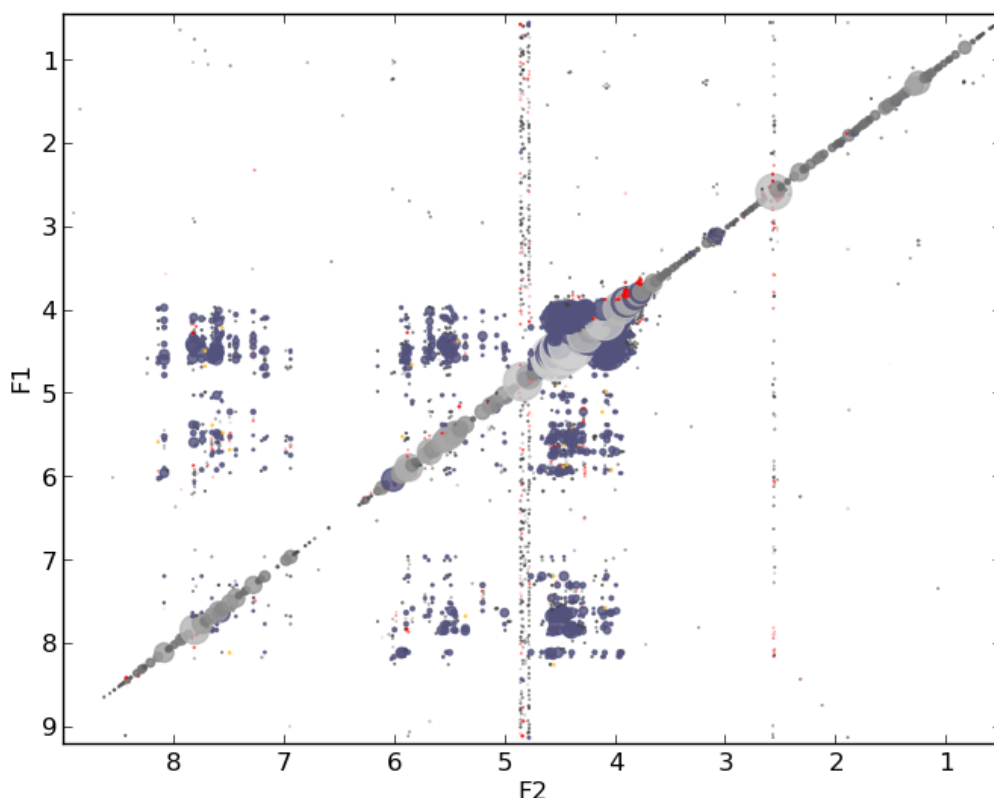
Rys. 56. Kopiowanie eksperymentu w panelu użytkownika

Dla listy sygnałów przygotowanej w formacie Felix potrzebna jest dodatkowo lista objętości tych sygnałów oraz specjalny plik, w którym umieszczone są informacje o przesunięciach chemicznych dwóch bardzo oddalonych od siebie sygnałów. Dane te są niezbędne do konwersji wartości używanych w obrębie oprogramowania Felix i do zmiany wartości położenia sygnałów w każdym z wymiarów na wartości wyrażone w ppm. Dodatkowo konieczna jest informacja w jakiej temperaturze, pH, jakim buforze i rozpuszczalniku została przygotowana próbka, co umożliwi automatyczny wybór widm do obliczeń.

Po starcie eksperymentu użytkownik jest informowany o postępach oraz zakończeniu obliczeń. Dane te są wyświetlane na stronie eksperymentu z wykorzystaniem technologii AJAX (ang. *Asynchronous JavaScript and XML*) dzięki czemu informacje te pojawiają się prawie natychmiastowo. Po zakończeniu eksperymentu użytkownik może pobrać pliki wynikowe, wśród których znajdują się pliki z wygenerowanymi ścieżkami oraz obliczonymi przypisaniami sygnałów.

Wśród wyników obliczeń są nie tylko dane w postaci tekstowej, lecz również diagramy i wykresy, które w przejrzysty sposób wizualizują dany etap obliczeń. Tak wygenerowane wykresy są przedstawiane użytkownikowi na stronie danego eksperymentu (rys. 57). Informacje o eksperymentach w fazie obliczeń oraz dane przedstawione na diagramach pozwalają na

przerwanie obliczeń, wprowadzenie stosownych zmian oraz ponowne uruchomienie obliczeń.



Rys. 57. Przykładowy diagram wizualizujący wyniki działania algorytmu eliminacji szumów. Niebieskim kolorem zaznaczone są pozostawione sygnały, czerwonym kolorem sygnały o ujemnej objętości, odcieniami szarości sygnały odrzucone na pierwszym etapie eliminacji szumów a pomarańczowym sygnały odrzucone na drugim etapie eliminacji szumów.

Panel użytkownika pozwala również na przejrzanie listy eksperymentów oraz listy eksperymentów oczekujących w kolejce na obliczenia. Informacje o wydarzeniach w systemie są przekazywane użytkownikowi poprzez system powiadomień a kontakt z administratorem jest umożliwiony poprzez panel kontaktowy przypominający wyglądem pocztę mailową. Webserwis wyposażylem również w funkcjonalność "blog" zapewniającą łatwe umieszczanie i modyfikowanie informacji o serwisie, oraz prosty samouczek.

W momencie pisania rozprawy doktorskiej trwają prace związane z umożliwieniem użytkownikowi wprowadzenie zmian do obliczeń praktycznie na każdym etapie. Przykładowo użytkownik powinien mieć możliwość edycji wartości przesunięć chemicznych definiujących regiony na widmie lub edycji automatycznie usuwanych sygnałów.

IV.5.2 Bezpieczeństwo serwisu

Konieczność zapewnienia bezwzględnego działania serwera RNA NOEpathfinder jest silnie związana z zagadnieniami bezpieczeństwa serwisu internetowego. Uzyskanie dostępu do

niepublikowanych danych przez osoby niepowołane mogłoby by utrudnić badania lub nawet umożliwić konkurencyjnemu laboratorium szybsze rozwiązanie danego problemu. W związku z tym podczas projektowania, tworzenia oraz wdrażania programu RNA NOEpathfinder wiele uwagi i wysiłku poświęciłem kwestiom bezpieczeństwa aplikacji i ochronie danych użytkownika.

Bezpieczeństwo systemu informatycznego, a w tym wypadku systemu RNA NOEpathfinder można traktować w odniesieniu do następujących kwestii:

1. bezpieczeństwo strony internetowej
2. bezpieczeństwo aplikacji uruchamianej na maszynie obliczeniowej
3. bezpieczeństwo fizyczne serwera i superkomputera
4. bezpieczeństwo związane z czynnikiem ludzkim

Bezpieczeństwo fizyczne jest zapewnione poprzez umieszczenie komputerów w pomieszczeniach do których dostęp jest ograniczony do wybranych osób i kontrolowany. Pozostałe kwestie omówiłem poniżej.

Bezpieczeństwo strony internetowej

Aby zabezpieczyć stronę internetową przed możliwymi atakami, przede wszystkim wykorzystałem istniejący wzorzec projektowy do tworzenia stron internetowych. Podejście takie ułatwia zachowanie bezpieczeństwa, gdyż wiele mechanizmów zabezpieczających jest wbudowanych we wzorzec projektowy i wystarczy z nich odpowiednio korzystać. Ponadto popularność wzorca projektowego zapewnia duże wsparcie w przypadku luk bezpieczeństwa, które są usuwane w kolejnych wersjach oprogramowania.

Bezpieczeństwo webserwisu obejmuje szereg kwestii. Zaliczają się do nich między innymi:

- zabezpieczenie komunikacji z użytkownikiem
- zabezpieczenia związane z logowaniem do systemu i dostępem do danych
- zabezpieczenie przed atakami typu CSRF i XSS
- zabezpieczenie przed atakami SQL injection
- zabezpieczenie przed nieuprawnionym dostępem do panelu administracyjnego
- zabezpieczenie przed usunięciem danych i skasowaniem konta

Poniżej omówiłem wybrane aspekty związane z bezpieczeństwem.

Zabezpieczenie komunikacji z użytkownikiem zapewniłem poprzez zaimplementowanie szyfrowanego połączenia ze stroną internetową za pośrednictwem protokołu https (ang. *Hypertext Transfer Protocol Secure*). Rozwiązanie takie bardzo utrudnia ataki typu „człowiek pomiędzy” (ang. *man in the middle*), w którym osoba mająca dostęp do komputera, łączy lub połączenia bezprzewodowego przez które odbywa się komunikacja może przechwycić komunikację oraz wysyłać własne pakiety danych.

Zabezpieczenia związane z logowaniem do systemu to przede wszystkim wprowadzenie kodu CAPTCHA, który pozwala na automatyczne odróżnienie człowieka i robota, co w pewnym

stopniu zabezpiecza przed automatycznym masowym zakładaniem nowych kont. Niemniej system ten nie jest odporny na ataki, w których kopia tego kodu jest przekazywana do innych serwisów w których rzeczywiste osoby wpisują poprawny kod, na zamawianie usługi łamania kodów CAPTCHA oraz złożone algorytmy rozpoznawania tekstu na wygenerowanym kodzie. Ze względu na charakter serwisu oraz ograniczone korzyści jakie można uzyskać, ataki tego typu wydają się mało prawdopodobne.

Hasło podawane przez użytkownika nie jest weryfikowane pod względem bezpieczeństwa, użytkownik jest sam odpowiedzialny za podanie bezpiecznego hasła.

Dostęp do przesyłanych plików jest zapewniany dzięki mechanizmowi zaimplementowanemu w aplikacji django-private-files. Aplikacja ta weryfikuje czy zalogowany i nieanonimowy użytkownik jest właścicielem danego pliku i tylko jeśli te warunki są spełnione plik może zostać wyświetlony pobrany lub usunięty przez użytkownika.

Ataki CSRF (ang. *Cross-site request forgery*) polegają na wykonywaniu czynności w serwisie bez wiedzy i zgody osoby atakowanej, np. usuwanie plików lub eksperymentów poprzez kod uruchomiony na innej stronie. W celu zabezpieczenia przed tego typu atakiem wykorzystałem dostępne w Django oprogramowanie pośredniczące (ang. *middleware*), które zabezpiecza przed tego typu atakami w przypadku danych przesyłanych metodą POST poprzez dodawanie specjalnego kodu. Ataki typu XSS (ang. *Cross-site scripting*) polegają na „wstrzyknięciu” niepożądanego kodu na stronę wyświetlaną użytkownikowi. Mechanizm szablonów (ang. *templates*) w Django ma domyślną opcję usuwania tagów, które są niezbędne do uruchomienia „wstrzykniętego” kodu. Zadbalem o to, by w webserwisie RNA NOEpathfinder wszystkie pola w formularzach wypełnianych przez użytkowników były chronione przez ten mechanizm, dodatkowo wprowadziłem własne zabezpieczenie filtrujące tagi w przypadku funkcjonalności „blog”, umożliwiając poprawne formatowanie tekstu.

Ataki typu SQL injection polegają na wywoływaniu poleceń SQL poprzez specjalnie skonstruowane wyrażenia wprowadzane do formularzy na stronie internetowej. Podobnie jak w poprzednim przypadku, aby uniknąć tego niebezpieczeństwa wykorzystałem domyślne zabezpieczenia wykorzystywane przez Django, które wykrywają i neutralizują ataki tego typu. Ponadto dane przesyłane w postaci tekstowej w formularzach lub w postaci plików i paczek nie są uruchamiane na maszynie obliczeniowej, tylko przetwarzane przez stworzone przeze mnie oprogramowanie, które w przypadku wysłania niebezpiecznego kodu poinformuje użytkownika że plik jest nieprawidłowy.

Ze względu na to, że administrator ma wgląd we wszystkie dane, konieczne było dodatkowe zabezpieczenie panelu administracyjnego. W tym celu ustawiłem pułapkę, tzw. honeypot na osoby chcące uzyskać nieuprawniony dostęp do panelu administracyjnego. Pułapka ta znajduje się pod domyślnym adresem panelu administracyjnego (<http://rmanoepathfinder.ibch.poznan.pl/admin/>), natomiast rzeczywisty system administracyjny został przesunięty pod inny adres (zabezpieczenie poprzez utajenie).

Skasowanie danych wynikowych, eksperymentu, oraz konta musi być potwierdzone

hasłem. Zabezpiecza to przed usunięciem konta i danych w przypadku niewylogowania się użytkownika z systemu. Niestety w takim przypadku osoba nieuprawniona, korzystająca z komputera użytkownika, który nie zdołał się wylogować może uzyskać dostęp do wrażliwych danych.

Bezpieczeństwo związane z czynnikiem ludzkim

Ta klasa problemów jest stosunkowo złożona i często jest wykorzystywana do nieautoryzowanego dostępu do systemu. Rozwiązaniem jakie wprowadziłem jest przede wszystkim spójna polityka bezpieczeństwa, która dokładnie określa zachowania które administrator powinien podjąć podczas pracy z programem RNA NOE*pathfinder*, zarówno przy logowaniu przez stronę internetową jak i łączeniu się bezpiecznym połączeniem z serwerem stron www lub maszyną obliczeniową. Do polityki tej należą między innymi:

- logowanie użyciem bezpiecznego połączenia
- długie trudne do złamania hasło
- monitoring serwera
- monitoring aplikacji (honeypot, analiza ruchu użytkowników, analiza liczby odwiedzin)
- testy bezpieczeństwa przed wprowadzeniem zmiany
- wszelkie testy prowadzone są przed wprowadzeniem zmian na serwer
- okresowe tworzenie i niszczenie kopii bezpieczeństwa

IV.6 Programy towarzyszące

Podczas moich wcześniejszych badań oraz na etapie tworzenia programu do analizy widm wielokierunkowych wielokrotnie pojawiała się potrzeba stworzenia skryptów i programów, które wykonywałyby często powtarzające się przekształcenia. W tym celu stworzyłem szereg programów przedstawiony w tabeli 7.

Tabela 7. Oprogramowanie

Nazwa programu	Wykonywane zadania
SPEC (SpectraExtractor)	Automatyczne pobranie położenia sygnałów z widm NMR zdefiniowanych w odpowiednim pliku
PDBinout	Konwersja i dopasowanie formatu PDB do jego wariantów wykorzystywanych i tworzonych przez różnego rodzaju oprogramowanie oraz tworzenie własnych odmian tego formatu
TotalRestaints	Analiza i przekształcanie więzów NMR, generowanie więzów dla istniejących modeli przestrzennych; zmiana formatu Felix na format XPLOR; dane wyjściowe wykorzystywane przez oprogramowanie XPLOR-NIH lub CNS Solve, a w szczególnych przypadkach również przez program HADDOCK
BrokenRestraintsAnalyser	Analiza niespełnionych więzów strukturalnych dla struktur obliczonych z użyciem programu XPLOR-NIH lub CNS Solve

IV.6.1 SPEC - skrypt do pobierania sygnałów z widm NMR z użyciem programu Felix

SpectraExtractor (SPEC) jest to skrypt, który upraszcza pobieranie listy sygnałów z widm NMR oraz tworzenie katalogu zawierającego poprawną strukturę danych. Do swojego działania skrypt potrzebuje listy eksperymentów wraz z parametrami w jakich były prowadzone i na jej podstawie uruchamia procedury w programie Felix wyszukujące sygnały na widmach. Na podstawie tej listy uruchamia program Felix i wykonuje odpowiednie komendy, które generują listy położenia sygnałów i ich objętości. Ze względu na złożoność wewnętrznego języka skryptowego wykorzystywanego w programie Felix oraz jego skromnej dokumentacji, skrypt ten wymaga od użytkownika ręcznego stworzenia plików do konwersji współrzędnych na wartości przesunięć chemicznych.

IV.6.2 PDBinout: narzędzie do konwersji i personalizowania plików PDB

Format PDB jest najpowszechniej wykorzystywanym formatem zawierającym koordynaty przestrzenne cząsteczek biologicznych. Rozwój metod krystalograficznych oraz metod spektroskopii NMR umożliwił poznanie struktur wielu biomolekuł. Badania te zapoczątkowały rozwój różnych metod obliczeniowych, które zostały zaimplementowane w wielu programach. Wiele z takich programów wykorzystuje wersje formatu PDB, które w pewnym stopniu odbiegają od wyznaczonych standardów (53). Pomimo istnienia narzędzi takich jak MMTSB (95) czy VEGA (96), które pozwalają na zmianę niektórych elementów plików PDB, do tej pory brakowało uniwersalnego rozwiązania tego problemu. Stworzony przeze mnie program *PDBinout* (97) pozwala na dokonanie automatycznej konwersji sekcji współrzędnych przestrzennych pomiędzy różnymi wersjami formatu PDB. Pozwala to na uniknięcie powtarzalnych operacji dopasowywania formatu plików wejściowych do wymogów konkretnego oprogramowania. Działanie konwertera uwzględnia między innymi różne nazewnictwo reszt nukleotydowych i aminokwasowych, atomów, znaczniki wykorzystywane w plikach, obecność lub brak atomów wodoru, możliwość występowania wielu modeli danej cząsteczki, łańcuchów w jednym modelu czy też pęknięć łańcucha. Dodatkowo *PDBinout* rozwiązuje problemy pojawiające się na przykład, gdy nazwy reszt lub atomów są tak długie że zajmują miejsce zarezerwowane na wskaźnik alternatywnego położenia, zapewnia wczytanie i generowanie orientacji nici definiowanej przez dodanie cyfr „5” i „3” lub liter „C” lub „N” w przypadku kwasów nukleinowych oraz białek odpowiednio, co jest charakterystyczne dla plików wykorzystywanych przez program Amber. Dodatkowo oprogramowanie to pozwala na łatwe tworzenie, nowych wersji formatu PDB, które są dopasowane do wymagań programów lub zadań nie uwzględnionych w predefiniowanych formatach.

Program *PDBinout* automatycznie wykrywa format wejściowy, w tym schemat nazewnictwa atomów, co upraszcza przetwarzanie plików PDB oraz umożliwia łatwą konwersję nawet w przypadku gdy format wejściowy nie znajduje się na liście znanych formatów. Dzięki tej właściwości, *PDBinout* może zostać wykorzystany przy tworzeniu programów lub webserwisów jako dane przyjmujących pliki PDB, gdyż ułatwi pracę z przesyłanymi przez

użytkowników plikami w różnych wersjach tego formatu. Wiele z elementów programu PDBinout zostało wykorzystanych do pobierania informacji o odległościach bezpośrednio z modeli cząsteczek RNA.

Program wraz z dokumentacją i tutorialiem jest dostępny na licencji LGPL (ang. *GNU Lesser General Public License*) na stronie <http://pdbinout.ibch.poznan.pl>

IV.6.3 TotalRestraints: Program ułatwiający pracę z więzami odległościowymi

Analiza i przygotowanie więzów strukturalnych jest dosyć czasochłonne. Istniejące rozwiązania (np. zaimplementowane w programie Felix) pomagają utworzyć pliki z więzami, ale nie rozwiązują wszystkich problemów. W celu ułatwienia tworzenia oraz pracy z więzami odległościowymi stworzyłem specjalny program - TotalRestraints. Na podstawie plików z oprogramowania Felix tworzy pliki w formacie XPLOR (83) a następnie umożliwia szereg operacji na danych wliczając uśrednienie wartości objętości symetrycznych sygnałów lub obliczenie odległości za pomocą podanych wartości wzorcowych (np. odległość pomiędzy atomami H5 i H6 pirymidyn oraz średnia wartość objętości dla sygnałów korelacyjnych H5/H6 na widmie NOESY). TotalRestraints pozwala również na wyznaczenie zakresu więzów odległościowych, sortowanie i grupowanie danych. Ponadto dzięki systemowi komentarzy oprogramowanie ułatwia przygotowanie więzów do obliczeń w przypadku ponownych obliczeń objętości oraz pozwala na zachowanie do obliczeń więzów pochodzących od tych samych atomów w danej cząsteczce.

Program TotalRestraints udostępniłem na licencji LGPL (ang. *Lesser General Public License*) na stronie <http://totalrestraints.ibch.poznan.pl>.

IV.6.4 Program do analizy niespełnionych więzów odległościowych „BrokenRestraintsAnalyser”

Jest to niewielki program do automatycznej analizy danych zwracanych przez oprogramowanie XPLOR. Ułatwia on wyszukanie błędnych więzów odległościowych, na podstawie analizy plików tworzonych przez oprogramowanie XPLOR. W przypadku gdy wygenerowana struktura nie spełnia jednego z zadanych więzów strukturalnych, informacja ta wypisywana jest do specjalnego pliku. BrokenRestraintsAnalyser analizuje ten plik oraz tworzy plik wynikowy, który można zaimportować do programu Excel lub LibreOffice Calc, co bardzo ułatwia wyszukiwanie problematycznych danych.

V. DYSKUSJA

Program RNA NOE*pathfinder* pozwala na analizę dwuwymiarowych widm NMR oraz wyszukanie ścieżek NOE w rejonie *aromatyczno-anomerycznym* widm 2D NOESY. Analiza obejmuje proces eliminacji szumów, wyszukanie regionów na widmach, dopasowywanie do siebie sygnałów z różnych widm, wybranie odpowiedniej ścieżki NOE do obliczeń oraz reguł do jej oceny a następnie wyszukanie ścieżek NOE i przypisanie sygnałów. W większości przypadków każdy z algorytmów składających się na proces analizy daje dobre rezultaty i w przypadku wstępnie przygotowanych danych pozwalają one na wyszukanie rzeczywistych ścieżek NOE. Niestety w przypadku jeśli ze względu na nakładanie się sygnałów, na widmie brakuje blisko 20% sygnałów wchodzących w skład ścieżki, opracowane przeze mnie rozwiązanie nie radzi sobie z oceną zarówno krótkich fragmentów ścieżek jak i całych ścieżek NOE. Poza tym, jeśli poziom szumów jest wysoki, lub na widmie występują bardzo słabe sygnały, bez ingerencji użytkownika program nie będzie w stanie rozróżnić sygnałów od szumów i może usunąć zbyt dużą liczbę sygnałów.

Wyniki testów wskazują, że RNA NOE*pathfinder* jest użytecznym narzędziem i posiada duży potencjał rozwoju. Dzięki porównaniu danych pochodzących z różnych widm, program jest w stanie lepiej ocenić generowane ścieżki NOE, niż bazując tylko na danych dla widma 2D NOESY. Tym samym zyskuje on znaczącą przewagę nad wcześniejszymi rozwiązaniami opartymi na generowaniu ścieżek NOE (1, 87, 88). Dodatkową przewagę daje wykorzystanie statystyk wartości przesunięć chemicznych, które dodatkowo poprawiają wartość oceny.

Niestety wykorzystanie techniki sekwencyjnych ścieżek NOE do uzyskania przypisań pozwala na uzyskanie takich przypisań tylko dla regularnych regionów helikalnych cząsteczek RNA. W innych przypadkach użytkownik musi sam przypisać brakujące sygnały. Problemu takiego nie posiada rozwiązanie oparte o algorytm FLYA (2). Niemniej bazując na wiedzy eksperymentatora można odpowiednio zmodyfikować plik z sekwencjami, tak by program traktował np. fragmenty pętli jako przedłużenie regionu helikalnego, uzyskując w ten sposób dodatkowe przypisania dla części niesparowanych nukleotydów. Rozwiązanie takie przetestowałem dla cząsteczki zawierającej niesparowane 3' końce.

Program stworzyłem z myślą o wykonywaniu analiz widm wielowymiarowych i w tym celu potrzebne są niewielkie prace wdrożeniowe oraz dokładne testy na prawdziwych danych. Wykorzystanie widm ^{13}C , ^{15}N NOESY-HSQC powinno wyeliminować problemy związane z nakładaniem się sygnałów, które pojawiają się w przypadku widm dwuwymiarowych. Wpłynie to pozytywnie zarówno na proces oceny ścieżek NOE jak i na czas oraz zasoby sprzętowe potrzebne do obliczeń.

Algorytmy przypisywania sygnałów i wyszukiwania ścieżek NOE można będzie wykorzystać do analizy innych regionów widma 2D NOESY. Etap ten wymagać będzie wybrania przez użytkownika jednej lub kilku ścieżek do dalszej analizy. Uruchomienie analiz dla

regionów innych niż region *aromatyczno-anomeryczny* wymaga również niewielkich prac wdrożeniowych oraz wielu testów i optymalizacji parametrów.

Oprogramowanie RNA NOE*pathfinder* można połączyć z programem RNAComposer stworzonym w naszym zespole. W ten sposób można szybko uzyskać modele przestrzenne cząsteczek RNA ułożone za pomocą danych eksperymentalnych.

RNA NOE*pathfinder* umożliwia również wykorzystanie jako szablonu modeli strukturalnych tworzonych przez RNAComposer. Na chwilę obecną dane te są wykorzystywane do wyszukiwania regularnych rejonów helikalnych. Niemniej istnieje możliwość wykorzystania algorytmów analizujących strukturę przestrzenną cząsteczki kwasu nukleinowego w celu uzyskania wartości przesunięć chemicznych dla poszczególnych atomów. Dane te można wykorzystać w funkcjach oceniających tworzone ścieżki NOE co może przełożyć się zarówno na poprawę jakości jak i szybkości obliczeń ścieżek NOE.

RNA NOE*pathfinder* jest udostępniony w postaci webserwisu pod adresem <http://rnanoepathfinder.ibch.poznan.pl>. Interfejs jest możliwie prosty i umożliwia szybkie przygotowanie eksperymentu obliczeniowego na kilka różnych sposobów. Wyniki są uzupełnione o schematy przedstawiające poszczególne etapy analizy. Pozwala to na szybką wizualną analizę i ocenę wyników obliczeń.

Program napisałem w sposób, który umożliwia jego dalsze modyfikacje i rozwój. Można zmodyfikować i udoskonalić praktycznie każdy element oprogramowania w zależności od potrzeb zgłaszanych przez jego użytkowników, od algorytmów wyszukiwania szumów po funkcje oceniające. Przykładowo można rozszerzyć obsługiwane formaty wejściowe o programy Sparky (98) czy Cara (99). Poza tym można wprowadzić metody analizy surowych widm NMR a nie tylko sygnałów na widmach. Można również ulepszyć funkcje oceniające tak, by uwzględniały typowe wartości przesunięć chemicznych dla danej reszty znajdującej się w określonym otoczeniu chemicznym. Kolejną możliwością jest rozwój algorytmów przypisujących sygnały NMR w różnych rejonach widm NMR. Struktura plików wewnętrznych programu umożliwia wprowadzenie danych dla różnych reszt nukleotydowych. Możliwe jest również uwzględnienie większego spektrum wykorzystywanych widm i technik NMR.

Wydaje się również, że po wprowadzeniu szeregu modyfikacji, program RNA NOE*pathfinder* mógłby być użyty do analizy kompleksów białko - RNA. W tym celu niezbędna jest integracja z jednym z istniejących rozwiązań pozwalających na automatyczne lub półautomatyczne przypisywanie sygnałów należących do poszczególnych reszt aminokwasowych. Następnie umiejętne wykorzystanie algorytmów nakładających widma pozwoliłoby na wyszukanie sygnałów należących do białka i RNA na podstawie widm wykonanych dla preparatów zawierających jedynie RNA lub jedynie białko.

VI. PODSUMOWANIE

- W wyniku mojej pracy powstało oprogramowanie RNA NOE*pathfinder*, służące do automatycznej analizy widm NMR dla cząsteczek kwasu nukleinowego.
- Analiza obejmuje przede wszystkim etapy redukcji szumów na widmie, wyszukiwania charakterystycznych regionów, korelację sygnałów pomiędzy poszczególnymi widmami NMR, generowanie i ocenianie ścieżek NOE oraz częściowe przypisywanie sygnałów.
- Redukcja szumów składa się z dwóch etapów, przy czym drugi etap jest różny dla widm homo- i heterojądrowych. Zastosowane rozwiązanie pozwala na znaczne wyeliminowanie fałszywie pozytywnych sygnałów co ułatwia dalsze obliczenia.
- Do wyszukiwania regionów na widmie wykorzystywane są wartości uzyskane z trzech różnych algorytmów, co pozwala na poprawne znalezienie regionów nawet w przypadku gdy jeden z algorytmów wskaże błędne wartości.
- Korelacja sygnałów między widmami umożliwia wykorzystanie informacji pochodzącej z różnych widm. Algorytm umożliwia dopasowanie sygnałów zarówno w przypadku gdy wszystkie wartości przesunięć chemicznych lub pojedyncza wartość przesunięcia chemicznego jest przesunięta w kierunku wyższych lub niższych wartości.
- Generowanie ścieżek NOE oraz przypisywania sygnałów opiera się o wieloetapowy proces wyszukiwania charakterystycznych ścieżek NOE oraz ich ocenę za pomocą specjalnych funkcji oceniających. Ścieżki NOE składane są z krótszych fragmentów, które zostały przypisane do określonego miejsca w sekwencji i ocenione. Funkcje oceniające uwzględniają zarówno zgodność wcześniejszych przypisań sygnałów z sekwencją, kształt ścieżki, wartości przesunięć chemicznych oraz obecność sygnałów w innych regionach widma NOESY.
- Podczas analizy wykonywane jest częściowe przypisywanie sygnałów wykorzystujące jednoznaczne informacje uzyskane przede wszystkim z korelacji pomiędzy widmami.
- Poszczególne algorytmy analizy widm NMR zostały przetestowane na danych rzeczywistych dla kilku różnych cząsteczek RNA.
- Cały program został przetestowany na dwóch cząsteczkach:
 - dla jednej z nich dane nie zostały poddane wstępnej obróbce, w tym przypadku została znaleziona prawidłowa ścieżka NOE, ale nie została ona najlepiej oceniona, ponadto konieczne było wprowadzenie pewnych ograniczeń do programu
 - w drugim przypadku dane zostały poddane wstępnemu przygotowaniu, program wygenerował prawidłową ścieżkę i została ona najlepiej oceniona.
- Oprogramowanie jest udostępnione przez stronę internetową z prostym i wygodnym interfejsem, który pozwala na uruchamianie obliczeń w systemie kolejkowym.
- Przypisania sygnałów na widmach NMR uzyskane dzięki analizom ścieżek NOE przez program RNA NOE*pathfinder* można wykorzystać do stworzenia więzów strukturalnych

niezbędnych dla procesu modelowania struktury przestrzennej cząsteczki RNA, lub ich przesyłanie do programu RNAComposer, gdzie mogą być one wykorzystywane do automatycznego udokładniania struktury przestrzennej modelu cząsteczki RNA.

- Program został w dużym stopniu dostosowany do analizy widm posiadających większą liczbę wymiarów, jednak do prawidłowego funkcjonowania niezbędne jest wprowadzenie niewielkich zmian oraz testy na danych rzeczywistych.
- W wyniku prac związanych z programem RNA NOE*pathfinder* powstały inne programy, z których warto wymienić dwa. Pierwszy to program PDB*inout* (97) służący do ułatwiania pracy z plikami PDB oraz automatycznego dopasowywania formatu plików PDB do własnych potrzeb. Drugim z programów jest TotalRestraints – program ułatwiający pracę z więzami odległościowymi w formacie XPLOR.

VII. LITERATURA

- 1 R.W. Adamiak, J. Blazewicz, P. Formanowicz, Z. Gdaniec, M. Kasprzak, M. Popenda, and M. Szachniuk (2004) *An algorithm for an automatic NOE pathways analysis of 2D NMR spectra of RNA duplexes* **J. Comput. Biol.** 11(1): 163-179
- 2 T. Aeschbacher, E. Schmidt, M. Blatter, C. Maris, O. Duss, F. H.-T. Allain, P. Güntert, and M. Schubert (2013) *Automated and assisted RNA resonance assignment using NMR chemical shift statistics* **Nucleic Acids Res.** 41(18): e172
- 3 B. Lewin, H.B.V. Zevallos, F.G. Roig (2008) „Genes”
- 4 H. Lodish, A. Berk, S.L. Zipursky, P. Matsudaira, D. Baltimore, J. Darnell. (2000) „Molecular Cell Biology. 4th edition.”
- 5 R.W. Holley, J. Apgar, G.A. Everett, J.T. Madison, M. Marquisee, S.H. Merrill, J.R. Penswick, A. Zamir (1965) Structure of a ribonucleic acid. **Science** 147(3664): 1462-5
- 6 R.M. Anderson, M. Kwon, S.A. Strobel (2007) *Toward ribosomal RNA catalytic activity in the absence of protein.* 64(4): 472-83
- 7 T. Auerbach, A. Bashan, J. Harms, F. Schlutzenzen, R. Zarivach, H. Bartels, I. Agmon, M. Kessler, M. Pioletti, F. Franceschi, A. Yonath (2002) *Antibiotics targeting ribosomes: crystallographic studies.* **Curr Drug Targets Infect Disord.** 2(2): 169-86.
- 8 M. U. Kaikkonen, M. T.Y. Lam, and C. K. Glass (2011) *Non-coding RNAs as regulators of gene expression and epigenetics* **Cardiovasc Res.** 90(3): 430-440
- 9 The ENCODE Project Consortium. (2004) *The ENCODE (ENCyclopedia Of DNA Elements) Project.* **Science.** 306: 636–640
- 10 J. Krol, I. Loedige, W. Filipowicz (2010) *The widespread regulation of microRNA biogenesis, function and decay.* **Nat Rev Genet.** 11: 597–610
- 11 Y. Lee, C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Rådmark, S. Kim, V. N. Kim (2003) *The nuclear RNase III Drosha initiates microRNA processing.* **Nature.** 425: 415–419
- 12 E. Bernstein, A.A. Caudy, S.M. Hammond, G.J. Hannon (2001) *Role for a bidentate ribonuclease in the initiation step of RNA interference.* **Nature.** 409: 363–366.
- 13 H. Siomi, M. C. Siomi (2009) *On the road to reading the RNA-interference code.* **Nature.** 457: 396–404
- 14 J. Brennecke, A.A. Aravin, A. Stark, M. Dus, M. Kellis, R. Sachidanandam, G.J. Hannon (2007) *Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila* **Cell.** 128: 1089–1103
- 15 R.W. Carthew, E.J. Sontheimer. (2009) *Origins and mechanisms of miRNAs and siRNAs* **Cell.** 136: 642–655
- 16 P. Kapranov, J. Cheng, S. Dike, D.A. Nix, R. Duttagupta, A.T. Willingham, P.F. Stadler, J. Hertel, J. Hackermüller, I. L. Hofacker, I. Bell, E. Cheung, J. Drenkow, E. Dumais, S. Patel, G. Helt, M. Ganesh, S. Ghosh, A. Piccolboni, V. Sementchenko, H. Tamma, T.R. Gingeras (2007) *RNA maps reveal new RNA classes and a possible function for pervasive transcription* **Science.** 316: 1484–1488.

- 17 F. De Santa, I. Barozzi, F. Mietton, S. Ghisletti, S. Polletti, B.K. Tusi, H. Muller, J. Ragoussis, C.L. Wei, G. Natoli (2010) *A large fraction of extragenic RNA pol II transcription sites overlap enhancers*. **PLoS Biol.** 8(5): e1000384
- 18 J. Han, D. Kim, K. V. Morris (2007) *Promoter-associated RNA is required for RNA-directed transcriptional gene silencing in human cells*. **Proc Natl Acad Sci U.S.A.** 104: 12422–12427
- 19 W. Filipowicz, V. Pogacic (2002) *Biogenesis of small nucleolar ribonucleoproteins*. **Curr Opin Cell Biol.** 14(3): 319-27
- 20 T. R. Cech (1990) *Self-splicing of group I introns* **Annu Rev Biochem.** 59: 435–461
- 21 A. Piekarowicz (2012) „Podstawy wirusologii molekularnej”
- 22 C. Altona, M. Sundaralingam (1972) *Conformational analysis of the sugar ring in nucleosides and nucleotides. A new description using the concept of pseudorotation*. **J. Am. Chem. Soc.** 94(23): 8205-12
- 23 C. Altona, M. Sundaralingam (1973) *Conformational analysis of the sugar ring in nucleosides and nucleotides. Improved method for the interpretation of proton magnetic resonance coupling constants*. **J Am Chem Soc.** 95(7): 2333–2344
- 24 W. Saenger. Springer-Verlag (1984) „Principles of Nucleic Acid Structure.”
- 25 U. Nagaswamy, N. Voss, Z. Zhang, G.E. Fox (2000) *Database of non-canonical base pairs found in known RNA structures*. **Nucleic Acids Res.** 28(1): 375-6
- 26 M. Popenda, J. Milecki, R.W. Adamiak (2004) *High salt solution structure of a left-handed RNA double helix*. **Nucleic Acids Res.** 32(13): 4044-54
- 27 L. Stryer (2003) „Biochemia”
- 28 R.T. Batey, R.P. Rambo, J.A. Doudna (1999) *Tertiary Motifs in RNA Structure and Folding*. **Angew Chem Int Ed Engl.** 38(16): 2326-2343
- 29 R. Thapar, A.P. Denmon, E.P. Nikonowicz (2014) *Recognition modes of RNA tetraloops and tetraloop-like motifs by RNA-binding proteins*. **Wiley Interdiscip Rev RNA.** 5(1): 49-67
- 30 M. de la Peña, D. Dufour, J. Gallego (2009) *Three-way RNA junctions with remote tertiary contacts: A recurrent and highly versatile fold* **RNA** 15(11): 1949–1964
- 31 E. Bindewald, R. Hayes, Y. Yingling, W. Kasprzak, B.A. Shapiro (2008) *RNAJunction: A Database of RNA Junctions and Kissing Loops For Three-Dimensional Structural Analysis and Nanodesign*. **Nucleic Acids Res.** 2008 (Database issue): D392-7
- 32 L. Popenda, L. Bielecki, Z. Gdaniec, R.W. Adamiak (2009) *Structure and dynamics of adenosine bulged RNA duplex reveals formation of the dinucleotide platform in the C:G-A triple* **ARKIVOC** 2009: 130-144
- 33 J.D. Robertus, J.E. Ladner, J.T. Finch, D. Rhodes, R.S. Brown, B.F. Clark, A. Klug (1974) *Structure of yeast phenylalanine tRNA at 3 Å resolution*. **Nature** 250(467): 546-51
- 34 K.Y. Chang, I. Tinoco Jr. (1997) *The structure of an RNA "kissing" hairpin complex of the HIV TAR hairpin loop and its complement*. **J Mol Biol.** 269(1): 52-66.

- 35 M. Taufer, A. Licon, R. Araiza, D. Mireles, F.H. van Batenburg, A.P. Gultyaev, M.Y. Leung (2009) *PseudoBase++: an extension of PseudoBase for easy searching, formatting and visualization of pseudoknots*. **Nucleic Acids Res.** 2009 (Database issue): D127-35
- 36 C. Tuerk, S. MacDougall, L. Gold (1992) *RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase*. **Proc Natl Acad Sci U S A.** 89(15): 6988-92
- 37 M. Małgowska, D. Gudanis, A. Teubert, G. Dominiak, Z. Gdaniec (2012) *How to study G-quadruplex structures* **BioTechnologia** 93(4): 381-390
- 38 D.D. Cash, O. Cohen-Zontag, N.K. Kim, K. Shefer, Y. Brown, N.B. Ulyanov, Y. Tzfati, J. Feigon (2013) *Pyrimidine motif triple helix in the Kluyveromyces lactis telomerase RNA pseudoknot is essential for function in vivo*. **Proc Natl Acad Sci U S A.** 110(27): 10970-5
- 39 C. Ehresmann, F. Baudin, M. Mougel, P. Romby, J.P. Ebel, B. Ehresmann (1987) *Probing the structure of RNAs in solution*. **Nucleic Acids Res.** 15(22): 9109–9128
- 40 C. Werner, B. Krebs, G. Keith, G. Dirheimer. (1976) *Specific cleavages of pure tRNAs by plumbous ions*. **Biochim Biophys Acta.** 432(2): 161-75
- 41 E.J. Merino, K.A. Wilkinson, J.L. Coughlan, K.M. Weeks. (2005) *RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE)* **J. Am. Chem. Soc.** 127(12): 4223–4231
- 42 L. Finzi, D.D. Dunlap (2010) *Single-molecule Approaches to Probe the Structure, Kinetics, and Thermodynamics of Nucleoprotein Complexes That Regulate Transcription* **J Biol Chem.** 285(25): 18973–18978
- 43 J. Esquivel-Rodríguez, D. Kihara (2013) *Computational methods for constructing protein structure models from 3D electron microscopy maps*. **J Struct Biol.** 184(1): 93-102
- 44 J. Kypr, I. Kejnovská, D. Renciuk, M. Vorlícková (2009) *Circular dichroism and conformational polymorphism of DNA*. **Nucleic Acids Res.** 37(6): 1713–25
- 45 C. Lecomte, B. Guillot, N. Muzet, V. Pichon-Pesme, C. Jelsch (2004) *Ultra-high-resolution X-ray structure of proteins*. **Cell Mol Life Sci.** 61(7-8): 774-82
- 46 N. Ban, P. Nissen, J. Hansen, P.B. Moore, T.A. Steitz (2000) *The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution* **Science.** 289(5481): 905-20
- 47 G.J. Pielak, C. Li, A.C. Miklos, A.P. Schlesinger, K.M. Slade, G. Wang, I.G. Zgoneanu (2009) *Protein Nuclear Magnetic Resonance under Physiological Conditions* **Biochemistry.** 48(2): 226–234
- 48 S.C. Keane, X. Heng, K. Lu, S. Kharytonchyk, V. Ramakrishnan, G. Carter, S. Barton, A. Hosic, A. Florwick, J. Santos, N.C. Bolden, S. McCowin, D.A. Case, B.A. Johnson, M. Salemi, A. Telesnitsky, M.F. Summers (2015) *Structure of the HIV-1 RNA packaging signal* **Science.** 348(6237): 917-921
- 49 J. M. Bujnicki (2009) „Prediction of Protein Structures, Functions, and Interactions”
- 50 M. A. Jonikas, R. J. Radmer, A. Laederach, R. Das, S. Pearlman, D. Herschlag, R. B. Altman (2009) *Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters*. **RNA** 15(2): 189-99

- 51 M. Popena, M. Szachniuk, M. Antczak, K. J. Purzycka, P. Lukasiak, N. Bartol, J. Blazewicz, R. W. Adamiak. (2012) *Automated 3D structure composition for large RNAs* **Nucleic Acids Res.** 40(14): e112
- 52 K.J. Purzycka, M. Popena, M. Szachniuk, M. Antczak, P. Lukasiak, J. Blazewicz, R.W. Adamiak (2015) *Automated 3D RNA structure prediction using the RNAComposer method for riboswitches.* **Methods Enzymol.** 553: 3-34
- 53 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 Res.** 28(1): 235–242
- 54 J. W. Hennel, J. Klinowski (2000) „Podstawy magnetycznego rezonansu jądrowego”
- 55 A. Ejchart, A. Gryff-Keller (2004) „NMR w cieczach Zarys teorii i metodologii”
- 56 K. Stott , J. Stonehouse , J. Keeler , T.-L. Hwang , A. J. Shaka (1995) *Excitation Sculpting in High-Resolution Nuclear Magnetic Resonance Spectroscopy: Application to Selective NOE Experiments .* **J. Am. Chem. Soc.**, 117(14): 4199–4200
- 57 K.V.R. Chary, G. Govil (2008). „NMR in Biological Systems”
- 58 E.P. Nikonowicz, A. Pardi (1992) *Three-dimensional heteronuclear NMR studies of RNA* **Nature** 355: 184-186
- 59 E.P. Nikonowicz, A. Pardi (1993) *An efficient procedure for assignment of the proton, carbon and nitrogen resonances* **J. Mol. Biol.** 232: 1141-1156
- 60 K.B. Hall, J.K. Kranz (1995) *Thermodynamics and mutations in RNA:protein interactions* **Methods in Enzymol.** 259: 261-281
- 61 A.H. Heus, S.S. Wijmenga, M.J.F. van de Ven (1994) *Sequential backbone assignment in ^{13}C -labeled RNA via through-bond coherence transfer using three-dimensional triple resonance spectroscopy (^1H , ^{13}C , ^{31}P) and two-dimensional hetero TOCSY* **J. Am. Chem. Soc.** 116: 4983-4984
- 62 W.P. Aue, E. Bartholdi, R.R. Ernst (1976) *Two-dimensional spectroscopy. Application to nuclear magnetic resonance* **J. Chem. Phys.** 64: 2229-2246
- 63 A. Bax, R. Freeman (1981) *Investigation of complex networks of spin-spin coupling by two-dimensional NMR.* **J. Magn. Reson.** 44: 542-561
- 64 A.J. Shaka, R. Freeman (1983) *Simplification of NMR spectra by filtration through multiple-quantum coherence* **J. Magn. Reson.** 51(1): 169-173
- 65 M. Rance, O.W. Sorensen, G. Bodenhausen, G. Wagner, R.R. Ernst, K. Wuthrich (1983) *Improved spectral resolution in cosy h-1-NMR spectra of proteins via double quantum filtering* **Biochemical and Biophysical Research Communications** 117: 479-485
- 66 N. Müller, R.R. Ernst, K. Wüthrich (1986) Multiple-quantum-filtered two-dimensional correlated NMR spectroscopy of proteins. **J. Am. Chem. Soc.** 108: 6482–6492
- 67 T. Facke, R. Wagner, S. Berger (1994) *Two-Dimensional ^{13}C , Heteroelement Correlation Spectroscopy* **Concepts in Magn. Reson.** 6: 293-306
- 68 G. Bodenhausen, D.J. Ruben (1980) *Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy* **Chem. Phys. Lett.** 69 (1): 185–189
- 69 J. Keeler, James (2010). „Understanding NMR Spectroscopy (2nd ed.)”

- 70 D.G. Davis, A. Bax (1985) *Assignment of complex ¹H NMR spectra via two-dimensional homonuclear Hartmann-Hahn spectroscopy*. **J. Am. Chem. Soc.** 107: 2820-2821 .
- 71 M.W. Edwards, A. Bax (1986) *Complete Proton and carbon-13 NMR assignment of the alkaloid gephyrotoxin through the use of homonuclear Hartmann-Hahn and two-dimensional spectroscopy*. **J. Am. Chem. Soc.** 108: 918-923
- 72 D. Piveteau, A. Delsuc, E. Guittet, J.Y. Lallemand (1987) *The „lock and return” experiment for water suppression with Hartmann-Hahn transfer* **J. Magn. Reson.** 71: 347-348
- 73 J. Jeener, B.H. Meier, P. Bachmann, R.R. Ernst (1979) *Investigation of exchange processes by two-dimensional NMR spectroscopy* **J. Chem Phys.** 71: 4546-4553
- 74 K. Lu, Y. Miyazaki, M.F. Summers (2010) *Isotope labeling strategies for NMR studies of RNA* **J Biomol NMR** 46(1): 113–125
- 75 O. Duss, P.J. Lukavsky, F.H. Allain (2012) *Isotope labeling and segmental labeling of larger RNAs for NMR structural studies*. **Adv Exp Med Biol** 992: 121-44
- 76 S.S. Wijmenga, B.N.M. van Buuren, (1998) *The use of NMR methods for conformational studies of nucleic acids*. **Prog. Nucl. Magn. Reson. Spectrosc.** 32: 287–387
- 77 G. Varani Jr. (1991) *RNA structure and NMR spectroscopy*. **Q. ReV. Biophys.** 24: 479–532
- 78 G. Varani, F. Aboulela, F.H.T. Allain (1996) *NMR investigation of RNA structure*. **Prog. Nucl. Magn. Reson. Spectrosc.** 29: 51–127
- 79 V.A. Roongta, C.R. Jones, D.G. Gorenstein (1990) *Effect of distortions in the deoxyribose phosphate backbone conformation of duplex oligodeoxyribonucleotide dodecamers containing GT, GG, GA, AC, and GU base-pair mismatches on ³¹P NMR spectra*. **Biochemistry** 29: 5245–5258
- 80 L. Popenda, R.W. Adamiak, Z. Gdaniec (2008) *Bulged Adenosine Influence on the RNA Duplex Conformation in Solution* **Biochemistry** 47(18): 5059-67
- 81 P. Güntert, C. Mumenthaler, K. Wüthrich (1997) *Torsion angle dynamics for NMR structure calculation with the new program DYANA*. **J Mol Biol.** 273(1): 283-98
- 82 E.G. Stein, L.M. Rice, A.T. Brünger (1997) *Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation*. **J. Magn. Reson.** 124: 154–164
- 83 C.D. Schwieters, J.J. Kuszewski, N. Tjandra, G.M. Clore (2003) *The Xplor-NIH NMR Molecular Structure Determination Package* **J. Magn. Res.** 160(1): 66-74
- 84 C.D. Schwieters, J.J. Kuszewski, G.M. Clore (2006) *Using Xplor-NIH for NMR molecular structure determination*. **Prog. Nucl. Magn. Reson. Spectrosc.** 48: 47–62.
- 85 P. Güntert (2004). *Automated NMR protein structure calculation with CYANA*. **Meth. Mol. Biol.** 278: 353-378
- 86 M.W. Roggenbuck, T.J. Hyman, P.N. Borex (1990) *Path analysis in NMR spectra: application to an RNA octamer*. **Structure Methods**, 3: 309–317.
- 87 J. Blazewicz, M. Szachniuk, A. Wojtowicz (2004) *Evolutionary approach to NOE paths assignment in RNA structure elucidation*. **Proceedings of the IEEE Symposium on Computer International in Bioinformatics and Computational Biology** 206–213

- 88 J. Blazewicz, M. Szachniuk, A. Wojtowicz (2005) *RNA tertiary structure determination: NOE pathways construction by tabu search*. **Bioinformatics**. 21(10): 2356-61
- 89 M. Szachniuk (2015) *Assigning NMR spectra of irregular RNAs by heuristic algorithms* **Bulletin of the Polish Academy of Sciences, Technical Sciences** 63(1): 329-338
- 90 E. Schmidt, P. Güntert (2012) *A new algorithm for reliable and general NMR resonance assignment*. **J Am Chem Soc**. 134(30): 12817-29
- 91 M. Szachniuk, M.C. De Cola, G. Felici, D. de Werra, J. Błażewicz (2015) *Optimal pathway reconstruction on 3D NMR maps* **Discrete Applied Mathematics** 182: 134-149
- 92 M. Szachniuk, M.C. De Cola, G. Felici, J. Blazewicz (2014) *The Orderly Colored Longest Path Problem - a survey of applications and new algorithms*. **RAIRO - Operations Research** 48(1): 25-51
- 93 D. Greenfeld, A. Roy (2014) „Two Scoops of Django: Best Practices for Django 1.6”
- 94 Z. Liu, A. Abbas, B.Y. Jing, X. Gao (2012) *WaVPeak: picking NMR peaks through wavelet-based smoothing and volume-based filtering*. **Bioinformatics**. 28(7): 914-20
- 95 M. Feig, J. Karanicolas, C.L. Brooks III (2004) *MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology*. **J Mol. Graph. Model** 22(5): 377–395.
- 96 A. Pedretti, L. Villa, G. Vistoli (2002) *VEGA: a versatile program to convert, handle and visualize molecular structure on Windows-based PCs*. **J Mol. Graph. Model** 21(1): 47–49
- 97 T. Woźniak, R.W. Adamiak (2013) *Personalization of structural PDB files*. **Acta Biochim Pol**. 60(4): 591-3
- 98 T.D. Goddard and D.G. Kneller, **SPARKY 3**, University of California, San Francisco
- 99 R. L.J. Keller (2004) „The Computer Aided Resonance Assignment Tutorial”
- 100 Barbara. Krähenbühl, P. Lukavsky, G. Wider (2014) *Strategy for automated NMR resonance assignment of RNA: application to 48-nucleotide K10* **Journal of Biomolecular NMR** 59(4): 231-240