Rozprawa doktorska p.t.:

Opracowanie nowych baz danych opisujących metabolizm kwasów nukleinowych

Mgr Kaja Milanowska

Pracownia Bioinformatyki
Instytut Biologii Molekularnej i Biotechnologii
Uniwersytet im. Adama Mickiewicza
Poznań

Promotor: prof. dr hab. Janusz M. Bujnicki

Pracownia Bioinformatyki
Instytut Biologii Molekularnej i Biotechnologii
Uniwersytet im. Adama Mickiewicza
Poznań

Laboratorium Bioinformatyki i Inżynierii Białka
Międzynarodowy Instytut Biologii Molekularnej i Komórkowej
Warszawa

Poznań 2013

Podziękowania

Serdecznie dziękuję mojemu promotorowi ***prof. dr hab. Januszowi M. Bujnickiemu*** *za niezwykle ciekawy okres studiów doktoranckich, za opiekę naukową, przekazaną wiedzę oraz inspirację do pracy badawczej.*

Serdecznie dziękuję Dyrektor Instytutu Biologii Molekularnej i Biotechnologii pani ***prof. dr hab. Zofii Szweykowskiej-Kulińskiej*** *za wszelką pomoc, wsparcie i dobrą radę.*

Dziękuję wszystkim współautorom projektów RNApathwaysDB oraz REPAIRtoire za przemiłą współpracę i wszelką pomoc, w szczególności ***dr Joannie Krwawicz, mgr Magdalenie Machnickiej***, ***dr Janowi Kosińskiemu*** *oraz* ***mgr Grzegorzowi Papajowi***.

Dziękuję ***mgr Annie Philips***, ***mgr Magdalenie Machnickiej, dr Joannie Kasprzak*** *oraz* ***mgr Markowi Muszyńskiemu*** *za opinie i uwagi merytoryczne oraz redakcyjne na temat niniejszej pracy.*

Dziękuję również adiunktom, doktorantom, magistrantom i licencjuszom Pracowni Bioinformatyki, za wspólnie przeżyte chwile, miłą atmosferę i wsparcie w trudnych chwilach. W szczególności ***dr Annie Czerwoniec***, ***dr Kristianowi Rotherowi***, ***dr Magdalenie Rother***, ***dr Tomaszowi Putonowi***, ***mgr Katarzynie Mikołajczak***, ***mgr Ewie Bielskiej***, ***mgr Annie Łukasik***, ***mgr Zuzannie Balcer*** *i* ***mgr Mateuszowi Korycińskiemu*** *za wspaniałą atmosferę w pracy, ciekawe dyskusje, zarówno naukowe jak i prywatne oraz wszystkie literówki, które pomogli usunąć z niniejszej pracy.*

Dziękuję także koleżankom i kolegom z Instytutu Biologii Molekularnej i Biotechnologii, Wydziału Biologii UAM oraz Pracowni Inżynierii Białka Międzynarodowego Instytutu Biologii Molekularnej i Komórkowej w Warszawie.

Pragnę z całego serca podziękować Mojej Rodzinie, Bliskim i Przyjaciołom za wsparcie i motywację.

Pracę dedykuję Siostrze i Rodzicom.

Spis treści

Finansowanie

Rozwój baz danych REPAIRtoire oraz RNApathwaysDB był finansowany z następujących źródeł:

REPAIRtoire:

- Ministerstwo Nauki i Szkolnictwa Wyższego (MNiSW) (grant Iuventus Plus 0360/IP1/2011/71 dla Kai Milanowskiej);
- 6 Europejski Program Ramowy (FP6, grant PLASTOMICS, numer kontraktu LSHG-CT-2003-503238 i DNA ENZYMES, numer kontraktu MRTN-CT-2005-019566 oraz FP7 HEALTHPROT, numer kontraktu 229676; kierownik zadań: Janusz M. Bujnicki);
- Grant polsko-norweski (PNRF-143-AI-1/07, subkontrakt dla Joanny Krwawicz);
- Ministerstwo Nauki i Szkolnictwa Wyższego (grant PBZ-MNiI-2/1/2005, subkontrakt dla Janusza M. Bujnickiego) oraz Deutscher Akademischer Austausch Dienst (stypendium D/09/42768 dla Kristiana Rothera, częściowo).

RNApathwaysDB:

- Narodowe Centrum Nauki (NCN) (N N301 072 640 dla Kai Milanowskiej);
- Fundusze statutowe Uniwersytetu im. Adama (dla Kai Milanowskiej);
- Fundacja na Rzecz Nauki Polskiej (FNP) (TEAM/2009-4/2 dla Janusza M. Bujnickiego);
- Ministerstwo Nauki i Szkolnictwa Wyższego (MNiSW) (POIG.02.03.00-00-003/09 dla Laboratorium Janusza M. Bujnickiego);
- 7 Europejski Program Ramowy (HEALTH-PROT, numer kontraktu 229676);
- German Academic Exchange Service (DAAD) (D/09/42768 dla Kristiana Rothera);
- European Research Council (ERC) (StG grant RNA+P=123D dla Janusza M. Bujnickiego).

Publikacje wynikające z pracy

Milanowska K, Mikolajczak K, Lukasik A, Skorupski M, Balcer Z, Machnicka MA, Nowacka M, Rother KM, Bujnicki JM ***RNApathwaysDB – a database of RNA maturation and decay pathways.*** *Nucleic Acids Res* 2013 Jan 1;41(D1):D268-72

Milanowska K, Krwawicz J, Papaj G, Kosinski J, Poleszak K, Lesiak J, Osinska E, Rother K, Bujnicki JM ***REPAIRtoire – a database of DNA repair pathways***. *Nucleic Acids Res* 2011 Jan;39 (Database issue):D788-92,

Milanowska K, Rother K, Bujnicki JM ***Databases and bioinformatics tools for the study of DNA repair.*** *Mol Biol Int.* 2011; 2011:475718. doi: 10.4061/2011/475718. Epub 2011 Jul 14.

Wykaz skrótów

16S and 23S Ribosomal RNA Mutation Database	zbiór informacji dotyczących 16S i 23S rRNA
3′-UTR	region niepodlegający translacji na końcu 3′ (ang. *3′-UnTranslated Region*)
5′-UTR	region niepodlegający translacji na końcu 5′ (ang. *5′-UnTranslated Region*)
5S Ribosomal RNA Database	źródło danych o 5S rRNA
ADP	adenozyno-5'-difosforan (ang. *Adenosine DiPhosphate*)
ATP	adenozyno-5'-trifosforan (ang. *Adenosine TriPhosphate*)
BER	naprawa przez wycinanie zasady (ang. *Base-Excision Repair*)
BioCyc	baza danych doświadczalnie przebadanych szlaków metabolicznych i enzymów z ponad 2500 organizmów
BioGrid	baza danych oddziaływań białko-białko (ang. *BIOlogical General Repository for Interaction Datasets*)
BioPAX	język tekstowej reprezentacji szlaków biologicznych (ang. *BIOlogical PAthway eXchange*)
BLAST	algorytm służący do analizy sekwencji nukleotydowych i aminokwasowych (ang. *Basic Local Alignment Search Tool*)
BRENDA	najobszerniejsza kolekcja informacji o enzymach (ang. *BRaunschweig ENzyme DAtabase*)
BSD	licencja zgodna z zasadami Wolnego Oprogramowania, powstała na Uniwersytecie Kalifornijskim w Berkeley (ang. *Berkeley Software Distribution License*)
C/D snoRNP	małe jąderkowe rybonukleoproteiny typu C/D (ang. *C/D box Small NucleOlar RiboNucleoProtein*)
CATH	półautomatyczna, hierarchiczna klasyfikacja domen białkowych (ang. *Class, Architecture, Topology, Homologous superfamily*)
CCR4-NOT	kompleks białek CCR4 oraz NOT (ang. *C-C chemokine Receptor type 4 - Negative regulator Of Transcription*)
ceRNA	współzawodniczące endogenne RNA (ang. *Competing Endogenous RNA*)
CM	model kowariancji (ang. *Covariance Model*)
crasiRNA	RNA związany z centromerem (ang. *centromere repeat associated short interacting RNA*)
CRW	projekt dostarczający informacji o strukturach i ewolucji RNA (ang. *the Comparative RNA Web*)
DAnCER	baza danych powiązanej z chorobami epigenetyki chromatyny (ang. *Disease-Annotated Chromatin Epigenetics Resource*)
DCP2	enzym usuwający czapeczkę mRNA (ang. *mRNA-DeCaPping enzyme 2*)
DDR	bezpośrednia naprawa DNA (ang. *Direct Reversal Repair*)
DDS	sygnalizowanie uszkodzenia DNA (ang. *DNA Damage Signaling*)
Deathbase	baza białek zaangażowanych w śmierć komórki.
Django	wzorzec projektowy, służący do tworzenia serwisów internetowych
DNA	kwas deoksyrybonukleinowy (ang. *DeoxyriboNucleic Acid*)
DNAtraffic	baza danych białek naprawy DNA, szlaków odpowiedzi na uszkodzenia i remodelowania chromatyny
DRY	„Nie Powtarzaj się”, reguła stosowana podczas programowania (ang. *Don’t Repeat Yourself*)
E.C.	Komisja Enzymatyczna (ang. *Enzyme Commission*)
EcoCyc	baza danych genów i matabolizmu *Escherichia coli K12*
EJC	„kompleks złącza egzonów” (ang. Exon Junction Complex)

eRF 1, 2	eukariotyczny czynnik terminacji (ang. *Eukaryotic Release Factor 1, 2*)
Exo 1	egzonukleaza 1 (ang. *Exonuclease 1*)
GGR–NER	globalna naprawa genomu przez wycinanie nukleotydu (ang. *Global Genome Repair- Nucleotide-Excision Repair*)
GPL	licencja wolnego oprogramowania (ang. General Public License)
H/ACA snoRNP	małe jąderkowe rybonukleoproteiny typu H/ACA (ang. *H/ACA box Small NucleOlar RiboNucleoProtein*)
HBV	wirus zapalenia wątroby typu B (ang. *Hepatitis B Virus*)
HCV	wirus zapalenia wątroby typu C (ang. *Hepatitis C Virus*)
HIV	wirus niedoboru odporności (ang. *Human Immunodeficiency Virus*)
HRR	naprawa rekombinacyjna = rekombinacja homologiczna (ang. *Homologous Recombination Repair*)
HTML	hipertekstowy język znaczników (ang. *HyperText Markup Language*)
HTTP	protokół przesyłania dokumentów hipertekstowych (ang. *HyperText Transfer Protocol*)
Human DNA Repair Genes	baza danych ludzkich genów naprawy DNA
InterPro	baza danych białek pogrupowanych w rodziny, z informacją o domenach i miejscach aktywnych.
JmolApplet	wtyczka do przeglądarki struktur przestrzennych
KEGG	encyklopedia genów i genomów (ang. *Kyoto Encyclopedia of Genes and Genomes*)
KISS	„nie komplikuj, głupku”, reguła utrzymania eleganckiej i przejrzystej struktury oprogramowania (ang. *Keep It Simple, Stupid*)
Lig1	DNA ligaza 1 (ang. *DNA Ligase 1*)
lncRNA	długi, regulatorowy, niekodujący RNA (ang. *Long Non-Coding RNA*)
lncRNAdb	baza danych eukariotycznych długich niekodujących RNA (ang. *Long Non-Coding RNA DataBase*)
LP-BER	długa naprawa przez wycinanie zasady (ang. *Long Patch Base-Excision Repair*)
MarvinSketch	zaawansowany edytor do rysowania struktur chemicznych
MetaCyc	baza danych szlaków metabolicznych z 2414 różnych organizmów
miRBase	baza poświecona miRNA (ang. *MIcroRNA DataBase*)
miRNA	mikroRNA (ang. *MIcroRNA*)
Mlh1	homolog białka MutL (ang. *MutL homolog 1*)
MMR	naprawa błędnie sparowanych zasad (ang. *MisMatch Repair*)
MODOMICS	baza danych o szlakach modyfikacji RNA
moRNA	RNA pochodzące z prekursorów mikroRNA (ang. *microRNA-offset RNA*)
mRNA	matrycowy RNA (ang. *messenger RNA*)
MSY-RNA	niezależny od białek MIWI RNA (ang. *MIWI-independent small RNA*)
MTV	Model – Szablon – Widok (ang. *Model-Template-View*)
MutLα	Hetero dimer o aktywności nukleazowej, biorący udział w naprawie DNA (ang. *MUTator L α*)
MutSα	heterodimer, wiążący się do niesparowań w DNA, ropoczynając tym samym naprawę DNA (ang. *MUTator S α*)
MVC	Model – Widok – Kontroler (ang. *Model – View – Controller*)
MVP	Model – Widok – Prezenter (ang. *Model – View – Presenter*)
MySQL	system zarządzania bazą danych
NCBI	Narodowe Centrum Informacji Biotechnologicznej (ang. *National Center for Biotechnology Information*)
ncRNA	niekodujący RNA (ang. *non-coding RNA*)
NER	naprawa przez wycinanie nukleotydu (ang. *Nucleotide-Excision Repair*)

NGD	degradacja transkryptów, w których z powodu występujących struktur drugorzędowych rybosom ulega zatrzymaniu (ang. *No-Go Decay*)
NHEJ	naprawa poprzez łączenie niehomologicznych końców (ang. *Non-Homologous End Joining repair*)
NMD	degradacja transkryptów, posiadających przedwczesny kodon STOP (ang. *Nonsense-Mediated Decay*)
NMR	spektroskopia magnetycznego rezonansu jądrowego (ang. *Nuclear Magnetic Resonance*)
NONCODE	baza danych niekodujących RNA
NRD	degradacją niefunkcjonalnego rRNA (ang. *Nonfunctional rRNA Decay*)
NSD	degradacja mRNA z brakującymi kodonami terminacji (ang. *NonStop decay*)
OMIM	baza danych o wszystkich opisanych chorobach uwarunkowanych genetycznie (ang. *Online Mendelian Inheritance in Man*)
OriDB	baza danych potwierdzonych i przewidzianych miejsc rozpoczęcia replikacji DNA (ang. *DNA replication ORIgin DataBase*)
ORM	mapowanie obiektowo-relacyjne (ang. *Object-Relational Mapping*)
PAN 2,3	nukleaza poli(A) 1,2 (ang. *Poly(A)-Nuclease 1,2*)
PAR	RNA związany z promotorem (ang. *Promoter-Associated RNA*)
Pathway Commons	zbiór danych dotyczących publicznie dostępnych szlaków z różnych organizmów.
PCNA	antygen jądrowy proliferujących komórek (ang. *Proliferating Cell Nuclear Antigen*)
PDB	baza danych struktur przestrzennych białek i kwasów nukleinowych (ang. *Protein Data Bank*)
PFAM	baza danych rodzin białek (ang. *Protein FAMily database*)
PIN	domena końca N białka PilT (ang. *PIlT N terminus*)
piRNA	RNA oddziałujący z białkami PIWI (ang. *Piwi-Interacting RNA*)
PMID	identyfikator bazy PubMed (ang. *PubMed IDentifier*)
Pms2	białko naprawy niesparowań w DNA (ang. *DNA MiSmatch repair Protein 2*)
PNAS	czasopismo Narodowej Akademia Nauk Stanów Zjednoczonych (ang. *Proceedings of the National Academy of Sciences*)
PNG	rastrowy format plików graficznych (ang. *Portable Network Graphics*)
PNPaza	fosforylaza polinukleotydowa (ang. *PNPase, PolyNucleotide PhosphorylASE*)
PTC	centrum peptydylotransferazy (ang. *Peptidyl Transferase Center*)
PubMed	baza danych publikacji
PyGraphviz	biblioteka języka Python służaca do rysowania grafów
Reactome	baza danych głównie ludzkich szlaków metabolicznych i składających się na te szlaki reakcji
REBASE	zbiór informacji o enzymach i genach związanych z restrykcją i modyfikacją DNA u Prokaryota (ang. *the Restriction Enzyme dataBASE*)
Repair-FunMap	baza danych oddziaływań pomiędzy białkami naprawy DNA a białkami zaangażowanymi w inne procesy.
ReplicationDomain	baza danych i narzędzia służące do analizy replikacji DNA.
RFAM	baza danych rodzin RNA (ang. *RNA FAMily database*)
RFC	czynnik replikacyjny C (ang. *Replication Factor C*)
RNA	kwas rybonukleinowy (ang. *RiboNucleic Acid*)
RNA Modification Database	baza danych modyfikacji RNA
RNaza	rybonukleaza (ang. *RNase, RiboNucleASE*)
RPA	białko replikacyjne A (ang. *Replacation Protein A*)
rrn A-E, H, G	operon rRNA A-E, H, G (ang. *Ribosomal RNA operoN*)
rRNA	rybosomowy RNA (ang. *ribosomal RNA*)

RTD	nagła degradacja tRNA (ang. *Rapid TRNA Decay*)
SCOP	strukturalna klasyfikacja domen białkowych (ang. *Structural Classification Of Proteins*)
sdRNA	RNA pochodzący z snoRNA (ang. *Small RNAs Derived from snoRNAs)*
siRNA	krótki interferujący RNA (ang. *Small Interfering RNA*)
SMBL	Język Znaczników Biologii Systemów (ang. *Systems Biology Markup Language*)
SMG 1, 5, 6, 7	kinaza serynowo/treoninowa 1, 5, 6, 7 (ang. *serine/threonine protein kinase 1, 5, 6, 7*)
SMG1C	białkowy kompleks kinaz
SMILES	sposób jednoznacznego zapisu struktury cząsteczek związków chemicznych z wykorzystaniem ciągu znaków ASCII (ang. *Simplified Molecular-Input Line-Entry System*)
snRNA	mały jądrowy RNA (ang. *small nucelar RNA*)
snoRNA	mały jąderkowy RNA (ang. *small nucleolar RNA*)
SP-BER	krótka naprawa przez wycinanie zasady (ang. *Short Patch Base-Excision Repair*)
SpliceDisease Database	baza danych mutacji i chorób związanych z nieprawidłowym wycinaniem intronów
SQL	strukturalny język zapytań (ang. Structured Query Language)
SQLite	system zarządzania bazą danych
SURF	kompleks białek SMG1-UPF1-eRF1-eRF2 (ang. *SMG1-UPF1-eRF1-eRF3 complex*)
SVG	format dwuwymiarowej grafiki wektorowej (ang. *Scalable Vector Graphics*)
TCR–NER	naprawa przez wycinanie nukleotydu sprzężona z transkrypcją (ang. *Transcription-Coupled Repair – Nucleotide-Excision Repair*)
Telomerase database	baza danych sekwencji i struktur podjednostek telomerazy wraz z informacjami o ich mutacjach.
tel-sRNA	krótki, telomerowy RNA (ang. *TELomeric-specific Small RNA*)
tiRNA	RNA pochodzący z tRNA (ang. *stress-induced tRNA-derived RNAs*)
TLS	synteza DNA ponad uszkodzonym miejscem na matrycy (ang. *TransLesion Synthesis*)
tmRNA	transferowo-matrycowy RNA (ang. *transfer-messenger RNA*)
tRNA	transferowy RNA (ang. *transfer RNA*)
tRNAdb	baza danych genów i sekwencji tRNA
U3 snoRNP	małe jąderkowe rybonukleoproteiny typu U3 (ang. *U3 Small NucleOlar RiboNucleoProtein*)
UniProt	baza danych sekwencji białkowych (ang. *the UNIversal PROTein resource*)
UPF 1,2,3	białko supresorowe przesunięcia ramki odczytu 1, 2, 3 (ang. *UP-Frameshift suppressor 1,2,3*)
URL	ujednolicony format adresowania zasobów (ang. *Uniform Resource Locator*)
UV	promieniowanie ultrafioletowe (ang. *UltraViolet*)
UvrA	białko naprawy DNA o aktywości ATPazy
VARNA	wtyczka dedykowana rysowaniu struktur drugorzędowych RNA
wiki	strona internetowa pozwalająca na dodawanie, modyfikowanie czy usuwanie treści poprzez przeglądarkę internetową
xiRNA	RNA odpowiedzialny za inaktywację chromosomu X, ang. *RNA responsible for X-inactivation*)
XML	rozszerzalny język znaczników (ang. *eXtensible Markup Language*)
XRN1	5′-3′ egzorybonukleaza (ang. *5′-3′ eXoRiboNuclease*)

Streszczenie

Głównym celem projektu było stworzenie elementów składowych ujednoliconego systemu bazodanowego opisującego cały metabolizm kwasów nukleinowych. Taki system pozwoli na szersze spojrzenie na biologię systemów tych związków, dogłębną analizę poszczególnych szlaków, systematyzację wiedzy, a także udostępnienie istniejących danych w sposób jasny, przejrzysty i czytelny szerokiemu gronu naukowemu. Ponadto umożliwi on w przyszłości symulację szlaków metabolicznych oraz odkrycie i opisanie nowych, nieznanych do tej pory elementów zaangażowanych w procesy dojrzewania, modyfikacji i degradacji kwasów nukleinowych. Otrzymana struktura danych powinna nie tylko umożliwić zastosowanie podejścia systemowego w analizie biologii kwasów nukleinowych, ale także wygenerować wiedzę, która będzie czymś więcej niż jedynie sumą zgromadzonych cząstkowych informacji. **Opracowanie tego typu podejścia będzie stanowiło krok milowy w dziedzinie biologii systemów kwasów nukleinowych, pozwoli lepiej zrozumieć wiele ważnych procesów biologicznych i da silny impuls do dalszych analiz doświadczalnych**.

W ramach niniejszego projektu doktorskiego opracowane zostały dwie bazy danych: REPAIRtoire – baza danych szlaków naprawy DNA i RNApathwaysDB – baza danych dojrzewania i degradacji RNA. Jest on projektem o charakterze bioinformatycznym, więc najważniejszymi stosowanymi podczas jego realizacji metodami były metody komputerowe - języki programowania, wzorzec projektowy (ang. *framework*) Django, metody eksploracji danych (ang. *data mining*) i eksploracji tekstu (ang. *text mining*).

Pierwszą z opracowanych baz była REPAIRtoire – baza danych szlaków naprawy DNA. Choć temat naprawy DNA jest poruszany przez wiele grup badawczych, a informacje znaleźć można w wielu źródłach, do tej pory nie powstała żadna specjalistyczna baza danych na ten temat. REPAIRtoire to pierwsza taka baza danych dostępna w Internecie. Szlaki naprawy DNA są niezwykle ważnymi procesami, ponieważ występowanie uszkodzeń w DNA jest czynnikiem etiologicznym wielu chorób (w tym wielu nowotworów oraz niedokrwistości Fanconiego). Celem istnienia tej bazy jest nie tylko zgromadzenie informacji na temat wszystkich systemów naprawy DNA, ale także usystematyzowanie wiedzy na temat powiązania ludzkich chorób z mutacjami występującymi w genach odpowiedzialnych za integralność i stabilność DNA oraz dostarczenie danych o źródłach i rodzajach toksycznych i mutagennych czynników uszkadzających DNA. REPAIRtoire to obszerne naukowe źródło informacji na temat szlaków naprawy DNA oraz cząsteczek biorących udział

w składających się na nie reakcjach (takich jak białka, cząsteczki RNA czy kompleksy enzymatyczne), z uwzględnieniem ich struktur i kodujących je genów. Co więcej wszystkie dane uzupełnione są o odsyłacze do odpowiedniej literatury i ogólnych baz danych. Baza ta ułatwia analizę systemów naprawy DNA i pozwala na łatwy dostęp do wiedzy na ten temat. Jest ona cały czas rozwijana i uzupełniana nowymi danymi. Została opublikowana w Nucleic Acid Research i jest dostępna pod adresem: http://repairtoire.genesilico.pl (Milanowska, Krwawicz i wsp. 2011).

Drugą opracowaną bazą jest RNApathwaysDB – baza danych dojrzewania i degradacji RNA. Celem tego projektu było stworzenie źródła wiedzy dotyczącej szlaków procesowania RNA, od dojrzewania po degradację. Baza ta gromadzi wysokiej jakości informacje dotyczące reakcji takich jak: wycinanie intronów (ang. *splicing*), dojrzewanie i degradacja tRNA, mRNA, rRNA, powstawanie tiRNA, a także dane o cząsteczkach biorących w nich udział: rybozymach, ryboprzełącznikach, czynnikach regulatorowych, procesowanych RNA, białkach i kompleksach enzymatycznych. Wszystkie informacje uzupełnione są odnośnikami literaturowymi oraz odsyłaczami do innych baz danych. Baza ta została także opublikowana w Nucleic Acid Research i jest dostępna pod adresem: http://genesilico.pl/rnapathwaysdb (Milanowska, Mikolajczak i wsp. 2013).

Najważniejsze narzędzia stosowane podczas realizacji części projektu dotyczącej tworzenia baz danych to języki programowania: Python oraz JavaScript, języki bazodanowe: MySQL oraz SQLite3, język znaczników HTML oraz wzorzec projektowy Django. Wszystkie te narzędzia pozwalają na zaprojektowanie oraz wdrożenie systemów bazodanowych.

Opisane bazy danych, razem z bazą danych szlaków modyfikacji RNA – Modomics (http://modomics.genesilico.pl), w przyszłości zostaną połączone w jeden spójny system. W systemie tym bazy będą nawzajem korzystać ze swoich danych, co spowoduje integrację zasobów i rozszerzenie funkcjonalności i możliwości stworzonych narzędzi.

1 Wstęp

1.1 Kwasy nukleinowe

Kwasy nukleinowe występują w komórkach wszystkich organizmów żywych na Ziemi i są niezmiernie ważnym elementem ich funkcjonowania. Pełnią funkcje nośników informacji genetycznej, pośredniczą w produkcji białek, działają jako regulatory ekspresji genów oraz przeprowadzają reakcje katalityczne (np. rybozymy). Do dogłębnego poznania i zrozumienia funkcjonowania żywych organizmów niezbędna jest wiedza nie tylko odnośnie samych kwasów nukleinowych i pełnionych przez nie funkcji, ale także o wszystkich procesach prowadzących do powstania tych cząsteczek, ich modyfikacji czy degradacji.

1.1.1 Budowa i funkcje DNA

Kwas deoksyrybonukleinowy (DNA) jest liniowym, nierozgałęzionym biopolimerem, który wraz z kwasem rybonukleinowym (RNA) i białkami, jest jedną z biologicznie najważniejszych makromolekuł. Koduje on genetyczne „instrukcje" niezbędne do rozwoju i funkcjonowania wszystkich żywych organizmów i niektórych wirusów. Kwas ten został po raz pierwszy wyizolowany w 1869 roku przez szwedzkiego fizyka Friedricha Mieschera jako składowa substancji, którą, z racji tego, że znaleziona została w jądrze komórkowym (ang. *nucleus*) Miescher nazwał „nukleiną" (Dahm 2005). Od tego czasu wielu badaczy po kolei wyjaśniało tajemnicę tej „substancji": od 1878 roku i izolacji kwasu nukleinowego z nukleiny przeprowadzonej przez Albrechta Kossela, przez 1919, gdy Levene (Levene 1917) zidentyfikował poszczególne jednostki budulcowe kwasu, po 1952, gdy Alfred Hershey i Martha Chase w eksperymencie Hershey–Chase udowodnili, że DNA posiada funkcję przekazywania informacji genetycznej (Hershey i Chase 1952). A następnie aż po rok 1953 i opisanie przez Jamesa Watsona i Francisa Cricka podwójnej helisy DNA (Watson i Crick 1953) na podstawie obrazu dyfrakcji rentgenowskiej uzyskanego przez Rosalind Franklin i Raymonda Goslinga w maju 1952 (nazwanego *„Photo 51"*) (Franklin i Gosling 1953), za co w 1962 roku, po śmierci Franklin, Watson, Crick i Wilkins otrzymali Nagrodę Nobla. Kolejnym krokiem milowym w badaniach nad DNA było przedstawienie przez Cricka w 1957 głównego założenia biologii molekularnej, mówiącego, że pomiędzy DNA, RNA i białkami występują wzajemne powiązania i sformułowania

„hipotezy adaptorowej" (ang. *adaptor hypothesis*), która wnioskowała, że informacja genetyczna zostaje przepisana z DNA na białka za pomocą miejsc adaptorowych na kwasie nukleinowym, do których przyłączają się odpowiednie aminokwasy. Dalsze prace Cricka i współpracowników pokazały, że kod genetyczny bazuje na nienakładających się trójkach nukleotydów, zwanych kodonami, co pozwoliło Har G. Khoranie, Robertowi W. Holleyowi i Marshallowi W. Nirenbergowi na rozszyfrowanie kodu genetycznego. Odkrycia te dały początek biologii molekularnej.

1.1.1.1 Budowa i organizacja DNA

Informacja genetyczna zakodowana jest jako sekwencja reszt nukleotydów, zbudowanych z pięciowęglowego cukru deoksyrybozy, którego grupa hydroksylowa znajdująca się przy ostatnim atomie węgla jest zestryfikowaną resztą fosforanową, a pierwszy atom węgla połączony jest wiązaniem N-glikozydowym z jedną z czterech zasad azotowych (guanina – G, adenina – A, tymina – T (lub bardzo rzadko uracyl – U, np. u bakteriofaga Yersinia piR1-37 (Kiljunen, Hakala i wsp. 2005)) i cytozyna – C). Zasady tworzą ze sobą pary połączone wiązaniami wodorowymi (dwoma pomiędzy A a T, trzema między G a C), dzięki czemu w większości cząsteczki DNA występują w postaci dwuniciowej prawoskrętnej helisy, w której na zewnątrz znajdują się rdzenie cukrowo-fosforanowe, a zasady skierowane są do wewnątrz. Każda z nici DNA ma na jednym końcu (oznaczanym jako koniec 5′) wolną grupę fosforanową przy węglu 5′ deoksyrybozy, a na drugim końcu (oznaczanym jako koniec 3′) wolną grupę hydroksylową przy węglu 3′ deoksyrybozy. Ze względu na to, że helisa dwóch nici DNA jest spleciona w taki sposób, że jedna z nici zaczyna się od końca 5′ a druga od końca 3′, mówi się, że obie nici są względem siebie antyrównoległe. Strukturę taką stabilizują nie tylko wiązania wodorowe między komplementarnymi zasadami, ale również tzw. oddziaływania warstwowe zasad ustawionych prostopadle do osi helisy (Yakovchuk, Protozanova i wsp. 2006). Struktura podwójnej helisy zapewnia również duplikację na każdej z nici informacji genetycznej, co jest niezwykle istotne dla wszystkich funkcji DNA w żywych komórkach.

W komórkach eukariotycznych DNA przechowywany jest w jądrze komórkowym oraz dodatkowo w niektórych organellach takich jak mitochondria czy plastydy, podczas gdy w komórkach prokariotycznych DNA w postaci pojedynczej, kolistej cząsteczki znajduje się w cytoplazmie. W każdym z przypadków, DNA upakowany jest w długie struktury zwane chromosomami. Podczas podziałów komórkowych chromosomy ulegają podwojeniu w procesie zwanym replikacją DNA, co pozwala zaopatrzyć każdą z nowych komórek

w pełen zestaw chromosomów. Chromosomy posiadają na końcach struktury zwane telomerami, których główną funkcją jest ochrona końców DNA i powstrzymanie procesów naprawy DNA przed traktowaniem tych końców jako uszkodzonych (Wright, Tesmer i wsp. 1997). Ekspresja genów zależy od tego, w jaki sposób DNA zorganizowany jest w chromosomach w wyższą strukturę zwaną chromatyną. Za tworzenie struktur wyższych, a więc pośrednio też za regulację ekspresji genów odpowiadać mogą np. modyfikacje zasad w DNA. Przykładem mogą tu być regiony o wysokiej zawartości zmetylowanych reszt cytozyny, np. 5-metylocytozyna jest ważna przy inaktywacji chromosomu X (Klose i Bird 2006). Pomimo swojej funkcji zmetylowane reszty cytozyny są bardzo podatne na mutacje, ze względu na możliwość ich deaminacji do tyminy. Za regulację ekspresji genów odpowiadać mogą także modyfikacje kowalencyjne białek histonowych, dookoła których DNA nawinięty jest w chromatynie, bądź procesy remodelowania chromatyny (Hu i Rosenfeld 2012).

1.1.1.2 Funkcja DNA

Kwas deoksyrybonukleinowy przechowuje informację genetyczną niezbędną do funkcjonowania, wzrostu i rozmnażania się wszystkich żywych organizmów.

Jak wcześniej wspomniano, DNA w komórkach występuje w postaci chromosomów: liniowych u Eukaryota albo cyrkularnych u Prokaryota. Pełen zestaw chromosomów tworzy genom, np. u człowieka genom składa się z 46 chromosomów zbudowanych z ok. 3 mld par zasad (pz) (Venter, Adams i wsp. 2001). Informacja genetyczna, określająca fenotyp organizmu, przechowywana jest w jednostkach zwanych genami, a jej przepisanie na białko możliwe jest dzięki komplementarności zasad w sekwencji mRNA, kodowanej przez gen, z resztami nukleotydowymi występującymi w antykodonie tRNA. Sekwencja mRNA określa sekwencję jednego i/lub kilku białek. Procesy odpowiedzialne za przepisywanie informacji z DNA na białko to transkrypcja i translacja. Funkcja przechowywania informacji genetycznej przez kwas deoksyrybonukleinowy jest kluczowa dla życia i dlatego też wszystkie procesy związane z jej obsługą są pod ścisłą kontrolą, gdyż błędy pojawiające się w ich czasie mogą mieć efekt prowadzący nawet do śmierci organizmu.

Podziały komórkowe są niezbędne do wzrostu organizmu. Podczas podziału, każda komórka potomna musi otrzymać swoją własną, pełną, identyczną do matczynej kopię genomu. Procesem odpowiedzialnym za powielanie informacji genetycznej jest replikacja DNA. Proces ten też podlega ścisłej kontroli.

Wszystkie funkcje DNA zależą od jego oddziaływań z białkami, które mogą być niespecyficzne, bądź ściśle wynikające z lokalnej sekwencji lub struktury DNA (białka wiążą się wtedy tylko do ściśle określonej sekwencji lub struktury). Oddziaływania DNA z białkami odpowiadają za jego strukturę i organizację w komórce (np. białka histonowe). Ponadto są one odpowiedzialne także za procesy replikacji, rekombinacji czy naprawy (np. wiążące jednoniciowy DNA ludzkie białko RPA (ang. *replication protein A*) (Iftode, Daniely i wsp. 1999)), procesy transkrypcji (np. czynniki transkrypcyjne, które są specyficzne względem sekwencji (Myers i Kornberg 2000, Spiegelman i Heinrich 2004), a zamiana aktywności jednego z nich może wpłynąć na ekspresję tysięcy genów (Li, Van Calcar i wsp. 2003)), czy modyfikacji DNA.

DNA wykorzystywany jest także w technologii. Inżynieria genetyczna dostarcza metod do izolacji DNA z komórek i manipulowania nim w laboratorium. Nowoczesna biologia i biochemia korzystają z tych metod w technologiach rekombinowanego DNA, dzięki którym uzyskiwać można sztucznie zaprojektowane cząsteczki, które po znalezieniu się w organizmie odpowiedzialne będą za produkcję np. zrekombinowanych białek, używanych m.in. w medycynie (Houdebine 2007). W sądownictwie badania DNA znalezionego w miejscach zbrodni dostarczają informacji o osobach zaangażowanych w zdarzenie. Do badań tych należy m.in. technika zwana profilowaniem DNA (ang. „*DNA profiling*") bądź „genetycznym odciskiem palca" (ang. *"genetic fingerprinting"*). W nanotechnologii DNA "zmusza" się cząsteczkę do przybierania odpowiednich form i struktur o określonych, pożądanych właściwościach (Rothemund 2006). Ponadto dzięki najnowszym osiągnięciom DNA może być wykorzystywany do przechowywania danych cyfrowych. Grupa Goldmana zakodowała za pomocą sekwencji reszt w DNA 739 kb danych tekstowych i dźwiękowych. Następnie DNA został zsyntetyzowany, a później zsekwencjonowany. Uzyskana informacja była w 100% zgodna z informacją zakodowaną pierwotnie (Goldman, Bertone i wsp. 2013).

1.1.2 Budowa, funkcje i rodzaje RNA

Kwas rybonukleinowy (RNA) występuje we wszystkich organizmach i do niedawna uważano, że większość jego cząsteczek funkcjonuje przede wszystkim w roli pośredników w przekazywaniu informacji genetycznej pomiędzy DNA a aktywnymi biologicznie białkami. Było to zgodne z główną doktryną biologii molekularnej, opracowaną na podstawie badań prostych organizmów takich jak *Escherichia coli* (Crick 1958; Crick, 1968). Jednakże pogląd

ten zmienił się dzięki odkryciu dokonanemu przez zespoły Sidneya Altmana (Guerrier-Takada, Gardiner i wsp. 1983) i Thomasa Cecha (Kruger, Grabowski i wsp. 1982) (naukowcy wyróżnieni nagrodą Nobla w dziedzinie chemii w 1989 roku) katalitycznych właściwości RNA. RNA jest obecnie uważany za jedyną cząsteczkę, która poza przechowywaniem informacji genetycznej, ma potencjał katalityczny (Westhof i Auffinger 2006) (DNAzymy, cząsteczki DNA o właściwościach katalitycznych, pochodzą z selekcji *in vitro* i nie zostały zaobserwowane w naturze (Mastroyiannopoulos, Uney i wsp. 2010)). Twierdzenie to wspiera hipotezę „świata RNA", według której w początkowych etapach ewolucji życia na Ziemi cząsteczki RNA pełniły rolę zarówno nośnika informacji genetycznej, jak i enzymu dokonującego jej powielenia. Termin ten został użyty po raz pierwszy w 1986 roku przez Waltera Gilberta (Walter 1986), jednak sama koncepcja została przedstawiona już we wcześniejszych pracach, m.in. Francisa Cricka (Crick 1968) oraz Leslie Orgela (Orgel 1968). Hipoteza ta mówi także o tym, że z czasem funkcję głównego nośnika informacji genetycznej przejęły cząsteczki kwasu deoksyrybonukleinowego, a białka zastąpiły cząsteczki RNA w funkcjach enzymatycznych.

By cząsteczki RNA mogły przeprowadzać reakcje katalityczne wymagane jest precyzyjne ułożenie atomów w przestrzeni. Dlatego też cząsteczki RNA charakteryzuje występowanie zróżnicowanej i jednocześnie dokładnie zdefiniowanej struktury (Shazman, Elber i wsp. 2011) (patrz: Podrozdział: 1.1.2.1). Znajomość struktury i jej związek z funkcją pozwalają również wnioskować o mechanizmach działania, np. rybozymów i RNA regulatorowych (Hoogstraten i Sumita 2007).

1.1.2.1 Budowa RNA

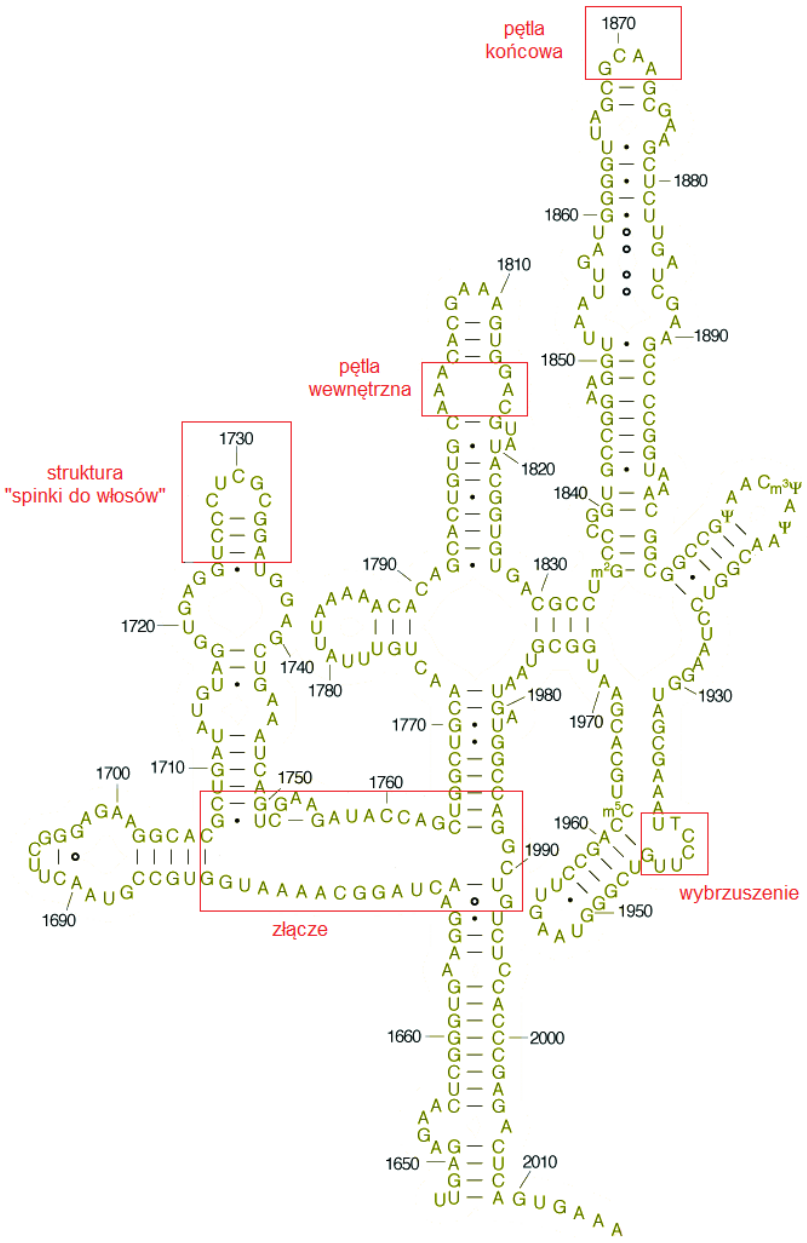
RNA jest liniowym polimerem, w którym monomerami są reszty rybonukleotydowe połączone za pomocą wiązań 5′-fosfodiestrowych. Każdy rybonukleotyd zbudowany jest z nukleozydu (złożony z pięciowęglowego cukru rybozy i jednej z czterech zasad azotowych (adeniny – A, guaniny – G, cytozyny – C lub uracylu – U) oraz reszty kwasu fosforowego. Pod względem budowy chemicznej RNA przypomina DNA, jednakże różni się od niego dwiema podstawowymi cechami. Po pierwsze w RNA zamiast tyminy występuje uracyl. Po drugie, cukrem wchodzącym w skład nukleozydu RNA jest ryboza, która w pozycji 2′ pierścienia pentozy posiada grupę hydroksylową zamiast wodoru charakterystycznego dla DNA. Grupa ta może, w odróżnieniu od wodoru, uczestniczyć w tworzeniu skomplikowanej sieci oddziaływań, powodując, że RNA łatwiej niż DNA tworzy skomplikowane struktury przestrzenne (Salazar, Fedoroff i wsp. 1993,

Westhof, 2006). Pod względem budowy przestrzennej RNA jest więc bardziej zbliżone do białek. Analogia do białek dotyczy również funkcji. Oprócz swojej roli w biosyntezie białek RNA jest katalizatorem wielu reakcji chemicznych oraz reguluje ekspresję genów. Ponadto, podobnie jak w białkach, w RNA często występują potranskrypcyjne modyfikacje reszt, które odgrywają rolę ważną dla struktury i funkcji cząsteczki.

Pełnienie przez cząsteczki RNA wyżej wymienionych funkcji wynika z przyjmowanych przez nie struktur przestrzennych. Przyjęcie odpowiedniej konformacji łańcucha polinukleotydowego jest niezbędnym warunkiem, który cząsteczki RNA muszą spełnić, aby przejawiały swą aktywność biologiczną. M.in. cięcie RNA przez RNazę P (Guerrier-Takada, Gardiner i wsp. 1983) oraz tworzenie wiązania peptydowego poprzez aktywność peptydylo-transferazy rRNA (Ban, Nissen i wsp. 2000, Schluenzen, 2000, Wimberly, 2000) jest możliwe dzięki temu, że te cząsteczki RNA przyjmują odpowiednią strukturę .

Struktura RNA jest hierarchiczna i przedstawia się następująco:

- struktura pierwszorzędowa – liniowe ułożenie reszt nukleotydów w cząsteczce;
- struktura drugorzędowa – układ regionów podwójnej helisy utworzonych przez pary zasad typu Watson-Crick, połączonych regionami niesparowanymi; do elementów struktury drugorzędowej zalicza się: pętlę końcową w strukturze „spinki do włosów”, pętlę wewnętrzną (dwa regiony jednoniciowe, łączące dwa segmenty podwójnej helisy), wybrzuszenie (w przypadku, gdy region niesparowany występuje tylko po jednej stronie), złącze (trzy lub więcej połączonych segmentów dwuniciowych) (patrz: Rycina 1) (Holbrook 2008);
- struktura trzeciorzędowa – wzajemne ułożenie przestrzenne elementów struktury drugorzędowej, na które wpływ mają oddziaływania Van der Waalsa, wiązania wodorowe pomiędzy zasadami azotowymi (w tym parowania niekanoniczne), oddziaływania elektrostatyczne oraz oddziaływania warstwowe (ang. *stacking*).



Rycina 1 Podstawowe typy elementów struktury drugorzędowej RNA na podstawie domeny IV 23S rRNA z *Escherichia coli.*

1.1.2.2 Rodzaje RNA

Organizmy prokariotyczne posiadają genom złożony w przeważającej ilości z sekwencji kodujących białka (80–95% całego genomu) (Mattick i Makunin 2006). Na uwagę zasługuje fakt, że proteom (zestaw białek danego organizmu) prokariotyczny różni się często wyraźnie nawet pomiędzy blisko spokrewnionymi szczepami bakterii (Hayashi, Makino i wsp. 2001). Zjawisko to związane jest głównie z horyzontalnym transferem genów. Natomiast w przypadku organizmów eukariotycznych proporcja sekwencji

DNA kodujących białka względem pozostałej części DNA zmniejsza się wraz ze wzrostem złożoności organizmu. Tutaj jako przykłady służyć mogą: drożdże (70% genomu stanowią sekwencje kodujące białka), *Arabidopsis thaliana* (29%), *Drosophila melanogaster* (20%), człowiek (tylko około 1,4%) (Sana, Faltejskova i wsp. 2012). Powyższa prawidłowość przekłada się na różnice w składzie transkryptomów mikroorganizmów i organizmów wielokomórkowych. U mikroorganizmów przeważa mRNA kodujący białka, podczas gdy u organizmów wielokomórkowych głównym składnikiem jest RNA niekodujący białek (ncRNA, ang. *non-coding RNA*). Ponadto różnicę widać także w stopniu skomplikowania maszynerii odpowiedzialnej za przetwarzanie RNA i przekazywanie sygnałów, która jest zdecydowanie bardziej złożona u organizmów eukariotycznych, co związane jest z bardziej złożonymi szlakami regulacji genów u Eukaryota. Przykładem tu służyć mogą: interferencja RNA (Fire, Xu i wsp. 1998), metylacja DNA (Cedar i Bergman 2012) czy naznaczenie genetyczne (ang. *genomic imprinting*) (Barlow 2011). Ponadto wyższe organizmy mają także większy zestaw białek oddziałujących z RNA, a ostatnio zaproponowano hipotezę o współzawodniczących endogennych RNA (ceRNA, ang. *competing endogenous RNA hypothesis*) (Salmena, Poliseno i wsp. 2011), według której zarówno kodujące jak i niekodujące cząsteczki RNA tworzą sieć regulatorową, która wpływa na ekspresję informacji genetycznej. Ta różnorodność i funkcje pełnione przez niekodujące RNA w komórkach wyższych organizmów eukariotycznych najlepiej odzwierciedlone są poprzez liczbę różnych klas RNA. Najważniejsze obecnie znane klasy niekodujących cząsteczek RNA, poza rybosomalnym RNA (rRNA) i transportującym RNA (tRNA), ssaków to:

- **crasiRNA** (RNA związany z centromerem, ang. *centromere repeat associated short interacting RNA*) – klasa niekodujących cząsteczek RNA o długości ok. 34–42 nt, które odpowiedzialne są za prowadzenie procesu modyfikacji chromatyny (Carone, Longo i wsp. 2009);
- **introny oraz rejony 3′ i 5′ UTR w mRNA** – grupa niekodujących fragmentów w mRNA; rejony UTR (ang. *untranslated regions*) zlokalizowane są na końcach 3′ oraz 5′ mRNA, z którymi oddziałują białka regulujące ekspresję informacji genetycznej (Mazumder, Seshadri i wsp. 2003, Pickering i Willis 2005);
- **lncRNA** (długi, regulatorowy, niekodujący RNA, ang. *long non-coding RNA*) – wszystkie cząsteczki RNA niekodujące białek i dłuższe niż 200 nt; ich rolą jest m.in. epigenetyczna regulacja ekspresji genów oraz funkcjonowanie jako

kotwice kompleksów białkowych (Mercer, Dinger i wsp. 2009, Wilusz, Sunwoo i wsp. 2009);

- **miRNA** (mikroRNA) – krótkie sekwencje RNA, długości ok. 22 nt, powstające w wyniku cięcia przez enzym Dicer pierwotnych transkryptów; przyjmujących struktury tzw. spinek do włosów; miRNA oddziałują z białkami z rodziny Argonaute i odpowiedzialne są głównie za potranskrypcyjną regulację ekspresji genów (Carthew i Sontheimer 2009, Ghildiyal i Zamore 2009, Winter, Jung i wsp. 2009);
- **moRNA** (ang. *microRNA-offset RNA*) – małe, ok. 20 nt. długości, cząsteczki RNA pochodzące z rejonów sąsiadujących z regionami kodującymi;
- **pre-miRNA** – funkcja nieznana (Langenberger, Bermudez-Santana i wsp. 2009, Shi, Hendrix i wsp. 2009);
- **MSY-RNA** (RNA niezależny od białek MIWI, ang. *MIWI-independent small RNA*) – RNA długości ok. 26–30 nt. o nieznanej funkcji; podobnie jak w przypadku piRNA jego występowanie jest w ogromnej większości ograniczone do komórek linii płciowej (Xu, Medvedev i wsp. 2009);
- **PAR** (RNA związany z promotorem, ang. *promoter-associated RNA*) – ogólny termin określający zarówno krótkie, jak i długie cząsteczki RNA związane z promotorem, a także RNA biorący udział w inicjacji transkrypcji; regulują ekspresję genów (Belostotsky 2009, Taft, Kaplan i wsp. 2009);
- **piRNA** (RNA oddziałujący z białkami PIWI, ang. *Piwi-interacting RNA*) - krótki RNA, dł. ok. 26–30 nt., występuje głównie w komórkach linii płciowej i komórkach somatycznych powstających bezpośrednio z komórek linii płciowej, oddziałuje z grupą białek z rodziny Argonaute (białka PIWI), wpływa na aktywność transpozonów oraz stopień kondensacji chromatyny (Ghildiyal i Zamore 2009, Malone i Hannon 2009);
- **sdRNA** (RNA pochodzący z snoRNA, ang. *small RNAs derived from snoRNAs*) – małe cząsteczki RNA, z których część powstaje w wyniku działania enzymu Dicer, pochodzące z małego, jąderkowego RNA (snoRNA); niektóre z sdRNA odgrywają, podobnie jak miRNA, role w regulacji translacji (Ender, Krek i wsp. 2008, Saraiya i Wang 2008, Taft, Glazov i wsp. 2009);
- **siRNA** (krótki, interferujący RNA, ang. *small interfering RNA*) – RNA o długości 21–22 nt, powstający w wyniku cięcia dwuniciowych cząsteczek

RNA przez enzym Dicer; tworzy kompleksy z białkami z rodziny Argonaute, których aktywność związana jest z regulacją ekspresji genów, kontrolą transpozonów, a także obroną przed wirusami (Brodersen i Voinnet 2006, Carthew i Sontheimer 2009, Ghildiyal i Zamore 2009, Malone i Hannon 2009);

- **snRNA** (małe jądrowe RNA ang. *small nuclear RNA*) występujący w jądrze komórkowym niekodujący RNA, długości ok. 150 nt, pełniący funkcję rybozymu w procesie wycinania intronów (Matera, Terns i wsp. 2007);
- **snoRNA** (mały, jąderkowy RNA, ang. *small nucleolar RNA*) – cząsteczki RNA (dł. od kilkudziesięciu do kilkuset nt.), wiążące się z metylacją i pseudourydylacją rRNA i snRNA; istnieją dowody świadczące o ich udziale w regulacji ekspresji genów (Matera, Terns i wsp. 2007);
- **tel-sRNA** (krótki, telomerowy RNA, ang. *telomeric-specific small RNA*) - cząsteczki RNA (dł. ok. 24 nt), których szlak biogenezy nie jest dobrze poznany, lecz wiadomo, że nie powstają w wyniku aktywności enzymu Dicer; mogą odgrywać rolę w utrzymywaniu integralności końców chromosomów (Cao, Li i wsp. 2009);
- **tiRNA (**RNA pochodzący z tRNA, ang. *stress-induced tRNA-derived RNAs*) - cząsteczki RNA funkcjonujące jako represory translacji, które powstają w wyniku cięcia tRNA przez angiogeninę (rybonukleazę 5) (Thompson i Parker 2009);
- **tmRNA** (transferowo-matrycowy RNA ang. *transfer-messenger RNA*), bakteryjny RNA posiadający dwie funkcjonalności: tRNA i mRNA (Ray i Apirion 1979);
- **xiRNA** (RNA odpowiedzialny za inaktywację chromosomu X, ang. *RNA responsible for X-inactivation*) – cząsteczki RNA powstające w wyniku działania enzymu Dicer, tworzone z dupleksów długich niekodujących RNA Xist i Tsix, które odpowiedzialne są za lionizację (inaktywację chromosomu X) u ssaków łożyskowych (Ogawa, Sun i wsp. 2008).

W przypadku człowieka liczba znanych sekwencji niekodujących RNA wynosi 91152, zgodnie z danymi zawartymi w bazie danych NONCODE (wersja 3.0) (Bu, Yu i wsp. 2012).

1.2 Metabolizm kwasów nukleinowych

1.2.1 Szlaki naprawy DNA

Szlaki naprawy DNA są sekwencjami reakcji enzymatycznych, mającymi kluczowe znaczenie dla utrzymania i zachowania informacji genetycznej wszystkich organizmów. Stabilność genomu jest stale zagrożona ze względu na negatywny wpływ czynników środowiskowych lub endogennych procesów metabolicznych czy błędy pojawiające się w trakcie procesów komórkowych. Modyfikacje DNA, zwłaszcza te, które zmieniają lub uniemożliwiają tworzenie się kanonicznych par typu Watson-Crick, mogą prowadzić do mutacji, zmieniających na stałe sekwencję kodującą kwasu nukleinowego, czego wynikiem mogą być różne choroby genetyczne. Uszkodzenia DNA często towarzyszą procesom komórkowymi, takim jak replikacja DNA czy transkrypcja (Maddukuri, Dudzinska i wsp. 2007, Jeggo i Lavin 2009). By zapobiec błędom w DNA organizmy w trakcie ewolucji wytworzyły różnego typu systemy zapobiegania i naprawy (Vaisman, Lehmann i wsp. 2004, Arczewska i Kusmierek 2007, Krwawicz, Arczewska i wsp. 2007, Brissett i Doherty 2009, Robertson, Klungland i wsp. 2009). Systemy te zapewniają stabilność DNA i prawidłowy przepływ informacji genetycznej poprzez ochronę genomu przed dużą liczbą różnych zmian chemicznych i strukturalnych. Warto podkreślić, że z drugiej strony uszkodzenia i błędy w replikacji DNA stanowią główne źródło zmienności genetycznej i tym samym siłę napędową ewolucji. W organizmach wielokomórkowych zmiany w sekwencji i strukturze DNA są odpowiedzialne np. za produkcję przeciwciał przez system immunologiczny (Slatter i Gennery 2010). Dlatego też mechanizmy naprawcze DNA muszą działać w sposób pozwalający zbalansować szkodliwe i korzystne wpływy zmian sekwencji i struktury genomu.

W DNA komórki ssaczej w ciągu dnia pojawia się ponad 20000 uszkodzeń spowodowanych przez czynniki zewnętrzne, z czego większość to deaminacje, spontaniczne hydrolizy wiązań N-glikozydowych, alkilacje i uszkodzenia wywołane przez reaktywne formy tlenu (Lindahl 1993, De Bont i van Larebeke 2004, Drablos, Feyzi i wsp. 2004, Olinski, Siomek i wsp. 2007, Tudek 2007). Uszkodzenia wywoływane są także błędami pojawiającymi się podczas procesów metabolicznych DNA, takimi jak: tworzenie się jedno- i dwuniciowych przerw, przerwanie widełek replikacyjnych czy wprowadzenie zmodyfikowanych nukleotydów podczas replikacji. Sumarycznie u dorosłego, zdrowego człowieka (10^{12} komórek) pojawia się dziennie 10^{16}-10^{18} zdarzeń naprawczych

(Friedberg, Walker i wsp. 2006). Niestety niektóre uszkodzenia „uciekają" mechanizmom naprawy i w konsekwencji prowadzą do mutacji, starzenia się i różnych chorób (w tym nowotworów i neurodegeneracji) (Hansen i Kelley 2000, Wilson i Barsky 2001, Wood, Mitchell i wsp. 2001, Wood, Mitchell i wsp. 2005, Raptis i Bapat 2006).

Naprawa DNA jest bardzo skomplikowanym procesem wymagającym wielu czynników. Przykładowo do tej pory zidentyfikowano w ludzkim genomie ponad 178 genów kodujących białka związane z naprawą DNA (Ronen i Glickman 2001, Wood, Mitchell i wsp. 2001, Wood, Mitchell i wsp. 2005, Wood, Mitchell i wsp. 2013). Są one zaangażowane w różne procesy, zaczynając od wykrywania miejsc uszkodzonych w DNA, przez kilka etapów transformacji enzymatycznej błędnego DNA, po rekombinację, zatrzymanie cyklu komórkowego, bądź zainicjowanie programowanej śmierci komórki (apoptozy). Inną formą reagowania na uszkodzenia DNA jest pominięcie ich, co ułatwia kontynuację replikacji nawet wtedy, gdy pojawiają się nieusuwalne modyfikacje, ale nie gwarantuje prawidłowego odtworzenia oryginalnej sekwencji i często prowadzi do wprowadzania mutacji przez polimerazy TLS (polimerazy syntezy DNA ponad uszkodzonym miejscem na matrycy, ang. *translesion synthesis polymerases*). Liczne chemiczne i strukturalne transformacje prowadzące od błędnego DNA do DNA poprawnego mogą być opisane jako ścieżki składające się z serii kroków. Obecnie znane szlaki naprawy DNA mogą być podzielone na osiem kategorii:

- DDS (ang. *DNA damage signaling*): szlak indukowany w odpowiedzi na uszkodzenia DNA spowodowane przez czynniki endogenne i środowiskowe;
- DDR – bezpośrednia naprawa DNA (ang. *direct reversal repair*): odtwarza natywne reszty nukleotydowe poprzez usunięcie nienatywnych modyfikacji chemicznych, takich jak np. O6-metyloguanozyna, 1-metylo-2′-deoksyadenozyno-5′-monofosforan, dimer cyklobutanu pirymidyny, 6-4 fotoprodukt (Lindahl i Wood 1999);
- BER – naprawa przez wycinanie zasady (ang. *base-excision repair*): rozpoczynana przez wycięcie zmodyfikowanych zasad z DNA. W zależności od długości resyntezy DNA podzielona jest na dwa podszlaki: krótki (SP-BER) i długi (LP-BER);
- NER – naprawa przez wycinanie nukleotydu (ang. *nucleotide-excision repair*): usuwa duże uszkodzenia z DNA. Istnieją dwa rodzaje NER:
    - (TCR)–NER (ang. *transcription-coupled repair*), naprawa sprzężona z transkrypcją, usuwa uszkodzenie z nici aktywnej transkrybowanego genu;

  - (GGR)–NER (ang. *global genome repair*), globalna naprawa genomu, usuwa uszkodzenia obecne w dowolnym miejscu w genomie;
- MMR – naprawa błędnie sparowanych zasad (ang. *mismatch repair*): poreplikacyjna naprawa DNA, która usuwa błędy wprowadzone podczas replikacji (źle wstawione nukleotydy, małe pętle, insercje, delecje);
- HRR - naprawa rekombinacyjna (znana również jako rekombinacja homologiczna) (ang. *homologous recombination repair*): naprawa dwuniciowych przerw w DNA z użyciem homologicznej nici DNA jako szablonu do resytnezy;
- NHEJ - łączenie niehomologicznych końców (ang. *non-homologous end joining repair*): ligacja końców dwuniciowych przerw w DNA;
- TLS – awaryjny mechanizm syntezy DNA, synteza DNA ponad uszkodzonym miejscem na matrycy (ang. *translesion synthesis*): szlak tolerancji uszkodzeń, wymaga wyspecjalizowanych polimeraz, umożliwiających przeprowadzenie replikacji pomimo uszkodzenia.

Każdy z tych szlaków może być zaprezentowany jako seria transformacji enzymatycznych pomiędzy różnymi strukturami DNA, katalizowanych przez określony zbiór białek. Należy podkreślić, że szlaki naprawy DNA są ze sobą połączone, co oznacza, że te same białka mogą brać udział w różnych rodzajach naprawy (Friedberg, Walker i wsp. 2006). W konsekwencji białka naprawcze nie działają osobno w komórce, a ich aktywność jest zależna od innych elementów systemów naprawy DNA. Dlatego też poznanie zarazem całych systemów naprawy DNA jak i wszystkich elementów biorących w nich udział jest kluczowe dla zrozumienia, w jaki sposób komórki kontrolują i naprawiają pojawiające się przez cały czas uszkodzenia ich genomów. Wiele zaangażowanych w naprawę DNA białek z różnych organizmów jest bardzo dobrze opisanych, włączając w to specyficzność substratową, kinetykę, mechanizm reakcji i strukturę przestrzenną w kompleksie z substratem i/lub produktem ich aktywności. Jednakże informacja o różnych czynnikach biorących udział w naprawie DNA w innych organizmach jest ograniczona bądź rozproszona w najróżniejszych bazach danych. Ponadto istnieją procesy, dla których podejrzewana lub znana jest aktywność enzymatyczna, ale nie zostały jeszcze określone geny czy białka z nią związane. Przykładem może być enzym zdolny do usuwania 5-hydroksymetylourydyny, produktu oksydacji tyminy, który musi istnieć, ale do tej pory nie został zidentyfikowany (Baker, Liu i wsp. 2002).

1.2.2 Szlaki dojrzewania i degradacji RNA

RNA odgrywa w komórkach wiele kluczowych ról, takich jak: przekazywanie informacji genetycznej pomiędzy DNA a białkami, regulacja wielu procesów czy kataliza. W każdym z tych przypadków, funkcja RNA zależy od sekwencji nukleotydowej. Sekwencja dojrzałej, funkcjonalnej cząsteczki kwasu rybonukleinowego jest często inna od tej, którą posiada pierwotny transkrypt, co związane jest z kombinacją różnorodnych wydarzeń przetwarzających ten transkrypt, wśród których znajdują się: wycinanie funkcjonalnych jednostek z cząsteczki prekursora, usuwanie intronów (ang. *splicing*), łączenie różnych elementów (ang. *trans-splicing*), dodanie nieobecnych w szablonie reszt nukleotydowych (poprzez dodanie czapeczki, edycję czy poliadenylację), bądź zmianę własności chemicznych pojedynczych reszt (poprzez edycję czy modyfikację) (artykuły przeglądowe: (Grosjean 2005, Arraiano, Andrade i wsp. 2010, Licatalosi i Darnell 2010, Lutz i Moreira 2011, Motorin i Helm 2011)). Funkcja RNA zależy także od jego lokalizacji, tworzenia struktur wyższego rzędu czy formowania kompleksów z innymi cząsteczkami, w szczególności białkami czy innymi cząsteczkami RNA (Fatica i Tollervey 2002, Iglesias i Stutz 2008, Holt i Bullock 2009, Hocine, Singer i wsp. 2010).

Funkcjonalne cząsteczki RNA, które nie są już potrzebne, gdyż spełniły swoją funkcję, bądź posiadają błędy związane z uszkodzeniami lub nieprawidłowym przetwarzaniem, zwijaniem się czy składaniem w funkcjonalne kompleksy, są eliminowane z komórek poprzez różnorodne szlaki degradacji (Coller i Parker 2004, Kushner 2004, Houseley i Tollervey 2009). Rozkład RNA usuwa także produkty uboczne ekspresji genów, w tym wycięte introny czy inne fragmenty RNA uwalniane podczas jego przeróbki. Wreszcie, szlaki degradacji RNA eliminują międzygenowe, wewnątrzgenowe, związane z promotorem i antysensowne RNA, które pojawiają się albo jako cząsteczki regulatorowe, albo jako szum transkrypcyjny. Efektywność i specyficzność degradacji RNA zapewniona jest głównie poprzez szerokie spektrum różnych endo- i egzorybonukleaz (Arraiano, Andrade i wsp. 2010, Tomecki i Dziembowski 2010), jednakże zależna jest od formowania się specyficznych struktur RNA czy kompleksów RNA-białko, zawierających czynniki regulatorowe (Alonso 2012) i może być regulowana zarówno na poziomie ekspresji genów jak i na poziomie białkowym.

Biogeneza funkcjonalnych cząsteczek RNA, tak samo jak ich rozkład, zarówno w komórkach prokariotycznych jak i eukariotycznych, wymaga serii chemicznych i strukturalnych zmian, podczas których RNA oddziałuje z licznymi czynnikami

komórkowymi, przede wszystkim z enzymami (Luna, Gaillard i wsp. 2008). Należy podkreślić, że szlaki przetwarzania RNA są ze sobą powiązane i przeplatają się nawzajem z innymi szlakami komórkowymi, poprzez np. występowanie w różnych szlakach tych samych białek, czy etapów (Beggs i Tollervey 2005, Arraiano, Andrade i wsp. 2010). Jedna cząsteczka RNA może także być przedmiotem kilku procesów enzymatycznych w tym samym czasie. Dlatego też, dokładnie tak samo jak w przypadku naprawy DNA, poznanie zarówno całej sieci zdarzeń przetwarzania RNA jak i białek odpowiedzialnych za poszczególne przekształcenia jest kluczowe dla zrozumienia metabolizmu RNA. Również, jak w przypadku naprawy DNA, wiele białek zaangażowanych w te procesy jest dobrze opisanych, ale informacja o nich jest rozproszona pomiędzy różnymi źródłami. Podobnie również istnieje wiele procesów, dla których znana jest aktywność, ale nieopisane zostały jeszcze geny/białka/enzymy za nią odpowiedzialne. Dlatego tak ważne są prace, których celem jest rekonstrukcja szlaków i sieci metabolicznych RNA, która mogłaby rzucić nowe światło na cały system i umożliwić odkrycie nowych elementów w niego zaangażowanych. Ponadto porównywanie szlaków pomiędzy różnymi gatunkami może wskazać białka homologiczne zaangażowane w podobne aktywności.

1.2.2.1 Dojrzewanie i degradacja mRNA

Proces biogenezy mRNA rozpoczyna się transkrypcją, a kończy degradacją. W czasie życia cząsteczki mRNA mogą ulegać modyfikacjom, edycjom czy transportowi. mRNA eukariotyczny, w odróżnieniu od prokariotycznego, często wymaga złożonego procesowania i transportu.

Proces transkrypcji rozpoczyna cykl życia cząsteczki mRNA i jest on podobny u Prokaryota i Eukaryota. Produktem transkrypcji najczęściej jest prekursorowy mRNA (pre-mRNA), który musi przejść proces dojrzewania, by stać się w pełni funkcjonalną cząsteczką. Proces dojrzewania nie jest identyczny u bakterii, archeowców i eukariontów. W komórkach prokariotycznych, poza nielicznymi przypadkami, mRNA najczęściej nie wymaga już obróbki. U Eukaryota natomiast proces ten obejmuje: dojrzewanie końca 5′, dojrzewanie końca 3′, wycinanie intronów oraz edycję, czyli zmianę składu nukleotydowego (przykładem mRNA ulegającego edycji może być mRNA ludzkiej apolipoproteiny B, który jest edytowany w niektórych tkankach, dzięki czemu wprowadzony zostaje wcześniejszy kodon STOP, umożliwiając otrzymanie krótszej wersji białka).

Dojrzała cząsteczka mRNA, która odegrała już swoją rolę ulega degradacji. Różne cząsteczki mRNA mają różny czas życia (stabilność) – u bakterii waha się on pomiędzy

sekundami a godziną, w komórkach ssaczych natomiast od kilku minut po dni. Ponadto istnieją mechanizmy usuwające błędnie przygotowane cząsteczki mRNA, chroniąc komórkę przed niewłaściwie utworzonymi białkami. Mechanizmy degradacji są różne u Prokaryota i Eukaryota. W komórkach prokariotycznych usuwanie cząsteczek mRNA zachodzi przy użyciu różnego typu rybonukleaz: endonukleaz, 3′ egzonukleaz i 5′ egzonukleaz. W niektórych przypadkach małe RNA mogą stymulować degradację mRNA przez parowanie się z sekwencjami komplementarnymi i aktywowanie cięcia przez RNazę III. Ponadto według ostatnich odkryć bakteryjny mRNA posiada na końcu 5′ trójfosforan, który po usunięciu dwóch fosforów staje się 5′ monofosforanem, będącym dla RNazy J sygnałem do degradacji mRNA w kierunku 5′ – 3′ (Deana, Celesnik i wsp. 2008).

W komórkach zarówno eukariotycznych, jak i prokariotycznych, procesy translacji i degradacji mRNA są zbalansowane. Najważniejszymi białkami i kompleksami enzymatycznymi biorącymi udział w degradacji u Eukaryota są białka usuwające czapeczkę z końca 5′ (np. białko DCP2), kompleksy usuwające ogon poliA, egzosom oraz egzonukleaza Xrn1. Szlakami występującymi w komórkach eukariotycznych i do tej pory opisanymi są m.in.:

- NMD (ang. *Nonsense-mediated decay*) – usuwanie transkryptów posiadających przedwczesny kodon STOP (Maquat 2005, Doma i Parker 2007, Isken i Maquat 2007);
- NGD (ang. *No-go decay*) – usuwanie transkryptów, w których z powodu występujących struktur drugorzędowych rybosom ulega zatrzymaniu (Doma i Parker 2007, Isken i Maquat 2007);
- NSD (ang. *Nonstop decay*) – mechanizm kontroli jakości mRNA, który usuwa mRNA z brakującymi kodonami terminacji (Doma i Parker 2007, Isken i Maquat 2007);
- degradacja zależna od deadenylacji (ang. *deadenylation-dependent decay*) – główny szlak usuwania dojrzałego mRNA, gdzie główną rolę odgrywają kompleksy usuwające ogon poliA (np. kompleks CCR4-NOT, czy PAN2-PAN3) (Garneau, Wilusz i wsp. 2007);
- degradacja niezależna od deadenylacji (ang. *deadenylation-independent decay*) (Garneau, Wilusz i wsp. 2007);
- degradacja zależna od enzonukleaz (ang. *endonuclease-mediated decay*) (Garneau, Wilusz i wsp. 2007);

- degradacja elementów bogatych w AU (ang. *AU-rich element decay*) (Chen i Shyu 1995).

1.2.2.2 Dojrzewanie i degradacja rRNA

Biosynteza rRNA jest podobna do biosyntezy mRNA. Różnice występują w dojrzewaniu prekursora rRNA (pre-rRNA). Bakterie syntetyzują trzy rodzaje rRNA: 5S rRNA, 16S rRNA i 23S rRNA, przy czym nazwy wskazują na wielkość cząstek mierzoną poprzez analizę sedymentacji. Ich geny połączone są w jedną jednostkę transkrypcyjną (operony rrn A-E, H, G), która zwykle występuje w wielu kopiach i dodatkowo zawiera geny tRNA w kopiach od jednej do czterech. Aby uwolnić dojrzałe rRNA, prekursorowa cząsteczka RNA musi wpierw ulec modyfikacjom, a później zostać pocięta. Cięcie przeprowadzają różne rybonukleazy w miejscach wyznaczonych przez regiony dwuniciowe powstałe przez łączenie komplementarnych zasad z różnych części pre-rRNA. Końce powstałe po cięciu są przycinane przez egzonukleazy (Deutscher 2009).

U Eukaryota istnieją cztery rodzaje rRNA. Jeden z nich, 5S rRNA nie podlega dojrzewaniu, a pozostałe: 5,8S; 18S i 28S (nomenklatura zastosowana w rozprawie doktorskiej, opisująca stałe sedymentacji rRNA, została opracowana w oparciu o cząsteczki rRNA występujące u człowieka) powstają z jednej jednostki transkrypcyjnej jako pre-rRNA. W pojedynczym locus występuje ponad 100 tandemowo połączonych genów. Prekursor wpierw ulega modyfikacjom – są to metylacje rybozy z udziałem C/D snoRNP (małe jąderkowe rybonukleoproteiny, ang. *small nucleolar ribonucleoprotein*) oraz modyfikacje urydyn do pseudourydyn przez H/ACA snoRNP. Następnie następuje cięcie i przycinanie końców (Lafontaine i Tollervey 2001).

Degradacja rRNA zachodzi poprzez różne szlaki. Jest ona głównie związana z usuwaniem niepoprawnie złożonych rybosomów. Procesy kontroli jakości występują zarówno u Prokaryota jak i Eukaryota. U *E. coli* nieprawidłowe prekursory RNA cięte są wpierw przez endorybonukleazy, a następnie powstałe fragmenty rRNA degradowane są przez RNazę R i/lub PNPazę (Deutscher 2009). Ponadto u *E. coli* występuje mechanizm degradacji rRNA w odpowiedzi na stres, który nie jest jeszcze do końca poznany (Kaplan i Apirion 1974, Cohen i Kaplan 1977). U drożdży występuje mechanizm kontroli jakości rybosomów zwany degradacją niefunkcjonalnego rRNA (NRD, ang. *nonfunctional rRNA decay*), który dzieli się na dwa szlaki:

- 18S NRD – usuwa rRNA, które posiadają szkodliwe mutacje w miejscu akceptorowym w rybosomie;
- 25S NRD – eliminuje rRNA, które posiadają szkodliwe mutacje w centrum peptydylotransferazy dużej podjednostki rybosomu (LaRiviere, Cole i wsp. 2006).

1.2.2.3 Dojrzewanie i degradacja tRNA

Sekwencje tRNA w komórkach prokariotycznych kodowane są najczęściej w zespołach składających się z: genów różnych bądź tych samych tRNA lub genów tRNA razem z genami rRNA lub genów tRNA i sekwencji kodujących białka. Prekursory tRNA (pre-tRNA) wymagają dojrzewania zarówno końca 5′ jak i 3′. U *E. coli* dojrzewanie przebiega w uporządkowanej serii etapów cięcia przez rybonukleazy (RNazy: P, D, BN, T, PH, II oraz PNPazę). Sekwencja CCA z końca 3′ dojrzałego tRNA kodowana jest bezpośrednio przez gen, a nie dodawana potranskrypcyjnie (Condon 2007). U Eukaryota cząsteczki tRNA kodowane są przez oddzielne geny, czasem zgrupowane w jednym miejscu. Transkrypty mogą posiadać introny. Dojrzewanie zachodzi także w uporządkowanej serii etapów składającej się z przycięcia końca 5′, przycięcia końca 3′, dodania sekwencji CCA na końcu 3′ przez transferazę nukleotydową, usunięcia intronów (mechanizm cięcia i ligacji) i modyfikacji reszt nukleotydów (Phizicky i Hopper 2010, Popow, Schleiffer i wsp. 2012).

Mechanizmy degradacji tRNA najlepiej opisane są dla drożdży i związane są z kontrolą jakości tRNA. Degradowane są tRNA, które nie mają w swojej strukturze odpowiednich modyfikacji. Przykładem może być tRNAVal(AAC), który przy braku m7G46 i m5C49 degradowany jest w procesie zwanym nagłą degradacją tRNA (RTD, ang *rapid tRNA decay*) (Chernyakov, Whipple i wsp. 2008).

1.3 Bazy danych – obecny stan wiedzy

1.3.1 Ogólne bazy danych

W czasach komputeryzacji wszelkiego rodzaju informacje biologiczne, poza literaturą, dostępne są w różnorakich źródłach cyfrowych, jakimi są np. bazy danych. Biologiczne i biomedyczne bazy danych mogą służyć jako „kopalnie” informacji z dziedziny nauk przyrodniczych, pochodzących z doświadczeń naukowych i projektów wielkoskalowych, artykułów naukowych i analiz komputerowych. Zawierają m. in. dane dotyczące sekwencji,

funkcji i struktury genów oraz białek, efektów klinicznych mutacji, szlaków metabolicznych jak również podobieństw między sekwencjami i strukturami.

1.3.1.1 KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/) (Kanehisa, Araki i wsp. 2008) to zbiór połączonych ze sobą baz danych, do których należą m.in: KEGG PATHWAY (szlaki metaboliczne), KEGG DISEASE (choroby ludzkie), KEGG GENES (geny i białka), czy KEGG ORGANISMS (organizmy). Jeśli chodzi o naprawę DNA i szlaki metaboliczne RNA to głównymi bazami, mogącymi służyć jako źródło informacji, są KEGG GENES (katalog genów zsekwencjonowanych genomów uzyskany z publicznie dostępnych źródeł, głównie z NCBI RefSeq) oraz KEGG PATHWAYS (zbiór ręcznie rysowanych map szlaków przedstawiających molekularne oddziaływania i sieci reakcji dla: metabolizmu, przetwarzania informacji genetycznej, przetwarzania informacji środowiskowych, procesów komórkowych, chorób ludzkich i reakcji tych systemów na leki.

Szlaki naprawy DNA, które znajdują się w KEGG to: BER, NER, NHEJ, MMR, i HRR. Nieuwzględnione są DDS, DDR i TLS. Ponadto ciekawą sekcją dotyczącą naprawy DNA jest *DNA repair and recombination proteins* (białka rekombinacji i naprawy DNA, identyfikator *ko034400*), w której wszystkie białka związane z naprawą DNA sklasyfikowane są względem ich funkcji. Jeśli chodzi o szlaki dojrzewania i degradacji RNA, to zdecydowanie trudniej je odnaleźć. Wiele informacji znajduje się w sekcjach: *Transcription factors* (czynniki transkrypcyjne), *Transcription machinery* (maszyneria transkrypcyjna), *Spliceosome*, *Ribosome* (rybosom), *Ribosome biogenesis* (biogeneza rybosomu), czy *Transfer RNA biogenesis* (biogeneza transferowego RNA). Graficzna reprezentacja kompleksów biorących udział w wyżej wymienionych procesach jest dostępna osobno dla Eukaryota i Prokaryota. KEGG BRITE to także baza danych, która zawiera informacje o interesujących nas szlakach. Jest to zbiór hierarchii funkcjonalnych i binarnych relacji pomiędzy obiektami biologicznymi (ang. *functional hierarchies and binary relationships of biological entities*), który poza molekularnymi oddziaływaniami czy reakcjami posiada także dane dotyczące różnego typu relacji pomiędzy tymi obiektami.

1.3.1.2 Reactome

Reactome (http://www.reactome.org) (Matthews, Gopinath i wsp. 2009) to darmowe źródło danych, zbierające dane o różnego typu szlakach głównie u człowieka, ale dostępne są też dane dla innych organizmów (np. *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Mus musculus* czy *Escherichia coli*), opracowane przez współpracujące ze sobą grupy specjalistów. Poza zbieraniem danych strona oferuje także narzędzia do analizy takie jak: *the Pathway Browser* (przeglądarka szlaków), *the Pathway and Expression Analysis tools* (narzędzia do analizy szlaków i ekspresji) czy *the Species Comparison tool* (narzędzie do porównywania organizmów). W odróżnieniu od KEGG, Reactome zawiera graficzne reprezentacje szlaków naprawy DNA oraz szlaków dojrzewania i degradacji RNA, które generowane są dla każdego organizmu osobno. Poza tym w bazie tej występuje dodatkowy podział szlaków na pod-szlaki (np. w GG-NER wyodrębniono cztery pod-szlaki) oraz łatwiej w niej znaleźć informacje dotyczące szlaków degradacji, dojrzewania i edycji RNA, które są wyodrębnione w przeglądarce szlaków. Ponadto narzędzie do analizy szlaków pozwala na przeprowadzenie analiz porównawczych, np. znalezienie połączenia pomiędzy transkrypcją a naprawą DNA (Stein 2004).

1.3.1.3 BioCyc

BioCyc (http://biocyc.org/) (Caspi, Altman i wsp. 2010) to zbiór 2920 baz danych szlaków/genomowych (na dzień 17.05.2013). Każda baza w BioCyc opisuje genom bądź szlaki metaboliczne pojedynczego organizmu. Jest to, jednakże, nie tylko zbiór baz danych, ale zasób udostępniający różnorakie narzędzia do analiz bioinformatycznych, takie jak: przeglądarka genomowa, narzędzie do wyświetlania pojedynczych szlaków metabolicznych lub pełnej mapy metabolicznej, narzędzie do wizualnej analizy danych „omowych" (ang. *„omics"*) użytkownika, które rysowane są na istniejących metabolicznych, regulatorowych czy genomowych mapach i narzędzie do analiz porównawczych. Istnieje także możliwość pobrania kopii BioCyc, która zawiera *Pathway Tools* (narzędzia do analiz szlaków) na dysk lokalny. Bazy danych BioCyc podzielone są na trzy poziomy (ang. *tier*), w zależności od jakości danych. Poziom pierwszy to bazy, które zawierają dane ręcznie przygotowane na podstawie przeszukiwania literatury i są najbardziej wiarygodne. Do tej grupy zaliczają się takie bazy jak: EcoCyc (http://ecocyc.org) (Keseler, Bonavides-Martinez i wsp. 2009) – baza danych biologii *Escherichia coli* K-12 MG1655 czy MetaCyc (http://metacyc.org/) – baza danych niepowtarzających się, doświadczalnie zbadanych

szlaków metabolicznych (Caspi, Altman i wsp. 2012). Dane znajdujące się w tych bazach podlegają procedurze sprawdzenia przez zewnętrznych ekspertów, którzy pracują na tych konkretnych systemach komórkowych. Ostatnio tego typu podejście zostało również zastosowane np. przy dodawaniu danych dotyczących procesów naprawy DNA, zarówno przy bezpośrednich mechanizmach naprawczych (np. fotoliza), jak i pośrednich szlakach naprawy (np. NER, BER, HHR) (Keseler, Bonavides-Martinez i wsp. 2009). Poziom drugi i trzeci to bazy danych komputerowo przewidzianych: szlaków metabolicznych, genów, brakujących w tych szlakach enzymów i operonów. BioCyc nie ma bezpośrednich sekcji dedykowanych naprawie DNA czy dojrzewaniu i degradacji RNA, jednakże wiedzę na ten temat znaleźć można wśród innych sekcji. Co więcej BioCyc pozwala na dogłębną analizę systemów biologicznych, co zostało przedstawione np. w pracy na temat zróżnicowanej odpowiedzi systemu na leki i stres Cabusory (Cabusora, Sutton i wsp. 2005), w której analizowano dane z BioCyc, dotyczące ekspresji znanych czynników odpowiedzi na stres i genów naprawy DNA w *Mycobacterium tuberculosis*.

1.3.1.4 BRENDA

BRENDA (BRaunschweig ENzyme Database, http://www.brenda-enzymes.org) (Scheer, Grote i wsp. 2011) to baza danych enzymów, która zawiera ręcznie przygotowany zestaw informacji odnośnie właściwości enzymów, włączając w to mutanty i warianty otrzymane na drodze metod inżynierii genetycznej (ang. *engineered variants*). Opisuje enzymy zaangażowane w procesy naprawcze oraz procesy dojrzewania i degradacji, które posiadają numer E.C. (ang. *Enzyme Commission number*, (1999)), np. UvrA o E.C. 3.1.25.1. Wśród dostępnych danych znajdują się: odnośniki do oryginalnych publikacji, klasyfikacja, nazewnictwo, typ reakcji, specyficzność substratowa, parametry biochemiczne, organizm, sekwencja i struktura przestrzenna, praktyczne zastosowanie, informacja o mutantach czy wariantach otrzymanych na drodze metod inżynierii genetycznej, stabilności, chorobach, izolacji i przygotowaniu. Duża cześć bazy poświęcona jest danym dotyczącym metabolitów czy małych cząsteczek, które oddziałują z enzymami jako substraty, produkty, inhibitory, elementy aktywujące, kofaktory czy wiążące się cząsteczki jonów metali.

1.3.1.5 Pathway Commons

Pathway Commons (http://www.pathwaycommons.org) to publiczna, dostępna darmowo kolekcja zawierająca informacje odnośnie szlaków z wielu organizmów (Cerami, Gross i wsp. 2011), która uwzględnia reakcje biochemiczne, składanie kompleksów, transport, zdarzenia katalityczne i fizyczne oddziaływania pomiędzy białkami, DNA, RNA, małymi cząsteczkami czy kompleksami. Jest to meta-baza danych, która zbiera informacje z innych baz takich jak Reactome czy BioGrid, umożliwiając tym samym analizę zestawów danych na poziomie systemowym pomiędzy różnymi organizmami. Pozwala użytkownikom na przeglądanie i przeszukiwanie szlaków w różnych publicznie dostępnych bazach danych, pobieranie na dysk lokalny zintegrowanych zestawów szlaków w formacie BioPAX (jest to język bazujący na XML, służący do opisywania szlaków biologicznych, ang. *Biological Pathways eXchange*, (Demir, Cary i wsp. 2010)). Posiada także narzędzia dla twórców oprogramowania, które umożliwiają przeprowadzanie bardziej zaawansowanych analiz.

1.3.2 Bazy danych dedykowane naprawie DNA

Jak wspomniano wcześniej, wiedza dotycząca systemów naprawy DNA jest niezwykle istotna dla pełnego zrozumienia jak komórki kontrolują integralność swojego genomu. Z roku na rok następuje systematyczny wzrost ilości informacji, która jest rozproszona w literaturze i w wielu elektronicznych źródłach danych. Systematyzacja tej wiedzy i jej zaprezentowanie w sposób jasny i przejrzysty to obecnie główne zadania internetowych baz danych. Zebranie, przetworzenie i udostępnienie danych jest niezwykle podczas poszukiwania odpowiedzi na różnego typu pytania biologiczne dotyczące podsystemów naprawy DNA, np. „które białka uczestniczą w szlaku MMR zarówno u ludzi, jak i roślin?”, „jaka natychmiastowa odpowiedź komórkowa jest wywoływana w reakcji na uszkodzenia spowodowane światłem UV?”, bądź: „w jaki sposób HHR różni się pomiędzy kręgowcami a roślinami?”. Temat naprawy DNA jest poruszany przez wiele elektronicznych źródeł informacji, jednakże, samych dedykowanych mu baz danych jest niewiele, a większość informacji jest rozsiana w ogólnych biologicznych bazach danych. Poniżej przedstawiono wybrane bazy, natomiast Tabela 1 zawiera listę źródeł, które mogą służyć pomocą podczas analiz szlaków naprawy DNA.

1.3.2.1 DNAtraffic

Baza danych **DNAtraffic** (http://dnatraffic.ibb.waw.pl/) (Kuchta, Barszcz i wsp. 2012) jest bazą dedykowaną dynamice genomu w trakcie życia komórki. Baza ta zawiera informacje odnośnie nazewnictwa, ontologii, struktury i funkcji białek związanych z mechanizmami zachowania spójności DNA, takimi jak: remodelowanie chromatyny, naprawa DNA czy odpowiedź na uszkodzenia z ośmiu modelowych organizmów, najczęściej wykorzystywanych w badaniach nad DNA (*Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Escherichia coli K-12*, *Arabidopsis thaliana*). Ponadto w bazie tej znaleźć można informacje odnośnie chorób genetycznych powiązanych z białkami ludzkimi.

Należy podkreślić, że baza ta powstała **po** opublikowaniu bazy danych szlaków naprawy DNA **REPAIRtoire**, stanowiącej jedno z opracowań w ramach niniejszej rozprawy doktorskiej.

1.3.2.2 Human DNA Repair Genes

Human DNA Repair Genes to załącznik dostępny elektronicznie do publikacji przeglądowej Wooda, Mitchella i Lindahla z roku 2005 (Wood, Mitchell i wsp. 2005), który aktualizowany jest regularnie (ostatnia aktualizacja: 4.03.2013) i dostępny pod adresem: http://sciencepark.mdanderson.org/labs/wood/DNA_Repair_Genes.html. Przedstawia on informacje w formie tabelarycznej, w której geny połączone są z zewnętrznymi bazami danych: Gene Cards Brytyjskiego Centrum Badań nad Rakiem (http://www.genecards.org/) (Safran, Dalah i wsp. 2010), OMIM (Online Mendelian Inheritance in Man (Sayers, Barrett i wsp. 2009)), oraz z narzędziem MapView dostępnym w NCBI i z serwerami Entrez NCBI.

1.3.2.3 Repair-FunMap

Baza danych **Repair-FunMap** (Wen i Feng 2004) to niedziałające obecnie źródło danych, które udostępniało informacje o sieci oddziaływań pomiędzy białkami zaangażowanymi w naprawę DNA a białkami biorącymi udział w innych procesach komórkowych.

1.3.2.4 Inne bazy danych

Istnieje szereg baz dedykowanych innym aspektom metabolizmu DNA. Przykładem tu mogą służyć: replikacja DNA (OriDB (Nieduszynski, Hiraga i wsp. 2007), ReplicationDomain (Weddington, Stuy i wsp. 2008)), apoptoza (Deathbase (Diez, Walter i wsp. 2010)), konserwacja telomerów (Telomerase database (Podlevsky, Bley i wsp. 2008)), restrykcja i modyfikacja DNA (REBASE (Roberts, Vincze i wsp. 2010)), czy epigenetyka/modyfikacje chromatyny (DAnCER (Turinsky, Turner i wsp. 2011)). Procesy te związane są z naprawą DNA, gdyż mogą przyczyniać się do pojawiania się uszkodzeń w DNA (replikacja) czy do regulacji innych procesów enzymatycznych (metylacja DNA, kontrola cyklu komórkowego, apoptoza).

1.3.3 Bazy danych dedykowane dojrzewaniu i degradacji RNA

Tak samo jak w przypadku naprawy DNA istnieją bazy danych zbierające informacje o cząsteczkach RNA. Jednakże brakuje baz dedykowanych konkretnym aspektom metabolizmu RNA. Poniżej przedstawiono najważniejsze bazy danych, udostępniające różnego typu informacje o RNA, natomiast Tabela 1 zawiera listę źródeł, które mogą służyć pomocą podczas analiz szlaków dojrzewania i degradacji RNA.

1.3.3.1 RFAM

Baza danych **Rfam** (http://rfam.sanger.ac.uk/) (Burge, Daub i wsp. 2013) jest źródłem informacji dotyczących rodzin RNA. Każda rodzina składa się ze zbioru sekwencji RNA, które są uważane za pochodzące od wspólnego przodka. Do każdej z tych rodzin przypisane są następujące elementy: przyrównanie wielu sekwencji, konsensusowa struktura drugorzędowa oraz model kowariancji (CM, ang. *covariance model*). Wszystkie te elementy są sprawdzane ręcznie przez kuratorów bazy i dodatkowo połączone z zewnętrznymi bazami danych, takimi jak PDB, bazy ontologii, Europejskie Archiwum Nukleotydowe (ang. *European Nucleotide Archive*, http://www.ebi.ac.uk/ena/, (Leinonen, Akhtar i wsp. 2011)) czy mirBase (patrz: Podrozdział 1.3.3.3).

Pierwsza wersja bazy (v1.0) została opublikowana w lipcu 2002 roku i zawierała 25 rodzin. Obecnie dostępna jest wersja v11.0, która udostępnia wiedzę o 2208 rodzinach RNA i pokrywa 6 milionów regionów sekwencyjnych (pierwsza wersja posiadała tych

regionów ok. 50 tysięcy). Baza ta pokrywa wszystkie królestwa organizmów żywych i wiele typów funkcjonalnych niekodujących RNA.

1.3.3.2 NONCODE

NONCODE (http://www.noncode.org/NONCODERv3/) (Bu, Yu i wsp. 2012) to baza danych również dedykowana niekodującym RNA (oprócz tRNA i rRNA). Posiada wiele sekwencji niekodujących RNA, z których wszystkie zostały sprawdzone ręcznie, a ponad 80% z nich pochodzi z doświadczeń. Dane poklasyfikowane są według procesów komórkowych, w których dana cząsteczka bierze udział. Obecnie baza ta jest dostępna w wersji v3.0, która posiada informacje odnośnie 411554 publicznie dostępnych sekwencji z 1239 organizmów, pokrywające wszystkie domeny życia. Pośród nich 73372 to długie niekodujące RNA, które zawierają niemal wszystkie opublikowane sekwencje z człowieka i myszy. W sumie w bazie występują 134 klasy RNA biorące udział w 26 procesach komórkowych. Dla każdej cząsteczki w bazie dostępna jest informacja o jej klasie, nazewnictwie, lokalizacji komórkowej, powiązanych publikacjach czy mechanizmach, poprzez które spełnia swoją funkcję.

1.3.3.3 Inne bazy danych

Tak jak w przypadku naprawy DNA, również w przypadku RNA istnieje szereg baz dedykowanych innym aspektom związanym z badaniami nad tymi cząsteczkami. Pomimo, że brakuje baz danych dedykowanych bezpośrednio metabolizmowi RNA znaleźć można wiele źródeł zawierających informacje o rodzajach RNA: lncRNAdb (Long Non-Coding RNA Database, http://www.lncrnadb.com, (Amaral, Clark i wsp. 2011) – baza danych eukariotycznych długich niekodujących RNA), miRBase (http://www.mirbase.org/, (Kozomara i Griffiths-Jones 2011) – baza poświecona miRNA), tRNAdb (http://trnadb.bioinf.uni-leipzig.de/) ((Juhling, Morl i wsp. 2009) – baza danych genów i sekwencji tRNA); o ich modyfikacjach: MODOMICS (http://modomics.genesilico.pl/, (Machnicka, Milanowska i wsp. 2013) – baza modyfikacji RNA), The RNA Modification Database (http://mods.rna.albany.edu/home) ((Cantara, Crain i wsp. 2011) – baza modyfikacji RNA); o RNA rybosomowych: 16S and 23S Ribosomal RNA Mutation Database (http://ribosome.fandm.edu/, (Triman, Peister i wsp. 1998)), 5S Ribosomal RNA Database (http://biobases.ibch.poznan.pl/5SData/, (Szymanski, Barciszewska i wsp. 2002));

czy o chorobach związanych z różnymi aspektami metabolizmu RNA, jak np. SpliceDisease Database (http://cmbi.bjmu.edu.cn/Sdisease, (Wang, Zhang i wsp. 2012)), która dostarcza informacji o schorzeniach związanych z nieprawidłowym wycinaniem intronów. Istnieje także projekt dostarczający informacji i strukturach i ewolucji RNA, otrzymanych w oparciu o analizy porównawcze sekwencji: The Comparative RNA Web (CRW, http://www.rna.icmb.utexas.edu/) (Cannone, Subramanian i wsp. 2002).

Tabela 1 Bazy danych związane z naprawą DNA, dojrzewaniem i degradacją RNA oraz ogólne bazy, w których można znaleźć informacje odnośnie wymienionych tematów.

Nazwa	Adres WWW	Źródło	Opis
Bazy danych poświęcone naprawie DNA.			
REPAIRtoire	http://repairtoire.genesilico.pl/	(Milanowska, Krwawicz i wsp. 2011)	Baza danych naprawy DNA.
DNAtraffic	http://dnatraffic.ibb.waw.pl/	(Kuchta, Barszcz i wsp. 2012)	Baza danych białek naprawy DNA, szlaków odpowiedzi na uszkodzenia i remodelowania chromatyny.
Human DNA Repair Genes	http://sciencepark.mdanderson.org/labs/wood/DNA_Repair_Genes.html	(Wood, Mitchell i wsp. 2005)	Baza danych ludzkich genów naprawy DNA.
Repair-FunMap	Obecnie niedostępna	(Wen i Feng 2004)	Baza danych oddziaływań pomiędzy białkami naprawy DNA a białkami zaangażowanymi w inne procesy.
Bazy danych poświęcone dojrzewaniu i degradacji RNA.			
RFAM	http://rfam.sanger.ac.uk/	(Burge, Daub i wsp. 2013)	Baza danych rodzin RNA. Wersja v11.0 posiada zbiór 2208 rodzin.
NONCODE	http://www.noncode.org/NONCODERv3/	(Bu, Yu i wsp. 2012)	Baza danych niekodujących RNA. Wersja v3.0.
lncRNAdb	http://www.lncrnadb.com	(Amaral, Clark i wsp. 2011)	Baza danych eukariotycznych długich niekodujących RNA.
miRBase	http://www.mirbase.org/	(Kozomara i Griffiths-Jones 2011)	Baza poświęcona miRNA.
MODOMICS	http://modomics.genesilico.pl/	(Machnicka, Milanowska i wsp. 2013)	Baza danych modyfikacji RNA.
16S and 23S Ribosomal RNA Mutation Database	http://ribosome.fandm.edu/	(Triman, Peister i wsp. 1998)	Zbiór informacji dotyczących 16S i 23S rRNA.
5S Ribosomal RNA Database	http://biobases.ibch.poznan.pl/5SData/	(Szymanski, Barciszewska i wsp. 2002)	Źródło danych o 5S rRNA.

Nazwa	Adres WWW	Źródło	Opis
Bazy danych poświęcone dojrzewaniu i degradacji RNA (c.d.).			
SpliceDisease Database	http://cmbi.bjmu.edu.cn/Sdisease	(Wang, Zhang i wsp. 2012)	Baza danych mutacji i chorób związanych z nieprawidłowym wycinaniem intronów.
tRNAdb	http://trnadb.bioinf.uni-leipzig.de/	(Juhling, Morl i wsp. 2009)	Baza danych genów i sekwencji tRNA.
CRW	http://www.rna.icmb.utexas.edu/	(Cannone, Subramanian i wsp. 2002).	Strona i projekt dostarczający informacji i strukturach i ewolucji RNA, otrzymanych w oparciu o analizy porównawcze sekwencji.
The RNA Modification Database	http://mods.rna.albany.edu/home	(Cantara, Crain i wsp. 2011)	Baza danych modyfikacji RNA.
Inne bazy danych związane z naprawą DNA oraz dojrzewaniem i degradacją RNA.			
KEGG	http://www.genome.jp/kegg/	(Kanehisa, Araki i wsp. 2008)	Encyklopedia genów i genomów (Kyoto Encyclopedia of Genes and Genomes)
Reactome	http://www.reactome.org/	(Matthews, Gopinath i wsp. 2009)	Baza danych głównie ludzkich szlaków metabolicznych i składających się na te szlaki reakcji.
BioCyc (EcoCyc, MetaCyc)	http://biocyc.org/	(Caspi, Altman i wsp. 2010)	Doświadczalnie przebadane szlaki metaboliczne i enzymy z ponad 2500 organizmów.
BRENDA	http://www.brenda-enzymes.org	(Scheer, Grote i wsp. 2011)	Najobszerniejsza kolekcja informacji o enzymach.
Pathway Commons	http://www.pathwaycommons.org	(Cerami, Gross i wsp. 2011)	Zbiór danych dotyczących publicznie dostępnych szlaków z różnych organizmów.
OriDB	http://www.oridb.org/	(Nieduszynski, Hiraga i wsp. 2007)	Baza danych potwierdzonych i przewidzianych miejsc rozpoczęcia replikacji DNA.
REBASE	http://rebase.neb.com/	(Roberts, Vincze i wsp. 2010)	Zbiór informacji o enzymach i genach związanych z restrykcją i modyfikacją DNA u Prokaryota.
DAnCER	http://wodaklab.org/dancer/	(Turinsky, Turner i wsp. 2011)	Baza danych powiązanej z chorobami epigenetyki chromatyny.
Telomerase database	http://telomerase.asu.edu/	(Podlevsky, Bley i wsp. 2008)	Sekwencje i struktury podjednostek telomerazy wraz z informacjami o ich mutacjach.
ReplicationDomain	http://www.replicationdomain.com/	(Weddington, Stuy i wsp. 2008)	Baza danych i narzędzia służące do analizy replikacji DNA.
Deathbase	http://deathbase.org/	(Diez, Walter i wsp. 2010)	Baza białek zaangażowanych w śmierć komórki.

2 Cel pracy

Głównym celem projektu było opracowanie architektury oraz stworzenie dwóch elementów ujednoliconego systemu bazodanowego opisującego cały metabolizm kwasów nukleinowych. Taki system pozwoli na szersze spojrzenie na biologię systemów tych cząsteczek, dogłębną analizę poszczególnych szlaków, systematyzację wiedzy, udostępnienie istniejących danych w sposób jasny, przejrzysty i czytelny, a także pozwoli w przyszłości na symulację szlaków metabolicznych oraz odkrycie i opisanie nowych, nieznanych do tej pory elementów zaangażowanych w procesy dojrzewania, modyfikacji i degradacji DNA i RNA. Otrzymany system powinien nie tylko umożliwić zastosowanie podejścia systemowego w analizie biologii kwasów nukleinowych, ale także wygenerować informacje, które są czymś więcej niż jedynie zwykłą sumą danych wejściowych. **Opracowanie tego typu systemu będzie stanowiło krok milowy w dziedzinie biologii systemów kwasów nukleinowych, pozwoli lepiej zrozumieć wiele ważnych procesów biologicznych i da silny impuls do dalszych analiz doświadczalnych.** W ramach niniejszego projektu opracowywane zostały dwie bazy danych: REPAIRtoire – baza danych szlaków naprawy DNA i RNApathwaysDB - baza danych dojrzewania i degradacji RNA.

3 Materiały

3.1 Sprzęt komputerowy

W celu zrealizowania założeń pracy doktorskiej zostały wykorzystane następujące zasoby obliczeniowe Pracowni Bioinformatyki Uniwersytetu im. Adama Mickiewicza w Poznaniu oraz Laboratorium Bioinformatyki i Inżynierii Białka w Międzynarodowym Instytucie Biologii Molekularnej i Komórkowej w Warszawie:

- Komputery PC w Pracowni Bioinformatyki Uniwersytetu im. Adama Mickiewicza w Poznaniu (stacje robocze podobnej klasy).

Procesor	INTEL Xeon E5645 2.4G 6-CORE 12MB, LGA1366
Płyta główna	INTEL Workstation Board S5520SC, 5520, DDR3-1333, SATA, RAID, 2xGBLAN
Pamięć RAM	Kingston ValueRAM DDR3 (16GB,1066MHz,ECC,Reg,QRx4 w/TS) CL7 (KVR1066D3Q4R7S/16G)
Dysk	Seagate 2 TB Barracuda (64MB, SATA/600) OCz Vertex 3 SSD 2,5cala 120GB 550 MB/s 500 MB/s
System operacyjny	Ubuntu 12.10

- Serwer w Laboratorium Bioinformatyki i Inżynierii Białka w Międzynarodowym Instytucie Biologii Molekularnej i Komórkowej w Warszawie o parametrach: 4 procesory AMD Opteron 6174 o częstotliwości 2,2 GHz (48 rdzeni), 96 GB pamięci RAM oraz 9 TB przestrzeni dyskowej.

4 Metody

4.1 Narzędzia programistyczne

Bazy danych rozwijane w ramach niniejszej pracy zostały zaimplementowane z użyciem wzorca projektowego Django (patrz: Podrozdział 4.1.4), napisanego w języku Python (patrz: Podrozdział 4.1.1). Ponadto wykorzystano wiele bibliotek samego języka Python do rozwiązania typowych problemów programistycznych (takich jak np. komunikacja z systemem operacyjnym czy rysowanie szlaków metabolicznych) (patrz: Podrozdział 4.1.2). Użyte zostały również biblioteki dedykowane problemom biologicznym, np. pakiet Entrez z biblioteki Biopython, który pozwolił na obsługę referencji literaturowych dodawanych do baz danych. Dzięki temu pakietowi możliwa była komunikacja z bazą danych PubMed (Sayers, Barrett i wsp. 2009) i automatyczne dodawanie informacji o publikacjach związanych z danymi rekordami tworzonych baz danych.

4.1.1 Język programowania Python

Język programowania Python to język skryptowy oparty o programowanie obiektowe, który powstał w latach dziewięćdziesiątych (twórcą jest Guid van Rossum), o nazwie pochodzącej od nazwy brytyjskiego serialu *„Latający cyrk Monty Pythona”*. Jest on rozwijany pod licencją otwartego oprogramowania, a implementacja i dokumentacja są dostępne na stronie organizacji *Python Software Fundation* (http://www.python.org/).

Najważniejsze cechy języka Python to, że jest on:

- **interpretowany** - kod jest na bieżąco tłumaczony i wykonywany przez interpreter bez konieczności kompilacji i konsolidacji, dzięki czemu jest on niezależny od platformy, choć nie tak wydajny jak kod języków kompilowanych (np. C++);
- **obiektowy** - zawiera klasy, wyjątki, metody, automatyczne zarządzanie pamięcią, dziedziczenie, choć styl programowania nie jest wymuszany i możliwe jest również programowanie strukturalne i funkcjonalne;
- **dynamicznie rozwijany i wspierany**, co więcej posiada solidnie przygotowaną dokumentację;

- **przenośny**, co daje możliwość uruchamiania skryptu niezależnie od platformy systemowej – dzięki temu rdzeń języka Python działa tak samo na wszystkich systemach operacyjnych;
- **integrowalny** – zapewnia prostą integrację z komponentami napisanymi w innych językach programowania, np. w C; dzięki temu można łączyć ze sobą fragmenty kodu napisane w różnych językach programowania (co niejednokrotnie odbywałoby się kosztem obniżenia wydajności kodu);
- **wydajny i prosty** – został zaopatrzony w szereg wbudowanych typów i narzędzi, posiada dostęp do licznych bibliotek standardowych oraz bibliotek umożliwiających pracę z danymi biologicznymi (Chapman i Chang 2000, Cock, Antao i wsp. 2009);
- **przejrzysty** - kod sterowany jest przez wcięcia (indentacje, standardowo cztery znaki spacji), przez co programy pisane w tym języku są przejrzyste i mają łatwą w czytaniu składnię, co ułatwia zewnętrznym użytkownikom wgląd w kod;
- **posiada szerokie możliwości zastosowania** – od narzędzi umożliwiających administrowanie systemem, poprzez skrypty internetowe (bezpłatny serwer WWW Zope) i bazy danych po biblioteki do programowania numerycznego, przetwarzania obrazów i sztucznej inteligencji;
- **jest językiem wysokiego poziomu** – pod względem składni i semantyki przypomina język ludzki.

Wybór języka programowania Python do zrealizowania projektu wiązał się z faktem, iż obecnie większość bibliotek naukowych i programów wykorzystywanych w bioinformatyce korzysta właśnie z niego.

4.1.2 Biblioteki programistyczne

4.1.2.1 Pygraphviz

PyGraphviz (http://networkx.lanl.gov/pygraphviz/) to biblioteka języka Python, będąca nakładką na pakiet Graphiviz (http://www.graphviz.org/), służący do rysowania i wizualizacji grafów. Jest ona dostępna na otwartej licencji BSD (ang. *Berkeley Software Distribution License*). Dzięki tej bibliotece możliwe jest tworzenie, edytowanie, czytanie, zapisywanie i rysowanie grafów za pomocą języka Python, który pośredniczy pomiędzy

użytkownikiem a algorytmami pakietu Graphviz. PyGraphviz został użyty w obu bazach danych w celu tworzenia i wizualizacji grafów, będących graficznym odwzorowaniem szlaków metabolicznych, zarówno naprawy DNA jak i dojrzewania i degradacji RNA. Kod 1 stanowi przykład wykorzystania opisywanej biblioteki w bazie danych RNApathwaysDB.

Kod 1 Przykład użycia biblioteki PyGraphviz w bazie danych RNApathwaysDB

```
import pygraphviz as pgv

def _createGraph(reactions_list, pathway, bgcolor="#e7f3e9", download=False):
    """
    creates a graph - uses pygraphviz
    node = RNAState
    edge = Reaction
    each node has it's own PNG picture
    graph has a graph picture with a html map that links all edges and nodes to their own pages
    """

    path = BASE_PATH + '/'
    pathway_name = pathway.name
    G=pgv.AGraph(comment = str(pathway.name), bgcolor=bgcolor, center="true", cyclic="true", \
                 splines="false",  ordering="out", nodesep="1.0", ranksep='0.1')

    for reaction in reactions_list:
        if reaction.step_before:
            node_1 = str(reaction.step_before.id)
            RNAState_1 = RNAState.objects.get(pk=node_1)
            label = str(RNAState_1.name)
            if len(label) > 80:
                labels = label.split()
                new_label = ""
                check_point, counter = float(len(labels)/2), 0
                for el in labels:
                    if counter == check_point: new_label = new_label + el + "\r"
                    else:new_label = new_label + el + " "
                    counter += 1
            else: new_label = label
            try:
                open(path + str(reaction.step_before.picture).replace('svg','png'))
                picture_1 = path + str(reaction.step_before.picture).replace('svg','png')
            except:
                try:
                    svg = '\n'.join(open(path + str(reaction.step_before.picture)).readlines())
                    width, height = 1000, 300
                    _svgToPNG(str(reaction.step_before.picture), width, height)
                    open(path + str(reaction.step_before.picture).replace('svg','png'))
                    picture_1 = path + str(reaction.step_before.picture).replace('svg','png')
                except:
                    picture_1 = path + 'images/step_pictures/no_picture.png'

            if not G.has_node(node_1): G.add_node(node_1, style="filled", color="white", \
                                                  image=picture_1, imagescale="true", \
                                                  label = new_label, labelloc="t", shape="box",\
                                                  fixedsize="true", width="7", height="3" )
        else: pass

        if reaction.step_after:
            node_2 = str(reaction.step_after.id)
            RNAState_2 = RNAState.objects.get(pk=node_2)
            label = str(RNAState_2.name)
            if len(label) > 80:
                labels = label.split()
                new_label = ""
                check_point, counter = float(len(labels)/2), 0
                for el in labels:
                    if counter == check_point: new_label = new_label + el + "\r"
                    else: new_label = new_label + el + " "
                    counter += 1
```

```
            else: new_label = label
            try:
                open(path + str(reaction.step_after.picture).replace('svg','png'))
                picture_2 = path + str(reaction.step_after.picture).replace('svg','png')
            except:
                try:
                    svg = '\n'.join(open(path + str(reaction.step_after.picture)).readlines())
                    width, height = 1000, 300
                    _svgToPNG(str(reaction.step_after.picture), width, height)
                    open(path + str(reaction.step_after.picture).replace('svg','png'))
                    picture_2 = path + str(reaction.step_after.picture).replace('svg','png')
                    #print '2', width, height
                except:
                    picture_2 = path + 'images/step_pictures/no_picture.png'
            if not G.has_node(node_2): G.add_node(node_2, style="filled", color="white", \
                                                  label = new_label, labelloc="t", shape='box',\
                                                  fixedsize="true", image=picture_2,\
                                                  imagescale="true", width="7", height="3")
        else: pass
        if reaction.step_before and reaction.step_after:
            if not G.has_edge(node_1, node_2): G.add_edge(node_1, node_2, key=str(reaction.id),\
                                                          comment=str(reaction.id),\
                                                          nojustify="true", label="\t",\
                                                          dir="forward", tailport="s",\
                                                          headport="n")
        else: pass

    G.layout(prog='dot')
    if download:
        graph_path = BASE_PATH + '/images/pathway_pictures/'+pathway_name+'_white.png'
        graph_access_name = '/images/pathway_pictures/'+pathway_name+'_white.png'
    else:
        graph_path = BASE_PATH + ' /images/pathway_pictures/'+pathway_name+'.png'
        graph_access_name = '/images/pathway_pictures/'+pathway_name+'.png'
    G.draw(graph_path)
    G = G.string().replace('strict graph (Alonso 2012)','').replace('\n','').split(';')

    return G, graph_access_name
```

4.1.2.2 Standardowe biblioteki języka Python

Bogaty zasób bibliotek dostępnych dla języka Python oferuje gotowe rozwiązania typowych zadań pojawiających się podczas rozwijania oprogramowania. Biblioteki *os* i *sys* służą do komunikacji z systemem i były używane np. w celu uzyskania dostępu do bazy danych sekwencji stworzonej w celu przeszukiwania tworzonych baz za pomocą sekwencji z wykorzystaniem algorytmu BLAST. Poszukiwanie błędów w kodzie zostało ułatwione dzięki wykorzystaniu biblioteki *pdb*, która pozwala przenieść się w dowolne miejsce wykonywanego kodu i na bieżąco sprawdzać rezultat działania poszczególnych linii kodu. Raportowanie błędów możliwe było dzięki bibliotece *traceback*. Operacje matematyczne, wykonywane między innymi podczas rysowania stanów DNA występujących w obrębie szlaków naprawy w bazie REPAIRtoire, były możliwe dzięki funkcjom biblioteki *math*. Do poszukiwania wzoru w tekście stosowana była biblioteka *re*, np. wykorzystano ją przy formatowaniu i wyświetlaniu wyników wyszukiwania za pomocą algorytmy BLAST. Biblioteki *rsvg* i *cairo* służyły do przekształcania rysunków w formacie SVG (ang. *scalable vector graphics*) do formatu PNG, który był obsługiwany przez przeglądarkę Internet

Explorer. *Pickle* pozwalała na zapisywanie obiektów Pythona do plików, podczas gdy *urllib* umożliwiała łączenie się z zewnętrznymi bazami danych w celu pobrania informacji odnośnie poszczególnych białek, cząsteczek RNA czy publikacji. Tabela 2 przedstawia informacje na temat wykorzystywanych bibliotek.

Tabela 2 Biblioteki języka Python wykorzystane podczas implementacji baz danych.

Biblioteka	Opis	Przykład wykorzystanych funkcji
sys	Pozwala na dostęp do obiektów i funkcji używanych przez interpreter języka Python.	path() , exc_info()
os	Daje możliwość korzystania z funkcji zależnych od systemu operacyjnego, w tym na manipulację ścieżkami czy czytanie danych z katalogu.	os.path(), os.listdir(), remove()
math	Zawiera liczne funkcje, operacje i stałe matematyczne. Znajdują się tu np. funkcje trygonometryczne, funkcja przekształcająca wartości kątów ze stopni na radiany i odwrotnie, funkcja pierwiastkująca, stała PI i wiele innych.	sqrt()
cairo, rsvg	Pozwalają na manipulację obrazkami 2D, w tym konwersję pomiędzy formatami, np. z SVG do PNG.	rsvg.props(), cairo.ImageSurface(), cairo.Context(), render_cairo(), write_to_png()
re	Umożliwia obsługę wyrażeń regularnych, pozwalając na wyszukiwanie wzorców w tekście oraz na manipulację nimi.	findall(), finditer(), sub()
pickle	Pozwala na zapisywanie dowolnych obiektów Pythona w pliku, i odwrotnie, do późniejszego odczytywania i rekreacji obiektów.	dump(), load()
urllib	Zawiera funkcje służące do pobierania informacji o danych, a także pobierania samych danych z Internetu na podstawie adresu URL (głównie strony internetowe).	urlopen()
traceback	Dostarcza standardowy interfejs służący do wydobycia, formatowania i wyświetlania błędów pojawiających się w trakcie wykonywania programu.	format_exc()
datetime	Moduł służący do manipulacji datami, ich wydobywania i formatowania.	datetime.now()
pdb	Pozwala na interaktywne wyszukiwanie błędów w kodzie.	set_trace()

4.1.2.3 Biblioteka Biopython

Biblioteką języka Python stworzoną do analizy danych biologicznych jest BioPython (Chapman i Chang 2000, Cock, Antao i wsp. 2009) (http://biopython.org). Pakiet ten zawiera: „parsery", czyli narzędzia ułatwiające przetwarzanie plików w formatach najczęściej używanych przez biologów (m.in. BLAST, FASTA, Clustal), narzędzia ułatwiające dostęp

do serwisów internetwych (NCBI, Expasy, PDB), interfejsy do popularnych programów (Clustal), metody do klasteryzacji sekwencji i wiele innych. W niniejszej pracy korzystano głównie z pakietów *Bio.Entrez* oraz *Bio.Medline*, które służą do łączenia się z bazami danych NCBI poprzez podanie adresu WWW. *Bio.Entrez* umożliwia połączenie się z bazą danych, w tym projekcie z bazą PubMed, wyszukanie interesujących rekordów i wydobycie z nich niezbędnych informacji. *Bio.Medline* jest parserem pozwalającym na wyodrębnienie z wyników wyszukiwania za pomocą *Bio.Entrez* wszystkich informacji odnośnie pożądanej publikacji z bazy danych PubMed w postaci słownika Pythona. Moduły te wykorzystywane były do wydobycia z baz danych NCBI informacji dotyczących poszczególnych białek czy cząsteczek RNA. Ponadto, dzięki nim możliwe było zaimplementowanie zautomatyzowanego sposobu dodawania do baz danych publikacji na podstawie ich numeru PMID.

4.1.3 Bazy Danych MySQL

4.1.3.1 MySQL

MySQL (www.mysql.com, (AB. 2005)) jest ogólnodostępnym systemem zarządzania relacyjnymi bazami danych, rozwijanym przez firmę Oracle. Jest obecnie najpopularniejszym silnikiem relacyjnych baz danych i zarazem jednym z szybciej i wydajniej działających. Udostępniany na otwartej licencji GPL (licencja wolnego i otwartego oprogramowania, ang. *GNU General Public License*) system ten ma otwarty dostęp do kodu źródłowego i jest wykorzystywany do tworzenia niedrogich, skutecznych i skalowalnych wbudowanych aplikacji bazodanowych w sieci Internet. Do komunikacji z innymi programami używa standardowego języka zapytań (SQL, *Structured Query Language*) wraz z własnymi rozszerzeniami jego funkcji.

Główne cechy systemu: :

- jest on napisany w C i C++ - co decyduje o jego wydajności;
- udostępnia API dla wielu języków programowania: C, C++, Eiffel, Java, Perl, PHP, Python, Ruby, Tcl;
- posiada wielowątkowość, korzystającą z wątków jądra, co oznacza, że możliwa jest praca na maszynie wieloprocesorowej;
- dostępna jest opcjonalna obsługa transakcji;
- daje możliwość „osadzenia" (ang. *embed*) serwera MySQL w pisanej aplikacji;

- udostępnia dużą liczbę typów danych w kolumnach (np. liczby, ciągi znakowe, obiekty binarne (BLOB), datę i czas, typy wyliczeniowe, zestawy);
- na uwagę zasługuje fakt, że w MySQL można daną kolumnę dostosować do typu danych, które zamierza się w niej przechowywać (np. TINYINT, a nie INT), tym samym uzyskuje się większą wydajność i mniejsze zużycie pamięci (również dyskowej);
- istnieje możliwość definiowania niektórych typów danych jako narodowych (różne standardy kodowania);
- posiada obsługę klauzul agregujących i grupujących SQL;
- udostępnia komendy typu: *SHOW* (pozwala przeglądać informacje na temat baz, tabel i indeksów), czy *EXPLAIN* (opisuje pracę optymalizatora zapytań);
- posiada bardzo prosty w obsłudze system zabezpieczeń, w którym wszystkie hasła są szyfrowane;
- dostępne są proste w obsłudze narzędzia administracyjne, m.in. phpMyAdmin (za pomocą przeglądarki internetowej), MySQL Workbench czy MySQL Query Browser.

Wszystkie wyżej wymienione cechy sprawiły, że system ten został wybrany do projektu bazy danych dojrzewania i degradacji RNA – RNApathwaysDB.

4.1.3.2 SQLite

SQLite (www.sqlite.org, (Newman 2005)) jest zarówno systemem zarządzania bazą danych jak i biblioteką C, implementującą taki system, która obsługuje język SQL. System ten został stworzony przez Richarda Hippa, a kod dostępny jest w domenie publicznej (ang. *public domain*).

SQLite posiada również API do innych niż C języków programowania, a mianowicie: ActionScript, Perl, PHP, Ruby, C++, Delphi, Python, Java, Tcl, Visual Basic, platformy .NET i wielu innych.

SQLite obsługuje między innymi: zapytania zagnieżdżone, widoki, klucze obce, transakcje, wyzwalacze (częściowo), definiowanie własnych funkcji, przechowywanie baz danych w pamięci RAM komputera, co znacznie przyspiesza działanie.

Ważną cechą systemu jest to, że zawartość bazy danych przetrzymywana jest w jednym binarnym pliku (do 2 TB), którego bezpieczeństwo oparte jest na zabezpieczeniach oferowanych przez używany system plików. Istnieje też pakiet oferujący szyfrowanie baz

danych SQlite na bieżąco. Cecha ta wpłynęła na wybranie tego systemu do projektu bazy danych napraw DNA – REPAIRtoire, ze względu na potrzebę szybkiego przenoszenia i tworzenia kopii zapasowych całej bazy danych.

Tak jak dla MySQL, dla SQLite również istnieją narzędzia do zarządzania bazami danych. Są to m.in. SQLite Manager (dodatek przeglądarki Mozilla Firefox), DaDaBIK Database Interface Kreator (Open Source), czy SQLite Database Browser (narzędzie graficzne).

4.1.4 Wzorzec projektowy Django

Przy wyborze narzędzia, które umożliwiłoby stworzenie dynamicznych stron WWW dla baz danych rozwijanych w niniejszym projekcie kierowano się kilkoma założeniami:

1. narzędzie bazuje na języku programowania Python, ze względów opisanych wcześniej (patrz: Rozdział 4.1.1);
2. narzędzie jest bardzo dobrze udokumentowane oraz wspierane;
3. narzędzie rozwijane jest pod licencją otwartego oprogramowania.

Wzorzec projektowy Django (https://www.djangoproject.com/), spełniający wyżej wymienione założenia, dostępny jest na licencji BSD (ang. *Berkeley Software Distribution License*), liberalnej, skupionej na prawach użytkownika. Powstał pod koniec 2003 roku jako ewolucyjne rozwinięcie aplikacji internetowych, tworzonych przez grupę programistów związanych z Lawrence Journal-World, a jego nazwa pochodzi od imienia gitarzysty Django Reinhardta. Oprócz sporych zasobów internetowych na jego temat, informacji o nim można zasięgnąć zarówno w literaturze anglojęzycznej jak i polskojęzycznej. Cecha ta jest bardzo istotna podczas rozwijania oprogramowania, gdyż daje możliwość szybkiego rozwiązywaniu różnego typu problemów i niejasności napotkanych w trakcie projektowania i implementacji kodu.

4.1.4.1 Architektura Django

Wzorzec projektowy Django różni się od innych tego typu narzędzi architekturą i ideologią działania. Projektując serwisy internetowe najczęściej ma się do czynienia z budową opartą na wzorcu Model-Widok-Kontroler (ang. *Model-View-Controller*) (MVC). Django natomiast posiada architekturę Model-Szablon-Widok (ang. *Model-Template-View*) (MTV), która w warstwie prezentacji jest znacznie bliższa wzorcowi Model-Widok-Prezenter (ang. *Model-View-Presenter*) (MVP). Jednakże wszystkie te podejścia łączy wspólna cecha:

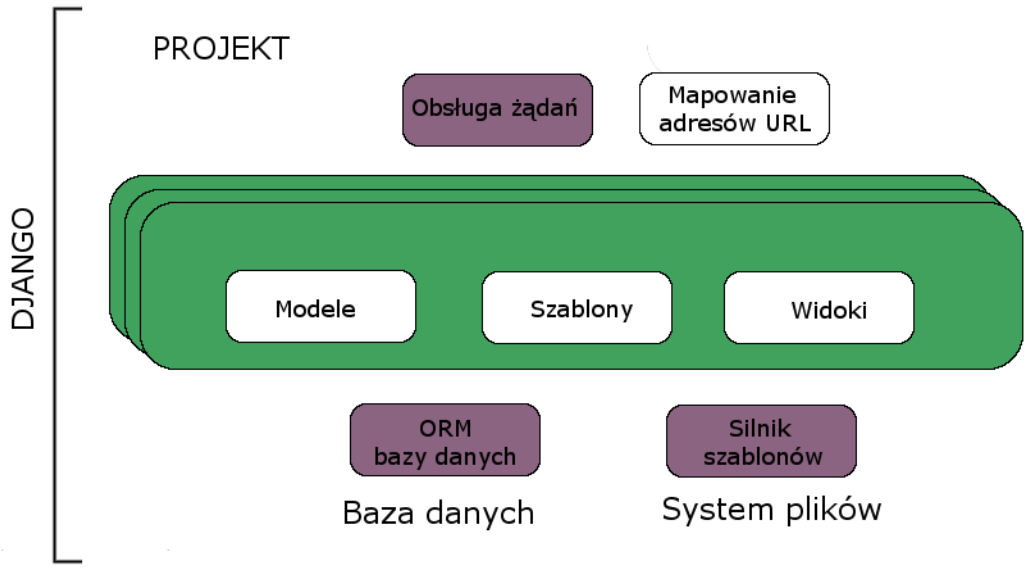
oddzielenie logiki biznesowej (*model*), logiki aplikacji (*widok*), wyglądu (*szablony*) oraz baz danych, co ułatwia projektowanie, implementowanie, przenoszenie i wdrażanie projektu. Architektura ta zostanie przedstawiona poniżej (patrz: Rycina 2) (Forcier, Bissex i wsp. 2008).

Warstwa *modelu* jest odpowiedzialna za wszelkie obiekty, które służą do wykonywania operacji związanych z implementacją funkcjonalności aplikacji. *Szablon* jest częścią warstwy prezentacji i decyduje o sposobie wyświetlania danych: otrzymuje on od *widoku* informacje, które dane zaprezentować i określa jak te dane reprezentować.

Widok stanowi warstwę tzw. logiki aplikacji. Jest on pomostem pomiędzy *modelami* i *szablonami*, który realizuje żądania pobrania danych z *modelu* i wyświetlania ich w *szablonach*. Warstwę tę przyrównuje się do powszechni znanej warstwy kontrolera w MVC. Różnicę stanowi fakt, że *widok* w Django jest odpowiedzialny jedynie za to, *co* ma być wyświetlane, a nie *jak*.

Do najistotniejszych zalet Django należą:

- kompletny i automatycznie generowany panel administracyjny, z możliwością dalszego dostosowywania jego wyglądu i funkcjonalności;
- funkcjonalny, lecz nieskomplikowany system szablonów, czytelny zarówno dla grafików jak i dla programistów;
- zoptymalizowana wydajności działania;
- bardzo przydatny i prosty w uruchomieniu serwer testowania aplikacji;
- zasada DRY (ang. *Don't Repeat Yourself*), czyli „nie powtarzaj się" w odniesieniu do tworzenia kodu aplikacji;
- zasada KISS (ang. *Keep It Simple, Stupid*), czyli dążenie do tworzenia rozwiązań, które są proste, przejrzyste i zrozumiałe;
- mapowanie obiektowo-relacyjne (ang. *Object-Relational Mapping*, ORM) wysokiego poziomu, pozwalające na łatwe i bezpieczne operowanie na bazach danych bez konieczności znajomości składni języka baz danych SQL;
- obsługa następujących bazy danych: PostgreSQL, MySQL, SQLite oraz Oracle.



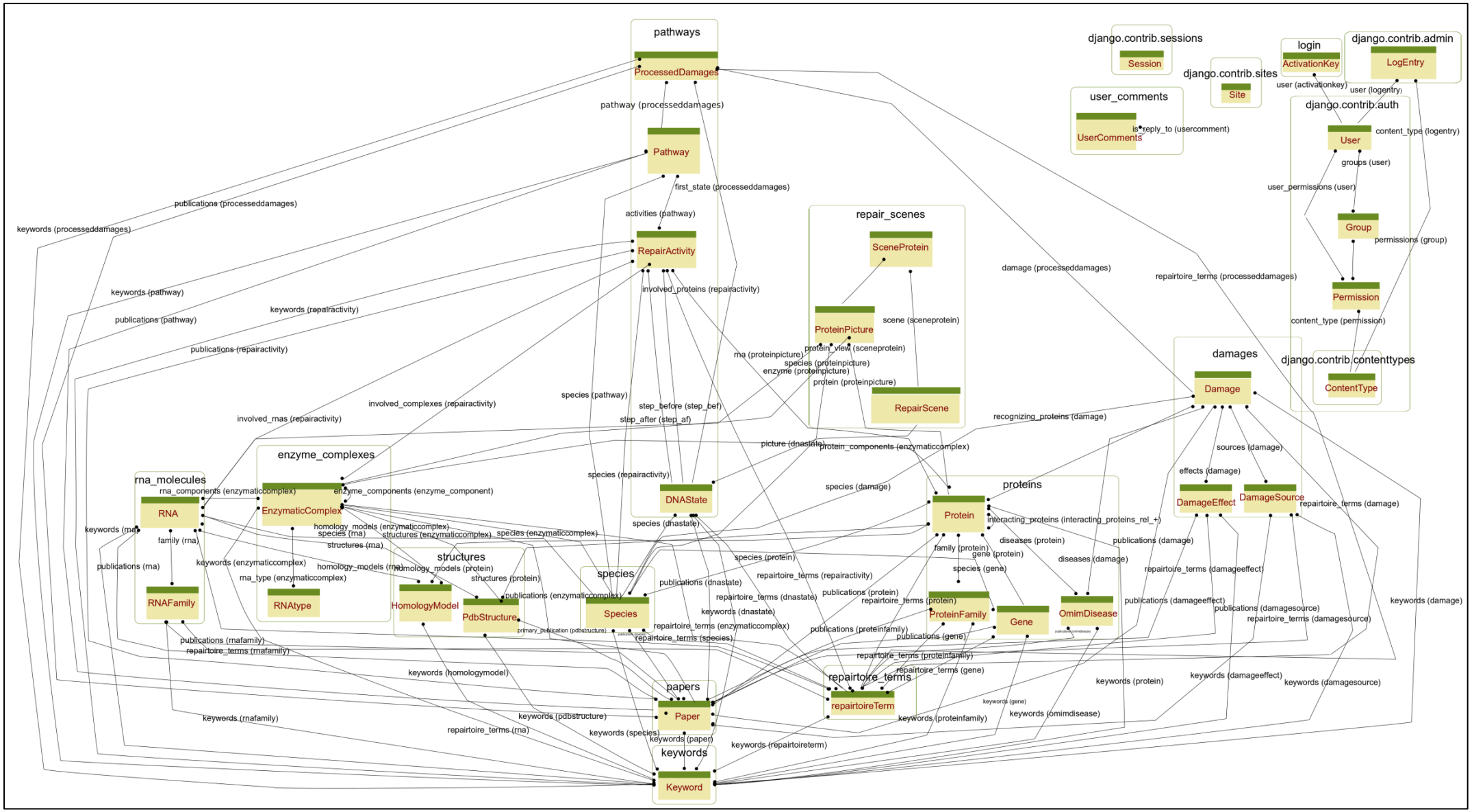
Rycina 2 Architektura typu Model-Szablon-Widok wykorzystywana w Django.

4.1.4.2 Projekt struktury baz

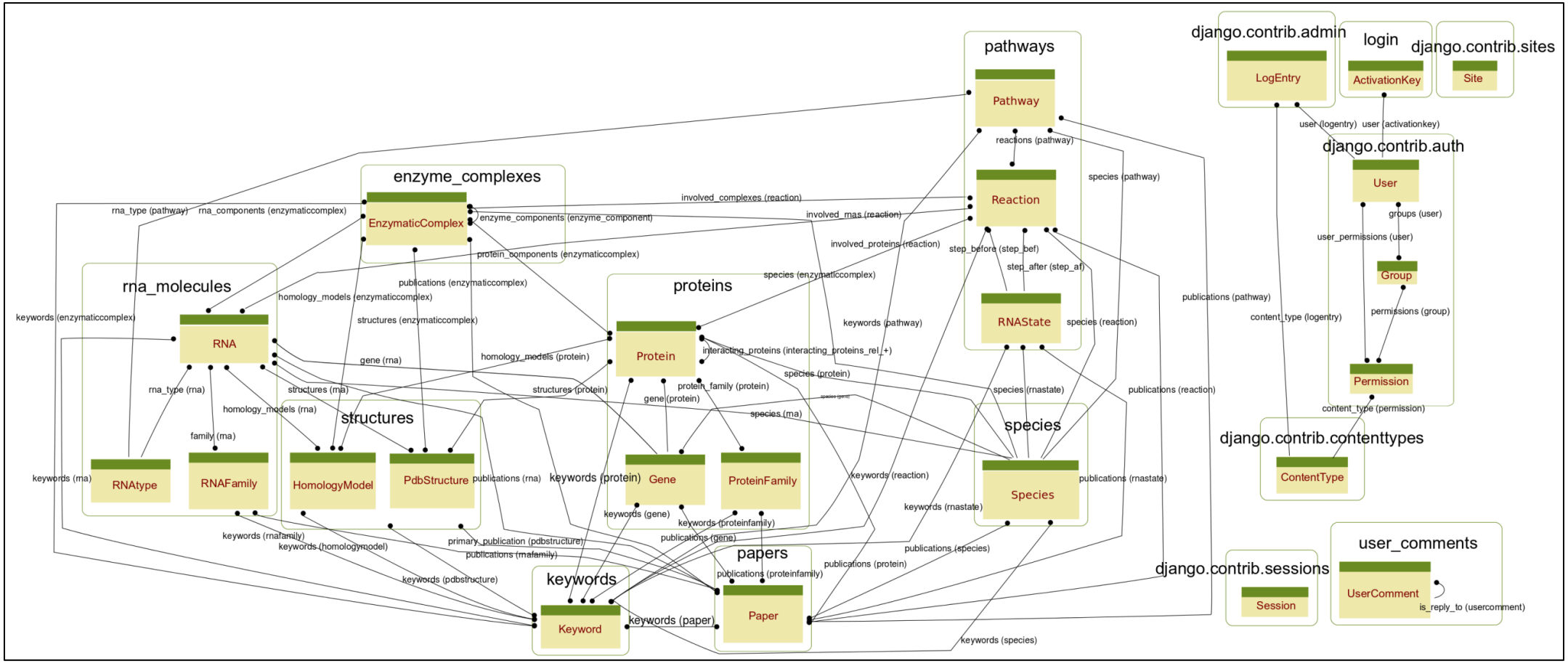
W trakcie pracy zaprojektowano dla powstających baz danych ogólną architekturę, która następnie do każdej bazy została odpowiednio dostosowana. Baza danych REPAIRtoire zawiera 26 tabel (nie wliczając standardowych tabel Django, dostępnych od razu po instalacji, których jest 7), natomiast baza danych RNApathwaysDB składa się z 17 tabel.

Rycina 3 przedstawia schematy obu baz danych.

A)


pathways
ProcessedDamages
pathway (processeddamages)
Pathway
first_state (processeddamages)
publications (processeddamages)
keywords (processeddamages)
activities (pathway)
RepairActivity
keywords (pathway)
keywords (repairactivity)
publications (pathway)
publications (repairactivity)
involved_proteins (repairactivity)
involved_complexes (repairactivity)
involved_rnas (repairactivity)
species (pathway)
step_before (step_bef)
step_after (step_af)
repair_scenes
SceneProtein
ProteinPicture
scene (sceneprotein)
protein_view (sceneprotein)
rna (proteinpicture)
species (proteinpicture)
enzyme (proteinpicture)
protein (proteinpicture)
RepairScene
damage (processeddamages)
repairtoire_terms (processeddamages)
recognizing_proteins (damage)
picture (dnastate)
protein_components (enzymaticcomplex)
species (repairactivity)
species (damage)
DNAState
species (dnastate)
enzyme_complexes
rna_components (enzymaticcomplex)
enzyme_components (enzyme_component)
EnzymaticComplex
homology_models (enzymaticcomplex)
species (rna)
structures (enzymaticcomplex)
species (enzymaticcomplex)
structures (rna)
rna_molecules
RNA
keywords (rna)
family (rna)
publications (rna)
keywords (enzymaticcomplex)
rna_type (enzymaticcomplex)
RNAFamily
RNAtype
structures
homology_models (rna)
homology_models (protein)
structures (protein)
HomologyModel
PdbStructure
publications (enzymaticcomplex)
primary_publication (pdbstructure)
species
Species
publications (dnastate)
repairtoire_terms (dnastate)
keywords (dnastate)
repairtoire_terms (enzymaticcomplex)
repairtoire_terms (species)
species (protein)
repairtoire_terms (repairactivity)
publications (rnafamily)
repairtoire_terms (rnafamily)
keywords (homologymodel)
keywords (rnafamily)
keywords (pdbstructure)
repairtoire_terms (rna)
keywords (species)
papers
Paper
keywords (paper)
keywords (repairtoireterm)
keywords
Keyword
proteins
Protein
interacting_proteins (interacting_proteins_rel_+)
family (protein)
gene (protein)
species (gene)
publications (protein)
repairtoire_terms (protein)
publications (proteinfamily)
repairtoire_terms (proteinfamily)
ProteinFamily
Gene
publications (gene)
repairtoire_terms (gene)
keywords (gene)
keywords (proteinfamily)
diseases (protein)
diseases (damage)
OmimDisease
keywords (omimdisease)
repairtoire_terms
repairtoireTerm
damages
Damage
sources (damage)
effects (damage)
DamageEffect
DamageSource
repairtoire_terms (damage)
publications (damage)
repairtoire_terms (damageeffect)
publications (damageeffect)
publications (damagesource)
repairtoire_terms (damagesource)
keywords (damage)
keywords (protein)
keywords (damageeffect)
keywords (damagesource)
django.contrib.sessions
Session
django.contrib.sites
Site
user_comments
UserComments
is_reply_to (usercomment)
login
ActivationKey
user (activationkey)
django.contrib.admin
LogEntry
user (logentry)
django.contrib.auth
User
content_type (logentry)
groups (user)
user_permissions (user)
Group
permissions (group)
Permission
content_type (permission)
django.contrib.contenttypes
ContentType

B)



Rycina 3 Schematy architektury baz danych. A) REPAIRtoire B) RNApathwaysDB

Jak wcześniej wspomniano, ze względu na różną tematykę podejmowaną przez powstałe bazy danych, poza wspólnym szkieletem, który może być wykorzystany we wszystkich bazach opisujących metabolizm kwasów nukleinowych, w obu bazach występują tabele charakterystyczne tylko dla danego tematu. Spis wszystkich tabel, które zostały zaimplementowane w obu bazach danych przedstawia Tabela 3.

Tabela 3 Tabele zaimplementowane w obu stworzonych bazach danych (wyłączając standardowe tabele Django).

Wspólne dla obu baz danych	**Występujące tylko w REPAIRtoire**	**Występujące tylko w RNApathwaysDB**
Aplikacja: pathways (szlaki)		
Pathway	RepairActivity, DNAState, ProcessedDamages	Reaction, RNAState
Aplikacja: species (organizmy)		
Species		
Aplikacja: proteins (białka)		
Protein, ProteinFamily, Gene	OmimDisease	
Aplikacja: rna_molecules (cząsteczki RNA)		
RNA, RNAFamily		RNAtype
Aplikacja: enzyme_complexes (kompleksy enzymatyczne)		
EnzymaticComplex	RNAtype	
Aplikacja: structures (struktury)		
PdbStructure, HomologyModel		
Aplikacja: keywords (słowa kluczowe)		
Keyword		
Aplikacja: papers (publikacje)		
Paper		
Aplikacja: user_comments (komentarze użytkowników)		
UserComments		
Aplikacja: login (logowanie)		
ActivationKey		
Aplikacja: repair_scenes (rysunki stanów naprawy DNA)		
	SceneProtein, ProteinPicture, RepairScene	
Aplikacja: damages (uszkodzenia)		
	Damage, DamageEffect, DamageSource	
Aplikacja: repairtoire_terms (terminy)		
	repairtoireTerm	

4.1.4.3 Implementacja bazy danych (*models*)

Django narzuca ustrukturyzowaną formę projektu, który składa się z aplikacji będących poszczególnymi częściami składowymi całości – modułami projektu. Każda aplikacja opisuje pojedynczy zbiór logicznie ze sobą związanych elementów składowych, jednakże nie narzuca sposobu organizacji tych elementów. Metoda podziału elementów na aplikacje uzależniona jest od twórcy, który dobiera ją w sposób najlepiej pasujący do rozwiązywanego problemu. Każda aplikacja składa się z systemu plików, które odpowiadają wcześniej wymienionym składowym architektury Django: modelom (pliki *models.py*), widokom (pliki *views.py*) i szablonom (pliki *.html* oraz *urls.py*). Kod niezbędny do implementacji wszystkich tych elementów jest zatem rozdzielony, dzięki czemu cała struktura staje się przejrzysta i uporządkowana.

Dzięki systemowi ORM, czyli mapowaniu obiektowo-relacyjnemu, wszystkie tabele w bazie danych zostały opisane za pomocą osobnych, wcześniej wymienionych plików *models.py*. *Models.py* zawierają opisy tabel bazy danych reprezentowanych przez klasy w języku Python, czyli modele. Modele określają w czytelny sposób strukturę tabel oraz relacje pomiędzy nimi. Na ich podstawie Django automatycznie generuje zapytania SQL. Te ostatnie dotyczą między innymi tworzenia samych tabel, ich edycji oraz pobierania danych. W wielu sytuacjach nie trzeba zatem w ogóle używać języka SQL, co zapewnia lepszą kompatybilność kodu (obsługę wielu systemów bazodanowych) i sprzyja zachowaniu wcześniej wymienionej zasady DRY. Ponadto system ORM stanowi wygodny mechanizm obsługi samych danych na poziomie poszczególnych wierszy i tabel, z możliwością definiowania własnych metod dla obiektów.

W niniejszej pracy zaimplementowano modele dla:

- 13 aplikacji w projekcie REPAIRtoire;
- 10 aplikacji w projekcie RNApathwaysDB,

które mają swoje odzwierciedlenie w bazach danych (patrz: Rycina 3). Tabela 4 przedstawia zestawienie aplikacji stworzonych baz danych wraz z wyróżnieniem klas, opisujących poszczególne tabele baz.

Tabela 4 Zestawienie aplikacji REPAIRtoire oraz RNApathwaysDB.

Nazwa aplikacji	Baza danych, w której aplikacja występuje	Funkcje
pathways	REPAIRtoire, RNApathwaysDB	Obiekty tej klasy reprezentują szlaki metaboliczne: naprawy DNA, dojrzewania czy degradacji RNA. Szlak składa się z aktywności (również będących obiektami tej klasy), które reprezentują zdarzenie i same składają się ze stanów (stan przed i po aktywności, który także jest obiektem tej klasy).
species	REPAIRtoire, RNApathwaysDB	Obiekty tej klasy reprezentują organizmy, dla których przedstawiane są informacje. Organizm połączony jest ze wszystkimi tabelami w bazach danych: od białek po aktywności czy szlaki.
proteins	REPAIRtoire, RNApathwaysDB	Obiekty przechowują dane o białkach i genach je kodujących.
rna_molecules	REPAIRtoire, RNApathwaysDB	Obiekty przechowują dane o cząsteczkach RNA i genach je kodujących.
enzyme_complexes	REPAIRtoire, RNApathwaysDB	Obiekty przechowują dane o kompleksach enzymatycznych, ich budowie i składzie podjednostkowym.
structures	REPAIRtoire, RNApathwaysDB	Obiekty przechowują dane o strukturach PDB powiązanych z cząsteczkami zebranymi w obu bazach danych.
keywords	REPAIRtoire, RNApathwaysDB	Obiekty przechowują dane o słowie kluczowym, jego nazwę i opis. Słowa kluczowe przypisane są do wszystkich rekordów występujących w bazach danych.
papers	REPAIRtoire, RNApathwaysDB	Obiekty przechowują informacje o publikacjach przypisanych do danego rekordu w bazie danych.
user_comments	REPAIRtoire, RNApathwaysDB	Obiekty przechowują dane o komentarzach pozostawionych przez użytkowników baz danych. Komentarze przypisane są do wszystkich rekordów.
login	REPAIRtoire, RNApathwaysDB	Obiekty tej klasy pozwalają na tworzenie konta użytkownika i logowanie się do bazy danych.
repair_scenes	REPAIRtoire	Obiekty przechowują informację o rysunkach opisujących poszczególne stany naprawy DNA. Obiekt ten występuje jedynie w bazie danych REPAIRtoire, że względu na to, że w bazie RNApathwaysDB zostało użyte inne rozwiązanie rysowania stanów poszczególnych aktywności dojrzewania i degradacji RNA.
damages	REPAIRtoire	Obiekty tej klasy zawierają dane dotyczące uszkodzeń DNA, łącznie z ich źródłem, które jest osobnym obiektem, jak i efektem jaki wywołują. Efekt jest także obiektem tej klasy.
repairtoire_terms	REPAIRtoire	Obiekty tej klasy zawierają informację odnośnie terminologii pozwalającej na stworzenie w przyszłości systemu ontologii do opisu zdarzeń naprawy DNA.

Przykładem aplikacji w Django może być aplikacja *proteins* (białka). Aplikacja ta jest wspólna dla obu baz danych i definiuje następujące klasy: *Protein*, *ProteinFamily*, *Gene*. *Protein* przechowuje informacje o białkach: ich nazewnictwie (w tym: nazwa zwyczajowa, skróty, nazwy alternatywne), aktywnościach, lokalizacji komórkowej, strukturach, odnośnikach do innych baz danych czy do piśmiennictwa na ich temat. Klasa *ProteinFamily* to dane dotyczące rodzin białkowych, a *Gene* zbiera informacje o genach kodujących dane białka. Wszystkie klasy są ze sobą powiązane odpowiednimi relacjami: *Protein* z *ProteinFamily* relacją typu *wiele do wielu* (ang. *many to many*), *Protein* z *Gene* relacją typu *wiele do jednego* (ang. *many to one*), gdyż białka mogą należeć do kilku rodzin, a jeden gen może kodować kilka rodzajów białek. Wszystkie te klasy kodują odpowiednie tabele w bazie danych. Kod 2 jest przykładem zapisu klasy języka Python odzwierciedlającej tabelę w bazie danych. Poszczególne atrybuty klasy odpowiadają kolumnom w tabeli.

Kod 2 Klasa języka Python odzwierciedlająca tabelę *Protein* w bazie danych RNApathwaysDB. Poszczególne wiersze opisują kolumny, które pojawią się w tabeli.

```
from django.db                    import models
from rnb.species.models           import Species
from rnb.papers.models            import Paper
from rnb.keywords.models          import Keyword
from rnb.structures.models        import PdbStructure, HomologyModel

class Protein(models.Model):
    """
    Protein is a protein from a specific organism. It can be a particular
    enzymatic component of a maturation or decay pathway, or any cofactor.
    """

    name                = models.CharField(max_length=100)
    ncbi_gi_numbers     = models.TextField(blank=True, help_text = 'GI numbers separated with
commas')  # ncbi_gi_numbers = GI number! -> NCBI ids
    uniprot             = models.CharField(max_length=12, blank=True)
    scop                = models.CharField(max_length=500, help_text = 'SCOP ids seperated with
commas / max_lenght=500', blank=True) # scop ids
    cath                = models.CharField(max_length=500, help_text = 'CATH ids seperated with
commas / max_lenght=500', blank=True) # cath ids
    pfam                = models.CharField(max_length=500, help_text = 'Pfam ids seperated with
commas / max_lenght=500', blank=True) # pfam id
    interpro            = models.CharField(max_length=500, help_text = 'InterPro ids seperated
with commas / max_lenght=500', blank=True) # interpro id
    brenda              = models.CharField(max_length=12, blank=True)
    kegg                = models.CharField(max_length=14, blank=True)
    alternative_names   = models.CharField(max_length=200, blank=True)
    abbreviations       = models.CharField(max_length=300, blank=True)
    interacting_proteins = models.ManyToManyField("self", related_name='interacting_protein',
symmetrical=True, null=True, blank=True)
    species             = models.ForeignKey(Species, null=True, blank=True)
    protein_family      = models.ManyToManyField(ProteinFamily, null=True, blank=True)
    structures          = models.ManyToManyField(PdbStructure, null=True, blank=True)
    homology_models     = models.ManyToManyField(HomologyModel, null=True, blank=True)
    publications        = models.ManyToManyField(Paper, null=True, blank=True,
help_text="test")
    protein_sequence    = models.TextField(blank=True)
    gene                = models.ForeignKey(Gene, null=True, blank=True)
    description         = models.TextField(blank=True)
    internal_comments   = models.TextField(blank=True)
    last_mod            = models.DateTimeField(auto_now=True)
    keywords            = models.ManyToManyField(Keyword, null=True, blank=True)

    def __family__(self):
        return list(self.protein_family.all())
    __family__.short_description = "family"

    def __unicode__(self):
        if self.name:
            descript = unicode(self.name)
        else:
            descript = 'Unnamed protein'
        if self.species:
            descript = '%s (%s)' %(descript, self.species)
               return descript
```

4.1.4.4 Mapowanie adresów URL (*urls*)

Po stworzeniu bazy danych niezbędne jest opracowanie sposobu wyświetlania tych danych na stronach WWW. Django oferuje gotowe rozwiązanie w postaci widoków, które stanowią pomost pomiędzy modelami a szablonami. Decydują one o tym, *co* ma być wyświetlane, ale nie *jak*. Jednakże by móc wywołać konkretny szablon poprzez funkcję

widoku, która jest kontrolerem, niezbędny jest mechanizm obsługi żądania-odpowiedzi protokołu HTTP. Django oferuje prosty mechanizm odwzorowania opartych na wyrażeniach regularnych schematów adresów URL na metody widoku, które łączą adres URL żądania i wygenerowaną odpowiedź. System ten pozwala także na generowanie adresów URL na podstawie nazw i parametrów, co pozwala całkowicie rozdzielić je od funkcji i szablonów. Odwzorowania te są przechowywane w specjalnych plikach *urls.py* (pliki *URLconf*, ang. *URL configuration*) (Forcier, Bissex i wsp. 2008). Plik *urls.py* określa, który widok zostanie wywołany dla wskazanego adresu URL. Kod 3 pokazuje przykład zakodowania za pomocą pliku *urls.py* reakcji serwera na wywołanie adresu http://genesilico.pl/rnapathwaysdb/proteins/. Adres ten powoduje uruchomienie funkcji *protein_list* z pliku *views.py* aplikacji *Proteins*, która decyduje o tym co zostanie wyświetlone pod tym adresem URL.

Kod 3 Kod wywołujący reakcję aplikacji po odwiedzeniu adresu „http://genesilico.pl/rnapathwaysdb/proteins/". Dzięki niemu następuje wywołanie funkcji widoku *protein_list*, wyświetlającej tabelę ze wszystkimi białkami w bazie danych RNApathwaysDB.

Plik *urls.py* w katalogu głównym, który definiuje ogólne zasady reakcji serwera na wywołanie poszczególnych adresów URL.

```
from django.conf.urls.defaults import *
from settings import URL_BASE

urlpatterns = patterns('',
    (r'^'+URL_BASE+'proteins/', include('rnb.proteins.urls')),
)
```

Plik *urls.py* w katalogu aplikacji Proteins, który definiuje szczegółowe zasady reakcji serwera na wywołanie poszczególnych adresów URL.

```
from django.conf.urls.defaults import *
from rnb.proteins.models import Protein

urlpatterns = patterns('',
     (r'^$', 'rnb.proteins.views.protein_list'),
)
```

Pliki *urls.py* przechowują również informacje na temat generowania adresów URL. Reguły te muszą być unikalne dla każdego przypadku i muszą jednoznacznie klasyfikować dany adres URL. Zapewniają to poprawnie skonstruowane, udostępnione przez język Python, wyrażenia regularne zapisane w postaci `r'^proteins/(?P<id>\d+)/$'`. Pierwsza część wpisu to wyrażenie regularne (zaczynające się od przedrostka *'r'*), które w tym przypadku interpretuje wszystkie adresy URL zaczynające się od słowa *„proteins/"*. Dalsza

część adresu, pasująca do wzorca (dowolny ciąg liczbowy, składający się z przynajmniej jednej cyfry) jest przypisywana do zmiennej „*id*", która automatycznie przekazana zostaje do odpowiedniej funkcji widoku.

4.1.4.5 Implementacja widoków (*views*)

Widoki (kontrolery) stanowią podstawę każdej aplikacji Django, ponieważ odpowiadają one za niemal całą logikę aplikacji. Widoki są standardowymi funkcjami języka Python, których jedynym zadaniem jest pobieranie obiektu żądania i zwracanie obiektu odpowiedzi lub zgłaszanie wyjątku. Takie podejście sprzyja bezpośrednio realizacji zasady KISS. Widoki są odpowiedzialne za przydział funkcji po wywołaniu adresu URL i jednocześnie przekazanie ewentualnych danych do szablonów. Metody obsługujące widoki zapisywane są w plikach *views.py* (Forcier, Bissex i wsp. 2008). Kod 4 pokazuje przykładową funkcję kontrolera, która przesyła do szablonu *proteins_list.html* zbiór wszystkich białek znajdujących się w bazie danych.

Kod 4 Kod kontrolera, który pozwala na pozyskanie z bazy danych RNApathwaysDB informacji odnośnie wszystkich białek bazy i przesyłający ją do szablonu *proteins_list.html*.

```
from django.shortcuts import render_to_response
from django.template  import RequestContext
from rnb.proteins.models          import Protein
from rnb.pathways.models          import Reaction

def protein_list(request):
    """ displays list of proteins from a database """
        try:
        if request.POST['extended'] == 'extended': extended = True
        else: extended = False
    except KeyError:
        extended = False

    try:
        proteins = Protein.objects.all().order_by('name')
    except KeyError:
        proteins = Protein.objects.all().order_by('id')

    proteins = list(proteins)
    proteins_and_reactions = []

    for protein in proteins:
        if protein.ncbi_gi_numbers:
            gi_numbers = protein.ncbi_gi_numbers.split(',')
        else:
            gi_numbers = []

        proteins_and_reactions.append(
            {'protein'            : protein,
             'reactions'          : Reaction.objects.filter(involved_proteins=protein),
             'gi_numbers'         : gi_numbers,
            })
    return render_to_response(
        'proteins/protein_list.html',
        {
        'object_list'    : proteins_and_reactions,
        'extended'       : extended,},
        context_instance=RequestContext(request), )
```

4.1.4.6 Implementacja szablonów (*templates*)

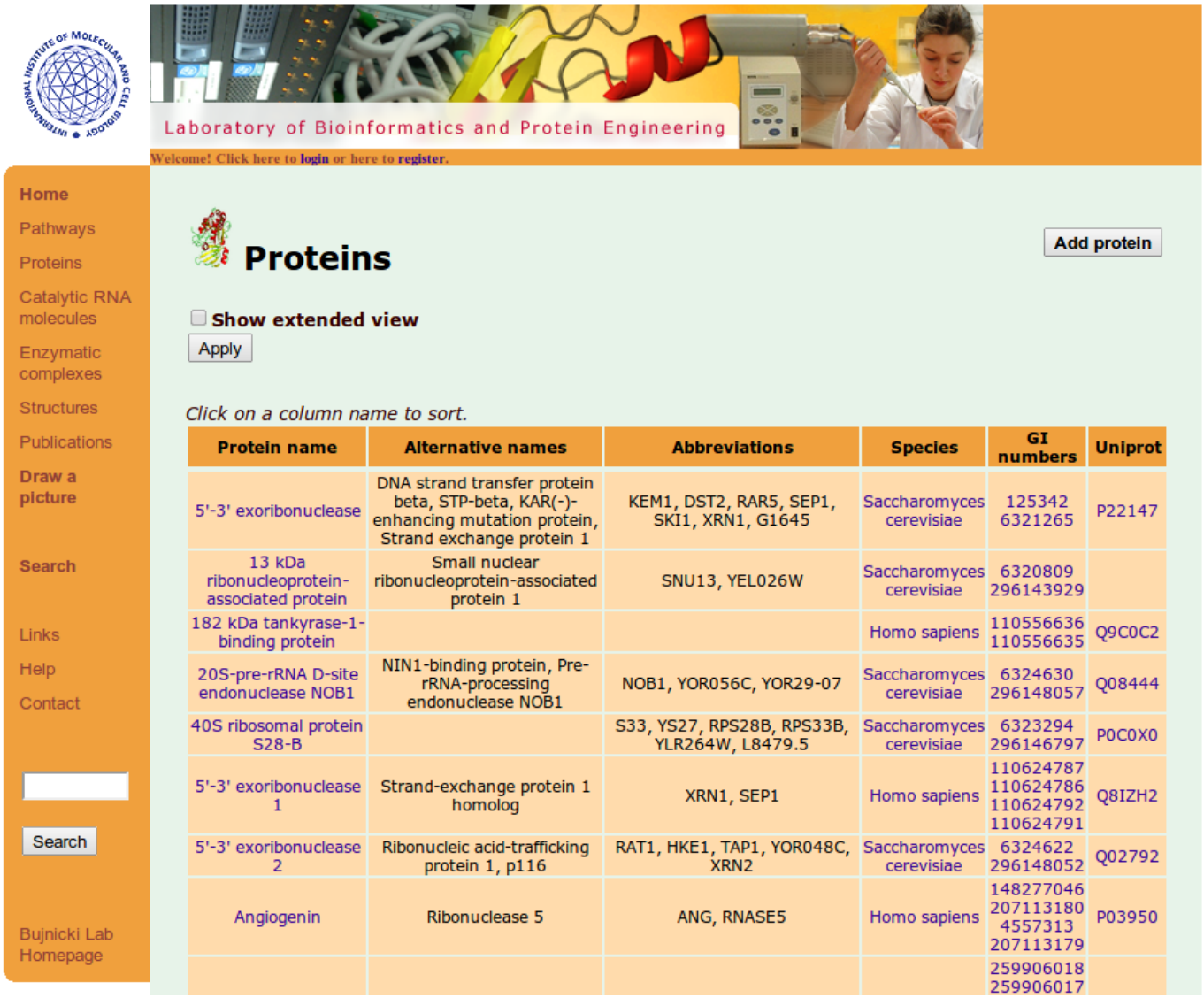
Ostatnim elementem opisanej struktury są szablony służące do wyświetlania końcowych wyników. Zbudowano je według uproszczonego schematu. Dzięki temu cała struktura jest zrozumiała zarówno dla projektantów jak i grafików oraz innych osób standardowo niezajmujących się programowaniem (zasada KISS). Jednocześnie obsługuje ona system dziedziczenia przypominający dziedziczenie klas, co minimalizuje powtórzenia kodu HTML i pomaga zastosować się do wskazówki wyrażonej akronimem DRY. Szablony są niezależnymi plikami tekstowymi, zawierającymi niezmienną treść statyczną (np. kod HTML) oraz dynamiczne znaczniki określające m. in. operacje logiczne, pętle i sposób wyświetlania danych. Wybór szablonu zastosowanego do utworzenia strony jest podejmowany w funkcji widoku lub w jej argumentach. Język szablonów Django został zaprojektowany z myślą o projektantach wyglądu strony, którzy nie muszą być programistami. Z tego względu wraz z odseparowaniem logiki od prezentacji język znaczników nie został zagnieżdżony w języku Python. Mimo to projektanci mogą korzystać z licznych znaczników i filtrów, aby stosować bardziej zaawansowane konstrukcje, wykorzystywane do tworzenia kodu HTML (Forcier, Bissex i wsp. 2008). System szablonów może także współpracować, poza kodem HTML, z innymi typami danych tekstowych, jak: e-mail, pliki CSV czy dzienniki zdarzeń. Pliki szablonów mogą mieć nadawane dowolne nazwy, ale zawsze kończą się rozszerzeniem *html*.

Wcześniej wymieniony plik *proteins_list.html* to szablon HTML opisujący wygląd strony WWW. Wykorzystuje język szablonowy z prostymi poleceniami logicznymi.

Kod 5 stanowi przykład szablonu, który decyduje o tym, w jaki sposób wyświetlana będzie lista białek znajdujących się w bazie danych. Natomiast Rycina 4 przedstawia wynik działania tego kodu.

Kod 5 Fragment kodu szablonu *proteins_list.html*, który wyświetla na stronie WWW listę wszystkich białek z bazy danych RNApathwaysDB

```
(Kushner)
(Milanowska, Mikolajczak i wsp.)
<table style="width:100%">
  <tr>
    <td>
      <h2>
        <img src="(Milanowska, Mikolajczak i wsp.)}images/proteins.png" alt=""
style="border:0px; height:50px;">
        Proteins</h2>
    </td>
    <td style="text-align:right;">
      <form action="{{MEDIA_URL}}add/protein/False">
      <button type="submit" value="Add protein" ><b>Add protein</b></button>
      </form>
    </td>
  </tr>
</table>
{% if object_list%}
<i>Click on a column name to sort.</i>
<center>
<table class="sortable">
  <thead>
  <tr>
    <th><b>Protein name</b></th>
    <th><b>Alternative names</b></th>
    <th><b>Abbreviations</b></th>
    <th><b>Species</b></th>
    <th><b>GI numbers</b></th>
    <th><b>Uniprot</b></th>
   {%if extended%}
    <th><b>Involved in reactions</b></th>
    <th><b>Families</b></th>
    <td class="results" align="center"><b>Solved structures</b></td>
   {%endif%}
  </tr>
    </thead>
    <tbody>
    {%for protein in object_list%}
      {%ifequal protein.protein.species.name 'Arabidopsis thaliana'%}
      {%else%}
   <tr>
      <td>
      (Milanowska, Mikolajczak i wsp.)
        <a href="(Milanowska, Mikolajczak i wsp.)">{{protein.protein.name}}</a>
                {%else%}
        <a href="(Milanowska, Mikolajczak i wsp.)"> Unnamed protein </a>
      {%endif%}
    </td>
(...)
    <td>
      {%for structure in protein.protein.structures.all%}
        <a href="{%url structures.views.detail structure.id%}">{{structure}}</a><br/>
      {%endfor%}
    </td>
    {%endif%}
      </tr>
   {%endifequal%}
    {%endfor %}
    </tbody>
  </table>
</center><br />
{% endif %}
{% endblock %}
```



Rycina 4 Zrzut ekranu pokazujący wynik działania kodu *protein_list.html*.

4.2 Bazy danych wykorzystane podczas zbierania danych

Tabela 5 Bazy danych wykorzystane podczas zbierania danych.

Baza danych	Adres WWW	Źródło	Opis
NCBI	http://www.ncbi.nlm.nih.gov/	(Sayers, Barrett i wsp. 2011)	Szereg baz danych związanych z biotechnologią i biomedycyną.
UniProt	http://www.uniprot.org/	(UniProt 2012)	Baza danych sekwencji białkowych (The Universal Protein Resource).
PFAM	http://pfam.sanger.ac.uk/	(Punta, Coggill i wsp. 2012)	Zbiór przyrównań wielu sekwencji i profili HMM (ang. *hidden Markov model*) reprezentujących rodziny białek bądź domeny.
InterPro	http://www.ebi.ac.uk/interpro/	(Hunter, Jones i wsp. 2012)	Baza danych białek pogrupowanych w rodziny, z informacją o domenach i miejscach aktywnych.
PDB	http://www.pdb.org	(Rose, Beran i wsp. 2011)	Baza danych struktur przestrzennych (ang. *Protein Data Bank*), w której gromadzone są współrzędne przestrzenne struktur białek, kwasów nukleinowych, a także ich kompleksów.
BRENDA	http://www.brenda-enzymes.org	(Scheer, Grote i wsp. 2011)	Najobszerniejsza kolekcja informacji o enzymach.
KEGG	http://www.genome.jp/kegg/	(Kanehisa, Araki i wsp. 2008)	Encyklopedia genów i genomów (*Kyoto Encyclopedia of Genes and Genomes*)
Reactome	http://www.reactome.org/	(Matthews, Gopinath i wsp. 2009)	Baza danych szlaków i reakcji metabolicznych, głównie ludzkich.
CATH	http://www.cathdb.info/	(Cuff, Sillitoe i wsp. 2011)	Półautomatyczna, hierarchiczna klasyfikacja domen białkowych.
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/	(Andreeva, Howorth i wsp. 2008)	Strukturalna klasyfikacja domen białkowych (ang. *Structural Classification of Proteins*) bazująca na podobieństwie strukturalnym i sekwencyjnym.
RFAM	http://rfam.sanger.ac.uk/	(Burge, Daub i wsp. 2013)	Baza danych rodzin RNA.

4.3 Wtyczki wykorzystane w bazach danych

Tabela 6 Wtyczki wykorzystane w bazach danych.

Wtyczka	Adres WWW	Źródło	Opis
JmolApplet	http://jmol.sourceforge.net/	(Herraez 2006)	Jmol to przeglądarka struktur przestrzennych napisana w języku Java, oparta na licencji wolnego oprogramowania. JmolApplet to wtyczka Jmola wykorzystywana w przeglądarkach internetowych do wizualizacji struktur przestrzennych. Wykorzystana została np. do wizualizacji struktur trójwymiarowych uszkodzeń w DNA.
VARNA	http://varna.lri.fr/	(Darty, Denise i wsp. 2009)	VARNA to wtyczka napisana w języku Java dedykowana rysowaniu struktur drugorzędowych RNA. Wtyczka ta służy również wizualizacji struktury drugorzędowej RNA na stronach internetowych.
MarvinSketch	http://www.chemaxon.com/products/marvin/	(ChemAxon 2013)	MarvinSketch to zaawansowany edytor chemiczny do rysowania struktur chemicznych czy reakcji. Został wykorzystany do narysowania struktur uszkodzeń występujących w DNA.

5 Wyniki

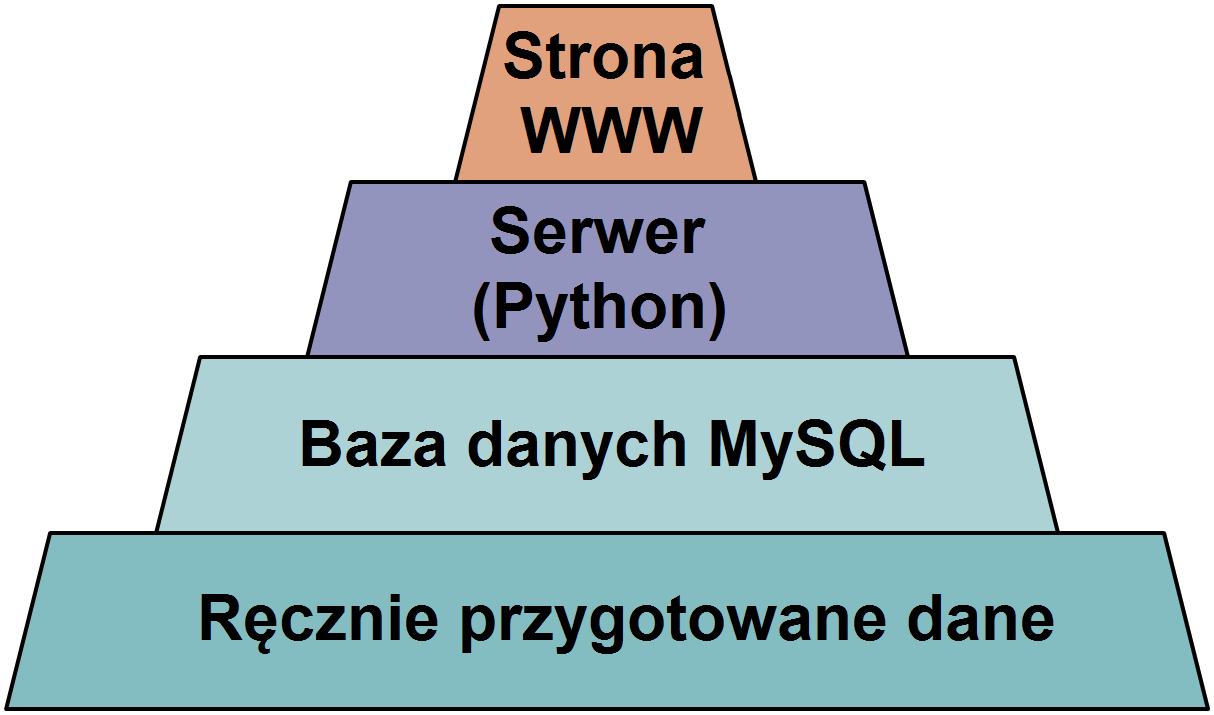
5.1 Ogólna charakterystyka stworzonych baz danych.

Bazy danych REPAIRtoire i RNApathwaysDB opierają się na wspólnym schemacie architektury, który został zaprojektowany przez autorkę niniejszej rozprawy i opisany w Rodziale 4.1.4 (patrz Podrodziały: 4.1.4.2, 4.1.4.3, 4.1.4.4, 4.1.4.5, 4.1.4.6). Pierwszą warstwę całego projektu stanowią ręcznie przygotowane dane, które zostały zebrane z dostępnych publikacji i internetowych źródeł danych, a następnie opracowane: uporządkowane i usystematyzowane. Wszystkie informacje zostały przygotowane w sposób pozwalający na ich późniejsze umieszczenie w bazach danych. Proces katalogowania danych był nakierowany na jak najpełniejsze zebranie dostępnych wiadomości i odtworzenie szlaków dojrzewania i degradacji RNA oraz naprawy DNA zgodnie z najnowszą wiedzą. W miejscach spornych, które pojawiły się w trakcie odtwarzania szlaków, jak np. w szlaku degradacji mRNA zawierających przedwczesny kodon STOP (PTC, ang. *premature termination codon*) – NMD (ang. *Nonsense-mediated decay*), w którym nieznany jest w pełni mechanizm działania białka SMG6 i różne publikacje podają sprzeczne informacje, wszystkie możliwe wersje zostały zebrane, przeanalizowane, możliwie ze sobą uzgodnione i następnie przedstawione w postaci najbardziej prawdopodobnego scenariusza.

Drugą warstwę stanowi w obu projektach baza danych, która została uzupełniona zebranymi danymi. Proces uzupełniania baz danymi w większości przeprowadzany był ręcznie, jednakże część danych została w bazach umieszczona automatycznie, ale wciąż na podstawie opracowanych wcześniej manualnie informacji.

Trzecią warstwę stanowi serwer napisany w języku Python, który stanowi połączenie pomiędzy danymi umieszczonymi w bazach a użytkownikiem. Dzięki serwerowi wszystkie dane wyświetlane są w czwartej warstwie, czyli na stronach WWW, które rozpowszechniają oba projekty i opracowane w nich informacje szerokiemu gronu naukowemu na całym świecie.

Rycina 5 przedstawia model opisanych powyżej czterech warstw, na którym opierają się obie bazy danych: REPAIRtoire oraz RNApathwaysDB.



Rycina 5 Model czterowarstwowego uporządkowania danych w bazach danych REPAIRtoire i RNApathwaysDB umożliwiający łatwą aktualizację baz, ze względu na ich ustrukturyzowany system.

5.2 REPAIRtoire - baza danych szlaków naprawy DNA

5.2.1 Ogólna charakterystyka bazy danych REPAIRtoire

Baza danych REPAIRtoire jest internetowym źródłem informacji o szlakach naprawy DNA, w obecnej wersji dane zgromadzono dla człowieka i dwóch organizmów modelowych: drożdży piekarniczych i pałeczki okrężnicy. Baza ta zawiera nie tylko informacje dotyczące samych szlaków naprawy, białek i kompleksów enzymatycznych biorących w nich udział, publikacji i zewnętrznych referencji, ale także informacje te systematyzuje, organizuje i przedstawia w przystępny sposób. Źródło to udostępnia także dane o powiązaniu ludzkich chorób z mutacjami w genach odpowiedzialnych za stabilność i integralność DNA, jak i informacje o toksycznych i mutagennych czynnikach wywołujących uszkodzenia w kwasie deoksyrybonukleinowym.

5.2.2 Zawartość bazy

Bazując na wyczerpującym przeszukiwaniu literatury zebrane zostały następujące zestawy danych:

- lista uszkodzeń DNA powiązana ze środowiskowymi czynnikami cytotoksycznymi i mutagennymi;
- szlaki naprawy, obejścia uszkodzenia, tolerancji uszkodzenia i systemy sygnalizacji uszkodzeń DNA, zawierające struktury uszkodzonego DNA i produktów pośrednich procesów naprawczych powiązanych z reakcjami katalizowanymi przez enzymy;
- białka zaangażowane w wyżej wymienione procesy;
- informacje o chorobach związanych z defektami białek naprawczych.

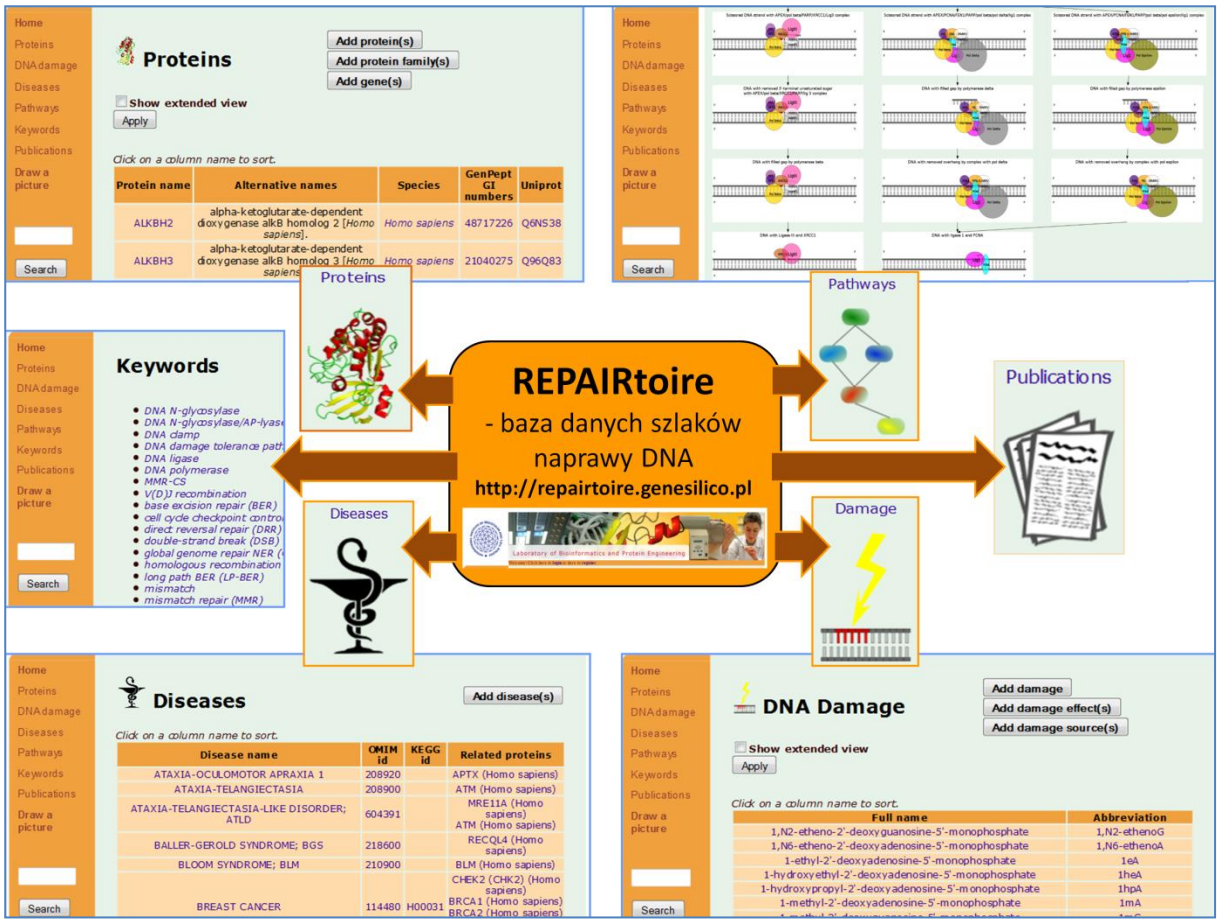
Wszystkie wymienione zestawy danych zostały przygotowane ręcznie i połączone z odpowiednimi referencjami do opublikowanych raportów doświadczalnych bądź zewnętrznych baz danych. Tabela 7 przedstawia spis typów danych oraz ich liczby w bazie.

Tabela 7 REPAIRtoire w liczbach

Aplikacja	Typ danych	Liczba rekordów w bazie
Uszkodzenia	Uszkodzenia	**87**
	Źródła uszkodzeń	**55**
	Efekty uszkodzeń	**21**
Białka	Białka	**301**
	Geny	**301**
	Choroby	**41**
Kompleksy	Kompleksy enzymatyczne	**12**
Szlaki	Szlaki	**24**
	Stany DNA	**255**
	Aktywności naprawy	**284**
Publikacje	Publikacje	**3819**
Słowa kluczowe	Słowa kluczowe	**23**
Struktury	Struktury	**79**

5.2.3 Struktura bazy danych i dostęp

REPAIRtoire to relacyjna baza danych, która łączy wcześniej wymienione zestawy danych. Dostępne są one poprzez pięć pozycji menu: *"DNA DAMAGE"* (uszkodzenia DNA), *"PATHWAYS"* (szlaki), *"PROTEINS"* (białka), *"DISEASES"* (choroby) i *"PUBLICATIONS"* (publikacje) (Rycina 6). Każde pozycja menu daje dostęp do osobnych podstron zawierających odpowiednie zestawy danych (patrz: Podrozdziały 5.2.3.1, 5.2.3.2, 5.2.3.3, 5.2.3.4, 5.2.3.5, 5.2.3.6)



Rycina 6 Schematyczne przedstawienie struktury bazy danych REPAIRtoire.

Rysunek przygotowano na podstawie publikacji (Milanowska, Krwawicz i wsp. 2011) - oryginał rysunku autorstwa dr Joanny Krwawicz.

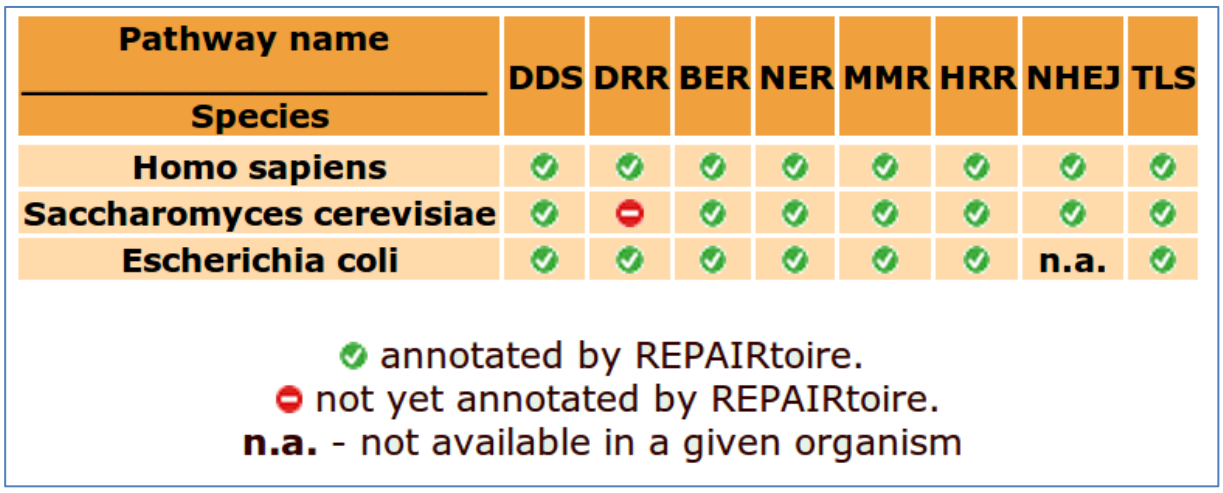
5.2.3.1 Uszkodzenia

Zebrano 87 różnych typów uszkodzeń DNA. Wiele z nich nie reprezentuje konkretnych struktur chemicznych, lecz szerzej rozumiane rodzaje struktur powodujących uszkodzenia, takie jak: pojedynczoniciowe pęknięcia czy niezależna od sekwencji utrata

zasady. 60 związków chemicznych wywołujących uszkodzenia DNA zostało połączonych z odpowiednimi typami uszkodzeń. Wśród wszystkich typów uszkodzeń, 36 zostało połączonych z pojedynczą strukturą molekularną, przykładem tu mogą być mutacje punktowe czy modyfikacje reszt nukleotydów takie jak 1-hydroksypropylo-adenina. Każdy typ uszkodzenia jest opisany na osobnej podstronie, zawierającej informacje o potencjalnym źródle uszkodzenia, (np. tworzenie się spontaniczne lub występowanie jako stan pośredni w procesie naprawy), białkach rozpoznających jego obecność, słowach kluczowych opisujących dany typ, które w przyszłości będą mogły posłużyć do stworzenia systemu ontologii, a także odnośniki do oryginalnych publikacji. Dla 36 typów uszkodzeń, które mają unikalne struktury chemiczne, REPAIRtoire udostępnia struktury pierwszorzędowe w postaci kodów SMILES, struktury drugorzędowe i trzeciorzędowe z użyciem aplikacji JMol i umożliwia ściągnięcie koordynatów przestrzennych atomów w formacie *.mol*.

5.2.3.2 Szlaki naprawy DNA

W bazie zaprezentowanych zostało 8 opisanych powyżej szlaków naprawy DNA: DDS, DDR, BER, NER, MMR, HHR, NHEJ, TLS z trzech organizmów modelowych: *Escherichia coli*, *Saccharomyces cerevisiae* i *Homo sapiens* (patrz: Rycina 7). Szlaki te prezentowane są jako grafy stworzone z użyciem wcześniej opisanego narzędzia PyGraphviz. Graf przedstawia szlaki jako układy stanów połączonych ze sobą za pomocą przejść (reakcji) (stany definiowane są jako struktura cząsteczki kwasu deoksyrybonukleinowego na danym etapie naprawy wraz ze wszystkimi czynnikami biorącymi udział w reakcjach prowadzących z i do danego stanu). Węzły są reprezentacją stanów, natomiast krawędzie, które są ukierunkowane (pokazują kierunki zachodzenia reakcji) to przejścia pomiędzy stanami. Wszystkie krawędzie (pod)grafów, czyli strzałki łączące rysunki, są powiązane ze statycznymi stronami zawierającymi szczegółowe informacje dotyczące danego etapu naprawy DNA. Wśród tyvh informacji uwzględnione są wszystkie zaangażowane w proces białka i kompleksy enzymatyczne, które również połączone są z odpowiednimi podstronami.



Pathway name / Species	DDS	DRR	BER	NER	MMR	HRR	NHEJ	TLS
Homo sapiens	✓	✓	✓	✓	✓	✓	✓	✓
Saccharomyces cerevisiae	✓	⊖	✓	✓	✓	✓	✓	✓
Escherichia coli	✓	✓	✓	✓	✓	✓	n.a.	✓

✓ annotated by REPAIRtoire.
⊖ not yet annotated by REPAIRtoire.
n.a. - not available in a given organism

Rycina 7 Spis szlaków dostępnych w bazie danych REPAIRtoire.

MMR jako przykład szlaku naprawy DNA:

Rycina 8 przedstawia przykład szlaku naprawy uwzględnionego w bazie REPAIRtoire. Jest to szlak MMR (naprawa niesparowanych zasad, ang. *mismatch repair*) u człowieka, który został opracowany w uproszczonej wersji.

Szlak ten jest odpowiedzialny za naprawę błędów rekombinacji i replikacji (niesparowania, małe insercje i delecje), które zostały pominięte podczas sprawdzania poprawności reakcji przez aktywność korekcyjną polimerazy DNA (Iyer, Pluciennik i wsp. 2006, Kunz, Saito i wsp. 2009). U człowieka proces ten rozpoczynają białka MutSα i MutLα. MutSα najpierw rozpoznaje błąd w DNA, następnie fizycznie oddziałuje z MutLα, które aktywuje białka odpowiedzialne za usunięcie fragmentu nici DNA z błędem (usunięciu ulec może od kilku do kilku tysięcy reszt) i syntetyzowanie nowej (Modrich 2006). Mutacje w genach kodujących ludzkie białka MutSα i MutLα zostały powiązane z syndromem Lyncha, który charakteryzuje się podwyższonym ryzykiem wystąpienia nowotworu (Plotz, Zeuzem i wsp. 2006).

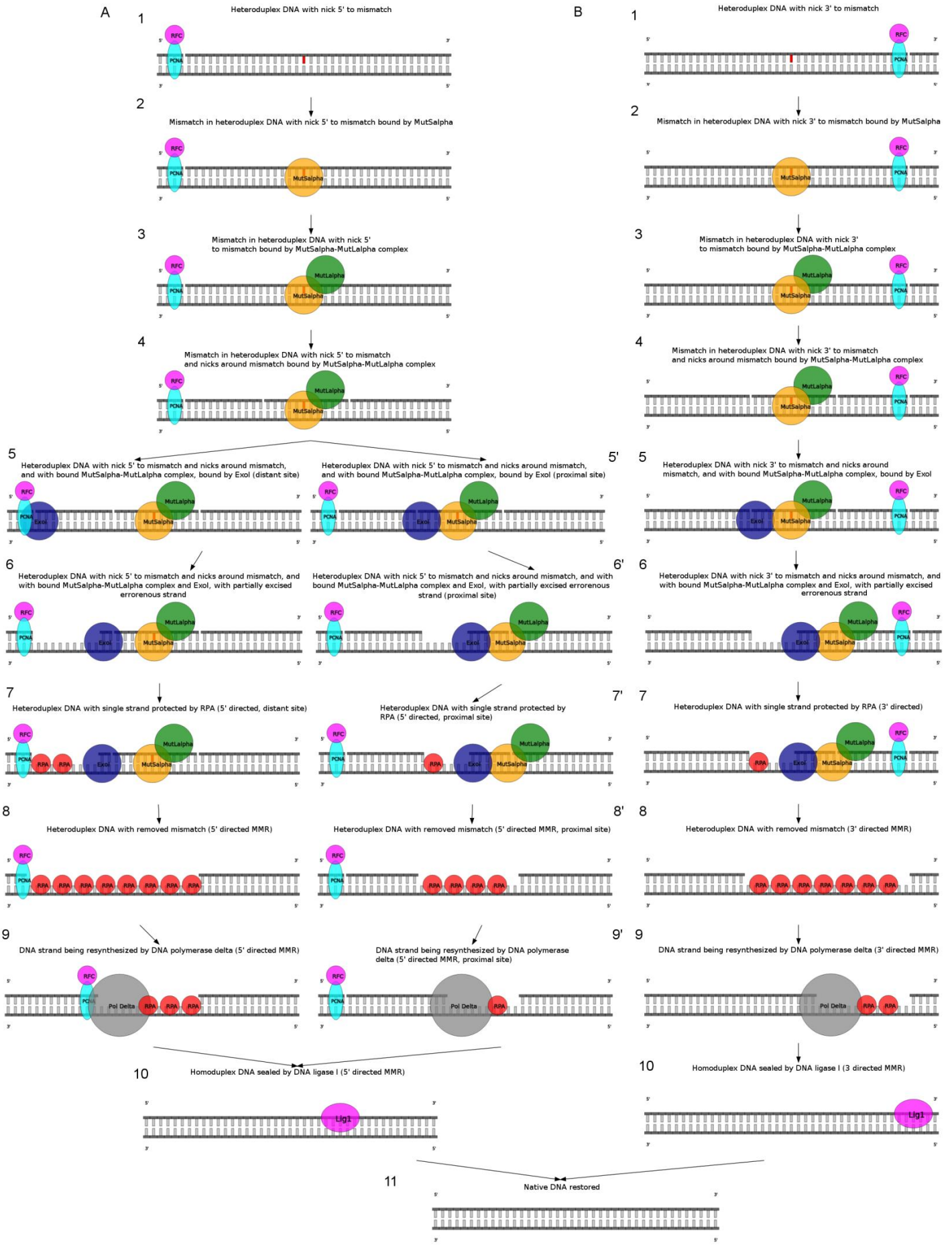
Naprawa niesparowanych zasad jest niciowo specyficzna – rozróżnia nowo syntetyzowaną nić od nici, która jest szablonem. Jednakże mechanizm tego rozróżniania u Eukaryota jest nieznany. Podejrzewa się, że nowo syntetyzowana nić opóźniona przejściowo zawiera przerwy (przed połączeniem przez ligazę DNA) i na tej podstawie jest rozpoznawana przez mechanizm sprawdzania niesparowań (Heller i Marians 2006). Według pracy z 2010 roku opublikowanej w PNAS (Pluciennik, Dzantiev i wsp. 2010) przerwy te są miejscem przyłączania zależnego od RFC białka PCNA (zwanego białkiem typu „ślizgowy zacisk pierścieniowy", ang. *sliding clamp*) w sposób określający jego orientację na nici.

W ten sposób zorientowane białko PCNA kieruje białko MutLα do jednej tylko nici, w której znajduje się niesparowanie.

W pracy rozróżnione zostały dwa mechanizmy naprawy niesparowań: A) mechanizm wybierany gdy przerwa znajduje się na końcu 5′ od niesparowania, B) mechanizm wybierany gdy przerwa znajduje się na końcu 3′ od niesparowania. Pierwsze cztery etapy dla obu tych mechanizmów są takie same (patrz: Rycina 8, etapy A1, A2, A3, A4, B1, B2, B3, B4). Najpierw, jak opisano powyżej, przyłącza się zależne od RFC białko PCNA, które kieruje heterodimer MutSα (składający się z białek Msh2 i Msh6) w miejsce uszkodzenia. Przyłączenie MutSα indukuje zgięcie DNA, co wywołuje zmianę konformacyjną białka, wymianę ADP na ATP i umożliwia przyłączenie kompleksu MutLα, składającego się z białek Mlh1 i Pms2. Kompleks MutSα-MutLα wprowadza przerwy po obu stronach uszkodzenia w nici syntetyzowanej dzięki aktywności nukleazowej MutLα. Następujące później zdarzenia różnią się od siebie w zależności od tego, po której stronie uszkodzenia występowała pierwotna przerwa w powstającej nici. Egzonukleaza I (Exo I) przyłącza się albo do kompleksu MutSα –MutLα (patrz: Rycina 8, etapy A5', B5) – tzw. miejsce bliższe (ang. *proximal site*), albo do białka PCNA (patrz: Rycina 8, etap A5) – tzw. miejsce dalsze (ang. *distant site*). W pierwszym przypadku Exo I wycina reszty nukleotydowe w pobliżu uszkodzenia, łącznie z uszkodzeniem i nukleotydami poza nim. Do powstającego jednoniciowego fragmentu przyłączają się białka RPA, które stabilizują jednoniciowy fragment DNA. Po wycięciu odpowiedniego fragmentu kompleks MutSα-MutLα-Exo I odłącza się, a do kompleksu DNA z białkami RPA przyłącza się polimeraza Delta, która resyntetyzuje wycięty fragment (patrz: Rycina 8, etapy A6', A7', A8', A9', B6, B7, B8, B9). W drugim przypadku Exo I wycina fragment od pierwotnej przerwy na końcu 5′ do przerwy wprowadzonej przez MutLα w pobliżu niesparowania. Również przyłączają się w tym czasie białka RPA stabilizujące nić DNA. Exo I wycina niesparowaną resztę i reszty otaczające. W tym przypadku usunięciu ulega znacznie dłuższy fragment niż w przypadku pierwszym. Po odłączeniu się kompleksu odpowiedzialnego za wycinanie polimeraza Delta resyntetyzuje nić DNA (patrz: Rycina 8, etapy A6, A7, A8, A9). Następnie w obu przypadkach dosyntetyzowana nić zostaje połączona ze starą nicią za pomocą ligazy I, by odtworzyć natywną cząsteczkę DNA (patrz: Rycina 8, etapy A10, A11, B10).

Opracowana wersja szlaku nie uwzględnia cyklów wymiany i hydrolizy ATP/ADP przez białka MutSα i MutLα oraz pasywnego przesuwania kompleksu MutSα-MutLα z miejsca niesparowania. Istnieją alternatywne modele wydarzeń, które następują

po przyłączeniu MutSα i/lub MutLα do niesparowania, w tym: model przełącznika molekularnego (ang. *molecular switch model*) – według którego występuje pasywne przesuwanie kompleksu; model aktywnej translokacji (ang. *active-translocation model*) i model wyginania DNA (ang. *DNA bending model*), w którym kompleks MutSα-MutLα pozostaje przy niesparowaniu. Najbardziej prawdopodobny jest model molekularnego przełącznika, jednakże jest on ciągle spekulatywny, dlatego nie został uwzględniony.



Rycina 8 Szlak naprawy niesparowanych zasad (MMR, ang. *mismatch repair*). A) pierwotna przerwa pozwalająca rozpoznać białku PCNA nić z niesparowaniem znajduje się na końcu 5′ od niesparowania, B) przerwa znajduje się na końcu 3′ od niesparowania. Oznaczenia białek: RFC – kolor różowy, PCNA – kolor niebieski, MutSα – kolor żółty, MutLα – kolor zielony, Exo I – kolor granatowy, RPA – kolor czerwony, Polimeraza Delta – kolor szary, Lig1 – kolor różowy.

Rysunek stworzony na podstawie oryginalnego rysunku z bazy danych REPAIRtoire.

5.2.3.3 Białka:

REPAIRtoire przechowuje obecnie informacje odnośnie odpowiednio 69, 78 i 154 białek i ich genów z *E. coli, S. cerevisiae and H. sapiens*. Każde białko może być wyświetlone albo z poziomu menu albo z poziomu szlaku, w którym bierze udział. Podczas procesu ręcznego zbierania danych uwzględnione zostały takie dane jak alternatywne i skrótowe nazwy genów i białek, sekwencje nukleotydowe i aminokwasowe pobrane z baz danych NCBI, struktury trzeciorzędowe z PDB. W opisach zamieszczono także informację o funkcji, lokalizacji komórkowej, specyficzności tkankowej i enzymatycznej, informację odnośnie oddziaływań, tworzenia kompleksów z innymi cząsteczkami czy obecności izoform, a także powiązania z chorobami. Całość wzbogacona jest w odnośniki do innych baz danych, w tym: PFAM, InterPro, UniProt, KEGG i BRENDA (patrz: Rycina 9). Obecnie w bazie znajdują się jedynie białka z *E. coli*, *S. cerevisiae*, i *H. sapiens*, ale w przyszłości zostanie ona rozszerzona i zawierać będzie także wszystkie ortologi scharakteryzowanych pod względem funkcji enzymów możliwe do zidentyfikowania w pełni zsekwencjonowanych genomach.

Home
Proteins
DNA damage
Diseases
Pathways
Keywords
Publications
Draw a picture

Search

Links
Help
Contact

Genesilico
Homepage

RPA1

Edit protein

Protein FULL name:

Replication protein A 70 kDa DNA-binding subunit, Replication factor A protein 1, Single-stranded DNA-binding protein.,

Protein SHORT name:

RP-A p70 RF-A 1

RPA1 (*Homo sapiens*) is product of expression of RPA1 gene.

RPA1 is involved in:

NER in Homo sapiens

- XPC complex dissotiates in Homo sapiens
- XPA recruits XPF/ERCC1 complex in Homo sapiens
- RPA binds to undamaged strand in Homo sapiens
- XPG joins the complex in Homo sapiens
- recruitment of XPF/ERCC1 complex in Homo sapiens
- RNAP dissociates in Homo sapiens
- ubiquitination of CSB in Homo sapiens
- Polymerase Delta fills the gap in Homo sapiens
- Polymerase Epsilon fills the gap in Homo sapiens

FUNCTION: Plays an essential role in several cellular processes in DNA metabolism including replication, recombination and DNA repair. Binds and subsequently stabilizes single-stranded DNA intermediates and thus prevents complementary DNA from reannealing.

SUBUNIT: Heterotrimer composed of RPA1, RPA2 and RPA3. The DNA- binding activity may reside exclusively on the RPA1 subunit. Interacts with RIP and XPA. Interacts with RPA4. Interacts with the polymerase alpha subunit POLA1/p180; this interaction stabilizes the replicative complex and reduces the misincorporation rate of DNA polymerase alpha by acting as a fidelity clamp.

INTERACTION: P54132:BLM; NbExp=1; IntAct=EBI-621389, EBI-621372; P04637:TP53; NbExp=1; IntAct=EBI-621389, EBI-366083; Q14191:WRN; NbExp=4; IntAct=EBI-621389, EBI-368417;

SUBCELLULAR LOCATION: Nucleus.

PTM: Phosphorylated upon DNA damage, probably by ATM or ATR.

SIMILARITY: Belongs to the replication factor A protein 1 family.

SEQUENCE CAUTION: Sequence=BAD92969.1; Type=Erroneous initiation;

WEB RESOURCE: Name=NIEHS-SNPs; [LINK]

NCBI GenPept GI number(s): 1350579, 4506583

Species: *Homo sapiens*

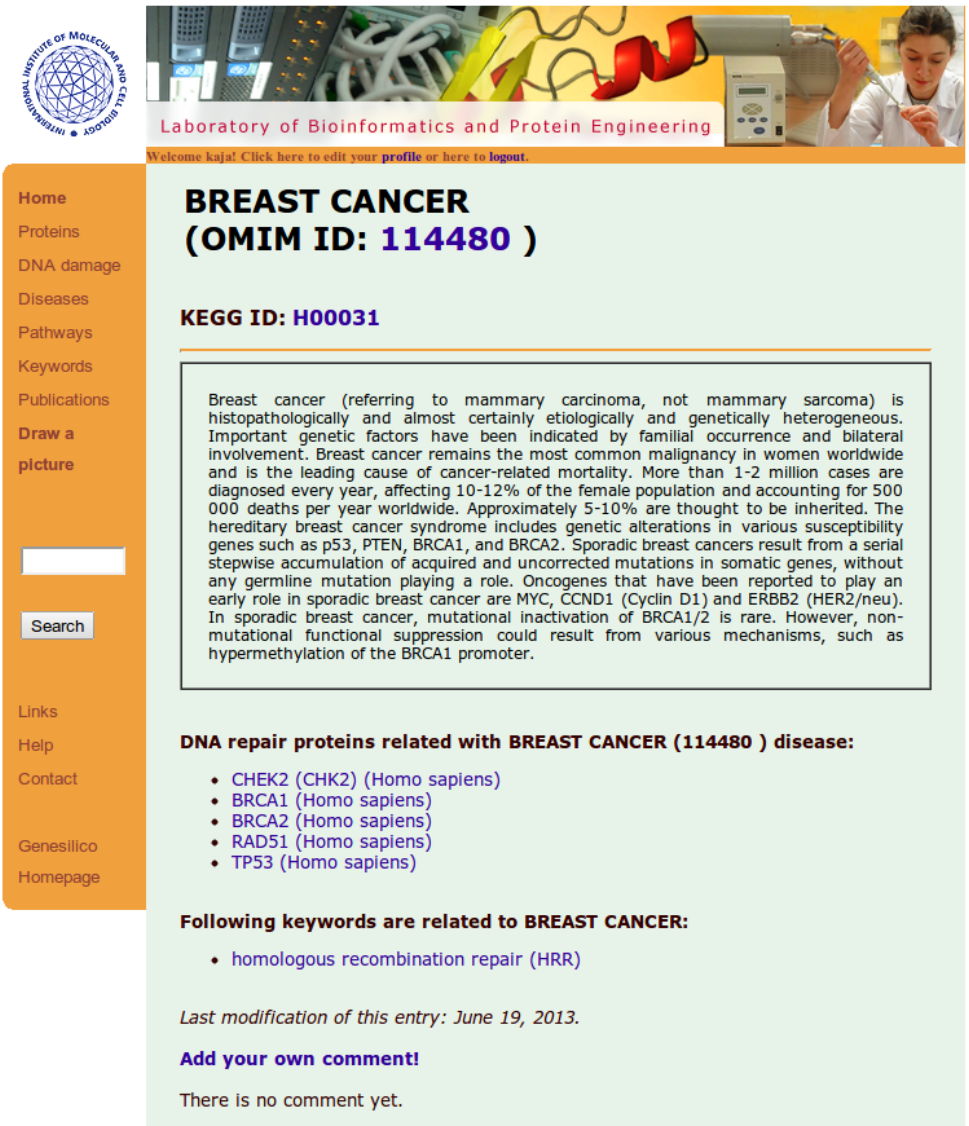
Links to other databases:

Database	ID	Link
Uniprot	P27694	P27694
PFAM:	PF04057 PF08646 PF01336	PF04057 PF08646 PF01336
InterPro:	IPR012340 IPR016027 IPR004365 IPR007199 IPR013955 IPR004591	IPR012340 IPR016027 IPR004365 IPR007199 IPR013955 IPR004591
CATH:	-	-
SCOP:	-	-
PDB:	-	-

Rycina 9 Przykład wyświetlania informacji o białku w bazie danych REPAIRtoire.

5.2.3.4 Choroby:

Obecnie w bazie zebranych jest 41 chorób wywoływanych przez mutacje w 32 genach, które powodują defekty w białkach naprawczych. Zbiór ten przedstawiony jest jako tabela z linkami do odpowiednich białek i do wpisów w bazie OMIM. Każda choroba ma swoją podstronę (patrz: Rycina 10) ze zwięzłym opisem i dodatkowymi odnośnikami, np. do słów kluczowych. Dostępne są także zwrotne odsyłacze do chorób na podstronach białek.



Rycina 10 Zrzut ekranu przedstawiający podstronę dla choroby w bazie danych REPAIRtoire.

5.2.3.5 Słowa kluczowe:

To menu umożliwia szybki dostęp do słów kluczowych przypisanych do rekordów bazy. Słowa kluczowe umożliwiają scharakteryzowanie i zaklasyfikowanie danego wpisu w bazie danych do odpowiedniej klasy pojęć. Przykładem tu może być uczestnictwo w określonym procesie lub przynależność do danej rodziny białek: odpowiedź na uszkodzenie DNA (ang. *the response to DNA damage*), punkt kontrolny cyklu komórkowego (ang. *the cell cycle checkpoint control*), N-glikozylaza DNA (ang. *the DNA N-glycosylase*), N-glikozylazo-AP-liaza DNA (ang. *the DNA_N-glycosylase/AP-lyase*) i wiele innych.

5.2.3.6 Publikacje:

Wszystkie rekordy w bazie posiadają odnośniki literaturowe do oryginalnych publikacji związanych z danym procesem, cząsteczką czy strukturą. Odsyłacze prowadzą do bazy danych PubMed. Obecnie w bazie znajduje się 3819 pozycji literaturowych.

5.2.4 Implementacja

Baza REPAIRtoire została zaimplementowany z użyciem Django (patrz: Podrozdział 4.1.4), a relacyjna baza danych przechowująca wszystkie informacje oparta jest o język SQlite3 (patrz: Podrozdział 4.1.3.2). Wśród dostępnych w serwisie opcji administracyjnych znajdują się: profil i logowanie, narzędzie do rysowania, edytowalne strony i narzędzie do wyszukiwania.

5.2.4.1 Profil, logowanie i edytowalne strony

Aby skorzystać z bazy danych rejestracja nie jest wymagana. Jest to opcja pozwalająca na dostęp do możliwości edycji zawartości bazy. Stworzenie profilu użytkownika i zalogowanie się do bazy danych pozwala na dostanie się zarówno do panelu administracyjnego jak i edytowalnych stron przedstawionych w postaci stron Wiki (ang. *Wiki-like pages*). Opcja ta jest dostępna dla współpracowników i użytkowników z prawami administracyjnymi, którzy są zainteresowani nie tylko dostępem do samych danych, ale także możliwością edytowania zawartości bazy. Poprzez te opcje można dodawać nowe dane (panel administracyjny), usuwać informacje (panel administracyjny) czy poprawiać błędy (panel administracyjny i strony edytowalne). Możliwe jest także komentowanie rekordów bazy i dodawanie nowych odnośników literaturowych.

5.2.4.2 Narzędzie do rysowania

Narzędzie od rysowania to dostępny przez osobną pozycję menu jako podstrona system służący do rysowania stanów w szlakach naprawy DNA, dzięki któremu stany przedstawione są jako kompleksy DNA-białka, w których białka reprezentowane są w postaci elipsoidalnych „modeli ziemniaczkowych" (ang. *potato models*) (patrz: Rycina 13). Narzędzie to posłużyło do narysowania wszystkich ilustracji zawartych w bazie REPAIRtoire, jednakże może ono służyć również jako niezależne narzędzie do tworzenia rysunków kompleksów białko-białko czy białko-DNA. System ten używa formatu grafiki wektorowej SVG (ang. *scalable vector graphics*), który umożliwia zmiany wymiarów rysunku bez utraty jakości, edycję tekstową rysunku, czy modyfikacje za pomocą zewnętrznych programów służących do obsługi grafiki wektorowej, np. Inkscape (http://inkscape.org).

Narzędzie do rysowania wykonane zostało przez mgr Grzegorza Papaja (Międzynarodowy Instytut Biologii Molekularnej i Komórkowej w Warszawie). Opis programu umieszczono w sekcji Wyniki, gdyż na jego podstawie powstawło narzędzie DrawBioPath użyte w bazie danych RNApathwaysDB (patrz: 5.3.4.1).

5.2.4.3 Narzędzie do wyszukiwania.

Baza danych REPAIRtoire może być przeszukiwana przy użyciu prostego mechanizmu wyszukiwania tekstowego, który dostępny jest z poziomu głównego menu. Narzędzie to zwraca wyniki w postaci ustrukturalizowanej listy rekordów bazy, które zawierają zapytanie (ang. *query*), np. „rak" (ang. *"cancer"*), "polimeraza RNA" (ang. *"RNA polymerase"*), „sieciowanie" (ang. *"crosslink"*), "adenina" (ang. *"adenine"),* imię i nazwisko autora publikacji, i tym podobne. W przyszłości do bazy dodane będą opcje filtrowania i opcje wyszukiwania zaawansowanego, by móc sortować wyniki w zależności od stopnia dopasowania zapytania do wyniku.

5.2.5 Dostępność

Zawartość bazy danych i narzędzie do tworzenia własnych, niestandardowych rysunków kompleksów DNA-białko są dostępne darmowo pod adresem http://repairtoire.genesilico.pl.

5.3 RNApathwaysDB – baza danych dojrzewania i degradacji RNA.

5.3.1 Ogólna charakterystyka bazy danych RNApathwaysDB

RNApathwaysDB to internetowe źródło informacji obejmujące szlaki dojrzewania i degradacji RNA. Obecnie w bazie znaleźć można informacje dotyczące: mRNA, tRNA i rRNA z dla człowieka i dwóch organizmów modelowych: drożdży piekarniczych i pałeczki okrężnicy. Dane te są obecnie (stan na dzień 7.06.2013) uzupełniane przez Adama Ustaszewskiego o szlaki dojrzewania i degradacji niekodujących RNA u człowieka. Baza ta, tak samo jak REPAIRtoire, nie tylko zbiera informacje dotyczące samych szlaków dojrzewania i degradacji, białek i kompleksów enzymatycznych biorących w nich udział, publikacji i zewnętrznych referencji, ale także informacje te systematyzuje, organizuje i przedstawia w prosty sposób.

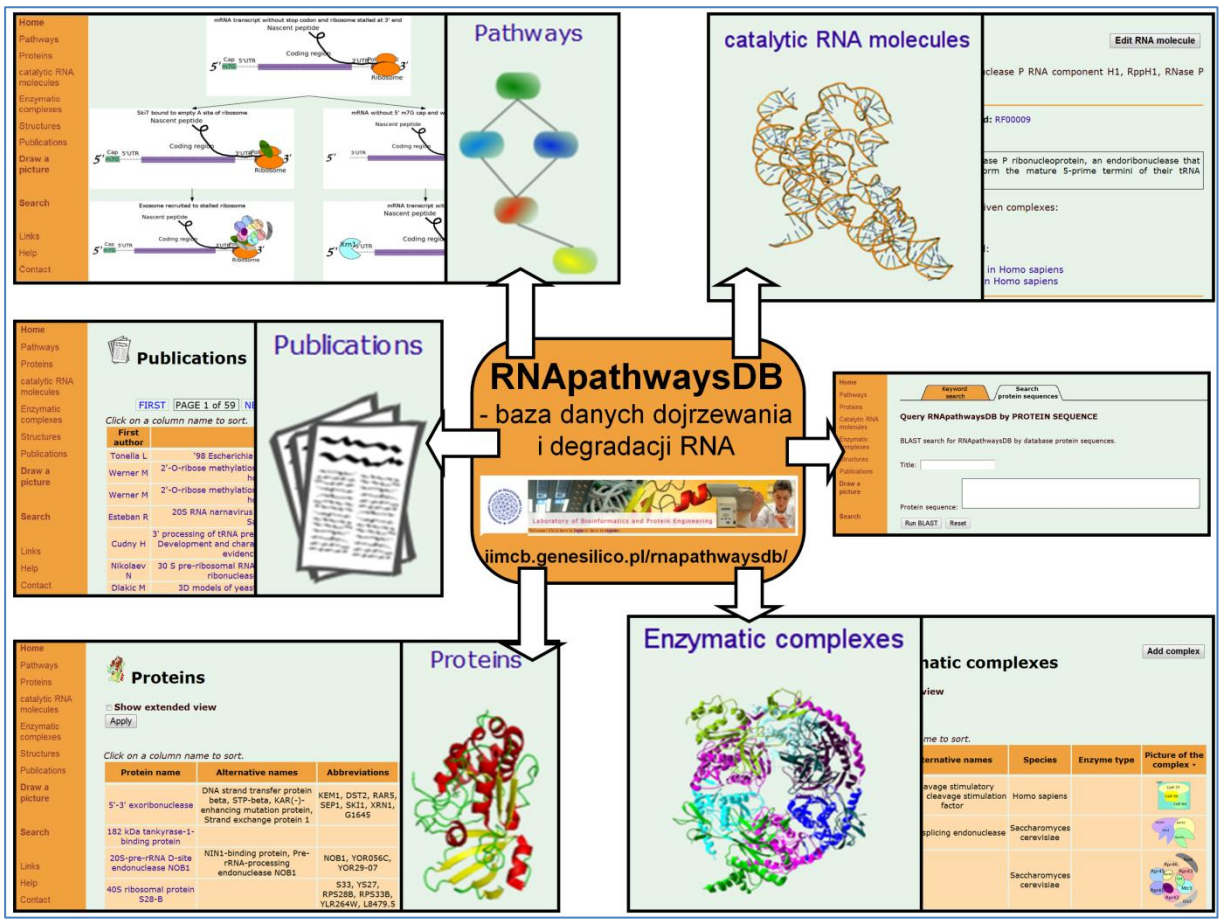
5.3.2 Zawartość bazy danych.

Na podstawie wyczerpującego przeszukiwania literatury i innych baz danych w bazie zostały zebrane następujące zestawy danych:

1. Szlaki dojrzewania i degradacji mRNA, tRNA i rRNA składające się ze stanów - RNA i produktów pośrednich procesów połączonych ze sobą za pomocą poszczególnych transformacji, takich jak reakcje katalizowane przez enzymy, wiązanie bądź uwalnianie poszczególnych czynników, bądź dobrze zdefiniowane, ważne funkcjonalnie zmiany konformacyjne.
2. Białka, kompleksy enzymatyczne oraz cząsteczki RNA posiadające właściwości katalityczne wraz ze znanymi strukturami trzeciorzędowymi, które są zaangażowane w wyżej wymienione procesy.
3. Odnośniki literaturowe do oryginalnych publikacji.
4. Odnośniki do zewnętrznych baz danych takich jak: KEGG lub Reactome w przypadku szlaków; UniProt, NCBI, BRENDA, PFAM i InterPro w przypadku białek; RFAM dla cząsteczek RNA oraz PDB dla struktur makromolekularnych.

5.3.3 Organizacja i dostęp.

Baza danych RNApathwaysDB to relacyjna baza danych, która łączy ze sobą zestawy danych wymienione powyżej, do których dostęp umożliwiają następujące sekcje menu głównego: *„PATHWAYS”* (szlaki), *„PROTEINS”* (białka), *„CATALYTIC RNA MOLECULES”* (cząsteczki RNA o właściwościach katalitycznych), *„ENZYMATIC COMPLEXES”* (kompleksy enzymatyczne), *„STRUCTURES”* (struktury) i *„PUBLICATIONS”* (publikacje) (patrz: Rycina 11). Tabela 8 przedstawia spis typów danych oraz ich liczby w bazie.



Rycina 11 Schematyczne przedstawienie struktury bazy danych RNApathwaysDB.

Rysunek przygotowano na podstawie publikacji (Milanowska, Mikolajczak i wsp. 2013); rysunek powstał w oparciu o grafikę autorstwa dr Joanny Krwawicz zaprezentowaną w publikacji opisującej bazę danych REPAIRtoire (Milanowska, Krwawicz i wsp. 2011).

Tabela 8 RNApathwaysDB w liczbach

Aplikacja	Typ danych	Liczba rekordów w bazie
Cząsteczki RNA	Katalityczne cząsteczki RNA	**9**
	Typy RNA	**13**
	Rodziny	**7**
Białka	Białka	**272**
	Geny (białka i cząsteczki RNA)	**278**
	Rodziny	**45**
Kompleksy	Kompleksy enzymatyczne	**42**
Szlaki	Szlaki	**33**
	Stany RNA	**319**
	Aktywności naprawy	**322**
Publikacje	Publikacje	**1635**
Słowa kluczowe	Słowa kluczowe	**8**
Struktury	Struktury	**241**

5.3.3.1 Szlaki:

Dostęp do informacji odnośnie szlaków dojrzewania (od pierwotnego transkryptu do formy dojrzałej) i degradacji RNA (rozkład zarówno formy dojrzałej jak i form pośrednich posiadających błędy) z wcześniej wymienionych organizmów umożliwia pozycja menu *"PATHWAYS"*. Tabela 9 przedstawia 33 (stan na dzień 7.06.2013) opracowane szlaki z trzech organizmów: człowieka, drożdży i pałeczki okrężnicy. Wszystkie szlaki wizualizowane są z użyciem narzędzia PyGraphviz. Stany układu (kompleksy cząsteczek RNA z białkami) reprezentowane są jako węzły grafu, podczas gdy przejścia pomiędzy nimi przedstawione są jako krawędzie grafu, czyli ukierunkowane strzałki łączące poszczególne stany ze sobą. Każdy węzeł i każda krawędź podłączone są do odpowiednich podstron, zawierających podstawowe informacje o każdym etapie procesowania RNA i reakcjach pomiędzy poszczególnymi etapami. Na podstronach znajdują się rysunki i dane dotyczące białek, kompleksów enzymatycznych i cząsteczek RNA biorących udział w danym procesie. Wszystkie informacje wzbogacone są o odsyłacze do oryginalnych publikacji. Z dniem 7.06.2013 w bazie znajduje się 319 stanów układu i 322 reakcje.

Tabela 9 Szlaki dojrzewania i degradacji RNA zebrane w bazie danych RNApathwaysDB.

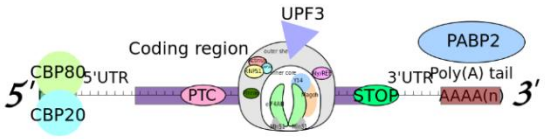
Nazwa	Organizm
Szlaki metaboliczne mRNA	
degradacja mRNA	*Escherichia coli*
degradacja mRNA zależna od deadenylacji	*Homo sapiens*
degradacja NMD	*Homo sapiens*
degradacja NSD	*Homo sapiens*
dojrzewanie końca 3′ histonowego mRNA	*Homo sapiens*
dodawanie czapeczki mRNA	*Homo sapiens*
poliadenylacja mRNA	*Homo sapiens*
wycinanie intronów z mRNA – główny szlak	*Homo sapiens*
wycinanie intronów z mRNA – szlak poboczny	*Homo sapiens*
degradacja mRNA zależna od deadenylacji	*Saccharomyces cerevisiae*
degradacja mRNA niezależna od deadenylacji	*Saccharomyces cerevisiae*
degradacja mRNA prowadzona przez endonukleazy	*Saccharomyces cerevisiae*
degradacja NGD	*Saccharomyces cerevisiae*
degradacja NMD	*Saccharomyces cerevisiae*
degradacja NSD	*Saccharomyces cerevisiae*
dodawanie czapeczki mRNA	*Saccharomyces cerevisiae*
Szlaki metaboliczne rRNA	
degradacja rRNA prowadzona przez RNazę I	*Escherichia coli*
degradacja rRNA w odpowiedzi na stres	*Escherichia coli*
dojrzewanie rRNA	*Escherichia coli*
kontrola jakości rRNA	*Escherichia coli*
dojrzewanie rRNA	*Homo sapiens*
degradacja 18S NRD	*Saccharomyces cerevisiae*
degradacja 25S NRD	*Saccharomyces cerevisiae*
dojrzewanie rRNA	*Saccharomyces cerevisiae*
Szlaki metaboliczne tRNA	
dojrzewanie tRNA	*Escherichia coli*
dojrzewanie tRNA	*Homo sapiens*
dojrzewanie tRNA	*Saccharomyces cerevisiae*
degradacja tRNA typu RTD	*Saccharomyces cerevisiae*
szlak sprawdzania poprawności tRNA	*Saccharomyces cerevisiae*
Szlaki metaboliczne małych RNA pochodzących z tRNA	
biogeneza tRFs (fragmenty RNA pochodzące z tRNA)	*Homo sapiens*
Szlaki metaboliczne w odpowiedzi na stres małych RNA pochodzących z tRNA	
biogeneza tiRNA (RNA tworzone z tRNA w odpowiedzi na stres)	*Homo sapiens*
biogeneza tiRNA (RNA tworzone z tRNA w odpowiedzi na stres)	*Saccharomyces cerevisiae*

NMD jako przykład szlaku degradacji mRNA:

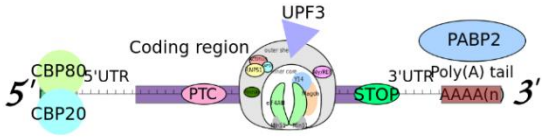
Rycina 12 przedstawia przykład takiego szlaku. Jest to szlak NMD u człowieka. Podczas tego szlaku dochodzi do wykrywania i eliminacji transkryptów, które posiadają przedwczesne kodony terminacji (PTC, ang. *premature termination codon*), które mogą być wynikiem np. mutacji czy błędów inicjacji translacji. U człowieka transkrypty takie posiadają przyłączony kompleks EJC (ang. *exon junction complex*), który jest pozostałością po wycinaniu intronów, umieszczoną 20-24 reszty nukleotydowe powyżej każdego miejsca łączenia egzonów. W warunkach naturalnych kompleksy EJC usuwane są poprzez przesuwający się rybosom, jednakże w mRNA, które posiadają PTC, kompleks EJC umieszczony jest nieprawidłowo poniżej PTC, gdzie jest wykrywany przez maszynerię naprawy błędów. Transkrypt posiadający PTC wraz z kompleksem EJC transportowany jest do cytoplazmy, gdzie przyłącza się rybosom rozpoczynający translację. Przed transportem do cytoplazmy do kompleksu EJC przyłącza się białko UPF3, będące „mostem" pomiędzy transkryptem a białkami maszynerii NMD. W normalnej terminacji translacji kodon STOP rozpoznawany jest przez czynniki uwalniające eRF1 i eRF3, które powodują odłączenie się tRNA. Gdy pojawia się PTC uwolnienie peptydu zostaje opóźnione, w związku z czym rybosom zatrzymuje się zablokowany w pobliżu kodonu STOP. W tym czasie zaczynają się tworzyć kompleksy związane z NMD. Wpierw do kompleksu EJC, oddziałując z UPF3, przyłącza się białko UPF2, następnie formuje się kompleks SMG1C, w skład którego wchodzi kinaza SMG1, odpowiedzialna za fosforylację helikazy UPF1, która oddziałując z czynnikami uwalniającymi tworzy kompleks SURF. Dzięki oddziaływaniu UPF1 z UPF2 tworzy się połączenie pomiędzy rybosomem a kompleksem EJC, co prowadzi do fosforylacji białka UPF1 (głównie w motywach [S/T]Q) przez kinazę SMG1 i uwolnienia czynników eRF. Ufosforylowane białko UPF1 służy jako platforma rekrutująca dla białek SMG5-SMG7 w NMD zależnym od SMG5-7 (ang. *SMG5-SMG7 mediated NMD*), bądź dla białka SMG6 w NMD zależnym od SMG6 (ang. *SMG6 mediated NMD*). Białka te są odpowiedzialne za uruchomienie maszynerii degradacji mRNA. Nie do końca jasne jest jak przebiegają następne etapy degradacji. W szlaku zależnym od SMG5-SMG7 następuje defosforylacja białka UPF1, co może prowadzić albo do usunięcia czapeczki i degradacji w kierunku 5′-3′ przez egzorybonukleazę XRN1, bądź do deadenylacji i degradacji w kierunku 3′-5′ przez egzosom. Nie jest jasne, który mechanizm jest preferowany i dlaczego. Białko SMG6 posiadające domenę PIN, która ma aktywność endonukleazową, tnie transkrypt pomiędzy PTC a kompleksem EJC. Fragment posiadający czapeczkę degradowany jest w kierunku

3′-5′ przez egzosom, natomiast fragment z ogonem poliA, po wcześniejszej defosforylacji UPF1 i odłączeniu kompleksów SURF i SMG1C, ulega degradacji w kierunku 5′-3′ przez egzorybonukleazę XRN1.

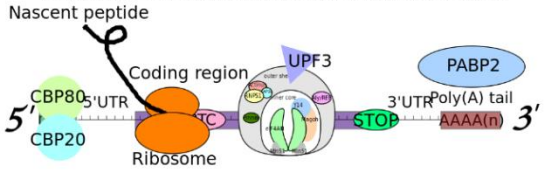
mRNA Complex with Premature Termination Codon Preceding Exon Junction



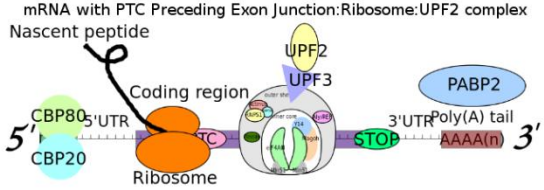
mRNA Complex with Premature Termination Codon Preceding Exon Junction



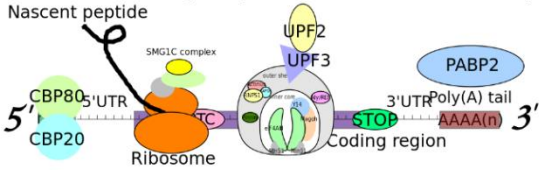
mRNA with PTC Preceding Exon Junction bound to Ribosome



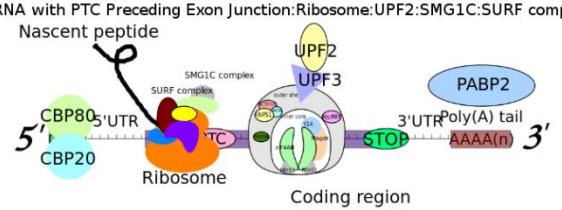
mRNA with PTC Preceding Exon Junction:Ribosome:UPF2 complex



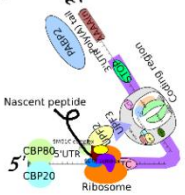
mRNA with PTC Preceding Exon Junction:Ribosome:UPF2:SMGC1 complex



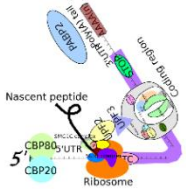
mRNA with PTC Preceding Exon Junction:Ribosome:UPF2:SMG1C:SURF complex



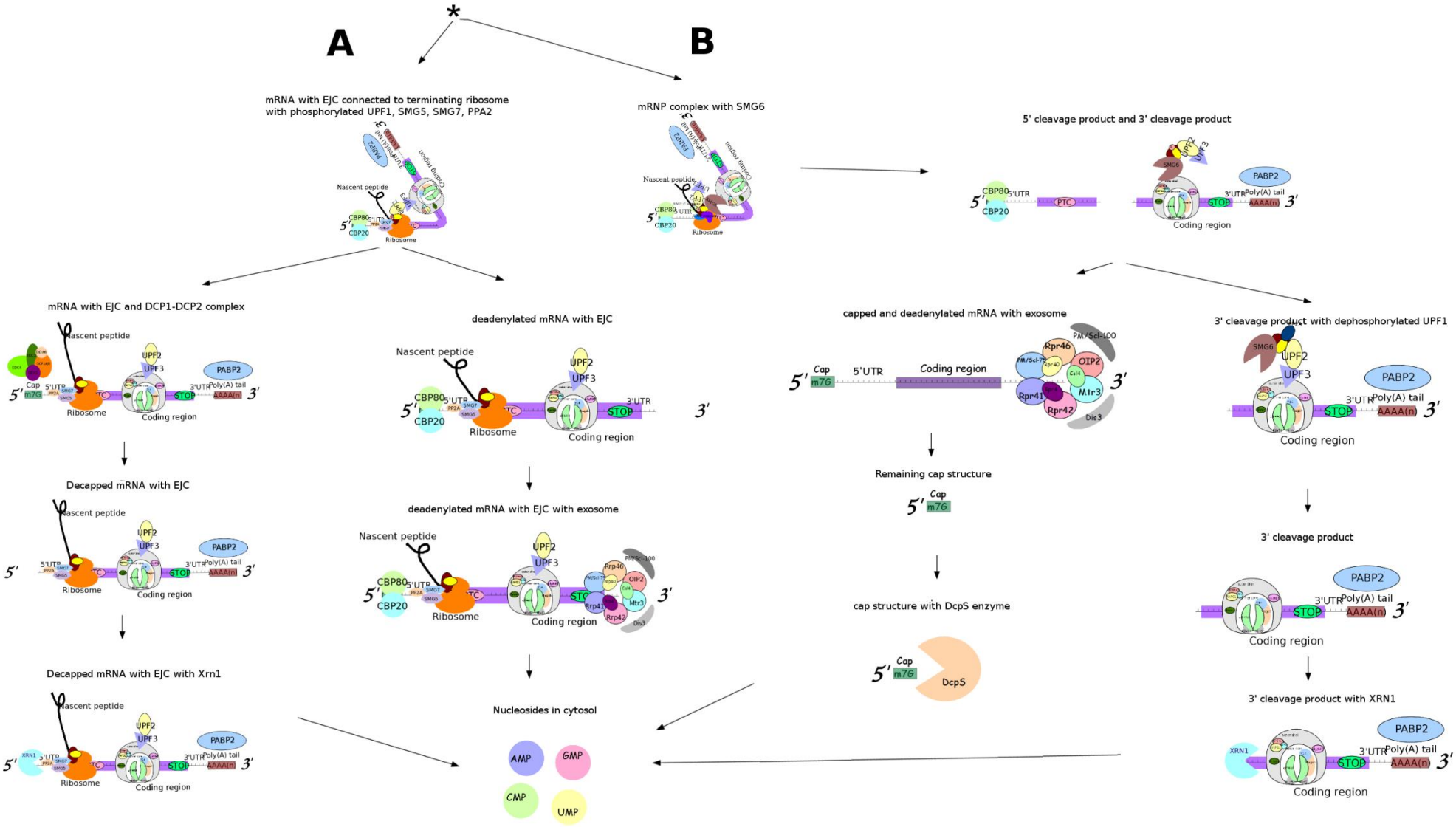
mRNA with EJC connected to terminating ribosome



mRNA with EJC connected to terminating ribosome with phosphorylated UPF1



*



Rycina 12 Ludzki szlak degradacji mRNA NMD (ang. *Nonsense-Mediated Decay*). A) szlak NMD zależny od SMG5-SMG7 B) szlak NMD zależny od SMG6.

Rysunek stworzony na podstawie oryginalnego rysunku z bazy danych RNApathwaysDB.

5.3.3.2 Białka:

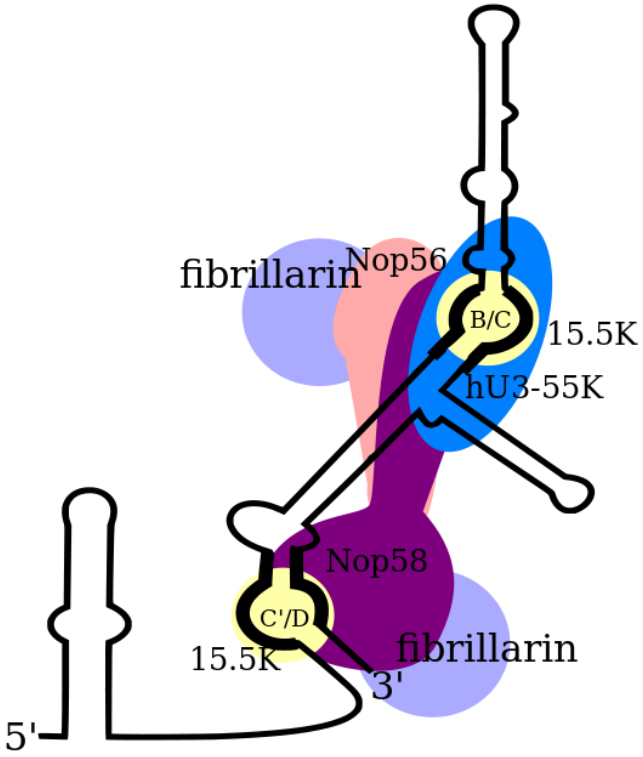
Na dzień 7.06.2013 w RNApathwaysDB znajduje się odpowiednio 17, 85 i 128 białek z *E. coli*, *S. cerevisiae* i *H. sapiens*. Zestaw ten nie zawiera wszystkich białek biorących udział w metabolizmie RNA. Jest ograniczony do tych białek, których rola została dobrze scharakteryzowana w procesach dojrzewania i degradacji RNA i które mogą być połączone z konkretnymi przejściami czy stanami. Każde białko posiada dedykowaną podstronę, która dostarcza informacji o kodującym je genie, podaje wszystkie nazwy i skróty, sekwencje z bazy danych NCBI, struktury trzeciorzędowe, jeśli są dostępne oraz łącza do zewnętrznych baz danych. Każda podstrona posiada także szczegółowy opis funkcji białka, aktywności enzymatycznej, informuje o obecności izoform, komórkowej/tkankowej specyficzności czy lokalizacji komórkowej. Wszystko wzbogacone jest o odnośniki literaturowe.

5.3.3.3 Cząsteczki RNA posiadające aktywność katalityczną

W tej sekcji zebrane zostały dane dotyczące cząsteczek RNA, które posiadają udokumentowaną aktywność katalityczną i mogą działać osobno bądź jako części składowe większych kompleksów. W bazie danych obecnie (7.06.2013) znajduje się 10 takich cząsteczek i dla każdej z nich dostępne są dane o nazewnictwie, sekwencji, rodzinie homologicznej z odnośnikiem do bazy RFAM, a także o oryginalnych publikacjach.

5.3.3.4 Kompleksy enzymatyczne

W bazie danych RNApathwaysDB zebranych zostało 39 kompleksów enzymatycznych: dwa z pałeczki okrężnicy, 25 z człowieka i 12 z drożdży (stan na dzień 7.06.2013). Dla każdego kompleksu poza opisem, budową podjednostkową oraz odnośnikami dostępny jest także rysunek przedstawiający uproszczoną reprezentację dwuwymiarową kompleksu w postaci „modelu ziemniaczkowego” (patrz: Rycina 13).



Rycina 13 Ludzki kompleks U3 snoRNP jako przykład reprezentacji dwuwymiarowej w postaci „modelu ziemniaczkowego”

5.3.3.5 Struktury

Dla białek, cząsteczek RNA i kompleksów enzymatycznych zawartych w RNApathwaysDB, dla których dostępne są struktury NMR bądź struktury krystaliczne podane są koordynaty przestrzenne atomów pobrane z bazy danych PDB. RNApathwaysDB przechowuje 241 modeli strukturalnych (stan na dzień 7.06.2013).

5.3.3.6 Publikacje

Odnośniki literaturowe do wszystkich rekordów w bazie zostały zebrane korzystając z bazy PubMed. Obecnie w bazie znajduje się 1635 pozycji.

5.3.4 Implementacja

Baza RNApathwaysDB napisana została z użyciem Django i używa systemu zarządzania relacyjnymi bazami danych MySQL. Serwis, podobnie jak REPAIRtoire, udostępnia następujące opcje: profil i logowanie, narzędzie do rysowania, edytowalne strony

i narzędzie do wyszukiwania. Profil i logowanie, a także strony edytowalne są zaimplementowane identycznie jak w opisanej wcześniej bazie danych. Różnice pojawiają się w narzędziu do rysowania oraz narzędziu do wyszukiwania, dlatego też zostaną one opisane osobno.

5.3.4.1 Narzędzie do rysowania

DrawBioPath to nowe narzędzie do rysowania schematów biologicznych (dostępne przez link *"draw a picture"* w głównym menu bazy, który przekierowuje do strony http://drawbiopath.genesilico.pl/). DrawBioPath, napisany w JavaScript, został opracowany w celu umożliwienia reprezentacji szlaków metabolicznych DNA i RNA w sposób przejrzysty i został użyty przy tworzeniu wszystkich rysunków do bazy danych RNApathwaysDB. Narzędzie to oparte jest na wolno dostępowym narzędziu do edytowania rysunków w formacie SVG – SVG-edit web editor (http://code.google.com/p/svg-edit/). System ten używa, tak samo jak narzędzie do rysowania udostępnione w bazie danych REPAIRtoire, formatu grafiki wektorowej SVG. Jednakże DrawBioPath, w odróżnieniu od rozwiązania dostępnego w REPAIRtoire, jest wygodniejsze i łatwiejsze w użyciu. Ponadto dostarcza biblioteki kształtów, które mogą być użyte do rysowania struktur cząsteczek RNA czy DNA, białek czy elementów nieregularnych. Istnieje także możliwość przesłania własnego pliku w formacie SVG i modyfikowanie go w edytorze graficznym albo tekstowym.

DrawBioPath wykonany został przez mgr Zuzannę Balcer. Pomysł oraz cechy programu opracowane zostały przez autorkę niniejszej rozprawy doktorskiej.

5.3.4.2 Narzędzie do wyszukiwania.

Baza danych RNApathwaysDB dostarcza dwóch możliwości przeszukiwania zasobów, które dostępne są w osobnej zakładce menu:

1. używając słów kluczowych;
2. używając sekwencji białkowej.

Wyszukiwanie tekstowe z użyciem słów kluczowych zwraca wyniki w postaci listy rekordów bazy, które zawierają zapytanie (podkreślone w wynikach na czerwono).

Przeszukiwanie bazy przy użyciu sekwencji białkowej wykorzystuje algorytm BLAST i bazę danych sekwencji białkowych przygotowaną na podstawie zasobów RNApathwaysDB. Ponadto z poziomu podstron z informacjami o białkach możliwe jest bezpośrednie wysłanie

sekwencji danego białka jako zapytania do programu BLAST dostępnego na serwerach NCBI.

5.3.5 Dostępność

Zawartość bazy RNApathwaysDB dostępna jest darmowo pod adresem http://iimcb.genesilico.pl/rnapathwaysdb. Narzędzie do generowania niestandardowych rysunków dostępne jest również darmowo pod adresem http://drawbiopath.genesilico.pl.

6 Dyskusja

Głównym celem projektu było stworzenie nowych biologicznych baz danych zbierających, porządkujących i opisujących dane dotyczące metabolizmu kwasów nukleinowych, które wejdą w skład ujednoliconego systemu bazodanowego opisującego cały metabolizm kwasów nukleinowych. W ramach projektu powstały dwie bazy danych - pierwsza baza danych poświęcona jest szlakom naprawy DNA (REPAIRtoire, opublikowana w Nucleic Acid Research i dostępna pod adresem: http://repairtoire.genesilico.pl), druga to baza danych szlaków metabolicznych RNA (RNApathwaysDB, również opublikowana w Nucleic Acid Research i dostępna pod adresem: http://iimcb.genesilico.pl/rnapathwaysdb).

REPAIRtoire – baza danych szlaków naprawy DNA

Temat naprawy DNA jest podejmowany przez wiele grup badawczych i udostępniany za pomocą wielu zasobów internetowych, jednakże do tej pory nie istniała specjalistyczna baza danych dedykowana jedynie szlakom naprawy DNA. Baza *repair*GENES (http://www.repairgenes.org/) zbiera informacje dotyczące genów kodujących białka zaangażowane w naprawę DNA, ale także dane uzyskane z zewnętrznych baz danych sekwencji i ontologii. Repair-FunMap (Wen i Feng 2004) była bazą dostarczającą wiedzy na temat sieci oddziaływań pomiędzy białkami zaangażowanymi w naprawę DNA, a innymi białkami, ale obecnie baza ta jest już niedostępna. Informacje odnośnie szlaków naprawy DNA oraz ich komponentów dostępne są także w ogólnych biologicznych bazach danych takich jak KEGG (Kanehisa, Araki i wsp. 2008) czy Reactome (Matthews, Gopinath i wsp. 2009). REPAIRtoire jest bazą danych dedykowaną naprawie DNA i zawiera bardziej szczegółowe i wyselekcjonowane informacje niż ogólne bazy danych. Ważnym elementem składowym bazy REPAIRtoire, który nie jest dostępny w innych źródłach, jest zestaw uszkodzeń DNA połączonych z potencjalnymi przyczynami każdego uszkodzenia oraz z efektami, które pojawiają się, gdy dane uszkodzenie nie jest usuwane. REPAIRtoire jest także źródłem unikalnym ze względu na to, że udostępnia informacje odnośnie wzajemnych połączeń pomiędzy uszkodzeniem (albo bardziej ogólnie – typem szkodzenia) a białkami, które takie uszkodzenia wykrywają i usuwają. Przeprowadzone w ramach tej pracy przeszukiwania danych literaturowych pozwoliły na zebranie zdecydowanie większej

liczby połączeń pomiędzy poszczególnymi uszkodzeniami DNA i odpowiednimi białkami, które je rozpoznają i eliminują, niż jest to dostępne w ogólnych bazach danych.

RNApathwaysDB – baza danych szlaków dojrzewania i degradacji RNA

To samo dotyczy szlaków dojrzewania i degradacji RNA. Metabolizm kwasów nukleinowych to temat stary i bardzo szeroki. Sieci wzajemnych powiązań i oddziaływań pomiędzy szlakami metabolicznymi kwasów rybonukleinowych są niezwykle złożone, liczba danych jest olbrzymia, jednakże informacje te są rozproszone po różnych źródłach, zarówno internetowych, jak i literaturowych. Tak samo jak w przypadku szlaków naprawy DNA, wiedzę odnośnie szlaków procesowania RNA znaleźć można w ogólnych bazach danych, takich jak: KEGG, Reactome, BioCyc (http://biocyc.org) (Caspi, Altman i wsp. 2012) czy WikiPathways (http://www.wikipathways.org/) (Kelder, van Iersel i wsp. 2012). Jednakże żadna z tych baz nie posiada kompletnego opisu dojrzewania i degradacji RNA. Olbrzymia ilość danych i brak pojedynczego bioinformatycznego źródła dedykowanego procesowaniu i degradacji RNA były inspiracją do stworzenia ogólnodostępnej bazy danych poświęconej tym właśnie zagadnieniom, która nie tylko zbiera informacje odnośnie szlaków obróbki RNA, ale także systematyzuje wiedzę i udostępnia ją szerokiemu gronu naukowemu.

RNApathwaysDB zawiera bardziej szczegółowe dane dotyczące szlaków procesowania RNA niż wymienione powyżej bazy danych, włączając w to informacje odnośnie różnych aspektów dojrzewania RNA, takich jak np. dodawanie czapeczki mRNA, wycinanie intronów, proces poliadenylacji, biosynteza tRNA czy dojrzewanie rRNA. Baza danych opisuje także szlaki degradacji RNA, takie jak szybka degradacja tRNA (ang *rapid tRNA decay*), szlak nadzorowania tRNA w jądrze (ang. *tRNA nuclear surveillance*), szlaki degradacji mRNA (ang. *mRNA decay*) czy szlaki kontroli jakości rRNA (ang. *rRNA quality control*). Ważną częścią składową RNApathwaysDB jest zbiór kompleksów enzymatycznych razem z ich strukturami makromolekularnymi przedstawionymi jako „modele ziemniaczkowe", które mogą być użyte np. do uproszczonych reprezentacji przebiegu procesów biologicznych, opisu mechanizmu ich działania czy do ilustrowania tych kompleksów, czy procesów podczas wykładów. Obecnie nie istnieje inne źródło dostarczające tak szerokiej wiedzy odnośnie metabolizmu RNA.

Potencjalne wykorzystanie stworzonych baz danych.

Dodatkową wartością projektu jest to, że stworzone bazy danych mogą znaleźć szereg zastosowań w różnych dziedzinach nauk - począwszy od biologii, przez biologię molekularną, biotechnologię, bioinformatykę po biologię systemów, medycynę i farmakologię. Dane zawarte w bazach danych są opracowane ręcznie, usystematyzowane i pokazane w sposób jasny, przejrzysty i czytelny dla użytkownika. Korzystać z nich mogą nie tylko naukowcy, ale także farmaceuci czy lekarze – co daje możliwość szybszego rozwoju nie tylko w środowisku ściśle naukowym, ale także klinicznym. Bardzo ważnym elementem niniejszej pracy było opracowanie zestawienia chorób genetycznych i powiązanych z nimi błędów pojawiających się podczas naprawy DNA. Zestawienie to zostało opracowane dla człowieka. Dane dotyczące chorób udostępnione są wraz z literaturą oraz powiązane zostały z odpowiednimi białkami oraz słowami kluczowymi wiążącymi je z odpowiednimi uszkodzeniami w cząsteczkach kwasu deoksyrybonukleinowego czy konkretnymi szlakami naprawy DNA. Udostępnienie danych w sposób przejrzysty i usystematyzowany w sieci Internet umożliwia korzystanie z nich naukowcom w Polsce i na świecie. Podczas rozwoju baz danych stworzone zostały również dwa programy do wizualizacji poszczególnych etapów ścieżek metabolicznych. Mogą one posłużyć do tworzenia rysunków ścieżek do publikacji, prezentacji czy prac naukowych.

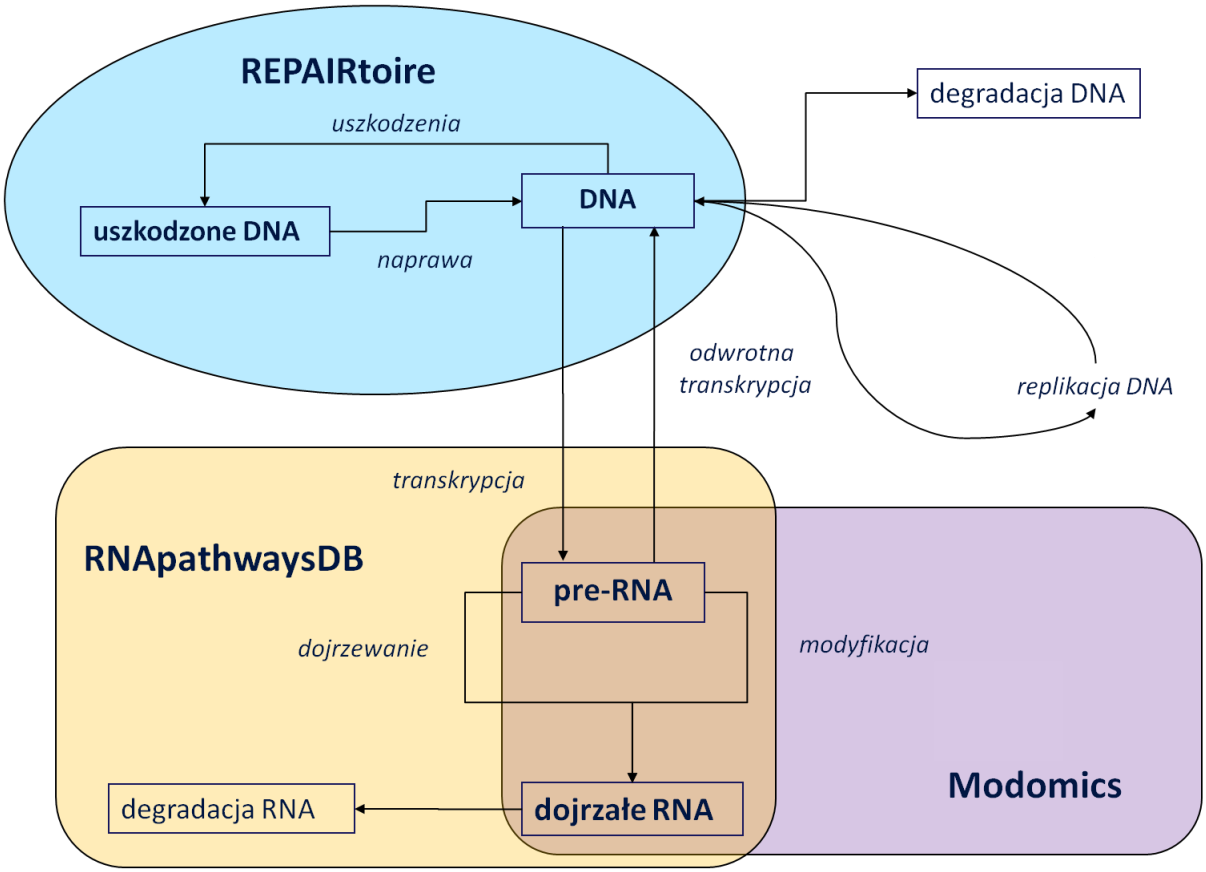
Baza danych szlaków dojrzewania i degradacji RNA również może być wykorzystana przez środowisko naukowe. Badanie funkcji RNA oraz enzymów biorących udział w metabolizmie RNA jest obecnie jednym z najintensywniej badanych obszarów biologii molekularnej. RNA pełni wiele kluczowych dla organizmu funkcji, m.in. przekazuje informację genetyczną pomiędzy DNA i białkiem. RNA ma także właściwości katalityczne i regulatorowe. Badania nad zjawiskami takimi jak interferencja RNA oraz działanie rybozymów i ryboprzełączników stwarzają ogromne perspektywy dla współczesnej medycyny. Dają potencjalną możliwość opracowania leków wyciszających geny kodujące niekorzystne białka, czy też ograniczających ekspresję genów wirusów takich jak HIV, HBV, HCV, czy polio. W każdym z tych przypadków spełnianie ww. funkcji jest możliwe dzięki odpowiednio przygotowanym (poprzez obróbkę enzymatyczną) cząsteczkom RNA. Koniecznie trzeba podkreślić, że w porównaniu z DNA i białkami cząsteczki RNA mają bardzo krótki czas życia i ich poziom w komórce (warunkujący odpowiednią funkcję) ulega często dużym zmianom; poziom ten regulowany jest przez enzymy biorące udział w syntezie, dojrzewaniu i degradacji. Dlatego też dokładne zrozumienie ścieżek metabolicznych

poszczególnych cząsteczek RNA, a także mechanizmów działania enzymów biorących udział w dojrzewaniu, degradacji czy modyfikacji RNA jest kluczowe dla wszelkich analiz funkcji RNA.

Tematyczne bazy danych poza tym, że systematyzują wiedzę, która obecnie rozproszona jest w wielu miejscach (począwszy od ogólnych baz danych po dane literaturowe), umożliwiają dotarcie do niezbędnych informacji w sposób szybki, oszczędzając pracy i pozwalając na skupienie się na samej analizie, a nie zbieraniu do niej danych. Zebrane informacje dostępne są w sieci Internet, co umożliwia naukowcom zajmującym się danym zagadnieniem na szybkie dotarcie do potrzebnych im wiadomości.

Ujednolicony system bazodanowy opisującego cały metabolizm kwasów nukleinowych.

Podczas rozwoju baz danych REPAIRtoire i RNApathwaysDB użyta została wiedza i doświadczenie z wcześniejszej pracy nad bazą danych szlaków modyfikacji RNA - MODOMICS (Dunin-Horkawicz, Czerwoniec i wsp. 2006, Czerwoniec, Dunin-Horkawicz i wsp. 2009, Machnicka, Milanowska i wsp. 2013). W przyszłości, wszystkie wymienione bazy danych posłużą jako elementy składowe większego systemu bazodanowego opisującego cały metabolizm kwasów nukleinowych. System taki pozwoli na szersze spojrzenie na biologię systemów DNA i RNA, dogłębną analizę poszczególnych szlaków, systematyzację wiedzy, udostępnienie istniejących danych w sposób jasny, przejrzysty i czytelny, a także pozwoli na symulacje szlaków i odkrycie nowych, nieznanych do tej pory elementów. Będzie to także stanowić punkt wyjścia do zaplanowania nowych badań doświadczalnych. Rycina 14 przedstawia dotychczasowe pokrycie tematów związanych z metabolizmem kwasów nukleinowych. W przyszłości system bazodanowy zostanie rozszerzony tak, by pokryć także tematy degradacji DNA, replikacji DNA oraz odwrotnej transkrypcji i stworzyć pełen obraz całego metabolizmu.



Rycina 14 Pokrycie tematu metabolizmu kwasów nukleinowych przez bazy danych stworzone w laboratorium prof. Janusza Bujnickiego.

Należy zwrócić uwagę, że tworzone bazy danych oraz wszelkie przeprowadzane analizy służą dogłębnemu poznaniu i zrozumieniu metabolizmu kwasów nukleinowych. Wszystkie uzyskane wyniki stanowią pierwszy etap do stworzenia meta-bazy danych oraz meta-metod, pozwalających na prowadzenie badań biologii systemów, w tym symulowanie poszczególnych szlaków, odkrywanie nowych elementów biorących udział w metabolizmie kwasów nukleinowych, a tym samym podjęcie próby opisania w pełni biologii kwasów nukleinowych i zaprojektowania doświadczeń badających właściwości nowo poznanych molekuł. Stworzone w ramach niniejszej pracy, publicznie dostępne bazy danych są nie tylko narzędziem informatycznym wspierającym wykorzystanie wiedzy i technologii powstających w naszym laboratorium, ale promują osiągnięcia i wyniki naukowe uzyskane w naszej pracowni bioinformatycznej na arenie międzynarodowej.

Największym wyzwaniem i najtrudniejszym elementem projektu było opracowanie wspólnej struktury danych i ujednoliconej architektury dla wszystkich baz danych, która

opisywałaby różne aspekty metabolizmu kwasów nukleinowych. Ponadto zebranie, opracowanie i udostępnienie wszystkich danych w sposób jasny i przejrzysty było również trudnym zadaniem. Z tymi aspektami rozprawy doktorskiej wiążą się perspektywy dalszego rozwoju całego systemu. Poza cechami, które już zostały opracowane, czyli:

- możliwość wprowadzania nowych danych, korygowania i edytowania już zawartych informacji przez ekspertów z poszczególnych dziedzin (redagowanie typu „wiki", moderowane przez twórców bazy danych, aby nie dopuścić do degradacji zawartości);
- integralność powstałych baz danych, dzięki wspólnej architekturze przy zachowaniu autonomiczności poszczególnych elementów tak, by stanowiły w pełni funkcjonalne źródło danych, dostarczające informacji z poszczególnych działów metabolizmu kwasów nukleinowych;

do systemu mają być wprowadzone w perspektywie czasu następujące usprawnienia:

- zintegrowanie wszystkich elementów w taki sposób, by możliwe było w przyszłości symulowanie skomplikowanych sieci reakcji biochemicznych albo z użyciem już istniejących symulatorów typu np. COPASI (Hoops, Sahle i wsp. 2006), albo z użyciem metodologii sieci Petriego (Peleg, Yeh i wsp. 2002);
- udostępnienie wszystkich zebranych informacji w stworzonych bazach danych w formatach, które mogą być stosowane w istniejących już narzędziach do analiz bioinformatycznych (np. BioPax (Demir, Cary i wsp. 2010) czy SMBL (ang. *System Biology Markup Language*) (Hucka, Finney i wsp. 2003)).

Poza rozwojem samej struktury systemu, czy dodawaniem nowych funkcjonalności do poszczególnych baz danych, olbrzymim wyzwaniem będzie utrzymanie aktualności wszystkich danych o stale pojawiające się nowe informacje literaturowe. W związku z dużym zmniejszeniem kosztów badań ilość danych z roku na rok przybywa w coraz większym tempie, co przyczynia się do zadania pytania czy da się opracować taką metodę aktualizacji, by w jakimś stopniu była ona automatyczna?, a jeśli tak, to na ile będzie można jej zaufać i w jakim stopniu można będzie zautomatyzować i tym samym przyspieszyć aktualizację danych? Ponadto bazy mają zostać uzupełnione o opisy szlaków w nowych organizmach oraz dane dotyczące nowych typów RNA, co może wiązać się z pojawieniem się konieczności reorganizacji struktury w taki sposób, by objąć dane nowego typu, jeśli takie się pojawią. Wszystkie wyżej wymienione zadania sprawiają, że niniejszy projekt jest perspektywiczny i będzie mógł być dalej rozwijany.

7 Wnioski

W wyniku prac, które opisane zostały w niniejszej rozprawie doktorskiej, opracowano:

- wspólną architekturę baz danych, dostosowaną do przechowywania informacji odnośnie szlaków metabolicznych kwasów nukleinowych;
- bazę danych szlaków naprawy DNA REPAIRtoire, która zbiera informacje odnośnie ośmiu typów szlaków naprawy DNA w trzech organizmach modelowych wraz ze wszystkimi elementami biorącymi w nich udział i oryginalnymi publikacjami opisującymi wszystkie elementy w bazie;
- bazę danych szlaków dojrzewania i degradacji RNA, która zbiera informacje odnośnie 33 szlaków dojrzewania i degradacji RNA w trzech organizmach modelowych wraz ze wszystkimi elementami biorącymi w nich udział i oryginalnymi publikacjami opisującymi wszystkie elementy w bazie.

Ponadto w wyniku prac związanych z powyższą rozprawą doktorską powstały:

- narzędzie do rysowania szlaków naprawy DNA, które dostępne jest w zakładce głównego menu bazy danych REPAIRtoire (wykonanie: mgr Grzegorz Papaj);
- narzędzie do rysowania rysunków biologicznych, które dostępne jest w zakładce głównego menu bazy danych RNApathwaysDB oraz na osobnej stronie dedykowanej narzędziu (wykonanie: mgr Zuzanna Balcer);

8 Wykaz załączników do pracy doktorskiej

Kod bazy danych szlaków naprawy DNA – REPAIRtoire. (CD)

Na płycie CD załączony został cały kod bazy danych szlaków naprawy DNA – REPAIRtoire.

Kod bazy danych dojrzewania i degradacji RNA – RNApathwaysDB. (CD)

Na płycie CD załączony został cały kod bazy dojrzewania i degradacji RNA – RNApathwaysDB.

9 Spis tabel

10 Spis rycin

11 Spis przykładów kodu

12 Bibliografia

(1999). "Nomenclature committee of the international union of biochemistry and molecular biology (NC-IUBMB), Enzyme Supplement 5 (1999)." Eur J Biochem **264**(2): 610-650.

AB., M. (2005). MySQL language reference, MySQL Press.

Alonso, C. R. (2012). "A complex 'mRNA degradation code' controls gene expression during animal development." Trends Genet **28**(2): 78-88.

Amaral, P. P., M. B. Clark, D. K. Gascoigne, M. E. Dinger i J. S. Mattick (2011). "lncRNAdb: a reference database for long noncoding RNAs." Nucleic Acids Res **39**(Database issue): D146-151.

Andreeva, A., D. Howorth, J. M. Chandonia, S. E. Brenner, T. J. Hubbard, C. Chothia i A. G. Murzin (2008). "Data growth and its impact on the SCOP database: new developments." Nucleic Acids Res **36**(Database issue): D419-425.

Arczewska, K. D. i J. T. Kusmierek (2007). "Bacterial DNA repair genes and their eukaryotic homologues: 2. Role of bacterial mutator gene homologues in human disease. Overview of nucleotide pool sanitization and mismatch repair systems." Acta Biochim Pol **54**(3): 435-457.

Arraiano, C. M., i wsp. (2010). "The critical role of RNA processing and degradation in the control of gene expression." FEMS Microbiol Rev **34**(5): 883-923.

Baker, D., P. Liu, A. Burdzy i L. C. Sowers (2002). "Characterization of the substrate specificity of a human 5-hydroxymethyluracil glycosylase activity." Chem Res Toxicol **15**(1): 33-39.

Ban, N., P. Nissen, J. Hansen, P. B. Moore i T. A. Steitz (2000). "The complete atomic structure of the large ribosomal subunit at 2.4 A resolution." Science **289**(5481): 905-920.

Barlow, D. P. (2011). "Genomic imprinting: a mammalian epigenetic discovery model." Annu Rev Genet **45**: 379-403.

Beggs, J. D. i D. Tollervey (2005). "Crosstalk between RNA metabolic pathways: an RNOMICS approach." Nat Rev Mol Cell Biol **6**(5): 423-429.

Belostotsky, D. (2009). "Exosome complex and pervasive transcription in eukaryotic genomes." Curr Opin Cell Biol **21**(3): 352-358.

Brissett, N. C. i A. J. Doherty (2009). "Repairing DNA double-strand breaks by the prokaryotic non-homologous end-joining pathway." Biochem Soc Trans **37**(Pt 3): 539-545.

Brodersen, P. i O. Voinnet (2006). "The diversity of RNA silencing pathways in plants." Trends Genet **22**(5): 268-280.

Bu, D., i wsp. (2012). "NONCODE v3.0: integrative annotation of long noncoding RNAs." Nucleic Acids Res **40**(Database issue): D210-215.

Burge, S. W., J. Daub, R. Eberhardt, J. Tate, L. Barquist, E. P. Nawrocki, S. R. Eddy, P. P. Gardner i A. Bateman (2013). "Rfam 11.0: 10 years of RNA families." Nucleic Acids Res **41**(Database issue): D226-232.

Cabusora, L., E. Sutton, A. Fulmer i C. V. Forst (2005). "Differential network expression during drug and stress response." Bioinformatics **21**(12): 2898-2905.

Cannone, J. J., i wsp. (2002). "The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs." BMC Bioinformatics **3**: 2.

Cantara, W. A., P. F. Crain, J. Rozenski, J. A. McCloskey, K. A. Harris, X. Zhang, F. A. Vendeix, D. Fabris i P. F. Agris (2011). "The RNA Modification Database, RNAMDB: 2011 update." Nucleic Acids Res **39**(Database issue): D195-201.

Cao, F., X. Li, S. Hiew, H. Brady, Y. Liu i Y. Dou (2009). "Dicer independent small RNAs associate with telomeric heterochromatin." Rna **15**(7): 1274-1281.

Carone, D. M., i wsp. (2009). "A new class of retroviral and satellite encoded small RNAs emanates from mammalian centromeres." Chromosoma **118**(1): 113-125.

Carthew, R. W. i E. J. Sontheimer (2009). "Origins and Mechanisms of miRNAs and siRNAs." Cell **136**(4): 642-655.

Caspi, R., i wsp. (2010). "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases." Nucleic Acids Res **38**(Database issue): D473-479.

Caspi, R., i wsp. (2012). "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases." Nucleic Acids Res **40**(Database issue): D742-753.

Cedar, H. i Y. Bergman (2012). "Programming of DNA methylation patterns." Annu Rev Biochem **81**: 97-117.

Cerami, E. G., B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader i C. Sander (2011). "Pathway Commons, a web resource for biological pathway data." Nucleic Acids Res **39**(Database issue): D685-690.

Chapman, B. i J. Chang (2000). "Biopython: python tools for computational biology." ACM SIGBIO Newslett. **20**: 15-19.

ChemAxon (2013). MarvinSketch 6.0.0.

Chen, C. Y. i A. B. Shyu (1995). "AU-rich elements: characterization and importance in mRNA degradation." Trends Biochem Sci **20**(11): 465-470.

Chernyakov, I., J. M. Whipple, L. Kotelawala, E. J. Grayhack i E. M. Phizicky (2008). "Degradation of several hypomodified mature tRNA species in Saccharomyces cerevisiae is mediated by Met22 and the 5'-3' exonucleases Rat1 and Xrn1." Genes Dev **22**(10): 1369-1380.

Cock, P. J., i wsp. (2009). "Biopython: freely available Python tools for computational molecular biology and bioinformatics." Bioinformatics **25**(11): 1422-1423.

Cohen, L. i R. Kaplan (1977). "Accumulation of nucleotides by starved Escherichia coli cells as a probe for the involvement of ribonucleases in ribonucleic acid degradation." J Bacteriol **129**(2): 651-657.

Coller, J. i R. Parker (2004). "Eukaryotic mRNA decapping." Annu Rev Biochem **73**: 861-890.

Condon, C. (2007). "Maturation and degradation of RNA in bacteria." Curr Opin Microbiol **10**(3): 271-278.

Crick, F. H. (1958). "On protein synthesis." Symp Soc Exp Biol **12**: 138-163.

Crick, F. H. (1968). "The origin of the genetic code." J Mol Biol **38**(3): 367-379.

Cuff, A. L., i wsp. (2011). "Extending CATH: increasing coverage of the protein structure universe and linking structure with function." Nucleic Acids Res **39**(Database issue): D420-426.

Czerwoniec, A., S. Dunin-Horkawicz, E. Purta, K. H. Kaminska, J. M. Kasprzak, J. M. Bujnicki, H. Grosjean i K. Rother (2009). "MODOMICS: a database of RNA modification pathways. 2008 update." Nucleic Acids Res **37**(Database issue): D118-121.

Dahm, R. (2005). "Friedrich Miescher and the discovery of DNA." Dev Biol **278**(2): 274-288.

Darty, K., A. Denise i Y. Ponty (2009). "VARNA: Interactive drawing and editing of the RNA secondary structure." Bioinformatics **25**(15): 1974-1975.

De Bont, R. i N. van Larebeke (2004). "Endogenous DNA damage in humans: a review of quantitative data." Mutagenesis **19**(3): 169-185.

Deana, A., H. Celesnik i J. G. Belasco (2008). "The bacterial enzyme RppH triggers messenger RNA degradation by 5' pyrophosphate removal." Nature **451**(7176): 355-358.

Demir, E., i wsp. (2010). "The BioPAX community standard for pathway data sharing." Nat Biotechnol **28**(9): 935-942.

Deutscher, M. P. (2009). "Maturation and degradation of ribosomal RNA in bacteria." Prog Mol Biol Transl Sci **85**: 369-391.

Diez, J., D. Walter, C. Munoz-Pinedo i T. Gabaldon (2010). "DeathBase: a database on structure, evolution and function of proteins involved in apoptosis and other forms of cell death." Cell Death Differ **17**(5): 735-736.

Doma, M. K. i R. Parker (2007). "RNA quality control in eukaryotes." Cell **131**(4): 660-668.

Drablos, F., i wsp. (2004). "Alkylation damage in DNA and RNA--repair mechanisms and medical significance." DNA Repair (Amst) **3**(11): 1389-1407.

Dunin-Horkawicz, S., A. Czerwoniec, M. J. Gajda, M. Feder, H. Grosjean i J. M. Bujnicki (2006). "MODOMICS: a database of RNA modification pathways." Nucleic Acids Res **34**(Database issue): D145-149.

Ender, C., A. Krek, M. R. Friedlander, M. Beitzinger, L. Weinmann, W. Chen, S. Pfeffer, N. Rajewsky i G. Meister (2008). "A human snoRNA with microRNA-like functions." Mol Cell **32**(4): 519-528.

Fatica, A. i D. Tollervey (2002). "Making ribosomes." Curr Opin Cell Biol **14**(3): 313-318.

Fire, A., S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver i C. C. Mello (1998). "Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans." Nature **391**(6669): 806-811.

Forcier, J., P. Bissex i W. Chun (2008). Python Web Development with Django, Addison-Wesley Professional.

Franklin, R. E. i R. G. Gosling (1953). "Molecular configuration in sodium thymonucleate." Nature **171**(4356): 740-741.

Friedberg, E. C., G. C. Walker, W. Siede, R. D. Wood, R. A. Schulz i T. Ellenberger (2006). "DNA repair and mutagenesis."

Garneau, N. L., J. Wilusz i C. J. Wilusz (2007). "The highways and byways of mRNA decay." Nat Rev Mol Cell Biol **8**(2): 113-126.

Ghildiyal, M. i P. D. Zamore (2009). "Small silencing RNAs: an expanding universe." Nat Rev Genet **10**(2): 94-108.

Goldman, N., P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos i E. Birney (2013). "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA." Nature **494**(7435): 77-80.

Grosjean, H. (2005). Fine-tuning of RNA functions by modification and editing. Berlin-Heidelberg, Springer-Verlag.

Guerrier-Takada, C., K. Gardiner, T. Marsh, N. Pace i S. Altman (1983). "The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme." Cell **35**(3 Pt 2): 849-857.

Hansen, W. K. i M. R. Kelley (2000). "Review of mammalian DNA repair and translational implications." J Pharmacol Exp Ther **295**(1): 1-9.

Hayashi, T., i wsp. (2001). "Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12." DNA Res **8**(1): 11-22.

Heller, R. C. i K. J. Marians (2006). "Replisome assembly and the direct restart of stalled replication forks." Nat Rev Mol Cell Biol **7**(12): 932-943.

Herraez, A. (2006). "Biomolecules in the computer: Jmol to the rescue." Biochem Mol Biol Educ **34**(4): 255-261.

Hershey, A. D. i M. Chase (1952). "Independent functions of viral protein and nucleic acid in growth of bacteriophage." J Gen Physiol **36**(1): 39-56.

Hocine, S., R. H. Singer i D. Grunwald (2010). "RNA processing and export." Cold Spring Harb Perspect Biol **2**(12): a000752.

Holbrook, S. R. (2008). "Structural principles from large RNAs." Annu Rev Biophys **37**: 445-464.

Holt, C. E. i S. L. Bullock (2009). "Subcellular mRNA localization in animal cells and why it matters." Science **326**(5957): 1212-1216.

Hoogstraten, C. G. i M. Sumita (2007). "Structure-function relationships in RNA and RNP enzymes: recent advances." Biopolymers **87**(5-6): 317-328.

Hoops, S., i wsp. (2006). "COPASI--a COmplex PAthway SImulator." Bioinformatics **22**(24): 3067-3074.

Houdebine, L. M. (2007). "Transgenic animal models in biomedical research." Methods Mol Biol **360**: 163-202.

Houseley, J. i D. Tollervey (2009). "The many pathways of RNA degradation." Cell **136**(4): 763-776.

Hu, Q. i M. G. Rosenfeld (2012). "Epigenetic regulation of human embryonic stem cells." Front Genet **3**: 238.

Hucka, M., i wsp. (2003). "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models." Bioinformatics **19**(4): 524-531.

Hunter, S., i wsp. (2012). "InterPro in 2011: new developments in the family and domain prediction database." Nucleic Acids Res **40**(Database issue): D306-312.

Iftode, C., Y. Daniely i J. A. Borowiec (1999). "Replication protein A (RPA): the eukaryotic SSB." Crit Rev Biochem Mol Biol **34**(3): 141-180.

Iglesias, N. i F. Stutz (2008). "Regulation of mRNP dynamics along the export pathway." FEBS Lett **582**(14): 1987-1996.

Isken, O. i L. E. Maquat (2007). "Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function." Genes Dev **21**(15): 1833-1856.

Iyer, R. R., A. Pluciennik, V. Burdett i P. L. Modrich (2006). "DNA mismatch repair: functions and mechanisms." Chem Rev **106**(2): 302-323.

Jeggo, P. i M. F. Lavin (2009). "Cellular radiosensitivity: how much better do we understand it?" Int J Radiat Biol **85**(12): 1061-1081.

Juhling, F., M. Morl, R. K. Hartmann, M. Sprinzl, P. F. Stadler i J. Putz (2009). "tRNAdb 2009: compilation of tRNA sequences and tRNA genes." Nucleic Acids Res **37**(Database issue): D159-162.

Kanehisa, M., i wsp. (2008). "KEGG for linking genomes to life and the environment." Nucleic Acids Res **36**(Database issue): D480-484.

Kaplan, R. i D. Apirion (1974). "The involvement of ribonuclease I, ribonuclease II, and polynucleotide phosphorylase in the degradation of stable ribonucleic acid during carbon starvation in Escherichia coli." J Biol Chem **249**(1): 149-151.

Kelder, T., M. P. van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo i A. R. Pico (2012). "WikiPathways: building research communities on biological pathways." Nucleic Acids Res **40**(Database issue): D1301-1307.

Keseler, I. M., i wsp. (2009). "EcoCyc: a comprehensive view of Escherichia coli biology." Nucleic Acids Res **37**(Database issue): D464-470.

Kiljunen, S., K. Hakala, E. Pinta, S. Huttunen, P. Pluta, A. Gador, H. Lonnberg i M. Skurnik (2005). "Yersiniophage phiR1-37 is a tailed bacteriophage having a 270 kb DNA genome with thymidine replaced by deoxyuridine." Microbiology **151**(Pt 12): 4093-4102.

Klose, R. J. i A. P. Bird (2006). "Genomic DNA methylation: the mark and its mediators." Trends Biochem Sci **31**(2): 89-97.

Kozomara, A. i S. Griffiths-Jones (2011). "miRBase: integrating microRNA annotation and deep-sequencing data." Nucleic Acids Res **39**(Database issue): D152-157.

Kruger, K., P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling i T. R. Cech (1982). "Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena." Cell **31**(1): 147-157.

Krwawicz, J., K. D. Arczewska, E. Speina, A. Maciejewska i E. Grzesiuk (2007). "Bacterial DNA repair genes and their eukaryotic homologues: 1. Mutations in genes involved in base excision repair (BER) and DNA-end processors and their implication in mutagenesis and human disease." Acta Biochim Pol **54**(3): 413-434.

Kuchta, K., D. Barszcz, E. Grzesiuk, P. Pomorski i J. Krwawicz (2012). "DNAtraffic--a new database for systems biology of DNA dynamics during the cell life." Nucleic Acids Res **40**(Database issue): D1235-1240.

Kunz, C., Y. Saito i P. Schar (2009). "DNA Repair in mammalian cells: Mismatched repair: variations on a theme." Cell Mol Life Sci **66**(6): 1021-1038.

Kushner, S. R. (2004). "mRNA decay in prokaryotes and eukaryotes: different approaches to a similar problem." IUBMB Life **56**(10): 585-594.

Lafontaine, D. L. J. i D. Tollervey (2001). "The function and synthesis of ribosomes." Nat Rev Mol Cell Biol **2**(7): 514-520.

Langenberger, D., C. Bermudez-Santana, J. Hertel, S. Hoffmann, P. Khaitovich i P. F. Stadler (2009). "Evidence for human microRNA-offset RNAs in small RNA sequencing data." Bioinformatics **25**(18): 2298-2301.

LaRiviere, F. J., S. E. Cole, D. J. Ferullo i M. J. Moore (2006). "A late-acting quality control process for mature eukaryotic rRNAs." Mol Cell **24**(4): 619-626.

Leinonen, R., i wsp. (2011). "The European Nucleotide Archive." Nucleic Acids Res **39**(Database issue): D28-31.

Levene, P. A. (1917). "THE STRUCTURE OF YEAST NUCLEIC ACID." Journal of Biological Chemistry **31**(3): 591-598.

Li, Z., S. Van Calcar, C. Qu, W. K. Cavenee, M. Q. Zhang i B. Ren (2003). "A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells." Proc Natl Acad Sci U S A **100**(14): 8164-8169.

Licatalosi, D. D. i R. B. Darnell (2010). "RNA processing and its regulation: global insights into biological networks." Nat Rev Genet **11**(1): 75-87.

Lindahl, T. (1993). "Instability and decay of the primary structure of DNA." Nature **362**(6422): 709-715.

Lindahl, T. i R. D. Wood (1999). "Quality control by DNA repair." Science **286**(5446): 1897-1905.

Luna, R., H. Gaillard, C. Gonzalez-Aguilera i A. Aguilera (2008). "Biogenesis of mRNPs: integrating different processes in the eukaryotic nucleus." Chromosoma **117**(4): 319-331.

Lutz, C. S. i A. Moreira (2011). "Alternative mRNA polyadenylation in eukaryotes: an effective regulator of gene expression." Wiley Interdiscip Rev RNA **2**(1): 22-31.

Machnicka, M. A., i wsp. (2013). "MODOMICS: a database of RNA modification pathways--2013 update." Nucleic Acids Res **41**(Database issue): D262-267.

Maddukuri, L., D. Dudzinska i B. Tudek (2007). "Bacterial DNA repair genes and their eukaryotic homologues: 4. The role of nucleotide excision DNA repair (NER) system in mammalian cells." Acta Biochim Pol **54**(3): 469-482.

Malone, C. D. i G. J. Hannon (2009). "Small RNAs as guardians of the genome." Cell **136**(4): 656-668.

Maquat, L. E. (2005). "Nonsense-mediated mRNA decay in mammals." J Cell Sci **118**(Pt 9): 1773-1776.

Mastroyiannopoulos, N. P., J. B. Uney i L. A. Phylactou (2010). "The application of ribozymes and DNAzymes in muscle and brain." Molecules **15**(8): 5460-5472.

Matera, A. G., R. M. Terns i M. P. Terns (2007). "Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs." Nat Rev Mol Cell Biol **8**(3): 209-220.

Matthews, L., i wsp. (2009). "Reactome knowledgebase of human biological pathways and processes." Nucleic Acids Res **37**(Database issue): D619-622.

Mattick, J. S. i I. V. Makunin (2006). "Non-coding RNA." Hum Mol Genet **15 Spec No 1**: R17-29.

Mazumder, B., V. Seshadri i P. L. Fox (2003). "Translational control by the 3'-UTR: the ends specify the means." Trends Biochem Sci **28**(2): 91-98.

Mercer, T. R., M. E. Dinger i J. S. Mattick (2009). "Long non-coding RNAs: insights into functions." Nat Rev Genet **10**(3): 155-159.

Milanowska, K., J. Krwawicz, G. Papaj, J. Kosinski, K. Poleszak, J. Lesiak, E. Osinska, K. Rother i J. M. Bujnicki (2011). "REPAIRtoire--a database of DNA repair pathways." Nucleic Acids Res **39**(Database issue): D788-792.

Milanowska, K., K. Mikolajczak, A. Lukasik, M. Skorupski, Z. Balcer, M. A. Machnicka, M. Nowacka, K. M. Rother i J. M. Bujnicki (2013). "RNApathwaysDB--a database of RNA maturation and decay pathways." Nucleic Acids Res **41**(Database issue): D268-272.

Modrich, P. (2006). "Mechanisms in eukaryotic mismatch repair." J Biol Chem **281**(41): 30305-30309.

Motorin, Y. i M. Helm (2011). "RNA nucleotide methylation." Wiley Interdiscip Rev RNA **2**(5): 611-631.

Myers, L. C. i R. D. Kornberg (2000). "Mediator of transcriptional regulation." Annu Rev Biochem **69**: 729-749.

Newman, C. (2005). SQLite, Sams.

Nieduszynski, C. A., S. Hiraga, P. Ak, C. J. Benham i A. D. Donaldson (2007). "OriDB: a DNA replication origin database." Nucleic Acids Res **35**(Database issue): D40-46.

Ogawa, Y., B. K. Sun i J. T. Lee (2008). "Intersection of the RNA interference and X-inactivation pathways." Science **320**(5881): 1336-1341.

Olinski, R., A. Siomek, R. Rozalski, D. Gackowski, M. Foksinski, J. Guz, T. Dziaman, A. Szpila i B. Tudek (2007). "Oxidative damage to DNA and antioxidant status in aging and age-related diseases." Acta Biochim Pol **54**(1): 11-26.

Orgel, L. E. (1968). "Evolution of the genetic apparatus." J Mol Biol **38**(3): 381-393.

Peleg, M., I. Yeh i R. B. Altman (2002). "Modelling biological processes using workflow and Petri Net models." Bioinformatics **18**(6): 825-837.

Phizicky, E. M. i A. K. Hopper (2010). "tRNA biology charges to the front." Genes Dev **24**(17): 1832-1860.

Pickering, B. M. i A. E. Willis (2005). "The implications of structured 5' untranslated regions on translation and disease." Semin Cell Dev Biol **16**(1): 39-47.

Plotz, G., S. Zeuzem i J. Raedle (2006). "DNA mismatch repair and Lynch syndrome." J Mol Histol **37**(5-7): 271-283.

Pluciennik, A., L. Dzantiev, R. R. Iyer, N. Constantin, F. A. Kadyrov i P. Modrich (2010). "PCNA function in the activation and strand direction of MutLalpha endonuclease in mismatch repair." Proc Natl Acad Sci U S A **107**(37): 16066-16071.

Podlevsky, J. D., C. J. Bley, R. V. Omana, X. Qi i J. J. Chen (2008). "The telomerase database." Nucleic Acids Res **36**(Database issue): D339-343.

Popow, J., A. Schleiffer i J. Martinez (2012). "Diversity and roles of (t)RNA ligases." Cell Mol Life Sci **69**(16): 2657-2670.

Punta, M., i wsp. (2012). "The Pfam protein families database." Nucleic Acids Res **40**(Database issue): D290-301.

Raptis, S. i B. Bapat (2006). "Genetic instability in human tumors." EXS(96): 303-320.

Ray, B. K. i D. Apirion (1979). "Characterization of 10S RNA: a new stable rna molecule from Escherichia coli." Mol Gen Genet **174**(1): 25-32.

Roberts, R. J., T. Vincze, J. Posfai i D. Macelis (2010). "REBASE--a database for DNA restriction and modification: enzymes, genes and genomes." Nucleic Acids Res **38**(Database issue): D234-236.

Robertson, A. B., A. Klungland, T. Rognes i I. Leiros (2009). "DNA repair in mammalian cells: Base excision repair: the long and short of it." Cell Mol Life Sci **66**(6): 981-993.

Ronen, A. i B. W. Glickman (2001). "Human DNA repair genes." Environ Mol Mutagen **37**(3): 241-283.

Rose, P. W., i wsp. (2011). "The RCSB Protein Data Bank: redesigned web site and web services." Nucleic Acids Res **39**(Database issue): D392-401.

Rothemund, P. W. (2006). "Folding DNA to create nanoscale shapes and patterns." Nature **440**(7082): 297-302.

Safran, M., i wsp. (2010). "GeneCards Version 3: the human gene integrator." Database : the journal of biological databases and curation **2010**: baq020.

Salazar, M., O. Y. Fedoroff, J. M. Miller, N. S. Ribeiro i B. R. Reid (1993). "The DNA strand in DNA.RNA hybrid duplexes is neither B-form nor A-form in solution." Biochemistry **32**(16): 4207-4215.

Salmena, L., L. Poliseno, Y. Tay, L. Kats i P. P. Pandolfi (2011). "A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?" Cell **146**(3): 353-358.

Sana, J., P. Faltejskova, M. Svoboda i O. Slaby (2012). "Novel classes of non-coding RNAs and cancer." J Transl Med **10**: 103.

Saraiya, A. A. i C. C. Wang (2008). "snoRNA, a novel precursor of microRNA in Giardia lamblia." PLoS Pathog **4**(11): e1000224.

Sayers, E. W., i wsp. (2011). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **39**(Database issue): D38-51.

Sayers, E. W., i wsp. (2009). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **37**(Database issue): D5-15.

Scheer, M., i wsp. (2011). "BRENDA, the enzyme information system in 2011." Nucleic Acids Res **39**(Database issue): D670-676.

Shazman, S., G. Elber i Y. Mandel-Gutfreund (2011). "From face to interface recognition: a differential geometric approach to distinguish DNA from RNA binding surfaces." Nucleic Acids Res **39**(17): 7390-7399.

Shi, W., D. Hendrix, M. Levine i B. Haley (2009). "A distinct class of small RNAs arises from pre-miRNA-proximal regions in a simple chordate." Nat Struct Mol Biol **16**(2): 183-189.

Slatter, M. A. i A. R. Gennery (2010). "Primary immunodeficiencies associated with DNA-repair disorders." Expert Rev Mol Med **12**: e9.

Spiegelman, B. M. i R. Heinrich (2004). "Biological control through regulated transcriptional coactivators." Cell **119**(2): 157-167.

Stein, L. D. (2004). "Using the Reactome database." Curr Protoc Bioinformatics **Chapter 8**: Unit 8 7.

Szymanski, M., M. Z. Barciszewska, V. A. Erdmann i J. Barciszewski (2002). "5S Ribosomal RNA Database." Nucleic Acids Res **30**(1): 176-178.

Taft, R. J., E. A. Glazov, T. Lassmann, Y. Hayashizaki, P. Carninci i J. S. Mattick (2009). "Small RNAs derived from snoRNAs." Rna **15**(7): 1233-1240.

Taft, R. J., C. D. Kaplan, C. Simons i J. S. Mattick (2009). "Evolution, biogenesis and function of promoter-associated RNAs." Cell Cycle **8**(15): 2332-2338.

Thompson, D. M. i R. Parker (2009). "Stressing out over tRNA cleavage." Cell **138**(2): 215-219.

Tomecki, R. i A. Dziembowski (2010). "Novel endoribonucleases as central players in various pathways of eukaryotic RNA metabolism." Rna **16**(9): 1692-1724.

Triman, K. L., A. Peister i R. A. Goel (1998). "Expanded versions of the 16S and 23S ribosomal RNA mutation databases (16SMDBexp and 23SMDBexp)." Nucleic Acids Res **26**(1): 280-284.

Tudek, B. (2007). "Base excision repair modulation as a risk factor for human cancers." Mol Aspects Med **28**(3-4): 258-275.

Turinsky, A. L., i wsp. (2011). "DAnCER: disease-annotated chromatin epigenetics resource." Nucleic Acids Res **39**(Database issue): D889-894.

UniProt, C. (2012). "Reorganizing the protein space at the Universal Protein Resource (UniProt)." Nucleic Acids Res **40**(Database issue): D71-75.

Vaisman, A., A. R. Lehmann i R. Woodgate (2004). "DNA polymerases eta and iota." Adv Protein Chem **69**: 205-228.

Venter, J. C., i wsp. (2001). "The sequence of the human genome." Science **291**(5507): 1304-1351.

Walter, G. (1986). "Origin of life: The RNA world." Nature **319**(6055): 618-618.

Wang, J., J. Zhang, K. Li, W. Zhao i Q. Cui (2012). "SpliceDisease database: linking RNA splicing and disease." Nucleic Acids Res **40**(Database issue): D1055-1059.

Watson, J. D. i F. H. Crick (1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid." Nature **171**(4356): 737-738.

Weddington, N., A. Stuy, I. Hiratani, T. Ryba, T. Yokochi i D. M. Gilbert (2008). "ReplicationDomain: a visualization tool and comparative database for genome-wide replication timing data." BMC Bioinformatics **9**: 530.

Wen, L. i J. A. Feng (2004). "Repair-FunMap: a functional database of proteins of the DNA repair systems." Bioinformatics **20**(13): 2135-2137.

Westhof, E. i P. Auffinger (2006). RNA Tertiary Structure. Encyclopedia of Analytical Chemistry, John Wiley & Sons, Ltd.

Wilson, D. M., 3rd i D. Barsky (2001). "The major human abasic endonuclease: formation, consequences and repair of abasic lesions in DNA." Mutat Res **485**(4): 283-307.

Wilusz, J. E., H. Sunwoo i D. L. Spector (2009). "Long noncoding RNAs: functional surprises from the RNA world." Genes Dev **23**(13): 1494-1504.

Winter, J., S. Jung, S. Keller, R. I. Gregory i S. Diederichs (2009). "Many roads to maturity: microRNA biogenesis pathways and their regulation." Nat Cell Biol **11**(3): 228-234.

Wood, R. D., M. Mitchell i T. Lindahl (2005). "Human DNA repair genes, 2005." Mutat Res **577**(1-2): 275-283.

Wood, R. D., M. Mitchell i T. Lindahl (2013). "Human DNA Repair Genes."

Wood, R. D., M. Mitchell, J. Sgouros i T. Lindahl (2001). "Human DNA repair genes." Science **291**(5507): 1284-1289.

Wright, W. E., V. M. Tesmer, K. E. Huffman, S. D. Levene i J. W. Shay (1997). "Normal human chromosomes have long G-rich telomeric overhangs at one end." Genes Dev **11**(21): 2801-2809.

Xu, M., S. Medvedev, J. Yang i N. B. Hecht (2009). "MIWI-independent small RNAs (MSY-RNAs) bind to the RNA-binding protein, MSY2, in male germ cells." Proc Natl Acad Sci U S A **106**(30): 12371-12376.

Yakovchuk, P., E. Protozanova i M. D. Frank-Kamenetskii (2006). "Base-stacking and base-pairing contributions into thermal stability of the DNA double helix." Nucleic Acids Res **34**(2): 564-574.