Wydział Biologii
Uniwersytet im. Adama Mickiewicza w Poznaniu

Rozprawa doktorska

Identyfikacja domen WG/GW zaangażowanych w wiązanie białek Argonaute oraz analiza mechanizmów molekularnych odpowiedzialnych za ich zmienność

Andrzej Zieleziński

Praca napisana pod kierunkiem
Prof. dr. hab. Wojciecha Karłowskiego
w Pracowni Bioinformatyki
Instytutu Biologii Molekularnej i Biotechnologii

Poznań, 2014

Podziękowania

Składam serdeczne podziękowania

mojemu Promotorowi
Panu prof. dr. hab. Wojciechowi Karłowskiemu
za intelektualne inspiracje oraz wiedzę,
jaką przekazał mi w okresie studiów doktoranckich,
a także twórcze rozwinięcie moich zainteresowań naukowych,
tak abym mógł przygotować niniejszą dysertację

Dziękuję również Koleżankom i Kolegom z Pracowni Bioinformatyki, w szczególności
dr. Maciejowi Szymańskiemu, dr. Markowi Żywickiemu oraz mgr Sylwii Alabie
za pomoc w kwestiach naukowych, a także pouczające dyskusje

Finansowanie

Niniejsza praca powstała przy finansowym udziale:

1. Narodowego Centrum Nauki (grant 2011/03/N/NZ2/01440 dla A.Z.)
2. Wydziału Biologii Uniwersytetu im. Adama Mickiewicza w Poznaniu (grant GDWB-09/2011, dla A.Z.)
3. Wojewódzkiego Urzędu Pracy (stypendium w ramach projektu „Wsparcie stypendialne dla doktorantów na kierunkach uznanych za strategiczne z punktu widzenia rozwoju Wielkopolski” w ramach programu 8.2.2. Programu Operacyjnego Kapitał Ludzki)

Publikacje autora związane z pracą doktorską

1. Zielezinski A & Karlowski WM.
Tyrosine-tryptophan substitution switches on the rapid evolution of WG/GW domain in RRM AGO-binding proteins from Arabidopsis and rice.
[publikacja w przygotowaniu]

2. Zielezinski A & Karlowski WM.
Whub: a comprehensive knowledgebase portal for AGO-binding protein research
Nucleic Acids Research
[artykuł wysłany do redakcji]

3. Zielezinski A & Karlowski WM.
Agos - -a universal web tool for GW Argonaute-binding domain prediction.
Bioinformatics. 27(9). 2011

4. Karlowski WM, Zielezinski A, Carrère J, Pontier D, Lagrange T, Cooke R.
Genome-wide computational identification of WG/GW
Argonaute-binding proteins in Arabidopsis.
Nucleic Acids Research. 38(13). 2010

Spis treści

Streszczenie

Wstęp. Domeny białkowe WG/GW złożone z licznie występujących par tryptofanu (W) i glicyny (G) są niezbędne do wiązania białek Argonaute (AGO) w procesie interferencji RNA (RNAi). Bardzo niski stopień podobieństwa sekwencji domen WG/GW, różna długość ich sekwencji (22-700 reszt), zmienna liczba powtórzeń motywu WG/GW (1-45) uniemożliwiają wiarygodne określenie ich relacji homologicznych, a także są źródłem trudności podczas ich identyfikacji tradycyjnymi metodami przewidywania domen i motywów białkowych.

Cel pracy. Celem niniejszej pracy jest identyfikacja nowych białek zawierających domenę WG/GW wiążącą AGO oraz zbadanie mechanizmów molekularnych warunkujących ich zróżnicowanie.

Metody. Stworzone programy adnotacji domen wiążących AGO zostały napisane w języku Python. Zakres badań filogenetycznych sprowadzono do konserwatywnych fragmentów białek oddziałujących z AGO.

Wyniki. Opracowano trzy metody identyfikacji *de novo* domen WG/GW (Agos, Wsearch, i-Wsearch) zaimplementowane w formie ogólnodostępnych aplikacji internetowych i programów przeznaczonych do uruchomienia na lokalnym komputerze, które stanowią pierwsze bioinformatyczne narzędzia służące do adnotacji domen wiążących AGO. Wynikiem ich zastosowania są listy rankingowe nowych genów kodujących potencjalne domeny WG/GW u Eukariota, z których część została już potwierdzona eksperymentalnie (WGRP1, SDE3, hnRNP). Skanowanie genomów Prokariota pozwoliło także zidentyfikować sekwencje pierwszych potencjalnych domen WG/GW w tym królestwie, które w większości występują u gatunków archeonów i bakterii kodujących białka Argonaute. Również wśród wirusów

infekujących rośliny i zwierzęta znalezione zostały statystycznie istotne motywy WG/GW, które mogą oddziaływać z białkami AGO komórki gospodarza w celu przełamania jego systemu odporności zależnego od RNAi.

Z rekonstrukcji filogenetycznej i analizy porównawczej domen WG/GW, przeprowadzonej w wielogenowej rodzinie białek hnRNP, wynika, że sekwencje genomowe domen wiążących AGO stanowią ekstremalny przypadek polimorfizmu genetycznego. Oprócz tandemowych i segmentowych duplikacji oraz alternatywnego splicingu, fragmenty sekwencji kodujących motywy WG/GW znajdują się pod działaniem pozytywnej selekcji przyspieszającej utrwalanie substytucji aminokwasowych, które dodatkowo podlegają licznym przetasowaniom między paralogami na drodze częstych rekombinacji i nierównego crossing-over. Przeprowadzając wirtualną symulację eksperymentu polegającego na wymianie domen WG/GW między niespokrewnionymi roślinnymi i zwierzęcymi białkami wiążącymi AGO wykazano, że mimo wysokiego zróżnicowania długości i stopnia zachowania sekwencji tych domen, ich specyficzna kompozycja aminokwasowa jest zachowana u wszystkich organizmów eukariotycznych. Filogenia domen WG/GW oraz analiza ich składu aminokwasowego sugerują, że zarówno zwierzęce, jak i roślinne domeny wiążące AGO, występujące w różnych rodzinach białkowych, powstają z regionów białek inherentnie nieuporządkowanych (IDP, ang. *intrinsically disordered proteins*) obejmujących także regiony glicynobogate. Ponadto wykazano, że uniwersalną cechą eukariotycznych domen wiążących AGO są wielokrotne - najczęściej tandemowe - powtórzenia motywów długości 10-20 reszt aminokwasowych, wewnątrz których znajduje się reszta Trp otoczona hydrofilowymi obszarami o nieskompensowanych ładunkach, co określa miejsca wiązania AGO w ograniczonym podzbiorze możliwych sekwencji.

Końcowym wynikiem tego projektu jest publicznie dostępny, przyjazny dla użytkownika portal internetowy (Whub), który oferuje zestaw narzędzi, takich jak: (i) katalog eksperymentalnie potwierdzonych białek wiążących AGO dostarczający informacji na temat funkcjonalnych regionów białkowych i wpływu mutagenezy na funkcjonowanie domeny oraz fenotyp, a także informacji o dostępnych publikacjach; (ii) interaktywne aplikacje internetowe przeznaczone do analizy pozycyjnie-zależnej kompozycji aminokwasowej funkcjonalnych motywów bogatych w Trp; (iii) wersje on-line programów Agos, Wsearch i i-Wsearch umożliwiające wizualizację wyników; (iv) interaktywną grę internetową, która polega na generowaniu *in silico* sekwencji wiążących białka AGO lub modyfikowaniu istniejących sekwencji poprzez wprowadzanie mutacji lub symulowanie eksperymentów laboratoryjnych związanych z mutagenezą tych białek.

1

Wstęp

Dominujący postulat biologii molekularnej o jednoznaczności sekwencji i struktury białka oraz pełnionej przez niego funkcji okazał się fundamentalny w wyjaśnieniu funkcji tysięcy różnych domen i rodzin białkowych [1,2]. Ostatnie doniesienia naukowe ujawniają jednak przypadki wychodzące poza ten klasyczny kanon. Przykładem są tu białka zawierające domenę WG/GW (Trp-Gly/Gly-Trp), która złożona jest z licznie powtórzonych par zawierających reszty tryptofanu W (Trp) i glicyny G (Gly). Obecność tego binarnego kodu jest niezbędna niemal u wszystkich organizmów eukariotycznych podczas procesu interferencji RNA (RNAi, ang. *RNA interference*) stanowiącego naturalny mechanizm regulacji ekspresji genów, w którym małe cząsteczki RNA znajdujące się w kompleksie z białkami Argonaute (AGO) i białkami zawierającymi domenę WG/GW, odgrywają rolę przewodników odnajdujących komplementarne do nich docelowe sekwencje.

Domeny WG/GW, nie tylko charakteryzują się różną długością i bardzo niskim stopniem podobieństwa sekwencji, ale także nie posiadają uporządkowanej struktury przestrzennej. Wyjątkowo zmienny charakter tych domen sprawia zatem, że ich identyfikacja i klasyfikacja wykraczają również poza ramy klasycznej bioinformatyki, której metody adnotacji - tj. przypisywania funkcji regionom sekwencji - opierają się w dużej mierze na ilościowym podobieństwie do znanych już funkcjonalnie spokrewnionych sekwencji. W związku z tym w praktyce badawczej, nieskuteczne okazują się przeszukiwania podobnych sekwencji z wykorzystaniem tradycyjnych programów opartych na przyrównywaniach sekwencji (np. BLAST [3]) czy algorytmach rozpoznających motywy (PSI-BLAST, HMMER [4]).

Z uwagi na wysoki poziom dywergencji sekwencji domen WG/GW, ich przyrównywanie uniemożliwia wiarygodne określenie ich relacji homologicznych. Toteż mimo zwiększającej się liczby zidentyfikowanych białek o doświadczalnie potwierdzonej aktywności wiązania AGO,

nadal bez odpowiedzi pozostają pytania dotyczące powstawania i różnicowania tych ekstremalnie zmiennych, a mimo to funkcjonalnych domen białkowych. Ponadto domena WG/GW, ze względu na niejasne kryteria klasyfikacji, nie została zdefiniowana w publicznie dostępnych bazach danych motywów i domen białkowych (np. Pfam [1], PROSITE [5], InterPro [6]). Dodatkowo utrudnione jest znalezienie pełnej listy białek wchodzących w interakcję z białkami AGO w oparciu o główne bazy danych sekwencji białkowych (UniProt, RefSeq), ponieważ domena WG/GW występuje w wielu niespokrewnionych rodzinach białkowych charakterystycznych tylko dla niektórych grup systematycznych, np. białko Tas3 występujące jedynie u drożdży z gatunku *S. pombe* [7,8], podjednostka NRPE1 polimerazy V u roślin wyższych [9] czy białka WAG1 i CnjB orzęska *T. thermophila* [10].

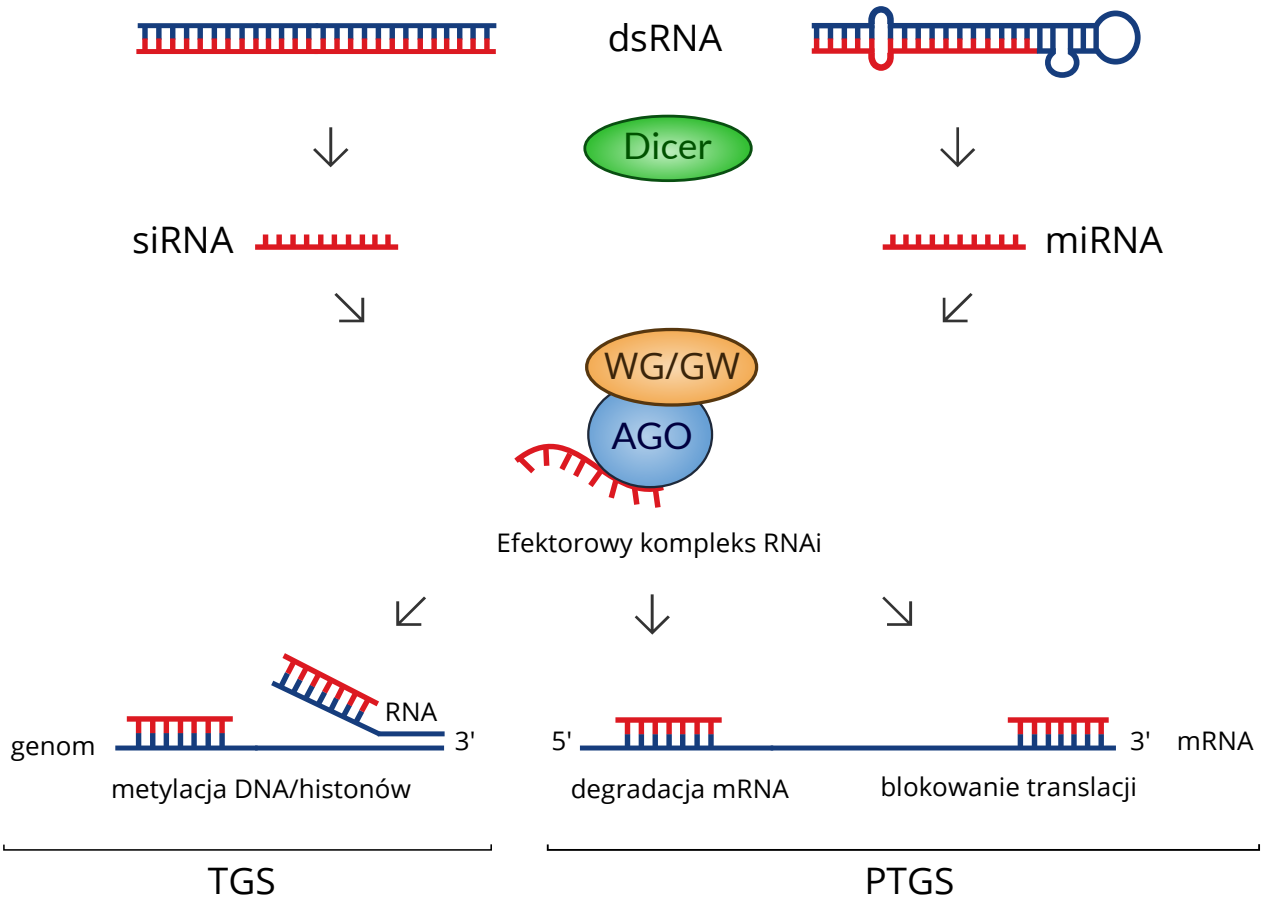
W niniejszej pracy zaprezentowane zostaną trzy podejścia komputerowe przeznaczone do przewidywania *de novo* domen wiążących białka z rodziny AGO w zadanym przez użytkownika zestawie sekwencji. Wykorzystanie opracowanego oprogramowania umożliwia identyfikację - wraz z oceną wiarygodności przewidywań - genów eukariotycznych, które kodują funkcjonalne motywy oddziałujące z białkami AGO. Następnie przedstawiona zostanie analiza porównawcza białek wiążących AGO, podczas której zbadane zostaną związki filogenetyczne kodujących je genów oraz mechanizmy molekularne zapewniające wysokie zróżnicowanie występujących w nich domen WG/GW. W ostatniej części tej pracy zostanie zaprezentowany publicznie dostępny portal internetowy, który z jednej strony stanowi repozytorium informacji na temat krótkich motywów zawierających tryptofan zaangażowanych w proces RNAi, a jednocześnie oferuje społeczności naukowej komplet sieciowych aplikacji wspomagających eksperymentalne badania nad tego typu domenami.

1.1. Białka WG/GW w procesie RNAi

Interferencja RNA jest naturalnym mechanizmem funkcjonującym w komórkach eukariotycznych polegającym na regulacji ekspresji genów przy udziale małych regulatorowych cząsteczek RNA (srRNA, ang. *small regulatory RNA*) o sekwencji identycznej lub podobnej do sekwencji DNA docelowego genu. Spośród wielu klas srRNA, rozróżnia się dwa główne typy tych cząsteczek: mikroRNA (miRNA, ang. *microRNA*) oraz małe interferujące RNA (siRNA, ang. *small interfering RNA*) [11]. Cząsteczki miRNA są krótkimi jednoniciowymi RNA kodowanymi przez genom komórki, które odpowiadają za regulację ekspresji genów podczas rozwoju i funkcjonowania organizmu. Funkcja ta zwykle realizowana jest przez cząsteczki miRNA wykazujące częściową identyczność do docelowej komplementarnej sekwencji mRNA [12]. Natomiast małe siRNA powstają z długich dwuniciowych cząsteczek (dsRNA, ang.

double-stranded RNA) syntetyzowanych przez polimerazę RNA zależną od RNA (RdRP, ang. *RNA-dependent RNA polymerase*) na jednoniciowej matrycy RNA transpozonów, elementów powtarzalnych oraz niektórych wirusów. Czynnikiem decydującym o uruchomieniu takiego obronnego mechanizmu przeciwwirusowego i kontroli ekspresji materiału genetycznego zawartego w transpozonach jest całkowita komplementarność interferującego RNA do sekwencji podlegającej wyciszeniu [13].

Zarówno siRNA jak i miRNA są produktem katalitycznej aktywności endorybonukleazy Dicer, która rozcina dwuniciowe cząsteczki RNA (rys. 1). Tak powstałe niskocząsteczkowe miRNA/siRNA są bezpośrednimi mediatorami kierującymi procesem RNAi, który może działać na dwóch poziomach: (i) DNA, poprzez transkrypcyjne wyciszanie genów (TGS, ang. *transcriptional gene silencing*) oraz (ii) RNA, na drodze post-transkrypcyjnego wyciszania genów (PTGS, ang. *post-transcriptional gene silencing*). Kontrola ekspresji genów na poziomie transkrypcji została zaobserwowana u drożdży, roślin oraz muszek owocowych i odbywa się



Rys. 1. Schemat szlaków RNAi w komórce. Krótkie regulatorowe RNA (srRNA), siRNA, miRNA, są produktem katalitycznej aktywności endorybonukleazy Dicer, który rozcina dwuniciowe cząsteczki RNA (dsRNA). sRNA wraz z AGO i bialkiem zawierającym domenę WG/GW tworzą rdzeń wielopodjednostkowych kompleksów efektorowych, np. RITS (ang. *RNA-induced transcriptional gene silencing*) lub RISC (ang. *RNA-induced silencing complex*). Kompleksy mogą działać na dwóch poziomach: (i) transkrypcyjnym (TGS), indukując wyciszanie epigenetyczne towarzyszące modyfikacji chromatyny, oraz (ii) potranskrypcyjnym (PTGS) degradując komplementany mRNA lub blokując jego translację.

przez epigenetyczne modyfikacje materiału genetycznego. Interferujący RNA oddziałując z RdRP i z metylotransferazą histonową promuje metylację histonów prowadząc do wyciszenia centromerowego DNA i/lub formowania heterochromatyny [14]. Z kolei efektem regulacji na poziomie potranskrypcyjnym może być enzymatyczne rozcięcie i degradacja mRNA lub zahamowanie translacji wynikające z bezpośredniego wiązania mRNA przez kompleks efektorowy RNAi [15], choć w pewnych warunkach oddziaływanie z miRNA prowadzić może do wzmożenia translacji [12].

Różne mechanizmy wyciszenia docelowego mRNA, operujące zarówno w szlaku TGS, jak i PTGS, determinowane są rodzajem efektorowego kompleksu białkowo-rybonukleinowego (RNP, ang. *ribonucleoprotein*), z którym dana klasa srRNA oddziałuje. Na przykład, rozcięcie docelowej nici RNA odbywa się z udziałem kompleksu RISC (ang., *RNA-induced silencing complex*), zahamowanie translacji wymaga kompleksu miRNP (ang. *microribonucleoprotein*), a regulacja ekspresji przez wpływ na strukturę chromatyny jest realizowana przez kompleks RITS (ang. *RNA-induced transcriptional gene silencing)* [16]. Rdzeń tych kompleksów stanowią białka z rodziny Argonaute (AGO) związane z cząsteczkami srRNA oraz białkami zawierającymi domenę WG/GW.

Członkowie rodziny AGO posiadają trzy charakterystyczne, zachowane ewolucyjnie domeny - sąsiadującą z domeną N-końcową domenę PAZ oraz środkową MID, odpowiadające za wiązanie odpowiednio 3’ i 5’ końca interferującego RNA oraz domenę C-końcową, PIWI, wykazującą aktywność RNazy H [17]. W toku ewolucji białka Argonaute podlegały licznym duplikacjom, szczególnie u roślin i zwierząt, które prowadziły do specjalizacji ich funkcji. Na przykład w genomie Arabidopsis, który koduje 10 przedstawicieli AGO, regulacja ekspresji na drodze miRNA zależy głównie od udziału AGO1, a także AGO2 i AGO5, natomiast cząsteczki siRNA zostają przyłączone do białek AGO4 lub AGO6 prowadząc do metylacji *de novo* sekwencji DNA w procesie metylacji DNA kierowanej przez RNA (RdDM, ang. *RNA-directed DNA methylation*). W badaniach *D. melanogaster*, zawierającej dwóch przedstawicieli rodziny AGO, wykazano, że miRNA są preferencyjnie wiązane przez AGO1 wpływając na supresję translacji, natomiast siRNA uruchamiają degradację mRNA dzięki oddziałującemu z nimi AGO2 [18]. U ssaków, u których rodzina Argonaute obejmuje czterech członków, siRNA są preferencyjnie wiązane przez AGO1 lub AGO2, natomiast miRNA wykorzystywać może każde z czterech AGO [19]. Z kolei genom *C. elegans* koduje przynajmniej 26 białek Argonaute, z których ALG-1 i ALG-2 wiążą miRNA, natomiast SAGO-1, SAGO2, RDE-1 oddziałują z siRNA. W przypadku drożdży *S. pombe*, w szlakach TGS i PTGS zaangażowany jest jeden przedstawiciel białek AGO, podobnie jak u orzęska *T. thermophila*, który posiada jedną kopię białka Argonaute (Twi1) zaangażowaną w rearanżację materiału genetycznego prowadzącą do eliminacji części sekwencji DNA [17].

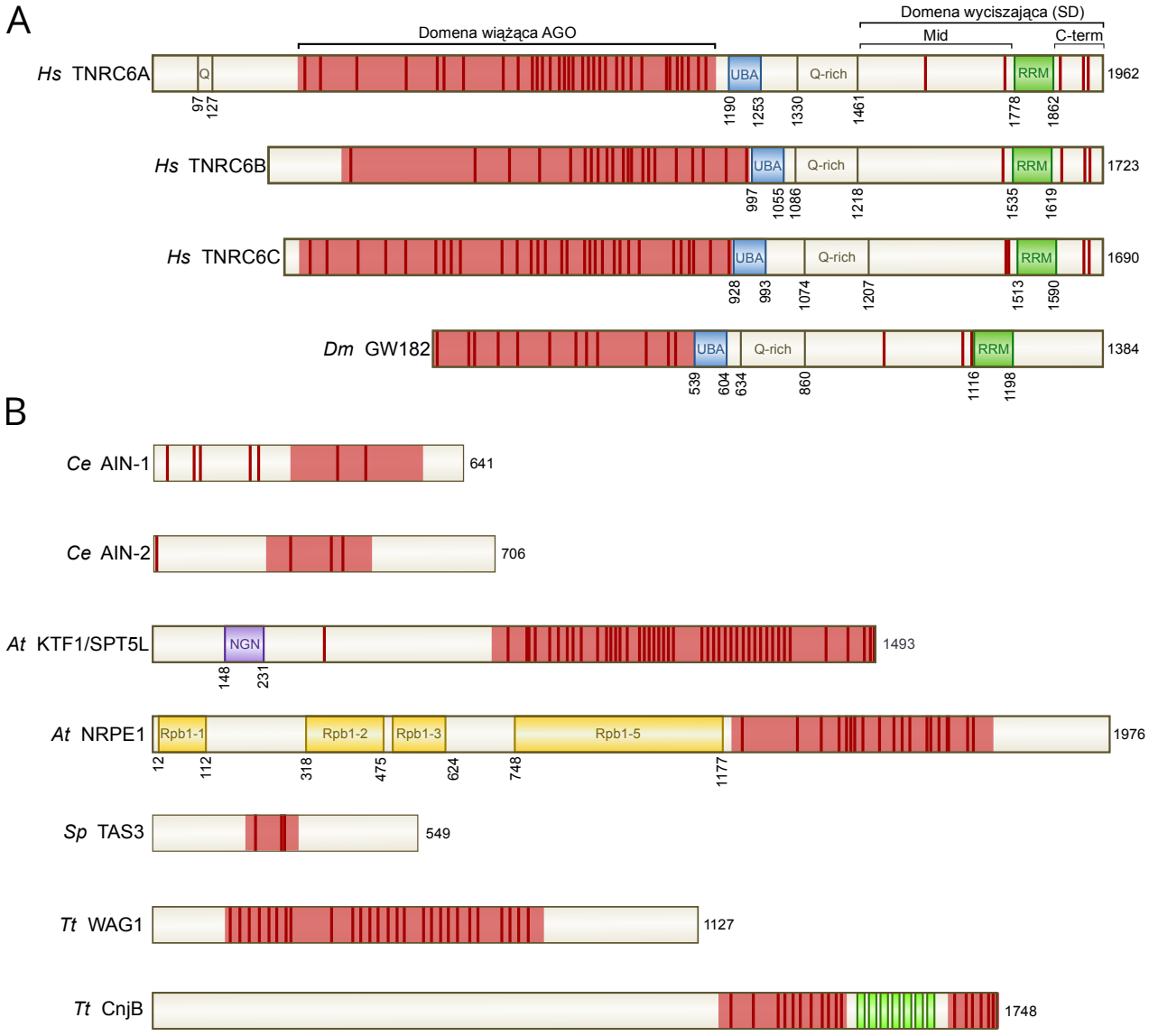
Jednak bez względu na rodzaj białka AGO, wspólną cechą efektorowych kompleksów RNAi jest obecność białek zawierających domenę WG/GW bezpośrednio związaną z domeną PIWI różnych przedstawicieli rodziny AGO. Domena WG/GW złożona z funkcjonalnych powtórzeń Trp i Gly stanowi przedmiot niniejszej analizy i będzie szczegółowo opisana w dalszych częściach pracy. Terminy: motyw i powtórzenie WG/GW, odnoszą się do wystąpienia takiego układu aminokwasów w sekwencji białka, zaś domena rozumiana jest jako dłuższy odcinek sekwencji wyodrębniony ze względu na niezależną od reszty białka zdolność zachowania aktywności wiązania białek AGO.

1.2. Rola domeny WG/GW w wiązaniu białek z rodziny Argonaute

Powtórzenia WG/GW zaobserwowano po raz pierwszy w ludzkich komórkach w białku GW182, które zidentyfikowano jako antygen rozpoznawany przez surowicę pacjenta chorego na neuropatię motoryczną i sensoryczną [20]. Jednak ich związek ze szlakiem RNAi został odkryty w kolejnych badaniach doświadczalnych, podczas których białka GW182 izolowano w kompleksach z białkami AGO lub wykazywano ich kluczowe znaczenie w szlakach regulacyjnych realizowanych z udziałem miRNA u zwierząt. Eksperymenty te obejmowały badania genetyczne u *C. elegans*, badania przesiewowe za pomocą RNAi (ang. *RNAi screening*) u *D. melanogaster* oraz oczyszczanie i analizę biochemiczną kompleksów zawierających białka AGO pochodzące z komórek człowieka [21–25]. Rodzina GW182 składa się z trzech paralogów genu GW182 (TNRC6A/GW182, TNRC6B i TNRC6C) u kręgowców i niektórych owadów [26], z wyjątkiem muszki owocowej, która koduje jedną kopię tego genu (DmGW182) [27]. Charakterystyczną cechą białek GW182 jest obecność dwóch dobrze zachowanych domen o zdefiniowanej strukturze przestrzennej obejmujących: centralnie położona domenę UBA (ang. *ubiquitin-associated domain*) i domenę wiążąca RNA (RRM, ang. *RNA-recognition motif*) zlokalizowaną na C-końcu białka (rys. 1A). Te dwie globularne domeny otoczone są regionami sekwencji, które na podstawie analiz komputerowych przewidziane zostały jako fragmenty nieustrukturyzowane [25,27]. Sekwencje te obejmują trzy części białka, N-koniec (N-term), środek (Mid) i C-koniec (C-term), zawierające liczne powtórzenia par WG/GW, a także region bogaty w reszty glutaminy (Q-rich) znajdujący się między domenami UBA i RRM [20,25,27,28]. Chociaż liczba motywów WG/GW znajdujących się w regionach N-, Mid- i C- różni się w obrębie paralogów grupy białek GW182, najwięcej powtórzeń występuje w regionie N-końca u wszystkich przedstawicieli tej rodziny, podczas gdy regiony Mid i C-term zawierają znacznie mniej lub pozbawione są powtórzeń par tryptofanu i glicyny. Metody koimmunoprecypitacji przeprowadzone w komórkach *D. melanogaster* wykazały, że obecność N-końcowego regionu

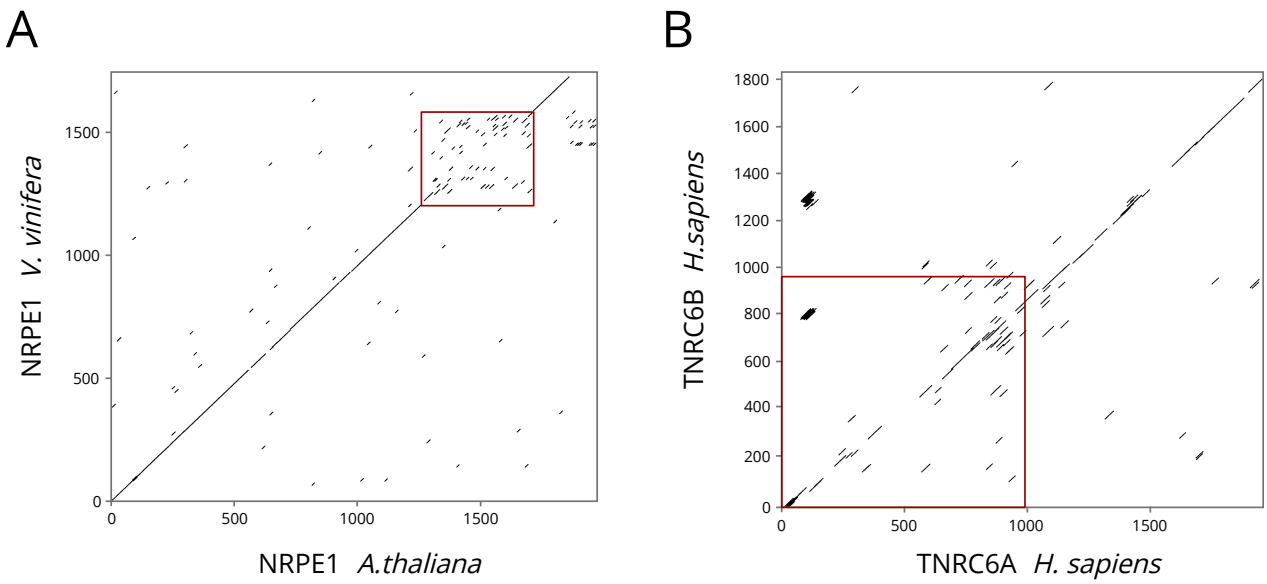
białka DmGW182 jest zarówno krytyczna, jak i wystarczająca do bezpośredniej asocjacji z białkami AGO [25]. Zastosowanie ukierunkowanej mutagenezy w celu oceny wkładu poszczególnych reszt aminokwasów w oddziaływanie z AGO pozwoliło zidentyfikować motywy WG/GW jako miejsca kontaktu, wewnątrz których tryptofan odgrywa krytyczną rolę [7,9,29–31]. Obecność i ułożenie pewnych reszt aminokwasowych znajdujących się w lokalnym otoczeniu zdefiniowanego motywu WG/GW ma również wpływ na stabilność kompleksu GW182-AGO. Nie wszystkie dwuliterowe powtórzenia tryptofanu i glicyny wykazują bowiem jednakową specyficzność podczas asocjacji [30,32]. Ponadto mutageneza reszt znajdujących się w najbliższym otoczeniu motywów WG/GW wpływała na siłę wiązanie białek AGO [7].

Obok rodziny GW182, w tym samym czasie domenę wiążącą AGO zaobserwowano



Rys. 2. Architektura domen w doświadczalnie potwierdzonych białkach wiążących AGO u Eukayota. Domena WG/GW oznaczona została czerwonym prostokątem. Wystąpienia motywów WG/GW wyróżniono pionową linią. **A.** Rodzina GW182 (TNRC6A, TNRC6B, TNRC6C). Hs (*H. sapines*), Dm (*D. melanogaster*). **B.** Inne rodziny białkowe. Ce (*C.elegans*), At (*A. thaliana*), Sp (*S. pombe*), Tt (*T. thermophila*).

u Arabidopsis w obrębie C-końca podjednostki NRPE1 polimerazy V (polV) (rys. 3A). Ta domena, podobnie jak w przypadku białka GW182, składa się z licznie występujących powtórzeń WG/GW, które tworzą molekularną platformę niezbędną do wiązania białek AGO4 podczas szlaku RdDM. Mutanty polV wewnątrz powtórzeń WG/GW, w których jeden tryptofan został zastąpiony różnymi pod względem fizykochemicznym resztami aminokwasowymi, np. znacznie mniejszą alaniną czy fenyloalaniną będącą dużym aromatycznym aminokwasem o podobnych właściwościach do tryptofanu, są niezdolne do wiązania białek AGO4, co w konsekwencji prowadzi do obniżenia poziomu metylacji DNA elementów powtarzalnych [9]. Sekwencja domeny wiążącej AGO w białku NRPE1 oprócz konserwatywnych powtórzeń glicyny i tryptofanu wykazuje niski stopień podobieństwa nawet między blisko spokrewnionymi roślinami np. *A. thaliana* i *V. vinifera* (rys. 3A). Tak wysoki poziom kumulowania mutacji w obrębie domeny WG/GW między ortologicznymi sekwencjami NRPE1 oraz paralogicznymi sekwencjami GW182 (rys. 3B) przy jednoczesnym zachowaniu wysokiego podobieństwa sekwencji pozostałych części białka sugeruje, że domena wiążąca AGO może podlegać silnej presji selekcyjnej.



Rys. 3. Porównanie metodą dot-matrix sekwencji białek wiążących AGO. Mało zachowany region odpowiadający domenie WG/GW został obramowany czerwonym prostokątem. **A.** Ortologi podjednostki pol V NRPE1 *A.thaliana* i *V.vinifera*. **B.** Paralogi TNRC6A i TNR6B *H. sapiens*.

W celu zbadania wpływu mutacji na funkcjonowanie domeny wiążącej AGO, El-Shami wraz z zespołem (2007) przeprowadził spektakularny eksperyment polegający na wymianie domen WG/GW pomiędzy dwoma niespokrewnionymi białkami NRPE1 Arabidopsis, a GW182 człowieka. Powstałe w tym doświadczeniu chimeryczne białka, mimo braku jakiekolwiek podobieństwa sekwencji oprócz par WG/GW, były nadal zdolne do wiązania zarówno AGO2 człowieka, jak i AGO4 Arabidopsis [9]. W tym samym czasie odnotowano także funkcjonalne

zachowanie motywów WG/GW między organizmami, które dzieli jeszcze większy dystans ewolucyjny: między białkiem AGO prokariotycznego archeonu *A. fulgidus*, a domeną wiążącą AGO ludzkiego białka GW182 [7].

Od momentu odkrycia dwójkowego kodu aminokwasów WG/GW w rodzinie GW182 i NRPE1, funkcjonalne powtórzenia WG/GW zostały zaobserwowane również w innych rodzinach białkowych we wszystkich czterech królestwach Eukariota: protista, grzyby, rośliny i zwierzęta (tabela 1). U orzęska *Tetrahymena thermophila* dwa białka, WAG1 i CnJB, należące do różnych rodzin białkowych (rys 2B) zawierają powtórzenia WG/GW, które wiążą białko z rodziny Argonaute (Twi1) podczas metylacji reszty lizyny histonów H3 [10]. U drożdży, wchodząca w skład kompleksu RITS podjednostka Tas3 posiada domenę długości 50 aminokwasów zawierającą dwa powtórzenia, WG i GWG, które bezpośrednio oddziałują z białkiem AGO1 prowadząc do transkrypcyjnego wyciszenia rejonów okołocentromerowych. Doświadczenie ukierunkowanej mutagenezy w obrębie jednego z dwóch powtórzeń WG, polegające na zamianie tryptofanu na alaninę lub fenyloalaninę i glicynę na alaninę, wykazało, że pojedynczy motyw jest niezbędny i wystarczający do asocjacji białka Tas3 z AGO1 [7,8].

Tabela 1. Eksperymentalnie potwierdzone białka wiążące AGO.

Organizm	Wirus*	Białko (motywy**)	Funkcja	AGO	Literatura
A. thaliana		NRPE1 (17)	RdDM	AGO4,6,9	[9,33]
A. thaliana		SPT5/KTF1 (44)	RdDM	AGO4	[34,35]
A. thaliana		SPT6/GTB1(12)	-	-	[9]
A. thaliana	TCV	P38 (2)	mimikra- wiązanie AGO	AGO1,4	[36]
S. tuberosum	SPMMV	P1 (3)	mimikra- wiązanieAGO1	AGO1	[37]
N. benthamiana	TSWV	NSs (1)	mimikra- wiązanieAGO1	AGO1	[38]
H. sapiens		TNRC6A (39), TNRC6B (34), TNRC6C (34)	PTGS	AGO1,2,3,4	[27]
C.elegans		AIN1/AIN2 (7/4)	PTGS	AGO1	[39]
D. melanogaster		DmGW182 (15)	PTGS	AGO1	[25,29]
T. thermophila		WAG1 (27)	rearanżacja genomu	Twi1	[10]
T. thermophila		cnjB (18)	rearanżacja genomu	Twi1	[10]
S.pombe		Tas3 (3)	TGS	AGO1	[7,8]

*TCV (ang. *Turnip crinkle virus*), SPMMV (ang. *Sweet Potato mild mottle virus*), TSWV (ang. *Tomato spotted wilt virus*) **liczba motywów WG/GW w białku

Z kolei u *A. thaliana* czynnik elongacyjny transkrypcji KTF1/SPT5, oprócz domeny KOW zawiera na swoim C-końcu ponad 40 powtórzeń WG/GW rozciągających się w sekwencji na

długość 700 aminokwasów, a mutacje w jego obrębie uniemożliwiają wiązanie białek AGO4 [34,35]. Genom nicienia *C. elegans* koduje dwa białka, AIN-1 i AIN-2, zawierające domenę składającą się odpowiednio z 7 i 4 powtórzeń motywów WG/GW, które oddziałują z białkami AGO oraz są kluczowe w szlaku miRNA [21,39,40].

W ostatnim czasie odnotowano również trzy przypadki wirusów infekujących rośliny, które kodują białka mogące funkcjonalnie upodobnić się do domen WG/GW gospodarza i dzięki temu przełamać jego naturalny system obronny i zainicjować infekcję [36–38]. Na przykład białko P38 kapsydu wirusa TCV (ang. *Turnip crinkle virus*) używa powtórzeń WG/GW jako przynęty rekrutującej białka AGO, przełamując w ten sposób system obronny RdDM Arabidopsis [36]. Również w białku proteazy serynowej P1 wirusa SPMMV (ang. *Sweet potato mild mottle virus*) powtórzenia WG/GW są niezbędne podczas wiązania i supresji białek AGO1 [37]. Niedawno wykazano, że mutacja w obrębie pojedynczego motywu WG/GW białka supresorowego NSs wirusa TSWV (ang. *Tomato spotted wilt virus*) całkowicie pozbawia funkcji supresorowych tego białka, co sugeruje potencjalną interakcję między motywem WG/GW a AGO1 [38]. Sekwencje WG/GW stanowią zatem uniwersalne narzędzie wykorzystywane przez komórkę do rekrutacji i wiązania białek AGO podczas realizacji różnych procesów RNAi zachodzących w organizmach znajdujących się na różnych poziomach organizacji życia.

1.3. Specyfika domen WG/GW

Domeny wiążące AGO wyróżniają się pięcioma charakterystycznymi właściwościami, które sprawiają, że procedura ich identyfikacji stanowi bardzo trudny element adnotacji oraz analiz porównawczych sekwencji białkowych.

Po pierwsze, poziom identyczności sekwencji domen WG/WG, nawet w obrębie blisko spokrewnionych organizmów (rys. 3), mieści się w zakresie od 20% do 35%. Ten przedział nazywa się "strefą mroku" przyrównań sekwencji (ang. *twilight zone*) [41], gdzie spokrewnione sekwencje mieszają się z niespokrewnionymi sekwencjami, których podobieństwo jest przypadkowe. Z kolei stopień identyczności sekwencji domen WG/GW pochodzących z różnych rodzin białkowych spada poniżej 20%, tym samym przenikając do tzw. "strefy ciemności" (ang. *midnight zone*), gdzie większość spośród przyrównywanych sekwencji całkowicie nie jest ze sobą spokrewniona. Uniemożliwia to wiarygodne określenie relacji homologicznych białek oddziałujących z AGO [41]. Z tych właśnie względów nie można odnaleźć pokrewieństw funkcjonalnych białek wiążących AGO wśród wyników prostych przeszukań baz danych programami BLAST [3] lub FASTA [42]. Brak możliwości wyznaczenia wiarygodnego przyrównania dwóch sekwencji domeny WG/GW stanowi poważne ograniczenie komputerowej

procedury klasyfikacji motywów i domen białkowych. U podstaw tych metod, leży bowiem założenie, że w obrębie dopasowań wielosekwencyjnych można identyfikować konserwatywne odcinki sekwencji, których istnienie ma uzasadnienie strukturalne lub funkcjonalne. Takie części dopasowań służą następnie jako cechy diagnostyczne danego zestawu sekwencji, które z kolei zostaną wykorzystane do wykrywania nowych członków odpowiednich rodzin białek. A zatem w przypadku białek WG/GW nieskuteczne okazują się klasyczne metody przewidywań domen i motywów, które obejmują stosowanie: (i) pojedynczych motywów zapisanych w formie wyrażeń regularnych (np. PROSTE [5]), (ii) wielu motywów reprezentowanych w postaci "śladów sekwencyjnych" rodzin białek (ang. *fingerprint*) (np. PRINTS [43], BLOCKS) czy całych domen reprezentowanych jako profile (PSI-BLAST, PROSITE), modele HMM (PFAM [44], InterPro [6]).

Drugi problem identyfikacji domen wiążących AGO wynika z diametralnie zróżnicowanych długości ich sekwencji, które w potwierdzonych białkach wahają się od 22 do ponad 700 aminokwasów. Brak jednoznacznie zdefiniowanej długości domeny nie uzasadnia również w tym przypadku zastosowania metod, które zamiast przeprowadzania przyrównania sekwencji, wykorzystują algorytmy kombinatoryczne np. próbkowanie Gibbsa [45] (Gibbs Motif Sampler [46]) lub maksymalizację wartości oczekiwanej (MEME [47]) oraz rozmaite warianty podejść dystansowych (ang. *alignment-free methods*) opierające się na analizie składu krótkich wyrazów sekwencyjnych [48]. Ponieważ programy te wymagają dysponowania informacjami o długościach domen, ich zastosowanie podczas identyfikacji domen WG/GW nie przynosi zadowalających rezultatów, nawet pomimo zastosowania technik przesuwającego się okna, z uwzględnieniem różnych jego wielkości.

Po trzecie, jak dotąd nieokreślona jest liczba powtórzeń WG/GW definiująca domenę wiążącą AGO. Biorąc pod uwagę fakt, że w białkach o potwierdzonej funkcji wiązania, liczba wystąpień WG/GW waha się od dwóch w domenie białka Tas3 do 45 kopii u SPT5/KTF1, szukanie białek, w których występują powtórzenia WG/GW również nie jest skutecznym rozwiązaniem. Mimo, że tryptofan jest najrzadziej występującym aminokwasem w białkach eukariotycznych (1.78% [49]), to sekwencja WG lub GW pojawia się średnio w co drugim białku. Tak więc samo rozpoznanie motywów WG/GW nie jest wiarygodne, z uwagi na ich niską specyficzność, a przy braku dodatkowych informacji sam fakt wystąpienia ich w sekwencji niczego nie oznacza, bowiem motyw jest zbyt krótki, by mógł być specyficzną cechą pozwalającą odróżnić białka wiążące AGO od innych sekwencji.

Czwartym utrudnieniem procesu identyfikacji domen WG/GW jest ich występowanie w wielu niespokrewnionych rodzinach białkowych charakteryzujących się odmienną architekturą domen (rys. 2). Brak przesłanek dotyczących jakichkolwiek korelacji współwystępowania domen

WG/GW w sąsiedztwie innych motywów uniemożliwia również prowadzenie przeszukiwań baz sekwencji w oparciu o inne dobrze zakonserwowane fragmenty sekwencji białka. Z tego względu również zastosowanie ostatnio rozwijanych programów przeznaczonych do odnajdywania białek o najbardziej podobnej architekturze domen (RADS/RAMPAGE [50] DoMosaics [51]) jest nieskuteczne w przypadku białek wiążących AGO.

Ostatnia, piąta charakterystyczna właściwość domen WG/GW utrudniająca proces adnotacji białek, wynika z faktu, że domeny wiążące AGO nie posiadają uporządkowanej struktury przestrzennej. Przewidywania elementów struktury drugorzędowej białek GW182 sugerują, że motywy WG/GW występują w regionach nieustrukturyzowanych. Uniemożliwia to prowadzenie porównań strukturalnych tych białek, a także klasyfikację domen na podstawie odpowiedniej klasy zwojów.

A zatem na podstawie opisanych powyżej pięciu cech domen wiążących AGO, stanowiących źródło problemów ich identyfikacji i analiz porównawczych, konieczne stało się utworzenie nowych metod bioinformatycznych które pozwalają uzyskać bardziej wiarygodne adnotacje domen WG/GW.

ROZDZIAŁ 2

Cel pracy

Celem niniejszej pracy jest realizacja trzech głównych zadań:

1. Opracowanie oraz implementacja nowej metody obliczeniowej, która umożliwi identyfikację i adnotację domen WG/GW.

2. Zaprojektowanie i przeprowadzenie analizy porównawczej białek wiążących AGO, która pozwoli poznać specyficzne mechanizmy molekularne odpowiedzialne za powstawanie i różnicowanie się domen WG/GW.

3. Stworzenie ogólnie dostępnego portalu internetowego poświęconego badaniom nad domenami zaangażowanymi w procesy RNAi, który oprócz systemu informacji na temat białek wiążących AGO, zawierać będzie komplet aplikacji internetowych umożliwiających użytkownikom analizowanie i przewidywanie domen WG/GW.

ROZDZIAŁ 3

Metody

3.1. Metody identyfikacji domen WG/GW

Metody identyfikacji domen wiążących AGO zostały zaimplementowane w języku programowania Python. Większość operacji numerycznych, np. obliczanie wartości punktacji przewidywanej domeny lub wyznaczanie wartości prawdopodobieństw dla odpowiedniej oceny punktacji, przeprowadzono przy użyciu dwóch bibliotek numpy (1.6.1) i scipy (0.9), przeznaczonych do obliczeń matematycznych i zastosowań naukowych [52].

3.1.1. Metoda kompozycyjna I i II generacji

Zestaw sekwencji źródłowych

Zestaw sekwencji, który posłużył do budowy matryc *dos* i *ics* obejmował w pierwszej wersji metody 26 białek roślinnych zawierających domenę WG/GW: NRPE1 u *A. thaliana* (NCBI GI: 79571777), *V. vinifera* (225465870), *S. lycopersicum* (68300841), *S. oleracea* (59939212), *S. lycopersicum* (68300841), *O. sativa* (222622188; błędnie opisane w bazie jako podjednostka pol II), *P. patens* (168027477), *Z. mays* (zidentyfikowana przez TBLASTN na sekwencji genomowej), Arabidopsis SPT5 (15237667) i SPT6 (42562972). W drugiej wersji metody zestaw źródłowy został rozszerzony do 38 sekwencji uwzględniając dodatkowe białka eukariotyczne o potwierdzonej funkcji wiązania AGO: WAG1 (213054510) i cnjB (161752) *T. thermophila*, Tas3 *S. pombe* (19112427), WGRP1 *A. thaliana* i GW182 *D. melanogaster* (24638679), człowieka (116805348, 241982729, 119609886) i ortologów ssaków: *B. taurus* (119916998, 194676322), *M. mulatta* (109094291), *E. caballus* (194219119), *P. troglodytes* (114661685), *C.*

familiaris (73964979). Sekwencje zostały znalezione w publicznie dostępnych bazach danych w oparciu o program PSI-BLAST oraz procedurę dwukierunkowego przeszukiwania BLAST (ang. *reciprocal best-hits BLAST*) używaną podczas identyfikacji ortologów [53].

Przygotowanie matrycy punktacji dos

Macierz punktacji *dos* została obliczona w oparciu o analizę składu aminokwasowego w zestawie źródłowych sekwencji i następnie użyta została podczas automatycznej detekcji przewidywania miejsca początku i końca domeny w sekwencji zapytania. Macierz punktacji w pierwszej generacji metody zawiera wartości logarytmów ilorazu szans (ang. *log odds ratio*) częstości występowania każdego aminokwasu. Wartości te odzwierciedlają stosunek szans wystąpienia danego aminokwasu w domenie WG/GW do wystąpienia tego samego aminokwasu w innej części białka nie wykazującej aktywności wiązania AGO. Logarytm ilorazu szans wyrażony jest wzorem:

$$D_i = 2 \times \log\left(\frac{P_{id}}{P_{ip}}\right) \quad (1)$$

gdzie:

i - dany aminokwas,

P_{id} i P_{ip} - częstości występowania aminokwasu i odpowiednio w sekwencjach domeny oraz pozostałych częściach białka.

W drugiej wersji metody macierz *dos* zawiera wartości *log odds ratio* częstości występowania wszystkich 400 możliwych kombinacji dipeptydów obecnych w domenie WG/GW w porównaniu do częstości pojawiania się tych sekwencji w odpowiadających im proteomach.

Obliczanie punktacji ics

Na procedurę obliczenia wartości punktacji *ics* składają się dwa kroki: (i) utworzenie tablicy zawierającej wartości *log odds ratio* dla wszystkich 200 kombinacji par aminokwasów w analizowanej domenie, oraz (ii) obliczenie różnic tych wartości między tablicą uzyskaną dla rzeczywistych domen wiążących AGO, a tablicą obliczoną dla analizowanej domeny. Końcowa wartość *ics* jest sumą wartości bezwzględnych wszystkich 200 różnic. W odróżnieniu od parametru *dos*, gdzie wyższe wartości reprezentują wyższą ocenę przewidywanej domeny, w przypadku *ics*, wartości bliższe zeru wskazują na większą zgodność zależności między aminokwasami w stosunku do sekwencji domen WG/GW o potwierdzonej funkcji. Wartości zawarte w macierzy *ics* wyznaczone dla każdego aminokwasu opisane są wzorem:

$$I_{ij} = \left| \log_2 \left(\frac{N_i}{N_j} \right) \right| \quad (2)$$

gdzie:

i, j - dwa aminokwasy znajdujące się w domenie,

N_i, N_j - iczba aminokwasów *i* oraz *j* znajdujących się w domenie.

Ocena istotności statystycznej punktacji dos

Modelowanie funkcji opisującej rozkład prawdopodobieństw punktacji *dos* przeprowadzono w programie EasyFit firmy Mathwave Technologies [54]. Procedura ta obejmowała analizę dopasowania empirycznych rozkładów punktacji *dos* (ang. *Distribution fitting*) do ponad 50 modeli teoretycznych, oraz ocenę zgodności każdego z dopasowań (ang. *Goodness of fit*). Test Kołmogorowa-Smirnowa (KS) [55] został wykorzystany do sprawdzenia odległości empirycznego rozkładu wartości *dos* z dystrybuantami teoretycznych modeli. Dodatkowo do oceny zgodności dopasowania modeli wykorzystano metody oparte na teorii informacji (SIC, AIC, HQIC) [56].

Wyznaczenie wartości granicznej ics

Źródłowy zestaw 26 sekwencji domen WG/GW został podzielony na dwie grupy: sekwencje testowe użyte do obliczenia wartości oceny *ics* oraz sekwencje referencyjne użyte do zbudowania macierzy punktacji *ics*, która z kolei posłużyła do oceny zestawu testowego. Pięć rund obliczeń zostało przeprowadzonych odpowiednio dla różnej liczby sekwencji referencyjnych w zakresie od 21 do 25. W każdej serii uwzględniano wszystkie kombinacje sekwencji. Przy użyciu liniowej regresji maksymalnych wartości *ics* uzyskanych w analizowanych pięciu punktach wyznaczono oczekiwaną wartość *ics* jako graniczną wartość identyfikującą domeny WG/GW. W drugiej generacji metody źródłowy zestaw sekwencji obejmował 38 białek, a pięć serii obliczeń prowadzonych było dla różnych kombinacji sekwencji referencyjnych w zakresie od 33 do 37.

3.1.2. Metoda profilu PSSM

Zestaw sekwencji źródłowych

Zestaw sekwencji źródłowych służący do budowy pozycyjnie-specyficznej macierzy wartościującej (PSSM, ang. *Position-specific scoring matrix*) obejmuje 195 ortologicznych białek wiążących AGO u Eukariota (http://www.comgen.pl/whub/download/files/). Z tego zbioru wyodrębniono 6999 nienakładających się sekwencji motywów zawierających pojedyncze

wystąpienie Trp. W motywach tych, długości sekwencji flankujących resztę Trp odpowiadają połowie odległości do kolejnej najbliższej reszty Trp na N- i C-końcu. W przypadku braku kolejnego wystąpienia Trp na N- i/lub C-końcu, pobrana zostaje sekwencja odpowiednio, od początku i/lub do końca białka.

Budowa macierzy PSSM

Na podstawie sekwencji motywów utworzone zostaje przyrównanie, które jest pozbawione przerw (ang. *ungapped alignment*). W takim przyrównaniu, zachowany we wszystkich sekwencjach Trp znajduje się w pozycji centralnej, a otaczające go inne reszty aminokwasowe rozchodzą się w dwóch kierunkach N- i C-końca. Następnie zliczane zostają częstości występowania poszczególnych aminokwasów na każdej pozycji krótkiego fragmentu, zgodnie ze wzorem:

$$p_{ia} = \frac{n_{ia}}{n_{seq}} \tag{3}$$

gdzie:

p_{ia} - obserwowana częstość występowania aminokwasu *a* na pozycji *i*,

n_{ia} - obserwowana liczba wystąpień aminokwasu *a* w dostępnym zbiorze motywów na pozycji *i*,

n_{seq} - liczbą sekwencji motywów.

W ten sam sposób obliczona zostaje wartość q_{ia}, rozumiana jako średnie częstości występowania aminokwasów w zestawie motywów białek eukariotycznych zawierających w centrum powtórzenie Trp. Następnie zostaje skonstruowany profil PSSM, którego komórki wypełnione są ocenami punktowymi logarytmów ilorazu szans występowania danego aminokwasu na określonej pozycji w motywie:

$$D_{ia} = 2\log_2\left(\frac{p_{ia}}{q_{ia}}\right) \tag{4}$$

Wówczas dodatnie wartości D_{ia} oznaczają, że aminokwas znajdujący się na pozycji *i* jest z większym prawdopodobieństwem elementem domeny wiążącej AGO niż fragmentem niefunkcjonalnego powtórzenia zawierającego Trp.

Identyfikacja rodzin białkowych zawierających potencjalną domenę WG/GW

Aminokwasowe sekwencje, które posłużyły, jako sekwencje zapytania podczas identyfikacji domen WG/GW u Eukariota, Prokariota i wirusów, zostały pobrane z bazy UniProt (wersja: 2013_11). Przynależność zidentyfikowanych sekwencji do odpowiednich rodzin białkowych

została wyznaczona w oparciu o podobieństwo sekwencji i składu domen białkowych. Program Cd-hit [57] został wykorzystany do utworzenia grup podobnych sekwencji, których procent identyczności wynosi powyżej 50% i obejmuje przynajmniej 40% długości krótszej sekwencji. Badanie składu domen białkowych przeprowadzane zostało w oparciu o pakiet HMMER3 [58] i bazę rodzin i domen białkowych Pfam-A (wydanie 26.0) [44]. Analiza regionów nieuporządkowanych w zidentyfikowanych białkach przeprowadzona została przy użyciu programu IUPred [59].

3.1.3. Metoda wykorzystująca nauczanie maszynowe

i-Wsearch realizowany jest przez algorytm lasów losowych (ang. *Random forest*) [60] i wykorzystuje klasyfikację binarną, która określa czy dana reszta Trp należy do klasy sekwencji wiążących AGO. Korzystając z techniki przesuwanego okna każdy aminokwas otaczający środkową resztę Trp kodowany jest za pomocą 6 właściwości fizykochemicznych: indeksu hydrofobowości [61], elastyczności [62] i hydrofilowości [63], masy, objętości [64] oraz względnej dostępności na powierzchni (w układzie G-X-G, gdzie X jest analizowanym aminokwasem) [65]. Jednocześnie dla analizowanego Trp brane są pod uwagę odległości w sekwencji do najbliższych reszt Trp na N- i C-końcu. Zatem o przynależności danej reszty tryptofanu do klasy motywów wiążących AGO decyduje kontekst fizykochemiczny oskrzydlających go sekwencji oraz kontekst występowania innych motywów sąsiadujących. Proces klasyfikacji motywów przeprowadzano w środowisku Python, korzystając z zaimplementowanego algorytmu lasów losowych, dostępnego w pakiecie do uczenia maszynowego scikit-learn [66].

Ewaluacja metody

Do wytrenowania metody użyto dwóch zbiorów uczących liczących po 6779 motywów znajdujących się w sekwencjach każdej z dwóch klas: sekwencji wiążących AGO oraz sekwencji nie wykazujących takiej aktywności. Podczas budowy klasyfikatora lasów losowych użyto 100 drzew. Skuteczność rozpoznawania motywów została zbadana w oparciu o 10-krotny sprawdzian krzyżowy (ang. *10-fold cross validation*). W technice tej zbiór uczący zostaje podzielony na 10 równych podzbiorów. Jedna z wydzielonych części stanowi zbiór testowy, natomiast zbiór obejmujący pozostałe 9 części służy do wytrenowania klasyfikatora. Algorytm dokonuje wyliczeń wskaźników trafności przewidywań. Analiza wykonywana jest 10-krotnie, a wskaźniki trafności są następnie uśredniane w celu uzyskania jednego wyniku.

Liczbowe wskaźniki skuteczności działania klasyfikatora, takie jak czułość (SN),

specyficzność (SP), trafność (ACC) i wartość F wyliczono na podstawie poniższych wzorów.

$$SN = \frac{TP}{TP + FN} \tag{5}$$

$$SP = \frac{TN}{TN + FP} \tag{6}$$

$$PPV = \frac{TP}{TP + FP} \tag{7}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$wartość\ F = \frac{2 \times SN \times SP}{SN + SP} \tag{9}$$

gdzie TP, TN, FP, FN są liczbami wyników, odpowiednio, prawdziwie dodatnich, prawdziwie ujemnych, fałszywie dodatnich i fałszywie ujemnych. Wartość F, będąca harmoniczną średnią czułości i specyficzności została wybrana, podobnie jak w innych pracach [67], jako główna miara skuteczności algorytmu.

Oceny skuteczności algorytmu przeprowadzono dla różnych wielkości przesuwającego się okna (od 3 do 25) i wszystkich kombinacji siedmiu zmiennych opisujących właściwości fizykochemiczne aminokwasów.

3.2. Analiza filogenetyczna

Sekwencje wejściowe i wybór białek z rodziny RRM

Wejściowe sekwencje aminokwasowe i nukleotydowe roślin *Arabidopsis thaliana* i *Oryza sativa* zostały pobrane odpowiednio z bazy TAIR (wersja 9) [68] i MSU Rice Genome Annoation Project Database and Resources (wersja 6.1) [69]. Z bazy DOE Joint Genome Institute (JGI) [70] uzyskano adnotacje sekwencji genomowych wraz z sekwencjami EST dla gatunków roślin: *Chlamydomonas reinhardtii* (wersja 4.1), *Chlorella NC64A* (wersja 1.0), *Ostreococcus lucimarinus* (wersja 2.0), *Micromonas RCC299* (wersja 3.0), *Volvox carteri* (wersja 2.0), *Selaginella moellendorffii* (wersja 1.0), *Physcomitrella patens* (wersja 1.1).

Do identyfikacji sekwencji białkowych zawierających domenę RRM w analizowanych proteomach roślinnych wykorzystano pakiet HMMER3 [58] i bazę rodzin i domen białkowych Pfam-A (wydanie 24.0) [44]. Do analizy filogenetycznej wybrano białka zawierające

przynajmniej jedną domenę RRM, której wartość oczekiwana była niższa lub równa od przyjętej przez kuratorów bazy Pfam wartości granicznej 0,001.

Przyrównania sekwencji i rekonstrukcja filogenetyczna

Zakres badań filogenetycznych został ograniczony do konserwatywnych fragmentów sekwencji domen RRM, gdyż sekwencje zawierające powtórzenia WG/GW są wysoce zmienne, co sprawia, że zawarta w nich informacja filogenetyczna jest mało wiarygodna. Sekwencje pojedynczego motywu RRM położonego najbliżej N-końca białka wyodrębnione zostały z sekwencji pełnej długości w oparciu o adnotacje bazy Pfam. Przyrównanie sekwencji domeny RRM przeprowadzono używając programu MAFFT 6.717 [71]. Do konstrukcji drzew filogenetycznych równolegle wykorzystano metody największej wiarygodności (ML, ang. *Maximum Likelihood*) zaimplementowane w programie PhyML 3.0 [72] oraz metody oparte na wnioskowaniu bayesowskim i symulacji Monte Carlo dla łańcuchów Markova (MCMC, ang. *Markov Chain Monte Carlo*), którą wykorzystuje aplikacja MrBayes 3.1.2 [73,74]. Dla obu metod model substytucji LG [75] został wybrany na podstawie porównania wartości największego prawdopodobieństwa (-lnL) oraz kryterium informacyjnego Akaike (AIC, AICc) określonych przy pomocy programu Prottest [76]. Model ewolucyjny w symulacji w programie MrBayes uwzględniał pozycje inwariantne oraz tempo podstawień opisane rozkładem gamma. Obliczenia zostały przeprowadzane dla 2 milionów pokoleń przy czterech łańcuchach MCMC, próbkowanie drzew następowało co 100 generacji. Analiza prowadzona była do osiągnięcia stabilnego stanu łańcucha i wymaganego poziomu odchylenia standardowego poniżej 1%.

Obliczanie tempa substytucji niesynonimicznych i synonimicznych

W celu porównania tempa zachodzenia substytucji niesynonimicznych i synonimicznych w obrębie domen RRM, GRP i WG/GW wyodrębniono odpowiadające im sekwencje aminokwasowe. W obrębie danego typu domeny, dla każdej z par przeprowadzono przyrównanie sekwencji kodujących w oparciu o wyznaczone wcześniej przyrównanie sekwencji aminokwasowych przy użyciu programu MAFFT [71] i TranslatorX [77]. Stosunek liczby substytucji niesynonimicznych (*Ka*) i synonimicznych (*Ks*) obliczono dla każdej z par sekwencji korzystając z programu KaKs_Calculator, który uwzględnia wyniki działania 10 algorytmów wykorzystujących różne modele ewolucyjne [78].

Analiza konwersji genów

Program GENECONV 1.81 został wykorzystany do identyfikacji potencjalnych zdarzeń wymiany sekwencji (rekombinacja i/lub konwersja genów i/lub nierówny crossing-over) [79].

Ze względu na wysoką liczbę substytucji niesynonimicznych, w obrębie wymienianego segmentu sekwencji uwzględniona została możliwości występowania niedopasowanych reszt. W analizie użyto domyślnych parametrów: 10.000 permutacji oraz globalna (znormalizowana ze względu na wiele porównywań) wartość $p < 0,05$. Jednocześnie, do detekcji potencjalnych zrekombinowanych sekwencji użyto algorytmów RDP, BOOTSCAN, MAXCHI, CHIMAERA, SISCAN, 3SEQ zaimplementowanych w programie RDP3 [80], Regiony sekwencji, które zostały zidentyfikowana przynajmniej przez trzy metody oraz mające potwierdzenie w rekonstrukcji filogenetycznej prowadzonej w RDP3 zostały uznane za potencjalne miejsca wymiany sekwencji.

3.3. Technologie wykorzystane w aplikacjach internetowych

Aplikacja Agos, jak i portal internetowy Whub zostały napisane w języku Python 3.2 przy przy użyciu biblioteki Django 1.5.2 [81] przeznaczonej do tworzenia aplikacji internetowych. Baza danych rekordów białek i ich adnotacji oraz publikacji zaimplementowana została w języku SQlite3 [82]. Strukturę elementów stron tworzono zgodnie z koncepcją stron responsywnych (ang. *responsive Design*) korzystając z technologii HTML5, CSS3 i systemu Twitter Bootstrap 3.0.3 [83]. Dynamiczna aktualizacja interfejsu użytkownika przeprowadzana jest przy użyciu języka JavaScript (biblioteka jQuery [84]) oraz technologii AJAX. Interaktywne wykresy generowano w oparciu o bibliotekę jQuery, HighCharts. Większość pozostałych wizualizacji (np. mapy termiczne, gra internetowa) tworzono przy pomocy technologi SVG w oparciu o bibliotekę d3.js [85].

ROZDZIAŁ 4

Wyniki

4.1. Identyfikacja *de novo* domen wiążących białka AGO

W tym podrozdziale zaprezentowane zostaną dwie nowe metody obliczeniowe służące do identyfikacji nowych białek zawierających potencjalne domeny wiążące AGO. Zastosowanie tych programów umożliwia otrzymanie list rankingowych białek zawierających potencjalną domenę WG/GW w genomach organizmów eukariotycznych, prokariotycznych oraz wirusów.

4.1.1. Metoda przewidywania domen WG/GW oparta na kompozycji aminokwasów

Podczas przeprowadzonego przez zespół El-Shamiego (2007) doświadczenia, które polegało na wymianie domen WG/GW między niespokrewnionymi białkami człowieka i Arabidopsis, zauważono, że obie domeny bogate są w reszty Gly, Ser, Try i w mniejszym stopniu Glu, Asp i Asn. Jednocześnie sekwencje obu domen wykazywały małą zawartość aminokwasów hydrofobowych: Cys, Phe, His, Met i Tyr [9]. Obserwacja ta pozwoliła opracować metodę w laboratorium prof. Karłowskiego w Pracowni Bioinformatyki Uniwersytetu im. Adama Mickiewicza w Poznaniu, która - w oparciu o analizę składu aminokwasowego sekwencji domen wiążących AGO - umożliwia zdefiniowanie wartości granicznych, pozwalających wyodrębnić funkcjonalne domeny WG/GW spośród innych białek [56].

Badanie specyficznej kompozycji aminokwasowej domen WG/GW zostało przeprowadzone na zestawie 26 sekwencji obejmujących 3 eksperymentalnie potwierdzone wówczas białka wiążące AGO u Arabidopsis - NRPE1 [9], SPT5/KTF1 [34] i SPT6/GTB1 - wraz z ich ortologami u innych roślin (patrz: Metody - Rozdział 3.1.1). Porównanie składu aminokwasowego sekwencji

domen WG/GW w tych białkach w odniesieniu do pozostałych części sekwencji przedstawiono w tabeli 2. Wartość punktacji danego aminokwasu (ang. *score*) wyraża logarytm stosunku obserwowanej częstości występowania danej reszty w domenie do oczekiwanego prawdopodobieństwa jej wystąpienia w innej części białka (patrz: Metody - Rozdział 3.1.1). Podobnie jak w macierzach substytucji BLOSUM i PAM, wartości logarytmu ilorazu szans odpowiadają preferencjom występowania pewnych reszt aminokwasowych w domenie. Wartości dodatnie i ujemne oznaczają występowanie danego aminokwasu w domenie WG/GW, odpowiednio częściej i rzadziej niż w pozostałych regionach białka pozbawionych aktywności wiązania białek AGO. Natomiast wartości zerowe wskazują, że dany aminokwas pojawiania się w domenie WG/GW z taką samą częstością jak w przypadku pozostałych sekwencji tła.

Tabela 2. Macierz punktacji domeny wiążącej białka AGO wykorzystywana przez algorytm podczas przewidywania wielkości domeny i obliczania całkowitej wartości *score*.

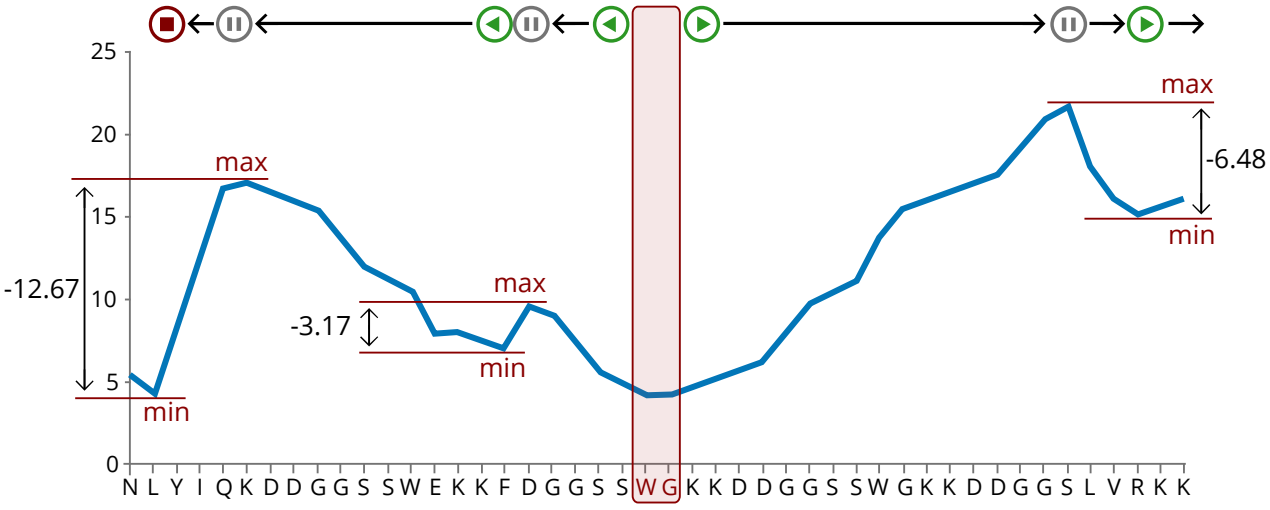
Aminokwas	Wartość punktacji [połowa bitu] ↓	Wartość punktacji [jeden bit]	Stosunek częstości	Częstość	Zliczenia
Trp (W)	2,666	1,333	2,520	0,063:0,025	743:1062
Gly (G)	2,068	1,034	2,048	0,213:0,104	2490:4447
Asn (N)	1,510	0,755	1,688	0,081:0,048	949:2051
Ser (S)	1,236	0,618	1,535	0,152:0,099	1774:4213
Ala (A)	0,280	0,140	1,102	0,065:0,059	762:2537
Asp (D)	0,184	0,092	1,066	0,081:0,076	950:3254
Thr (T)	0,000	0,000	1,000	0,040:0,040	467:1718
Gln (Q)	-0,076	-0,038	0,974	0,038:0,039	440:1686
Lys (K)	-0,120	-0,060	0,959	0,070:0,073	821:3136
Arg (R)	-0,590	-0,295	0,815	0,044:0,054	518:2319
Pro (P)	-0,644	-0,322	0,800	0,032:0,040	373:1726
Glu (E)	-1,288	-0,644	0,640	0,048:0,075	560:3219
Val (V)	-2,408	-1,204	0,434	0,023:0,053	274:2260
Phe (F)	-2,558	-1,279	0,412	0,014:0,034	169:1443
His (H)	-3,324	-1,662	0,316	0,006:0,019	76:796
Cys (C)	-3,398	-1,699	0,308	0,004:0,013	43:568
Tyr (Y)	-4,792	-2,396	0,190	0,004:0,021	50:890
Met (M)	-5,012	-2,506	0,176	0,003:0,017	39:743
Ile (I)	-5,030	-2,515	0,175	0,007:0,040	79:1705
Leu (L)	-5,252	-2,626	0,162	0,011:0,068	132:2925

Aminokwasy są uszeregowane ze względu na malejące wartości punktacji. Druga i trzecia kolumna wykorzystywane są przez algorytm do identyfikacji domeny WG/GW w sekwencji oraz obliczenia sumarycznej wartości jej punktacji. Ostatnie dwie kolumny zawierają zliczenia i częstości występowania danego aminokwasu w sekwencjach odpowiednio domen WG/GW i pozostałych częściach białka.

Zgodnie z oczekiwaniami najwyżej punktowanymi aminokwasami są Trp i Gly, a także dostrzegalna jest podwyższona zawartość reszt Asp, Ser, Ala oraz Asn. Najmniej preferowanymi aminokwasami w domenach WG/GW są kolejno Leu, Ile, Met, Tyr i Cys. Natomiast, udział reszt

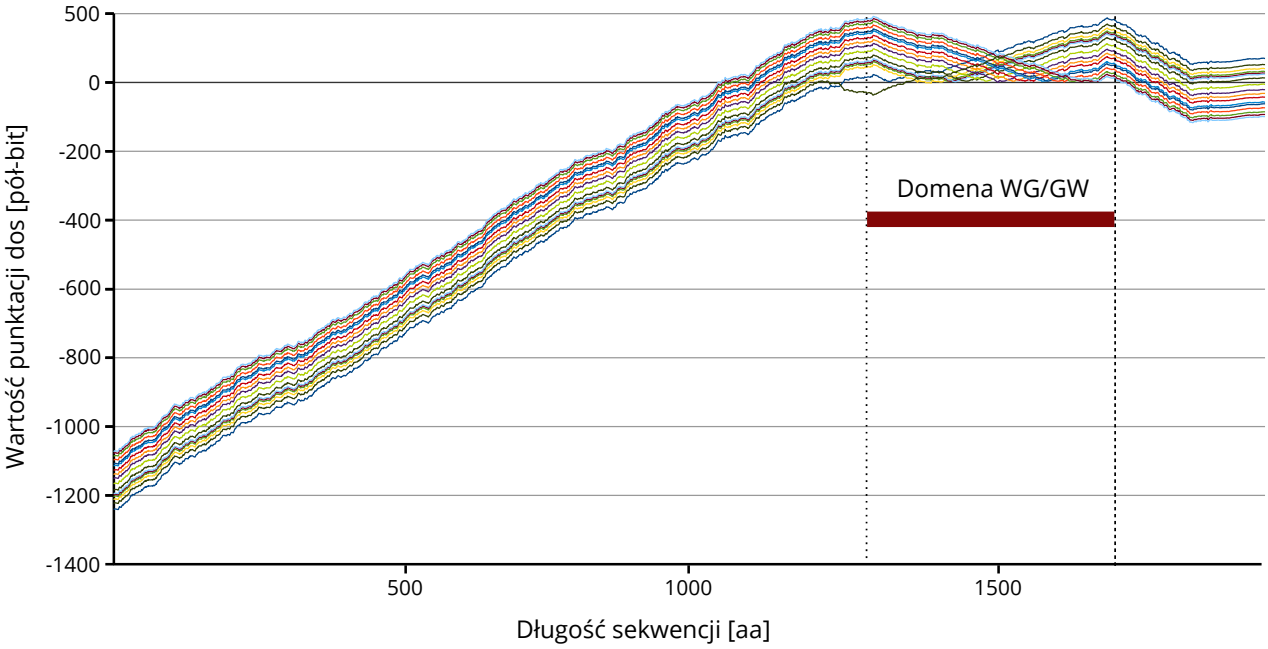
Thr w sekwencji domeny WG/GW jest taki sam jak w przypadku sekwencji niewiążących AGO. Właściwości fizykochemiczne nadreprezentowanych aminokwasów otaczających reszty tryptofanu w sekwencjach domen WG/GW, wskazują preferencje w kierunku małych, hydrofilowych i naładowanych reszt bocznych. Jednak dwukrotnie większa liczba reszt aminokwasowych o niskiej częstości występowania (13 aminokwasów posiada ujemne wartości *score*), przy jednoczesnym dwukrotnie większym zakresie ujemnych wartości punktacji tych reszt (-5,252 w przypadku Leu i 2,666 dla Trp) sugerują większy wpływ działania negatywnej selekcji skierowanej przeciwko występowaniu pewnych aminokwasów w domenie, niż pozytywnej selekcji działającej w kierunku utrzymywania innych aminokwasów.

Procedura przewidywania domen WG/GW w dowolnej sekwencji aminokwasowej przeprowadzana jest z wykorzystaniem uzyskanej macierzy punktującej (tabela 2). Algorytm rozpoczyna działanie od wyszukania w sekwencji zapytania wszystkich wystąpień dwuliterowych motywów WG i GW. Następnie wyznaczony zostaje dłuższy region sekwencji przez rozbudowę prowadzoną na obu końcach każdego znalezionego motywu, z jednoczesnym obliczaniem wartości punktacji całego regionu (rys. 4).



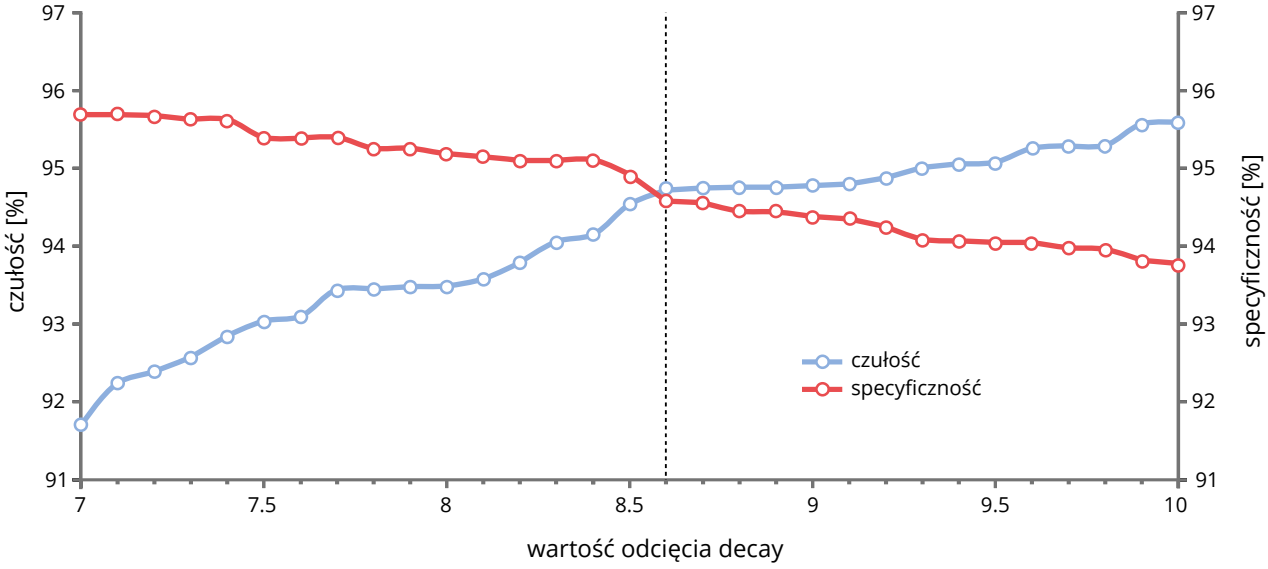
Rys. 4. Procedura rozbudowy sekwencji motywów WG i GW. Sekwencja WG lub GW ulega rozszerzeniu w dwóch kierunkach, z jednoczesnym wyliczaniem skumulowanej wartości punktacji dla każdej kolejnej reszty w sekwencji. Podczas wyznaczania miejsca początku potencjalnej domeny wystąpił spadek wartości jej punktacji o 12,67 pół-bitów. Ponieważ wartość ta jest większa od założonej wartości *decay* 8,6 pół-bitów, proces rozbudowy sekwencji N-końca zostaje zatrzymany.

Niezależnie od kierunku rozbudowy, wydłużanie sekwencji jest przerywane, w chwili, kiedy wielkość oceny obniży się względem wcześniej osiągniętego maksimum o więcej niż pewna ustaloną wartość progową *decay*. W przypadku uzyskania nakładających się sekwencji, łączy się je, tworząc pełnej długości segment sekwencji i obliczając dla niego sumaryczną wartość punktacji *dos*. Otrzymane w ten sposób nienakładające się, wysoko ocenione segmenty odpowiadają potencjalnym domenom WG/GW (rys. 5).



Rys. 5. Wizualizacja wyznaczania miejsca początku i końca domeny WG/GW w sekwencji NRPE1 Arabidopsis. Rozbudowa sekwencji prowadzona jest oddzielnie dla każdego motywu WG i GW (kolorowe linie). Nakładające się sekwencje zostają połączone w jedną domenę, dla której obliczona zostaje wartość punktacji *dos*.

Wartość *decay* wpływa bezpośrednio na długość wyznaczanej domeny: jej zwiększenie powoduje dopuszczanie większej liczby aminokwasów o ujemnych wartościach *dos* podczas rozbudowy motywów, tym samym zwiększając długość sekwencji potencjalnej domeny. Uruchomienie procedury na zestawie pełnej długości sekwencji białek wiążących AGO, przy zastosowaniu serii różnych wartości *decay* z zakresu od 7 do 10 pół-bitów, pozwoliło uzyskać wysoką czułość (94,7%) i specyficzność (94,6%) dla wartości *decay* wynoszącej 8.6 (rys. 6).



Rys. 6. Zdolność przewidywania domen WG/GW. Czułość i selektywność metody zostały wyznaczone na poziomie pojedynczych aminokwasów dla różnych wartości parametru *decay* w zakresie 7-10. Najwyższe wartości obu parametrów (94,7% i 94,6%) uzyskano dla wartości *decay* = 8,6 pół bitów.

Ocena identyfikowanej sekwencji zależy bezpośrednio od jej długości, a zatem wnioskowanie o potencjalnych domenach wiążących AGO wyłącznie na podstawie uzyskanej wartości punktacji *dos* jest niewystarczające do pełnego rozpoznania funkcjonalnych domen WG/GW. Ponieważ długość rzeczywistej domeny WG/GW jest nieznana, do oceny funkcjonalności zidentyfikowanych sekwencji wprowadzono dodatkowy parametr, *ics* (ang. *internal domain composition score*), który podobnie jak punktacja *dos* charakteryzuje kompozycję aminokwasową domeny, lecz nie zależy od długości jej sekwencji. Parametr *ics* opisuje wzajemne relacje między częstościami występowania aminokwasów w obrębie zidentyfikowanej domeny (patrz: Metody - Rozdział 3.1.1). Wyraża on stopień odchylenia tych zależności w stosunku do macierzy opisującej zależności występowania aminokwasów w eksperymentalnie potwierdzonych domenach WG/GW (tabela 3).

Tabela 3. Macierz zależności współwystępowania aminokwasów w domenach wiążących białka AGO wykorzystywana przez algorytm do obliczenia wartości punktacji parametru *ics*.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.0																			
C	2.4	0.0																		
D	0.4	2.6	0.0																	
E	0.8	2.3	0.9	0.0																
F	2.2	1.1	2.4	1.7	0.0															
G	1.8	3.2	1.5	2.1	4.0	0.0														
H	2.2	0.7	2.4	1.9	0.7	3.4	0.0													
I	2.5	0.8	2.7	1.7	1.0	3.7	0.5	0.0												
K	0.8	2.5	0.7	0.6	1.8	2.0	2.2	2.3	0.0											
L	1.9	1.1	2.1	1.4	0.6	3.1	0.6	0.5	1.8	0.0										
M	2.3	0.5	2.6	2.2	1.1	3.3	0.5	0.6	2.5	1.0	0.0									
N	0.5	2.6	0.4	1.0	2.4	1.6	2.4	2.6	0.7	2.1	2.6	0.0								
P	1.0	1.9	1.3	0.8	1.3	2.8	1.5	1.6	0.9	1.2	1.8	1.3	0.0							
Q	1.5	2.0	1.6	0.6	1.3	3.2	1.6	1.5	1.0	1.1	1.9	1.6	1.0	0.0						
R	0.9	1.9	0.9	1.1	1.7	2.3	1.7	2.0	1.3	1.4	2.0	0.9	1.1	1.4	0.0					
S	1.2	3.1	0.9	1.7	3.4	0.8	3.1	3.3	1.5	2.8	3.2	1.0	2.2	2.6	1.7	0.0				
T	0.9	2.1	1.1	0.6	1.3	2.6	1.7	1.8	0.7	1.4	2.0	1.1	0.5	0.8	1.0	2.0	0.0			
V	1.5	1.7	1.6	0.8	0.9	3.0	1.1	1.2	1.2	0.8	1.5	1.6	0.6	0.7	1.3	2.5	0.8	0.0		
W	0.4	2.4	0.4	0.7	2.2	1.9	2.1	2.4	0.6	1.9	2.4	0.3	1.0	1.3	0.8	1.3	0.9	1.4	0.0	
Y	2.2	0.5	2.4	1.9	1.0	3.5	0.6	0.5	2.1	0.9	0.4	2.4	1.6	1.6	2.0	3.1	1.7	1.3	2.2	0.0

Zgodnie z tabelą, największa różnica w częstości występowania dwóch aminokwasów w domenie WG/GW wynosi 4 pół-bity i zachodzi między Gly i Phe. Natomiast najbardziej zbliżony rozkład

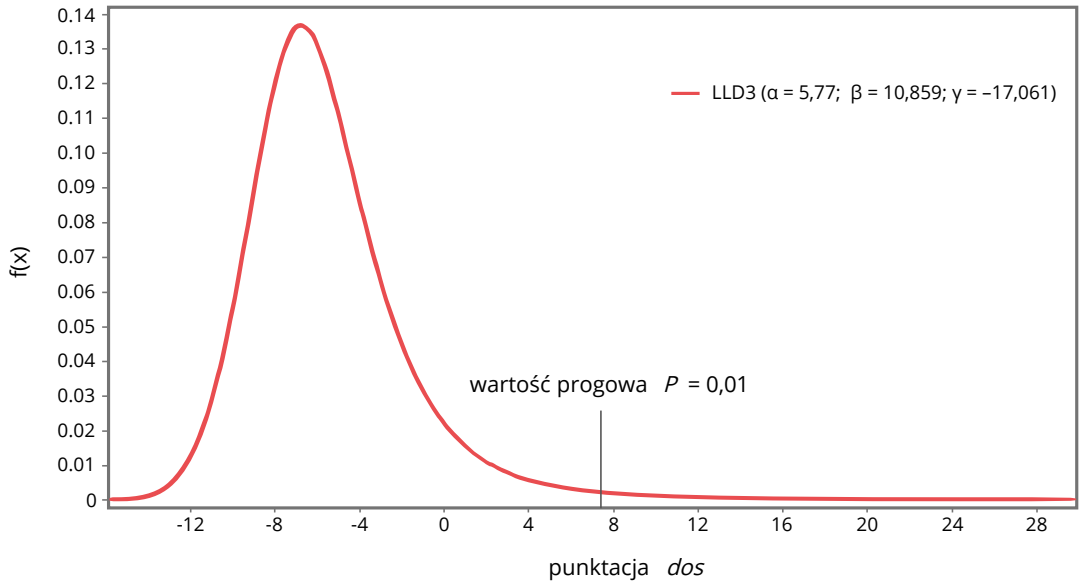
częstości występowania mają Trp i Asn (0,3 pół-bitu). Z kolei, Trp i Gly, kluczowe komponenty domen wiążących AGO, wykazują umiarkowaną różnicę w częstości wynoszącą 1,9 pół-bitów, co wskazuje, że jeden z tych aminokwasów (Gly) występuje z większą częstością w domenie wiążącej AGO niż drugi aminokwas (Trp).

Wprowadzenie dwuparamterowego systemu punktacji - *dos* i *ics* - umożliwia numeryczną ocenę zgodności kompozycji aminokwasowej przewidywanych domen w odniesieniu do wejściowego zestawu potwierdzonych sekwencji wiążących AGO, które zostały użyte do wygenerowania obu macierzy punktujących (tabele 2- 3).

4.1.2. Nowe białka wiążące AGO w genomie *Arabidopsis thaliana*

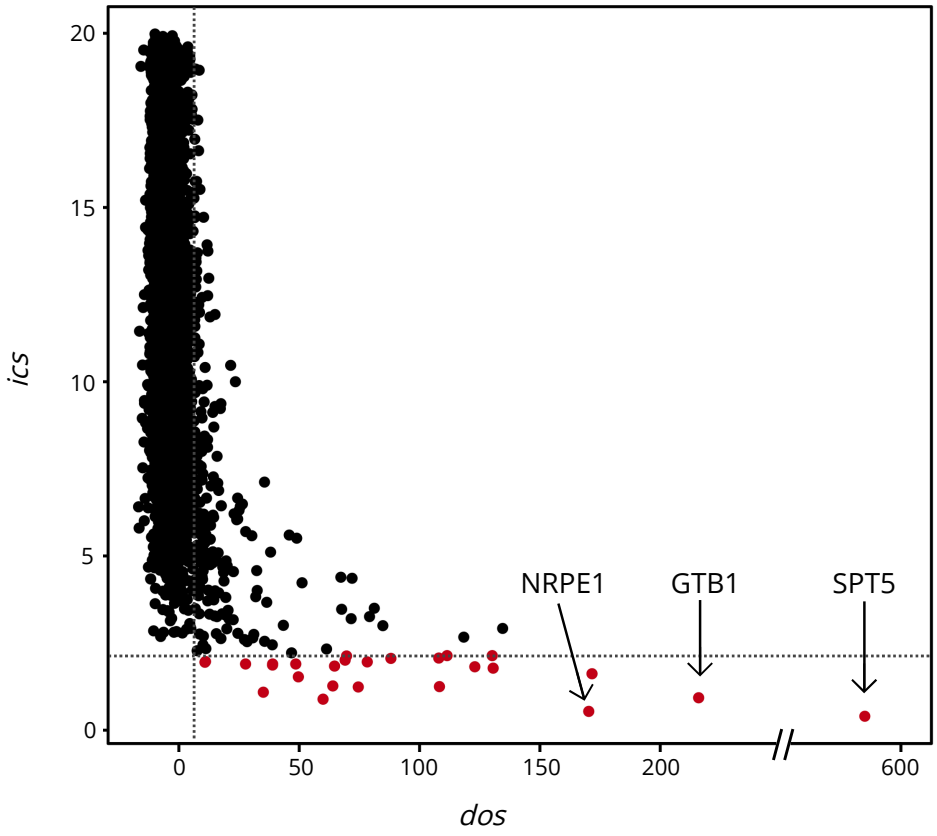
Zastosowanie opracowanej metody do przeszukiwań dowolnego zestawu sekwencji aminokwasowych umożliwia zatem (i) znalezienie wszystkich białek posiadających przynajmniej jedno powtórzenie WG lub GW, (ii) wyznaczenie lokalizacji potencjalnych domen WG/GW w ich sekwencjach w oparciu o macierz punktacji *dos* oraz (iii) dokonanie oceny kompatybilności składu aminokwasowego przewidzianych domen w stosunku do rzeczywistych domen wiążących AGO w oparciu o macierz punktacji *ics*. Genom *Arabidopsis thaliana* stanowi dobry model do takich przeszukiwań nie tylko ze względu na wysoką jakość sekwencji i adnotacji oraz bogatą bibliotekę sekwencji EST i cDNA, ale głównie dlatego, że w krótkim czasie odkryto w tym genomie trzy białka oddziałujące z AGO. Nasuwa to pytania dotyczące powszechności występowania i istotności tych domen białkowych w tym genomie.

Kryteria wartości obu parametrów identyfikujące domenę WG/GW wyznaczono poprzez analizę statystyczną prowadzoną oddzielnie dla obu systemów punktacji. Wyznaczenie wartości progowej parametru *dos* obejmowało analizę dopasowania rozkładów i polegało na modelowaniu funkcji opisującej rozkład prawdopodobieństw tej zmiennej we wszystkich białkach Arabidopsis (patrz: Materiały i Metody: Rozdział 3.1.1). Spośród ponad 40 ciągłych rozkładów teoretycznych, które zostały dopasowane do rozkładu wartości oceny *dos* w genomie Arabidopsis najlepiej dopasowanym do danych modelem jest trójparametrowy ($\alpha = 5{,}77$; $\beta = 10{,}859$; $\gamma = -17{,}061$) rozkład log-logistyczny (rys. 7). Zgodnie z jego dystrybuantą, wartość progowa punktacji *dos* w genomie Arabidopsis wybrana została na poziomie istotności 0,01 wynoszącym 6,99 pół-bitów. Z kolei wartość graniczna parametru *ics* równa 2,14 pół-bitów została wyznaczona w oparciu o estymację parametrów liniowej funkcji regresji prowadzonej na różnej liczbie kombinacji doświadczalnie potwierdzonych sekwencji domeny WG/GW (patrz: Materiały i metody). Zatem statystycznie istotny sygnał identyfikujący domenę WG/GW mają białka, które uzyskały wartości $dos \geq 6{,}99$ oraz $ics \leq 2.14$.



Rys. 7. Rozkład prawdopodobieństwa LLD3 dla punktacji *dos* w genomie Arabidopsis. Wartości $p = 0{,}01$ odpowiada wartość punktacji = 6,99 pół-bitów.

Na rys. 8 pokazano rozkład wartości dwóch systemów punktowania, *dos* i *ics*, dla wszystkich sekwencji aminokwasowych Arabidopsis posiadających przynajmniej jedno powtórzenie WG/GW. Kolorem czerwonym oznaczone zostały białka zawierające regiony sekwencji



Rys. 8. Rozkład wartości punktacji *dos* i *ics* dla wszystkich białek Arabidopsis zawierających przynajmniej jedno powtórzenie WG lub GW. Linie przerywane wyznaczają wartości graniczne dla dwóch punktacji określając potencjalne białka wiążące AGO zaznaczone kolorem czerwonym. Wśród najwyżej ocenionych białek znajdują się trzy białka o potwierdzonej funkcji wiązania AGO: NRPE1, SPT6/GTB1 i SPT5/KTF1. Opublikowano w [56]

spełniające statystyczne kryteria obu parametrów i tym samym, reprezentujące potencjalne białka oddziałujące z AGO.

W sumie zidentyfikowanych zostało 20 różnych genów kodujących potencjalną domenę WG/GW (tabela 4). Zgodnie z oczekiwaniami, geny kodujące białka o potwierdzonej funkcji wiązania AGO - KTF1/SPT5 (p = 8.37E-11), NRPE1 (p = 7.15E-8) i GTB1/SPT6 (p = 2.03E-8) - znajdują się wśród najwyżej ocenionych białek. Spośród zidentyfikowanych genów, dwa transkrypty - czynnika transkrypcyjnego SPT6/GTB1 (AT1G65440) i białka wiążącego RNA (AT4G16830) - powstają na drodze alternatywnego splicingu kodując dwie izoformy białkowe różniące się wartością oceny powtórzeń WG/GW. Z kolei białko SDE3, o którym wiadomo było, że bierze udział w PTGS u Arabidopsis [56], zostało rok później doświadczalnie potwierdzone, jako czynnik oddziałujący z AGO podczas mechanizmów RNAi obrony przeciwwirusowej i kontroli ekspresji materiału genetycznego zawartego w transpozonach [86].

Generalnie biorąc, wielkość sekwencji uzyskanych podczas adnotacji domen WG/GW waha się między 92 i 654 resztami aminokwasowymi (odpowiednio w At3g51940 i At5g04290). Zidentyfikowane białka należą w dużej części do kilku rodzin: wiążących RNA (AT2G16485, AT2G33410, AT2G40030, AT4G16830), czynników transkrypcyjnych (AT1G65440, AT5G04290), białek glicynobogatych (AT1G04800, AT2G15780, AT4G33930, AT4G38710, AT5G07540, AT5G61660), czynników inicjujących translację (AT1G13020, AT3G26400) i białek związanych z wyciszaniem genów (AT1G05460). Niezbędne było zatem przeprowadzenie badań doświadczalnych, które pozwoliły zweryfikować funkcjonalność powtórzeń WG/GW tych białek.

Eksperymentalna weryfikacja komputerowych przewidywań

W celu weryfikacji wiarygodności uzyskanych adnotacji funkcjonalnych białek zawierających potencjalną domenę WG/GW, do dalszych badań eksperymentalnych nad aktywnością wiązania AGO, wybrane zostało białko AT3G51940.1, które uzyskało najniższą wartość oceny powtórzeń WG/GW wśród zidentyfikowanych białek kandydujących (*dos* = 10,83, *ics* = 1,95; p = 4,28E-03). Białko to, posiadające długi region sekwencji zawierający 10 powtórzeń motywu WG/GW, opisane zostało w bazie TAIR jako oksydoreduktaza (ang. *oxidoreductase/transition metal ion binding protein*) (tabela 4). Jednak ta adnotacja nie znajduje potwierdzenia w potencjalnych sekwencjach ortologicznych tego białka u innych roślin, ani w analizie potencjalnych domen białkowych. Dodatkowo niski stopień podobieństwa do domeny reduktazy rybonukleotydowej wskazuje na to, że białko to zostało najprawdopodobniej błędnie scharakteryzowane. Dlatego w trakcie realizacji niniejszej pracy nadano nową nazwę dla genu kodującego to białko - WGRP1 (ang. *WG/GW-Rich Protein 1*).

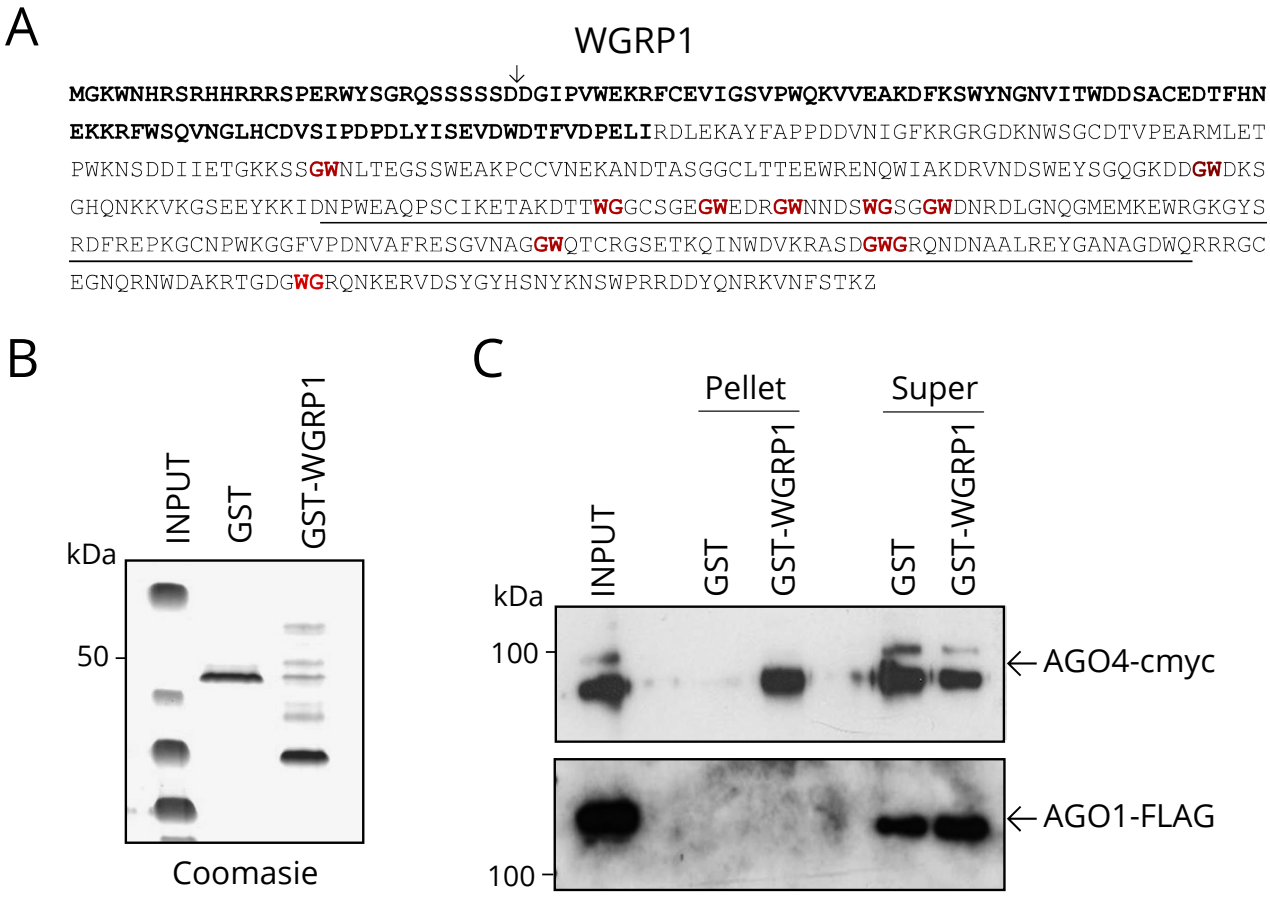
Tabela 4. Białka zawierające potencjalną domeną wiążąca AGO w genomie *Arabidopsis thaliana*. Opublikowano w [56].

AGI locus ↓	*dos*	wartość p	*ics*	adnotacja w bazie TAIR (częściowo)
AT1G04800.1	78.26	3.55E-6	1.96	glycine-rich protein; INVOLVED IN: N-terminal protein myristoylation; LOCATED IN: endomembrane system; EXPRESSED IN: 17 plant structures;
AT1G05460.1	74.53	4.47E-6	1.14	SDE3 – SILENCING DEFECTIVE: a protein with similarity to RNA helicases; mutantsare defective in post-transcriptional gene silencing.
AT1G10270.1	108.26	7.30E-7	1.25	GRP23 – GLUTAMINE-RICH PROTEIN 23: InterPro IPR011990 - tetratricopeptide-like helical domain; InterPro IPR002885 - pentatricopeptiderepeat; InterPro IPR013026 - tetratricopeptideregion.
AT1G13020.1	63.96	9.07E-6	1.27	EIF4B2 - eukaryotic initiation factor 4B2; Plant specific eukaryotic initiation factor 4B:
AT1G15840.1	88.09	2.01E-6	2.06	unknown protein; LOCATED IN: cellular_component unknown; EXPRESSED IN: 11 plant structures
AT1G65440.1	215.95	2.03E-8	0.96	GTB1 – GLOBAL TRANSCRIPTION FACTOR GROUP B1: related to yeast Spt6 protein, whichfunctions as part of a
AT1G65440.2	69.07	6.37E-6	2.01	protein complex in transcription initiation and also plays a role in chromatinstructure / assembly.
AT2G16485.1	59.91	1.22E-5	0.89	DNA binding / nucleic acid binding / protein binding / zinc ion binding; Zinc finger (CCCH-type) family protein / GYF domain-containing protein: InterPro:IPR000571
AT2G33410.1	27.71	2.79E-4	1.9	heterogeneous nuclear ribonucleoprotein, / hnRNP: contains InterPro domain RNA recognition motif, RNP-1;
AT2G15780.1	107.99	7.39E-7	2.07	glycine-rich protein; FUNCTIONS IN: electron carrier activity, copper ion binding; LOCATED IN: endomembrane system; Plastocyanin-like (InterPro:IPR003245), Cupredoxin
AT2G40030.1	170.3	7.15E-8	0.54	NRPE1 - the largest subunit of nuclear DNA-dependent RNA polymerase V;
AT3G26400.1	49.64	2.79E-5	1.53	EIF4B - eukaryotic initiation factor 4B; Plant specific eukaryotic initiation factor 4B:InterPro:IPR010433
AT3G51940.1	10.83	4.28E-3	1.95	oxidoreductase / transition metal ion binding: InterPro domain Ferritin/ribonucleotide reductase-like;
AT4G16830.1	38.91	7.69E-5	1.87	nuclear RNA-binding protein (RGGA): InterPro domain
AT4G16830.3	38.95	7.66E-5	1.9	Hyaluronan/mRNA binding protein
AT4G33930.1	130.58	2.83E-7	1.78	glycine-rich protein; LOCATED IN: endomembrane system; CONTAINS InterPro DOMAIN/s: Cupredoxin
AT4G36230.1	171.65	6.86E-8	1.62	unknown protein; hypothetical protein
AT4G38710.1	11.05	4.09E-3	1.97	glycine-rich protein: InterPro domain Plant specific eukaryotic initiation factor 4B (InterPro:IPR010433)
AT5G03990.1	35.08	1.16E-4	1.09	similar to oxidoreductase/ transition metal ion binding
AT5G04290.1	585.79	8.37E-11	0.4	KTF1 - KOW DOMAIN-CONTAINING TRANSCRIPTION FACTOR 1; SPT5-Like,
AT5G07540.1	122.96	3.85E-7	1.82	GLYCINE-RICH PROTEIN 16 (GRP16); Oleosin (InterPro:IPR000136); FUNCTIONS IN: lipid binding, nutrient reservoir activity; INVOLVED IN: sexual reproduction, lipid storage;
AT5G61660.1	64.68	8.62E-6	1.84	glycine-rich protein;

Geny uszeregowane zostały według identyfikatora AGI (ze względu na lokalizację w genomie).

Gen WGRP1 ulega ekspresji niemal we wszystkich tkankach Arabidopsis i składa się z dwóch egzonów, które kodują białko długości 454 reszt aminokwasowych (rys. 9). Potencjalne białka ortologiczne znalezione u tasznika (RefSeq AC: XP_006292811), topoli (XP_006368584) i kakaowca (XP_007008766) wykazują wysokie zachowanie sekwencji N-końca na odcinku długości około 120 aminokwasów oraz występowanie licznych powtórzeń WG/GW w obrębie bardziej zmiennej sekwencji C-końca białka.

Doświadczenia laboratoryjne mające na celu weryfikację zdolności białka WGRP1 Arabidopsis do wiązania AGO przeprowadzone zostało przez grupę badawczą z Laboratorium Genomu i Rozwoju Roślin na Uniwersytecie w Perpignan we Francji. Skonstruowane zostało białko fuzyjne złożone z regionu białka WGRP1 bogatego w powtórzenia WG/GW (258 - 396) oraz białka GST (rys. 9B; GST-WGRP).



Rys. 9. Zidentyfikowane białko WGRP1 Arabidopsis posiada aktywność wiązania AGO4.
A. Sekwencja aminokwasowa WGRP1 z Arabidopsis. Ewolucyjnie zachowany N-koniec sekwencji wyróżniono pogrubioną czcionką. Pionowa strzałka wskazuje lokalizację intronu względem otwartej ramki odczytu. Występujące w sekwencji motywy WG/GW wskazano czerwoną czcionką. Region aminokwasowy domeny WG/GW, który zostały przyłączony do znacznika GST został podkreślony w sekwencji. **B.** Rozdział białek GST i GST-WGRP na żelu akryloamidowym. **C.** Analiza preferencyjnego wiązania AGO4 do białka WGRP1. Opublikowano w [56].

Metoda koimmunoprecypitacji wykazała, że konstrukt GST-WGRP1, w odróżnieniu od kontrolnego białka GST, oddziałuje z białkiem AGO4. Nie stwierdzono natomiast oddziaływania

białka GST-WGRP1 z białkiem AGO1, co sugeruje funkcjonalną specyficzność powtórzeń WG/GW w oddziaływaniu *in vitro* z białkami AGO4. Warto podkreślić, że podobną specyficzność wiązania z białkiem AGO4 odnotowano w białku SPT5 Arabidopsis [35]. Z całą pewnością określenie roli białka WGRP1 *in vivo* w szlaku RNAi wymaga dalszych badań eksperymentalnych, jednak już potwierdzenie interakcji *in vitro* dla białka, które uzyskało najniższą ocenę spośród zidentyfikowanych sekwencji, istotnie zwiększa wiarygodność przewidywań dla wszystkich pozostałych zaproponowanych sekwencji.

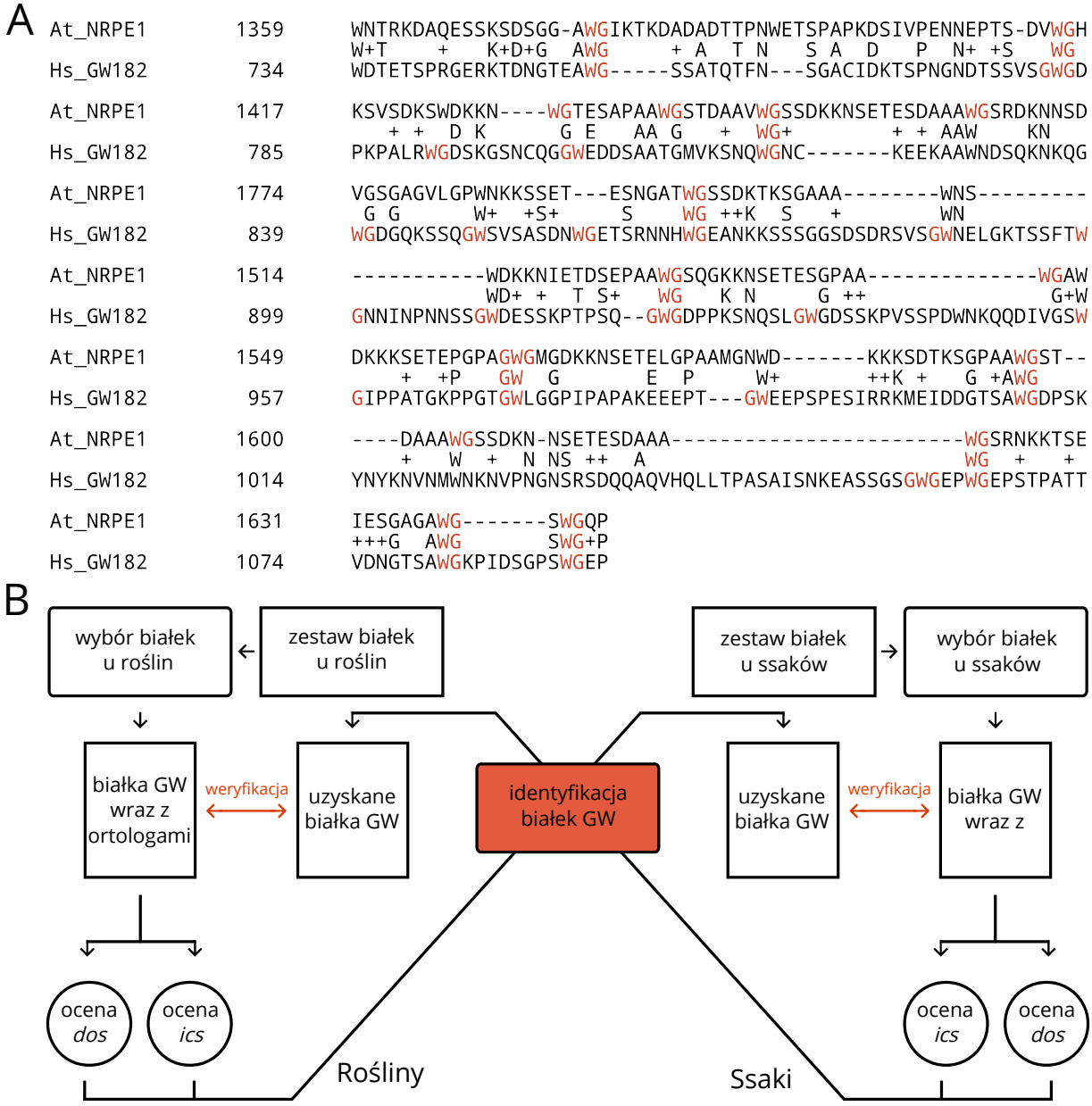
4.1.3. Wirtualna symulacja eksperymentu wymiany domen WG/GW

W oryginalnym eksperymencie przeprowadzonym przez El-Shami i in. (2007) polegającym na wymianie domen WG/GW między dwoma niespokrewnionymi białkami, GW182 człowieka i NRPE1 Arabidopsis wykazano, że powtórzenia WG/GW są funkcjonalnie zachowane u roślin i zwierząt. Oba powstałe chimeryczne białka miały bowiem zdolność oddziaływania z AGO4 Arabidopsis, a ich obecność w komórce warunkowała przynajmniej częściową metylację DNA wybranych elementów powtarzalnych. Ponadto ukierunkowana mutageneza reszt tryptofanu całkowicie zaburzała zdolność obu domen do asocjacji z białkami AGO *in vitro* [9].

Aby zbadać poziom zachowania kompozycji aminokwasowej domen WG/GW wśród roślin i zwierząt, w niniejszej pracy zaprojektowana została symulacja rozszerzająca pole badań eksperymentu El-Shamiego (2007). W symulacji tej wykorzystano macierze punktacji *dos* i *ics* zbudowane oddzielnie na zestawie roślinnych i zwierzęcych białek wiążących AGO do adnotacji sekwencji w układzie krzyżowym (rys. 10). W ten sposób, zastosowanie dwóch różnych tablic punktujących podczas przewidywania funkcjonalnych sekwencji WG/GW, emuluje *in silico* doświadczalną procedurę wymiany domen.

W pierwszym etapie, macierze punktacji zbudowane na roślinnych sekwencjach o eksperymentalnie potwierdzonej funkcji wiązania AGO zostały użyte do przeszukania wybranych sekwencji genomowych ssaków, obejmujących człowieka, szympansa, mysz, szczura, krowę, konia, oposa, dziobaka i makaka. Białka, które uzyskały najwyższą wartość oceny domeny WG/GW podczas tego skanowania obejmują koortologi wszystkich trzech przedstawicieli rodziny GW182 człowieka, które wiążą białka AGO - tj.TNRC6A (p = 1.94E–06), TNRC6B (p = 7.11E–07) oraz TNRC6C (p = 1.24E–06). U pozostałych wytypowanych sekwencji zdolność wiązania AGO nie była testowana eksperymentalnie. Sekwencje te obejmowały grupę białek związanych z keratynocytami, np. wysoko punktowane białko horneryna (p = 9.46E–8) i dermokina (p = 4.94E–8), których cechą charakterystyczną jest tworzenie dużych kompleksów białkowych. Inne zidentyfikowane grupy białek obejmowały czynniki splicingowe, palce

cynkowe, członków rodziny homeobox oraz kinazy. Warto podkreślić, że ponad 40% spośród najwyżej ocenionych białek nie posiada eksperymentalnie zdefiniowanej funkcji i większość z nich zadnotowana jest w bazie danych jako potencjalny lub hipotetyczny produkt białkowy, natomiast w pozostałych przypadkach opis sekwencji jest automatycznie zaimportowany z innych rekordów białkowych wykazujących największe podobieństwo sekwencji.



Rys. 10. Wirtualna symulacja eksperymentu wymiany domen WG/GW między białkami roślin i ssaków. A. Przyrównanie sekwencji domeny WG/GW białka NRPE1 Arabidopsis i białka GW182 człowieka. **B.** Macierze dos/ics zostały wygenerowane w oparciu o doświadczalnie potwierdzone domeny WG/GW roślin/ssaków. Następnie użyte zostały do przeszukania proteomów ssaków/roślin. Zidentyfikowane domeny WG/GW porównano z rzeczywistymi białkami wiążącymi AGO. Symetryczny wynik tej symulacji świadczy o zachowanej kompozycji aminokwasowej roślinnych i zwierzęcych domen WG/GW. Opublikowano w [56].

W drugim etapie, w analogiczny sposób przeprowadzono symetryczne przeszukanie genomu Arabidopsis w oparciu o macierze punktacji *dos* i *ics* wygenerowane na zestawie

eksperymentalnie potwierdzonych sekwencji WG/GW pochodzących od ssaków. Podobnie jak w tym przypadku, na szczycie listy rankingowej predykcji znalazły się znane białka wiążące AGO: NRPE1 (p = 3.29E–07), SPT5/KTF1 (p = 1.10E–08), GTB1 (p = 2.21E– 06) i SDE3 (p = 1.31E–05). Wynik tego dwukierunkowego testu, w którym sygnał kompozycji aminokwasowej białek WG/GW roślinnych rozpoznaje białka WG/GW zwierzęce, i odwrotnie - białka roślinne WG/GW zostają rozpoznane przez sygnał kompozycji białek zwierzęcych, sugeruje, że mimo ekstremalnego stopnia zróżnicowania sekwencji, skład aminokwasowy domeny wiążącej AGO jest ewolucyjnie zachowany pomiędzy królestwami roślin i zwierząt.

4.1.4. Metoda detekcji pojedynczych motywów wiążących AGO

Wirtualna symulacja eksperymentu wymiany domen WG/GW między białkami roślinnymi i zwierzęcymi, wykazała, że po pierwsze, skład aminokwasowy tych domen jest zachowany w obu królestwach, i po drugie, analiza kompozycji domeny posiada wystarczającą zdolność rozpoznawczą przynależności sekwencji do rodziny białek oddziałujących z AGO. Jednak w ciągu ostatnich trzech lat pojawiły się doniesienia, w których wykazano, że powtórzenia reszt aminokwasowych Trp i Gly w ludzkich białkach GW182 są niezbędne do tworzenia innych kompleksów szlaku RNAi nie związanych z białkami AGO1-4 człowieka. Obecność tryptofanu w sąsiedztwie jednego z aminokwasów, glicyny, seryny lub treoniny (G/S/TW lub WT/S/G), okazała się bowiem kluczowa podczas wiązania białek PAN3 lub NOT1, które są podjednostkami kompleksów deadenylacji, odpowiednio, PAN2-PAN3 i CCR4-NOT [87–89]. Brak istotnych różnic w kompozycji aminokwasowej między domeną wchodzącą w interakcję z AGO oraz domeną wiążącą kompleksy deadenylujące ($\chi 2$ = 23.7, p = 0.21) stanowi poważne utrudnienie procesu klasyfikacji tych domen w oparciu o opisaną w poprzednich rozdziałach 4.1.1, 4.1.2, 4.1.3 metodę kompozycyjną. Kolejnym nierozwiązanym problemem podczas przewidywania miejsc wiązania białek AGO jest niejednakowy udział powtórzeń WG/GW w zakresie interakcji z białkami AGO [29,30], który w dużym stopniu zależy od reszt znajdujących się w najbliższym otoczeniu tryptofanu [7]. W świetle tych obserwacji bardziej zasadna wydaje się zatem identyfikacja miejsc wiążących AGO prowadzona oddzielnie dla każdego lokalnego motywu posiadającego tryptofan (nazywanego dalej W-motywem), aniżeli poszukiwanie długich regionów sekwencji wykazujących istotne odchylenie kompozycji aminokwasów.

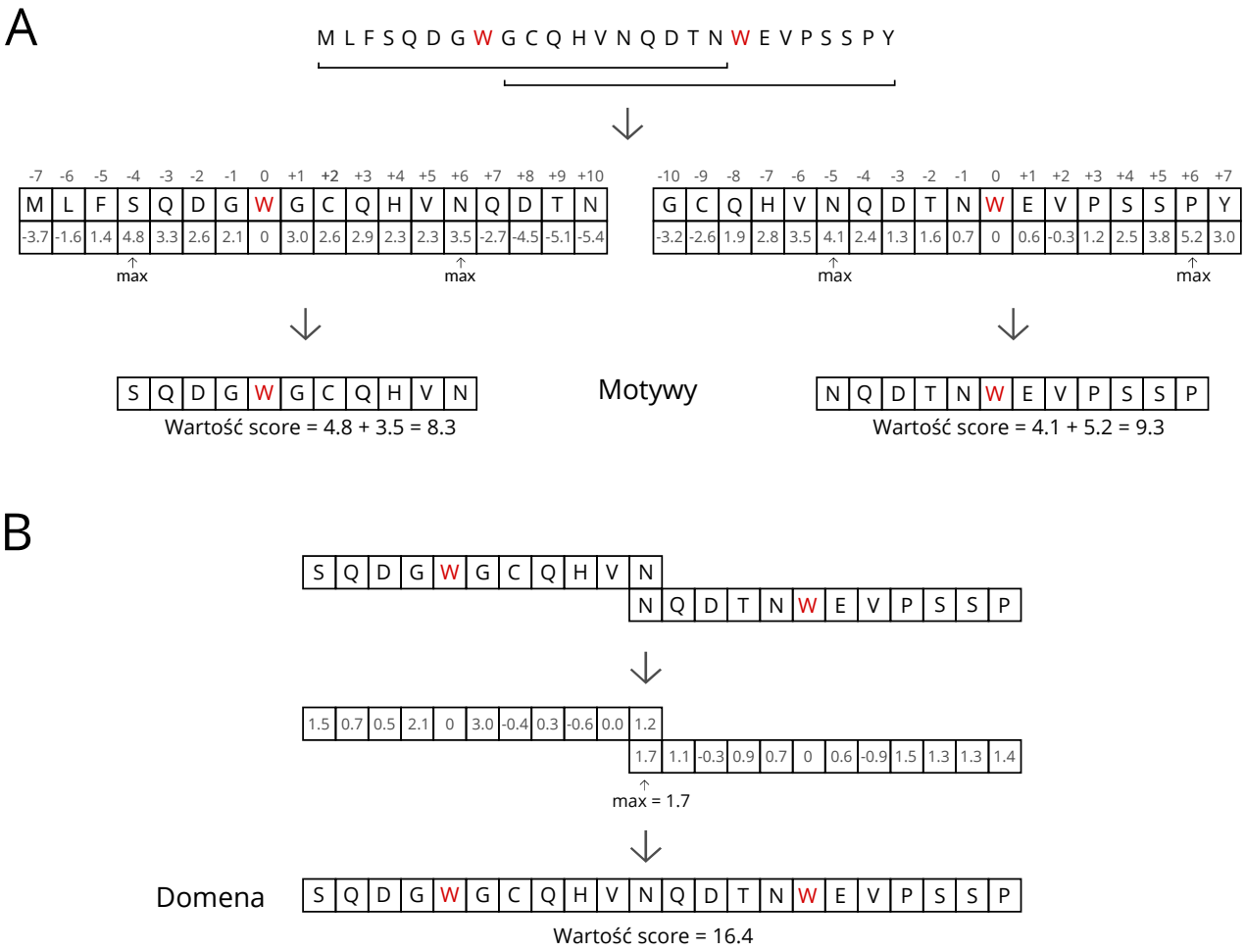
Ponieważ Trp jest niezbędny podczas wszystkich dotychczas potwierdzonych interakcji z białkami AGO [90], można założyć, że stanowi on centrum takiego motywu. Zgodnie z tym założeniem, bez konieczności przeprowadzania dopasowania sekwencji pojedynczych motywów, wyznaczona została pozycyjnie-specyficzna macierz wartościująca (PSSM, ang. *Position-Specific*

Scoring Matrix), która opisuje zależność oceny wystąpienia dowolnej reszty aminokwasowej od pozycji centralnej w macierzy - od reszty tryptofanu (patrz: Materiały i metody: Rozdział 3.1.2). Warto zwrócić uwagę, że długość pojedynczych motywów, użytych do zbudowania tego profilu, nie przekracza zwykle kilkunastu aminokwasów. Oznacza to, że wartości logarytmu ilorazu szans dla aminokwasów znajdujących się w dalszych pozycjach od Trp mają mniejszy wpływ podczas wyznaczania oceny motywu, niż reszty aminokwasowe znajdujące się w najbliższym sąsiedztwie tryptofanu. Dodatkową zaletą proponowanego podejścia jest rozważenie możliwości funkcjonowania wszystkich motywów zawierających tryptofan (WX/XW, gdzie X jest dowolnym aminokwasem), w odróżnieniu od metody kompozycyjnej, której predykcje ograniczone są wyłącznie do motywów WG/GW.

Algorytm predykcji polega na wyznaczaniu sekwencji W-motywu, która uzyska najwyższą ocenę przy zastosowaniu danej macierzy PSSM. W tym celu z sekwencji zapytania wyodrębnione zostają fragmenty sekwencji obejmujące pojedyncze wystąpienie Trp, który otoczony jest resztami aminokwasowymi innymi niż Trp (rys 11A). W obrębie takiej sekwencji, rozpoczynając od pozycji 0 zawierającej Trp, dla każdej kolejnej pozycji obliczona zostaje skumulowana wartość punktacji. W ten sposób wyznaczone zostają sekwencje flankujące o najwyższej wartości. Na koniec zwracana jest lista wszystkich najwyżej punktowanych motywów. W przypadku ocenionych motywów, których sekwencje nakładają się na siebie, algorytm łączy je, dokonując konstrukcji większych domen (rys 11B). Całkowita wartość oceny punktacji domeny jest wówczas sumą punktacji każdego aminokwasu na danej pozycji zgodnie z przyjętym profilem PSSM. W przypadku aminokwasów, których przynależność została przypisana jednocześnie do dwóch motywów, wyższa wartość punktacji zostaje wybrana, jako składowa wartości *score* domeny.

Jak wspomniano w Rozdziale 4.1.1, metoda kompozycyjna podczas rozbudowy motywu WG/GW, kontynuuje rozszerzanie sekwencji dopóki sumaryczna wartość oceny nie obniży się względem osiągniętego wcześniej maksimum o więcej niż pewną zadaną wartość krytyczną. Używanie wartości odcięcia w metodzie kompozycyjnej, podobnie jak w algorytmie BLAST, jest więc podejściem heurystycznym, które nie gwarantuje znalezienia regionu o najwyższej wartości *score* zgodnie z przyjętym systemem punktacji. Choć z drugiej strony, zrezygnowanie z jakiekolwiek wartości odcięcia powoduje w wielu przypadkach nadmiarowe wydłużanie domen - o regiony niskiej złożoności składu aminokwasów - ponieważ algorytm dąży do zmaksymalizowania sumarycznej wartości punktacji. W nowym podejściu, tendencja do nadmiarowego rozbudowania domeny ograniczona jest, z jednej strony przez wystąpienie następnej reszty Trp w analizowanej sekwencji, a z drugiej strony wynika z właściwości macierzy PSSM, w której wartości *score* maleją wraz ze wzrostem odległości od tryptofanu. Takie

rozwiązanie gwarantuje wyznaczenie najwyższej wartości oceny sekwencji pojedynczego motywu przy założonym systemie punktacji. Wynikiem zastosowania metody jest identyfikacja wszystkich najwyżej ocenionych motywów zawierających Trp wraz z wartością ich punktacji oraz lokalizacją w sekwencji.



Rys. 11. Procedura wyznaczania motywów oraz konstruowania potencjalnych domen WG/GW. A. Z sekwencji zapytania zostają wyodrębnione sekwencje zawierające pojedyncze wystąpienie reszty W. Następnie w oparciu o profil PSSM wyliczone zostają skumulowane wartości punktacji dla aminokwasów znajdujących się na wszystkich pozycjach w motywie. Miejsca początku i końca motywu zostają określone przez znalezienie dwóch najwyższych wartości na N- i C-końcu analizowanego motywu. Wartość oceny motywu jest sumą obu tych punktacji. **B.** Wyznaczone motywy, których sekwencje nakładają się na siebie zostają połączone w celu skonstruowania potencjalnej domeny. Do każdego aminokwasu występującego na określonej pozycji w motywie zostają przypisane wartości punktacji z profilu PSSM. W przypadku aminokwasów, które mają przypisane dwie wartości punktacji, ponieważ zostały zidentyfikowane w obu motywach, z obu ocen wybrana zostaje wartość najwyższa. Końcowa wartość oceny domeny jest sumą wartości znajdujących się na wszystkich pozycjach w nakładających się motywach.

Statystyczna istotność ocen zidentyfikowanych motywów i/lub domen jest wyrażana względem rozkładu ocen motywów i domen zawartych w 9 milionach sekwencji eukariotycznych. Oceny domen w tym zestawie, podobnie jak w przypadku metody kompozycyjnej, tworzą rozkład LLD3 (α = 4.40E+8; β = -1.04E+9; γ = 1.04E+9). Z kolei, dystrybucja wartości punktacji motywów przyjmuje czteroparametrowy rozkład Johnson SU

(γ = 1.88; δ = 3.30; ξ = 5.42, λ = 10.98). Zgodnie z wymodelowanymi funkcjami prawdopodobieństw obu rozkładów, wartości p = 0,001 odpowiadają wartościom punktacji: 9,53 dla sekwencji motywów i 16,29 dla domen.

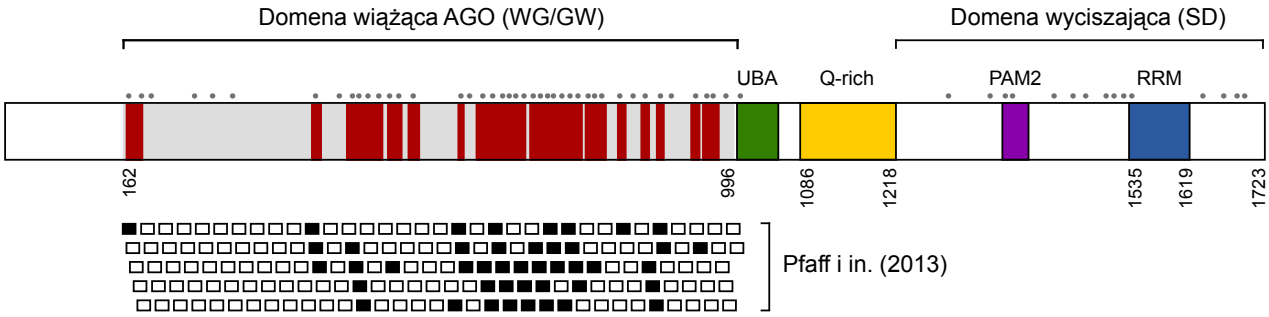
Ewaluacja metody

W celu oceny zdolności proponowanego algorytmu do klasyfikowania przynależności W-motywów do miejsc wiązania AGO, wyniki jego predykcji zostały porównane z opublikowanymi przez inne grupy badawcze doświadczeniami laboratoryjnymi, które wykorzystywały techniki mutagenezy prowadzone w obrębie domen wiążących AGO.

Badania laboratoryjne nad białkiem DmGW182 wykazały, że sekwencja N-końca w pozycji 1-35, nazwana motywem I, jest głównym miejscem wiązania białek AGO1 obejmującym dwie reszty Trp [25,29]. Delecja tego regionu lub substytucja pojedynczego motywu GWG resztami alaniny (G10W - AA) niemal całkowicie zaburza interakcję tego białka z białkami Argonaute. Z kolei mutacja powtórzeń WG/GW, zlokalizowanych w sekwencji poza motywem I, wykazuje niewielki lub brak wpływu na wiązanie [29]. W przewidywaniach nową metodą, sekwencja motywu I znajdująca się w pozycji 7-28 białka, uzyskała najwyższą wartość oceny (16.52 pół-bitów; p = 9.09E-04). Podobnie jak w przeprowadzonych doświadczeniach laboratoryjnych [29], symulacja mutacji w obrębie motywu I (G10W - AA) obniża wartość oceny do -6.44 pół-bitów i klasyfikuje taką sekwencję jako niefunkcjonalną. Wśród pozostałych 33 reszt Trp w tym białku, trzy motywy znajdujące się w obrębie domeny WG/GW uzyskały statystycznie istotną ocenę ($p < 0.001$).

Ostatnio zespół Pfaffa (2013) przeprowadził szczegółowe badanie powinowactwa krótkich motywów ludzkiego białka TNRC6B z białkiem AGO2. W eksperymencie tym, poruszając się wzdłuż domeny wiążącej AGO (162–996) sprawdzona została zdolność wszystkich nakładających się motywów długości 20 reszt aminokwasowych do interakcji z białkiem AGO2. Stwierdzono kilkadziesiąt motywów wykazujących różny stopień specyficzności wiązania białek AGO2 [91]. Na rys. 12 przedstawiono schematyczną reprezentację zidentyfikowanych domen w oparciu o metodę adnotacji *in silico*, zestawioną z wynikami eksperymentu zespołu Pfaffa. Motywy, których zdolność wiązania AGO2 została potwierdzona eksperymentalnie zaznaczono na rys. 12 jako wypełnione prostokąty. Wszystkie te sekwencje zostały zidentyfikowane również przy użyciu metody pozycyjnej, której wyniki przewidywań ($p < 0.001$) przedstawiono na rysunku w formie czerwonych prostokątów. Wśród nich szczególną uwagę zwraca region wiążący AGO2 (472 - 491) [91], który zamiast kanonicznych powtórzeń WG/GW, zawiera motywy SWD i SWN. Algorytm przewidział jednak dwa dodatkowe regiony, których udział w wiązaniu AGO2 nie został potwierdzony w eksperymencie laboratoryjnym. Niewykluczone, że sekwencje te

oddziałują specyficznie z pozostałymi trzema członkami rodziny AGO. Rozpoznanie wszystkich motywów wiążących AGO świadczy o wysokiej czułości algorytmu. Jednocześnie niewielka liczba sekwencji dodatkowo zaklasyfikowanych, jako miejsca wiązania AGO potwierdza również wysoką selektywność algorytmu.



Rys. 12. Porównanie wyników predykcji miejsc wiążących AGO w białku TNRC6B człowieka z wynikami doświadczalnymi Pfaff i in. (2013). Potencjalne miejsca wiązania AGO zaznaczone zostały na schemacie białka czerwonymi prostokątami. Reszty Trp występujące na całej długości sekwencji oznaczono szarymi punktami nad schematem białka. Testowane przez Pfaff i in. 20-aminokwasowe motywy, oznaczone zostały przez prostokąty, które umieszczono pod schematem. Wypełnione prostokąty reprezentują motywy o stwierdzonej zdolności do interakcji z białkiem AGO2 [91].

Liczne powtórzenia G/S/TW i WG/S/T w obrębie C-końca białka GW182 (rys. 12) zostały opisane przez Lian i in. (2009) jako miejsca interakcji z białkami AGO2 [92]. W niniejszych predykcjach motywy zawierające te powtórzenia nie wykazują istotnego podobieństwa do regionów wiążących AGO. Znajduje to również potwierdzenie w wielu innych pracach eksperymentalnych, w których wykazano, że powtórzenia Trp na C-końcu białek GW182 *D. melanogaster* i człowieka nie biorą udziału w wiązaniu białek z rodziny Argonaute [27,30,31,87–89,93]. Wynik ten potwierdza zdolność algorytmu do rozróżnienia motywów uczestniczących w tworzeniu różnych kompleksów białkowych, mimo podobnej kompozycji ich sekwencji.

4.1.5. Nowe białka wiążące AGO u Eukariota

Proponowana metoda pozycyjnie-specyficzna została następnie wykorzystana do identyfikacji potencjalnych białek wiążących AGO u eukariontów. W tabeli 5 przedstawiono listę rankingową rodzin białkowych, które uzyskały najwyższą wartość oceny domeny WG/GW (p < 1e-05) wraz z odpowiadającą im informacją dotyczącą liczby gatunków oraz jednostki taksonomicznej, w której występują.

Wśród przewidywań znajdują się rodziny białek o potwierdzonej funkcji wiązania AGO. Co szczególnie interesujące, domeny WG/GW białek SPT5, SPT6, GW182, jak dotąd badane jedynie u pojedynczych organizmów, mają najprawdopodobniej szerszy zasięg filogenetyczny. Na

przykład, domena WG/GW czynników transkrypcyjnych SPT5 i SPT6, oprócz Arabidopsis [34,35,56], występuje powszechnie u innych roślin wyższych oraz grzybów. Powtórzenia WG/GW reprezentantów rodziny GW182, charakterystycznej dla kręgowców [31] i niektórych owadów [27], zidentyfikowane zostały u niektórych stawonogów, mięczaków i parzydełkowców (patrz: Rozdział 4.3.5).

Wśród białek zidentyfikowanych podczas wcześniejszych przeszukiwań metodą kompozycyjną kilkanaście uzyskało wysoką wartość punktacji ze względu na występujące w ich sekwencjach regiony powtarzalne złożone z par aminokwasów występujących na szczycie macierzy punktacji (tabele 4). W takich przypadkach wartość stosunku sygnału kompozycji aminokwasów do szumu charakterystycznego dla białek zawierających liczne powtórzenia (np. glicynobogatych, serynopogadaych) jest zbyt niska aby prawidłowo rozróżnić te domeny. Na przykład hipotetyczne białko człowieka (GI: 239758013) uzyskało bardzo wiarygodną ocenę domeny WG/GW podczas wcześniejszego skanowania (p = 1.21e-08) [56], ponieważ w ponad połowie sekwencji składa się wyłącznie z najwyżej punktowanych aminokwasów (W, G, N, S, A, D). Jednak szczegółowa analiza sekwencji najbliższego otoczenia Trp przeprowadzona metodą pozycyjnie-specyficzną nie potwierdza zdolności interakcji tego białka z AGO. Podobne przykłady sekwencji o niskiej złożoności składu aminokwasowego, prawdopodobnie błędnie zaklasyfikowane metodą kompozycyjną jako funkcjonalne, obejmują białka koagulacji u ssaków: łańcuch alfa fibrynogenu (GI:194208383), kolagen typu VI (GI:156616290) i prokolagen typu VII (GI:157819015). Z kolei, ciekawymi przykładami sekwencji o niskiej złożoności aminokwasów, które zostały zidentyfikowane również w tym przeszukaniu, są dwa białka związane z keratocytami: dermatokina (DMKN; GI:297276784) i horneryna (HRNR; GI:57864582) [56]. Sekwencja DMKN jest nieustrukturyzowana i składa się w 60% z aminokwasów, których obecność w domenach WG/GW jest faworyzowana [56]. Nowe podejście, spośród 12 motywów zawierających tryptofan, identyfikuje jeden jako potencjalnie wiążący białka AGO, natomiast pozostałe sekwencje nie spełniają statystycznych kryteriów i złożenie ich w większą funkcjonalną domenę jest w tym przypadku niemożliwe. W białku HRNR, 8 spośród 12 motywów zawierających Trp przewidzianych zostało jako potencjalne miejsca wiążące AGO (p < 0.001). Pośrednią przesłankę o możliwej interakcji powtórzeń WG/GW w tej rodzinie białek z białkami AGO dostarcza przeprowadzana analiza spektrometrii mas, w której białko HRNR zostało zidentyfikowane w kompleksie białkowym człowieka zawierającym AGO2 [94]. Natomiast jednym z przykładów domeny WG/GW z potwierdzoną funkcją wiązania AGO, które we wcześniejszym skanowaniu nie zostało zidentyfikowane, jest nieustrukturyzowany region N-końca białka prionowego człowieka (PrP) składający się z pięciu tandemowo powtórzonych wariantów sekwencji ośmioaminokwasowej [95]. Metoda pozycyjna

poprawnie identyfikuje wszystkie motywy (p < 0.001) łącząc je w większą domenę (p = 6.83E-11). Ponadto białka PrP zawierające potencjalną domenę WG/GW zostały zidentyfikowane wśród 152 gatunków ssaków (tabela 5). Sekwencje tych białek wykazują istotną zmienność zarówno pod względem liczby powtórzeń W-motywów (od 2 do 12 kopii), jak i długości ich sekwencji, co sugeruje genetyczną niestabilność tego regionu.

Wśród przewidzianych białek potencjalnie wiążących AGO, na szczególną uwagę zasługują domeny WG/GW obecne w dwóch rodzinach białkowych, helikazach RNA typu DEAD-box i heterogennych rybonukleoproteinach jądrowych (hnRNPs, ang, *heterogeneous nuclear ribonucleoproteins*). Wykazują one najszerszy zasięg ewolucyjny występując jednocześnie w kilku królestwach. W rodzinie helikaz, potencjalna domena wiążąca AGO występuje u 76 przedstawicieli grzybów, 18 zwierząt, 21 roślin i 5 protistów. W białkach tych, domena helikazy RNA występuje w otoczeniu zmiennych sekwencji zawierających powtórzenia WG/GW. Ze względu na udział tej rodziny białek w procesach inicjacji i supresji translacji oraz potwierdzoną obecność w strukturach cytoplazmatycznych (ang. *P-bodies*) u drożdży [96], a także w kompleksie RISC człowieka [97], białka te są dobrymi kandydatami wiązania białek AGO. Z kolei, rodzinę hnRNP charakteryzuje występowanie jednej lub dwóch kopii dobrze zachowanej domeny wiążącej RNA (RRM, ang. *RNA-recognition motifs*) oraz słabo zachowanego regionu sekwencji bogatego w wystąpienia par WG/GW. Udział w wyciszaniu genów na drodze RNAi niektórych członków tej rodziny został stwierdzony u nicieni i stawonogów. Domena RRM białka A1 *C. elegans* wiąże strukturę pętli pri-miR-18a [98], a dwa białka *D. melanogaster* - Hrb98DE i Hrb87F - zostały zidentyfikowane w kompleksie deadenylującym CCR4-NOT1 [99].

Pozostałe zidentyfikowane domeny WG/GW (tabela 5) występują w różnych niespokrewnionych rodzinach białkowych, i wykazują specyficzność w wąskich grupach systematycznych. Na przykład, sekwencje czynników inicjacji translacji IF-2 u dwóch gatunków grzybów z rodzaju *Cryptococcus*, zawierające 54 wystąpienia tryptofanu, nie wykazują znaczącego podobieństwa do białek pochodzących z innych organizmów. Podobnie dwa białka, Nowa1p i Nowa2p, charakterystyczne są jedynie dla pantofelka *Paramecium tetraurelia* i biorą udział w rearanżacji materiału genetycznego podczas procesów płciowych, który polega na precyzyjnym wycięciu jednokopijnych niekodujących sekwencji (IES, ang. *Internal Eliminated Sequence*) [100]. Ze względu na występowanie licznych motywów WG/GW w sekwencjach tych białek, zostały one zaproponowane jako potencjalne białka wchodzące w interakcję z AGO [9], co znalazło również dodatkowe potwierdzenie w prezentowanej analizie. Inne przykłady słabo zachowanych sekwencji obejmują hipotetyczne białka o nieznanej funkcji, białka glicynobogate (GRP, ang. *glycine-rich proteins*) w całym królestwie Eukariota oraz sekwencje *C. elegans* podobne do białek prionowych (tabela 5)

Tabela 5. Lista rankingowa rodzin białkowych potencjalnie wiążących AGO u Eukariota.

Zasięg filogenetyczny	Gat.	Nazwa białka	Opis białka (częściowy)	L. białek	Asocjacja z AGO*	Najwyżej ocenione białko		
						UniProtA	Ocena ↓	Gatunek
Drosophila	5	-	unknown protein	5	NT	B4P508	486.97	*D. yakuba*
Eukaryota	122	SPT5	Transcription elongation factor Spt5	171	AGO4 [34]	M4CZ23	404.9	*B. rapa*
Eukaryota	84	DEAD/DEAH box	DEAD/DEAH box helicase family	120	NT	A2D755	353.62	*T. vaginalis*
Oikopleura	1	-	Uncharacterized	2	NT	E4Z1X2	318.7	*O. dioica*
Metazoa	93	GW182	Trinucleotide repeat-containing gene family	285	AGO1-4 [27]	M7B672	279.41	*C. agassiz*
Dictyosteliida	1	SODC	Copper/zinc superoxide dismutase (SODC)	1	NT	D3BJC0	263.1	*P. pallidum*
Dictyosteliida	1	-	Uncharacterized	1	NT	D3BJ75	262.37	*P. pallidum*
Eukaryota	2	xylA-like	Bifunctional endo-1,4-beta-xylanase	2	NT	A2FUM7	258.85	*T. vaginalis*
Glycine max	1	-	Uncharacterized	1	NT	I1NFI4	252.82	*G. max*
Plants, Fungi	34	SPT6	Transcription elongation factor Spt6	48	AGO [9]	G7ZZF8	202.3	*M. truncatula*
Glarea	1	-	Uncharacterized protein	1	NT	S3DME1	195.3	*G. lozoyensis*
Tetrahymena	1	CnjB	Zinc knuckle family	2	Twi1p	Q24BQ3	194.4	*T. thermophila*
Sclerotiniaceae	2	-	Uncharacterized	3	NT	A7EU40	183.79	*S. sclerotiorum*
Eukaryota	32	GRP	Uncharacterized glycine rich proteins	62	NT	B4GYA0	181.21	*D. persimilis*
Nannochloropsis	1	PAP2	Pap2 haloperoxidase domain-containing protein	1	NT	K8YVY0	172.52	*N. gaditana*
Caenorhabditis	5	PQN	Prion-like-(Q/N-rich)- domain-bearing protein	11	NT	A8XSU8	171.48	*C. briggsae*
Paramecium	1	NOWA	NOWA1, NOWA2	3	kandydat [100]	A0CDB6	161.8	*P. tetraurelia*
Cryptococcus	2	IF2	Translation initiation factor IF-2	4	NT	J9VH01	154.66	*C. neoformans*
Trichoplax	1	-	Predicted protein	1	NT	B3S6S9	154.27	*T. adhaerens*
Tetrahymena	1	WAG1	GW repeat protein	2	Twi1 [10]	B8XQC5	148.77	*T. thermophila*
Magnoliophyta	13	NRPE1	Polymerase V subunit	19	AGO4 [9]	Q5D868	142.18	*S. oleracea*
Oxytricha	1	Nucleoporin	Nucleoporin	1	NT	J9J4X8	132.44	*O. trifallax*
Magnaporthaceae	2	-	Uncharacterized	2	NT	M4G6T7	127.85	*M. poae*
Muscidae	1	-	Uncharacterized	1	NT	T1PKW5	125.96	*M. domestica*
Vitis	1	-	Uncharacterized	1	NT	F6HND0	124.29	*V. vinifera*
Agaricomycotina	8	RNA_pol	DNA-directed RNA polymerase	11	NT	S8FH34	122.47	*F. pinicola*
Marssonina	1	-	Uncharacterized	1	NT	K1WUC6	117.62	*M. brunnea*
Eukaryota	58	hnRNP / RRM	Heterogeneous nuclear ribonucleoprotein	109	NT	Q4Q4J4	117.48	*L. major*
Ichthyophthirius	1	-	Universal minicircle sequence binding protein	1	NT	G0QMY8	114.7	*I. multifiliis*
Panicoideae	2	-	Uncharacterized protein	2	NT	K3Z3H0	110.74	*S. italica*
Coprinopsis	1	-	Putative uncharacterized protein	1	NT	A8N305	110.12	*C. cinerea*
Mammalia	152	PrP	Prion Protein	401	AGO1,2 [95]	Q16409	105.12	*H. sapiens*

Białka uszeregowano według malejącej sumarycznej oceny domen WG/GW.
*NT - nie testowano aktywności wiązania AGO.

4.1.6. Meta-genomowe przewidywanie domen WG/GW u Prokariota i wirusów

W przeciwieństwie do organizmów eukariotycznych, eubakterie i archeony wykształciły własny system regulacji ekspresji genów - CRISPR (ang. *Clustered Regularly Interspaced Short Palindromic Repeats*), który jakkolwiek zależny od RNA, nie wymaga udziału białek AGO [101]. Jednak ortologi eukariotycznych białek Argonaute oraz białek posiadających domenę PIWI są obecne u niektórych bakterii i archeonów, chociaż pozostałe komponenty RNAi nie zostały znalezione w tych genomach [102]. Pomimo, że funkcja prokariotycznych białek AGO (pAGO) nadal pozostaje nieznana, to wykazują one wysoki stopień podobieństwa do białek eukariotycznych. Ostatnio, Koonin wraz z zespołem (2013) zaproponował, że białka pAGO sa kluczowymi komponentami nowego systemu obrony organizmów prokariotycznych, który wykorzystuje DNA lub RNA do wyciszania docelowych sekwencji (np. transkryptów wirusów lub plazmidów) [103]. Jak dotąd białka posiadające domenę WG/GW nie zostały zidentyfikowane w przypadku żadnego organizmu prokariotycznego. W ramach tego projektu przeszukane zostały genomy archeonów i bakterii wykorzystując profil PSSM zbudowany w oparciu o eukariotyczne domeny WG/GW.

Archeony

W królestwie Archaea, zidentyfikowane zostały 23 grupy białek zawierające przypuszczalne miejsca tworzące domenę WG/GW (tabela 6). Białka te pochodzą z 36 gatunków należących do 19 rodzajów systematycznych. Szczególną uwagę zwracają zidentyfikowane domeny WG/GW, które znajdują się u wszystkich 7 gatunków archeonów posiadających ortologiczne białka Argonaute organizmów eukariotycznych. Reprezentatywnym przykładem jest *Archaeoglobus fulgidus*, u którego struktura przestrzenna białka Mid-Piwi została rozwiązana krystalograficznie [104]. W prezentowanym skanowaniu, zidentyfikowane zostało statystycznie istotne powtórzenie WG w sekwencji białka fosfodiestereazy (p = 1.75E-004). Eksperyment laboratoryjny przeprowadzony przez Till i in. (2007) wykazał, że białko pAGO *A. fulgidus* oddziałuje bezpośrednio z motywem drożdżowego białka Tas3 [7]. Sugeruje to zatem, że *A. fulgidus* koduje białko zawierające funkcjonalne powtórzenia WG/GW, a fakt zachowania wysoko ocenionego motywu WG w fosfodiestereazie także u pozostałych 5 blisko spokrewnionych gatunków (tabela 6) może świadczyć o potencjalnej roli tego motywu w wiązaniu białek pAGO. Kolejnymi archeonami, posiadającymi białka Argonaute, u których zidentyfikowano potencjalne miejsca z nimi oddziałujące są *Methanopyrus kandleri* [105,106], *Halogeometricum borinquense* (p = 6.47e-04), *Natrinema gari* i *Natrinema pellirubrum* [105] i *Pyrococcus furiosus* [102], u których domeny WG/GW zidentyfikowane zostały w niescharakteryzowanych białkach.

Tabela 6. Rodziny białkowe archeonów posiadające potencjalną domenę WG/GW.

Zasieg filogenetyczny	L. Gat.	Nazwa rodziny	Opis bialka	L. bialek	AGO/PIWI*	Najwyżej ocenione białko		
						UniProt	*p* ↑	Gatunek
Euryarchaeota	2	-	Uncharacterized protein	2	-	F8AKM4	2.33E-11	*M. okinawensis*
Euryarchaeota	2	-	Uncharacterized protein	6	-	D1YWK8	1.41E-07	*M. paludicola*
Halobacteriaceae	3	-	Uncharacterized protein	3	+	L9XEE3	1.95E-06	*N.innermongolicus*
Archaeoglobi, Thermoprotei	6	-	Phosphodiesterase	6	+	D2RFK9	5.27E-06	*A. profundus*
Thermococcaceae	4	-	Putative dinitrogenase iron-molybdenum cofactor biosynthesis protein 2	5	+	I3ZW97	7.24E-06	*Thermococcus sp.*
Methanosarcina	1	-	Uncharacterized protein	1	-	Q466K1	2.43E-05	*M. barkeri*
Pyrobaculum	2	-	Putative uncharacterized protein	2	-	A1RQH1	2.43E-05	*P. islandicum*
Euryarchaeota	2	-	Uncharacterized protein	2	-	D1Z2J8	4.10E-05	*M. paludicola*
Methanosaeta	1	-	Uncharacterized protein	1	-	F4BWP8	1.20E-04	*M. concilii*
Methanocella	1	-	Uncharacterized protein	1	-	D1YXV4	1.38E-04	*M. paludicola*
Euryarchaeota	2	-	Uncharacterized protein	2	-	H8IAN2	1.76E-04	*M. conradii*
Natrinema	3	-	Uncharacterized protein	3	+	I7CF71	1.76E-04	*Natrinema sp.*
Halococcus	1	-	Uncharacterized protein	1	-	M0MUF5	1.93E-04	*H. thailandensis*
Natrinema	2	-	Flagellin domain protein	2	+	L9ZAS5	3.82E-04	*N. altunense*
Methanosarcina	1	-	Uncharacterized protein	1	-	Q468S4	4.16E-04	*M. barkeri*
Thermococcaceae		SBP_bac	Putative perplasmic sugar binding protein	6	+	F4HLT5	5.64E-04	*P. furiosus*
Halopiger	1	-	Uncharacterized protein	1	+	F8D382	6.14E-04	*H. xanaduensis*
Thermoplasmatales	1	-	Uncharacterized protein	1	-	M7TZJ9	6.26E-04	*T. archaeon*
Halogeometricum	1	-	Uncharacterized protein	1	+	E4NM73	6.47E-04	*H. borinquense*
Methanopyrus	1	-	Uncharacterized; specific for M.kandleri	1	+	Q8TVT4	8.17e-04	*M. kandleri*
Methanosphaerula	1	G-D-S-L	Lipolytic protein G-D-S-L family	1	-	B8GE17	8.58E-04	*M. palustris*
Methanosarcina	1	-	Possible beta-lactamase class C and other pencillin-binding proteins	1	-	Q46D73	8.61E-04	*M. barkeri*
Natronococcus	1	-	Pyrrolo-quinoline quinone	1	-	L9XG02	8.74E-04	*N. amylolyticus*
Methanosarcina	1	-	Uncharacterized protein	1	-	Q46GJ7	9.71E-04	*M. barkeri*
Aciduliprofundum	1	-	Peptidase S53 propeptide	3	-	D3TC53	9.78E-04	*A. boonei*

Białka uszeregowano według malejącej sumarycznej oceny domen WG/GW.

*Symbol +/- oznacza, że w obrębie zidentyfikowanej rodziny białek znajduje się przynajmniej jeden gatunek odpowiednio posiadający lub pozbawiony pełnej długości białka Argonaute lub pozbawionego domeny PAZ.

Z kolei u *Thermococcus barophilus*, również zawierającego białko Argonaute [105], statystycznie istotny pojedynczy motyw WG występuje w białkach odpowiedzialnych za import oligosacharydów (p = 5.64E-04).

Potencjalne domeny WG/GW zidentyfikowano również wśród 25 gatunków, które najprawdopodobniej nie posiadają białek Argonaute. Przykładem są gatunki z rodzajów *Pyrococcus* i *Thermococcus*, u których potencjalna domena WG/GW znajduje się w sekwencji kofaktora dinitrogenezy żelazowo-molibdenowej (p = 7.24e-06). Warto podkreślić, że region sekwencji odpowiadający domenie WG/GW w tym enzymie uległ częściowej delecji w odpowiadających białkach ortologicznych gatunków *P. furiosus* i *T. barophilus*, które posiadają białka Argonaute. Statystycznie istotne powtórzenia WG/GW znaleziono również w kilku przypadkach sekwencji innych enzymów np. w peptydazie p53 *Alstonia boonei* (p =9.78E-04), β-Laktamaza *Methanosarcina barkeri*(p = 8.61E-04). Większość pozostałych zidentyfikowaych białek opisana jest w bazie danych jako niescharakteryzowane.

Bakterie

U bakterii potencjalna domena WG/GW została znaleziona wśród 558 białek należących do 135 rodzin. Ze względu na tak dużą liczbę zidentyfikowanych białek w tabeli 7 pokazano pierwsze 35 najwyżej ocenionych rodzin białkowych (p < 1e-08). Natomiast wszystkie znalezione białka objęto analizą terminów ontologii genów (GO, ang. *Gene Ontology*), w której częstość występowania funkcji molekularnej przedstawiona została na rys. 13A formie mapy tagów.

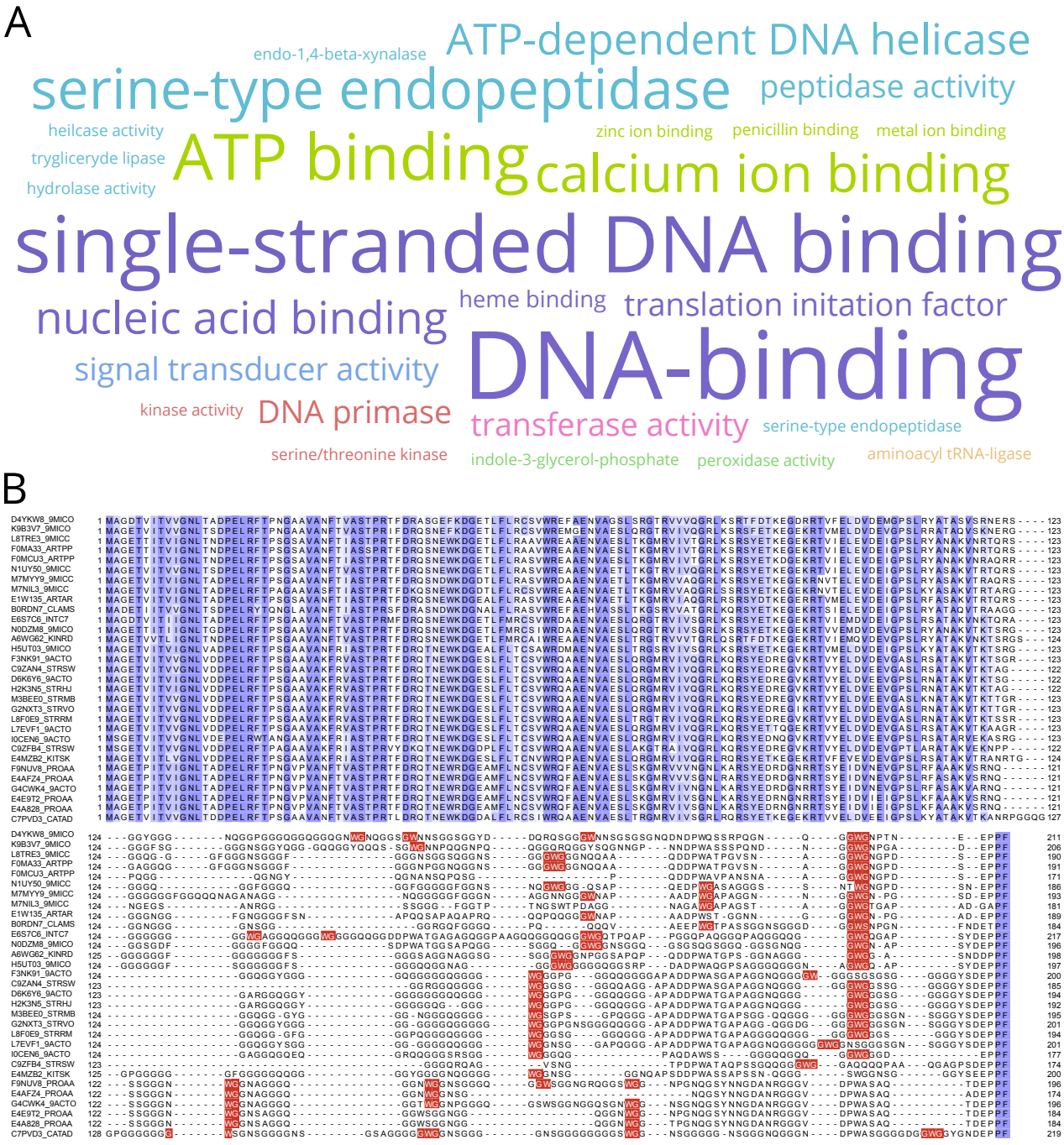
Zidentyfikowane potencjalne domeny wiążące AGO występują w różnych rodzinach białkowych u 65 spośród 101 gatunków bakterii, które kodują pełnej długości lub pozbawione domeny PAZ, białka Argonaute. Przykładem są funkcjonalnie niescharakteryzowane białka *Cyanothece sp.*, *Ktedonobacter racemifer* , *Brucella melitensis* , *Saccharopolyspora erythraea*, *Clostridium saccharoperbutylacet*, oraz *Nocardia brasiliensis*, które uzyskały najwyższą ocenę w przeprowadzonej adnotacji funkcjonalnej (p < 1e-10). Potencjalne W-motywy zidentyfikowano również w białkach glicynobogatych u fotosyntetycznych sinic wiążących azot *Cyanothece sp.* oraz symbiotycznych bakterii roślin *Frankia alni* (p < 1e-10), które również posiadają białka z rodziny Argonaute.

Potencjalne domeny WG/GW zidentyfikowano także w białkach pochodzących z organizmów pozbawionych przedstawicieli rodziny Argonaute. Szczególnym przykładem jest rodzina 91 białek wiążących jednoniciowy DNA (SSB, ang. *single-stranded DNA-binding*), najliczniej występująca w rodzaju *Propionibacterium* obejmującym komensali i fakultatywnych pasożytów człowieka i innych zwierząt. Białka SSB odgrywają główną rolę na różnych poziomach metabolizmu DNA: replikacji, rekombinacji i naprawy.

Tabela 7. Pierwszych 35 najwyżej ocenionych (p < 1e-08) rodzin białkowych bakterii posiadających potencjalną domenę WG/GW.

Zasieg filogenetyczny	L. Gat.	Nazwa bialka	Opis bialka	L. bialek	AGO/PIWI*	Najwyżej ocenione białko		
						UniprotAC	p ↑	Gatunek
Ruminococcus	1	xynA	Bifunctional endo-1,4-beta-xylanase XylA	1	-	P29126	0.00E+00	*R. flavefaciens*
Bacteria	8	-	Uncharacterized protein	14	+	B7K9S1	0.00E+00	*Cyanothece sp.*
Bacteria	4	-	Uncharacterized protein	4	+	N7NN64	0.00E+00	*B. melitensis*
Bacteria	15	-	Uncharacterized protein	17	+	D6TKC9	0.00E+00	*K. racemifer*
Bacteria	8	PE-GRPS	PE-PGRS family protein	9	+	B7KGF5	0.00E+00	*Cyanothece sp.*
Proteobacteria	2	SH3	SH3 type 3 domain protein	2	-	D7A229	0.00E+00	*S. novella*
Bacteria	2	-	Animal heme peroxidase	2	-	I3BRV8	0.00E+00	*T. nivea*
Coprococcus	1	-	Uncharacterized protein	1	-	C0BD12	1.47E-14	*C. comes*
Bacteria	3	-	Uncharacterized protein	3	+	D7DVB6	7.31E-14	*N. azollae*
Bacteria	3	-	Putative uncharacterized protein	3	-	F5J3V3	3.29E-13	*D. gadei*
Peptoniphilus	3	-	Putative uncharacterized protein	3	-	E0NMG7	5.91E-13	*P. duerdenii*
Actinomycetales	2	-	Uncharacterized protein	2	+	A4FIT1	2.71E-12	*S erythraea*
Bacteria	2	-	Uncharacterized protein	2	+	B7KDL5	6.43E-12	*Cyanothece sp.*
Clostridium	1	-	Uncharacterized protein	1	+	M1MED1	2.09E-11	*C. saccharoperbutylacet*
Bacteria	6	-	Hypothetical glycine-rich protein	7	+	Q0RHR1	2.51E-11	*F. alni*
Bacteria	3	-	Uncharacterized protein	3	-	A0LCQ1	3.16E-11	*Magnetococcus sp.*
Bacteria	3	-	Uncharacterized protein	3	+	K0FAQ7	9.36E-11	*N. brasiliensis*
Paenibacillus	1	-	Uncharacterized protein	1	-	K4ZTA1	2.54E-10	*P. alvei*
Bacillus	1	-	FimV protein	1	+	Q2B2Z1	2.84E-10	*Bacillus sp.*
Clostridium	1	-	Dockerin-like protein	1	+	G8LXM6	2.87E-10	*C. clariflavum*
Actinomycetales	2	-	Peptidase, S8/S53 family	2	-	J1LLT9	3.17E-10	*Actinomyces sp.*
Bacteria	3	-	integral membrane protein	3	-	J7LQM9	6.15E-10	*Arthrobacter sp.*
Brevundimonas	1	-	Beta and gamma crystallin	1	-	D9QGA8	6.62E-10	*B. subvibrioides*
Bacteria	2	-	Putative uncharacterized protein	2	-	D3D5H5	7.09E-10	*Frankia sp.*
Bacteria	2	-	Uncharacterized protein	2	-	B8EP03	8.62E-10	*M. silvestris*
Bacteria	5	-	Uncharacterized protein	5	-	A8HWL7	9.04E-10	*A. caulinodans*
Bacteria	2	-	Iron Transport-associated superfamily	2	-	D1VTA7	9.63E-10	*P. lacrimalis*
Corynebacterium	1	-	Putative uncharacterized protein	1	-	C4LL67	9.78E-10	*C. kroppenstedtii*
Bacteria	3	-	Hypothetical conserved protein	3	+	F2J3J7	1.18E-09	*P. gilvum*
Thioalkalivibrio	1	-	Putative uncharacterized protein	1	+	B8GSI4	1.68E-09	*Thioalkalivibrio sp.*
Actinomycetales	4	-	Uncharacterized protein	6	-	Q5YUY3	1.89E-09	*N. farcinica*
Bacteria	5	-	Uncharacterized protein	46	+	I8CRK4	2.28E-09	*M abscessus*
Actinomycetales	26	SSB	Single-stranded DNA-binding protein	91	-	F9NUV8	3.53E-09	*P. acnes*
Corynebacterineae	2	-	Putative uncharacterized protein	2	-	D7WEG7	6.55E-09	*C. genitalium*
Microcystis	1	-	Uncharacterized protein	1	+	I4IVS5	8.73E-09	*M.aeruginosa*

Dodatkowo, nieustrukturyzowany region C-końca sekwencji SSB bierze udział w rekrutacji i wiązaniu wielu różnych partnerów białkowych [107]. Region ten, zawiera powtórzenia WG/GW i odpowiada potencjalnej domenie wiążącej AGO u przedstawicieli rodzin bakterii *Propionibacterineae*, *Streptomycineae*, *Enterobacteriales* i innych (rys. 13B). Należy podkreślić, że bakteryjne RNA pochodzące z dwóch gatunków: *Propionibacterium acnes* and *Brevibacterium mcbrellneri*, zostały znalezione w kompleksach RISC człowieka [108].



Rys. 13. Potencjalne domeny WG/GW występujące w białkach bakteryjnych. A. Mapa tagów częstości występowania terminów Gene Ontology związanych z funkcją molekularną wśród 558 zidentyfikowanych białek. **B.** Przyrównanie sekwencji rodziny SSB zawierających potencjalną domenę WG/GW. Niebieskim kolorem podświetlono pozycje w przyrównaniu, których procent identyczności ≥ 40%. Czerwonym kolorem zaznaczono motywy WG/GW.

Kolejną grupą organizmów, w której zidentyfikowane zostały potencjalne domeny WG/GW są bakterie chorobotwórcze. Na przykład, region zawierający 5 wystąpień GW w białku o nieznanej funkcji (p = 2.28E-09) pochodzący z prątków *Mycobacterium abscessus* powodujących zapalenie płuc człowieka. Powtórzenia WG/GW występują również w obrębie wysoce zmiennego C-końca białek zawierających dobrze zachowaną domenę TPM u krętków rodzaju *Leptospira* (p = 1.24E-06), wywołujących wysoce zakaźną i zaraźliwą chorobę zwierząt i ludzi, oraz paciorkowców *S. tigurinus* (p = 5.72E-007) , *S. oralis* (p = 2.27E-008), *S. mitis* (p = 7.09E-007). Potencjalna domena oddziałująca z AGO, zbudowana z sześciu powtórzeń Trp, zidentyfikowana została w ATP-zależnej helikazie (p = 1.13E-008) opornego na wiele klas antybiotyków (ang. *multidrug-resistance*) pasożyta człowieka *Corynebacterium resistens*. Innymi przykładami białek potencjalnie wiążących AGO komórek gospodarza są: transportery ABC patogenu człowieka *Clostridium hathewayi* (p = 1.42E-008), hipotetyczny attenuator transkrypcji (ang. *cell envelope-related transcriptional attenuator domain*) szczepu *Lachnospiraceae bacterium 1_4_56FAA* (p = 1.13E-007), białko zawierające domenę SSB (p = 6.34E-007) bakterii *Erwinia amylovora* wywołującą zarazę ogniową, chorobę wielu roślin głównie z rodziny różowatych

Wirusy

Chcąc dokonać adnotacji funkcjonalnej wirusowych powtórzeń WG/GW, które na drodze mimikry molekularnej mogą oddziaływać z białkami AGO gospodarza i w ten sposób wymykać się spod działania systemu obronnego RNAi, przeprowadzono podobne skanowanie dostępnych białek wirusowych wykorzystujące profil PSSM eukariotycznych domen WG/GW (tabela 8). Wprawdzie W-motywy w czterech wirusowych białkach o potwierdzonej funkcji wiązania AGO - P1 (p = 0,1) i P38 (p = 0,01) oraz NSs (p = 0,005), NEF (p = 0,06) - uzyskują dodatnią ocenę punktacji odzwierciedlającą podobieństwo do eukariotycznych motywów, to jednak wynik ten nie spełnia statystycznych kryteriów założonych w prowadzonej analizie. Równocześnie jednak zidentyfikowano 488 białek tworzących 32 rodzin białkowych, które uzyskują bardziej wiarygodną ocenę powtórzeń WG/GW.

Potencjalnie miejsca wiązania AGO zidentyfikowane zostały wśród 14 rodzin białkowych wirusów infekujących organizmy zwierzęce. Szczególną uwagę zwracają trzy rodziny białkowe wirusów, które kodują cząsteczki miRNA o potwierdzonej funkcji. Na przykład wirus SGIV (ang. *Singapore grouper iridovirus*) infekujący ryby, zawiera 10 powtórzeń WG/GW na N-końcu niescharakteryzowanego białka, które uzyskało najwyższą ocenę (p = 3.39E-011) Ostatnio Yang i in. (2011) wykazał, że SGIV koduje przynajmniej 11 miRNA, które ulegają komórkowym szlakom procesowania miRNA i są funkcjonalne w zainfekowanych komórkach [109]. Przypuszczalnie funkcjonalne motywy WG/GW znaleziono również w białku ORFRU4-R

najbliżej spokrewnionego z ludzkim wirusem mięsaka Kaposiego (KSHV), wirusa RRV (ang. *Rhesus monkey rhadinovirus*) należącego do rodziny herpeswirusów, który koduje 15 miRNA ulegających ekspresji w wywołanych przez wirusach nowotworach lymphomas naczelnych [110]. Również należące do herpeswirusów wirusy - *Suid herpesvirus 1* i *Human herpesvirus 3* - oprócz funkcjonalnych miRNA oddziałujących z kompleksem RISC gospodarzy, kodują potencjalny motyw WG oddziałujący z AGO. Kolejnym przykładem potencjalnych miejsc wiązania ludzkich białek AGO jest fragment sekwencji składający się z pięciu powtórzeń Trp (p = 3.29E-005) znajdujący się w prekursorowym białku gp160 osłonki zewnętrznej wirusa HIV, którego genom koduje 4 miRNA wchodzące w interakcje z białkiem AGO człowieka [111]. Co dość zaskakujące, niedawno Aqil i in. (2013) eksperymentalnie wykazali, że którekolwiek z dwóch powtórzeń GW w innym białku wirusa HIV - NEF - odpowiedzialne są za wiązanie z białkami AGO2 człowieka, uniemożliwiając rekrutację białek GW182 i tym samym, obniżając wydajność komórkowej odpowiedzi zależnej od miRNA [112]. Mimo, że ocena punktacji tych dwóch motywów w przeprowadzonym skanowaniu uzyskuje wartości dodatnie (p = 0.089), jednak nie spełnia założonych statystycznych kryteriów ($p < 0{,}001$).

Potencjalne miejsca wiązania AGO zidentyfikowano również wśród infekujących zwierzęta wirusów, u których nie stwierdzono obecności cząsteczek miRNA. Warto zauważyć, że wirusy te wywołują odpowiedź RNAi gospodarza. Na przykład niescharakteryzowane białko długości 75 reszt aminokwasowych (p = 4.00E-04) zawiera potencjalne miejsce wiążące AGO gospodarza u wirusa IIV-6 (ang. *Invertebrate iridescent virus 6*), który infekuje stawonogi, w szczególności owady żyjące w wilgotnych i wodnych środowiskach na całym świecie. Ostatnio wykazano, że w odpowiedzi na infekcję IIV-6, *D. melanogaster* uruchamia proces RNAi jako system obronny przeciwwirusowej [113]. W zainfekowanych wirusem IIV-6 muszkach, wykryto bowiem wiele cząsteczek vsiRNA powstałych w wyniku aktywności Dcr-2. Ponadto, mutanty *D. melanogaster*, w których wyłączono białka Dcr2 i AGO, wykazują szczególną podatność na infekcje wirusem IIV-6, w porównaniu do grupy kontrolnej [113,114]. Innym przykładem jest wirus zapalenia wątroby typu C (HCV, ang. *Hepatitis C virus*), który koduje pojedynczy motyw GW mogący potencjalnie oddziaływać z białek AGO (p = 8.18E-005). Niedawno Conrad i in (2013) wykazali, że kodowane przez człowieka miR-122 ulegające ekspresji w wątrobie, w kompleksie z białkiem AGO2 bezpośrednio oddziałuje z 5-UTR RNA wirusa HCV [115],

Potencjalne domeny WG/GW zidentyfikowano również w 8 rodzinach białkowych wirusów, których gospodarzami są organizmy roślinne. Przykładem jest potencjalna domena WG/GW (p = 9.26E-008) wirusa PBCV-1 (ang. *Paramecium bursaria chlorella virus*), który infekuje glony z gromady zielenic. Niedawno zespół Rowe'a (2013) odnotował, że wirus PBCV-1 uruchamia procesy RNAi *C. variabilis* w ciągu mniej niż poł godziny od momentu infekcji.

Tabela 8. Lista rankingowa rodzin białkowych wirusów posiadających potencjalną domenę WG/GW.

Zasieg filogenetyczny	L. gat.	Nazwa rodziny	Opis bialka	L. bialek	UniprotAC	$p \uparrow$	Gatunek
Ranavirus	1	-	Putative uncharacterized protein	1	Q5YFK2	3.39E-11	Singapore grouper iridovirus
Viruses	6	-	Putative uncharacterized protein M442R	13	A7IUH2	7.13E-10	Paramecium bursaria Chlorella virus PBCV
Rhadinovirus	2	-	ORFRU4-R	2	Q9J2I0	7.45E-08	Rhesus monkey rhadinovirus
Myoviridae	1	-	Secreted cell wall DL-endopeptidase	6	R4JF66	3.08E-06	Bacillus phage SIOphi
Herpesviridae	1	-	Putative ORF-3 protein	1	Q5PPC8	5.23E-06	Suid herpesvirus 1
Herpesviridae	1	-	Immediate-early transactivator	1	Q6X674	7.75E-06	Human herpesvirus 3
Caudovirales	7	SSB	Single-stranded binding proteins	7	D5LH30	1.03E-06	Escherichia phage
Mimiviridae	2	-	Putative uncharacterized protein	3	G8EDF7	1.08E-06	Acanthamoeba castellanii mamavirus
Myoviridae	1	-	Putative uncharacterized protein	1	G9I9V9	1.08E-06	Pseudomonas phage
-	1	-	Putative membrane protein	1	M4HNG2	1.65E-05	Bacillus phage
Siphoviridae	1	-	Gp57	3	Q855F3	1.74E-05	Mycobacterium phage
Lentivirus	1	GP160/GP120	Envelope glycoprotein GP160, GP120	368	B3VFF2	3.29E-05	Human immunodeficiency virus 1 (HIV)
Podoviridae	1	-	Structural protein	2	H2EI58	5.45E-05	Brucella phage
Mimiviridae	1	-	Uncharacterized protein	1	M1PGD8	6.48E-05	Moumouvirus goulette
Hepacivirus	1	-	Genome polyprotein	2	I7AZ50	8.18E-05	Hepatitis C virus (HCV)
Tepovirus	1	-	Movement protein	1	K9ML32	8.54E-05	Potato virus T (PVT)
Betaflexiviridae	1	-	RNA-dependant RNA polymerase	1	Q993S7	9.21E-05	Banana mild mosaic virus (BMMV)
Luteoviridae	2		Aphid transmission protein	2	A5A2W2	1.01E-04	Barley yellow dwarf virus (BYDV)
Myoviridae	1	-	Putative uncharacterized protein orf40T	1	Q6VT41	1.14E-04	Vibrio phage
-	1	-	Putative uncharacterized protein	1	G8DGQ9	2.43E-04	Emiliania huxleyi virus
Mimiviridae	1	-	Putative uncharacterized protein mg762	1	G5CRF4	3.50E-04	Megavirus chiliensis
Iridovirus	1	-	uncharacterized protein	1	Q91G16	4.00E-04	Invertebrate iridescent virus 6
Luteoviridae	1	P0	P0	1	Q8QYP7	4.09E-04	Potato leafroll virus
Parapoxvirus	1	-	Putative uncharacterized protein	1	Q6TVI3	4.59E-04	Bovine papular stomatitis virus
-	1	-	Fascin-like domain motif-containing	2	S4VTH3	4.75E-04	Pandoravirus salinus
Gammaretrovirus	1	-	Surface envelope protein (Fragment)	1	S5RPW8	5.09E-04	Avian reticuloendotheliosis virus
Tenuivirus	1	RdRp	RNA polymerase	1	D7NNC5	6.23E-04	Rice stripe virus
Percavirus	1	CLAP	CARD-like apoptotic protein	1	Q9YJN5	6.36E-04	Equid herpesvirus
Siphoviridae	1	-	Tail fiber protein	1	Q6UYJ9	6.66E-04	Burkholderia phage
Begomovirus	2	-	Gemini_V-containing domain	8	I6LI24	7.49E-04	Sweet potato leaf curl
Tospovirus	3	-	Glycoprotein	10	Q2TI70	7.59E-04	Tomato spotted wilt virus

Białka uszeregowano według malejącej sumarycznej oceny domen WG/GW.

Ponadto, spośród 180 genów zaangażowanych w odpowiedź przeciwwiursową na drodze RNAi, 31 wykazuje przynajmniej dwukrotnie zwiększony poziom ekspresji, obejmując białka AGO i podobnych do DCL przedstawicieli rybonukleaz III [116]. Przykładami pojedynczych motywów WG/GW, przypuszczalnie zaangażowanych w wiązanie AGO są białka odpowiedzialne za ruch (p = 8.54E-005) wirusa PVT (ang. *Potato virus T*), a także polimeraza RNA zależnej od RNA (p = 9.21E-005) kodowana przez wirus BBMV (ang. *Banana mild mosaic virus*). Z kolei u wirusów BYDV (ang. *Barley yellow dwarf virus*) i CYDV (ang. *Cereal yellow dwarf virus*) powodujących żółtą karłowatość zbóż, przypuszczalne miejsce wiązania AGO zidentyfikowano w białkach odpowiedzialnych za transmisje wirusa. Warto podkreślić, że blisko spokrewnione z wirusami BYDV i CYDV wirusy z tej samej rodziny Luteoviridae , np.: BWYV (ang. *Beet western yellows virus*) lub PLRV (ang. *Potato leafroll virus)*, kodują białka dezaktywujące domenę PAZ białek AGO1 gospodarza powodując ich degradację [117]. Według przewidywań pokazanych w tabeli 8, wirus PLRV w białku P0 posiada potencjalną domenę wiążącą AGO (p = 4.09E-004). RSV jest kolejnym roślinnym wirusem, który posiada potencjalnie funkcjonalny motyw WG (p = 6.23E-004) obecny w polimerazie RNA. Motyw WG jest także obecny u dwóch wirusów z rodziny *Begomovirus* w białkach Geminivirus V1, niezbędnych podczas infekcji, a także u wirusa TSWV (ang. *Tomato spotted wilt virus*) w białku glikoproteiny (p = 7.59E-004).

Potencjalne miejsca wiążące AGO zidentyfikowane zostały u kilku przedstawicieli bakteriofagów z rzędu *Caudovirales*. Na przykład motywy WG/GW obecne w hipotetycznym białku ogona fagu Gp23, a także innych białkach fagów, których gospodarzami są *Bacillus*, *Mycobacterium*, *Pseudomonas*. Według przeprowadzonego skanowania gatunki tych bakterii kodują potencjalne domeny WG/GW oraz białka Piwi/Argonaute. Co charakterystyczne, motyw WG/GW faga *Escherichia* (p = 1.03E-06) występuje w białku SSB, które jest homologiczne do białka SSB *E. coli* zawierającego jedno powtórzenie GWG.

4.2. Programy do adnotacji i analizy domen WG/GW

Mimo, że powtórzenia WG/GW pełnią kluczową funkcję w wiązaniu białek z rodziny Argonaute, do tej pory nie zostały opisane w dostępnych bazach danych motywów i domen białkowych. Również ze względu na występowanie tych domen w różnych rodzinach białkowych dodatkowo utrudnione jest wyszukanie w centralnych bazach białkowych listy białek oddziałujących z AGO. Podobnie przegląd fachowej literatury nie dostarcza pełnego obrazu domen WG/GW, ponieważ różne grupy badawcze prowadzą badania w obrębie wybranych pojedynczych białek charakterystycznych tylko dla danego organizmu. Dlatego w ramach niniejszego projektu stworzony został publicznie dostępny portal internetowy, Whub

(http://www.comgen.pl/whub), który z jednej strony stanowi aktualną bazę informacji na temat krótkich motywów zawierających tryptofan zaangażowanych w proces RNAi, a jednocześnie oferuje użytkownikowi komplet sieciowych aplikacji wspomagających badania tego typu sekwencji.

4.2.1. Whub - portal internetowy do badań nad motywami zaangażowanymi w RNAi

Portal zaprojektowano w sposób pozwalający na najbardziej optymalne, szybkie i wygodnie dla użytkownika korzystanie z zawartych w serwisie aplikacji i informacji. Układ elementów na stronach portalu jest elastyczny (ang. *responsive Web Design*) - tak by automatycznie modyfikował się w zależności od używanego urządzenia (np. komputer, tablet, smartfon), umożliwiając pełną optymalizację treści. Zarówno podczas prezentowania informacji zawartych w bazie danych, jak i wyników analiz, zastosowano wiele interaktywnych rozwiązań, takich jak: (i) dynamicznie generowane wykresy, (ii) wielofunkcyjne tabele umożliwiające sortowanie, stronicowanie i przeszukiwanie, (iii) rozbudowany system komentowania zintegrowany z popularnymi serwisami społecznościowymi (np. Facebook, Twitter) pozwalający użytkownikom prowadzenie konwersacji w wątkach oraz opiniowanie informacji umieszczonych w portalu.

Whub oferuje cztery drogi pozyskiwania informacji: (i) katalog doświadczalnie potwierdzonych białek wiążących AGO i związaną z nimi (ii) przeglądarkę publikacji, (iii) serwis przeznaczony do analizy składu aminokwasowego domen wiążących AGO/CCR4 oraz (iv) narzędzia skanujące służące do przewidywania miejsc wiązania AGO lub funkcjonalnych W-motywów.

Katalog białek wiążących AGO

W katalogu białek GW opisane zostały wszystkie dotychczas potwierdzone białka wiążące AGO pochodzące z: *A. thaliana* (NRPE1, SPT5/KTF1, WGRP1, SDE3), *H. sapiens* (TNRC6A, TNRC6B, TNRC6C, Prp), *D. melanogaster* (DmGW182/Gawky), *S. pombe* (Tas3), *C. elegans* (AIN-1, AIN-2), *T. thermophila* (WAG1, CnjB), oraz białka wirusów: TCV (P38), SPMMV (P1), TSWV (NSs) i HIV1 (NEF). Rekord białka, oprócz podstawowych informacji o sekwencji (rys. 14A), zawiera dodatkowe części tematyczne. Panel *Domain configuration* przedstawia graficzną reprezentację architektury domen białkowych w kontekście wystąpień reszt tryptofanu (rys. 14B). Jego rozszerzenie stanowi panel *Regions*, który umożliwia przegląd biochemicznie potwierdzonych regionów białka, szczególnie istotnych dla pełnionej przez niego funkcji np. miejsca wystarczające lub niezbędne podczas tworzenia kompleksów z białkami AGO lub CCR4-NOT (rys. 14C).

Rys. 14. Przykładowy rekord ludzkiego białka TNRC6A w portalu Whub podzielony na panele tematyczne: **A.** Basic info - podstawowe informacje o białku (nazwa, opis, długość). **B.** Domain configuration - architektura domen w kontekście W-motywów. **C.** Regions - eksperymentalnie potwierdzone funkcjonalne regiony białka. **D.** Lista opublikowanych wyników mutagenezy. **E.** Sekwencja pełnej długości białka oraz odnośniki literaturowe.

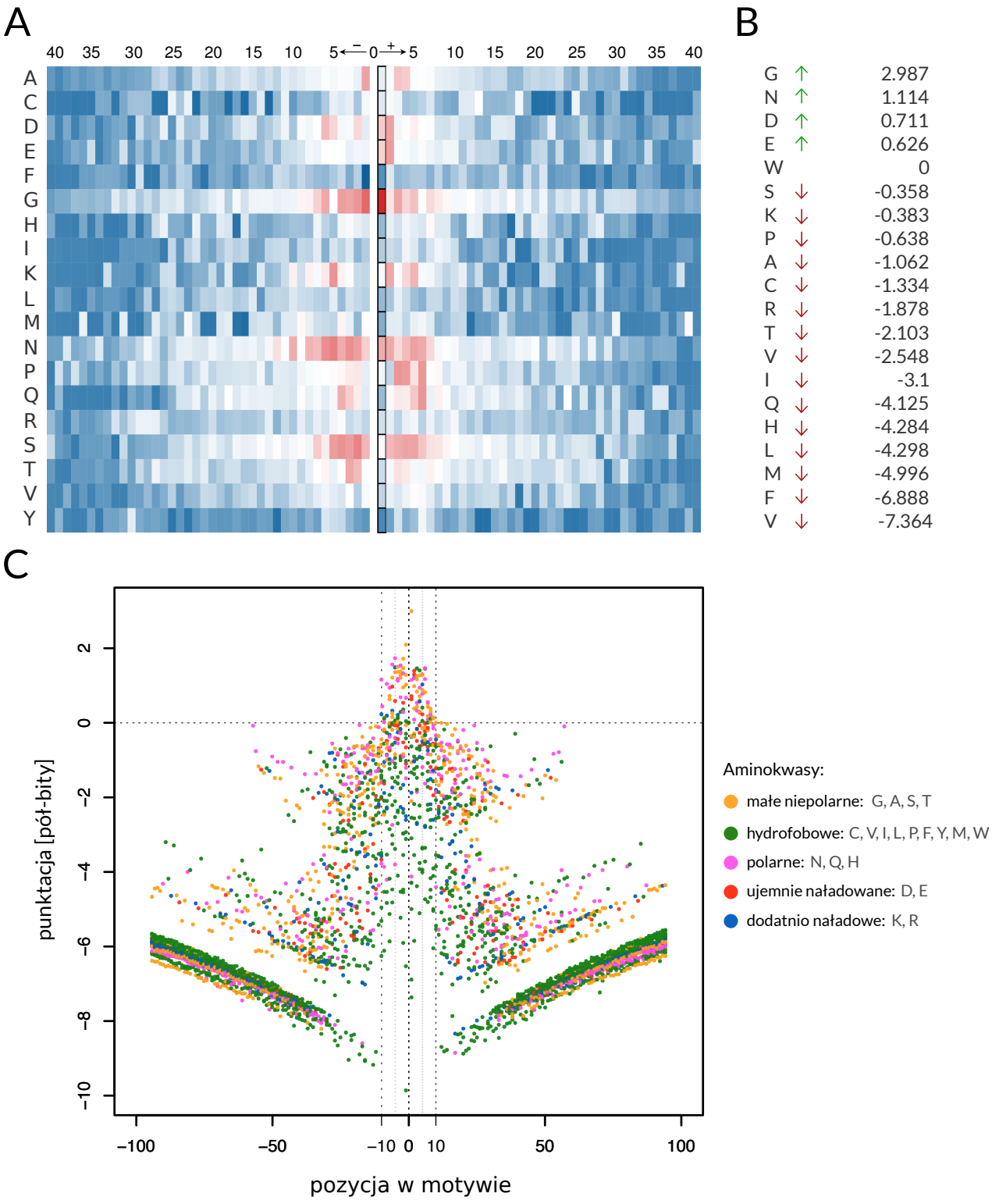
Panel ten jest uzupełniony o szczegółowe adnotacje pochodzące z opublikowanych eksperymentów mutagenezy, przedstawione w formie tabeli, która zapewnia szybki dostęp do informacji poprzez opcje sortowania i przeszukiwania zbiorów (rys. 14D). Każdy wiersz tabeli zawiera odpowiednio typ mutacji (substytucja, delecja, insercja), miejsce w sekwencji oraz wpływ na fenotyp. W dwóch ostatnich częściach rekordu umieszczone zostały odpowiednio pełnej długości sekwencja oraz odnośniki literaturowe (rys. 14E).

Dodatkowo, Whub pozwala na graficzne wyświetlanie cytowanych w portalu publikacji, które prezentowane są w formie kolekcji kart, umożliwiając wizualne sortowanie oraz zawężanie liczby wyświetlanych artykułów przez nakładanie odpowiednich filtrów (np. typ artykułu, funkcja W-motywów). Ponadto, każda pozycja literaturowa powiązana jest z jednym, lub większą liczbą rekordów białek, które stanowiły przedmiot opublikowanych badań oraz wskazuje liczbę eksperymentów przeprowadzonych w obrębie tych białek. Kliknięcie na dowolną kartę spowoduje pobranie z bazy PubMed aktualnych danych literaturowych na temat danej publikacji.

Analiza składu aminokwasowego domen

Portal posiada dział (*Domain analysis*) poświęcony szczegółowemu badaniu składu aminokwasowego sekwencji potwierdzonych domen wiążących AGO i CCR4, które dostępne są w bazie danych. Analiza 6799 W-motywów zaangażowanych w wiązanie AGO u Eukariota ujawnia wiele istotnych właściwości domen WG/GW. Profil PSSM reprezentujący sekwencje tych motywów przedstawiony jest w formie mapy termicznej (ang. *heatmap*), w której preferencja występowania każdego z aminokwasów na danej pozycji została wyrażona odpowiednim kolorem i intensywnością (rys. 15A). W najbliższym otoczeniu Trp tj. w pozycjach -1 i +1, udział glicyny jest odpowiednio cztero- i pięciokrotnie większy w motywach wiążących AGO, niż na tych samych pozycjach w sekwencjach tła. Ponadto motyw w pozycjach -1 i +1 jest dwukrotnie bogatszy w reszty aminokwasowe: Ser, Asn, Asp i Glu. Dodatkowo, jak przedstawiono na rys. 15, można zaobserwować znacznie silniejszą negatywną selekcję przeciwko pojawianiu się dużych hydrofobowych (Ile, Leu, Val) i aromatycznych reszt aminokwasowych (Phe, Tyr, His) na danych pozycjach w motywie. Skrajnym przykładem jest Phe, której udział jest 50 i 900 razy rzadszy odpowiednio w pozycji +1 i -1. Drugim najmniej preferowanym aminokwasem na tych pozycjach w motywie jest Tyr, której częstość występowania jest 50 i 90 rzadsza. Reszty His, Ile, Leu i Val pojawiają się od 2 do 20 razy rzadziej niż na pozycjach -1 i +1 motywów nie wykazujących aktywności wiązania AGO.

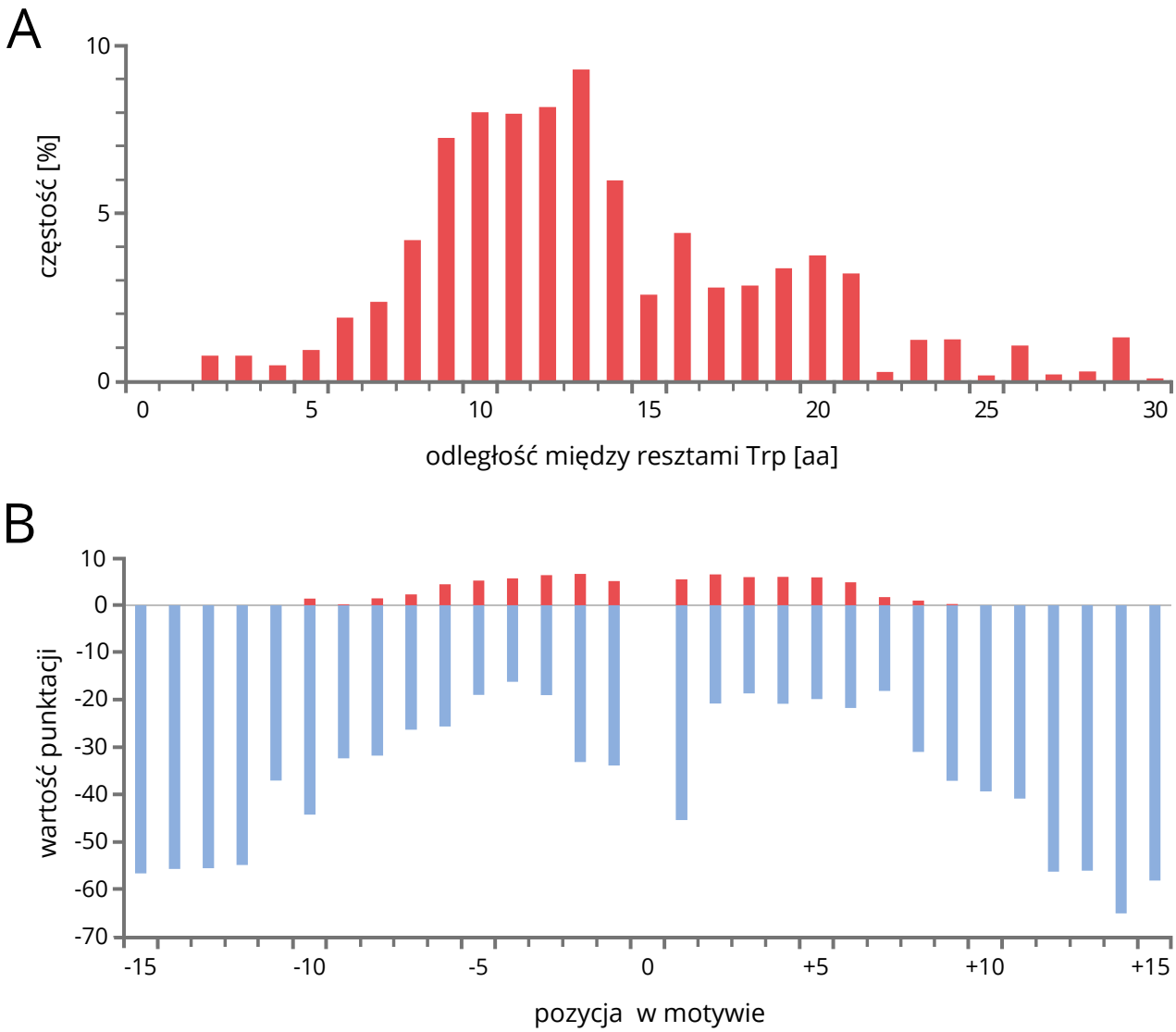
Z rozkładu częstości odległości między resztami Trp w domenach wiążących AGO u Eukariota wynika, że w obrębie tandemowo powtórzonych motywów reszty Trp rozłożone są w większości przypadków co 8-13 reszt aminokwasowych (rys. 16A).



Rys. 15. Kompozycja aminokwasowa 6799 W-motywów występujących w białkach wiążących AGO u Eukariota. A. Profil PSSM przedstawiony w formie mapy termicznej. Czerwone i niebieskie komórki reprezentują wartości log odds ratio, odpowiednio dodatnie i ujemne. **B.** Wartość punktacji każdego aminokwasu na pozycji +1 względem Trp w motywie. **C.** Wykres punktowy przedstawiający właściwości fizykochemiczne aminokwasów na każdej pozycji w motywie. Właściwości aminokwasów zostały oznaczone kolorem zgodnie ze schematem kolorów zaproponowanym w [187].

Schirle i MacRae (2012) [118] rozwiązali ostatnio strukturę przestrzenną kompleksu ludzkiego białka AGO2 wraz z dwoma cząsteczkami wolnego tryptofanu. Co ciekawe, odległość wzdłuż

powierzchni białka AGO2 między dwoma kieszeniami, w których znajdowały się cząsteczki tryptofanu wynosi ok. 24 Å, co w sekwencji odpowiada dystansowi 8-14 reszt aminokwasowych (rys. 16B). Ponadto, przeprowadzone niedawno badania biochemiczne nad oddziaływaniem AGO2 z powtórzeniami zawierającymi Trp w ludzkim białku TNRC6B wykazały, że optymalna długość sekwencji łącznika (ang. *linker*), tj. znajdującej się między dwoma resztami Trp, wynosi przynajmniej 10 reszt aminokwasowych [91].



Rys. 16. Analiza długości W-motywów w białkach wiążących AGO u Eukariota. A. Odległość w sekwencji między sąsiadującymi resztami Trp wyrażona liczbą reszt aminokwasowych. **B.** Rozkład sumarycznych wartości dodatnich (czerwone) i ujemnych (niebieskie) logarytmów ilorazu szans dla każdej pozycji w motywie.

Dominująca przewaga występowania sekwencji łącznika długości 8-13 reszt, nie tylko w obrębie ludzkich białek GW182, lecz także w domenach WG/GW innych rodzin białkowych protistów, grzybów i roślin, sugeruje, że jest to uniwersalna właściwość tej domeny zachowana u organizmów eukariotycznych. Należy zauważyć, że spośród analizowanych W-motywów kilkaset wykazuje jednak większe rozproszenie w sekwencji. Wtedy długość sekwencji łącznika

przekracza 8-14 reszt sięgając w najbardziej ekstremalnych przypadkach ponad 100 aminokwasów.

W celu wyznaczenia optymalnego zakresu długości sekwencji W-motywów na rys. 16B dla każdej pozycji motywu w zakresie od -15 do +15 przedstawiono oddzielnie sumy dodatnich i ujemnych wartości punktacji w macierzy PSSM dla wszystkich aminokwasów. Wartości te odzwierciedlają działanie pozytywnej i negatywnej selekcji osobno dla każdej pozycji w motywie. Pozycje w zakresie od -6 do 6 charakteryzują się najwyższymi sumami dodatnich wartości punktacji, co sugeruje, że reszty aminokwasowe występujące w tych pozycjach są kluczowe dla funkcjonowania motywu. Mimo, że w pozycji ±7 od centrum odnotowano spadek o połowę dodatnich wartości punktacji w porównaniu do pozycji ±6, to jednak aminokwasy występujące na pozycjach -7, -8, -9 i -10 oraz 7, 8 i 9 nadal wykazują pewien stopień zachowania sekwencji. Natomiast w dalszych pozycjach nie zaobserwowano jakichkolwiek dodatnich wartości punktacji aminokwasów. Wynik ten pozwala stwierdzić, że optymalna długość sekwencji oskrzydlających centralnie położony Trp nie przekracza 10 reszt. W tym zakresie, pozycje motywu ±1 i -2, znajdujące się w najbliższym sąsiedztwie Trp podlegają najsilniejszej negatywnej selekcji (rys. 16B). Oznacza to, że w tych miejscach preferowana jest nieobecność pewnych aminokwasów, które najprawdopodobniej zakłócają lub uniemożliwiają uzyskanie pożądanej konformacji.

Podsumowując, domeny wiążące AGO u eukariotów - mimo, że wykazują wysoką zmienność w długości i zachowaniu sekwencji - zbudowane są z wielokrotnie powtórzonych motywów długości 10 - 20 reszt, wewnątrz których znajduje się Trp otoczony hydrofilowymi obszarami o nieskompensowanych ładunkach. Sugeruje to, że domeny wiążące AGO określone są w ograniczonym podzbiorze możliwych sekwencji, które najprawdopodobniej wynikają ze sterycznych restrykcji oddziaływania W-motywów z białkami AGO.

4.2.2. Agos - skaner on-line identyfikacji potencjalnych miejsc wiązania AGO

Wśród trzech dostępnych narzędzi adnotacji domen wiążących AGO, portal Whub zawiera odnośnik do aplikacji internetowej Agos (ang. *Argonaute-binding domain screener*), która jest pierwszym stworzonym w ramach tego projektu programem, dostępnym pod adresem: http://www.comgen.pl/agos. Aplikacja ta jest implementacją metody kompozycyjnej [56] opisanej w Rozdziale 4.1.1, która została opublikowana w czasopiśmie Bioinformatics [119]. W porównaniu do oryginalnej metody kompozycyjnej program Agos wykorzysuje matryce punktujące zbudowane również w oparciu o sekwencje wiążące AGO u zwierząt (Materiały i Metody: Rozdział 3.1.1) oraz uwzględnia dwie modyfikacje algorytmu dotyczące obliczania

macierzy punktującej *dos*, które poprawiły jakość predykcji zwiększając zarówno czułość (96,9%), jak i selektywność metody (97%). W odróżnieniu od pierwszej generacji metody, która wykorzystywała wartości logarytmu ilorazu szans dla pojedynczych aminokwasów, macierz w programie Agos tworzona jest na podstawie wszystkich 400 możliwych kombinacji dipeptydów. Zaletą tego rozwiązania jest uwzględnienie dodatkowej informacji o pewnych zależnościach pozycyjnych w występowaniu aminokwasów w domenie, bez konieczności przeprowadzania przyrównania sekwencji. Na przykład przy wartościach punktacji *dos* stosowanych w pierwszej wersji metody (patrz: Rozdział 4.1.1 - tabela 2), wystąpienia obok siebie reszt tryptofanu i fenyloalaniny (WF/FW) przyjmują wartość dodatnią (0.108 pół-bitów) będącą sumą punktacji *dos* dwóch aminokwasów (W = 2.666, F = -2.558). Mimo, że te dwa duże hydrofobowe aminokwasy występują 13 razy rzadziej w domenie WG/GW niż w niefunkcjonalnych sekwencjach, to pojawienie się ich w analizowanej sekwencji jest preferowane w pierwszej wersji metody - tzn. algorytm będzie zwiększać sumaryczną wartość oceny analizowanej domeny i jednocześnie wydłużać jej sekwencję. Natomiast w drugiej generacji metody Agos ocena powtórzenia WF/FW przyjmuje wartość ujemną -7.232 pół-bitów, odpowiadającą wprowadzeniu kary za znacznie rzadsze występowanie tego powtórzenia, co w efekcie spowoduje obniżenie sumarycznej punktacji domeny. Druga modyfikacja zwiększająca zdolność funkcjonalną metody Agos polegała na odniesieniu częstości występowania dipeptydów w stosunku do generalnej częstości występowania tych motywów we wszystkich sekwencjach eukariotycznych. Inaczej więc niż w przypadku początkowej wersji metody, w której analiza ograniczona była jedynie do regionów wąskiej grupy specyficznych białek. Takie uwzględnienie uniwersalnego rozkładu aminokwasów wzmacnia sygnał kompozycyjny pochodzący z domen WG/GW i powoduje wykrywanie bardziej odległych członków rodziny WG/GW.

Program Agos służy do przewidywania potencjalnych domen wiążących AGO u Eukariota w zadanej przez użytkownika sekwencji nukleotydowej lub aminokwasowej. Specyficzna kompozycja aminokwasowa domen WG/GW wykazująca niski stopień złożoności aminokwasów jest uwarunkowana nietypową preferencją kodonów. Ponieważ programy *ab initio* służące do przewidywania genów podczas wyznaczania struktury genu opierają się na statystycznej charakterystyce regionu kodującego, w niektórych przypadkach dochodzi do pominięcia regionów domen WG/GW, powodując niekiedy utratę całych egzonów. Dlatego sekwencja DNA, podana jako zapytanie w programie Agos, zostaje najpierw przetłumaczona na sekwencję aminokwasową w sześciu ramkach odczytu. co gwarantuje znalezienie wszystkich potencjalnych wystąpień motywu WG/GW. Następnie uruchomiona zostaje procedura identyfikacji potencjalnych domen wiążących AGO, która w oparciu o dwuparamterowy system punktacji (*dos* i *ics*) wyznacza regiony sekwencji o wysokiej punktacji i dokonuje oceny ich wiarygodności.

Wyniki tych przewidywań prezentowane są użytkownikowi bezpośrednio w oknie przeglądarki internetowej i mogą być zapisane w pliku tekstowym.

Rysunek 17 przedstawia wyniki przewidywania aplikacji Agos w sekwencji GW182 u *D. melanogaster*. Raport wynikowy podzielony jest na trzy części. Pierwsza z nich (*Data info*)

Results

Data info:

id:	sp|Q8SY33|GAWKY_DROME
description:	Protein Gawky OS=Drosophila melanogaster GN=gw PE=1 SV=1
length:	1384 aa
WG/GW no:	18

WG/GW domains:

no.	start	end	length	motifs	dos	p-value	ics
1	5	179	175	7	234.53	5.28E-05	0.98
2	216	244	29	1	-8.28	7.76E-01	9.55
3	249	324	76	2	26.19	3.23E-02	3.42
4	327	363	37	1	-10.29	8.58E-01	6.35
5	418	511	94	3	127.91	4.25E-04	2.57
6	590	599	10	1	-9.66	8.43E-01	35.03
7	911	953	43	1	46.35	8.71E-03	6.14
8	1081	1119	39	2	56.12	5.23E-03	7.29

Sequence:

```
mrealfsqdgwgcqhvnqdtnwevpsspepankdapgppmwkpsinngtdlwesnlrnggqpaaqqvpkps
wghtpssnlggtwgedddgadsssvwtggavsnagsgaavgvnqagvnvgpggvvssggpqwgqgvvgvgl
gstggngssnitgssgvatgssgnssnagngwgdpreirplgvggsmdirnvehrggngsgatssdprdir
midprdpirgdprgisgrlngtsemwghhpqmshnqlqginkmvgqsvatastsvgtsgsgigpggpgpst
vsgniptqwgpaqpvsvgvsgpkdmskqisgweepspppqrrsipnyddgtslwgqqtrvpaasghwkdmt
dsigrsshlmrgqsqtggigiagvgnsnvpvganpsnpissvvgpqaripsvggvqhkpdggamwvhsgnv
ggrnnvaavttwgddthsvnvgapssgsvssnnwvddksnstlaqnswsdpapvgvswgnkqskppsnsas
sgwstaagvvdgvdlgsewnthggiigksqqqqklaglnvgmvnvinaeiikqskqyrilvengfkkedve
ralvianmnieeaadmlransslsmdgwrrhdeslgsyadhnsstssggfagrypvnsgqpsmsfphnnlm
nnmggtavtggnnntnmtalqvqkylnqgqhgvavgpqavgnssavsvgfgqntsnaavagaasvniaant
nnqpsgqqirmlgqqiqlaihsgfissqiltqpltqttlnllnqllsnikhlqaaqqsltrggnvnpmavn
vaiskykqqiqnlqnqinaqqavyvkqqnmqptsqqqqpqqqqlpsvhlsnsgndylrghdainnlqsnfs
elninkpsgyqgasnqqsrlnqwklpvldkeinsdstefsrapgatkqnltantsninslglqndstwstg
rsigdgwpdpssdnenkdwsvaqptsaataytdlvqefepgkpwkgsqiksieddpsitpgsvarsplsin
stpkdadifantgknsptdlppslsssstwsfnpnqnypshswsdnsqqctatselwtsplnkssrgppp
gltansnksansnastpttitggangwlqprsggvqttntnwtggnttwgsswlllknltaqidgptlrtl
cmqhgplvsfhpylnqgialckyttreeankaqmalnncvlanttifaespsenevqsimqhlpqtpssts
ssgtsggnvggvgtsannansgsaaclsgnnsgngngsasgagsgnngnsscnnsaagggsssnntittva
nsnlvgssgsvsnssgvtansstvsvvsctasgnsingagtanssgskssannlasgqssasnltnstnst
wrqtsqnqalqsqsrpsgreadfdyislvysivdd
```

Rys. 17. Wynik działania programu Agos na przykładzie białka Gw182 u *D. melanogaster.*

oprócz podstawowych informacji o sekwencji użytej w zapytaniu (numer dostępu, opis i długość sekwencji, liczba motywów WG/GW), zawiera graficzną reprezentację sekwencji uwzględniającą zidentyfikowane regiony WG/GW. Wysokość oceny potencjalnych domen WG/GW, przedstawionych na schemacie w formie bloków, oznaczona została trzema kolorami. Zielone bloki wskazują regiony sekwencji o najwyższej wiarygodności, które spełniają statystyczne kryteria systemu punktacji (*dos* > 43 poł-bitów i *ics* < 1.53) i mogą być uznawane za potencjalne miejsca wiązania białek AGO. Regiony sekwencji oznaczone żółtymi blokami spełniają jedynie kryterium punktacji *dos* i wykazują wyraźnie zbliżoną kompozycję aminokwasową do białek wiążących AGO. Natomiast czerwone bloki reprezentują sekwencje, które wprawdzie posiadają przynajmniej jeden motyw WG/GW, lecz oprócz tego nie wykazują rozpoznawalnego podobieństwa do kompozycji aminokwasowej potwierdzonych białek wiążących AGO. Druga część raportu (*WG/GW domains*) przedstawia wyniki uzyskanych przewidywań w formie tabeli, uszeregowanej ze względu na pozycję wystąpienia znalezionej domeny w sekwencji. Każda z potencjalnych domen, odpowiadających wierszom tabeli, zawiera informacje o lokalizacji domeny w sekwencji zapytania, długości, liczbie motywów WG/GW, wartości punktacji *dos* i *ics* oraz oceny wiarygodności (ang. *p-value*). Trzecia sekcja raportu wyświetla sekwencję użytą w zapytaniu. Skierowanie kursora na dowolny blok znajdujący się w graficznej reprezentacji przewidywań spowoduje podświetlenie odpowiadającego mu regionu sekwencji oraz wiersza w tabeli.

Analizowane na rys. 17 białko DmGW182 stanowi szczególnie interesujący przykład funkcjonalnego różnicowania powtórzeń WG/GW. W badaniach zespołu Behm-Ansmant (2006) wykazano eksperymentalnie, że nieustrukturyzowany region znajdujący się w pozycji 1-539 tego białka zawierający 12 powtórzeń WG/GW oddziałuje z białkiem AGO1 [25]. Natomiast, w wynikach predykcji programu Agos jedyna potencjalna domena (*dos* = 234,5, *p* = 5.28E-05, *ics* = 0,98) znajduje się w pozycjach 5-179 i obejmuje 7 powtórzeń dipeptydów WG/GW. Wynik ten znajduje wyraźne potwierdzenie w ostatnich doniesieniach, w których wykazano, że konstrukt białka GW182 pozbawiony 1-205 aminokwasów N-końca jest niezdolny do wiązania białek AGO1 *in vitro*, natomiast powtórzenia WG/GW w regionie 205-490 biorą udział w procesie wyciszania genów, nie związanym z tworzeniem kompleksów z białkami AGO [120].

4.2.3. Wsearch / i-Wsearch - programy identyfikacji funkcjonalnych W-motywów

Obok programu Agos, portal Whub wyposażony jest w dwa dodatkowe narzędzia przeznaczone do przewidywania *de novo* potencjalnych miejsc wiązania białek AGO w dowolnej sekwencji aminokwasowej wprowadzonej przez użytkownika.

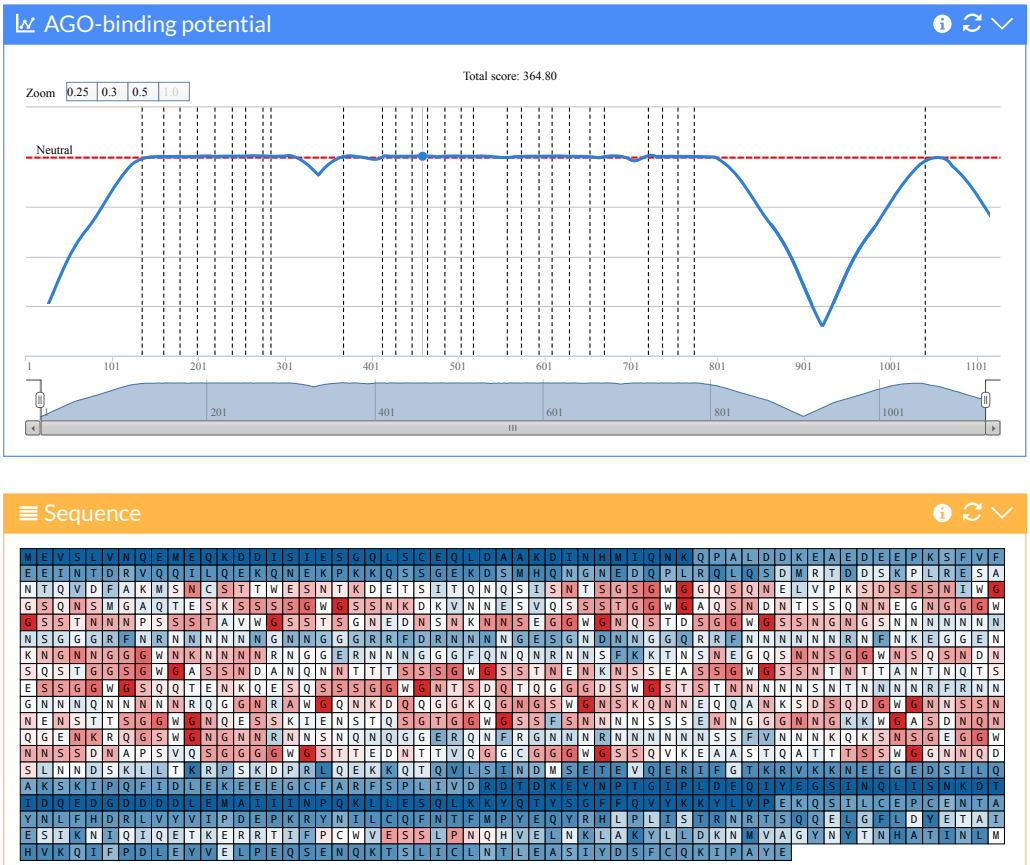
Wsearch

Pierwsze narzędzie - Wsearch jest implementacją metody pozycyjnie-specyficznej, opisanej w Rozdziale 4.1.4 i znajduje się pod adresem: http://www.comgen.pl/whub/wsearch. Program skanuje sekwencję wejściową stosując wybrany przez użytkownika profil PSSM, a następnie wyznacza najwyżej punktowane sekwencje W-motywów wraz z ich lokalizacją w sekwencji oraz oceną istotności statystycznej. W przypadku nakładania się sekwencji zidentyfikowanych motywów, program łączy je w większą domenę, ponownie dokonując oceny jej punktacji i zwracając ocenę wiarygodności. Wyniki przedstawione są użytkownikowi w dwóch częściach (rys. 18). Pierwsza z nich obejmuje interaktywny wykres liniowy oraz mapę sekwencji, w których zawarta jest informacja o wpływie wartości punktacji każdego aminokwasu wzdłuż całej długości białka. Na wykresie aminokwasy, których obecności jest faworyzowana w W-motywach przyjmują wartości dodatnie. W przypadku długich sekwencji zapytania wykres można dynamicznie skalować i powiększać stosując przeciągnięcia kursora lub korzystając z przycisków ze zdefiniowanymi skalami powiększenia. Skierowanie kursora w dowolne miejsce na wykresie spowoduje wyświetlenie informacji o skumulowanej wartości *score* aminokwasu znajdującego się w danej pozycji. W odróżnieniu od wykresu, mapa sekwencji dostarcza informacji o wpływie aminokwasu w wyznaczanie motywów, który zakodowany jest w kolorze odpowiadającym wartościom z matrycy PSSM. Druga część wynikowego raportu obejmuje tabele zawierające listy sekwencji zidentyfikowanych W-motywów i domen wraz z informacją o uzyskanej wartości oceny, istotności statystycznej oraz lokalizacji sekwencji.

Dodatkową funkcjonalnością programu Wsearch jest możliwość przewidywania W-motywów zaangażowanych w oddziaływanie z kompleksem deadenylacji CCR4-NOT poprzez wybranie odpowiedniego profilu PSSM. Należy jednak podkreślić, że sekwencje wiążące CCR4-NOT zidentyfikowane zostały dotychczas jedynie w białkach GW182 *H. sapiens* i *D. melanogaster*. Pomimo uwzględnienia w matrycy PSSM W-motywów białek ortologicznych, adnotacja nowych przedstawicieli oddziałujących z AGO na podstawie tylko jednej rodziny białek może nie dostarczać w pełni wiarygodnych wyników.

Oprócz aplikacji internetowej, która przeszukuje tylko jedną sekwencje jednocześnie, utworzona została lokalna wersja programu Wsearch umożliwiająca przeszukiwanie dużych zbiorów sekwencji. Można ją można pobrać wraz z dokumentacją oraz dostępnymi macierzamie PSSM ze strony: http://www.comgen.pl/whub/download/software. Program współpracuje z popularnymi systemami operacyjnymi (Windows, UNIX, MacOS) i do prawidłowego działania wymaga instalacji Python w wersji $\geq$ 3.3. Wyniki adnotacji W-motywów oraz domen prezentowane są użytkownikowi w pliku tekstowym w formacie tabularnym. Dodatkowo program umożliwia tworzenie macierzy PSSM dla sekwencji podanych przez użytkownika.

A



B

Single motifs

5 records Search:

Start	End	Score	P-val	Motif
130	141	6.22	0.017	SNCSTTWESNTK
153	166	14.88	4.17E-06	SNTSGSGWGGQSQN
171	185	11.17	2.00E-04	KSDSSSNIWGGSQNS
192	204	16.13	1.10E-06	SKSSSSGWGSSNK
214	229	14.11	9.41E-06	SSSTGGWGAQSNDNTS

Showing 1 to 5 of 29 entries

1 2 3 4 5

Assembled motifs

5 records Search:

Start	End	Score	P-val	Motif
232	262	27.22	9.73E-06	NNEGNGGGWGSSTNNNPSSSTAVWGSSTSGN
265	293	28.39	5.93E-06	NSNKNNSEGGWGNQSTDSGGWGSSNGNGS
499	525	26.98	1.08E-05	SSSGGWGNTSDQTQGGGDSWGSTSTNN
586	616	26.41	1.37E-05	NKSDSQDGWGNNSSNNENSTTSGGWGNQESS
643	675	23.44	4.82E-05	NNGGGNNGKKWGASDNQNQGENKRQGSWGNGNN

Showing 1 to 5 of 5 entries

1

Rys. 18. Raport wynikowy aplikacji Wsearch na przykładzie białka WAG1 *T. thermophila*. A. Wizualizacja wartości punktacji każdego aminokwasu w sekwencji przedstawiona w formie wykresu liniowego i mapy sekwencji. **B.** Listy zidentyfikowanych sekwencji W-motywów i domen zawierające informacje o uzyskanej liczbowej ocenie przewidywań, istotności statystycznej i lokalizacji w sekwencji zapytania.

i-Wsearch

Kolejną aplikacją internetową służącą do przewidywania potencjalnych miejsc wiązania AGO jest implementacja metody pozycyjnie-specyficzej realizowana przez algorytm nauczania maszynowego, lasów losowych [60] (patrz Metody - Rozdział 3.1.3). Program i-Wsearch znajdujący się pod adresem: http://www.comgen.pl/whub/i-wsearch/ w podanej przez użytkownika sekwencji białkowej analizuje wszystkie wystąpienia 21-aminokwasowych motywów zawierających w centrum resztę tryptofanu. Następnie dokonuje przypisania każdego z nich do jednej z dwóch klas: miejsc asocjacji z AGO lub sekwencji nie mających związku z białkami AGO. Zaletą tego podejścia jest wskazanie użytkownikowi jednoznacznej odpowiedzi czy dany Trp oddziałuje z białkami AGO.

Podczas klasyfikacji przynależności danego Trp do miejsca oddziaływania z AGO brane są pod uwagę zmienne dotyczące właściwości fizykochemicznych każdego aminokwasu z otoczenia Trp oraz dystans w sekwencji, wyrażony liczbą reszt aminokwasowych, od analizowanego Trp do kolejnych wystąpień tego aminokwasu w kierunku N- i C- końca białka. Najwyższą czułość (94,29%) i specyficzność (99,68%) uzyskano dla motywów długości 21 (tabela 9).

Tabela 9. Ocena dokładności działania programu i-Wsearch przeprowadzona dla różnych długości motywu podczas 10-krotnego sprawdzianu krzyżowego.

Długość motywu ↓	SN (%)	SP (%)	PPV (%)	ACC (%)	miara F
5	93,77	93,22	98,70	97,91	0,9617
7	94,12	95,21	99,09	98,33	0,9667
9	93,77	95,76	99,20	98,32	0,9641
11	93,47	97,22	99,48	98,50	0,9638
13	92,56	97,80	99,59	98,45	0,9595
15	94,05	99,06	99,82	98,88	0,9667
17	93,75	98,75	99,77	98,77	0,9667
19	94,03	99,68	99,94	98,96	0,9689
21	**94,29**	**99,68**	**99,94**	**99,01**	**0,9703**
23	93,96	99,68	99,94	98,95	0,9686
25	93,87	99,30	99,88	98,90	0,968

Pogrubioną czcionką zaznaczono wartość okna 21, która uzyskała najwyższą wartość F.

Zastosowanie trzech deskryptorów aminokwasów - objętości oraz indeksów hydrofilowości i elastyczności nieznacznie zwiększyło czułość (95.2%) metody (patrz Metody - Rozdział 3.1.3).

Na rys. 19 przedstawiono wynik działania programu i-Wsearch na przykładzie białka Tas3 *S. pombe*. Spośród dziewięciu reszt Trp występujących w sekwencji, trzy zostały zaklasyfikowane jako miejsca wiążące AGO.

10 records Search:

Trp(W) position	Motif	AGO-binding activity	Probab. of AGO-binding
46	ervhelftehwlsnlknrrek	no	0.001
175	kssgsmeieswdnstsdsiie	no	0.104
211	irssdsksvgwddnstgfres	no	0.098
254	frpkyqaksswfapddpeasw	no	0.062
264	wfapddpeaswgnlddgwget	yes	0.815
271	easwgnlddgwgetngswsst	yes	0.855
278	ddgwgetngswsstddtkhyk	yes	0.843
291	tddtkhyknewaesinldnfn	no	0.465
441	ttdkndlyinwlkslsffqtn	no	0.005

Showing 1 to 9 of 9 entries

1

Rys. 19. Raport wynikowy działania programu i-Wsearch na przykładzie białka Tas3 *S. pombe*.

4.2.4. Projektowanie *in silico* sekwencji domen w formie gry internetowej

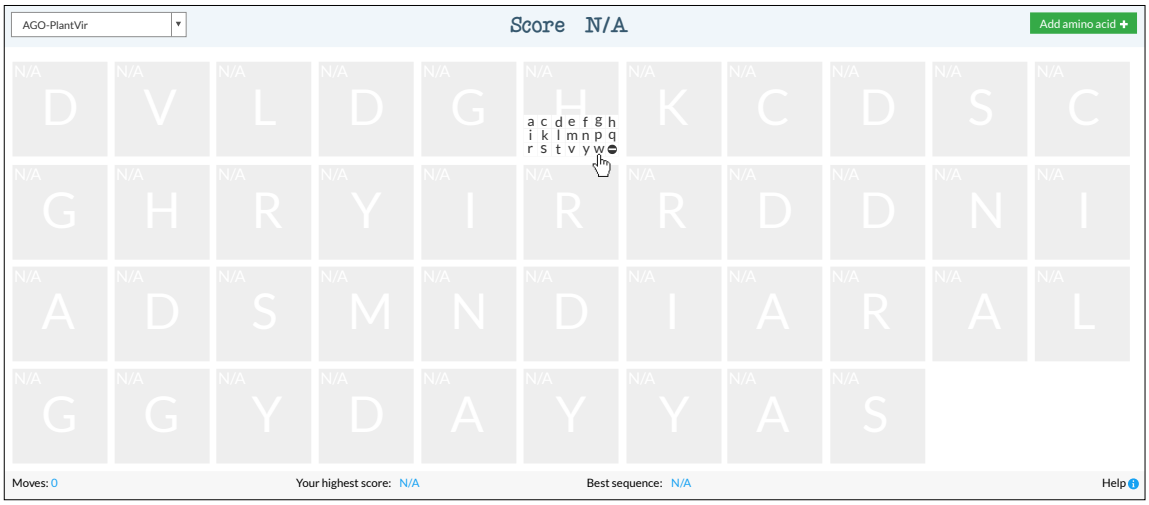
W ciągu ostatnich lat nastąpiła intensyfikacja badań biochemicznych nad szczegółowymi właściwościami domen bogatych w tryptofan, które zaangażowane są w procesy RNAi. Na przykład Szabó i in. (2012) poprzez wprowadzenie dwóch dipeptydów WG i GW do sekwencji niefunkcjonalnej części białka P1 wirusa SPFMV, dokonali jego transformacji w funkcjonalny supresor wyciszający, wykazujący zdolność wiązania białek AGO gospodarza [121]. W innym doświadczeniu grupa badawcza Filipowicza (2011) wykazała, że wprowadzenie przynajmniej czterech reszt Trp do nieustrukturyzowanego regionu białka Sic1p21, które nie bierze udziału w RNAi, spowodowało, że region ten okazał się wystarczający do oddziaływania z kompleksem CCR4-NOT tym samym powodując degradację docelowego mRNA [89]. Pokazano również, że wprowadzanie coraz większej liczby podstawień Trp na Ala w regionie Mid i C-term białek GW182 człowieka i muszki owocowej ma addytywny wpływ na zmniejszanie wydajności procesu wyciszania genów przez kompleksy deadenylacji CCR4-NOT [89].

Aby wspomóc prowadzone badania mutagenezy poprzez dostarczenie informacji na temat spodziewanego wpływu planowanej w eksperymencie mutacji, na stronach Whub wprowadzano interaktywną grę (http://www.comgen.pl/whub/design), która pozwala użytkowniki projektować *de novo* lub modyfikować istniejące domeny WG/GW. Sekwencja białkowa podana przez użytkownika przedstawiona jest w formie kolorowych bloków, które reprezentują poszczególne aminokwasy (rys. 20B-D). Kolor i intensywność bloku odpowiadają wartościom macierzy PSSM dla danego aminokwasu. W ten sposób, preferowane w domenie aminokwasy przybierają barwy czerwone, a reszty o bardzo niskim prawdopodobieństwie wystąpienia oznaczone są kolorem niebieskim.

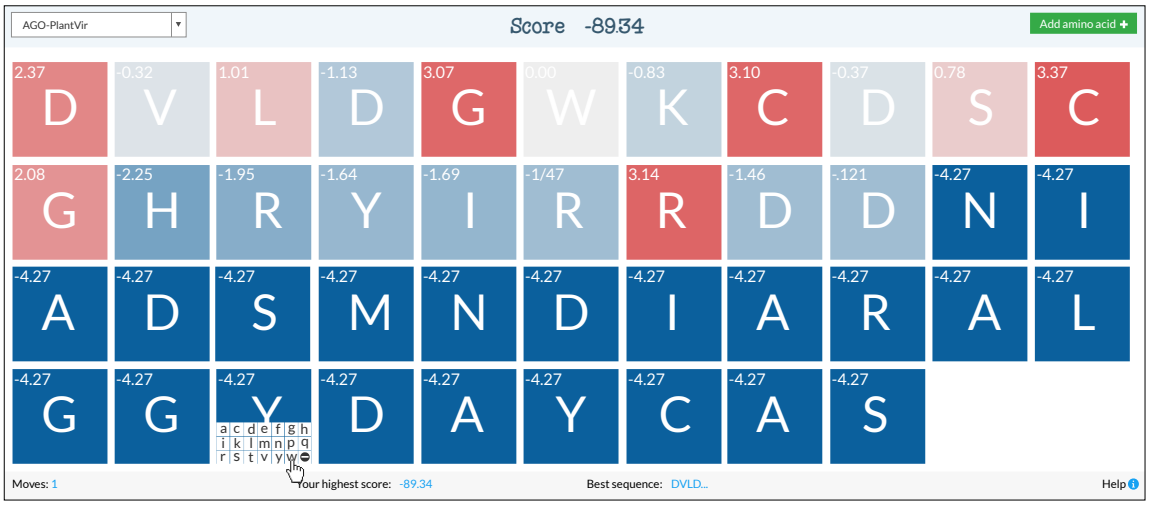
A

```
SPFMV      1 DVLDGHKCDSCGHRYIRRDDNIADSMNDIARALGGYDAYCAS  42
             |..:|..|:.||.:::.:||||:.|:.||:|||||||::.|..:
SPMMV      1 DSEEGWFCEDCGSQFNKRDDNVLDNKNDVARALGGWNEYEDA  42
```

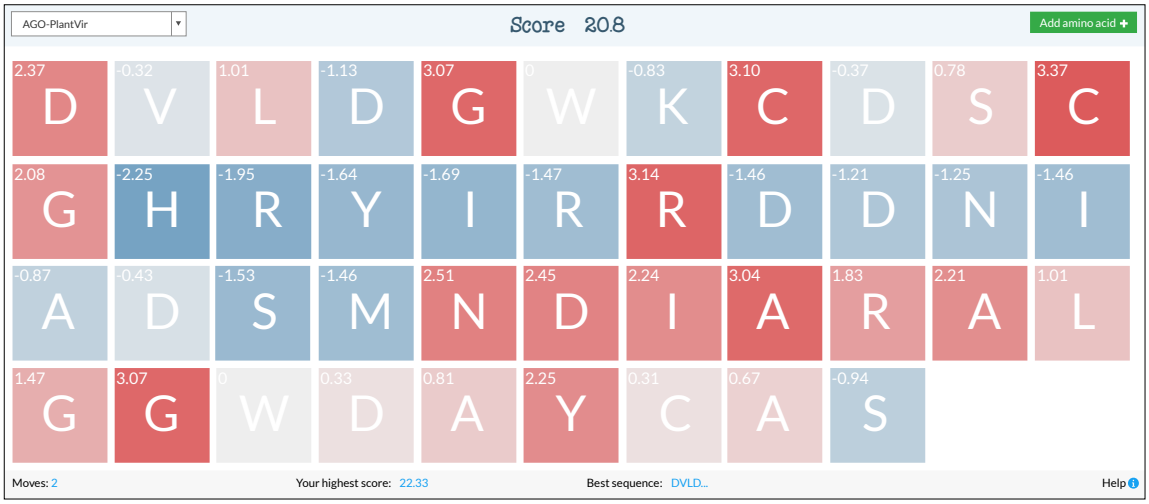
B



C



D



Rys. 20. Odtworzenie eksperymentu Szabó i in. (2012) w formie interaktywnej gry. A. Homologiczne sekwencje białka P1 SPFMV odpowiadające sekwencjom WG/GW i wirusa SPMMV. **B.** Pierwsza substytucja H>W w obrębie sekwencji P1 SPFMV powoduje obliczenie wartości punktacji wynoszącej -89,34. **C.** Bardzo niska wartość ujemna świadczy o barku podobieństwa analizowanej sekwencji z rzeczywistymi domenami wiążącymi AGO. **D.** Druga mutacja Y>W zwiększa wartość *score* do 20,8.

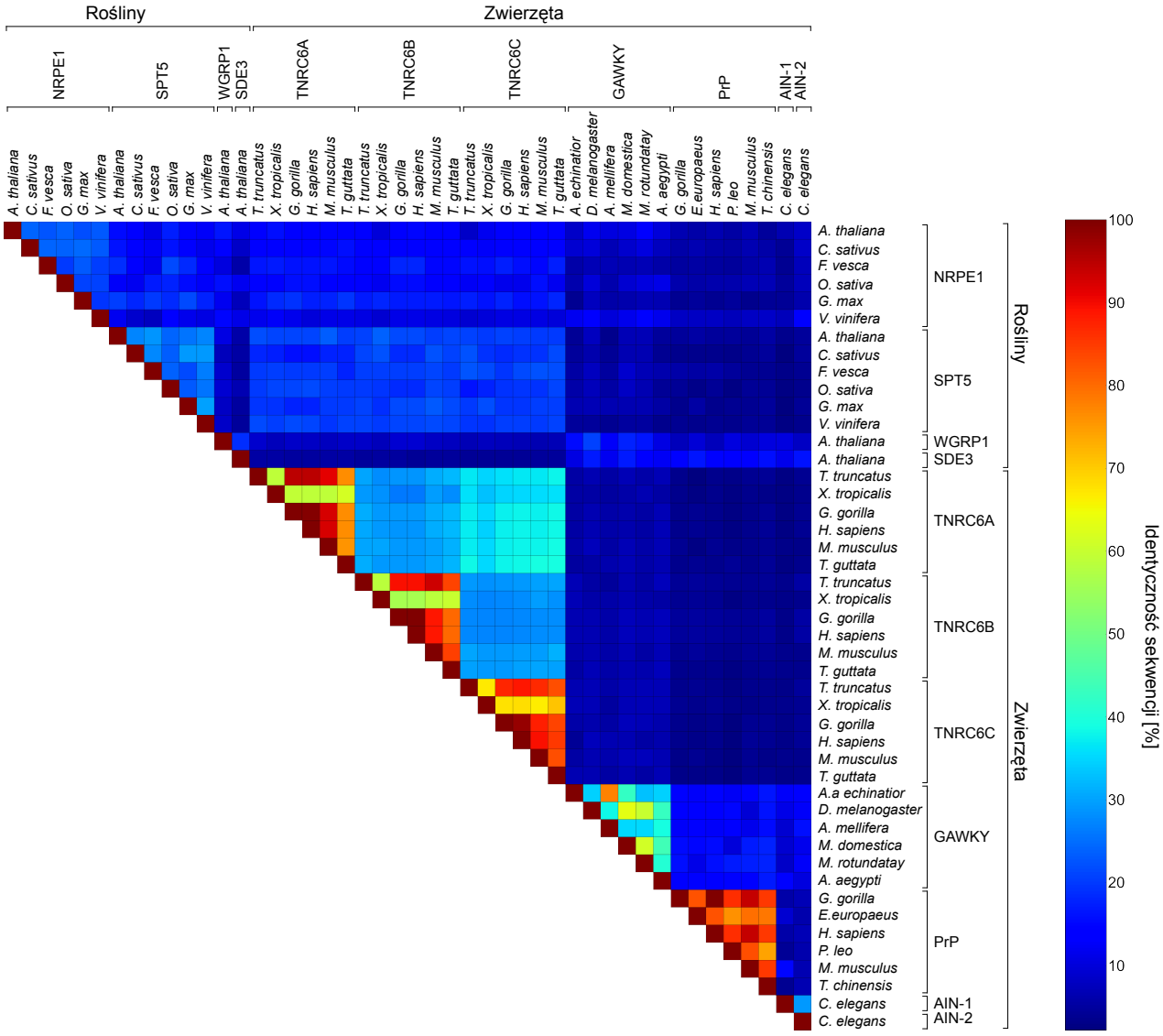
Za pomocą kliknięć i przeciągnięć wybranych bloków użytkownik może w czasie rzeczywistym symulować dowolne mutacje w analizowanej sekwencji: insercje/delecje, substytucje i różnego rodzaju rearanżacje aminokwasów. Każda zmiana w sekwencji wprowadzona przez użytkownika spowoduje dynamiczne obliczenie nowej wartości *score* dla całej domeny oraz odpowiednią zmianę barw i/lub intensywności kolorów bloków. Zatem projektowanie sekwencji funkcjonalnych domen polegać będzie na takim ułożeniu bloków aminokwasów, aby zwiększać natężenie czerwonego koloru w projektowanej domenie i tym samym maksymalizować wartość oceny punktacji domeny.

Na rysunku 20 przedstawiono rekonstrukcję eksperymentu Szabó i in. (2012), który polegał na wymodelowaniu domeny wykazującej aktywność oddziaływania z AGO. Sekwencja wejściowa pokazana na rys. 20B pochodzi z regionu białka P1 wirusa SPFMV, które nie posiada zdolności wiązania białek AGO. Ta pozbawiona reszt tryptofanu sekwencja odpowiada regionowi białka P1 wirusa SPMMV, w obrębie którego znajdują się dwa powtórzenia WG i GW biorące udział w wiązaniu białek AGO1 u różnych gatunków tytoniu (rys. 20A). Eksperyment polegał na wprowadzeniu tych powtórzeń do niefunkcjonalnej sekwencji P1 wirusa SPFMV w odpowiadające im pozycje homologiczne. Wprowadzenie pojedynczego powtórzenia GW było niewystarczające do wiązania białek AGO, dopiero dwa podstawienia umożliwiły wiązanie białek AGO i realizację funkcji supresorowych. W grze, początkowa sekwencja białka P1 SPFMV nie posiada przypisanej oceny punktacji, ponieważ pozbawiona jest reszt Trp. Wprowadzenie pierwszej substytucji His na Trp powoduje obliczenie wartości punktacji. Bardzo niska wartość ujemna -89,13 świadczy o wysokim stopniu niekompatybilności analizowanej sekwencji z rzeczywistymi sekwencjami wiążącymi AGO. Drugie podstawienie Tyr na Try sprawia, że wartości oceny sekwencji wzrasta do 20.8 pół-bitów, gdzie kompozycja sekwencji upodabnia się do W-motywów wirusów infekujących rośliny. A zatem tego typu gra symulacyjna może stanowić przydatne narzędzie prognozowania i planowania laboratoryjnych badań mutagenezy.

4.3. Molekularne mechanizmy powstawania i zmienności domen WG/GW

Najbardziej charakterystyczną cechą domen wiążących AGO jest wysokie zróżnicowanie ich sekwencji, mimo pełnionych przez te białka ściśle określonych funkcji molekularnych w komórce. Na rys. 21 przedstawiono przyrównanie parami sekwencji domen WG/GW na całej długości w obrębie białek roślinnych i zwierzęcych. Stopień zachowania sekwencji jest różny dla roślinnych i zwierzęcych ortologów. U blisko spokrewnionych roślinnych gatunków w obrębie białek NRPE1 i SPT5 średni procent identyczności sekwencji domeny wiążącej AGO4 wynosi odpowiednio 21% i 29%, podczas gdy sekwencja pozostałych części białka wykazuje trzykrotnie

wyższe podobieństwo Oznacza to, że roślinne domeny WG/GW ewoluują szybciej niż pozostałe regiony białkowe, a stopień dywergencji ich sekwencji może osiągnąć punkt, który uniemożliwia wiarygodne określenie ich relacji homologicznych. Z kolei u zwierząt, sekwencje domen wiążących AGO kodowane przez ortologiczne geny są znacznie bardziej zakonserwowane - np. domena WG/GW prionowego białka PrP lub GW182 ssaków utrzymuje identyczność sekwencji na poziomie 87-91%. Jednak zarówno roślinne, jak i zwierzęce domeny wiążące AGO - mimo, że podlegają różnej presji mutacyjnej w obrębie sekwencji ortologicznych - wykazują brak lub bardzo niskie podobieństwo sekwencji między różnymi rodzinami białkowymi nawet w obrębie tego samego gatunku (np. domena WG/GW między białkiem NRPE1 i SPT5 Arabidopsis lub między PrP a GW182 człowieka).



Rys. 21. Mapa termiczna przedstawiająca stopień zachowania sekwencji domen WG/GW obrębie roślinnych i zwierzęcych białek wiążących AGO. Przyrównanie sekwencji domen WG/GW przeprowadzono na całej długości domeny.

Te różnice mogą sugerować działanie różnych mechanizmów molekularnych w tych dwóch

grupach taksonomicznych, a także w obrębie wielu rodzin białkowych, prowadząc do powstawania funkcjonalnie tożsamych domen. Obserwacje te nasuwają wiele pytań dotyczących mechanizmów powstawania tych domen, dynamiki ich ewolucji oraz przyczyn odpowiedzialnych za ich zmienny charakter.

Główny problem prowadzenia jakichkolwiek analiz porównawczych domen WG/GW wynika z ich zmiennej sekwencji. Sprawia to, że zawarta w nich informacja filogenetyczna jest trudna do wykorzystania, a wyniki są mało wiarygodne. Jednym z pomysłów na prowadzenie badań filogenetycznych domeny WG/GW jest rekonstrukcja filogenetyczna tych białek w oparciu o inny dobrze zakonserwowany region sekwencji. Modelem do takich badań mogłaby być wielogenowa rodzina, charakteryzująca się szerokim spektrum kompozycji domen WG/GW w wielu białkach, która dodatkowo posiadałaby inną, dobrze zakonserwowaną domenę białkową. Może ona wtedy posłużyć jako wektor stanowiący podstawę do rekonstrukcji filogenetycznej domen WG/GW. Podobne metody użycia sekwencji wysoko zakonserwowanej domeny do analiz filogenetycznych innej, słabo zakonserwowanej domeny, stosowane są w coraz większej liczbie badań [122].

Wśród białek potencjalnie wiążących AGO (patrz: Rozdział 4.1 - tabela 4 i 5) zidentyfikowane zostały sekwencje kodowane przez wielogenową rodzinę hnRNP. Rodzina ta stanowi idealny model do badań zmienności domeny wiążącej AGO, ponieważ jej członkowie, oprócz wysoko zmiennej sekwencji WG/GW, zawierają domenę RRM, która jest najczęściej występującym motywem wiązania RNA wykazującym wysokie zachowanie sekwencji u niemal wszystkich organizmów żyjących na Ziemi [123]. Domena RRM obecna jest także w wielu innych rodzinach białkowych biorących udział w procesach RNAi, np. w zwierzęcych białkach GW182 [32], ludzkim białku RBM4, które wchodzi w skład kompleksów zawierających Ago2 [124,125] oraz białkach regulujących kwitnienie u Arabidopsis, FCA i FPA [126]. Ponadto funkcja wiązania AGO została niedawno potwierdzona doświadczalnie w zaproponowanym w tabeli 4 białku hnRNP Arabidopsis [Lagrange i in., dane nieopublikowane]. Dlatego analiza mechanizmów molekularnych zaangażowanych w różnicowanie się domen WG/GW została przeprowadzona w oparciu o rekonstrukcję filogenetyczną domeny RRM w proteomach wybranych przedstawicieli roślin: dwuliściennych (*A. thaliana*), jednoliściennych (*O. sativa*), mszaków (*P. patens*) oraz zielenic (*C. reinhardtii*).

4.3.1. Tandemowe i segmentowe duplikacje genów oraz alternatywny splicing

Wraz ze wzrostem poziomu złożoności budowy organizmów roślinnych zwiększa się liczba białek zawierających domenę RRM, a także liczba kopii motywu RRM występujących w białkach (tabela 10). Wzrostowi liczby białek zawierających 2 i 3 kopie domeny RRM towarzyszy

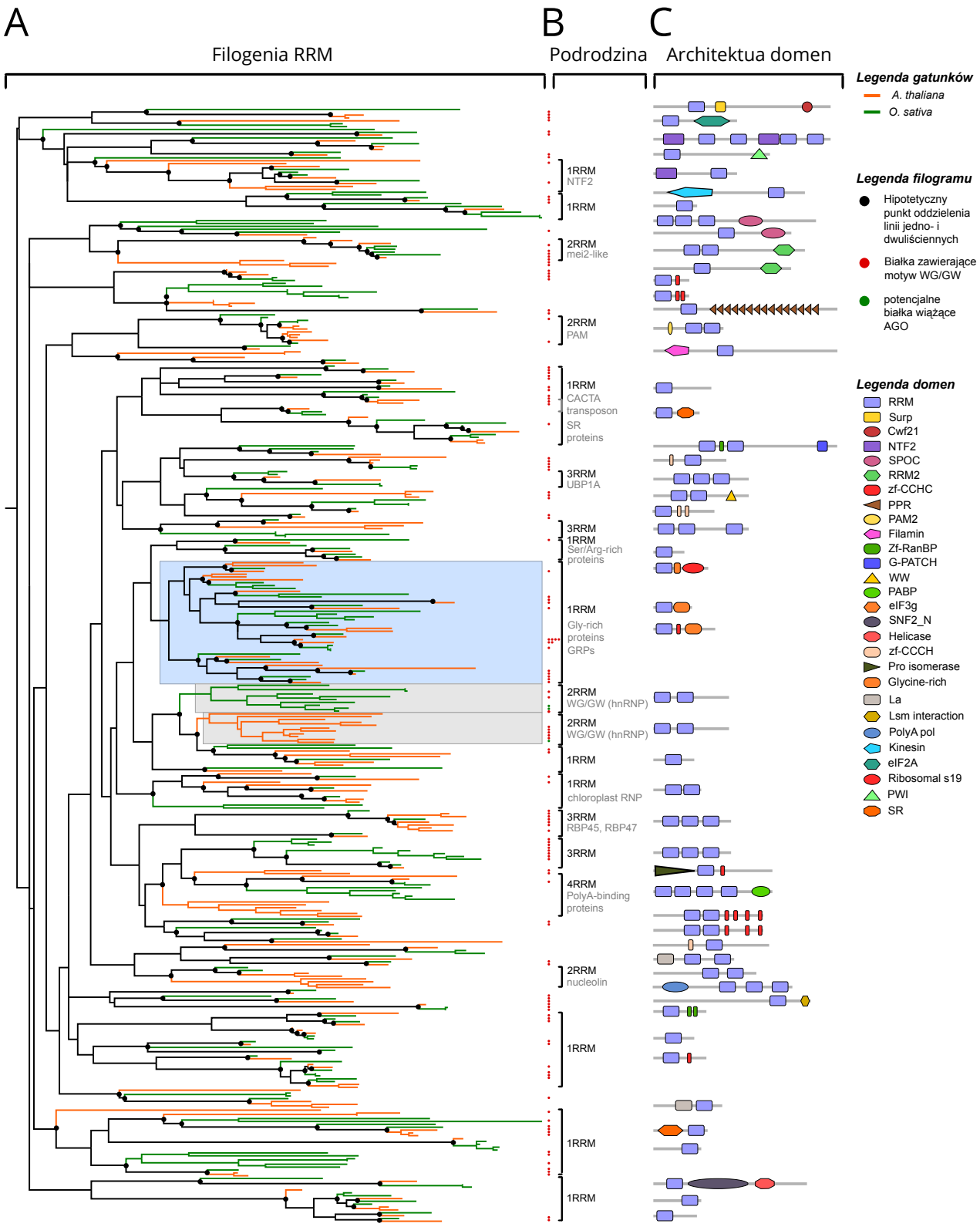
zmniejszenie liczby białek z pojedynczym motywem RRM. U ponad 1/3 białek zawierających domenę RRM u Arabidopsis i ryżu, motyw ten występuje w sąsiedztwie innych domen białkowych. Większość z nich bierze udział w wiązaniu DNA/RNA lub białek - motyw palca cynkowego typu CCHC, RanBP i CCCH: 30.7%, NTF2: 10.4%, PABP: 8.7%, PAM2: 7.8%, SPOC: 4.3%. Niemal 1/3 białek RRM posiadających inne domeny, zawiera na C-końcu krótkie powtórzenia kilkuaminokwasowe, które tworzą oddziałujące z innymi białkami modularne domeny takie jak: domeny glicynobogate (26.1%), seryno/arginino-bogate domeny (8%), domeny WW (2%) i domeny złożone z powtórzeń ankirynowych (ang. *ankyrin repeat domains*) (2%).

Tabela 10. Białka zawierające co najmniej jedną domenę RRM u wybranych gatunków roślin.

Gatunek roślinny	Liczba genów RRM	Liczba wszystkich genów	Liczba białek z różną liczbą kopii domeny RRM [%]					
			1	2	3	4	5	6
C. reinhardtti	85	16 709	72.9	20.0	4.7	2.4	0.0	0.0
P. patens	182	35 938	67.0	20.9	8.8	2.7	0.5	0.0
O. sativa	241	40 577	63.1	25.8	9.0	1.7	0.2	0.2
A, thaliana	230	27 380	57.8	28.6	10.7	2.6	0.3	0.0

Regiony glicynobogate występują we wszystkich organizmach - od sinic aż po człowieka - i zawierają duże ilości glicyny (20-70%), która obecna jest w powtarzających się motywach aminokwasowych tworzących elastyczne miejsca interakcji z innymi białkami. Domeny SR, bogate w powtórzenia Ser i Arg, pośredniczą w oddziaływaniu z innymi białkami, również posiadającymi domeny SR podczas formowania spliceosomu, eksportu mRNA z jądra komórkowego i w translacji [127]. Domena WW charakteryzuje się obecnością dwóch reszt Trp rozdzielonych w sekwencji o kilka aminokwasów, które biorą udział w wielu interakcjach z różnymi białkami zawierającymi krótkie motywy bogate w prolinę [128]. Również powtórzenia ankirynowe, występujące powszechnie w wielu rodzinach białkowych, składają się z tandemowo powtórzonych 33-aminokwasowych motywów bogatych w dipeptydy AR, które tworzą rusztowanie dla wielu interakcji białko-białko [129].

W rodzinie RRM występują także powtórzenia WG/GW zarówno o potwierdzonej, jak i przypuszczalnej funkcji wiązania AGO. Przeszukanie genomów roślinnych pod kątem potencjalnych domen WG/GW, przeprowadzone przy użyciu programu Wsearch, pozwoliło na ustalenie pierwszego wystąpienia w królestwie roślin białka RRM-WG/GW u *Physcomitrelli patens*. Wyniki te sugerują, że zarówno powstanie domeny WG/GW w rodzinie białek RRM, jak i generalny wzrost złożoności tej rodziny, miał miejsce po rozdzieleniu się linii ewolucyjnych glonów i roślin naczyniowych. co najprawdopodobniej towarzyszyło ekspansji roślin na ląd.



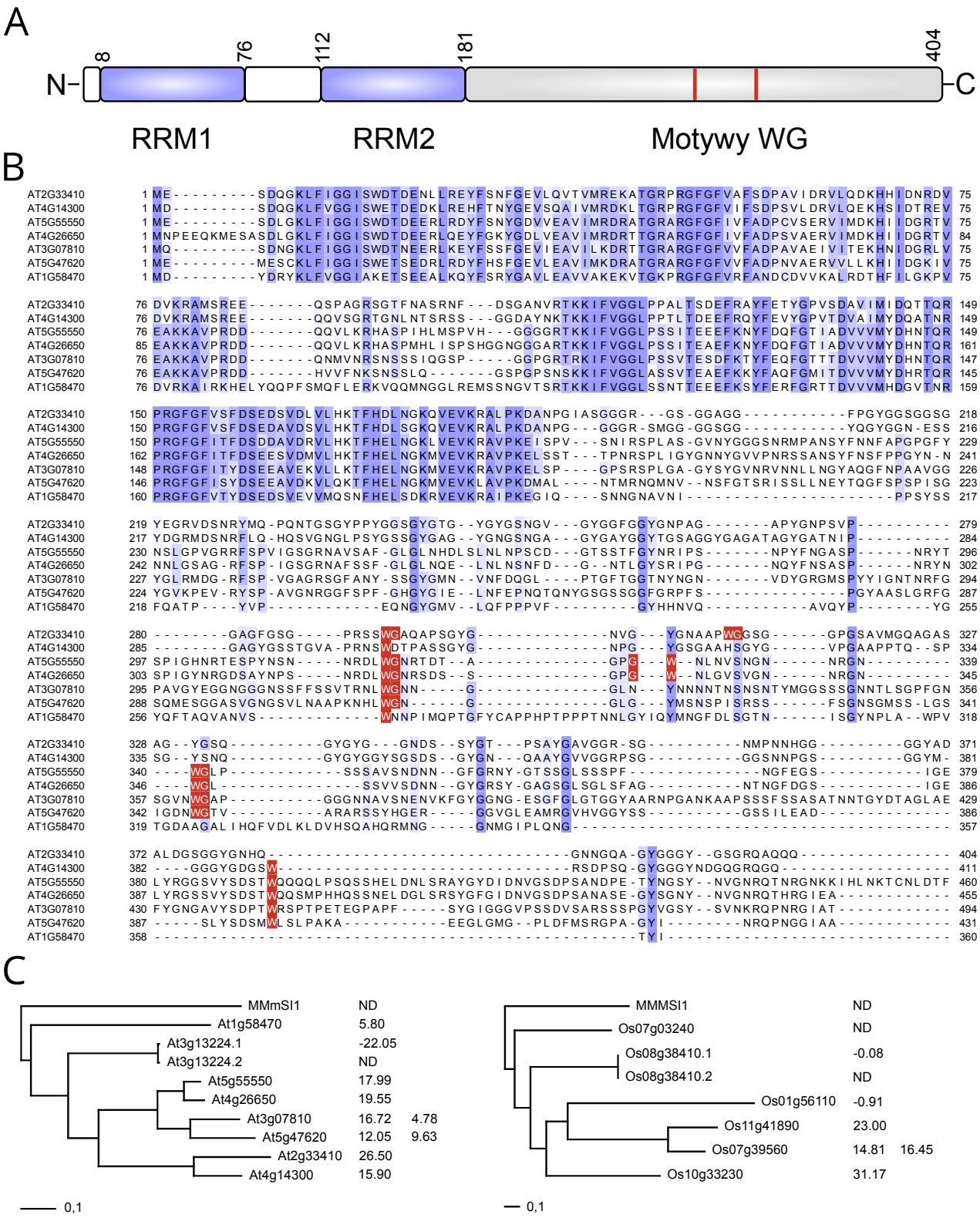
Rys. 22. Rekonstrukcja filogenetyczna oraz architektura domen białek posiadających domenę RRM w genomie *A. thaliana* i *O. sativa*. A. Filogram skonstruowany został na bazie dopasowania domeny RRM N-końca kodowanej przez geny Arabidopsis (pomarańczowe) i ryżu (zielone). Wypełnione czarne punkty znajdujące się na węzłach drzewa wyznaczają hipotetyczny punkt rozejścia się linii ewolucyjnych roślin jedno- i dwuliściennych. **B.** Opisy podrodzin przygotowano w oparciu o liczbę kopii domeny RRM oraz adnotacje genów zawarte w bazach TAIR i MSU. Kwadraty oznaczają wystąpienie motywu WG/GW w sekwencji w innym regionie niż zadnotowane domeny. Czerwone kwadraty oznaczają białka, które uzyskały niską punktację domeny WG/GW, natomiast zielone kwadraty wskazują na białka o potwierdzonej funkcji wiązania AGO lub zawierające statystycznie istotną domenę WG/GW. **C.** Schematyczna architektura domen sekwencji wyznaczona w oparciu o bazę Pfam.

Ze względu na zmienną liczbę kopii domeny RRM oraz występowanie różnych domen białkowych w tej rodzinie, do rekonstrukcji filogenetycznej między białkami *A. thaliana* i *O. sativa* została wykorzystana sekwencja pierwszej domeny RRM na N-końcu (rys. 22A). Większość sekwencji ze względu na liczbę posiadanych kopii motywu wiążącego RNA grupuje się na drzewie tworząc odrębne grupy monofiletyczne (rys. 22B). Wynika z tego, że sekwencje analizowanych białek posiadających dwie lub więcej domen RRM pochodzą od genu wspólnego przodka, który obecny był przed rozejściem się linii ewolucyjnych prowadzących do *A. thaliana* i *O. sativa*. Według przedstawionego na filogramie ciągu zdarzeń ewolucyjnych, obecnych jest 186 hipotetycznych punktów rozejścia się linii Arabidopsis i ryżu, co w przybliżeniu odpowiada liczbie białek RRM u wspólnego przodka roślin jedno- i dwuliściennych (rys. 22A). Liczba ta, bardzo zbliżona do 182 białek RRM występujących u *P. patens*, sugeruje, że rodzina białek zawierających domenę RRMs była względnie duża przed podziałem roślin na jedno- i dwuliścienne i nadal ulega dynamicznej ekspansji zarówno u *A. thaliana*, jak i *O. sativa*. Powtórzenia sekwencji WG/GW są na tyle krótkie, że występują przypadkowo niemal we wszystkich grupach monofiletycznych niższego rzędu (rys. 22B). Jednak w obu roślinach - Arabidopsis i ryżu - dwie siostrzane grupy białek RRM-WG/GW (hnRNP), oznaczone na rys. 22A szarymi polami, obejmują zidentyfikowane białka wiążące AGO - białko Arabidopsis (At2g33410) i dwa białka ryżu (Os10g33230 i Os11g41890). Obie rodziny RRM-WG/GW składają się z podobnej liczby członków i powstały na drodze duplikacji genów po rozdzieleniu się roślin jedno- i dwuliściennych. Zbieżne układy ekspansji rodziny RRM-WG/GW zachodzące niezależnie między ryżem i Arabidopsis mogą sugerować podobieństwa w funkcjonowaniu tej rodziny oraz analogicznych mechanizmach selekcji w obu roślinach.

Ekspansja rodziny genów RRM-WG/GW u A. thaliana i O. sativa

Układ domen w białkach RRM-WG/GW tworzą dwie kopie motywu RRM, oddzielone od siebie mniej zachowaną sekwencją łącznika długości ok. 20-40 reszt aminokwasowych oraz występującą na C-końcu białka nieustrukturyzowaną sekwencją (rys. 23A). Sekwencja C-końca, w odróżnieniu od zakonserwowanych sekwencji dwóch domen RRM, wykazuje wysoką zmienność i zawiera powtórzenia WG/GW (rys. 23B). Region N-końca, obejmujący dwie kopie domeny RRM, wykazuje największe podobieństwo sekwencji do białka Musashi, które występuje u wielu organizmów zwierzęcych i zaangażowane jest w regulację translacji poprzez wiązanie do specyficznych sekwencji znajdujących się w 3' UTR docelowych transkryptów [130].

Ponieważ rodziny RRM-WG/GW u Arabidopsis i ryżu dzieli różna historia ewolucyjna, przedstawiony filogram może w pełni nie odzwierciedlać relacji homologicznych między tymi genami, a także prawidłowego kierunku zachodzenia zmian ewolucyjnych.



Rys. 23. Rodzina białek RRM-WG/GW (hnRNP). A. Schematyczne ułożenie domen białkowych AT2G33410. Dwie kopie domeny RRM rozdzielone są sekwencją łącznika. Domena WG/GW występuje na C-końcu białka. Pionowymi liniami zaznaczono wystąpienie powtórzeń WG/GW. **B.** Przyrównanie pełnej długości sekwencji członków rodziny RRM-WG/GW Arabidopsis. Niebieskim kolorem podświetlono pozycje dopasowania, w których poziom identyczności >= 40%. Region ten odpowiada domenom RRM. Czerwonym kolorem zaznaczono powtórzenia tryptofanu występujące w sekwencjach. **C.** Filogenia rodziny RRM-WG/GW odpowiednio u *A.thaliana* i *O. sativa*. Sumaryczne wartość *score* uzyskano wykorzystując program Wsearch.

W celu przeprowadzenia bardziej dokładnej analizy, skonstruowano dwa drzewa filogenetyczne rodziny RRM-WG/GW oddzielnie dla obu gatunków roślin (rys. 23C). Topologie dwóch otrzymanych filogramów wykazują duże podobieństwo - sześć genów powstałych na drodze duplikacji umiejscowionych jest w stosunku jeden do jeden w podobnej lokalizacji w kladzie. Ponadto podobny układ powtórzeń WG/GW występujących między inparalogami jest również zachowany między dwoma gatunkami, gdzie wartości *score* W-motywów są bardzo zbliżone między koortologami. Dwa mechanizmy, duplikacje tandemowe i całego genomu (WGD, ang. *Whole Genome Duplication*), przyczyniły się do ekspansji rodziny RRM-WG/GW u Arabidopsis i ryżu. Filogram Arabidopsis (rys. 23C) określa trzy pary genów: At4g26650/At5g55550, At3g07810/At5g47620 i At2g33410/At4g14300. Mapa duplikacji genów Arabidopsis [131] wskazuje, że pierwsza para powstała podczas ostatniego zdarzenia WGD, natomiast pozostałe dwie pary są wynikiem wcześniejszego WGD. W przypadku ryżu, według ostatnio przeprowadzonej analizy duplikacji segmentowych [132] tylko jeden członek rodziny hnRNP, Os07g03240, który pozbawiony jest powtórzeń WG/GW, powstał na drodze wielkoskalowych duplikacji. Pozostałe geny RRM-WG/GW ryżu są najprawdopodobniej wynikiem lokalnych duplikacji, które odegrały znaczną rolę w różnicowaniu się tej rodziny (rys. 23D). Nawet w obrębie najbliżej spokrewnionych genów sekwencja C-końca białek RRM-WG/GW jest bardzo zmienna przy jednoczesnym zachowaniu domen RRM - na przykład w obrębie dwóch par inparalogów, Os11g41890/Os07g39560 i At2g33410/At4g14300, uzyskane wartości *score* W-motywów są różne. Warto zauważyć, że podobną sytuację, w której mechanizm duplikacji genów zapewnia wariantywność jedynie w obrębie powtórzeń WG/GW obserwuje się wśród paralogów zwierzęcych białek GW182. Szczególnie interesującym przykładem par paralogów są roślinne białka NRPE1, SPT5/KTF1 i SPT6/GTB, w obrębie których jeden gen nie koduje domeny WG/GW, natomiast druga kopia genu zawiera domenę wiążącą AGO. W tym układzie domena wiążąca AGO kodowana jest w całości przez odrębny egzon, co sugeruje, że mogła ona powstać na drodze mechanizmu tzw. "tasowania egzonów" (ang. *exon shuffling*).

Alternatywny splicing jest kolejnym mechanizmem, który warunkuje zmienność białkom RRM-WG/GW w obrębie powtórzeń WG/GW przy jednoczesnym stałym zachowaniu domen RRM. W przypadku dwóch par wariantów transkryptów Arabidopsis (At3g13224.1/At3g13224.2) i ryżu (Os08g3841.1/Os08g3841.2), krótszy produkt białkowy pozbawiony jest motywów WG/GW. Również w przypadku trzech innych genów, At3g07810, At5g47620 oraz Os07g39560, alternatywy splicing ma wpływ na występowanie motywów WG/GW na C-końcu tych białek. Przekłada się to na zmienne wartości *score* tych powtórzeń. Podobny mechanizm występuje u dwóch wariantów splicingowych - mające swoje potwierdzenie w sekwencjach EST i cDNA - genu GTB1, należącego do rodziny czynników transkrypcyjnych SPT6. Produkty białkowe

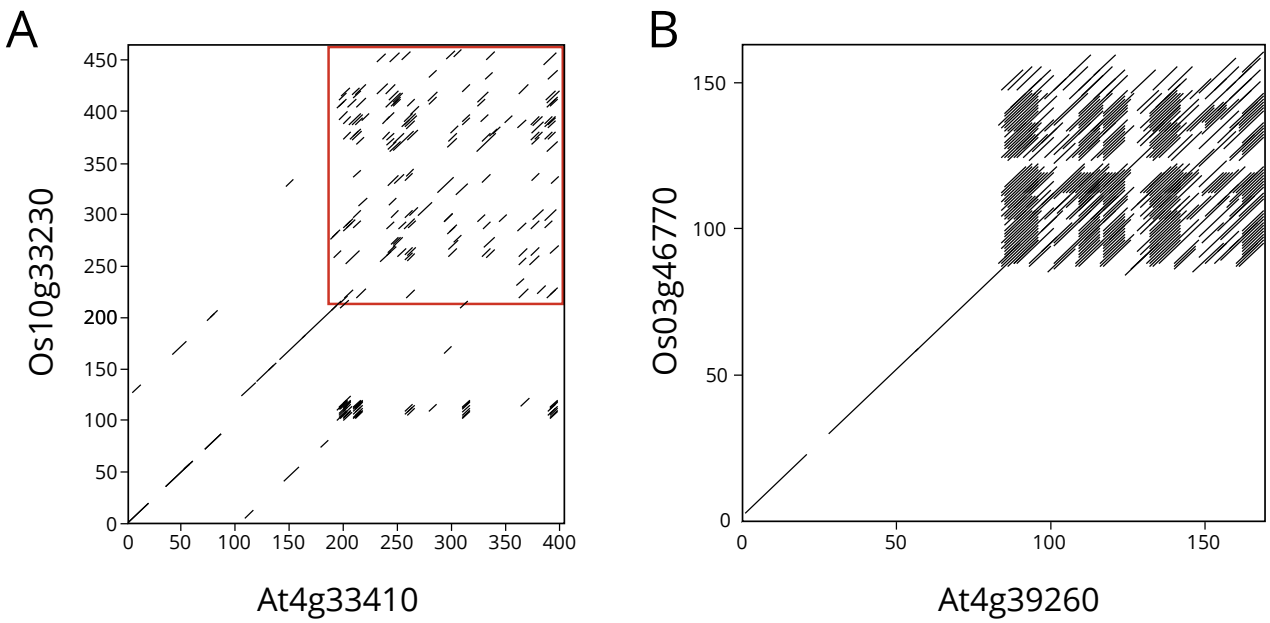
dwóch wersji tych transkryptów są identyczne na N-końcu, lecz różnią się pod względem długości sekwencji jedynie w obrębie domeny WG/GW.

Podsumowując, ekspansja rodziny RRM-WG/GW u Arabidopsis i ryżu wydaje się zależeć od dwóch przeciwstawnych procesów ewolucyjnych - kombinacji ewolucji dywergentnej i konwergentnej. Z jednej strony, wysoka dywergencja sekwencji białek RRM-WG/GW prowadzi do punktu, który uniemożliwia wiarygodne określenie relacji homologicznych sekwencji C-końca przy jednoczesnym wysokim stopniu zachowaniu domen RRM. Z drugiej strony, podobna liczebność rodziny oraz topologia drzewa, a także podobieństwo występowania powtórzeń WG/GW odzwierciedlają zakonserwowany układ ekspansji tej rodziny, który z dużym prawdopodobieństwem może pełnić bardzo podobne funkcje w obu organizmach. Układ taki może być wynikiem wspólnej konieczności adaptacji sekwencji, w której nacisk selekcyjny powoduje powstawanie szerokiego spektrum domen WG/GW reprezentujących różną charakterystykę kompozycji aminokwasów.

4.3.2. Tempo mutacji niesynonimicznych i synonimicznych

Zgodnie z topologią drzewa filogenetycznego białek RRM u Arabidopsis i ryżu (rys. 22A) domena RRM białek RRM-WG/GW jest najbliżej spokrewniona z białkami glicynobogatymi (RRM-GRP), które oznaczono na drzewie filogenetycznym niebieskim polem (rys. 22A). Kilka z tych białek otrzymało dodatnie wartości oceny domeny WG/GW, co odzwierciedla wysoki stopień podobieństwa składu aminokwasów tych sekwencji z sekwencjami rzeczywistych domen WG/GW. Ponadto zidentyfikowane w poprzednich rozdziałach potencjalne domeny WG/GW występują w wielu innych sekwencjach GRP zarówno wśród białek roślinnych, jak i zwierzęcych (patrz: Rozdział 4.1. - tabela 4 i 5).

Podobnie jak domeny WG/GW, regiony bogate w glicynę występują w różnych rodzinach białkowych w niemal wszystkich organizmach. Reszty glicyn poprzeplatane w sekwencji między nieglicynowymi resztami tworzą w przestrzeni elastyczne konformacje pozwalając na interakcję z wieloma innymi białkami. Podobnie domeny zawierające Trp mogą oddziaływać w różnych kompleksach efektorowych RNAi (np. RISC, RITS, CCR4-NOT). Również tak, jak w przypadku rodziny RRM-WG/GW, w obrębie białek RRM-GRP domena RRM wykazuje wysokie podobieństwo, podczas gdy C-końcowy region sekwencji jest mniej zachowawczy i jest złożony z licznych powtórzonych sekwencji (rys 24). W domenach GRP, powtórzone sekwencje długości ok. 7-9 reszt aminokwasowych występują tandemowo wzdłuż C-końca białka. Jednak w obrębie domen WG/GW, powtórzenia sekwencji długości 10-20 reszt aminokwasowych rozproszone są w mniej regularny sposób (rys. 24A).



Rys. 24. Porównanie metodą dot-matrix par sekwencji białkowych *A. thaliana* i *O. sativa*. W rodzinach **A.** RRM-WG/GW i **B.** RRM-GRP.

Chcąc dokładniej opisać kompozycję aminokwasów w obu domenach, w tabeli 11 przedstawiono częstości występowania aminokwasów w regionach GRP i WG/GW białek RRM oraz dla wszystkich białek Arabidopsis i ryżu.

Tabela 11. Skład aminokwasowy domen GRP, WG/GW w białkach RRM oraz w zestawie białek TAIR.

Charakter reszt	Aminokwas	GRP [%]	WG/GW [%]	TAIR9 [%]
Małe niepolarne	G	36,2	21,3	6,4
	A	6	6,5	6,4
	S	9,7	15,2	9,1
	T	1,8	4	5,1
Polarne	N	3,6	8,1	4,4
	Q	1,9	3,8	3,5
	H	0,6	0,6	2,3
Naładowane	D	6,2	8,1	5,4
	E	3,5	4,8	6,7
	K	1,8	7	6,4
	R	9,2	4,4	5,4
Hydrofobowe	C	0,3	0,4	1,8
	V	2,7	2,3	6,7
	I	0,7	0,7	5,3
	L	1,8	1,1	9,5
	P	2,6	3,2	4,8
	F	2,4	1,4	4,5
	M	0,7	0,3	2,5
	Y	7,8*	0,4	2,8
	W	0,5	6,3*	1,2

* Odwrotne proporcje występowania reszt Trp i Tyr w domenie GRP i WG/GW.

Domeny GRP i WG/GW wykazują podobną kompozycję aminokwasów ($\chi 2 = 25{,}84$; $p = 0{,}13$), która istotnie odbiega ($\chi 2 = 20{,}58$; $p < 0.005$) od generalnego składu aminokwasów w białkach roślinnych. Obie domeny wykazują wysoką zawartość polarnych reszt (GRP: 59.8%, WG/GW: 59.5%) oraz znaczny spadek aminokwasów hydrofobowych (GRP: 19.5%, WGRP: 16.1%). Zgodnie z oczekiwaniami Gly jest najczęściej występującym aminokwasem w obu domenach, GRP (36,2%) i WG/GW (21,3%), podczas gdy w niespokrewnionych białkach występuje 3-5 razy rzadziej. Najmniej oczekiwane aminokwasy w obu domenach, które występują kilkakrotnie częściej w innych białkach, obejmują reszty Cys (0,3% - 0,4%), Met (0,33% - 0,7%), His (0.6% - 0,7%), Ile (0,7% - 0,7%), Leu (1,1% - 1,8%) i Val (2,3% - 2,7%).

Pomimo tych ewidentnych podobieństw w kompozycji aminokwasów obu domen, istnieją jednak istotne różnice. Chociaż w obu domenach często pojawiają się hydrofilowe reszty, to kierunek ten jest jednak bardziej wyraźny w domenach WG/GW, w których zawartość reszt aminokwasowych Ser, Lys, Gln, Asn jest przynajmniej dwukrotnie większa niż w regionach bogatych w glicynę. Spośród hydrofobowych aminokwasów, które stanowią jedną piątą kompozycji domen GRP i WG/GW, Trp jest najczęstszym aminokwasem (39%) w domenach WG/GW, podczas gdy jego udział w domenach GRP jest kilkanaście razy mniejszy i stanowi zaledwie 2,5% wszystkich hydrofobowych reszt. W analogiczny sposób Tyr jest najczęstszym hydrofobowym aminokwasem w domenach GRP (40%), natomiast w domenach WG/GW pojawia się rzadko stanowiąc 2,5% hydrofobowych reszt. Tryptofan kodowany przez jeden kodon, jest najrzadziej występującym i ulegającym substytucjom aminokwasem. Według macierzy substytucji BLOSUM62, najczęściej wykorzystywanej podczas przyrównywania sekwencji spokrewnionych białek [133], szansa, że Trp nie ulegnie podstawieniu na żaden inny aminokwas wynosi 82%, a podstawienia Trp na Tyr lub Phe są jedynymi substytucjami bardziej prawdopodobnymi w toku ewolucji białek niż wynikałoby to z przypadku. Wystąpienie mutacji niesynonimicznej w kodonie tryptofanu, najprawdopodobniej (20% szans) wywoła substytucję Trp-Tyr. Podobnie, reszty Phe, Trp i His najczęściej wymieniają Tyr w przypadku substytucji niesynonimicznych w obrębie jej kodonu.

Podobna kompozycja aminokwasów wchodzących w skład obu domen, WG/GW oraz GRP, sugeruje zachowanie specyficznych właściwości fizykochemicznych, które tworzą bardziej plastyczne regiony sekwencji dla powtórzeń Trp i Gly w domenach WG/GW, lub w przypadku domen GRP reszt Tyr i Gly. Skład aminokwasów obu domen, istotnie odbiegający od generalnych częstości występowania aminokwasów może również sugerować, że selekcja odgrywa bardziej znaczącą rolę w utrzymywaniu specyficznej kompozycji aminokwasowej, niż mechanizmy ewolucji neutralnej.

W celu zbadania tempa ewolucji sekwencji domen WG/GW, obliczone zostały częstości

zmian synonimicznych (*Ks*) oraz niesynonimicznych (*Ka*) dla sekwencji DNA kodujących tę domenę. Tempo zachodzenia substytucji obliczono również w sekwencjach nukleotydowych domen GRP oraz RRM. Ze względu na fakt, że domena WG/GW została znaleziona u kilku przedstawicieli rodziny hnRNP, podczas gdy pozostali członkowie nie posiadają powtórzeń WG/GW lub wykazują niższy stopień podobieństwa kompozycji do domen wiążących AGO, stosunek *Ka/Ks* został obliczony dla tych grup oddzielnie. Ponieważ nie stwierdzono istotnych różnic w obu rozkładach *Ka/Ks* (Test t Welcha dla dwóch prób: $p = 0.9$; F test: $p = 0.78$), nie ma podstaw do oddzielnego traktowania obu tych grup, co dodatkowo świadczy o tym, że ewolucja wszystkich członków rodziny hnRNP podlega zbliżonej selekcji mutacyjnej.

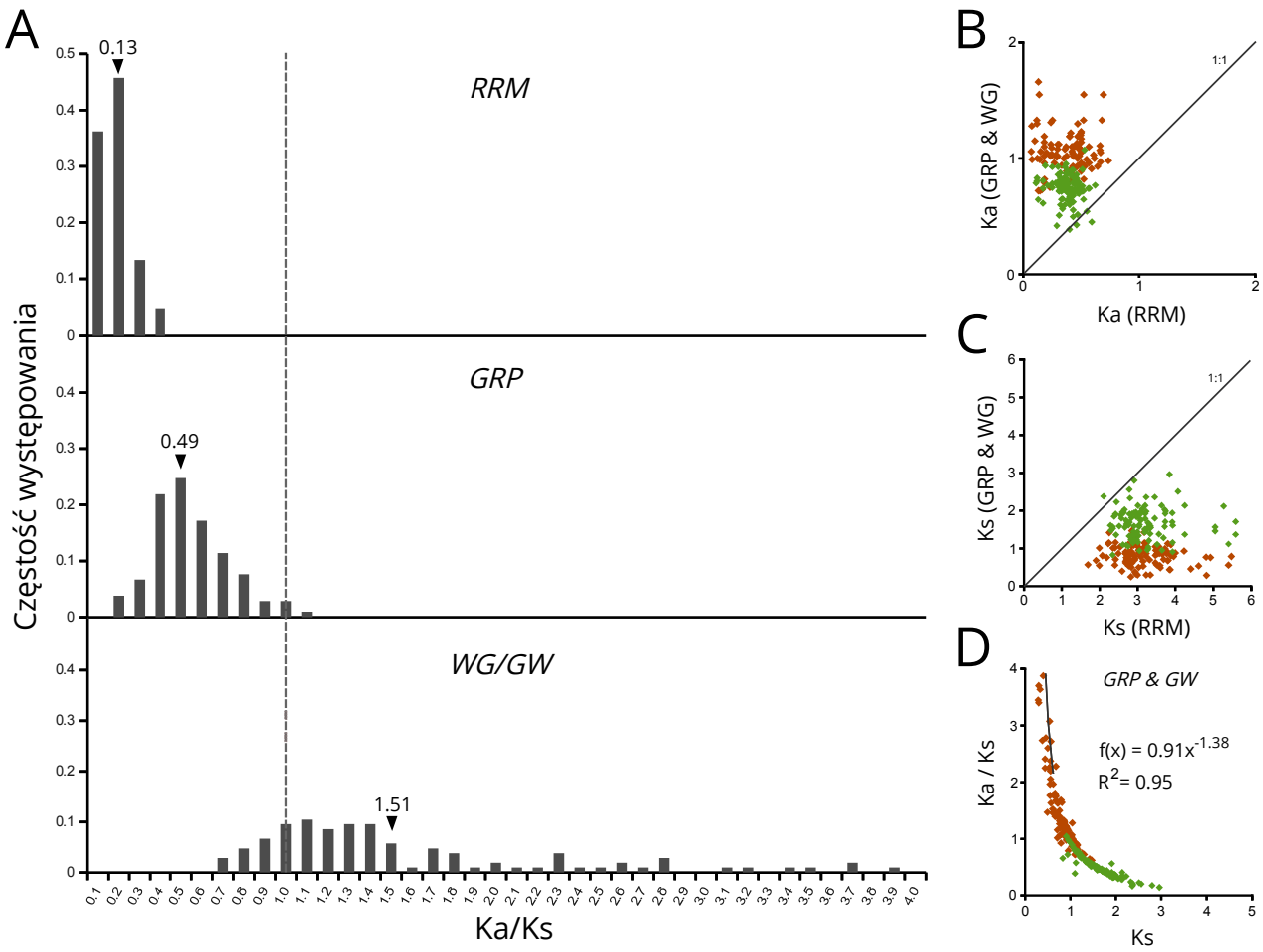
Rozkłady wartości *Ka/Ks* między różnymi domenami, RRM, GRP i WG/GW, wykazują istotne różnice (tabela 12). Wynika stąd, że te trzy regiony znajdują się pod działaniem różnej presji selekcyjnej.

Tabela 12. Porównanie rozkładów wartości *Ka/Ks* między domenami RRM, GRP i WG/GW.

	Test *t* Welcha	Test F
RRM versus GW	1.79E-023	3.41E-077
RRM versus GRP	4.74E-039	1.93E-031
GRP versus GW	1.79E-023	2.62E-040
RRM versus RRM	0.39028	0.00003

W obrębie wszystkich przyrównań par sekwencji domeny RRM, średnia wartość stosunku tempa mutacji niesynonimicznych i synonimicznych wynosi 0,13 w białkach GRP i hnRNP. Ponadto, nie odnotowano statystycznie istotnych różnic (tabela 25) w wartościach *Ka/Ks* domeny RRM między dwoma rodzinami białek GRP i WG/GW, dlatego rozkład ich wartości umieszczono na jednym wykresie (rys. 25A). Region ten charakteryzuje niemal 8-krotnie wyższe tempo zachodzenia substytucji synonimicznych niż niesynonimicznych, tak więc domena RRM podlega selekcji stabilizującej, w wyniku której eliminowana jest większość zmian niesynonimicznych i jedynie nieliczne z nich zostają utrwalone w genomie. Ta negatywna selekcja działa na domeny RRM w obu rodzinach białkowych (SD = 0,05), najprawdopodobniej ze względu na zakonserwowaną funkcję wiązania RNA przez te białka. Ponadto rozkład wartości *Ka/Ks* dla domeny RRM przyjmuje bardzo wąski zakres, w którym wartość mediany zbliżona jest wartości średniej, co świadczy o ustabilizowanym wpływie selekcji mutacyjnej. W porównaniu do domeny RRM, tempo zachodzenia zmian niesynonimicznych wewnątrz sekwencji GRP i WG/GW jest wyższe (rys. 25B), a jednocześnie tempo zmian synonimicznych jest niższe (rys. 25C). Rozkład wartości *Ka/Ks* domeny GRP, podobnie jak w przypadku domeny RRM odznacza się zbliżonymi wartościami mediany i średniej (0,49), lecz jest on prawostronnie skośny (SD = 0,19) obejmując

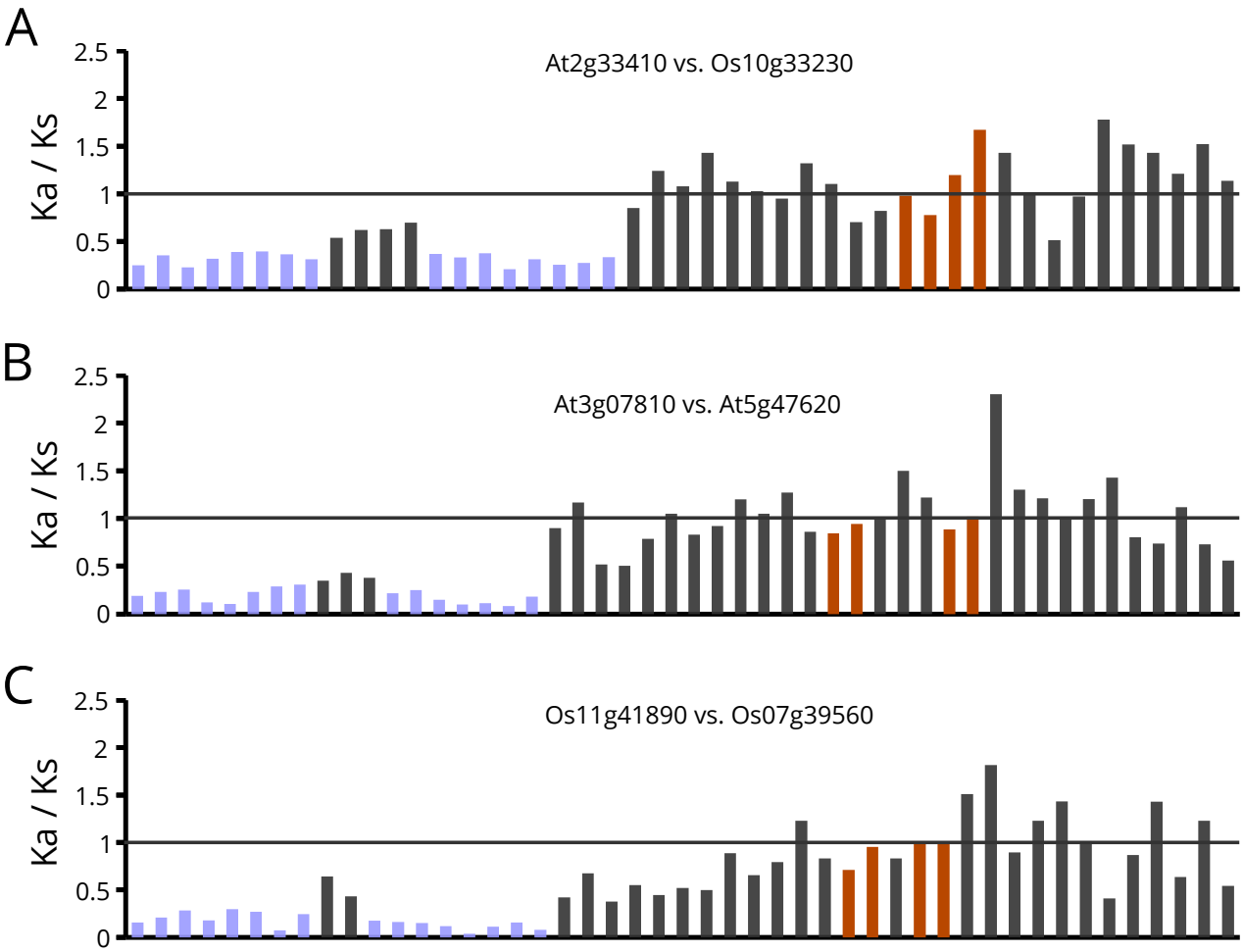
wydłużony ciąg wartości w kierunku selekcji neutralnej (*Ka/Ks* = 1). Z kolei rozkład wartości *Ka/Ks* regionów związanych z powtórzeniami WG/GW w porównaniu do domen RRM i GRP, jest znacznie bardziej wydłużony w kierunku wyższych wartości, a średni stosunek *Ka/Ks* wynosi 1,51. Świadczy to o znacznie szybciej ewolucji domen WG/GW niż w przypadku domen GRP i RRM. Mimo, że kształt histogramu wartości *Ka/Ks* domeny WG/GW jest płaski i szeroki, to najczęstsze wartości *Ka/Ks* są równomiernie rozłożone w przedziale 1-1,5.



Rys. 25. Tempo *Ka/Ks* dla domen RRM, GRP i WG/GW w białkach RRM-GRP i RRM-WG/GW *A. thalina* i *O. sativa*. A. Histogram wartości *Ka/Ks* dla domeny RRM, GRP i WG/GW. Średnia wartość *Ka/Ks* została zaznaczona strzałką. **B., C.** Wykres punktowy wartości *Ka* lub *Ks* domen GRP (zielone) i WG/GW (czerwone) zestawionych z wartościami *Ka* lub *Ks* domeny RRM w tym samym białku. Dwusieczna wskazuje relację jeden-do-jeden między wartościami *Ka* lub *Ks* obu domen. **D.** Wykres punktowy przedstawiający związek wartości *Ka/Ks* domen GRP (zielone) i GW (czerwone) z wartościami *Ks* w tych domenach. Wartości *Ka/Ks* ulegają spadkowi wraz ze zwiększającymi wartościami *Ks*, przy czym spadek ten jest szczególnie wyraźny w przypadku domen WG/GW.

Wynika z tego, że domena wiążąca AGO podlega pozytywnej selekcji (*Ka/Ks* > 1), której działanie przyspiesza utrwalanie nowych mutacji. Wysokie odchylenie standardowe *Ka/Ks* (SD = 0,83) jest widoczne na rys. 25A w postaci długiego prawego ogona histogramu. Wysoki stosunek wartość *Ka* i *Ks* w domenie WG/GW, przewyższający tempo zmian GRP, wynika ze zwiększenia tempa substytucji niesynonimicznych (rys. 25B) przy jednoczesnym zmniejszeniu liczby

substytucji synonimicznych (rys. 25C-D). Działanie pozytywnej selekcji w regionach WG/GW zostało również zaobserwowane korzystając z techniki przesuwającego się okna wzdłuż całej sekwencji u większości przypadków par genów ortologicznych i paralogicznych (rys. 26).



Rys. 26. Detekcja selekcji pozytywnej w białkach rodziny RRM-WG/GW między ortologami i paralogami *A. thaliana* i *O. sativa*. Wielkość okna i krok wynoszą odpowiednio 60 i 30. W obrębie przesuwającego się okna obliczano stosunek liczby substytucji niesynonimicznych (*Ka*) do substytucji synonimicznych (*Ks*), które liczono dla całej sekwencji. Miało to na celu zmniejszenie liczby fałszywych pozytywów, ponieważ w obrębie okna wartości Ks są niskie. Pozioma linia wskazuje na wartość *Ka/Ks* = 1. Niebieskie i czerwone słupki oznaczają sekwencje znajdujące się w obrębie, odpowiednio, domeny RRM i powtórzeń WG/GW.

Wysokie wartości *Ka/Ks* charakterystyczne dla domen WG/GW białek hnRNP sugerują, że domeny te podlegają pozytywnej selekcji, która przejawia się zwiększonym tempem akumulowania niesynonimicznych substytucji. Jednak rozrzut wartości *Ka/Ks* jest bardzo duży - waha się od 0,62 do 3,88 - co bardzo często obserwuje się w przypadku rodzin genów, które oprócz pozytywnej selekcji, podlegały działaniu mechanizmów konwersji genów i/lub rekombinacji [134–136]. Na przykład, Chen i in. (2010) odnotowali równie wysoką różnorodność wartości *Ka/Ks* w rodzinie genów odporności R u Arabidopsis, które znajdują się pod działaniem pozytywnej selekcji i ewoluują na drodze częstych zdarzeń polegających na wymianie sekwencji, obejmujących rekombinacje i konwersje genów [137].

4.3.3. Analiza konwersji genów i/lub rekombinacji

Aby dokładniej prześledzić mechanizmy molekularne warunkujące różnicowanie się domen WG/GW w rodzinie RRM-WG/GW przeprowadzona została analiza potencjalnych zdarzeń wymiany materiału genetycznego, obejmująca: rekombinację, konwersję genów i nierówny crossing-over. W tym celu wykorzystano zestaw siedmiu algorytmów identyfikacji potencjalnych miejsc rekombinacji (patrz: Metody - Rozdział 3.3).

Wśród członków rodziny genów RRM-WG/GW *A. thalina* i *O. sativa* zidentyfikowane zostały 43 prawdopodobne ($p < 0,05$) zdarzenia wymiany DNA, które nastąpiły zarówno przed, jak i po rozdzieleniu się linii ewolucyjnych obu roślin (tabela 13). Natomiast w rodzinie RRM-GRP nie stwierdzono tego typu zdarzeń. Długość przewidzianych odcinków sekwencji genomowych RRM-WG/GW, które podlegały potencjalnym rekombinacjom waha się znacznie, w przedziale od 24 pz do 725 pz. Niemal wszystkie z tych fragmentów zlokalizowane są w obrębie sekwencji kodujących. Spośród nich, około 67% sekwencji dotyczy regionów WG/GW, podczas gdy pozostałe 29% i 4% zidentyfikowane zostały w sekwencji genomowej odpowiednio w domenie RRM i regionach UTR. Dominująca część wymienianego DNA w regionach WG/GW sugeruje, że pozytywna selekcja może odgrywać znaczącą rolę w przenoszeniu i utrwalaniu rekombinowanych fragmentów pomiędzy członkami rodziny RRM-WG/GW.

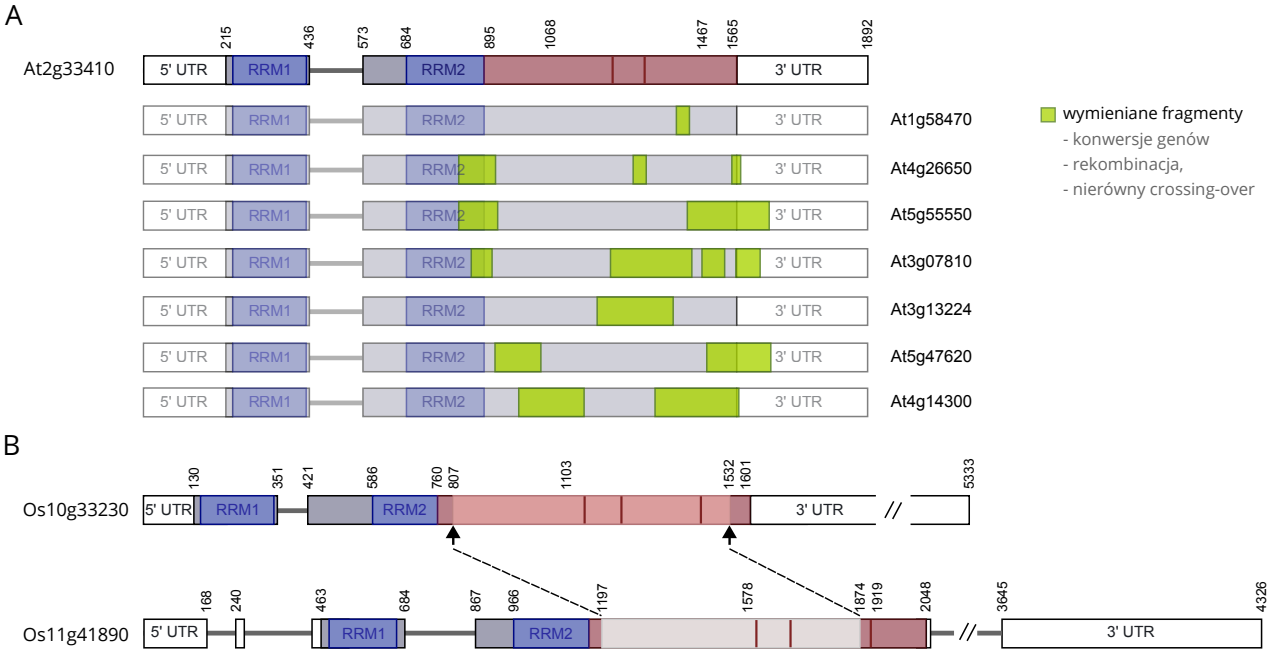
Generalnie, konwersja genów zachodząca między paralogami odpowiedzialna jest za homogenizację ich sekwencji, co w rezultacie z jednej strony może prowadzić do zwiększonego stopnia zachowania sekwencji u danego gatunku, pomimo duplikacji, a z drugiej strony zwiększa stopień dywergencji sekwencji między gatunkami [138,139]. W przypadku rodziny genów RRM-WG/GW (tabela 13), między ortologami doszło do 20 potencjalnych zdarzeń wymiany DNA, które odpowiadają zdarzeniom rekombinacji, jakie mogły zachodzić między dawnymi zduplikowanymi genami występującymi u wspólnego przodka roślin jedno- i dwuliściennych. Z kolei, w obrębie paralogów Arabidopsis i ryżu zidentyfikowano odpowiednio 19 i 4 potencjalne miejsca wymiany DNA. Wynik ten sugeruje różne tempo zachodzenia zdarzeń rekombinacji i konwersji genów RRM-WG/GW między dwoma gatunkami od czasu oddzielenia się ich linii ewolucyjnych. Ponadto długość wymienianych fragmentów DNA u Arabidopsis jest różna i mieści się w zakresie od 26 pz do 222 pz, natomiast u ryżu średnia długość rekombinowanych odcinków wynosi 300 pz. Na przykład, 15 fragmentów potencjalnie podlegających konwersji genów zostało zidentyfikowanych pomiędzy genem At2g33410 kodującym domenę WG/GW i jego paralogami (rys. 27A). Jednak w przypadku ryżu, odcinek sekwencji długości 725 pz, który odpowiada domenie WG/GW, uległ najprawdopodobniej tylko pojedynczej rekombinacji między genami Os11g41890 i Os10g33230 (rys. 27B).

Tabela 13. Potencjalne regiony sekwencji podlegające konwersji genów w rodzinie RRM-WG/GW *A. thaliana* i *O. sativa*.

seq1	seq2	seq1 start pz [aa]	seq1 endt pz [aa]	seq2 start pz [aa]	seq end pz [aa]	wartość *p**
AT2G33410	AT1G58470	1392 [348]	1430 [360]	1466 [345]	1513 [360]	7.64E-05
OS07G39560	AT3G07810	868 [133]	1026 [185]	1217 [135]	1375 [187]	1.30E-03
AT2G33410	AT3G07810	858 [177]	916 [188]	1316 [168]	1374 [186]	0.00E+00
OS11G41890	AT3G07810	658 [121]	996 [233]	1181 [123]	1519 [235]	0.00E+00
OS11G41890	AT3G07810	670 [125]	858 [187]	1193 [127]	1381 [189]	2.90E-03
OS10G33230	AT3G07810	594 [155]	740 [203]	1217 [135]	1363 [183]	1.10E-03
AT5G47620	AT3G07810	1194 [133]	1349 [184]	1217 [135]	1372 [186]	4.00E-04
AT2G33410	AT3G07810	1561 [404-]	1628 [-]	1782 [324]	1854 [346]	1.99E-03
AT2G33410	AT3G07810	1461 [371]	1522 [401]	1710 [300]	1786 [324]	1.97E-06
AT2G33410	AT3G07810	1220 [291]	1328 [326]	1952 [380]	2071 [419]	0.00E+00
AT2G33410	AT3G13224	1187 [280]	1391 [302]	1709 [298]	1862 [358]	8.02E-09
AT2G33410	AT4G14300	981 [211]	1157 [269]	1122 [197]	1340[269]	1.84E-02
AT2G33410	AT4G14300	1338 [330]	1560 [400]	1536 [335]	1781 [411]	2.24E-02
AT4G26650	OS07G39560	1819 [181]	1882 [203]	964 [165]	1027 [185]	0.00E+00
AT2G33410	AT4G26650	822 [158]	914 [188]	1786 [170]	1878 [200]	1.30E-03
OS11G41890	AT4G26650	679 [128]	858 [187]	1708 [144]	1887 [203]	0.00E+00
OS10G33230	AT4G26650	597 [156]	740 [203]	1726 [150]	1869 [197]	0.00E+00
AT4G26650	OS11G41890	1636 [120]	1887 [203]	607 [104]	858 [187]	0.00E+00
AT3G07810	AT4G26650	1220 [136]	1381 [189]	1726 [150]	1887 [203]	6.00E-04
AT2G33410	AT4G26650	1539 [397]	1565 [404-]	2584 [436]	2635 [455]	6.05E-04
AT2G33410	AT4G26650	1279 [311]	1318 [322]	2314 [347]	2361 [361]	0.00E+00
OS07G39560	AT5G47620	865 [132]	1023 [184]	1191 [132]	1349 [184]	2.50E-02
OS11G41890	AT5G47620	691 [132]	849 [184]	1191 [132]	1349 [184]	3.50E-03
OS10G33230	AT5G47620	660 [177]	734 [201]	1260 [155]	1334 [179]	4.00E-04
AT2G33410	AT5G47620	920 [191]	1046 [232]	1383 [196]	1517 [240]	5.74E-04
AT2G33410	AT5G47620	1489 [381]	1660 [404 -]	1696 [301]	1857 [353]	1.78E-07
AT2G33410	AT5G47620	1220 [290]	1439 [363]	1653 [286]	1997 [400]	8.36E-08
OS07G39560	AT5G55550	871 [134]	1026 [185]	906 [138]	1061 [189]	1.30E-03
AT2G33410	AT5G55550	822 [158]	914 [188]	966 [158]	1058 [188]	2.00E-03
OS11G41890	AT5G55550	679 [128]	852 [185]	888 [132]	1061 [189]	9.80E-03
OS10G33230	AT5G55550	597 [156]	740 [203]	906 [138]	1049 [185]	0.00E+00
OS11G41890	AT5G55550	391 [32]	555 [86]	447 [32]	752 [86]	1.10E-03
AT4G14300	OS10G33230	1036 [169]	1066 [177]	688 [187]	718 [195]	2.50E-03
OS10G33230	AT5G55550	138 [1]	176 [15]	354 [1]	398 [15]	4.19E-02
AT4G14300	AT5G55550	1005 [158]	1097 [188]	966 [158]	1058 [188]	2.02E-02
AT2G33410	AT5G55550	1420 [358]	1639 [404-]	1403 [304]	1749 [418]	1.40E-06
OS11G41890	OS07G39560	670 [125]	852 [185]	844 [125]	1026 [185]	9.00E-04
OS10G33230	OS07G39560	594 [155]	740 [203]	868 [133]	1014 [181]	0.00E+00
AT4G14300	OS07G39560	172 [4]	234 [24]	481 [4]	543 [24]	1.64E-02
OS10G33230	OS11G41890	594 [155]	740 [203]	694 [133]	840 [181]	8.00E-04
OS10G33230	OS11G41890	807 [226]	1532 [464]	1197 [301]	1578 [427]	1.33E-18
AT4G14300	OS11G41890	1040 [170]	1099 [188]	792 [166]	851 [184]	1.53E-04
AT4G14300	OS10G33230	1257 [242]	1281 [249]	1356 [409]	1380 [416]	1.00E-04

Spośród siedmiu użytych algorytmów, za potencjalne miejsca rekombinacji uznano te, które zostały zidentyfikowane przez przynajmniej trzy metody, a także miały potwierdzenie filogenetyczne. *Najniższa wartość *p* została pokazana.

Zatem po rozdzieleniu się roślin jedno- i dwuliściennych, w obu liniach zachodziły dwa odmienne mechanizmy konwersji genów mające wpływ na różnicowanie się domeny wiążącej AGO. U paralogów Arabidopsis, krótkie fragmenty średnio 32-aminokwasowe (od 7 aa do 73 aa) podlegały częstym przetasowaniom wzdłuż wysoko zmiennej sekwencji C-końca białek. W obrębie paralogów ryżu konwersja genów RRM-WG/GW zachodziła rzadziej i polegała na wymianie dłuższych sekwencji, średnio 99-aminokwasowych (od 48 aa do 238 aa).



Rys. 27. Konwersja genów w rodzinie RRM-WG/GW *A. thaliana* i *O. sativa*. A. 15 statystycznie istotnych regionów ($p < 0.05$) podlegających konwersji genów (27 - 222 pz) między genem zawierającym domenę wiążącą AGO a jego paralogami. **B.** Konwersja sekwencji długości 726 pz między genami Os11g41890 i Os10g33230 w obrębie C-końca białka, który odpowiada domenie wiążącej AGO. Strzałkami oznaczono potencjalne pozycje rekombinacji (ang. *breakpoints*).

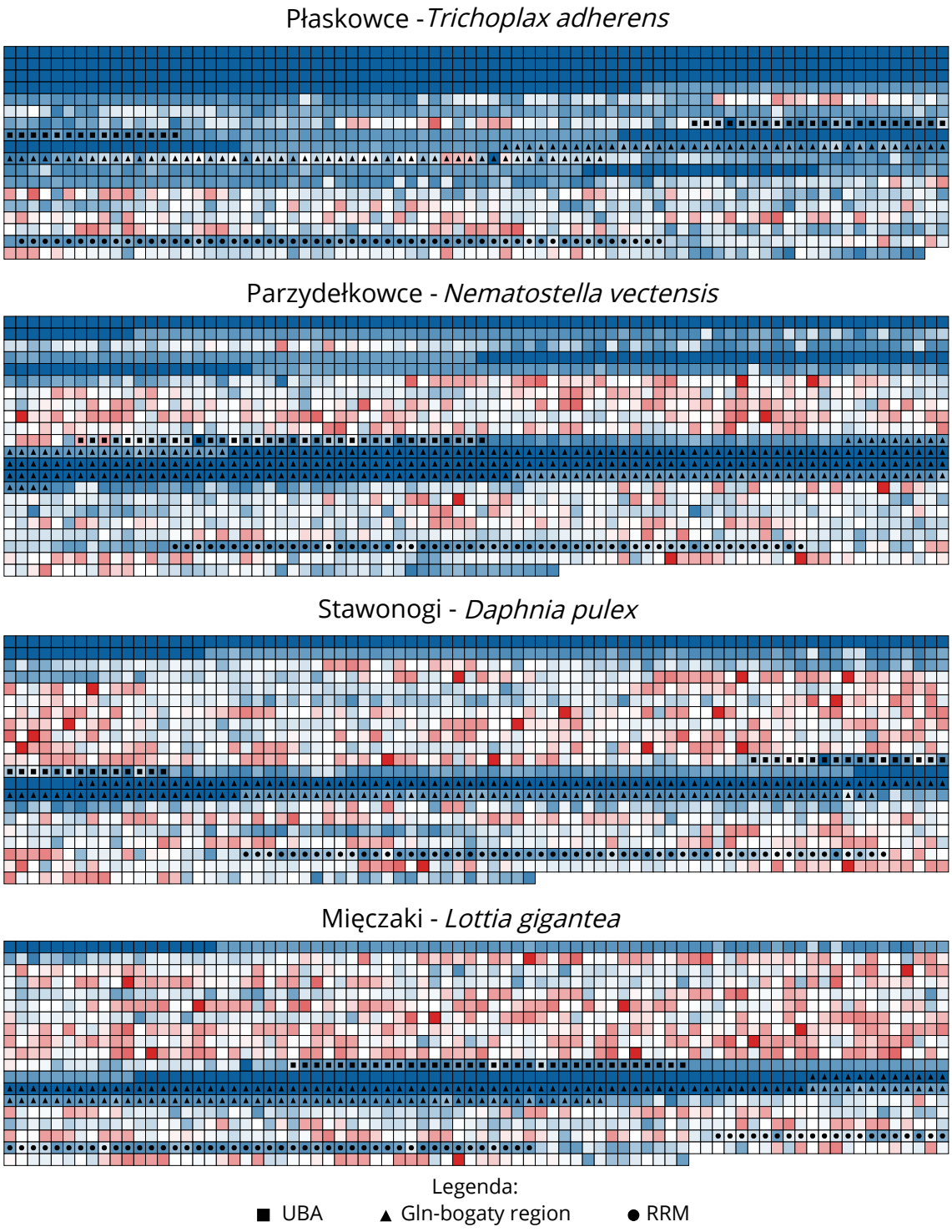
Przedstawione rezultaty wskazują, że mechanizm konwersji genów odgrywa ważną rolę w historii powstawania domen WG/GW w rodzinie RRM-WG/GW, zarówno u *A. thaliana*, jak i *O. sativa*. Współistnienie dwóch mechanizmów ewolucyjnych, homogenizującego wpływu konwersji genów i dywergencyjnego wpływu pozytywnej selekcji jest zjawiskiem coraz częściej dokumentowanym w rodzinach genów związanych z funkcjonowaniem mechanizmów odpornościowych organizmów [140–143], np. głównego układu zgodności tkankowej (MHC, ang. *Major Histocompatibility Complex*) [144], zwierzęcych immunoglobulin i genów odporności R u roślin [137,145].

4.3.4. Powstawanie *de novo* domeny WG/GW

Domena RRM, oprócz badanej w poprzednich podrozdziałach rodziny hnRNP, jest również dobrze zachowana w białkach GW182, które do tej pory zostały znalezione u kręgowców

i niektórych owadów [26]. Zastosowanie programu Wsearch podczas przeszukiwań dostępnych białek eukariotycznych (patrz: Rozdział 4.1.4 - tabela 5) umożliwiło identyfikację pojedynczych motywów WG/GW na wczesnych etapach ewolucji bezkręgowców w białkach, które wykazują wspólną architekturę domen z białkami GW182 wyższych organizmów. Pierwsze wystąpienie potencjalnego białka GW182 odnotowano u wielokomórkowca morskiego *Trichoplax adhaernes* należącego do typu płaskowców (*Placozoa*), który podobnie jak ludzkie białko GW182 posiada domenę UBA, region bogaty w glutaminę oraz motyw RRM. Mimo, że kilka krótkich W-motywów tego białka wykazuje niewielkie podobieństwo do miejsc wiążących AGO, to tylko jeden z nich może być rozpatrywany jako potencjalnie oddziałujący z białkami AGO (p = 7.78E-04). U parzydełkowców z gatunku *Nematostella vectensis* w ortologicznym białku GW182 spośród 39 reszt Trp, cztery mogą tworzyć potencjalne miejsca wiązania AGO (p < 0,001). Z kolei u rozwielitki *Daphnia pulex*, 10 spośród 49 pojedynczych W-motywów spełnia statystyczne kryteria (p < 0,001) i tworzy trzy większe domeny znajdujące się na N-końcu białka GW182 (p = 4.84E-05). Z kolei w białku GW182 mięczaka skałoczepa *Lottia gigantea*, które zawiera 57 wystąpień Trp, zidentyfikowane zostały 4 duże domeny WG/GW w obrębie N-końca. Co charakterystyczne, genomy tych czterech organizmów kodują podstawowe komponenty procesu RNAi - Argonaute, Dicer i Drosha - jednak u *T. adhaernes* nie zidentyfikowano cząsteczek miRNA i białka Pasha, które bierze udział w ich biogenezie [146].

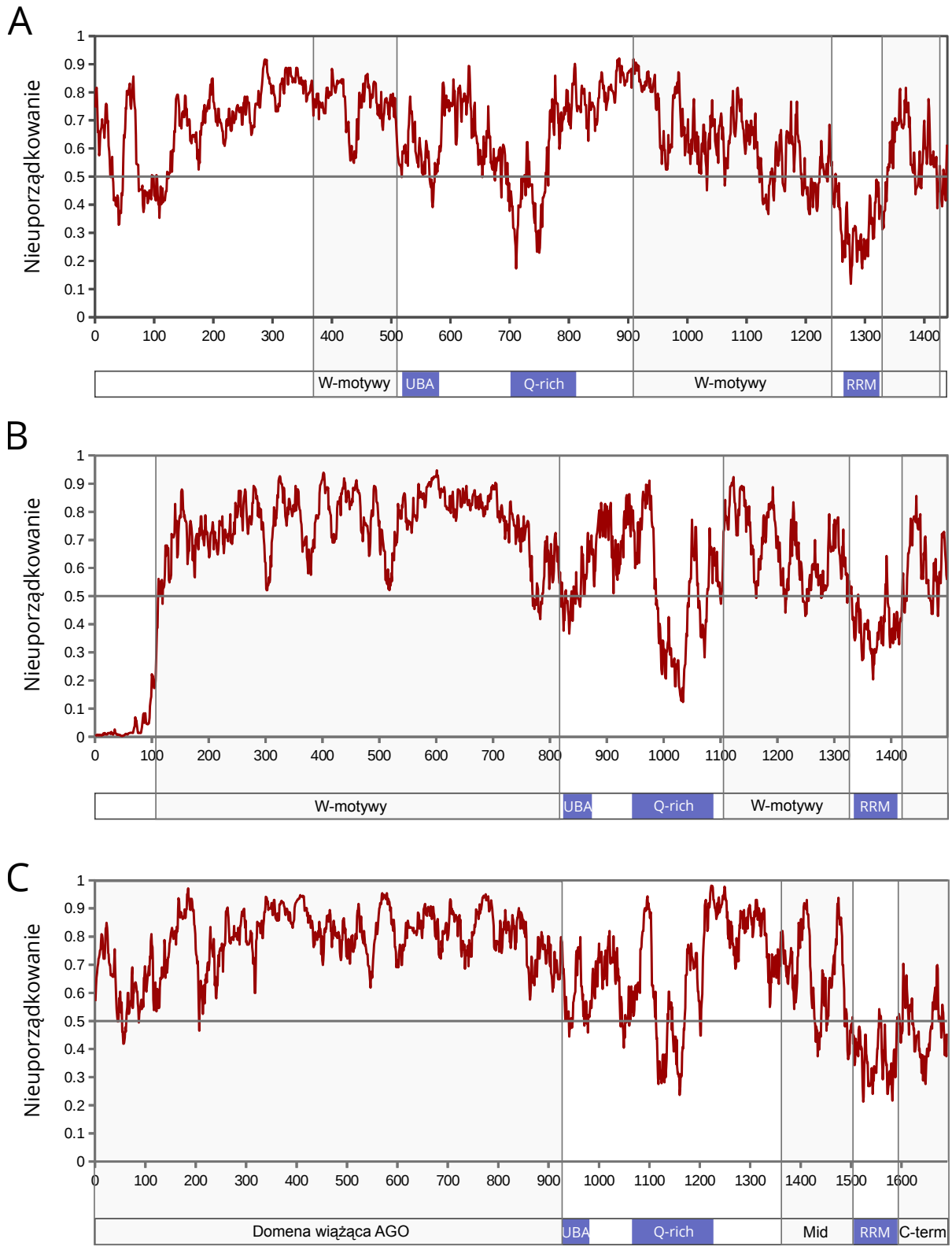
Zastosowanie narzędzia do wizualizacji W-motywów, dostępnego w portalu Whub pozwala uchwycić dynamikę różnicowania się sekwencji W-motywów w potencjalnych ortologach białka GW182 u przedstawicieli: prostych wielokomórkowców (*T. adhaernes*), parzydełkowców (*N. vectensis*), stawonogów (*D. pulex*) i mięczaków (*L. gigantea*) (rys. 28). We wszystkich białkach zachowany jest charakterystyczny dla GW182 układ domen UBA (kwadrat), Q-rich (trójkąt) i RRM (koło), które stanowią punkty odniesienia do badania domeny WG/GW. W obrębie regionów otaczających domeny UBA, Q-rich i RRM znajdują się powtórzenia Trp, gdzie sekwencja N-końca znajdująca się przed domeną UBA odpowiada położeniu domeny wiążącej AGO u kręgowców. Region ten, zawierający dwa krótkie motywy Trp u *T. adhaernes,* podlega licznym zmianom w toku ewolucji, które mają bezpośrednie przełożenie w zwiększającej się wartości *score* domeny WG/GW i prowadzą do sukcesywnego zwiększania liczby wystąpień Trp i tym samym formowania się domeny wiążącej AGO. Wraz ze wzrostem poziomu złożoności organizmów zwiększa się również długość regionu N-końca bogatego w Trp, przy jednoczesnym zmniejszeniu długości sekwencji C-końca flankujących domenę RRM, które u kręgowców odpowiadają miejscom wiązania kompleksu CCR4-NOT. Proporcje długości sekwencji bogatych w powtórzenia Trp w białku GW182 *L. gigantea* są zbliżone do odpowiadających im sekwencji w białku TNRC6C człowieka i innych ssaków (rys. 28 i 29).



Rys. 28. Zastosowanie wizualizacji aplikacji internetowej Wsearch do przedstawienia sekwencji potencjalnych ortologów białek GW182 u przedstawicieli bezkręgowców: płaskowców, parzydełkowców, stawonogów i mięczaków. Aminokwasy wchodzące w skład domen UBA, Q-rich i RRM zostały odpowiednio oznaczone symbolem: kwadratu, trójkąta i koła.

Zgodnie z wynikami analizy przy pomocy programu IUPred, sekwencje, w których znajdują się reszty Trp charakteryzują się wysokim stopniem nieuporządkowania w postaci długich regionów pozbawionych struktury trzeciorzędowej (rys. 29). Zorganizowane w ten sposób elastyczne i niesfałdowane fragmenty odseparowują niezależnie sfałdowane domeny w obrębie

cząsteczki białkowej, stanowiąc miejsca formowania się domen złożonych z licznych powtórzeń Trp wyspecjalizowanych w oddziaływaniu z róznymi kompleksami wielopodjenostkowymi (np. RISC, CCR4-NOT).



Rys. 29. Regiony nieuporządkowane (wg programu IUPred [59]) w białkach GW182. A. *T. adherens*. **B.** *L. gigantea* i **C.** *H. sapiens* (TNRC6C). Stopień nieuporządkowania powyżej 0,5 wskazuje na potencjalne regiony białka pozbawione struktury trzeciorzędowej.

ROZDZIAŁ 5

Dyskusja

W ciągu ostatnich pięciu lat wiedza o domenach WG/GW ewoluowała od obrazu wielkich platform molekularnych składających się z kilkudziesięciu motywów WG/GW, które tworzą sztywne rusztowania dla oddziaływań z białkami AGO. Okazało się jednak, że krótkie motywy zawierające Trp, w zależności od charakteru otaczających je reszt aminokwasowych, pełnią wyspecjalizowane funkcje w wiązaniu różnych kompleksów efektorowych RNAi, o zmiennym składzie u różnych organizmów. W rezultacie pojawiania się wciąż nowych informacji, w trakcie realizacji niniejszej pracy doktorskiej opracowano i zaimplementowano trzy różne metody identyfikacji *de novo* domen wiążących AGO - od poszukiwań długich, bogatych w pary WG/GW, sekwencji o specyficznej kompozycji aminokwasowej (program Agos), po identyfikację krótkich motywów zawierających Trp w oparciu o pozycyjnie-specyficzną analizę reszt występujących w jego otoczeniu (programy Wsearch i i-Wsearch). Wynikiem zastosowania stworzonych programów są listy nowych białek potencjalnie oddziałujących z AGO (tabela 4, 7 i 5), z których część została już potwierdzona eksperymentalnie (WGRP1, SDE3, RRM).

Powszechne występowanie domeny WG/GW w białkach wiążących RNA

Funkcjonalne powtórzenia WG/GW występują niemal u wszystkich organizmów eukariotycznych - od jednokomórkowych orzęsków do człowieka (tabela 5). Ponadto potencjalne domeny WG/GW zidentyfikowane zostały w niniejszej pracy wśród organizmów prokariotycznych posiadających białka Argonaute (tabela 6). Wyniki te sugerują wspólne pochodzenie domen WG/GW u wspólnego przodka Eukariota i Prokariota, którego genom najprawdopodobniej kodował białko Argonaute.

Domeny WG/GW zostały znalezione w wielu niespokrewnionych rodzinach białkowych,

które w większości są charakterystyczne jedynie dla pewnych linii ewolucyjnych np. ssaków (PrP), roślin wyższych (NRPE1), orzęsków (CnjB), drożdży (Tas3) (tabela 5). Jednak wspólną cechą białek wiążących AGO wśród różnych grup systematycznych jest ich współwystępowanie z motywami lub domenami wiążącymi RNA. Na przykład, u pantofelka białka Nowa1p i Nowa2p, są odpowiedzialne za wiązanie i transport scnRNA w procesie rearanżacji materiału genetycznego podczas koniugacji [100]. Z kolei u innego orzęska *T. termophila*, białko wiążące AGO, CnjB zawiera siedem tandemowo ułożonych motywów palca cynkowego typu CCHC [10]. U Arabidopsis czynnik transkrypcyjny SPT5/KTF1 posiada domenę wiążącą RNA, która jest zdolna do wiązania transkryptów syntetyzowanych przez polV zarówno w warunkach *in vitro* jak i *in vivo*. [34]. Również polimeraza V wykazuje zdolność wiązania transkryptów RNA [34]. Ostatnio wykazano, że białko wiążące AGO, SDE3 Arabidopsis posiada aktywność helikazy RNA, która jest niezbędna podczas TGS i PTGS [86]. Wprawdzie Tas3 drożdży nie posiada zdefiniowanej domeny wiążącej RNA, jednak sztuczne sprzężenie tego białka z powstającymi transkryptami wywołuje powstawanie heterochromatyny w obrębie *locus* tego transkryptu [147]. U ssaków, białka prionowe posiadające domenę WG/GW oddziałują z RNA [148], a mechanizm konwersji prionów stymulowany jest przez RNA [149]. Chociaż rola zwierzęcych białek GW182 w oddziaływaniu z RNA jak dotąd pozostaje niewyjaśniona [26], to białka te posiadają domenę RRM dobrze zachowaną u wszystkich członków rodziny (rys. 28). Również powszechnie występujące helikazy RNA typu DEAD-box oraz białka hnRNP, zaproponowane w niniejszej analizie jako najlepsze białka kandydujące do wiązania AGO, realizują swoje funkcje na drodze oddziaływań z RNA.

Interpretacja tych obserwacji według kryterium parsymonii sugeruje, że gen wspólnego przodka mógł kodować białko posiadające zarówno domenę wiążącą AGO, jak i motyw wiążący RNA, a w specyficznych liniach ewolucyjnych doszło do niezależnej utraty domeny WG/GW zgodnie z zasadą "wykorzystaj albo porzuć" (ang. *use it or lose it*). Z drugiej strony, ze względu na wysoki stopień dywergencji zachodzącej w toku ewolucji domen WG/GW, trudno w pełni rozstrzygnąć czy architektura białek wiążących AGO została zapożyczona w różnych grupach taksonomicznych od genu wspólnego przodka, czy też powstała niezależnie więcej niż jeden raz, aby umożliwić realizację funkcji w tworzeniu efektorowych kompleksów wyciszających w komórkach różnych organizmów. Brak statystycznie istotnego podobieństwa między sekwencjami domen WG/GW w tych białkach, różna długość ich sekwencji (od 22 w białku Tas3 do ponad 700 reszt aminokwasowych w SPT5/KTF1) oraz zmienna liczba powtórzeń motywu WG/GW (od 1 do ponad 45) nie pozwalają ani potwierdzić homologii, ani wykluczyć analogii między tymi białkami. Warto jednak zwrócić uwagę, że tak wyjątkowo zmienny charakter domen wiążących AGO stanowi pewnego rodzaju problem samej tożsamości domeny.

Wspólny wzór ekspansji rodziny RRM-WG/GW u A. thaliana i O. sativa

Spośród roślinnych genomów, białka RRM-WG/GW zostały zidentyfikowane w niniejszej pracy u mchów i roślin naczyniowych. U roślin niższych nie znaleziono domeny WG/GW w obrębie genów kodujących motyw RRM. Sugeruje to, że powstanie domeny wiążącej AGO w roślinnych białkach wiążących RNA nastąpiło po rozdzieleniu się linii rodowych glonów i roślin lądowych. Jednocześnie, różnicowaniu się domeny WG/GW w roślinach wyższych towarzyszył wzrost liczebności i złożoności rodziny RRM - na przykład genom mchu *P. patens*, uważanego za jedną z pierwszych roślin lądowych, koduje dwa razy więcej białek zawierających domenę RRM w porównaniu do zielenic (tabela 10).

U roślin jedną z najbardziej licznych regulatorowych cząsteczek RNA stanowią siRNA odpowiadające transpozonom, retrotranspozonom i retrowirusom [150], które powodują epigenetyczne wyciszanie sekwencji poprzez metylację *de novo* DNA w procesie RdDM. W tym procesie niezbędny jest udział białek zawierających domenę WG/GW. W przypadku glonu *C. reinhardtii* około 6,4% genomu stanowią elementy powtarzalne [151]. Ponadto do tej pory znaleziono tylko 26 cząsteczek siRNA [152]. Metylacja w genomowym DNA *Chlamydomonas* zachodzi na dużo niższym poziomie niż u roślin lądowych i częściej dotyczy egzonów niż elementów powtarzalnych i transpozonów [153]. Może to sugerować, że RdDM nie jest dominującym mechanizmem wyciszania elementów powtarzalnych u *Chlamydomonas*, co mogłoby również przekładać się na mniejszy udział białek WG/GW w szlakach RNAi. Nieobecność domen WG/GW odnotowano nie tylko w rodzinie RRM, ale także w przypadku innych dobrze scharakteryzowanych białek wiążących AGO u roślin - na przykład białko SPT6/GTB1 *C. reinhardtii* pozbawione jest powtórzeń WG/GW. Nie zidentyfikowano również ortologów białka SPT5/KTF1 i dużej podjednostki pol V, NRPE1. Z kolei genom *P. patens* koduje białko NRPE1, które posiada potencjalną domenę WG/GW oraz jedno białko RRM-WG/GW. Obecność białek WG/GW może być związana z bardziej złożoną strukturą genomów mszaków, w których niemal 19% jest wynikiem wielokrotnych serii powielania i integracji LTR-retrotranspozonów do genomu [154]. Ponadto w odróżnieniu od glonów, u mszaków siRNA odgrywają ważną rolę w odpowiedzi na transpozycję bądź niepożądaną rearanżację DNA [155]. Biorąc pod uwagę fakt, że początkowa maszyneria białkowa RNAi u roślin lądowych znacznie powiększyła się i zróżnicowała [156,157], ekspansja białek WG/GW przy jednoczesnej duplikacji komponentów RNAi, mogła umożliwić powstanie bardziej efektywnych strategii przeciwdziałania transpozycji i infekcji wirusowych.

Analiza białek RRM-WG/GW u roślin wyższych, *O. sativa* i *A. thaliana* , wykazała, że kodowane są one przez wielogenową rodzinę hnRNP (rys. 22, 23). Mimo, odmiennej ewolucyjnej historii, jaka dzieli te dwa gatunki oraz różnic w składzie genomu, oba organizmy wykazują

znaczne podobieństwo, zarówno w liczbie genów RRM-WG/GW, jak i układzie rozprzestrzeniania się członków tej rodziny (rys. 23C-D). Wydaje się więc prawdopodobne, że zduplikowane geny paralogiczne, kodujące niemal identyczne domeny RRM, lecz różniące się istotne pod względem sekwencji WG/GW, różnicowały się w celu dalszego udoskonalania funkcji lub subfunkcjonalizacji genu wspólnego przodka, który kodował białko wiążące AGO.

Można zatem skonstruować hipotezę, że odpowiadające sobie sekwencje tworzące tego typu układy, ściśle ze sobą współpracują w komórce podczas wiązania z białkami z rodziny AGO. Być może dzieje się tak poprzez regulację intensywności tego procesu lub zwiększenie specyficzności interferencji RNA do niektórych *loci* DNA lub transkryptów RNA. Podobne układy białek paralogicznych, które istotnie różniły się pod względem sekwencji domeny WG/GW, odnotowano w przypadku innych, znanych już białek niezbędnych do wiązania AGO. Na przykład, dwa białka, u orzęska *T. termophila* - Wag1p i CnjBp - zawierające różne warianty domeny WG/GW pełnią zbliżone funkcje podczas rearanżacji materiału genetycznego na drodze RNAi. Jedynie równoczesne mutacje w obrębie obu tych genów powodowały zaburzenia w usuwaniu elementów występujących w wielu kopiach, m.in. transpozonów i minisatelitów oraz precyzyjnym wycięciu pojedynczych, niekodujących sekwencji IES [10]. Również u *C. elegans*, podwójne mutanty dwóch białek wiążących AGO - AIN-1 i AIN-2, zaburzały proces RNAi bardziej skutecznie, niż wyłączanie jednego z genów. Świadczy to o pełnieniu tożsamych funkcji przez te białka [39,40]. Podobieństwo funkcjonalne wykazują również paralogi rodziny GW182, które oddziałują u człowieka ze wszystkimi czterema białkami AGO oraz uczestniczą w wyciszaniu podobnego zestawu docelowych genów [22,23,30,31,92,93,158,159]. Ponadto mutacje w obrębie któregokolwiek z paralogów TNRC6A, TNRC6B lub TRNC6C, częściowo zaburzają wyciszanie docelowych transkryptów mRNA w komórkach człowieka [22,23,93,97,158,159]. Z kolei u Arabidopsis podjednostka polimerazy V NRPE1, w odróżnieniu od swojej sekwencji paralogicznej NRPD1 wchodzącej w skład polimerazy IV, posiada domenę C-terminalną (CTD) bogatą w powtórzenia reszt WG/GW, które są niezbędne do wiązania białek AGO4. Również dwa czynniki transkrypcyjne - SPT6/GTB i SPT5/KTF1, zaangażowane w wiązanie białek AGO, różnią się istotnie w sekwencji domen WG/GW.

Zatem białka RRM-WG/GW, które uzyskały wysoką wartość oceny domeny WG/GW mogą być białkami niezbędnymi do asocjacji AGO, natomiast ich paralogi zawierające słabszą ocenę domeny mogą stanowić dodatkowe moduły regulujące formowanie efektorowych kompleksów różnych ścieżek RNAi. Takie rozwiązanie niesie kilka istotnych następstw: po pierwsze: umożliwia formowanie podstawowych funkcjonalnych kompleksów wyciszających za pomocą niewielkiej liczby wysokopunktowanych białek rdzenia, co zapewnia komórce dużą elastyczność. Po drugie, podczas stresu komórkowego np. wywołanego wniknięciem wirusa do komórki, nie

ma potrzeby produkowania skomplikowanych kompleksów RISC *de novo*, a konieczne jest jedynie dołączenie kilku niezbędnych białek posiadających słabszą domeną WG/GW, co umożliwia, w relatywnie prosty sposób, kontrolowanie dość złożonych maszynerii białkowych jedynie przez dostarczanie lub blokowanie istotnych elementów. Można zatem przyjąć, że bliskie występowanie takich układów paralogicznych białek, w niektórych sytuacjach może być konieczne do niezależnego różnicowania się domen wiążących AGO.

Mechanizmy molekularne różnicowania się domeny WG/GW

Jedną z charakterystycznych cech domen WG/GW jest wyjątkowa zmienność ich sekwencji nawet w obrębie blisko spokrewnionych genów (rys. 21). W niniejszej pracy wykazano wpływ różnych mechanizmów wewnątrz- i międzygenowych, które warunkują zróżnicowanie w obrębie domeny WG/GW w białkach RRM. W rodzinie RRM-WG/GW, podobnie jak w przypadku wielu innych wielogenowych rodzin, segmentowe i tandemowe duplikacje oraz alternatywny splicing są powszechnie występującymi mechanizmami odpowiedzialnymi za zróżnicowanie sekwencji domeny wiążącej AGO. Ponadto ewolucja domeny WG/GW następuje poprzez pozytywną selekcję (rys. 25 i 26) podlegając przy tym częstym zdarzeniom wymiany sekwencji (rys. 27). Tak szybkie tempo zmienności w połączeniu z częstymi rekombinacjami, konwersjami genów i nierównym crossing-over, składają się na główne mechanizmy zapewniające wysokie zróżnicowanie genetyczne domen WG/GW.

Współwystępowanie pozytywnej selekcji i częstych konwersji genów obserwuje się powszechnie wśród roślinnych i zwierzęcych rodzin genów odpornościowych, które kodują białka biorące udział w rozpoznaniu patogenów [140–143]. Na przykład konwersja genów generuje zróżnicowanie w należących do układu immunologicznego białkach głównego układu zgodności tkankowej [144,160], a także w immunoglobulinach [154], poprzez wprowadzanie i sortowanie punktowych mutacji wśród sekwencji paralogów [162,163]. Również związane z odpornością na patogeny, geny Cf pomidora [164] oraz trzy rodziny genów Arabidopsis (NBS-LRR, RLK, RLP) podlegają pozytywnej selekcji oraz konwersji genów [137,145,165,166].

Obecność domen WG/GW jest niezbędna w tworzeniu efektorowych kompleksów w procesie RNAi, który u roślin stanowi naturalny system obrony przeciwko wirusom [167]. W toku ewolucji wiele wirusów infekujących rośliny wykształciło mechanizmy chroniące je przed działaniem RNAi gospodarza np. przez wytwarzanie białek supresorowych bezpośrednio oddziałujących z komponentami kompleksu RISC gospodarza i zaburzających jego funkcję, wymykając się w ten sposób spod RNAi-zależnego systemu obronnego [168]. Na przykład białko 2b Trp kodowane przez wirusy z rodzaju *Cucumovirus* oddziałuje bezpośrednio z białkami AGO1 uniemożliwiając kompleksowi RISC przeprowadzenie degradacji docelowego mRNA [169].

Podobnie białko P0 wirusów z rodzajów *Polerovirus* asocjuje z białkami AGO i doprowadza do ich degradacji [170,171]. Ostatnie doniesienia sugerują, że niektóre wirusy mogą kodować własne białka, które funkcjonalnie upodobniają się do domen WG/GW gospodarza. Przez oddziaływanie tych białek z białkami AGO gospodarza zaburzone zostaje prawidłowe składanie kompleksu RISC. Na przykład białko P38 kapsydu wirusa TCV (ang. *Turnip crinkle virus*) używa powtórzeń WG/GW, jako przynęty wiążącej białka AGO, przełamując w ten sposób system obronny RdDM Arabidopsis [36]. Również w białku proteazy serynowej P1 wirusa SPMMV (ang. *Sweet potato mild mottle virus*) powtórzenia WG/GW są niezbędne podczas wiązania i supresji białek AGO1 [37]. Niedawno wykazano, że mutacja w obrębie pojedynczego motywu WG/GW białka supresorowego NSs wirusa TSWV (ang. *Tomato spotted wilt virus*) całkowicie pozbawia funkcji supresorowych to białko, co sugeruje potencjalną interakcję między sekwencjami WG/GW a AGO1 [38].

Uzasadniony zatem wydaje się wniosek, że szybka ewolucja domeny WG/GW w białkach RRM, w której różnicowanie zaangażowany jest złożony zestaw mechanizmów selektywnie akumulujących niesynonimmiczne substytucje, odzwierciedla w dużym stopniu dynamiczną koewolucję w systemie pasożyt-gospodarz. Domeny WG/GW gospodarza rekrutują białka AGO w celu degradacji wirusowych transkryptów RNA. Z kolei dynamicznie zmieniający się genom wirusowy może kodować białka, które w oparciu o mimikrę domen WG/GW wymykają się spod obronnej maszynerii RNAi gospodarza w celu zainicjowania infekcji. Pogląd ten potwierdzają ostatnie badania, które wskazują, że komponenty biorące udział w RNAi są kodowane przez jedne z najszybciej ewoluujących genów związanych z odpornością [172]. U podstaw tych szybkich zmian leży koncepcja Czerwonej Królowej (ang. *The Red Quenn Hipothesis*), zgodnie z którą między gospodarzem a patogenem dochodzi do ciągłego wyścigu molekularnego, który mógł bezpośrednio przyczynić się wykształcenia RNAi, jako mechanizmu obronnego komórki przed wirusami o genomie w postaci dwuniciowego RNA [167,173–175].

Przedstawiona w niniejszej pracy równoległa ekspansja rodziny RRM-WG/GW przebiegająca niezależnie u *O. sativa* i *A. thaliana*, lecz nie u glonów i mchów, może świadczyć o wspólnej konieczności adaptacyjnej procesu różnicowania się sekwencji. Czynnikiem selekcyjnym może tu być wspólna grupa wirusów infekująca obie rośliny. Większość poznanych wirusów roślinnych (75%) stanowią wirusy zawierające jednoniciowy RNA (ssRNA) oraz wirusy infekujące rośliny okrytonasienne, zwłaszcza jedno- i dwuliścienne [176]. Wspólny zbiór potencjalnych patogenów może być zatem siłą napędową powodującą konwergentny układ ekspansji białek RRM-WG/GW przy jednoczesnym zaangażowaniu mechanizmów molekularnych zapewniających wysoką dywergencję ich sekwencji.

Substytucja Tyr > Trp w białkach GRP uruchamiająca szybką ewolucję domen WG/GW

Domeny WG/GW posiadają wiele wspólnych cech z białkami glicynobogatymi. Po pierwsze, na drzewie filogenetycznym białek zawierających domenę RRM *O. sativa* i *A. thaliana* (rys. 22), białka GRP tworzą siostrzaną grupę do białek WG/GW, co sugeruje ich wspólne pochodzenie. Po drugie, obie rodziny białkowe wykazują podobną kompozycję aminokwasów, która znacznie odbiega od składu aminokwasowego innych niespokrewnionych białek (tabela 11). Po trzecie, domeny WG/GW i GRP składają się tandemowych powtórzeń krótkich sekwencji tworzących wydłużone modularne domeny (rys. 24,16). Po czwarte, obie domeny najprawdopodobniej pozbawione są stabilnej struktury trzeciorzędowej, co pozwala przyjmować im elastyczne konformacje i wykazywać uniwersalną zdolność wiązania.

Jednak oprócz tych podobieństw, domeny GRP i WG/GW wykazują istotne różnice. Jedną z nich jest odwrotna zawartość tryptofanu i tyrozyny w obu domenach przy jednoczesnym zachowaniu ogólnej kompozycji reszt hydrofobowych (tabela 11). W domenach WG/GW, wzrost zawartości Trp jest wprost proporcjonalny do spadku ilości Tyr, i odwrotnie w domenach GRP, wzrostowi ilości Tyr towarzyszy odpowiedni spadek zawartości Trp. Drugą właściwością różniącą domeny WG/GW i GRP jest udział odmiennych mechanizmów różnicowania oraz selekcji. Geny GRP wykazują bowiem niższe tempo ewolucji, niż geny kodujące domenę WG/GW, a stosunek częstości zachodzenia zmian niesynonimicznych i synonimicznych sugeruje, że niektóre regiony sekwencji GRP podlegają mechanizmom ewolucji neutralnej (rys. 25). Ponadto, w odróżnieniu od genów RRM-WG/GW, w obrębie rodziny RRM-GRP nie zaobserwowano potencjalnych zdarzeń rekombinacji i/lub konwersji genów. Na podstawie tych obserwacji można sugerować, że substytucja reszty tyrozyny na tryptofan w obrębie regionów glicynobogatych, może być kluczowym czynnikiem kierującym te geny na nową dynamiczną ścieżkę ewolucyjną, w której zarówno powstające mutacje, jak i motywy WG/GW ulegają częstym przetasowaniom w obrębie zduplikowanych genów. Wysoka efektywność takiej selekcji kierunkowej, została zarejestrowana w przypadku wielu enzymów, gdzie wystąpienie korzystnej mutacji uruchamia mechanizmy polegające na zwiększeniu tempa kolejnych substytucji przy jednoczesnej rekombinacji i wzroście presji selekcyjnej [177].

Ewolucyjny przełącznik polegający na substytucji tyrozyny i tryptofanu może reprezentować mutację polegającą na nabyciu przez białko nowej funkcji (ang. *gain-of-function*). Obie domeny, WG/GW i GRP, podobnie jak inne typy domen biorące udział w interakcjach białko-białko, mogą stanowić miejsca mutacji, w których podstawienia niesynonimiczne umożliwiają białkom oddziaływanie z nowym zestawem partnerów, potencjalnie zmieniając pełnione przez nie funkcje. Ponieważ mechanizm ten obejmuje możliwość utraty, jak i nabycia funkcji, z dużym prawdopodobieństwem będzie zachodził w kontekście duplikacji genów, tak jak w przypadku

białek GRP i WG/GW (rys. 22). Chociaż kombinacja reszt Trp i Gly w większości przypadków jest niezbędna podczas interakcji z białkami AGO, to nie wszystkie powtórzenia tego motywu mają jednakowy wpływ na wiązanie. Jednak pojedyncza substytucja polegająca na wprowadzeniu tryptofanu może zwiększyć specyficzność wiązania białek AGO [8,9,27]. Taka mutacja polegająca na powstaniu powtórzenia WG/GW może więc być uznawana za inicjujący moment powstawania nowego komponentu procesów RNAi. Taka zmiana funkcji może wówczas uruchomić działanie dodatkowych mechanizmów molekularnych, angażujących pozytywną selekcję i konwersję genów, które powodowałyby dalszą optymalizację aktywności wiązania białek AGO w celu przeciwdziałania wirusom i elementom mobilnym. Warto zwrócić uwagę, że gdy dochodzi do punktowej mutacji w białkach glicynobogatych, polegającej na substytucji dowolnego aminokwasu na tryptofan, można oczekiwać, że doprowadzi ona do pojawienia się właśnie jednego z tych powtórzeń – WG/GW/GWG. Biorąc pod uwagę fakt, że w kilku doświadczalnie potwierdzonych białkach wiążących AGO, tylko jedno powtórzenie WG lub GW jest wystarczające do bezpośredniej interakcji z białkiem AGO [7,29,38], glicynobogate obszary są zatem szczególnie predestynowane do formowania potencjalnych domen wiążących AGO we względnie niewielkiej liczbie substytucji. Istnieje zatem większa szansa, że w toku ewolucji domena WG/GW może powstawać wielokrotnie i niezależnie w różnych białkach glicynobogatych.

Domeny WG/GW tworzące klasę domen wewnętrznie nieuporządkowanych

Prezentowana w niniejszej pracy analiza pojedynczych motywów zawierających Trp pochodzących z eukariotycznych białek wiążących AGO, ujawnia kilka nowych uniwersalnych właściwości tych domen, które sugerują, że domeny WG/GW należą do klasy białek inherentnie nieuporządkowanych (IDP, ang. *Intrinsically Disordered Proteins*). W przeciwieństwie do białek globularnych, w warunkach uznanych za natywne, funkcjonalne IDP pozbawione są stabilnej struktury trzeciorzędowej. Charakteryzuje je natomiast wyjątkowa plastyczność i dynamika konformacji w reakcji na zmianę warunków środowiska lub na skutek oddziaływania z odmiennymi partnerami [178]. Z tego powodu IDP zaangażowane są w różnego typu szlaki regulacyjne i procesy, podczas których dochodzi do składania supramolekularnych kompleksów.

Domeny WG/GW, podobnie jak regiony nieuporządkowane, ewoluują znacznie szybciej niż regiony globularne białka (rys. 3,10,21,26,25). Dlatego w większości przypadków, stopień zróżnicowania sekwencji WG/GW jest zbyt wysoki, aby wiarygodnie określić ich relacje homologiczne (rys. 21). Domeny IDP wykazują zwiększone tempo zmian ewolucyjnych, ponieważ stan nieuporządkowania pozwala obejść restrykcje związane ze sztywną strukturą trzeciorzędową. Jednak badania powierzchni oddziaływań białko-białko, polegające na analizie

sekwencji nieuporządkowanych, za pośrednictwem których odbywa się molekularne rozpoznanie i oddziaływanie z partnerem, wykazały istnienie tzw. krótkich liniowych motywów zaangażowanych w inicjowanie powierzchni kontaktu białko-białko. Motywy te wykazują wysokie zakonserwowanie sekwencji [179,180]. W podobny sposób, słabo zachowana sekwencja domen WG/GW zbudowana jest z krótkich motywów, w obrębie których aminokwasy znajdujące się w najbliższym otoczeniu Trp zaangażowane są w wiązanie białek AGO (rys. 16).

Drugą właściwością sekwencji domen WG/GW, która upodobnia je do IDP, jest wysoki stosunek wypadkowego ładunku białka do hydropatii. W sekwencjach domen wiążących AGO obserwuje się wyraźny brak tolerancji obecności pewnych reszt aminokwasowych, przy równoczesnym pewnym faworyzowaniu innych aminokwasów (tabela 2, rys. 15). Sekwencje otaczające Trp cechuje znaczący niedomiar reszt promujących strukturę uporządkowaną (ang. *order-promoting residues*): Tyr, Phe, Ile, Leu, które tworzą hydrofobowy rdzeń białek globularnych. Motywy zawierające Trp wykazują również bardzo niską zawartość reszt Cys, która odgrywa istotną rolę w utrzymywaniu stabilnej struktury trzeciorzędowej poprzez tworzenie mostków disiarczkowych oraz grup prostetycznych. Chociaż domeny wiążące AGO są wyjątkowo bogate w reszty Trp, który jest hydrofobowym aminokwasem, to kontekst flankujących reszt wykazuje niską złożoność składu aminokwasowego i wysoki poziom polarnych i naładowanych aminokwasów (rys. 15C). Podobną zależność odnotowano w wielu domenach inherentnie nieuporządkowanych (IDD, ang. *intrinsically disordered domains*) [181–183], w których zwiększona hydrofobowość sekwencji wywołana obecnością reszt aminokwasowych, kluczowych dla interakcji, jest kompensowana przez zwiększony wypadkowy ładunek pozostałych regionów białka. W ten sposób utrzymana zostaje tendencja do powstawania struktury nieuporządkowanej, mimo obecności w jej sekwencji reszt promujących strukturę uporządkowaną.

Po trzecie, domeny wiążące AGO, podobnie jak IDD wykazują wysokie zrónicowanie długości sekwencji wynikające zarówno z różnej liczby, jak i długości W-motywów. Z ostatnich doniesień wynika, że IDD, które w większości podlegają licznym substytucjom i częstym zdarzeniom nierównego crossing-over, ewoluują poprzez rozprzestrzenianie się określonego, dla danego białka, motywu [184]. Podobnie domeny wiążące AGO cechuje zmienna liczba kopii W-motywów, zarówno między blisko spokrewnionymi ortologami NRPE1, jak i między paralogami GW182. W ekstremalnych przypadkach, pojedynczy motyw w białku TAS3 jest wystarczający do tworzenia interakcji z białkiem AGO u *S. pombe* [7,8], z kolei w białku SPT5/KTF1 *Manihot esculenta* motywy te są powtórzone nawet do 75 razy i rozciągają się w sekwencji na długości 1000 reszt aminokwasowych. Mimo wyjątkowo zmiennej długości sekwencji, domeny wiążące AGO zbudowane są ze zwielokrotnionych pojedynczych, zawierających tryptofan motywów długości od 10 do 20 aminokwasów. Zgodnie z matrycą PSSM

zbudowaną w oparciu o eukariotyczne domeny wiążące AGO, najbardziej preferowanym W-motywem jest 20-aminokwasowa sekwencja NGNNNNSNSGWGEPPNQNSS. Jednak motyw ten do tej pory nie został znaleziony w znanych sekwencjach białek.

Po czwarte, w eksperymentalnie rozwiązanej strukturze ludzkiego białka prionowego, powtórzenia WG/GW występują w regionie niestrukturyzowanym [185,186]. Również w białkach GW182, przewidywania struktur drugorzędowych sugerują, że domena wiążąca AGO nie posiada stabilnej struktury przestrzennej (rys. 29). Plastyczność sekwencji domeny wiążącej AGO najprawdopodobniej pozwala odsłaniać reszty Trp, będące miejscem bezpośredniego oddziaływania. Zatem brak strukturalnego uporządkowania może zapewniać uniwersalną zdolność domeny do wiązania białek AGO. Na przykład w doświadczeniu przeprowadzonym przez El-Shamiego i in. (2007), polegającym na wymianie domen WG/GW między dwoma niespokrewnionymi białkami, oba białka chimeryczne miały zdolność oddziaływania z białkami AGO mimo braku podobieństwa sekwencji [9]. Podobna interakcja WG/GW-AGO została zarejestrowana między organizmami, które dzieli jeszcze większa odległość ewolucyjną - domeną WG/GW człowieka i białkiem AGO prokariotycznego archeonu *A. fulgidus* [7]. Ze względu na uniwersalną funkcjonalność domen WG/GW pochodzących z różnych rodzin białkowych i organizmów oraz możliwość wygenerowania syntetycznej domeny przez wprowadzenie kilku motywów WG/GW, można więc przyjąć, że w domenach WG/GW, bardziej preferowany jest stan nieuporządkowania struktury, niż dążenie do utrzymania sztywnego zwoju. Daje to, w pewien sposób, przewagę regionów nieustrukturyzwanych nad białkami całkowicie sfałdowanymi, poprzez obejście sterycznych ograniczeń, gwarantując powstanie większej powierzchni kontaktów w oddziaływaniach kompleksów. Ponadto brak strukturalnego uporządkowania umożliwia tworzenie słabo zasocjowanych, wielopodjednostkowych kompleksów białkowych, które mogą podlegać takim przekształceniom konformacyjnym, jakie są preferowane w określonym stanie funkcjonowania komórki.

Podsumowanie

W ramach niniejszego projekty zrealizowane zostały następujące cele:

1. Opracowano trzy metody adnotacji białek zawierających domenę wiążącą AGO, które zaimplementowano w formie ogólnodostępnych aplikacji internetowych i/lub programów typu desktop (program Agos, Wsearch i i-Wsearch).
2. Zaprezentowano listę wraz z oceną wiarygodności nowych rodzin białkowych zawierających domenę WG/GW we wszystkich domenach życia oraz wirusach. Aktywność wiązania AGO została już potwierdzona eksperymentalnie w trzech zidentyfikowanych białkach (SDE3, WGRP1, hnRNP).
3. Wykazano, że mimo wysokiego zróżnicowania długości i stopnia zachowania sekwencji domen WG/GW, ich specyficzna kompozycja aminokwasowa jest zachowana u wszystkich organizmów eukariotycznych. Ponadto domeny te składają się z wielokrotnie, powtórzonych motywów długości 10-20 reszt aminokwasowych, wewnątrz których znajduje się reszta Trp otoczona hydrofilowymi resztami.
4. Zidentyfikowano wpływ mechanizmów molekularnych warunkujących zmienność domen wiążących AGO. Oprócz tandemowych i segmentowych duplikacji oraz alternatywnego splicingu, fragmenty sekwencji kodujące motywy WG/GW znajdują się pod działaniem pozytywnej selekcji przyspieszającej utrwalanie substytucji aminokwasowych, które dodatkowo podlegają licznym przetasowaniom między paralogami na drodze częstych rekombinacji i nierównego crossing-over. Ponadto zarówno zwierzęce, jak i roślinne domeny wiążące AGO, występujące w różnych rodzinach białkowych, powstają z regionów białek IDP, obejmujących także regiony glicynobogate.

Wykaz skrótów

AGO	białko Argonaute
dsRNA	dwuniciowy RNA (ang. *double-stranded RNA*)
hnRNP	heterogenne rybonukleoproteiny, (ang. *heterogeneous nuclear ribonucleoproteins*)
IDP	białka inherentnie nieuporządkowane (ang. *Intrinsically Disordered Protein*)
miRNA	cząsteczka mikroRNA
PSSM	pozycyjnie specyficzna macierz punktacji (ang. *Position-Specific Scoring Matrix*)
PTGS	post-transkrypcyjne wyciszanie genów (ang. *post-transcriptional gene silencing*).
RdDM	metylacja DNA kierowana przez RNA (ang. *RNA-directed DNA methylation*)
RdRP	polimeraza RNA zależna od RNA (ang. *RNA-dependent RNA polymerase*)
RISC	kompleks efektorowy RNAi (ang. *RNA-induced silencing complex*)
RITS	kompleks efektorowy RNAi związany z wyciszaniem transkrypcyjnym (ang. *RNA-induced transcriptional gene silencing*)
RNAi	interferencja RNA (ang. *RNA interference*)
RNP	kompleks białkowo-rybonukleinowy (ang. *ribonucleoprotein*)
RRM	domena wiążąca RNA (ang. *RNA-recognition motif*)
siRNA	cząsteczka interferującego RNA (ang. *small interfering RNA*)
srRNA	małe regulatorowe cząsteczki RNA(ang. *small regulatory RNA*)a
SSB	białko wiążące jednodniciowy DNA (ang. *single-stranded DNA-binding*)
TGS	transkrypcyjne wyciszanie genów (ang. *transcriptional gene silencing*)
W-motyw	sekwencja zawierająca pojedyncze wystąpienie Trp
WG/GW	sekwencja WG lub GW

Spis rysunków i tabel

Bibliografia

[1] Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res **2012**;40:D290–301.

[2] Fox NK, Brenner SE, Chandonia J-M. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res **2014**;42:D304–9.

[3] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **1997**;25:3389–402.

[4] Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res **2011**;39:W29–37.

[5] Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, et al. New and continuing developments at PROSITE. Nucleic Acids Res **2013**;41:D344–7.

[6] Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, et al. InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res **2012**;40:D306–12.

[7] Till S, Lejeune E, Thermann R, Bortfeld M, Hothorn M, Enderle D, et al. A conserved motif in Argonaute-interacting proteins mediates functional interactions through the Argonaute PIWI domain. Nat Struct Mol Biol **2007**;14:897–903.

[8] Partridge JF, DeBeauchamp JL, Kosinski AM, Ulrich DL, Hadler MJ, Noffsinger VJP. Functional separation of the requirements for establishment and maintenance of centromeric heterochromatin. Mol Cell **2007**;26:593–602.

[9] El-Shami M, Pontier D, Lahmy S, Braun L, Picart C, Vega D, et al. Reiterated WG/GW motifs form functionally and evolutionarily conserved ARGONAUTE-binding platforms in RNAi-related components. Genes Dev **2007**;21:2539–44.

[10] Bednenko J, Noto T, DeSouza L V, Siu KWM, Pearlman RE, Mochizuki K, et al. Two GW repeat proteins interact with Tetrahymena thermophila argonaute and promote genome rearrangement. Mol Cell Biol **2009**;29:5020–30.

[11] Czech B, Hannon GJ. Small RNA sorting: matchmaking for Argonautes. Nat Rev Genet **2011**;12:19–31.

[12] Vasudevan S, Tong Y, Steitz JA. Switching from repression to activation: microRNAs can up-regulate translation. Science **2007**;318:1931–4.

[13] Hannon GJ. RNA interference. Nature **2002**;418:244–51.

[14] Meister G, Tuschl T. Mechanisms of gene silencing by double-stranded RNA. Nature **2004**;431:343–9.

[15] Djuranovic S, Nahvi A, Green R. A parsimonious model for gene regulation by miRNAs. Science **2011**;331:550–3.

[16] Meister G, Tuschl T. Mechanisms of gene silencing by double-stranded RNA. Nature **2004**;431:343–9.

[17] Hutvagner G, Simard MJ. Argonaute proteins: key players in RNA silencing. Nat Rev Mol Cell Biol **2008**;9:22–32.

[18] Ghildiyal M, Xu J, Seitz H, Weng Z, Zamore PD. Sorting of Drosophila small silencing RNAs partitions microRNA* strands into the RNA interference pathway. RNA **2010**;16:43–56.

[19] Su H, Trombly MI, Chen J, Wang X. Essential and overlapping functions for mammalian Argonautes in microRNA silencing. Genes Dev **2009**;23:304–17.

[20] Eystathioy T, Chan EKL, Tenenbaum SA, Keene JD, Griffith K, Fritzler MJ. A phosphorylated cytoplasmic autoantigen, GW182, associates with a unique population of human mRNAs within novel cytoplasmic speckles. Mol Biol Cell **2002**;13:1338–51.

[21] Ding L, Spencer A, Morita K, Han M. The developmental timing regulator AIN-1 interacts with miRISCs and may target the argonaute protein ALG-1 to cytoplasmic P bodies in C. elegans. Mol Cell **2005**;19:437–47.

[22] Liu J, Rivas F V, Wohlschlegel J, Yates JR, Parker R, Hannon GJ. A role for the P-body component GW182 in microRNA function. Nat Cell Biol **2005**;7:1261–6.

[23] Meister G, Landthaler M, Peters L, Chen PY, Urlaub H, Lührmann R, et al. Identification of novel argonaute-associated proteins. Curr Biol **2005**;15:2149–55.

[24] Rehwinkel J, Behm-Ansmant I, Gatfield D, Izaurralde E. A crucial role for GW182 and the DCP1:DCP2 decapping complex in miRNA-mediated gene silencing. RNA **2005**;11:1640–7.

[25] Behm-Ansmant I, Rehwinkel J, Doerks T, Stark A, Bork P, Izaurralde E. mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. Genes Dev **2006**;20:1885–98.

[26] Braun JE, Huntzinger E, Izaurralde E. Ten Years of Progress in GW/P Body Research. Adv Exp Med Biol **2013**;738.

[27] Eulalio A, Tritschler F, Izaurralde E. The GW182 protein family in animal cells: new insights into domains required for miRNA-mediated gene silencing. RNA **2009**;15:1433–42.

[28] Ding L, Han M. GW182 family proteins are crucial for microRNA-mediated gene silencing. Trends Cell Biol **2007**;17:411–6.

[29] Eulalio A, Helms S, Fritzsch C, Fauser M, Izaurralde E. A C-terminal silencing domain in GW182 is essential for miRNA function. RNA **2009**;15:1067–77.

[30] Lazzaretti D, Tournier I, Izaurralde E. The C-terminal domains of human TNRC6A, TNRC6B, and TNRC6C silence bound transcripts independently of Argonaute proteins. RNA **2009**;15:1059–66.

[31] Takimoto K, Wakiyama M, Yokoyama S. Mammalian GW182 contains multiple Argonaute-binding sites and functions in microRNA-mediated translational repression. RNA **2009**;15:1078–89.

[32] Eulalio A, Tritschler F, Büttner R, Weichenrieder O, Izaurralde E, Truffault V. The RRM domain in GW182 proteins contributes to miRNA-mediated gene silencing. Nucleic Acids Res **2009**;37:2974–83.

[33] Havecker ER, Wallbridge LM, Hardcastle TJ, Bush MS, Kelly KA, Dunn RM, et al. The Arabidopsis RNA-directed DNA methylation argonautes functionally diverge based on their expression and interaction with target loci. Plant Cell **2010**;22:321–34.

[34] He X-J, Hsu Y-F, Zhu S, Wierzbicki AT, Pontes O, Pikaard CS, et al. An effector of RNA-directed DNA methylation in arabidopsis is an ARGONAUTE 4- and RNA-binding protein. Cell **2009**;137:498–508.

[35] Bies-Etheve N, Pontier D, Lahmy S, Picart C, Vega D, Cooke R, et al. RNA-directed DNA methylation requires an AGO4-interacting member of the SPT5 elongation factor family. EMBO Rep **2009**;10:649–54.

[36] Azevedo J, Garcia D, Pontier D, Ohnesorge S, Yu A, Garcia S, et al. Argonaute quenching and global changes in Dicer homeostasis caused by a pathogen-encoded GW repeat protein. Genes Dev **2010**;24:904–15.

[37] Giner A, Lakatos L, García-Chapa M, López-Moya JJ, Burgyán J. Viral protein inhibits RISC activity by argonaute binding through conserved WG/GW motifs. PLoS Pathog **2010**;6:e1000996.

[38] De Ronde D, Pasquier A, Ying S, Butterbach P, Lohuis D, Kormelink R. Analysis of Tomato spotted wilt virus NSs protein indicates the importance of the N-terminal domain for avirulence and RNA silencing suppression. Mol Plant Pathol **2014**;15:185–95.

[39] Zhang L, Ding L, Cheung TH, Dong M-Q, Chen J, Sewell AK, et al. Systematic identification of C. elegans miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2. Mol Cell **2007**;28:598–613.

[40] Ding XC, Grosshans H. Repression of C. elegans microRNA targets at the initiation level of translation requires GW182 proteins. EMBO J **2009**;28:213–22.

[41] Rost B. Twilight zone of protein sequence alignments. Protein Eng **1999**;12:85–94.

[42] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A **1988**;85:2444–8.

[43] Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, et al. PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res **2003**;31:400–2.

[44] Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, et al. The Pfam protein families database. Nucleic Acids Res **2010**;38:D211–22.

[45] Lawrence C, Altschul S, Boguski M, Liu J, Neuwald A, Wootton J. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science **1993**;262:208–14.

[46] Thompson WA, Newberg LA, Conlan S, McCue LA, Lawrence CE. The Gibbs Centroid Sampler. Nucleic Acids Res **2007**;35:W232–7.

[47] Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res **2006**;34:W369–73.

[48] Bonham-Carter O, Steele J, Bastola D. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. Brief Bioinform **2013**.

[49] The UniProt Consortium. UniProtKB/Swiss-Prot protein knowledgebase release 2014_04.

[50] Terrapon N, Weiner J, Grath S, Moore AD, Bornberg-Bauer E. Rapid similarity search of proteins using alignments of domain arrangements. Bioinformatics **2014**;30:274–81.

[51] Moore AD, Held A, Terrapon N, Weiner J, Bornberg-Bauer E. DoMosaics: software for domain arrangement visualization and domain-centric analysis of proteins. Bioinformatics **2014**;30:282–3.

[52] Oliphant TE. Python for Scientific Computing. Comput Sci Eng **2007**;9:10–20.

[53] Hirsh AE, Fraser HB. Protein dispensability and rate of evolution. Nature **2001**;411:1046–9.

[54] EasyFit. http://www.mathwave.com/.

[55] Hollander M. Nonparametric statistical methods. 2nd ed. New York: Wiley; **1999**.

[56] Karlowski WM, Zielezinski A, Carrère J, Pontier D, Lagrange T, Cooke R. Genome-wide computational identification of WG/GW Argonaute-binding proteins in Arabidopsis. Nucleic Acids Res **2010**;38:4231–45.

[57] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **2006**;22:1658–9.

[58] Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol **2011**;7:e1002195.

[59] Dosztányi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics **2005**;21:3433–4.

[60] Breiman L. Random Forests. Mach Learn **2001**;45:5–32.

[61] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol **1982**;157:105–32.

[62] Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. Proteins **1994**;19:141–9.

[63] Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci U S A **1981**;78:3824–8.

[64] Zamyatnin AA. Protein volume in solution. Prog Biophys Mol Biol **1972**;24:107–23.

[65] Chothia C. The nature of the accessible and buried surfaces in proteins. J Mol Biol **1976**;105:1–12.

[66] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel V, and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer P, and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos A and, Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay E. Scikit-learn: Machine Learning in Python. J Mach Learn Res **2011**;12:2825–30.

[67] Liu Z-P, Wu L-Y, Wang Y, Zhang X-S, Chen L. Prediction of protein-RNA binding sites by a random forest method with combined features. Bioinformatics **2010**;26:1616–22.

[68] Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, et al. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res **2008**;36:D1009–14.

[69] Ouyang S, Zhu W, Hamilton J, Lin H, Campbell et al. The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res **2007**;35:D883–7.

[70] Grigoriev I V, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, et al. The genome portal of the Department of Energy Joint Genome Institute. Nucleic Acids Res **2012**;40:D26–32.

[71] Katoh K, Asimenos G, Toh H. Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol **2009**;537:39–64.

[72] Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol **2010**;59:307–21.

[73] Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. Bayesian inference of phylogeny and its impact on evolutionary biology. Science **2001**;294:2310–4.

[74] Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **2003**;19:1572–4.

[75] Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol Biol Evol **2008**;25:1307–20.

[76] Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics **2005**;21:2104–5.

[77] Abascal F, Zardoya R, Telford MJ. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res **2010**;38:W7–13.

[78] Zhang Z, Li J, Zhao X-Q, Wang J, Wong GK-S, Yu J. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics **2006**;4:259–63.

[79] Sawyer S. GENECONV: a computer package for the statistical detection of gene conversion **1999**.

[80] Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics **2010**;26:2462–3.

[81] Project D. Django 2013.

[82] Hipp R. SQlite3 2000. http://sqlite.org/

[83] Twitter inc. Bootstrap. http://getbootstrap.com/

[84] Resig J. jQuery 2006.

[85] Bostock M. D3 for Data-Driven Documents.

[86] Garcia D, Garcia S, Pontier D, Marchais A, Renou JP, Lagrange T, et al. Ago hook and RNA helicase motifs underpin dual roles for SDE3 in antiviral defense and silencing of nonconserved intergenic regions. Mol Cell **2012**;48:109–20.

[87] Braun JE, Huntzinger E, Fauser M, Izaurralde E. GW182 proteins directly recruit cytoplasmic deadenylase complexes to miRNA targets. Mol Cell **2011**;44:120–33.

[88] Fabian MR, Cieplak MK, Frank F, Morita M, Green J, Srikumar T, et al. miRNA-mediated deadenylation is orchestrated by GW182 through two conserved motifs that interact with CCR4-NOT. Nat Struct Mol Biol **2011**;18:1211–7.

[89] Chekulaeva M, Mathys H, Zipprich JT, Attig J, Colic M, Parker R, et al. miRNA repression involves GW182-mediated recruitment of CCR4-NOT through conserved W-containing motifs. Nat Struct Mol Biol **2011**;18:1218–26.

[90] Azevedo J, Cooke R, Lagrange T. Taking RISCs with Ago hookers. Curr Opin Plant Biol **2011**;14:594–600.

[91] Pfaff J, Hennig J, Herzog F, Aebersold R, Sattler M, Niessing D, et al. Structural features of Argonaute-GW182 protein interactions. Proc Natl Acad Sci U S A **2013**;110(40).

[92] Lian SL, Li S, Abadal GX, Pauley BA, Fritzler MJ, Chan EKL. The C-terminal half of human Ago2 binds to multiple GW-rich regions of GW182 and requires GW182 to mediate silencing. RNA **2009**;15:804–13.

[93] Zipprich JT, Bhattacharyya S, Mathys H, Filipowicz W. Importance of the C-terminal domain of the human GW182 protein TNRC6C for translational repression. RNA **2009**;15:781–93.

[94] Stöhr JR. Proteomic and functional characterization of human Argonaute complexes. Ludwig-Maximilians-Universität München, **2011**.

[95] Gibbings D, Leblanc P, Jay F, Pontier D, Michel F, Schwab Y, et al. Human prion protein binds Argonaute and promotes accumulation of microRNA effector complexes. Nat Struct Mol Biol **2012**;19:517–24, S1.

[96] Beckham C, Hilliker A, Cziko A-M, Noueiry A, Ramaswami M, Parker R. The DEAD-box RNA helicase Ded1p affects and accumulates in Saccharomyces cerevisiae P-bodies. Mol Biol Cell **2008**;19:984–93.

[97] Chu C, Rana TM. Translation repression in human cells by microRNA-induced gene silencing requires RCK/p54. PLoS Biol **2006**;4:e210.

[98] Michlewski G, Guil S, Semple CA, Cáceres JF. Posttranscriptional regulation of miRNAs harboring conserved terminal loops. Mol Cell **2008**;32:383–93.

[99] Temme C, Zhang L, Kremmer E, Ihling C, Chartier A, Sinz A, et al. Subunits of the Drosophila CCR4-NOT complex and their roles in mRNA deadenylation. RNA **2010**;16:1356–70.

[100] Nowacki M, Zagorski-Ostoja W, Meyer E. Nowa1p and Nowa2p: novel putative RNA binding proteins involved in trans-nuclear crosstalk in Paramecium tetraurelia. Curr Biol **2005**;15:1616–28.

[101] Wiedenheft B, Sternberg SH, Doudna J a. RNA-guided genetic silencing systems in bacteria and archaea. Nature **2012**;482:331–8.

[102] Song J-J, Smith SK, Hannon GJ, Joshua-Tor L. Crystal structure of Argonaute and its implications for RISC slicer activity. Science **2004**;305:1434–7.

[103] Makarova KS, Wolf YI, Koonin E V. Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Res **2013**;41:4360–77.

[104] Ma J-B, Yuan Y-R, Meister G, Pei Y, Tuschl T, Patel DJ. Structural basis for 5'-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. Nature **2005**;434:666–70.

[105] Makarova KS, Wolf YI, van der Oost J, Koonin E V. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. Biol Direct **2009**;4:29.

[106] Li Y, Liu X, Huang L, Guo H, Wang X-J. Potential coexistence of both bacterial and eukaryotic small RNA biogenesis and functional related protein homologs in Archaea. J Genet Genomics **2010**;37:493–503.

[107] Mijakovic I, Petranovic D, Macek B, Cepo T, Mann M, Davies J, et al. Bacterial single-stranded DNA-binding proteins are phosphorylated on tyrosine. Nucleic Acids Res **2006**;34:1588–96.

[108] Wang K, Li H, Yuan Y, Etheridge A, Zhou Y, Huang D, et al. The complex exogenous RNA spectra in human plasma: an interface with human gut biota? PLoS One **2012**;7:e51009.

[109] Yan Y, Cui H, Jiang S, Huang Y, Huang X, Wei S, et al. Identification of a novel marine fish virus, Singapore grouper iridovirus-encoded microRNAs expressed in grouper cells by

Solexa sequencing. PLoS One **2011**;6:e19148.

[110] Umbach JL, Strelow LI, Wong SW, Cullen BR. Analysis of rhesus rhadinovirus microRNAs expressed in virus-induced tumors from infected rhesus macaques. Virology **2010**;405:592–9.

[111] Ouellet DL, Vigneault-Edwards J, Létourneau K, Gobeil L-A, Plante I, Burnett JC, et al. Regulation of host gene expression by HIV-1 TAR microRNAs. Retrovirology **2013**;10:86.

[112] Aqil M, Naqvi AR, Bano AS, Jameel S. The HIV-1 Nef Protein Binds Argonaute-2 and Functions as a Viral Suppressor of RNA Interference. PLoS One **2013**;8:e74472.

[113] Bronkhorst AW, van Cleef KWR, Vodovar N, Ince IA, Blanc H, Vlak JM, et al. The DNA virus Invertebrate iridescent virus 6 is a target of the Drosophila RNAi machinery. Proc Natl Acad Sci U S A **2012**;109:E3604–13.

[114] Kemp C, Mueller S, Goto A, Barbier V, Paro S, Bonnay F, Dostert C, Troxler L, Hetru C, Meignin C, Pfeffer S, Hoffmann JA IJ. Broad RNA interference-mediated antiviral immunity and virus-specific inducible responses in Drosophila. J Immunol **2013**;190.

[115] Conrad KD, Giering F, Erfurth C, Neumann A, Fehr C, Meister G, et al. MicroRNA-122 dependent binding of Ago2 protein to hepatitis C virus RNA is associated with enhanced RNA stability and translation stimulation. PLoS One **2013**;8:e56272.

[116] Rowe JM, Dunigan DD, Blanc G, Gurnon JR, Xia Y, Van Etten JL. Evaluation of higher plant virus resistance genes in the green alga, Chlorella variabilis NC64A, during the early phase of infection with Paramecium bursaria chlorella virus-1. Virology **2013**;442:101–13.

[117] Baumberger N, Tsai C-H, Lie M, Havecker E, Baulcombe DC. The Polerovirus silencing suppressor P0 targets ARGONAUTE proteins for degradation. Curr Biol **2007**;17:1609–14.

[118] Schirle NT, MacRae IJ. The crystal structure of human Argonaute2. Science **2012**;336:1037–40.

[119] Zielezinski A, Karlowski WM. Agos--a universal web tool for GW Argonaute-binding domain prediction. Bioinformatics **2011**;27:1318–9.

[120] Chekulaeva M, Parker R, Filipowicz W. The GW/WG repeats of Drosophila GW182 function as effector motifs for miRNA-mediated repression. Nucleic Acids Res **2010**;38:6673–83.

[121] Szabó EZ, Manczinger M, Göblös A, Kemény L, Lakatos L. Switching on RNA silencing suppressor activity by restoring argonaute binding to a viral protein. J Virol **2012**;86:8324–7.

[122] Shiu S, Karlowski WM, Pan R, Tzeng Y, Mayer KFX, Li W. Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. Plant Cell **2004**;16:1220–34.

[123] Cléry A, Blatter M, Allain FH-T. RNA recognition motifs: boring? Not quite. Curr Opin Struct Biol **2008**;18:290–8.

[124] Höck J, Weinmann L, Ender C, Rüdel S, Kremmer E, Raabe M, et al. Proteomic and

functional analysis of Argonaute-containing mRNA-protein complexes in human cells. EMBO Rep **2007**;8:1052–60.

[125] Lin J-C, Tarn W-Y. RNA-binding motif protein 4 translocates to cytoplasmic granules and suppresses translation via argonaute2 during muscle cell differentiation. J Biol Chem **2009**;284:34658–65.

[126] Bäurle I, Dean C. Differential interactions of the autonomous pathway RRM proteins and chromatin regulators in the silencing of Arabidopsis targets. PLoS One **2008**;3:e2733.

[127] Shepard PJ, Hertel KJ. Protein family review The SR protein family. Genome Biol **2009**:1–9.

[128] Maitra R, Sadofsky MJ. A WW-like module in the RAG1 N-terminal domain contributes to previously unidentified protein-protein interactions. Nucleic Acids Res **2009**;37:3301–9.

[129] Guo Y, Yuan C, Tian F, Huang K, Weghorst CM, Tsai M-D, et al. Contributions of conserved TPLH tetrapeptides to the conformational stability of ankyrin repeat proteins. J Mol Biol **2010**;399:168–81.

[130] Burgler C, Macdonald PM. Prediction and verification of microRNA targets by MovingTargets, a highly adaptable prediction method. BMC Genomics **2005**;6:88.

[131] Blanc G, Hokamp K, Wolfe KH. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res **2003**;13:137–44.

[132] Project Rice Genome Annotation. FTP serwer.

[133] Eddy SR. Where did the BLOSUM62 alignment score matrix come from? Nat Biotechnol **2004**;22:1035–6.

[134] Anisimova M, Nielsen R, Yang Z. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics **2003**;164:1229–36.

[135] Suzuki Y, Nei M. False-positive selection identified by ML-based methods: examples from the Sig1 gene of the diatom Thalassiosira weissflogii and the tax gene of a human T-cell lymphotropic virus. Mol Biol Evol **2004**;21:914–21.

[136] Wang GL, Ruan DL, Song WY, Sideris S, Chen L, Pi LY, et al. Xa21D encodes a receptor-like molecule with a leucine-rich repeat domain that determines race-specific recognition and is subject to adaptive evolution. Plant Cell **1998**;10:765–79.

[137] Chen Q, Han Z, Jiang H, Tian D, Yang S. Strong positive selection drives rapid diversification of R-genes in Arabidopsis relatives. J Mol Evol **2010**;70:137–48.

[138] Hurles M. Gene duplication: the genomic trade in spare parts. PLoS Biol **2004**;2:E206.

[139] Nei M, Rooney AP. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet **2005**;39:121–52.

[140] Beisswanger S, Stephan W. Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated polyhomeotic genes in Drosophila. Proc Natl Acad Sci U S A **2008**;105:5447–52.

[141] Goldstone HMH, Stegeman JJ. A revised evolutionary history of the CYP1A subfamily: gene duplication, gene conversion, and positive selection. J Mol Evol **2006**;62:708–17.

[142] Liu Q. Gene conversion and positive selection driving the evolution of the Caenorhabditis ssp. ZIM/HIM-8 protein family. J Mol Evol **2009**;68:217–26.

[143] Zhang L. Adaptive evolution and frequent gene conversion in the brain expressed X-linked gene family in mammals. Biochem Genet **2008**;46:293–311.

[144] Aguilar A, Garza JC. Patterns of historical balancing selection on the salmonid major histocompatibility complex class II beta gene. J Mol Evol **2007**;65:34–43.

[145] Mondragon-Palomino M, Gaut BS. Gene conversion and the evolution of three leucine-rich repeat gene families in Arabidopsis thaliana. Mol Biol Evol **2005**;22:2444–56.

[146] Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, et al. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature **2008**;455:1193–7.

[147] Bühler M, Verdel A, Moazed D. Tethering RITS to a nascent transcript initiates RNAi- and heterochromatin-dependent gene silencing. Cell **2006**;125:873–86.

[148] Gomes MPB, Cordeiro Y, Silva JL. The peculiar interaction between mammalian prion protein and RNA. Prion **2008**;2:64–6.

[149] Deleault NR, Lucassen RW, Supattapone S. RNA molecules stimulate prion protein conversion. Nature **2003**;425:717–20.

[150] Mosher RA, Schwach F, Studholme D, Baulcombe DC. PolIVb influences RNA-directed DNA methylation independently of its role in siRNA biogenesis. Proc Natl Acad Sci U S A **2008**;105:3145–50.

[151] Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, et al. Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. Science **2010**;329:223–6.

[152] Zhao T, Li G, Mi S, Li S, Hannon GJ, Wang X-J, et al. A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii. Genes Dev **2007**;21:1190–203.

[153] Feng S, Cokus SJ, Zhang X, Chen P-Y, Bostick M, Goll MG, et al. Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A **2010**;107:8689–94.

[154] Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, et al. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science **2008**;319:64–9.

[155] Cho SH, Addo-Quaye C, Coruh C, Arif MA, Ma Z, Frank W, et al. Physcomitrella patens DCL3 is required for 22-24 nt siRNA accumulation, suppression of retrotransposon-derived transcripts, and normal development. PLoS Genet **2008**;4:e1000314.

[156] Cerutti H, Casas-Mollano JA. On the origin and functions of RNA-mediated silencing: from

protists to man. Curr Genet **2006**;50:81–99.

[157] Margis R, Fusaro AF, Smith NA, Curtin SJ, Watson JM, Finnegan EJ, et al. The evolution and diversification of Dicers in plants. FEBS Lett **2006**;580:2442–50.

[158] Jakymiw A, Lian S, Eystathioy T, Li S, Satoh M, Hamel JC, et al. Disruption of GW bodies impairs mammalian RNA interference. Nat Cell Biol **2005**;7:1167–74.

[159] Landthaler M, Gaidatzis D, Rothballer A, Chen PY, Soll SJ, Dinic L, et al. Molecular characterization of human Argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. RNA **2008**;14:2580–96.

[160] Parham P, Ohta T. Population biology of antigen presentation by MHC class I molecules. Science **1996**;272:67–74.

[161] Ohta T. Role of diversifying selection and gene conversion in evolution of major histocompatibility complex loci. Proc Natl Acad Sci U S A **1991**;88:6716–20.

[162] Gyllensten U, Sundvall M, Ezcurra I, Erlich HA. Genetic diversity at class II DRB loci of the primate MHC. J Immunol **1991**;146:4368–76.

[163] Ohta T. Role of gene conversion in generating polymorphisms at major histocompatibility complex loci. Hereditas **1997**;127:97–103.

[164] Parniske M, Hammond-Kosack KE, Golstein C, Thomas CM, Jones DA, Harrison K, et al. Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the Cf-4/9 locus of tomato. Cell **1997**;91:821–32.

[165] Baumgarten A, Cannon S, Spangler R, May G. Genome-level evolution of resistance genes in Arabidopsis thaliana. Genetics **2003**;165:309–19.

[166] Kuang H, Woo S-S, Meyers BC, Nevo E, Michelmore RW. Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. Plant Cell **2004**;16:2870–94.

[167] Ding S-W, Voinnet O. Antiviral immunity directed by small RNAs. Cell **2007**;130:413–26.

[168] Bivalkar-Mehla S, Vakharia J, Mehla R, Abreha M, Kanwar JR, Tikoo A, et al. Viral RNA silencing suppressors (RSS): novel strategy of viruses to ablate the host RNA interference (RNAi) defense system. Virus Res **2011**;155:1–9.

[169] Zhang X, Yuan Y-R, Pei Y, Lin S-S, Tuschl T, Patel DJ, et al. Cucumber mosaic virus-encoded 2b suppressor inhibits Arabidopsis Argonaute1 cleavage activity to counter plant defense. Genes Dev **2006**;20:3255–68.

[170] Bortolamiol D, Pazhouhandeh M, Ziegler-Graff V. Viral suppression of RNA silencing by destabilisation of ARGONAUTE 1. Plant Signal Behav **2008**;3:657–9.

[171] Csorba T, Lózsa R, Hutvágner G, Burgyán J. Polerovirus protein P0 prevents the assembly of small RNA-containing RISC complexes and leads to degradation of ARGONAUTE1. Plant J **2010**;62:463–72.

[172] Obbard DJ, Jiggins FM, Halligan DL, Little TJ. Natural selection drives extremely rapid

evolution in antiviral RNAi genes. Curr Biol **2006**;16:580–5.

[173] Aravin AA, Hannon GJ, Brennecke J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science **2007**;318:761–4.

[174] Marques JT, Carthew RW. A call to arms: coevolution of animal viruses and host innate immune responses. Trends Genet **2007**;23:359–64.

[175] Moissiard G, Voinnet O. RNA silencing of host transcripts by cauliflower mosaic virus requires coordinated action of the four Arabidopsis Dicer-like proteins. Proc Natl Acad Sci U S A **2006**;103:19593–8.

[176] Hull R. Comparative Plant Virology, Second Edition. Second Edition. Academic Press; **2009**.

[177] Tracewell C a, Arnold FH. Directed enzyme evolution: climbing fitness peaks one amino acid at a time. Curr Opin Chem Biol **2009**;13:3–9.

[178] Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol **2005**;6:197–208.

[179] Davey NE, Shields DC, Edwards RJ. Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. Bioinformatics **2009**;25:443–50.

[180] Davey NE, Cowan JL, Shields DC, Gibson TJ, Coldwell MJ, Edwards RJ. SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. Nucleic Acids Res **2012**;40:10628–41.

[181] Fuxreiter M, Tompa P, Simon I. Local structural disorder imparts plasticity on linear motifs. Bioinformatics **2007**;23:950–6.

[182] Dinkel H, Michael S, Weatheritt RJ, Davey NE, Van Roey K, Altenberg B, et al. ELM--the database of eukaryotic linear motifs. Nucleic Acids Res **2012**;40:D242–51.

[183] Dinkel H, Van Roey K, Michael S, Davey NE, Weatheritt RJ, Born D, et al. The eukaryotic linear motif resource ELM: 10 years and counting. Nucleic Acids Res **2014**;42:D259–66.

[184] Tompa P. Intrinsically unstructured proteins evolve by repeat expansion. BioEssays **2003**;25:847–55.

[185] Zahn R, Liu A, Luhrs T, Riek R, von Schroetter C, Lopez Garcia F, et al. NMR solution structure of the human prion protein. Proc Natl Acad Sci **2000**;97:145–50.

[186] Knaus KJ, Morillas M, Swietnicki W, Malone M, Surewicz WK, Yee VC. Crystal structure of the human prion protein reveals a mechanism for oligomerization. Nat Struct Biol **2001**;8:770–4.

[187] Lesk A. Introduction to Bioinformatics. 3rd ed. Oxford University Press, USA; **2008**.