Transkryptomika porównawcza wybranych gatunków zwierząt morskich
mgr Magdalena Małachowicz

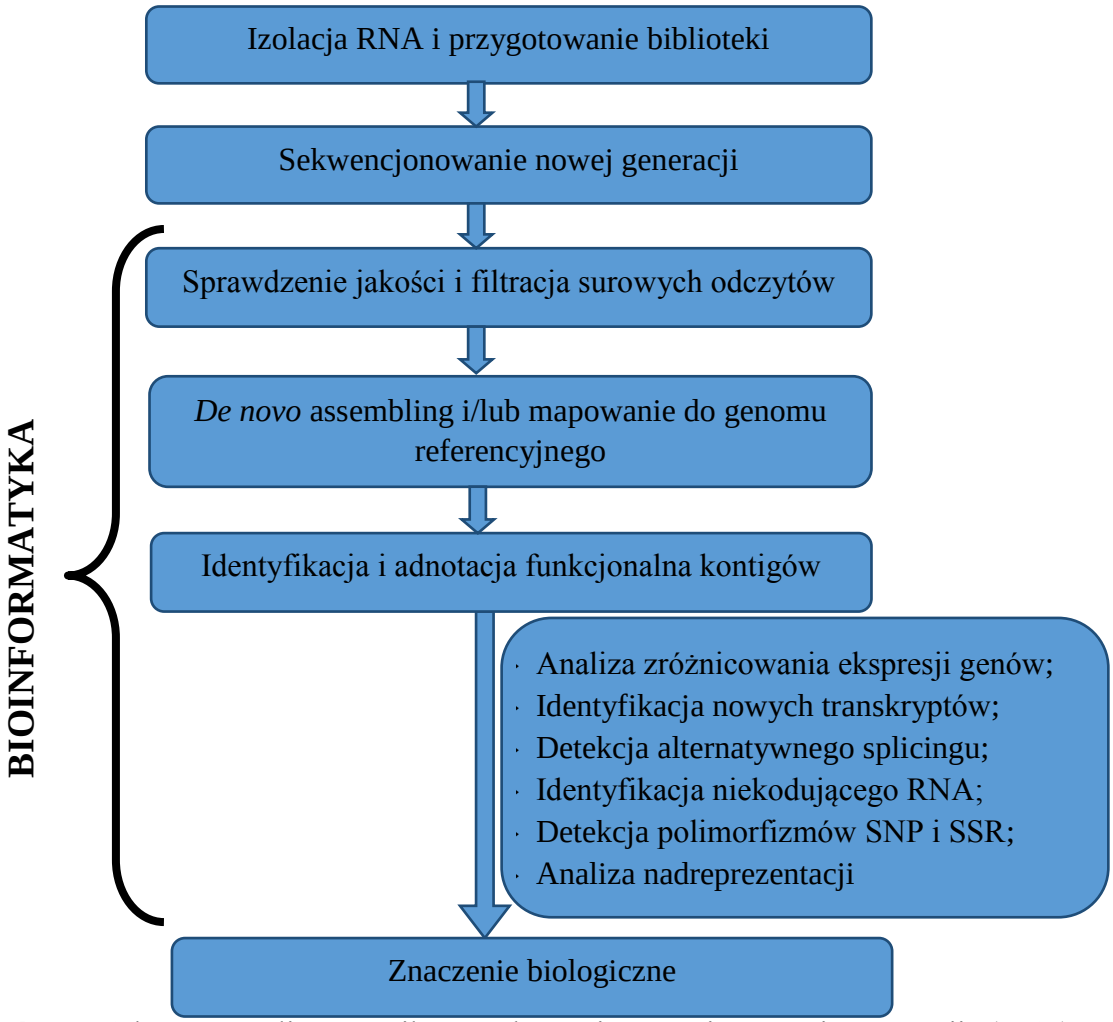
Transkryptomika dotyczy badania pełnego zestawu transkryptów (cząsteczek RNA syntetyzowanych na matrycy DNA na drodze transkrypcji) obecnego w danej komórce lub tkance w określonym momencie. Transkryptom obejmuje kodujące białka matrycowe RNA (mRNA) oraz niekodujące RNA (ncRNA), istotne w prawidłowym funkcjonowaniu komórek, w tym mikroRNA (miRNA), rybosomalne RNA (rRNA), transferowe RNA (tRNA) i inne ncRNA. W przeciwieństwie do genomu, transkryptom różni się w zależności od badanego typu tkanki, stadium rozwoju organizmu, stanu fizjologicznego oraz warunków środowiska. Analiza transkryptomu dostarcza informacji o ekspresji, funkcji i regulacji genów, pozwala na zrozumienie funkcjonowania poszczególnych tkanek oraz korelacji genotyp-fenotyp. Umożliwia to zrozumienie mechanizmów molekularnych leżących u podstaw różnorodności genetycznej organizmów oraz reakcji fizjologicznych na środowisko.

Obecnie do analizy transkryptomu wykorzystywane są dwie techniki: hybrydyzacja wyznakowanego cDNA/cRNA do sond cDNA/oligonukleotydowych umieszczonych na macierzy oraz sekwencjonowanie. Ze względu na ograniczenia występujące w technologii mikromacierzy (saturacja sygnału w przypadku transkryptów o wysokiej ekspresji, szum tła spowodowany niespecyficzną hybrydyzacją oraz zależność od posiadania danych genomowych), w ostatnich latach wzrasta zastosowanie metody sekwencjonowania do analizy transkryptomu. Technika ta jest szczególnie użyteczna zwłaszcza, gdy obiektem badań są organizmy niemodelowe, brakuje referencyjnego genomu lub zasoby genomowe są ograniczone. Oprócz analizy ekspresji genów, sekwencjonowanie transkryptomu umożliwia detekcję znanych transkryptów/izoform. Dostarcza informacji potrzebnych do adnotacji genomu – umożliwia lokalizację granicy intron-ekson. Ponadto pozwala na identyfikację nowych transkryptów, niekodującego RNA oraz alternatywnych wariantów składania (ang. alternative splicing variants), powstałych na drodze potranskrypcyjnych modyfikacji mRNA. Umożliwia przewidywanie potencjalnej struktury i funkcji białek bez wcześniejszej wiedzy o genach. Sekwencjonowanie transkryptomu jest także efektywnym sposobem na uzyskanie dużej liczby markerów molekularnych, takich jak polimorfizmy krótkich powtórzeń tandemowych (SSRs, ang. simple sequence repeats) oraz polimorfizmy pojedynczych nukleotydów (SNPs, ang. single nucleotide polymorphisms), występujących w regionach kodujących, określających funkcjonalną zmienność genetyczną.

Szybki rozwój transkryptomiki w ostatnim dziesięcioleciu został spowodowany rozwojem technologii sekwencjonowania nowej generacji (NGS, ang. next generation sequencing). NGS opiera się na równoległym, masowym sekwencjonowaniu od kilku tysięcy do miliardów odczytów w jednym przebiegu. Metody NGS można podzielić na dwie główne kategorie: sekwencjonowanie drugiej generacji (SGS, ang. second generation sequencing) oraz sekwencjonowanie trzeciej generacji (TGS, ang. third generation sequencing). Pierwsza grupa obejmuje technologie oparte o metodę PCR, tj. Roche/454 pirosekwencjonowanie (2005), sekwencjonowanie Illumina (2007) i Ion Torrent (2010) opierające się na sekwencjonowaniu przez syntezę (SBS, ang. sequencing by synthesis) oraz AB SOLiD (2007) - sekwencjonowanie przez ligację (SBL, ang. sequencing by ligation). Drugą grupę stanowią techniki umożliwiające bezpośredni odczyt sekwencji z pojedynczej cząsteczki DNA (SMS, ang. single-molecule sequencing) takie jak HeliScope (2008), system PacBio RS SMRT (2010) oraz sekwenator oparty na nanoporach - MinION Oxford Nanopore (2014). Główną zaletą technologii NGS, określanej również jako wysokoprzepustowe sekwencjonowanie (ang. high-throughput sequencing) jest możliwość taniego wytwarzania ogromnej ilości danych w krótkim czasie. Postęp ten zrewolucjonizował badania molekularne nad organizmami, w tym również morskimi. Rozwój wysokoprzepustowych technik i generacja dużej ilości danych w tym transkryptomicznych, doprowadziła do połączenia nauk biologicznych takich jak biologia molekularna, genomika, transkryptomika, z matematyką i informatyką, prowadząc do powstania interdyscyplinarnej dziedziny nauki – bioinformatyki. Analiza wyników sekwencjonowania składa się z wielu etapów, od ekstrakcji RNA, złożenia pojedynczych odczytów (ang. assembling), poprzez klasyfikację funkcjonalną i adnotację transkryptów po określanie znaczenia biologicznego (Rysunek 1). W tym celu opracowano liczne programy bioinformatyczne oparte na różnych algorytmach i modelach. Różnorodność oprogramowania umożliwia dostosowanie protokołu analizy do określonego celu.

Transkryptomika w połączeniu z sekwencjonowaniem nowej generacji, bioinformatyką i biostatystyką zapewniają nowe możliwości zrozumienia zagadnień biologicznych, w tym: procesów fizjologicznych i rozwoju organizmu, struktury i zmienności genetycznej populacji, adaptacji do warunków środowiskowych, ewolucji, filogenetyki i immunologii. Dzięki szerokiemu zastosowaniu stała się głównym narzędziem biologii ewolucyjnej. Poznanie różnorodności biologicznej poprzez analizę transkryptomu jest cennym źródłem informacji dla zrozumienia, w jaki sposób globalne zmiany klimatu, działania antropogeniczne i patogeny wpływają na organizmy. Poznanie tych zagadnień przyczynia się

do skutecznego zarządzania ochroną ważnych ekologicznie i ekonomicznie gatunków morskich.



Rysunek 1. Schemat analizy wyników sekwencjonowanie nowej generacji (NGS), od ekstrakcji RNA po określenie znaczenia biologicznego.

Głównym celem mojej pracy doktorskiej było opisanie różnorodności genetycznej wybranych gatunków morskich odgrywających ważną rolę w zaspokajaniu światowego zapotrzebowania na żywność, tj. ryb i małży o istotnym znaczeniu ekonomicznym i ekologicznym, w tym: dorsza atlantyckiego (*Gadus morhua*) z Morza Bałtyckiego, troci wędrownej (*Salmo trutta* m. *trutta*), małży z rodzaju *Mytilus* (*Mytilus edulis*, *Mytilus galloprovincialis*, *Mytilus trossulus* oraz *Mytilus chilensis*). W tym celu przeprowadziłam analizę transkryptomów z użyciem metody sekwencjonowania nowej generacji. Do badania zostały wybrane tkanki o kluczowym znaczeniu w kontakcie i odpowiedzi na warunki środowiskowe (skrzela - odpowiedź na zasolenie, skóra - pierwsza linia obrony organizmu, płaszcz - biomineralizacja). We wszystkich prezentowanych publikacjach, po izolacji RNA,

uzyskane próby były sekwencjonowanie przy użyciu techniki pirosekwencjonowania 454 (Roche GS-FLX). Następnie uzyskane sekwencje zostały wstępnie przetworzone przez usunięcie adapterów i sekwencji niskiej jakości za pomocą oprogramowania CLC Genomics Workbench (wer. 7.5.5, CLC Bio, Qiagen, Aarhus, Dania). Odczyty o dobrej jakości zostały złożone w kontigi przy użyciu algorytmu konstruującego wykres de Bruijna w programie CLC Genomics Workbench [1, 2, 3]. Uzyskane kontigi zidentyfikowałam przez przeszukiwanie publicznych baz danych, np. NCBI (ang. the National Center of Biotechnology Information) za pomocą narzędzi BLAST (ang. Basic Local Alignment Search Tool). Następnie przeprowadziłam klasyfikację funkcjonalną (ontologia genów, GO; program Blast2GO), przypisując kontigi do trzech głównych kategorii ontologicznych: funkcje molekularne (MF, ang. molecular function), procesy biologiczne (BP, ang. biological process) i elementy komórkowe (CC, ang. cellular component). Za pomocą odpowiednich narzędzi bioinformatycznych tj. KAAS (ang. KEGG Automatic Annotation Server) przeprowadziłam analizę szlaków metabolicznych i niemetabolicznych z użyciem bazy danych KEGG (Kyoto Encyclopedia of Genes and Genomes). Uzyskane wyniki porównałam z sekwencjami dostępnymi w bazach danych, pochodzącymi z różnych taksonów i tkanek (transkryptomika porównawcza) [1, 2, 3].

Pierwszy rozdział mojej pracy doktorskiej przedstawia porównanie transkryptomów dorsza atlantyckiego (*Gadus morhua* L.) z Oceanu Atlantyckiego i Morza Bałtyckiego [1]. Dorsz atlantycki jest ważnym, pod względem ekologicznym i ekonomicznym gatunkiem ryb morskich na Północnym Oceanie Atlantyckim. Dorsz w Morzu Bałtyckim żyje w warunkach niskiego zasolenia (7-12 PSU), jednakże do rozrodu wymaga wody dobrze natlenionej, o zasoleniu powyżej 14 PSU, co ogranicza jego tarliska do niewielu obszarów. W wyniku pirosekwencjonowania otrzymano 962516 odczytów, reprezentujących 379 Mbp transkryptomu dorsza bałtyckiego. Uzyskane sekwencje złożyłam w 14029 kontigi reprezentujące 13585 genów, z których 75.39% posiadało definicję ontologiczną, z dominującą kategorią funkcji molekularnych. Ponadto, uzyskane kontigi porównałam z transkryptomem dorsza z Oceanu Atlantyckiego (przy użyciu BLASTn), uzyskanym z bazy danych Ensembl. Pomimo wysokiego podobieństwa między transkryptomami, 202 kontigi, reprezentujące 170 genów dorsza bałtyckiego wykazały znaczną odrębność od dorsza atlantyckiego. Wśród różnicujących genów 100% stanowiły sekwencje kodujące białka, a do 68.82% zostały przypisane kategorie ontologiczne. Zidentyfikowane różnice mogą być spowodowane różnym pochodzeniem geograficznym i warunkami środowiska np. zasoleniem [1]. Dalsza analiza uzyskanego transkryptomu wykazała potencjalną rolę alternatywnego

splicingu w procesie adaptacji do niskiego zasolenia [4] (publikacja ta nie wchodzi w skład przedstawionej rozprawy doktorskiej).

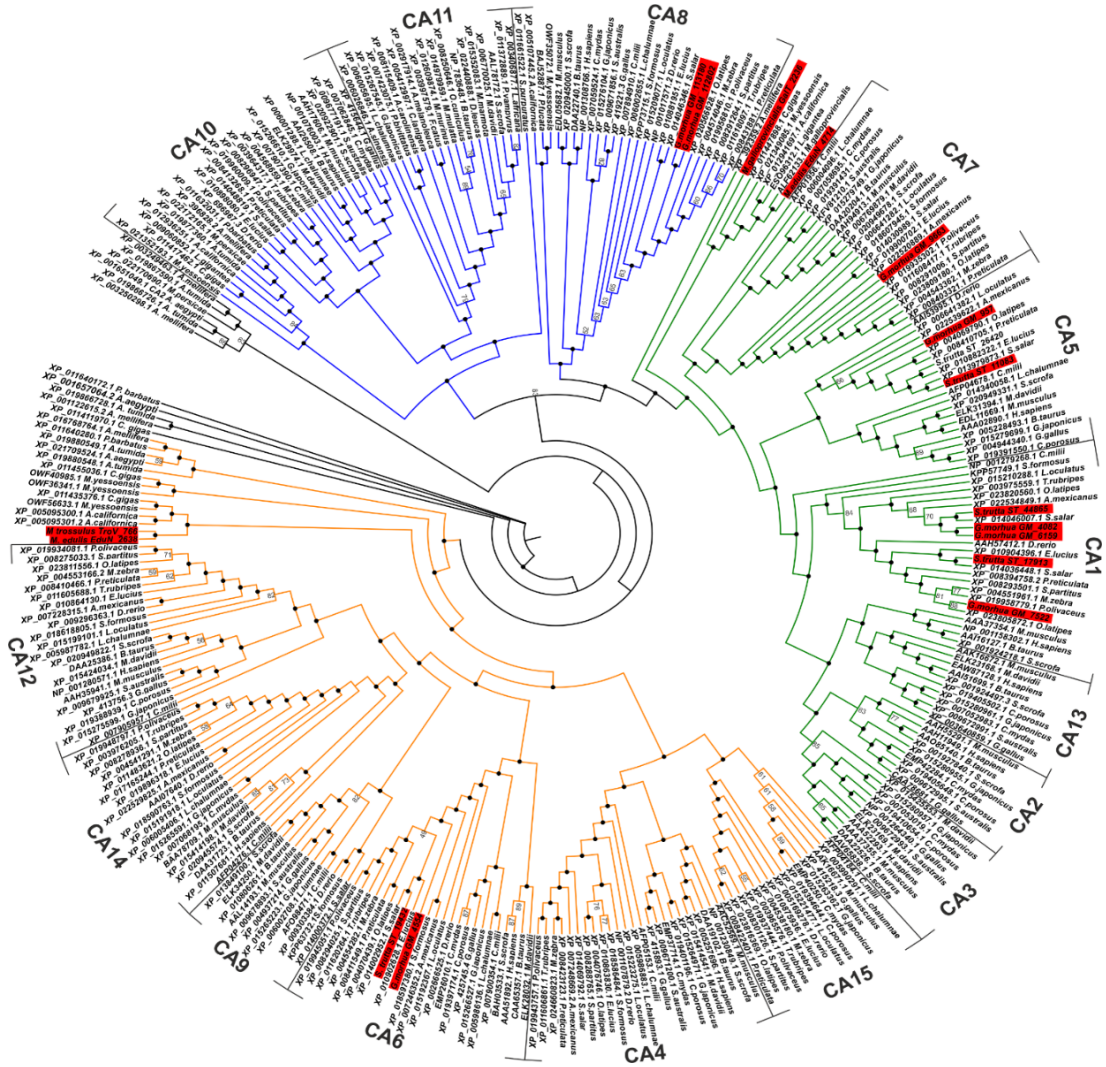
W drugiej części pracy porównałam transkryptom skóry troci wędrownej (*Salmo trutta* m. *trutta*) z innymi tkankami (sekwencje uzyskałam z literatury), wskazując na ekspresję licznych genów potencjalnie zaangażowanych w odpowiedź immunologiczną i wydzielanie śluzu, w tym mucyn [2]. Troć wędrowna jest anadromiczną formą troci (*Salmo trutta*), o dużym znaczeniu gospodarczym i ekologicznym w Europie. Populacja troci wędrownej zmniejszyła się w wyniku budowy licznych przegród i zapór na rzekach, co utrudniło lub uniemożliwiło wędrówkę tarlaków w górę rzek w celu przeprowadzenia rozrodu, a także z powodu zanieczyszczenia i degradacji środowiska oraz chorób. Łososiowate różnią się pod względem podatności na mikroorganizmy ze względu na zróżnicowaną morfologię skóry, która jest pierwszą barierą przeciwko patogenom i wzorce ekspresji genów. W związku z tym, scharakteryzowałam transkryptom skóry troci wędrownej za pomocą NGS. W wyniku sekwencjonowania otrzymano łącznie 1348306 dobrej jakości odczytów, które złożyłam do 75970 kontigów. Spośród nich 48.57% zostało zidentyfikowanych poprzez przeszukiwanie białkowych baz danych m.in. NCBI. Szlaki KEGG oraz ontologie genów zostały przypisane do odpowiednio 13.40% i 34.57% transkryptów, reprezentujących różne procesy i funkcje biologiczne. Wśród zidentyfikowanych kategorii KEGG, najlepiej reprezentowane były transdukcja sygnału (23.28%) i układ immunologiczny (8.82%), z licznymi genami zaangażowanymi w szlaki odpornościowe, implikując różnorodność odpowiedzi immunologicznych w skórze troci wędrownej. Analiza nadreprezentacji GO ujawniła wiele genów zaangażowanych w reakcję stresową i odpowiedź immunologiczną przeciwko patogenom w tkance skórnej troci w porównaniu do innych tkanek. Ponadto 85% homologicznych kontigów wykazywało podobieństwo >95%, a 9.57% stanowiło nowe transkrypty, potencjalnie specyficzne dla tkanki skóry. Dodatkowo zidentyfikowałam i scharakteryzowałam po raz pierwszy u troci 8 rodzajów mucyn – glikoprotein, głównych składników śluzu ryb oraz 140 genów zaangażowanych w ich biosyntezę. Zidentyfikowałam również 1119 potencjalnych polimorfizmów SSR [2].

Trzeci rozdział został poświęcony analizie porównawczej transkryptomów uzyskanych z tkanki płaszcza małży z rodzaju *Mytilus* i identyfikacji genów potencjalnie zaangażowanych w proces biomineralizacji i pigmentacji [3]. Omułki to morskie gatunki małży występujące w wodach przybrzeżnych zarówno półkuli północnej jak i południowej, mające duże znaczenie gospodarze oraz ekologiczne. Omułki są przedmiotem szerokich badań biogeograficznych prowadzonych w Pracowni Genetyki Organizmów Morskich w Instytucie Oceanologii PAN z

zastosowaniem genotypowania SNP. Z uwagi na rozpowszechnienie oraz zróżnicowanie genetyczne, małże są dobrym modelem do badań dotyczących adaptacji do warunków środowiskowych. Ponadto, dzięki zdolnościom filtrującym, pełnią istotną rolę jako naturalne wskaźniki zanieczyszczeń. U mięczaków muszla wydzielana przez tkankę płaszcza podczas procesu biomineralizacji stanowi pierwszą barierę przeciwko drapieżnikom i uszkodzeniom mechanicznym. Zmieniające się warunki środowiskowe, takie jak wzrastający poziom CO_2, z czym związane są zmiany pH wody morskiej, wpływają na wytrzymałość muszli, a tym samym na zdolność ochrony tkanek miękkich ciała. Aby lepiej zrozumieć powyższe zagadnienia scharakteryzowałam i porównałam transkryptomy czterech taksonów małży z różnych lokalizacji geograficznych (*M. edulis* – Morze Północne, *M. galloprovincialis* – Morze Śródziemne oraz Tasmania, *M. chilensis* - Chile, *M. trossulus* - Vancouver), jako reprezentantów czystych taksonów omułków, za pomocą sekwencjonowania nowej generacji. W sumie 743773 sekwencji dobrej jakości zostało złożonych w 20982 kontigi. Spośród nich odpowiednio 46.57%, 37.28% i 17.53% adnotowałam przy użyciu baz NCBI NR, GO i KEGG. Zidentyfikowałam 1292 kontigi potencjalnie związane z biomineralizacją i melanogenezą oraz wybrałam i przeanalizowałam sekwencje genów związanych z odpowiedzią na zakwaszanie wód, takich jak anhydraza węglanowa, chitynaza i tyrozynaza. Zidentyfikowałam potencjalne polimorfizmy SSR (483) i SNP (1497), które zapewniają zasoby do dalszych badań nad markerami molekularnymi w genach biomineralizacji, wiążąc je z różnymi warunkami środowiska i taksonami. W celu porównania uzyskanych transkryptomów przeprowadziłam analizę wzbogacania GO (Blast2GO, test Fishera) oraz zidentyfikowałam ortologi (program OrthoMCL) z wykorzystaniem dostępnych baz danych. Analiza nadreprezentacji GO wykazała reakcję stresową na temperaturę i pH u *M. edulis* z Morza Północnego i *M. galloprovincialis* z Morza Śródziemnego. Analiza filogenetyczna między sześcioma gatunkami *Mytilus* przy użyciu danych transkryptomicznych wykazała, że taksony *M. californianus* i *M. coruscus*, są genetycznie bardziej odległe (tworzą oddzielny klad) od innych taksonów: *M. edulis*, *M. chilensis*, *M. galloprovincialis* i *M. trossulus* [3].

Transkryptomika może być wykorzystywana do scharakteryzowania ewolucyjnych zależności wybranych genów między gatunkami za pomocą analizy filogenetycznej. W tym celu, uzyskane z prezentowanej rozprawy dane transkryptomiczne komercyjnie istotnych gatunków morskich (dorsz bałtycki, troć wędrowna i małże), wykorzystałam do analizy anhydraz węglanowych (CAs, ang. carbonic anhydrases). Nadrodzina CA, złożona z trzech głównych rodzin: α-CA (wielokomórkowce, bakterie, grzyby), β-CA (rośliny) oraz γ-CA (archeony i bakterie) zaliczana jest do grupy metaloenzymów, katalizujących odwracalną

reakcję hydratacji dwutlenku węgla (CO_2), tworząc jon wodorowęglanowy i proton ($CO_2 + H_2O \longleftrightarrow H_2CO_3 \longleftrightarrow HCO_3^- + H^+$). U ssaków rodzina α-CA składa się 13 enzymatycznie aktywnych izoform i 3 białek niewykazujących aktywności enzymatycznej (CARPs, ang. CA-related proteins). CA odgrywa kluczową rolę w regulacji jonów, wydalaniu CO_2, utrzymywaniu równowagi kwasowo-zasadowej i biomineralizacji u ryb i skorupiaków. Charakterystyka izoform CA z uwzględnieniem międzygatunkowej różnorodności i ekspresji w poszczególnych tkankach u organizmów morskich jest istotna ze względu na rosnącą emisję atmosferycznego dwutlenku węgla, oraz związanego z tym zakwaszania oceanów i globalnego ocieplenia. W celu zrozumienia ewolucji i porównania α-CA u tkankowców właściwych (Eumetazoa) porównałam kontigi zidentyfikowane jako CA uzyskane w prezentowanych pracach oraz białka dostępne w bazie NCBI (Rysunek 2).



Rysunek 2. Drzewo filogenetyczne anhydraz węglanowych z rodziny α (α-CA) u różnych gatunków zwierząt. Schemat kolorów: czerwony - sekwencje uzyskane z prezentowanych prac; niebieski: CARPs; zielony: wewnątrzkomórkowe CA; pomarańczowy: zewnątrzkomórkowe CA.

Niekompletne, pozbawione domen i/lub wysoce dywergentne sekwencje CA, zostały wykluczone z analizy. Sekwencje homologiczne (w sumie 319 białek) przyrównano za pomocą programu do konstrukcji alignmentów wielu sekwencji - MUSCLE i przycięto przy użyciu TrimAl (wartość progowa 50%). Do konstrukcji drzewa filogenetycznego użyłam programu IQ-TREE - model substytucji aminokwasów Whelan-and-Goldman (WAG + R9). Sekwencje CA uzyskane z prezentowanych transkryptomów zostały zaklasyfikowane głównie jako wewnątrzkomórkowe CA (ang. intracellular CA), CARP oraz wydzielnicze CA6 (Rysunek 2). Analiza filogenetyczna ujawniła podział na dwa główne klastry: wewnątrzkomórkowe CA (CA1, CA2, CA3, CA5, CA7, CA13) oraz pozakomórkowe (ang. extracellular) CA (CA4, CA6, CA9, CA12, CA14, CA15) (Rysunek 2). Uzyskane wyniki sugerują, że u ryb i małży nie występują izoformy CA3, CA11 i CA13, jednakże może być to spowodowane małą bazą sekwencji CA. Ponadto, CARP wykazały większe podobieństwo do wewnątrzkomórkowego CA i wydaje się, że są dobrze zakonserwowane między gatunkami, co sugeruje, że mogą odgrywać ważną rolę biologiczną.

Oryginalne artykuły, które stały się podstawą niniejszej rozprawy doktorskiej, poszerzyły zakres wiedzy na temat różnorodności genetycznej oraz bioróżnorodności kilku ważnych gatunków morskich, w tym dorsza z Morza Bałtyckiego, troci wędrownej oraz małży z rodzaju *Mytilus*. Uzyskane wyniki rozbudowują bazę sekwencji dla ww. gatunków oraz dostarczają cennych zasobów do dalszych badań ewolucyjnych, genomowych i filogenetycznych. Wyniki zaprezentowane w mojej rozprawie doktorskiej pokazują, że analiza porównawcza transkryptomów blisko spokrewnionych gatunków, z różnej lokalizacji geograficznej oraz między tkankami przy użyciu sekwencjonowania nowej generacji jest efektywną i skuteczną metodą na odkrycie zmienności genetycznej i jej znaczenia funkcjonalnego oraz daje możliwość zrozumienia jak organizmy adaptują się do różnych warunków środowiska. Jest to użyteczne w zarządzaniu ochroną gatunkową. Główną implikacją przedstawionych badań jest utworzenie transkryptomicznej bazy danych dla kluczowych organizmów morskich. Bardzo ważne w analizach wyników sekwencjonowania jest upublicznienie uzyskanych danych. Sekwencje wygenerowane w ramach prezentowanych publikacji zostały zdeponowane w bazach NCBI: Sequence Read Archive (SRA) numery akcesyjne: PRJNA273805 [1], PRJNA323793 [2] i PRJNA419475 [3] oraz GenBank: KY328727-KY328740 [2], MG827120–MG827134 [3]. Jako przykład przydatności i wszechstronnego zastosowania utworzonych transkryptomów, na podstawie uzyskanych sekwencji przeprowadziłam analizę filogenetyczną anhydraz węglanowych u zwierząt. Przedstawione publikacje powstały w efekcie realizacji grantów uzyskanych przez prof. dr

hab. R. Wenne z Narodowego Centrum Nauki, numery: 2011/01/M/NZ9/07207 i 2011/01/B/NZ9/04352 oraz Ministerstwa Nauki i Szkolnictwa Wyższego, numer: 397/N-cGRASP/2009/0.

Literatura

1. **Malachowicz M**, Kijewska A, Wenne R (2015) Transcriptome analysis of gill tissue of Atlantic cod (*Gadus morhua* L.) from the Baltic Sea. Marine Genomics, 23: 37-40. doi: 10.1016/j.margen.2015.04.005
2. **Malachowicz M**, Wenne R, Burzynski A (2017) *De novo* assembly of the sea trout (*Salmo trutta* m. *trutta*) skin transcriptome to identify putative genes involved in the immune response and epidermal mucus secretion. PLoS One, 12(2): e0172282. doi: 10.1371/journal.pone.0172282
3. **Malachowicz M**, Wenne R (2019) Mantle transcriptome sequencing of *Mytilus* spp. and identification of putative biomineralization genes. PeerJ, 6:e6245. doi: 10.7717/peerj.6245
4. Kijewska A, **Malachowicz M**, Wenne R (2018) Alternatively spliced variants in Atlantic cod (*Gadus morhua*) support response to variable salinity environment. Scientific Reports, 8: 11607. doi: 10.1038/s41598-018-29723-w (publikacja ta nie wchodzi w skład przedstawionej rozprawy doktorskiej)