Marcin Jąkalski

Wydział Biologii

Uniwersytetu im. Adama Mickiewicza w Poznaniu

Kierunek studiów: BIOLOGIA. Nr albumu: 291110

Specjalność: BIOINFORMATYKA

Analiza porównawcza nakładających się genów u dwunastu gatunków *Drosophila.*

Comparative analysis of overlapping genes in 12 *Drosophila* species.

Praca magisterska
wykonana na Wydziale Medycznym
Westfalskiego Uniwersytetu Wilhelma
w Münster
pod kierunkiem prof. dr Wojciecha Makałowskiego

Poznań 2009

Składam serdeczne podziękowania
Pani dr Izabeli Makałowskiej
za opiekę naukową, wszelką pomoc
oraz cenne uwagi merytoryczne
w czasie realizacji pracy.

Bardzo dziękuję
mgr Joannie Ciomborowskiej
za życzliwość oraz
wszechstronną pomoc
i cenne uwagi.

Dziękuję
wszystkim koleżankom i kolegom
ze studiów, a w szczególności
Andrzejowi Zielezińskiemu
za pomoc i owocną współpracę
w całym okresie studiów.

SPIS TREŚCI

1. WSTĘP.....7
1.1. Charakterystyka genów nakładających się.....7
1.2. Hipotezy na temat ewolucji genów nakładających się.....8
1.2.1. Ewolucja genów nakładających się na drodze *overprintingu* 8
1.2.2. Ewolucja genów nakładających się na drodze translokacji i adopcji sygnału 9
1.2.3. Powstanie nakładających się genów wyciekiem transkrypcyjnym? 11
1.3. Muszka owocowa jako organizm modelowy.....12
1.3.1. Genomy 12 gatunków muszek rodzaju *Drosophila* 13
1.4. Opublikowane analizy nakładania się genów u *Drosophila melanogaster* 14
1.5. Bioinformatyka w badaniach nad nakładającymi się genami.....15
1.5.1. EVOG.....15
1.5.2. OGtree 15
1.5.3. BPhyOG 16
2. CEL PRACY 17
3. MATERIAŁY, NARZĘDZIA I METODY 18
3.1. Materiały 18
3.2. Narzędzia.....19
3.2.1. UCSC Genome Browser.....19
3.2.2. GALAXY 20
3.2.2.1. Format BED.....22
3.2.3. BLAST 23
3.2.3.1. *Reciprocal* BLAST 24
3.2.4. PYTHON.....24
3.3. Metody 25
3.3.1. Identyfikacja genów nakładających się.....25
3.3.2. Identyfikacja ortologów u 11 gatunków *Drosophila*.....27
3.3.3. Identyfikacja ortologów pośród genów 6 organizmów modelowych 29
4. WYNIKI.....32
4.1. Geny nakładające się zidentyfikowane w genomie *D. melanogaster*.....32
4.2. Geny nakładające się u pozostałych 11 gatunków rodzaju *Drosophila*.....33
4.3. Zidentyfikowane geny nakładające się u 6 organizmów modelowych.....36

4.4. Geny nakładające się wspólne dla *D. melanogaster* i pozostałych analizowanych organizmów.....38
5. DYSKUSJA 40
6. STRESZCZENIE.....46
7. SUMMARY 49
8. BIBLIOGRAFIA.....51
9. SPIS ILUSTRACJI.....55
10. SPIS TABEL.....56

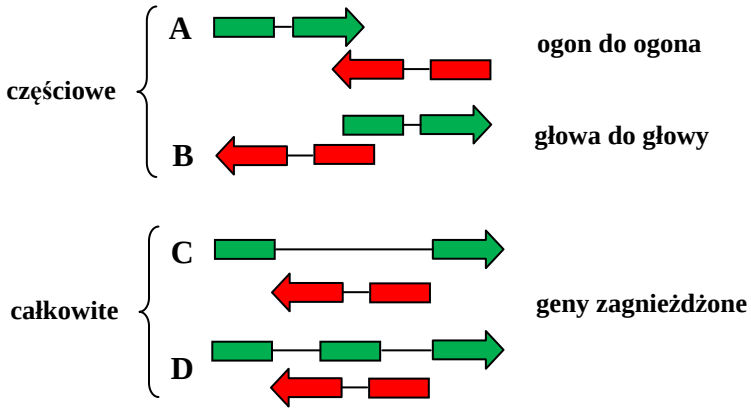
1. WSTĘP

1.1. Charakterystyka genów nakładających się

Tomasz Morgan w 1911 roku badając geny muszki owocowej (*Drosophila melanogaster*) doszedł do wniosku, że są one ułożone na chromosomie liniowo jak koraliki w naszyjniku [1]. Opis ten musiał jednak zostać poprawiony kiedy zsekwencjonowano pierwszy genom wirusowy - genom bakteriofaga ΦX174 [2]. Podczas analizy zauważono, że oprócz tandemowego ułożenia genów w genomie tym istnieją na przeciwległych niciach dwie pary genów kodujących białka, których translacja w dwóch różnych ramkach odczytu odbywała się na wspólnej sekwencji DNA [3]. Obserwacja ta była pierwszym udokumentowanym przypadkiem występowania zjawiska nakładania się genów.

Nakładające się geny, opisuje się jako pary różnych genów o pokrywających się całkowicie lub do pewnego stopnia sekwencjach kodujących bądź niekodujących. W regionie nakładania egzon lub intron jednego genu zawiera intron lub egzon drugiego genu. Zjawisko to obserwowane jest często w genomach wirusowych i prokariotycznych, jak również w DNA mitochondrialnym. Uważa się, że jest to powszechna strategia organizacji genomu i regulacji genowej u bakterii. W 1986 roku opisano pierwsze przypadki występowania genów nakładających się u organizmów eukariotycznych – myszy i muszki owocowej [4, 5]. Jednakże zjawisko nakładających się genów u wyższych organizmów przez długi jeszcze czas uważano za bardzo rzadkie [6, 7]. Ukończenie sekwencjonowania pierwszych genomów eukariotycznych i analizy całych tranksryptomów wykazały, że posiadają one wysoką liczbę wystąpień nakładających się jednostek transkrypcyjnych [8, 9, 10, 11, 12, 13]. Te antysensowne transkrypty uczestniczą w wielu procesach komórkowych, takich jak imprinting genomowy, inaktywacja chromosomu X, splicing alternatywny, wyciszanie genów i metylacja, redagowanie RNA oraz translacja [14, 15, 16, 17, 18, 19].

Geny nakładające się można podzielić na kilka różnych kategorii w oparciu o kierunek transkrypcji, jak również ze względu na fragmenty sekwencji dzielone pomiędzy nakładającymi się rejonami. Na potrzeby przeprowadzonych w niniejszej pracy analiz przyjęto, że podział genów nakładających się, składa się z trzech głównych typów: głowa do głowy (*head-to-head*), ogon do ogona (*tail-to-tail*) oraz geny zagnieżdżone (*nested*) (Ryc.1).



Ryc. 1. Ogólny schemat sposobu nakładania się genów, skategoryzowany w trzy główne typy: **A.** nakładanie typu ogon do ogona (*tail-to-tail*, nakładanie w regionie końców 3'); **B.** głowa do głowy (*head-to-head*, przy końcach 5' genów); **C, D.** geny zagnieżdżone (*nested*). Kolorowe bloki oznaczają egzony, cienkie linie introny.

1.2. Hipotezy na temat ewolucji genów nakładających się

Kolejne badania wskazują na coraz to nowe przypadki występowania zjawiska nakładania się genów u różnych gatunków. Jednakże ich pochodzenie ewolucyjne nadal pozostaje niejasne. W celu wytłumaczenia sposobu, w jaki doszło do powstania tego fenomenu zaproponowano kilka mechanizmów. Poniżej zaprezentowano trzy główne podejścia do tego tematu.

1.2.1. Ewolucja genów nakładających się na drodze *overprintingu*

W badaniach nad ewolucją molekularną zaproponowano hipotezę wielkiego wybuchu (*big bang*), która zakłada wystąpienie swego rodzaju genetycznej „eksplozji". Miała ona doprowadzić do szeroko rozpowszechnionego występowania spokrewnionych ze sobą cząsteczek biologicznych oraz szlaków biosyntetycznych, które są wspólne dla wszystkich organizmów [20]. Zgodnie więc z tą hipotezą za większość obserwowanej obecnie różnorodności molekularnej odpowiadają procesy takie jak mutacje, duplikacje DNA i jego rearanżacje, tasowanie egzonów, transpozycje itp. [21].

Nie można jednak pominąć faktu powstawania genów *de novo* w ewolucji genomów. Nowe geny mogą być tworzone na dwa różne sposoby. Cząsteczki polinukleotydów są polimeryzowane od nowa [22, 23] bądź też mogą być generowane na drodze translacji przy użyciu wcześniej niewykorzystywanych ramek odczytu lub istniejących kodujących i niekodujących fragmentów genomu. Możliwość utworzenia nowych genów z wcześniej

istniejących sekwencji nukleotydowych została określona mianem *overprintingu* [24]. Nowe geny lub regiony kodujące powstałe w wyniku tego procesu są często wykrywane w nakładających się genach.

Być może najbardziej przejrzystym przykładem powstawania genów *de novo* na drodze *overprintingu* wraz z jednoczesnym utworzeniem nakładającej się pary jest gen α receptora hormonu tyroidowego (*TR*). Należy on do rodziny genów kodujących receptory jądrowe, których ligandy (witamina D, retinoidy, hormony tyroidowe) zawierają steroidy [25]. Gen receptora *TRα* posiada dwie formy podlegające alternatywnemu splicingowi. Pierwsze osiem egzonów jest wspólne dla obu form, natomiast egzon 9-ty jest unikatowy dla formy TRα1, a egzon 10-ty dla *TRα2*. Formy te różnią się sekwencją i długością ich C-końca, który funkcjonuje jako domena wiążąca ligand. Aminokwasy C-końcowe formy *TRα2* kodowane przez nukleotydy tworzące region nakładania z genem receptora tyroidowego (*ear-1*) nie wykazują żadnego znaczącego podobieństwa do domen wiążących ligand pochodzących z innych członków wymienionej rodziny genowej. Natomiast sekwencje genów TRα1 i *ear-1* wykazują wyraźne podobieństwo sekwencji do receptorów innych członków rodziny genowej, dlatego też są oryginalnymi genami. Egzon 10-ty *TRα2* powstał później, *de novo* [26].

Stosując się do zaproponowanej powyżej hipotezy można założyć, że organizmy posiadają dwie klasy genów, które biorą udział z zjawisku nakładania się: stare ewolucyjnie geny, z których większość powstała przed oddzieleniem się organizmów prokariotycznych od eukariotycznych (np. kodujące rRNA, tRNA, białka rybosomalne) oraz geny młode, nowe, które filogenetycznie ograniczone są do organizmów, w których występują i które kodują białka o zróżnicowanych funkcjach dostosowanych do warunków życia danego organizmu. Ta druga klasa genów często zaangażowana jest w zjawisko nakładania się.

1.2.2. Ewolucja genów nakładających się na drodze translokacji i adopcji sygnału

Shintani [27] rozważa dwie drogi, na których teoretycznie mogło dojść do powstania zjawiska nakładania się genów. Po pierwsze, w danym obszarze DNA może zdarzyć się, że więcej niż jedna ramka odczytu ma potencjalną zdolność kodowania białka. Jeśli kodon startowy transkrypcji lub też miejsce inicjacji transkrypcji powstanie na drodze przypadku w obrębie tego odcinka DNA to z wykorzystaniem dodatkowej ramki odczytu, z tego samego *locus* genowego może powstać dwa lub więcej typów mRNA. Ma to miejsce

na przykład u ssaków z udziałem genów kodujących białka XLαs i ALEX [28]. Alternatywnie, dwa powstałe niezależnie geny mogą ulec translokacji, a każdy z nich uzyskać część swojego transkryptu z tego samego lub komplementarnego fragmentu DNA, co poskutkuje wytworzeniem nakładania.

Dla zbadania pochodzenia eukariotycznych nakładających się genów autorzy koncepcji przeanalizowali parę genów *ACAT2* i *TCP1*. Ludzki gen acetylotransferazy 2 acetylo-CoA (*ACAT2*) koduje tiolazę – enzym zaangażowany w metabolizm lipidów. Gen *TCP1* (T-comlpex 1) koduje molekularny chaperon, będący członkiem rodziny białek opiekuńczych [29]. Oba geny nakładają się swoimi niepodlegającymi translacji regionami 3' (*tail-to-tail*). W celu stwierdzenia w jaki sposób nakładanie to mogło powstać w toku ewolucji zidentyfikowano i przebadano homologiczne geny organizmów takich, jak danio pręgowany, afrykańska ropucha szponiasta, kajman, dziobak, opos i walabia [27].

Rezultaty badań wykazały, że przez większość historii ewolucyjnej kręgowców geny *ACAT2* i *TCP1* istniały jako niezależne jednostki. Ich nałożenia, które zaobserwowano w przebadanych na potrzeby przeprowadzonej analizy ssakach, wliczając w to stekowce i torbacze, powstało podczas tranzycji z ssakokształtnych gadów do ssaków. Mogło do tego dojść w dwojaki sposób. Po pierwsze, na drodze rearanżacji, której towarzyszyła utrata części 3' UTR wraz z sygnałem poliadenylacji, dla przykładu z genu *TCP1*. Jednocześnie nowy sąsiad, jakim został gen *ACAT2* zawierał niekodującą nić ze wszystkimi sygnałami niezbędnymi do zakończenia translacji oraz obróbki transkryptu. Dlatego też gen *TCP1* mógł zachować swoją normalną funkcjonalność. Drugi, bardziej prawdopodobny sposób, to taki, że dwa geny zostały sąsiadami na drodze rearanżacji, ale z początku nie nakładały się. W późniejszym czasie jeden z genów utracił swój sygnał poliadenylacji i zaczął wykorzystywać ten obecny w niekodującej nici innego genu i w ten sposób para została utrwalona. Niekodujące nici genów *ACAT2* z danio pręgowanego, ropuchy i kajmana zawierają w rzeczywistości jeden lub więcej potencjalnych, poprawnie zorientowanych i umiejscowionych sygnałów poliadenylacyjnych w regionach 3' UTR, które to mogły zostać wykorzystane przez geny *TCP1*.

Rearanżacja wygenerowała w ten sposób parę, która od momentu powstania dziedziczyła się jako jednostka. Ograniczenie nakładania do pojedynczej filogenetycznej linii, jaką są ssaki sugeruje, że połączenie dwóch genów nastąpiło tylko raz i utrzymuje się od 200 milionów lat. Wniosek ten popiera ponadto obserwacja konserwacji sekwencji w regionie 3' UTR genów *ACAT2* i *TCP1*. U ssaków ich podobieństwo jest stosunkowo

niskie w nakładającej się części, która oskrzydla region genu *ACAT2*. Jednakże w części otaczającej region genu *TCP1*, który ulega translacji, podobieństwo to jest uderzające – cały blok sekwencji uległ konserwacji w ewolucji stekowców, torbaczy i łożyskowców z ich wspólnego przodka.

Powód, dla którego wymienione geny pozostały sprzężone ze sobą przez około 200 milionów lat pozostaje niejasny. Nawet jeśli w niektórych gatunkach nastąpiła drugorzędna separacja tych genów, prawdopodobnie u większości ssaków pozostały one razem. Nie istnieje dowód na to, że geny *ACAT2* i *TCP1* są w jakikolwiek sposób spokrewnione ze sobą ewolucyjnie, strukturalnie, czy funkcjonalnie. Trwałość nałożonych genów może być konsekwencją konserwatywnej natury procesów ewolucyjnych albo utrudnień związanych z separacją [27].

1.2.3. Powstanie nakładających się genów wyciekiem transkrypcyjnym?

W większości przypadków nałożeń pomiędzy dwoma genami kodującymi białka zjawisko to ograniczone jest do ich regionów niepodlegających translacji (UTR) [30]. Dodatkowo w wielu przypadkach w nakładanie zaangażowana jest alternatywna poliadenylacja, co stwarza kilka wariantów transkryptu, które różnią się w swojej długości końca 3'. Zasugerowano, że przynajmniej połowa z wszystkich ludzkich genów koduje wiele transkryptów z alternatywnym końcem 3' [31]. Nie ustalono jednakże, czy ta alternatywna obróbka jest nieprzypadkowa i prowadzi do kontrolowanego nakładania pomiędzy transkryptami, czy też stanowi „wyciek" z transkrypcyjnej maszynerii RNA. Niepowodzenie mechanizmów transkrypcji w rozpoznaniu poprawnego miejsca poliadenylacji (na przykład z powodu mutacji) może w rzeczywistości doprowadzić do czytania genów w złym kierunku [32]. Dodatkowo kiedy kilka blisko położonych regionów poli-A znajduje się w tym samym transkrypcie konkurują one ze sobą o poliadenylację [33]. Takie ogony poli-A mogą z łatwością zostać dodane w ewolucji, czego wynikiem jest powstanie przeciwnie zorientowanych genów nakładających się ze sobą.

Dahary i współautorzy na podstawie porównania genomów człowieka i rozdymki (*Fugu rubripes*) stwierdzili, że w toku ewolucji antysensowne pary genów zachowywane są znacząco wyraźniej niż pary nie-antysensowne. Pociąga to za sobą wnioski, że nakładanie genów u człowieka mogło ulec konserwacji w toku ewolucji kręgowców. Stąd też takie nałożenia są raczej prawdziwe, a nie są po prostu swego rodzaju „wyciekiem"

transkrypcyjnym. Separacja genów z pary dotknęłaby oba z nich (w przeciwieństwie do neutralnego efektu separacji pomiędzy nienakładającymi się genami) i stąd też obserwowana jest negatywna selekcja [30].

1.3. Muszka owocowa jako organizm modelowy

Muszka owocowa, *Drosophila melanogaster*, jest niewielkim owadem, zaklasyfikowanym do rzędu muchówek (*Diptera*). Wielkość jej ciała waha się od 2 do 3 mm [34]. Zaliczana jest do ważnych bezkręgowych organizmów modelowych, a badania nad nią trwają już ponad 100 lat. *D. melanogaster* jest złożonym wielokomórkowym organizmem z wieloma cechami rozwojowymi i behawioralnymi, które są wspólne z innymi gatunkami, w tym także z człowiekiem. Dzięki temu prowadzone na niej badania poczyniły znaczący wkład w zrozumienie fundamentalnych procesów biologicznych.

Hodowla muszki owocowej wymaga niewielkiego wkładu i wysiłku - nawet w przypadku wykorzystywania większych ilości osobników nie ma zapotrzebowania na dużą przestrzeń hodowlaną. Jej morfologia jest łatwa do identyfikacji. Czas wytwarzania nowego pokolenia jest krótki, około 10 dni w temperaturze pokojowej, stąd w ciągu zaledwie paru tygodni można badać kilka pokoleń. *Drosophilę* cechuje wysoka płodność – samice mogą złożyć ponad 800 jajeczek w ciągu swojego życia. Samce i samice są łatwo rozróżnialne, a niezapłodnione osobniki można z łatwością izolować w celu przeprowadzenia genetycznych krzyżówek. Dorosłe larwy posiadają ogromne chromosomy w gruczołach ślinowych zwane chromosomami politenicznymi. Dające się z łatwością zaobserwować nabrzmienia (*puffs*) stanowią regiony o intensywnej przebiegającej transkrypcji. Samce nie wykazują rekombinacji mejotycznych [35].

Haploidalny genom muszki zawiera ok. 131-165 milionów par nukleotydów, a liczba występujących genów kodujących białka to ponad 13 tysięcy. *Drosophila* posiada jedynie 4 pary chromosomów - 3 pary autosomów i 1 parę chromosomów płciowych: zestaw XY dla samców i XX dla samic [34].

Drosophila melanogaster jest jednym z pierwszych eukariotycznych organizmów, których DNA całkowicie zsekwencjonowano. Genom muszki owocowej został opublikowany w 2000 roku [36].

1.3.1. Genomy 12 gatunków muszek rodzaju *Drosophila*

Gatunki z rodzaju *Drosophila* różnią się istotnie pod względem morfologii, ekologii i schematów zachowania [37]. Dwanaście zsekwencjonowanych gatunków pochodzi z Afryki, Azji, Ameryki Północnej i Południowej oraz z wysp Australii i Oceanii. Wśród nich są gatunki kosmopolityczne (*D. melanogaster, D. simulans*), jak i te zamieszkujące pojedyncze wyspy (*D. sechellia*) [38]. Reprezentują one mnogość strategii behawioralnych, od generalistów pokarmowych, jak *D. ananassae* do wyspecjalizowanych pokarmowo, takich jak *D. sechellia*, która żywi się wyłącznie owocami jednej rośliny.

Pomimo tej różnorodności fenotypowej gatunki z rodzaju *Drosophila* współdzielą charakterystyczny plan budowy i plan cyklu życiowego. Chociaż do tej pory ekstensywnie przebadano tylko *D. melanogaster*, wydaje się, że większość głównych aspektów komórkowych, molekularnych i rozwojowych jest zakonserwowana wśród tych gatunków. Dlatego też, oprócz badań nad powiązaniami pomiędzy sekwencjami, a różnorodnością fenotypową, genomy tych gatunków stanowią również doskonały model do analiz ewolucyjnych [39].

Kompletny genom wszystkich przebadanych muszek cechuje się podobną długością oraz liczbą genów – w przybliżeniu 14 tysięcy. Większość z genów kodujących białka u *D. melanogaster* jest zakonserwowana u pozostałych zsekwencjonowanych gatunków: 77% genów ma możliwe do zidentyfikowania homologi we wszystkich pozostałych genomach, 62% może być zidentyfikowane jako unikatowe ortologi w 6 genomach grupy *melanogaster*, a 49% jako unikatowe ortologi we wszystkich 12 genomach [39].

Jak pokazały badania najszybciej zmianom podlega ta frakcja genów muszek, która związana jest z ich rozrodem, odpornością na patogeny oraz zmysłami węchu i smaku. *D. sechellia*, która odżywia się wyłącznie jednym rodzajem pokarmu w toku ewolucji gubiła receptory smaku kilkakrotnie szybciej od innych gatunków. *D. willistoni* utraciła gen odpowiedzialny za syntezę selenoproteiny, który obecny jest u pozostałych gatunków [39].

Pierwszymi opublikowanymi genomami muszek z rodzaju *Drosophila* były genomy *D. melanogaster* [36] i *D. pseudoobscura* [40]. W 2007 roku Konsorcjum 12 genomów *Drosophili* (*Drosophila 12 Genomes Consortium*) opublikowało genomy 10 innych – *D. sechellia, D. simulans, D. yakuba, D. erecta, D. ananassae, D. persimilis, D. willistoni, D. mojavensis, D. virilis* oraz *D. grimshawi* wraz z analizą porównawczą wszystkich 12 gatunków [39].

1.4. Opublikowane analizy nakładania się genów u *Drosophila melanogaster*

Zagadnienie nakładania się genów u *Drosophila melanogaster* jest mało poznane. Jak dotąd wydano jedynie kilka artykułów na ten temat. Pierwszy opisany przez S. Henikoffa przypadek wystąpienia pary nakładających się genów opublikowano w 1986 roku [41]. Dotyczył on zagnieżdżenia genu kodującego białko kutykuli poczwarki (*Pcp*) wewnątrz intronu genu białka zaangażowanego w szlak biosyntezy puryn. W tym samym roku C.A. Spencer i współautorzy opisali przypadek nałożenia genu kodującego enzym DOPA-dekarboksylazy (*Dcd*) z genem o nieznanej funkcji [5]. Znajdują się one w orientacji ogon do ogona i współdzielą region genomowy o długości 88 par zasad. Na podstawie badań ekspresji obu genów oraz czasowej i przestrzennej dystrybucji ich transkryptów w komórkach autorzy wnioskowali o możliwości interferencji transkrypcyjnej wraz z towarzyszącymi jej implikacjami regulatorowymi jako skutek nałożenia wspomnianych genów.

Inny artykuł, opublikowany w 1989 roku przez R.A. Schulza i B.A. Butlera dotyczył organizacji klastera genowego *z600-gdl-Eip28/29* [42]. Autorzy stwierdzili w swoich badaniach, że wspomniane nakładające się geny mogą być przedmiotem transkrypcyjnej interferencji - negatywny wpływ na transkrypcję w kierunku od promotora. Drugą i całkiem odmienną konsekwencją organizacji nakładających się genów może być wykorzystanie wspólnych elementów regulatorowych, tak by współregulować ekspresję dwóch genów. Trzecią możliwością jest to, że geny w klasterze mogą ulegać ekspresji wyraźnie z powodu kombinacji funkcji wielu elementów regulatorowych zlokalizowanych w regionie genów nakładających się [42].

G. Pápai w swojej publikacji z 2002 roku opisał badania nad dwoma nakładającymi się genami *D. melanogaster*: *dada2a/drpb4* i *dtat* (w orientacji głowa do głowy) [43]. Obserwowane nakładanie jest tu rozpatrywane jako prawdopodobnie zaangażowanie w zjawisko wyciszania genów, interferencję RNA (RNAi) w regulacji transkrypcyjnej tych genów.

1.5. Bioinformatyka w badaniach nad nakładającymi się genami

Bioinformatyka stanowi zastosowanie technologii informatycznej w dziedzinie biologii, zwłaszcza molekularnej. Obejmuje zarówno tworzenie i rozwój baz danych, algorytmów, technik obliczeniowych i statystycznych, jak również zarządzanie i analizy danych biologicznych. Zadania, z jakimi mają do czynienia bioinformatycy, to między innymi mapowanie i analizowanie sekwencji DNA i białek, dopasowywanie różnych sekwencji DNA i/lub białkowych, przewidywanie i tworzenie trójwymiarowych modeli struktur białkowych.

Bioinformatyka znajduje także swoje zastosowanie w badaniach nad genami nakładającymi się i ich ewolucją. Poniżej przedstawiono przykłady trzech narzędzi bioinformatycznych wykorzystywanych w tego typu analizach.

1.5.1. EVOG

EVOG (*Evolution Visualizer for Overlapping Genes*), to internetowa baza danych do ewolucyjnych analiz genów nakładających się wraz z interfejsem do wizualizacji [44]. Zawiera zbiory nakładających się genów człowieka, szympansa, orangutana, rezusa, krowy, konia, dziobaka, psa, kota, myszy, kurczaka, szczura, danio pręgowanego, oposa oraz żaby szponiastej. Korzystając z tej bazy można w prosty sposób uzyskać zestawy nakładających się genów, które wspólne są dla kilku z wyżej wymienionych genomów. EVOG jest bardzo łatwa w użyciu w analizach porównawczych mających na celu zbadanie procesu ewolucji genów nakładających się. Wspomniany interfejs do wizualizacji służy do konfiguracji i prezentacji na chromosomach wyniku analizy porównawczej nakładających się genów. EVOG został zaimplementowany w języku JAVA i dostępny jest pod adresem http://neobio.cs.pusan.ac.kr/evog/. Niestety w jednej z ostatnich wersji usunięto zestaw danych dla genomu *Drosophili melanogaster*.

1.5.2. OGtree

OGtree to narzędzie do konstruowania drzew genomowych dla gatunków prokariotycznych w oparciu o pomiary zawartości nakładających się genów i ich uporządkowanie w genomie [45]. Geny nakładające się są wszechobecne w genomach bakteryjnych i o wiele bardziej zakonserwowane pomiędzy różnymi ich gatunkami niż geny

nienakładające się. W oparciu o te właściwości wykorzystano geny nakładające się jako cechy filogenetyczne, które mogą posłużyć do rekonstruowania ewolucyjnego powiązania wśród genomów bakteryjnych.

Na podstawie wprowadzonych numerów akcesyjnych genomów prokariotycznych OGtree pobiera ich kompletną sekwencję z bazy NCBI i identyfikuje geny nakładające się w każdym z genomów i ich ortologiczne pary w innych genomach. Następnie oblicza dystanse ewolucyjne pomiędzy każdą parą z wejściowych genomów na podstawie kombinacji ich zawartości genów nakładających się i kolejności ortologicznych par. Na koniec wykorzystuje opartą na dystansach metodę konstruowania drzew do stworzenia drzew genomowych dla wejściowych genomów prokariotycznych zgodnie z ich dystansami obliczonymi dla ortologicznych par genów nakładających się.

OGtree dostępny jest pod adresem http://bioalgorithm.life.nctu.edu.tw/OGtree/.

1.5.3. BPhyOG

BPhyOG to internetowy interaktywny serwer do rekonstruowania filogenii genomów bakteryjnych w oparciu o współdzielone przez nie nakładające się geny [46]. Do budowania drzew wykorzystuje dwie metody: *Neigbor Joining* oraz UPGMA (*Unweighted Pair-Group Method using Arithmetic averages*). Użytkownicy mogą zastosować wybraną metodę do wygenerowania drzew filogenetycznych, które oparte są na ewolucyjnej matrycy dystansów dla wybranego genomu. Dystans pomiędzy dwoma genomami określony jest przez znormalizowaną liczbę wspólnych dla nich nakładających się par genów. ByPhyOG pozwala użytkownikowi przeglądać takie pary, które zostały wykorzystane do wnioskowania powiązań filogenetycznych. Dostarcza także dokładną adnotację dla każdej pary nakładających się genów oraz cechy pojedynczych genów.

ByPhyOG zawiera obecnie 177 kompletnych genomów bakteryjnych z 79,855 parami nakładających się genów i ich adnotacje. Dostępny jest za darmo pod adresem http://cmb.bnu.edu.cn/BPhyOG.

2. CEL PRACY

Celem niniejszej pracy była analiza genomu muszki owocowej (*Drosophila melanogaster*) pod kątem występujących w nim genów nakładających się, zbadanie ich ewolucyjnego pochodzenia oraz przetestowanie, na przykładzie rodzaju *Drosophila*, istniejących hipotez ewolucji nakładania pomiędzy genami.

Etap pierwszy analizy miał za zadanie wyłonienie spośród genomu *D. melanogaster* par nakładających się genów z wykorzystaniem narzędzia GALAXY.

Etap drugi opierał się na zidentyfikowaniu znalezionych w etapie pierwszym genów wśród pozostałych gatunków rodzaju *Drosophila* (*D. pseudoobscura, D. simulans, D. sechellia, D. yakuba, D. erecta, D. ananassae, D. persimilis, D. willistoni, D. virilis, D. mojavensis, D. grimshawi*). Ideą takiego działania było zbadanie stopnia konserwacji nakładających się genów u bardzo blisko spokrewnionych gatunków i jednocześnie sprawdzenie hipotez Keese'a i Gibbsa o powstawaniu nakładających się genów na drodze zjawiska overprintingu oraz Shintani'ego o translokacjach genomowych prowadzących do tworzenia nakładających się par (patrz Wstęp).

Etap trzeci projektu polegał na znalezieniu pierwotnych nałożeń genów, które byłyby wspólne dla owadów i kręgowców na podstawie porównania genów znalezionych w etapie pierwszym z genami sześciu organizmów modelowych (*Annopheles gambiae, Apis mellifera, Danio rerio, Gallus gallus, Homo sapiens, Mus musculus*). Wynik miał posłużyć do potwierdzenia lub obalenia hipotezy autorstwa Dahary'ego (patrz Wstęp) o starym ewolucyjnie pochodzeniu genów nakładających się.

3. MATERIAŁY, NARZĘDZIA I METODY

3.1. Materiały

Wyjściowy zestaw danych do analiz to adnotacje genomu *Drosophila melanogaster* z bazy danych FlyBase. Na ich podstawie z użyciem oprogramowania GALAXY zidentyfikowano w pierwszym etapie pracy kolekcję par nakładających się genów, dla których pobrano sekwencje kodowanych przez nie białek z wykorzystaniem UCSC Genome Browser.

Dane do etapu drugiego pochodziły ze zbiorów bazy danych FlyBase (http://flybase.org/static_pages/downloads/bulkdata7.html). Wykorzystano dwa zbiory pochodzących z tej bazy. Pierwszy (gene_summaries.tsv) zawierał podstawowe informacje o genach przechowywanych w bazie (identyfikatory genu, lokalizacja na mapie genowej, funkcja molekularna, proces biologiczny, liczba alleli i ich fenotypy, liczba transkryptów i białek). Dane pochodziły z dnia 5 listopada 2008. Drugi zbiór (gene_orthologs_fb_2009_02.tsv) to zestaw genów muszki owocowej wraz z podanymi dla nich ortologami z innych organizmów i ich lokalizacją genomową. Zbiór zawiera dane z lutego 2009 roku. Na jego podstawie utworzone zostały osobne zbiory zawierające ortologi genów *D. melanogaster* pochodzące z pozostałych 11 gatunków muszek z rodzaju *Drosophila*.

Zestawy sekwencji białkowych sześciu organizmów modelowych do trzeciego etapu analiz pobrano przy użyciu bazy danych NCBI Taxonomy Browser (http://www.ncbi.nlm.nih.gov/Taxonomy/). Wielkości poszczególnych zestawów sekwencji to: *Annopheles gambiae* – 31,940, *Apis mellifera* – 10,008, *Danio rerio* – 66,660, *Gallus gallus* – 32,556, *Homo sapiens* – 473,348, *Mus musculus* – 249,113 białek. Zbiór 21,243 transkryptów *D. melanogaster* pobrano przy użyciu UCSC Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables)

Dodatkowy zbiór, gene2accesion (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/), został wykorzystany w celu uzyskania unikalnych identyfikatorów genów (*GeneID*) na podstawie identyfikatorów białek (*protein GI*) analizowanych organizmów. Plik ten stanowił kompleksowy raport numerów dostępu powiązanych z identyfikatorami genowymi, których źródłem były bazy RefSeq oraz Swiss-Prot.

3.2. Narzędzia

Analizy wykonano wykorzystując ogólnie dostępne i znane oprogramowanie bioinformatyczne, jak i liczne programy własne stworzone przy użyciu języka skryptowego Python (http://www.python.org/). Opis zastosowanych narzędzi znajduje się poniżej.

3.2.1. UCSC Genome Browser

Przeglądarka genomowa UCSC Genome Browser zapewnia szybki wgląd w dowolną część genomu wraz z dużą ilością adnotacji takich jak znane geny, geny z procesu przewidywania, EST, mRNA, wyspy CpG, informacje z procesu składania (*assembling*) sekwencji, geny homologiczne i wiele innych [47]. Połowa z tych adnotacji jest generowana w UCSC z publicznie dostępnych danych sekwencyjnych. Pozostałe pochodzą od współpracowników z całego świata. Użytkownicy mogą także dodawać swoje własne informacje na potrzeby własnych projektów naukowych.

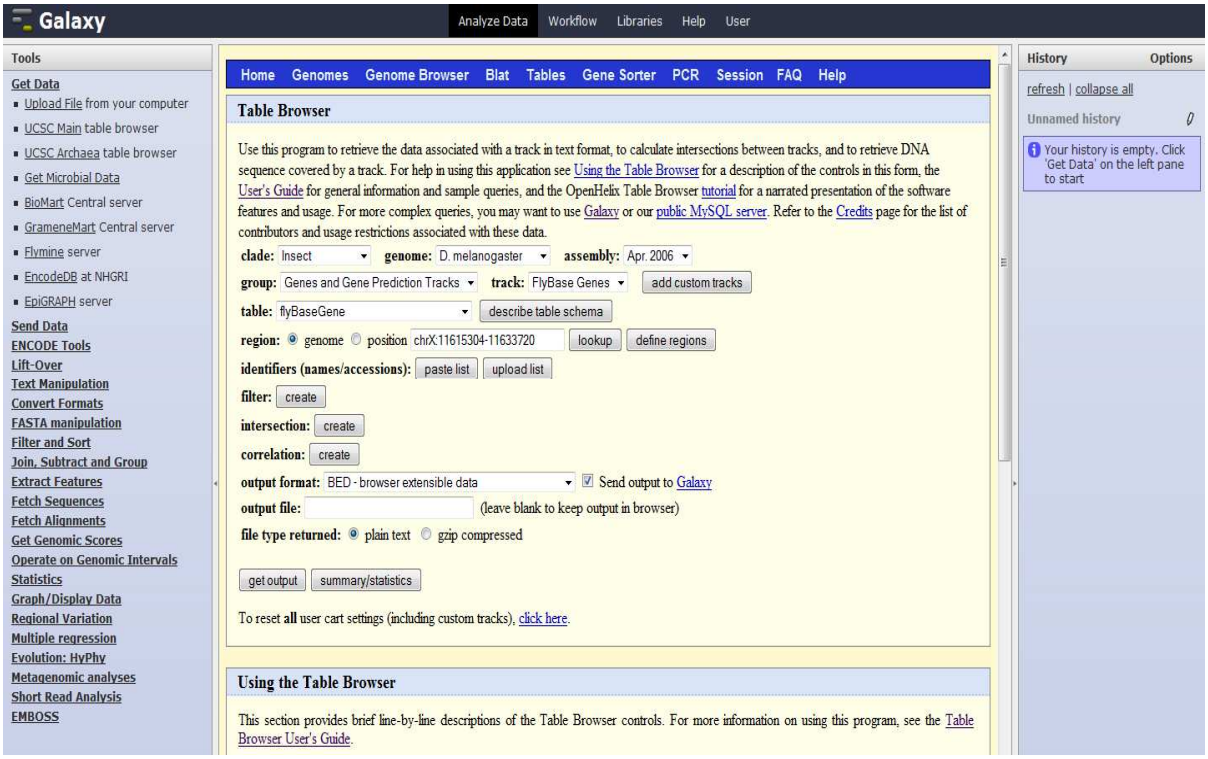
Genome Browser przechowuje adnotacje do sekwencji w powiązaniu z ich koordynatami przez co pozwala na błyskawiczną wizualizację różnych typów informacji. Użytkownik ma wgląd w cały chromosom, może powiększać jego wybrany fragment w celu na przykład obejrzenia zmapowanych EST i zbadania możliwości alternatywnego splicingu. Przeglądarka integruje wszystkie dostępne informacje w jednym miejscu dając użytkownikowi duże możliwości ich eksploracji.

Przeglądarka Genome Browser wspiera procesy przeszukiwania oparte na tekście oraz na sekwencjach zapewniając szybki i precyzyjny dostęp do dowolnego region zainteresowania. Drugorzędne linki z poszczególnych rekordów w obrębie adnotacji prowadzą do szczegółowych informacji odnoszących się do danej sekwencji oraz do dodatkowych baz danych poza przeglądarką, takich jak PubMed, GenBank, Entrez, i OMIM [46].

UCSC Genome Browser dostępny jest pod adresem http://genome.ucsc.edu/cgi-bin/hgGateway.

3.2.2. GALAXY

GALAXY to platforma do interaktywnych analiz genomowych na dużą skalę [48]. Jest to system służący do integracji sekwencji genomowych, ich dopasowań (*alignments*) i adnotacji funkcjonalnych. Galaxy nie jest przeglądarką. Zamiast tego pozwala użytkownikowi na zbieranie i manipulowanie danymi z istniejących już zasobów (np. z UCSC Genome Browser). Każda akcja użytkownika jest zapisywana i przechowywana w historii systemu, kluczowego elementu GALAXY. Pozwala to użytkownikowi na przeprowadzanie niezależnych zapytań (*queries*) dla danych genomowych pochodzących z różnych źródeł, a następnie wykorzystanie GALAXY w celu ich połączenia lub filtrowania, przeprowadzenia obliczeń, czy też wyciągnięcia i wizualizacji odpowiadających sobie sekwencji lub dopasowań. Działania te mogą zostać wykonane z wykorzystaniem prostego interfejsu internetowego widocznego na Ryc. 2.



Ryc. 2. Zrzut ekranowy interfejsu internetowego GALAXY.

Obecnie Galaxy składa się z trzech zasadniczych klas stworzonych do manipulowania danymi: klasa operacji nad zapytaniami, klasa stanowiąca narzędzie do analizy sekwencji oraz klasa do wyświetlania wyniku. Pierwsza z nich zawiera standardowy zestaw operacji, takich jak dodawanie, sumowanie, odejmowanie, jak również filtry opierające się na rozmiarze regionu, sąsiedztwie do regionu z innego zapytania (*query*)

oraz grupowanie na podstawie dystansów pomiędzy regionami w obrębie pojedynczego zapytania. Narzędzie do analizy sekwencji to niezależne moduły zaprojektowane do przeprowadzania zorientowanych biologicznie działań takich jak szukanie ortologicznych regionów u innego gatunku, wydobywanie dopasowań genomowych, otrzymywanie zawartości par GC lub konserwacji w regionach zainteresowania. Ostatni moduł, służący do prezentacji wyników, pozwala na wyświetlanie rezultatów wygenerowanych przez użytkownika w wybranym przez niego formacie. Użytkownik może również pobrać plik tekstowy z wynikami w wielu różnych formatach – standardowym BED (*Browser Extensible Data*), Ensembl upload lub też jako zwykły „surowy" tekst.

W celu zaprezentowania zasadniczej funkcjonalności GALAXY można posłużyć się przykładem. Korzystając z systemu, który zapisuje historię aktywności użytkownika można wystosować niezależne zapytania mające za zadanie znalezienia pojedynczych polimorfizmów nukleotydowych (*SNP*) w egzonach danego genu. Na początku użytkownik wybiera pożądany region genomowy (*locus* z badanym genem) z przeglądarki UCSC Table Browser [49], a w kolejnym kroku rezultaty jego zapytania zostają przekazane bezpośrednio do GALAXY. Strona z historią prezentuje jedno zapytanie z koordynatami genomowymi dla każdego egzonu należącego do badanego genu. Następnie użytkownik wykorzystując ponownie Table Browser żąda wszystkich SNP, które zawierają się w regionie genomowym analizowanego genu. Po takim działaniu na stronie historii pojawi się drugie zapytanie zawierające zażądane przez użytkownika SNP. Ponieważ głównym celem jest znalezienie SNP zawartych w kodujących egzonach należy wykonać operację skrzyżowania (*intersection*) dwóch zapytań z historii. Otrzymany wynik prezentowany jest jako nowa pozycja na stronie historii, a użytkownik może pobrać rezultaty lub wyświetlić je w zdefiniowanym przez siebie formacie.

System GALAXY pozwala użytkownikom na wykorzystanie istniejących algorytmów ewolucji molekularnej bezpośrednio do analizy sekwencji otrzymanych z zapytań. Wewnątrz GALAXY zawarte zostało narzędzie do obliczania wskaźników synonimicznych (Ks) i niesynonimicznych (Ka) substytucji z wykorzystaniem algorytmu Yanga-Neilsena [50]. Narzędzie to pozwala na wykonywanie szacowań zarówno dla sekwencji pełnej długości, jak i z wykorzystaniem podejścia ruchomego okna (*sliding window*). Wyniki mogą zostać użyte do prostego testu wskaźnika Ka/Ks, jako predykatora oddziaływań doboru naturalnego na region kodujący białko [51]. Użytkownicy GALAXY mogą zastosować taką analizę dla każdej kodującej sekwencji dostępnej w UCSC Table Browser.

Rdzeń GALAXY jak i biblioteki operacji zostały napisane w języku C i stworzone zgodnie ze standardami grupy bioinformatycznej w UCSC. Interfejs użytkownika napisany został w języku programowania Perl dla wygodnego manipulowania tekstem oraz dostępu CGI (*Common Gateway Interface*), czyli interfejsu, który pozwala na komunikację oprogramowania i serwera WWW z innymi programami zainstalowanymi na serwerze [52].

GALAXY dostępny jest pod adresem http://main.g2.bx.psu.edu/.

3.2.2.1. Format BED

Podstawowym formatem wykorzystywanym przez GALAXY w celu przechowywania wyników zapytań jest format BED. Jest on używany w UCS dla rekordów w Genome Browser oraz jest akceptowany przez różne inne strony. Format ten składa się z tekstu oddzielonego tabulacjami z możliwością czytania go zarówno przez programy komputerowe, jak i przez człowieka. Przykład formatu BED znajduje się poniżej:

chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512,
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601,

gdzie w pierwszej kolumnie (*chrom*) zawarta jest nazwa chromosomu lub rusztowania (scaffold), w drugiej (*chromStart*) i trzeciej (*chromEnd*) odpowiednio pozycja początkowa i końcowa danej cechy ba chromosomie. Kolumna czwarta (*name*) zawiera nazwę danego genu/cechy, piąta (*score*) to wartość wyniku, przyjmującego wielkości pomiędzy 0 a 1000. W szóstej kolumnie (*strand*) zawarto oznaczenie znaku nici (+ lub -), kolumny numer siedem (*thickStart*) i osiem (*thickEnd*) to odpowiednio początek i koniec pozycji, gdzie dana cecha rysowana jest pogrubioną linią (w UCSC Genome Browser), na przykład kodon start/stop w genie. Kolumna dziewiąta (*itemRgb*) zawiera informację na temat koloru wyświetlania danych zawartych w danej linijce formatu BED. Następna (*blockCount*) to liczba bloków - egzonów. Jedenasta (*blockSizes*) to oddzielona przecinkami lista rozmiarów tych bloków. Ostatnia, dwunasta kolumna (*blockStarts*) zawiera odseparowaną znakami przecinka listę pozycji startowych dla bloków [53]. Wszystkie pozycje dla bloków powinny być obliczane w odniesieniu do pozycji podanej w kolumnie zawierającej pozycję początkową na chromosomie (*chromStart*).

3.2.3. BLAST

Algorytm BLAST (*Basic Local Alignment Search Tool*) pierwszej generacji [54] opiera się na heurystycznej metodzie wyznaczania dopasowań lokalnych par sekwencji. Wykorzystanie metod heurystycznych ma na celu przyspieszenie przeszukiwania baz danych sekwencji, czy to nukleotydowych, czy białkowych, których wielkość rośnie w szybkim tempie. W ten sposób nadaje się bardziej do tego typu zadań niż metody wykorzystujące programowanie dynamiczne. Podejście to jest co prawda stosowane w programie BLAST, ale jedynie do stworzenia ostatecznych dopasowań sekwencji. Wybór metody heurystycznej pociąga za sobą koszt utraty gwarancji otrzymania najlepszego dopasowania sekwencji, ale za to pozwala na znaczne przyspieszenie przeszukiwania baz danych. BLAST mając podane jako zapytanie daną sekwencję przeszukuje z jej pomocą wskazaną bazę danych i wyznacza najlepiej ocenione, lokalne dopasowanie bez wprowadzonych przerw oraz kilka dopasowań o niższej punktacji.

Algorytm, na którym opiera się metoda BLAST korzysta z określonej domyślnie lub przez użytkownika wartości parametru nazywanego słowem (*w*). Długość takiego słowa jest zazwyczaj równa 3 dla sekwencji aminokwasowych, a 12 dla sekwencji nukleotydowych. BLAST szuka w bazie danych sekwencji, które posiadają zadaną długość *w*, a w kolejnym kroku rozszerza lokalne dopasowanie przez rozbudowę początku i końca szukanego słowa. Jeżeli na którymś z rozbudowywanych końców wartość oceny dopasowania obniży się w porównaniu z wcześniejszą wartością maksymalną o pewną ustaloną wcześniej wartość, to wydłużanie zostaje przerwane. Następnie spośród tak wyznaczonych lokalnych dopasowań algorytm wybiera to najlepiej ocenione.

Dla oceny wiarygodności działania algorytmu BLAST i otrzymanych wyników wprowadzono wartość *E* (*expected value*), która reprezentuje liczbę tzw. sekwencji *HSP* (*high-scoring segment pair*), które mogłyby być znalezione na drodze przypadku. Im więc niższa jest wartość E, tym bardziej znaczący jest wynik. HSP to nic innego, jak znalezione przy pomocy BLAST znaczące dopasowanie. Dla dwóch badanych sekwencji możliwe jest znalezienie więcej niż jednego HSP [55].

Z uwagi na różne potrzeby przeszukiwania baz danych algorytm BLAST został zaimplementowany na kilka sposobów:

- **BLASTP** – wykorzystywany do przeszukiwania bazy danych sekwencji białkowych na podstawie zapytania będącego również sekwencją aminokwasową,

- **BLASTN** – dla zadanej sekwencji nukleotydowej szuka dopasowań w bazie danych sekwencji nukleotydowych,
- **BLASTX** – zadaną sekwencję nukleotydową zamienia na sekwencję aminokwasową i dla uzyskanej sekwencji szuka dopasowań w bazie danych sekwencji aminokwasowych,
- **TBLASTN** – używany do przeszukiwania za pomocą zadanej sekwencji aminokwasowej bazy danych sekwencji aminokwasowych uzyskanych z tłumaczenia (translacji) sekwencji nukleotydowych.

Program BLAST można zainstalować na lokalnym komputerze niezależnie od wykorzystywanego systemu operacyjnego i uruchamiać z wiersza polecenia/konsoli, bądź też korzystać z jego implementacji na serwerach NCBI pod adresem http://blast.ncbi.nlm.nih.gov/Blast.cgi.

3.2.3.1. *Reciprocal* BLAST

Reciprocal BLAST (odwzajemniony BLAST) to powszechna metoda obliczeniowa wykorzystywana do przewidywania prawdopodobnych ortologów. W metodzie tej wybrany gen zostaje wykorzystany do przeszukiwania algorytmem BLAST bazy danych sekwencji analizowanego organizmu. Z wyników wybierane są geny, które uzyskały najwyższy wynik (*bit score*). Następnie zostają one użyte do przeszukania tym samym algorytmem BLAST bazy danych sekwencji genowych organizmu, z którego pochodził wybrany początkowo gen. Jeżeli w wyniku tej procedury otrzymamy najwyższą wartość dopasowania (*bit score*) dla genu, który został pierwotnie użyty w analizie, oznacza to, że te dwa geny można uważać jako przypuszczalne ortologi.

Oczywiście metoda ta nie dowodzi ortologii. Co więcej, *reciprocal* BLAST nie bierze w pełni pod uwagę sytuacji, gdzie historia genu jest skomplikowana przez duplikacje genowe [56].

3.2.4. PYTHON

Python to język programowania wysokiego poziomu przeznaczony do generalnego zastosowania. Jego filozofia opiera się na czytelności kodu. Podstawowa składnia i semantyka są minimalistyczne, podczas gdy standardowa biblioteka jest dość duża

i wszechstronna. To, co wyróżnia go spośród innych popularnych języków programowania to użycie białych znaków (spacji, tabulatorów) jako separatorów bloków [57].

Python został stworzony w 1990 roku przez Guido van Rossuma. Charakteryzuje się w pełni dynamicznym systemem typów oraz samoczynnym zarządzaniem pamięcią. Jest więc w tym względzie podobny do takich języków, jak Perl, czy Ruby. Rozwijany jest jako projekt otwartego oprogramowania (*Open Source*) i zarządzany przez *Python Software Foundation* (http://www.python.org/psf/) [58]. Jest to język w pełni obiektowy (zorientowanym obiektowo), a pomimo tego faktu nie wymaga od użytkownika obiektowego stylu programowania. Zmienne są w pełni dynamiczne i nie posiadają typów, lecz tylko wartości. Stąd też ta sama zmienna może raz przechowywać liczbę, a innym razem ciąg znaków.

Python po instalacji oferuje bogaty zestaw bibliotek. Ponieważ często wykorzystuje dziedziczenie w stosunkowo łatwy sposób można nie tylko implementować biblioteki we własnych programach, ale także stosować ich dziedziczenie, rozszerzać i tworzyć nowe. Środowisko Pythona jest w pełni interaktywne. Korzystając z interpretera można na bieżąco wprowadzać kolejne polecenia i oglądać wyniki, co bardzo ułatwia tworzenie programów i usuwanie z nich błędów [59].

Najnowszą wersję języka (Python 3.0.1 z 13 lutego 2009) przeznaczoną na różne systemy operacyjne można pobrać na stronie http://www.python.org/download/.

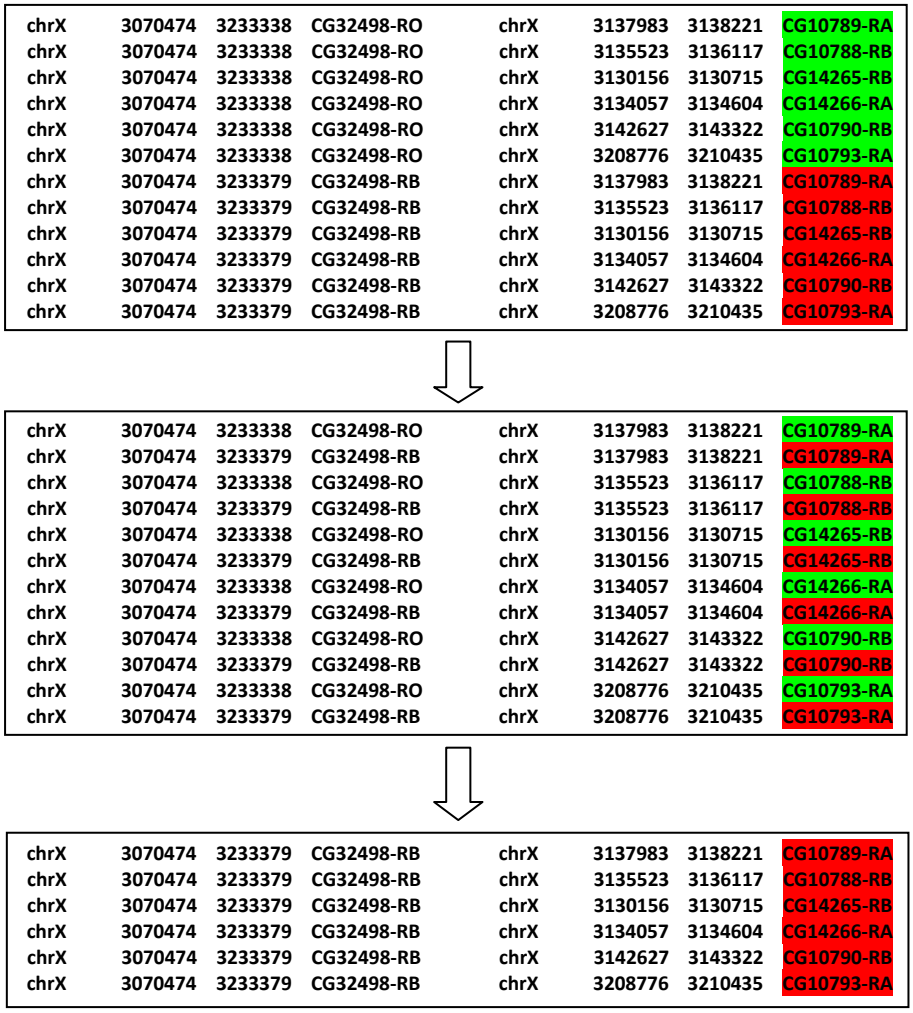
3.3. Metody

3.3.1. Identyfikacja genów nakładających się

W pierwszym etapie analiz z użyciem platformy GALAXY (http://main.g2.bx.psu.edu/) wyłoniono zbiór nakładających się genów z genomu *D. melanogaster*. Początkowym krokiem było wybranie genomu muszki owocowej z UCSC Table Browser jako wyjściowego zestawu do analiz. Wynik został przesłany do GALAXY, a następnie z użyciem polecenia *'Filter'* z zakładki *'Filter and Sort'* podzielony na dwa podzbiory względem zajmowanej nici DNA (kolumna 6 formatu BED). Otrzymane podzbiory z nici sens i antysens porównano ze sobą. Spośród dostępnych w GALAXY narzędzi, najbardziej przydatne okazało się JOIN z zakładki *'Operate on Genomic Intervals'*. Bierze ono na wejściu dwa zapytania (sekwencje) oraz parametr minimalnej długości nałożenia, a na wyjściu zwraca wynik w czterech możliwych formatach. Najprzydatniejszy

z nich to tzw. INNER JOIN – dostajemy tylko te rekordy z obu zapytań, które rzeczywiście się nakładają. W przypadku innych opcji zwracane są wszystkie rekordy pierwszego zapytania, wszystkie rekordy drugiego zapytania lub też rekordy z dwóch zapytań, bez względu na to, czy nakładają się, czy nie.

Otrzymany plik wynikowy z GALAXY w rozszerzonym formacie BED (24 kolumny, po 12 na jeden gen z pary nakładającej się) został następnie przefiltrowany pod kątem unikalnych nałożeń. Ponieważ zbiór zawierał geny, dla których istnieje po kilka wariantów splicingowych to samo nałożenie pomiędzy genami występowało kilka razy. Plik został początkowo posortowany alfabetycznie według kolumny zawierającej nazwę drugiego genu, a następnie na podstawie porównywania nazw obu genów w kolejnych wierszach usunięto powtarzające się nakładania. Schemat powyższego działania prezentuje Ryc. 3.



Ryc. 3. Przykład danych wyjściowych z GALAXY zawierających powtórzenia nałożeń spowodowane wariantami splicingowymi tego samego genu**.** Blok oznaczony na czerwono zawiera dokładnie te same nazwy genów, co blok zielony. Oba bloki nakładają się z tymi samymi genami, jednakże z ich różnymi wariantami splicingowymi. W przykładzie pominięto niektóre kolumny formatu BED.

W tak przefiltrowanym zbiorze na podstawie koordynatów ustalono dla poszczególnych par genów jeden z trzech typów nakładania (*head-to-head*, *tail-to-tail*, *nested*). Wszystkie powyższe analizy zostały przeprowadzone z użyciem autorskiego programu GalaxyParser, napisanego w języku programowania Python.

Kolejnym krokiem było wyłonienie unikatowych genów, które tworzą nakładające się pary o ustalonym już wcześniej typie nakładania. Z pliku otrzymanego w wyniku działania programu GalaxyParser wybrano tylko kolumny zawierające nazwy genów, które następnie posortowano i porównywano ze sobą. Powtórzenia zostały usunięte i w rezultacie otrzymano plik zawierający listę unikatowych elementów (nazw genów). Omówione analizy również wykonano z wykorzystaniem programu napisanego w języku Python.

Dla uzyskanej w poprzednim etapie listy nazw genów muszki owocowej pobrano manualnie sekwencje białkowe wykorzystując do tego celu UCSC Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables) i polecenie *Paste list*.

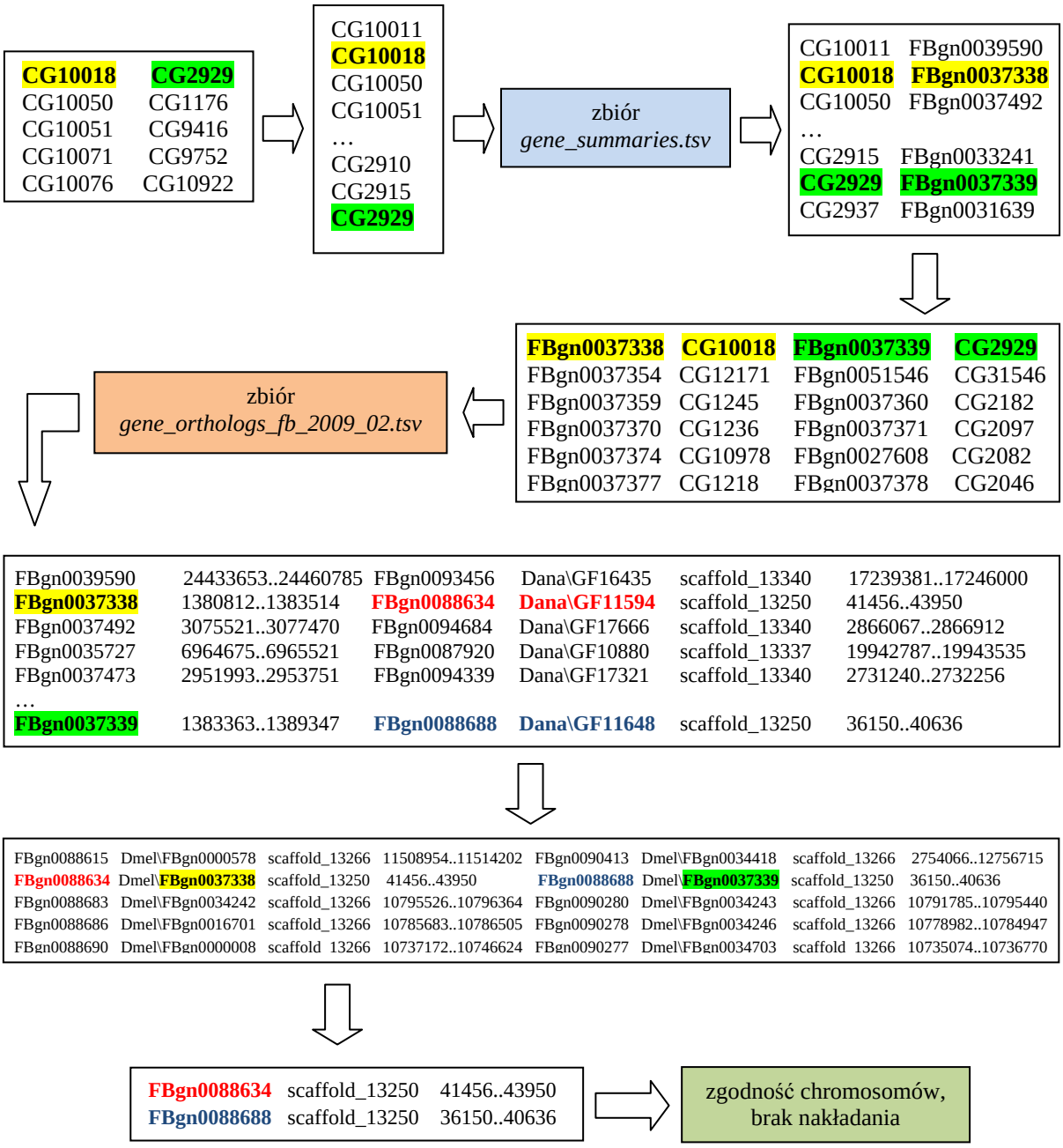
3.3.2. Identyfikacja ortologów u 11 gatunków *Drosophila*

Na początku drugiego etapu analiz na podstawie zbioru *gene_summaries.tsv* z bazy danych FlyBase (patrz Materiały) i listy unikatowych genów zaangażowanych w nakładanie utworzono listę zawierającą podwójne nazewnictwo każdego z tych genów (dla przykładu: *CG10071, FBgn0016726*). Dodatkowa druga nazwa to identyfikator genu w bazie FlyBase (tzw. FlyBase ID), gdyż pierwsza stanowiła tylko symbol adnotacji genu w tej bazie (*annotation symbol*). Działanie takie spowodowane było tym, że zbiór zawierający ortologi dla genów z *D. melanogaster* wykorzystywał, w odróżnieniu od GALAXY, numer FlyBase ID jako oznaczenia genów muszki.

Z wykorzystaniem tak utworzonej listy przeszukano zbiór pobrany z bazy FlyBase - *gene_orthologs_fb_2009_02.tsv*. Pozwoliło to na wybranie ortologów dla genów *D. melanogaster* spośród pozostałych 11 gatunków muszek z rodzaju *Drosophila*. Nowoutworzony zbiór posiadał także lokalizację chromosomową każdego ortologa oraz jego koordynaty.

Kolejnym krokiem było utworzenie listy par nakładających się ortologów dla każdego z tych 11 gatunków na podstawie listy par nakładających się genów *Drosophila*

melanogaster. Następnie sprawdzono na podstawie koordynatów utworzonych par ortologów, czy zachodzi pomiędzy nimi zjawisko nakładania oraz jakiego typu jest to nakładanie. Porównywano także, czy dane ortologi z pary znajdują się na tym samym chromosomie (lokalizacja na różnych chromosomach może świadczyć o translokacjach, co potwierdzałoby hipotezę Shintani'ego – patrz Wstęp). Wszystkie powyższe analizy także przeprowadzono z użyciem niewielkich autorskich programów stworzonych w języku Python. Rycina 4 prezentuje schemat postępowania w drugim etapie analiz.



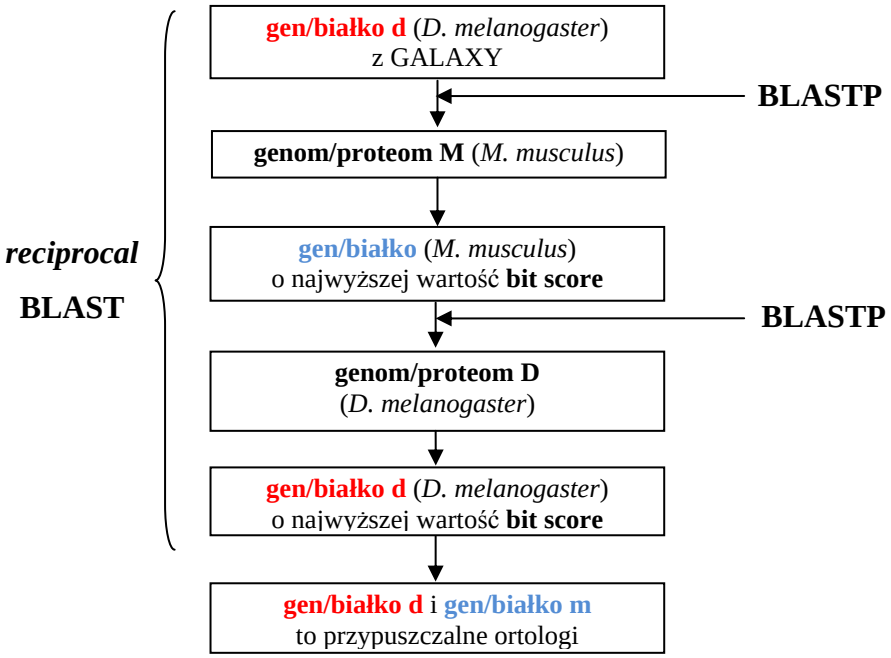
Ryc. 4. Schemat analizy i obróbki danych w drugim etapie pracy zaprezentowany na przykładzie pary nakładających się genów z *D. melanogaster* w orientacji *tail-to-tail* – *CG10018* i *CG2929*. Zidentyfikowane oba ortologi z genomu *D. ananassae* znajdują się na tym samym chromosomie, ale nie odtwarzają nakładania.

Końcowym krokiem tego etapu było zbadanie, które pary nakładających się genów *D. melanogaster* występują u wszystkich pozostałych 11 gatunków z rodzaju *Drosophila*. Analiza ta również została przeprowadzona za pomocą własnego programu stworzonego w języku Python.

3.3.3. Identyfikacja ortologów pośród genów 6 organizmów modelowych

W trzecim etapie zestaw białek z *D. melanogaster* kodowanych przez zidentyfikowane przy użyciu GALAXY geny nakładające się porównano z zestawami białek sześciu organizmów modelowych (*Annopheles gambiae, Apis mellifera, Danio rerio, Gallus gallus, Homo sapiens, Mus musculus*) w celu znalezienia ortologów dla genów muszki. Sekwencje białek organizmów modelowych pobrano z bazy danych NCBI Taxonomy Browser. Porównania dokonano z wykorzystaniem programu BLASTP, który uruchamiany był lokalnie na komputerze z systemem operacyjnym MS Windows®. Parametr *E* na potrzeby analiz ustalono na poziomie 0.0001. Wyniki ograniczone zostały do jednego, najwyżej punktowanego trafienia (opcja –b 1), a jako formę prezentacji wyniku wybrano format tabularny z opisami (opcja –m 9). Następnie przy użyciu autorskiego programu napisanego w języku Python wyodrębniono z pliku pary utworzone przez nazwę genu muszki owocowej i numer GI białka danego organizmu modelowego, które uzyskało najwyższy wynik w procesie dopasowania algorytmem BLAST.

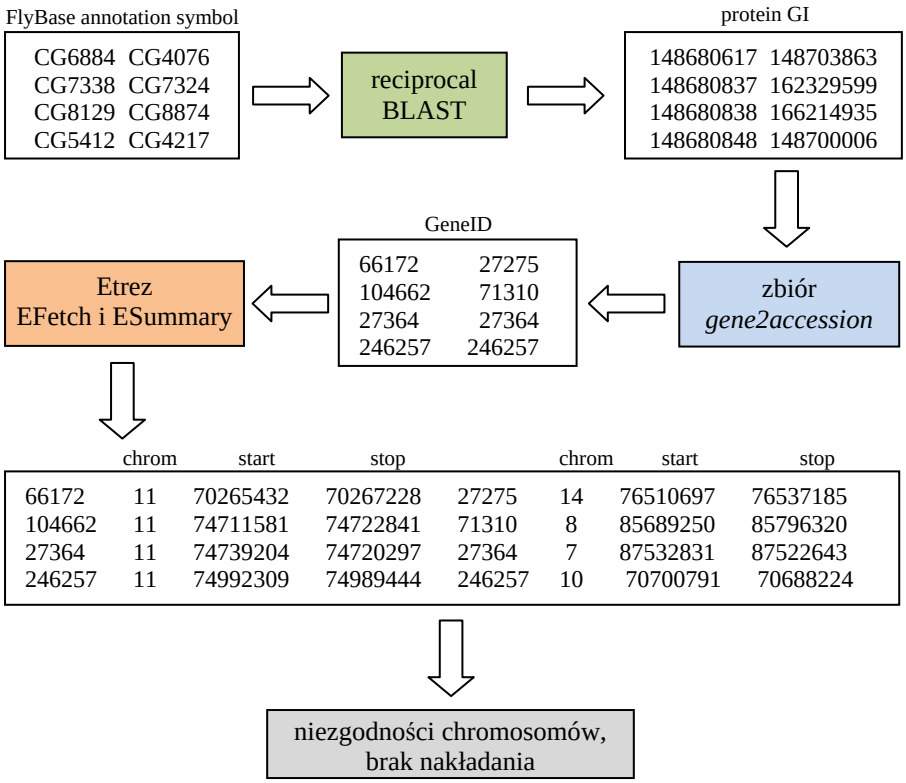
Dla listy numerów GI białek organizmów modelowych pobrano ich sekwencje z wykorzystaniem własnego programu, który korzysta z narzędzia EFetch z pakietu Entrez Programming Utilities (eUtils, http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html), pozwalającego na dostęp do zbiorów bazy NCBI poza przeglądarką internetową. Tak pobrane zbiory wykorzystano w przeszukiwaniu zbioru transkryptów *D. melanogaster* w celu potwierdzenia znalezionych ortologów. Ponownie jako metodę przeszukiwania wykorzystano algorytm BLASTP. Zbiór transkryptów muszki owocowej stanowiący bazę do przeszukiwania pobrano z UCSC Table Browser. Opisany powyżej sposób znajdowania ortologów zwany metodą odwzajemnionego BLASTa (*reciprocal* BLAST) omówiono już w rozdziale Narzędzia. Jego schemat na przykładzie genów muszki owocowej i myszy prezentuje rycina na kolejnej stronie (Ryc. 5).



Ryc. 5. Schemat postępowania w metodzie *reciprocal* BLAST na przykładzie *D. melanogaster* i *M. musculus*.

Na podstawie zbioru zawierającego pary nakładających się genów *D. melanogaster* utworzono pary stworzone ze zidentyfikowanych w metodzie reciprocal BLAST ortologów dla tych genów. Następnie wykorzystując zbiór gene2accesion uzyskano unikalne identyfikatory genów (*GeneID*) na podstawie identyfikatorów białek (*protein GI*) analizowanych sześciu organizmów modelowych.

W kolejnym etapie z wykorzystaniem autorskiego programu integrującego narzędzia EFetch i ESummary (http://www.ncbi.nlm.nih.gov/corehtml/query/static/esummary_help.html) pobrano dla każdego identyfikatora genowego informacje o jego lokalizacji chromosomowej oraz koordynaty genu. Ostatecznym krokiem było sprawdzenie, czy dana para genów odtwarza nakładanie wyjściowej pary genów z muszki owocowej. Sprawdzano zgodność chromosomów, na jakich dane dwa ortologi się znajdują, a w przypadku jej stwierdzenia oceniano na podstawie koordynatów genowych, czy istnieje między nimi jakiekolwiek pokrycie (nałożenie sekwencji). Schemat działania przedstawiono na kolejnej rycinie (Ryc. 6).



Ryc. 6. Schemat identyfikacji ortologów dla par genów nakładających się z *D. melanogaster* oraz weryfikacji, czy odtwarzają one dane nałożenie. Schemat zaprezentowano na przykładzie danych z *M. musculus*.

Na koniec sprawdzono, czy pomiędzy zanalizowanymi nałożeniami wykrytymi u 12 gatunków muszek z rodzaju *Drosophila*, a nałożeniami z 6 organizmów modelowych istnieje część wspólna, stanowiąca przykład starej ewolucyjnie pary nakładających się genów.

4. WYNIKI

4.1. Geny nakładające się zidentyfikowane w genomie D. melanogaster

W pierwszym etapie analiz, który miał na celu wyłonienie zestawu par nakładających się genów *D. melanogaster*, platforma GALAXY spośród przeanalizowanych 21243 transkryptów zwróciła plik wynikowy zawierający 3929 par nakładających się genów. 1928 wpisy usunięto z uwagi na powtarzalność związaną z występowaniem wariantów splicingowych, co zawyżało liczbę faktycznych nałożeń między genami. Dało to ostateczny zbiór 2001 unikalnych par nakładających się genów u muszki owocowej. Nakładania te utworzone były przez 3504 unikalne geny (1751 na nici DNA -, 1753 na nici +), co stanowi 16,49% wszystkich transkryptów muszki owocowej. Liczba zidentyfikowanych genów nie jest po prostu równa dwukrotności liczby par (2 razy 2001) z uwagi na to, że niektóre geny nakładają się z więcej niż jednym genem. Na przykład gen *CG15221* może tworzyć trzy pary nakładające się (Ryc. 7), które nie są równoznaczne sześciu genom zaangażowanym w nakładanie, a czterem.

Ryc. 7. Gen *CG15221* i trzy geny w nim zagnieżdżone: *CG34322*, *CG34321*, *CG2457*. Niebieskie bloki oznaczają egzony, strzałki kierunek genu na nici DNA.

Poniższa tabela zawiera podsumowanie wyników omówionego powyżej pierwszego etapu analiz (Tab. 1).

Tab. 1. Wyniki pierwszego etapu analiz nakładających się genów w genomie *D. melanogaster*.

Transkrypty *D. melanogaster*	Znalezionych par nakładających się	Unikatowych par	Unikatowych genów w parach
21243	3929	2001	3504

Spośród zidentyfikowanych 3504 unikatowych genów większość z nich znajdowała się w orientacji *tail-to-tail* - 1017. Liczba zagnieżdżonych genów to 838, a genów w pozycji *head-to-head* – 173 (Tab. 2).

Tab. 2. Wyniki analizy typu nakładania się dla genów *D. melanogaster* w pierwszym etapie analiz.

Nałożenia typu *nested*	Nałożenia typu *tail-to-tail*	Nałożenia typu *head-to-head*
838	1017	173

4.2. Geny nakładające się u pozostałych 11 gatunków rodzaju Drosophila

Etap drugi pracy miał na celu zidentyfikowanie ortologów dla genów *D. melanogaster* wśród pozostałych gatunków rodzaju *Drosophila* (*D. pseudoobscura, D. simulans, D. sechellia, D. yakuba, D. erecta, D. ananassae, D. persimilis, D. willistoni, D. virilis, D. mojavensis, D. grimshawi*).

Dla 3504 unikalnych genów zaangażowanych w nakładanie u *D. melanogaster* najwięcej ortologów (zakonserwowanych genów) zidentyfikowano w genomie *D. yakuba* – 3022, a najmniej u *D. persimilis* – 2712. W przypadku zidentyfikowanych w pierwszym etapie 2001 unikalnych par nakładających się najwięcej zakonserwowanych par zaobserwowano w genomie *D. yakuba* – 1537, a najmniej w genomie *D. persimilis* – 1250. Tabela podsumowująca wyniki tego etapu analiz znajduje się poniżej (Tab. 3).

Tab. 3. Wynik analiz genów nakładających się dla 11 gatunków muszek z rodzaju *Drosophila*.

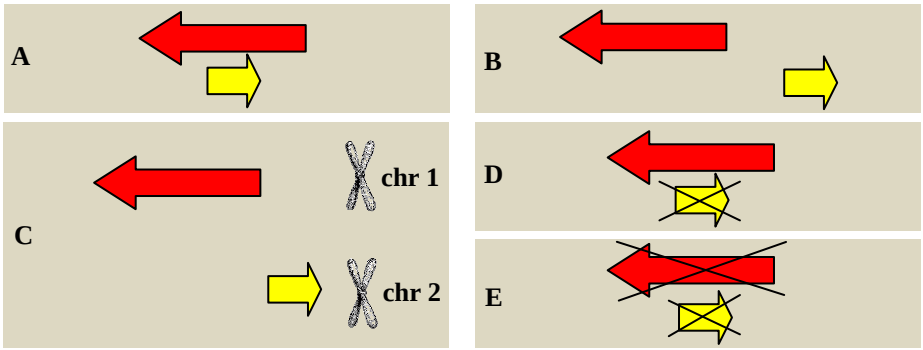
Nazwa gatunku	Geny ortologiczne	Zakonserwowane pary genów	Unikatowe geny w parach
D. ananassae	2882	1415	2561
D. erecta	3017	1525	2733
D. grimshawi	2714	1270	2323
D. mojavensis	2736	1284	2350
D. persimilis	2712	1250	2284
D. pseudoobscura	2807	1354	2457
D. sechellia	2955	1467	2639
D. simulans	2782	1300	2351
D. virilis	2750	1301	2377
D. willistoni	2762	1302	2388
D. yakuba	3022	1537	2752

W kolejnym kroku analizy ustalano ile ze zidentyfikowanych genów ortologicznych nie odtwarza pary nakładającej się (występuje tylko jeden gen z pary) Największa liczba takich genów zlokalizowana została w genomie *D. simulans* – 431, a najmniejsza u *D. yakuba* - 270. Ustalono liczbę genów biorących udział w zjawisku nakładania u *D. melanogaster*, która nie jest zakonserwowana w analizowanych 11 gatunkach muszek z rodzaju *Drosophila* z największym wynikiem dla *D. persimilis* – 792, a najmniejszym dla *D. sechellia* - 549. Ponadto wyznaczono liczbę par nakładających się genów, która nie uległa konserwacji w przebadanych muszkach. Największa liczba niezakonserwowanych par genów zidentyfikowana została w genomie *D. persimilis* – 751, a najmniejsza u *D. sechellia* – 534. Sumaryczne wyniki omówionych analiz zaprezentowano w tabeli 4.

Tab. 4. Wynik analizy braku konserwacji genów ortologicznych oraz par nakładających się genów. **A.** Liczba zidentyfikowanych genów ortologicznych pomniejszona o liczbę unikatowych genów w zakonserwowanych parach; **B.** Liczba unikatowych genów z *D. melanogaster* pomniejszona o liczbę zidentyfikowanych ortologów; **C.** Liczba unikatowych par z *D. melanogaster* pomniejszona o liczbę zakonserwowanych par genów.

Nazwa gatunku	A. Ortologi, które nie odtwarzają nakładania (brakujący1 gen z pary)	B. Niezakonserwowane geny (brakujące 2 geny z pary)	C. Niezakonserwowane pary
D. ananassae	2882-2561 = 321	3504 – 2882 = 622	2001 – 1415 = 586
D. erecta	3017-2733 = 284	3504 – 3017 = 487	2001 – 1525 = 476
D. grimshawi	2714-2323 = 391	3504 – 2714 = 790	2001 – 1270 = 731
D. mojavensis	2736-2350 = 386	3504 – 2736 = 768	2001 – 1284 = 717
D. persimilis	2712-2284 = 428	3504 – 2712 = 792	2001 – 1250 = 751
D. pseudoobscura	2807-2457 = 350	3504 – 2807 = 697	2001 – 1354 = 647
D. sechellia	2955-2639 = 316	3504 – 2955 = 549	2001 – 1467 = 534
D. simulans	2782-2351 = 431	3504 - 2782 = 722	2001 – 1300 = 701
D. virilis	2750-2377 = 373	3504 - 2750 = 754	2001 – 1301 = 700
D. willistoni	2762-2388 = 374	3504 - 2762 = 742	2001 – 1302 = 699
D. yakuba	3022-2752 = 270	3504 – 3022 = 482	2001 – 1537 = 464

Rycina poniżej prezentuje wszystkie typy przeanalizowanych przypadków konserwacji par nakładających się genów (Ryc. 8).



Ryc. 8. Typy analizowanych przypadków konserwacji par nakładających się genów. **A.** Zakonserwowane nałożenie; **B.** Zakonserwowana para genów, nie odtwarzająca nałożenia; **C.** Zakonserwowana para genów zlokalizowanych na różnych chromosomach; **D.** Zakonserwowany tylko jeden gen z pary; **E.** Brak konserwacji obu genów z pary.

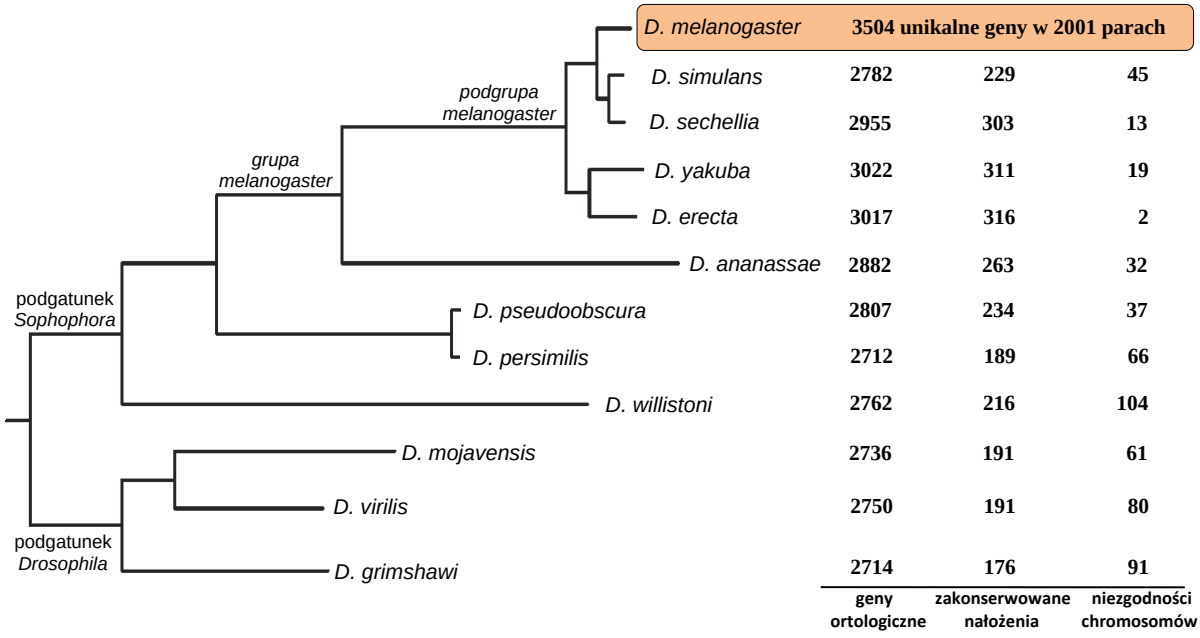
W kolejnym kroku analizowano, czy zakonserwowane pary genów odtwarzają nakładanie. Najwięcej zachowanych par nakładających się zidentyfikowano w genomie *D. erecta* – 316, a najmniej u *D. grimshawi* – 176. W wyniku analizy sposobu nakładania się genów w tych parach u 11 gatunków muszek z rodzaju *Drosophila* najwięcej zakonserwowanych nałożeń typu *head-to-head* zidentyfikowano u *D. mojavensis* i *D. pseudoobscura* – 5, typu *tail-to-tail* u *D. erecta* – 6, a nałożeń typu *nested* także w genomie *D. erecta* – 306. Bardzo niska liczba wykrytych nałożeń typu głowa do głowy,

czy ogon do ogona jest prawdopodobnie związana ze słabą adnotacją regionów 3’ UTR i 5’ UTR w genomach 11 muszek, w których to regionach występują tego typu nakładania. Podczas analizy konserwacji nałożenia zbadano także, czy oba geny z pary zlokalizowane są na tym samym chromosomie. W etapie tym najwięcej niezgodności chromosomów wykazały pary zakonserwowanych genów z genomu *D. willistoni* – 104, a najmniej *D. erecta* – 2. Podsumowanie wyników analiz omówionych powyżej prezentuje poniższa tabela (Tab. 5).

Tab. 5. Wyniki analizy sposobu nakładania się genów u 11 gatunków muszek z rodzaju *Drosophila*.

Nazwa gatunku	Zakonserwowane nałożenia	Nałożenia *head-to-head*	Nałożenia *tail-to-tail*	Nałożenia *nested*	Niezgodności chromosomów
D. ananassae	263	4	2	257	32
D. erecta	316	4	6	306	2
D. grimshawi	176	2	3	171	91
D. mojavensis	191	5	0	186	61
D. persimilis	189	4	1	184	66
D. pseudoobscura	234	5	3	226	37
D. sechellia	303	4	5	294	13
D. simulans	229	4	1	224	45
D. virilis	191	0	2	189	80
D. willistoni	216	4	2	210	104
D. yakuba	311	4	3	304	19

Rycina poniżej prezentuje częściowe wyniki drugiego etapu analizy na drzewie filogenetycznym rodzaju *Drosophila* (Ryc. 9).



Ryc. 9. Reprezentacja wyników analizy nakładających się genów wśród 12 gatunków z rodzaju *Drosophila* na jego drzewie filogenetycznym. Zmieniono za Nature 2007 [39].

W dalszych analizach poszukiwano par nakładających się genów, które występują zarówno u *D. melanogaster*, jak i u 11 przebadanych gatunków muszek z rodzaju *Drosophila*. Wykryto 61 par genów nakładających się, które utworzone są przez 116 unikatowych genów, sklasyfikowanych jako geny kodujące białka.

4.3. Zidentyfikowane geny nakładające się u 6 organizmów modelowych

W trzecim etapie pracy poszukiwano pierwotnych nałożeń genów wspólnych dla owadów i kręgowców na podstawie porównania genów znalezionych w etapie pierwszym z genami sześciu organizmów modelowych (*Annopheles gambiae, Apis mellifera, Danio rerio, Gallus gallus, Homo sapiens, Mus musculus*).

Dla unikalnych 3504 genów tworzących pary nakładające się u *D. melanogaster* najwięcej ortologicznych genów zidentyfikowano u komara (*A. gambiae*) - 2064, a najmniej u myszy – 1147. Geny te tworzyły najwięcej zakonserwowanych par u *A. gambiae* – 745, a najmniej u kurczaka (*G. gallus*) – 437. Sumaryczne wyniki przedstawiono w poniższej tabeli.

Tab. 6. Wyniki analizy ortologów oraz par nakładających się genów u 6 modelowych organizmów.

Nazwa gatunku	Geny ortologiczne	Zakonserwowane pary genów	Unikatowe geny w parach
A. gambiae	2064	745	1417
A. mellifera	1826	598	1145
D. rerio	1683	528	1017
G. gallus	1548	437	841
H. sapiens	1709	543	1039
M. musculus	1147	575	1093

Najwyższa liczba genów, które nie odtwarzają pary nakładającej się występuje u kurczaka – 707, a najniższa u myszy – 572. W przypadku identyfikacji genów zaangażowanych w nakładanie u *D. melanogaster*, które nie posiadają swoich ortologów u analizowanych organizmów modelowych najwięcej takich genów znaleziono w genomie *M. musculus* – 2357, a najmniej u *A. gambiae* – 1440. Kolejny krok analizy mający na celu weryfikację konserwacji par nakładających się wyłonił największą liczbę niezakonserwowanych par genów u *G. gallus* – 1564, a najmniejszą u *A. gambiae* – 1256. Podsumowanie wyników analiz omówionych powyżej zaprezentowano w tabeli na kolejnej stronie (Tab. 7).

Tab. 7. Wyniki analizy braku konserwacji genów ortologicznych oraz par nakładających się genów u 6 organizmów modelowych. **A.** Liczba zidentyfikowanych genów ortologicznych pomniejszona o liczbę unikatowych genów w zakonserwowanych parach; **B.** Liczba unikatowych genów z *D. melanogaster* pomniejszona o liczbę zidentyfikowanych ortologów; **C.** Liczba unikatowych par z *D. melanogaster* pomniejszona o liczbę zakonserwowanych par genów.

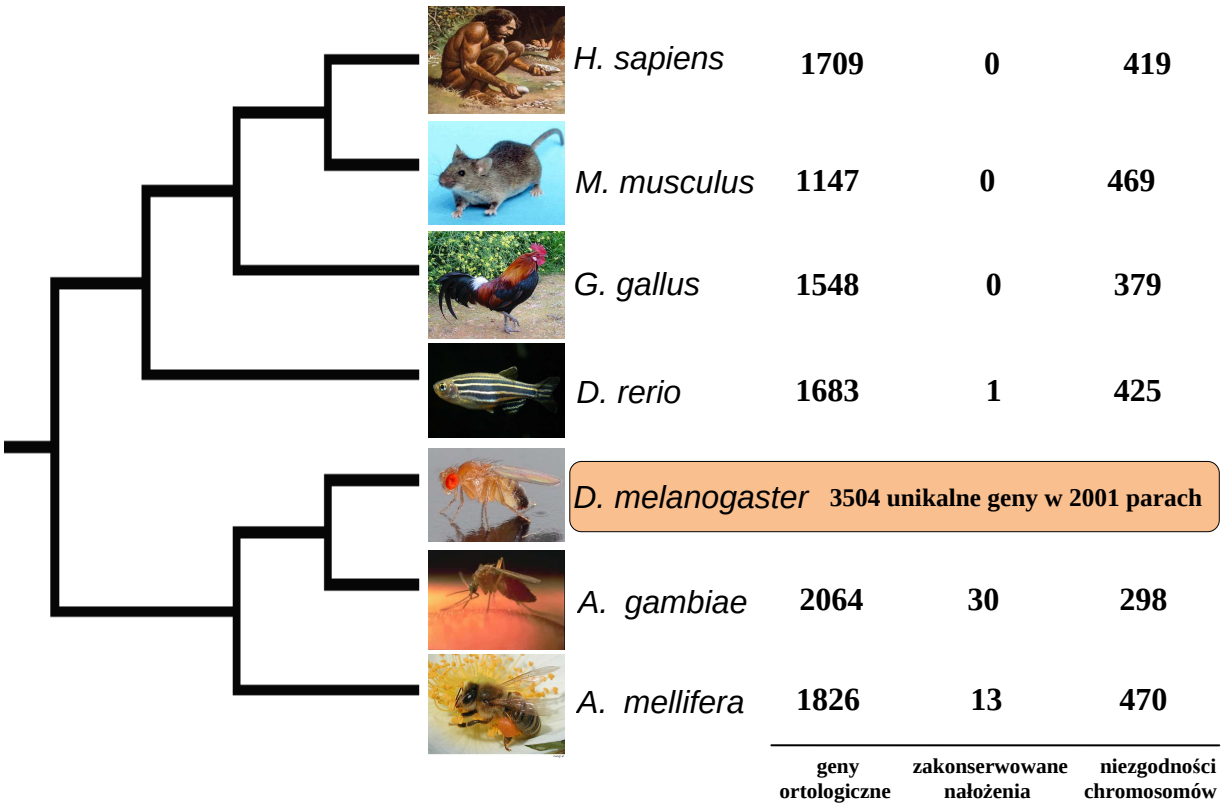
Nazwa gatunku	A. Ortologi, które nie odtwarzają nakładania (brakujący1 gen z pary)	B. Niezakonserwowane geny (brakujące 2 geny z pary)	C. Niezakonserwowane pary
A. gambiae	2064 – 1417 = 647	3504 – 2064 = 1440	2001 – 745 = 1256
A. mellifera	1826 – 1145 = 681	3504 – 1826 = 1678	2001 – 598 = 1403
D. rerio	1683 – 1017 = 666	3504 – 1683 = 1821	2001 – 528 = 1473
G. gallus	1548 – 841 = 707	3504 – 1548 = 1956	2001 – 437 = 1564
H. sapiens	1709 – 1039 = 670	3504 – 1709 = 1795	2001 – 543 = 1458
M. musculus	1147 – 1093 = 572	3504 – 1147 = 2357	2001 – 575 = 1426

W przypadku analizy konserwacji nałożeń zidentyfikowano ich 30 u komara, 13 u pszczoły i 1 u danio pręgowanego. Nie znaleziono żadnego zakonserwowanego nałożenia w genomie kurczaka, myszy i człowieka. Następnie wykonano analizę typu zidentyfikowanych powyżej zakonserwowanych nałożeń oraz ustalono liczbę niezgodności chromosomów pośród zakonserwowanych par ortologicznych genów. Najwięcej rozbieżności chromosomowych dały pary genów z *A. mellifera* – 470, a najmniej z *A. gambiae* – 298. Ilość niezgodności chromosomów jest znacznie wyższa niż w przypadku poprzedniej analizy dotyczącej 11 gatunków muszek rodzaju *Drosophila*. Podsumowanie wyników omówionych analiz przedstawiono w poniższej tabeli.

Tab. 8. Wynik analiz sposobu nakładania się genów dla 6 organizmów modelowych.

Nazwa gatunku	Zakonserwowane nałożenia	Nałożenia *head-to-head*	Nałożenia *tail-to-tail*	Nałożenia *nested*	Niezgodności chromosomów
A. gambiae	30	2	6	22	298
A. mellifera	13	6	2	5	470
D. rerio	1	0	0	1	425
G. gallus	0	0	0	0	379
H. sapiens	0	0	0	0	419
M. musculus	0	0	0	0	469

Na kolejnej rycinie zaprezentowano częściowe wyniki analiz wykonanych w trzecim etapie pracy w formie drzewa filogenetycznego uwzględniającego 6 badanych organizmów modelowych oraz *D. melanogaster* (Ryc. 10).



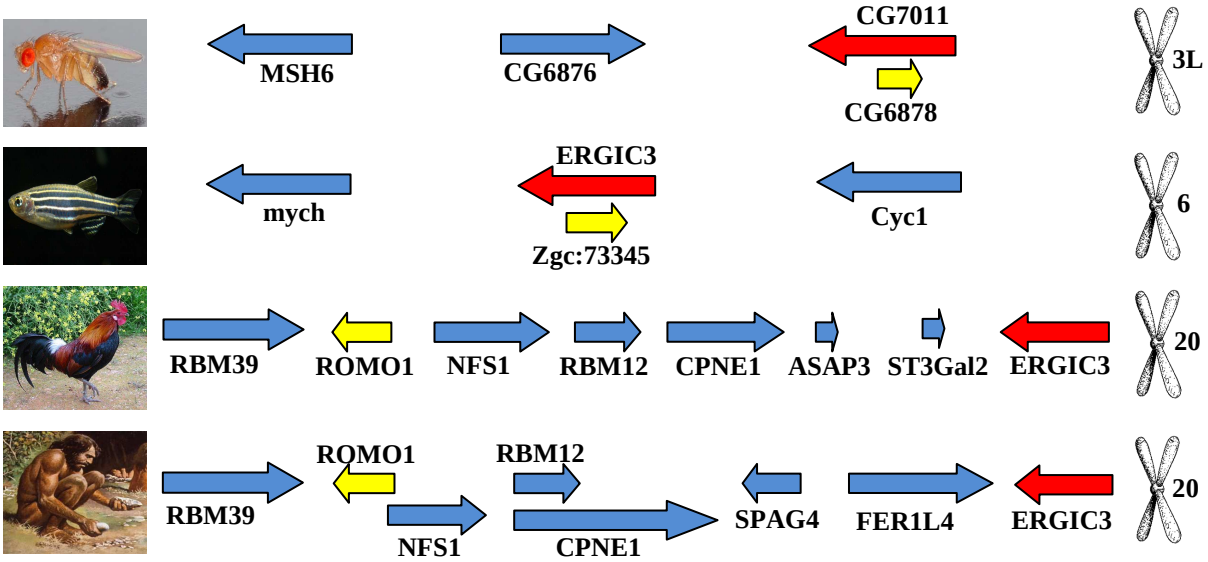
Ryc. 10. Reprezentacja wyników analizy nakładających się genów u 6 organizmów modelowych w formie drzewa rodowego badanych organizmów.

4.4. Geny nakładające się wspólne dla *D. melanogaster* i pozostałych analizowanych organizmów

W przeprowadzonych badaniach nie wykryto żadnej pary genów nakładających się, która byłaby wspólna zarówno dla 12 gatunków muszek z rodzaju *Drosophila*, jak i 6 przeanalizowanych organizmów modelowych – *A. gambiae*, *A. mellifera*, *D. rerio*, *G. gallus*, *H. sapiens*, *M. musculus*. Jedyna para wykazująca zakonserwowanie pomiędzy owadami i kręgowcami to para genów *D. melanogaster* - *CG6878*, *CG7011* – która wspólna jest dla pozostałych 11 gatunków muszek, ale również można zaobserwować parę ortologicznych genów u komara, pszczoły oraz danio pręgowanego. Para ta reprezentuje typ nałożenia *nested*. Ortologi tych genów występują także w genomach kurczaka, myszy

i człowieka, ale nie odtwarzają już nakładania. Znajdują się one na tych samych chromosomach, jednakże rozdzielone są przez kilka innych genów.

Geny *CG6878* i *CG7011* zlokalizowane są u muszki na chromosomie 3L. Zostały sklasyfikowane w bazie FlyBase jako geny kodujące białka o nieznanej funkcji molekularnej. Ich ortologi u przeanalizowanych kręgowców to odpowiednio geny *ERGIC3* i *zgc:73345* (*ROMO1*) u *D. rerio* zlokalizowane na chromosomie 6 oraz geny *ERGIC3* i *ROMO1* u kurczaka na chromosomie 20, myszy na chromosomie 2 i człowieka na chromosomie 20. U człowieka gen *ERGIC3* pełni rolę we wzroście komórki oraz indukowanej stresem śmierci komórkowej. Jest składnikiem błony retikulum endoplazmatycznego i aparatu Golgiego [60]. Gen *ROMO1* (*reactive oxygen species modulator 1*) odpowiedzialny jest za zwiększony poziom reaktywnych form tlenu związany z wiekiem i może pełnić ważną rolę w sygnalizacji redoks w komórkach rakowych [61, 62]. Poniższa rycina prezentuje omówione zidentyfikowane geny u muszki owocowej, danio pręgowanego, kurczaka i człowieka (Ryc. 11).



Ryc. 11. Geny *CG6878* i *CG7011* nakładające się u *D. melanogaster* oraz ich ortologi z *D. rerio*, *G. gallus* i *H. sapiens*. Na rysunku zaprezentowano ułożenie genów ortologicznych wraz z konserwacją ich nałożenia lub jej brakiem jak ma to miejsce u kurczaka i człowieka.

5. DYSKUSJA

Geny nakładające się stanowią niezwykle interesujące zjawisko obserwowane zarówno w genomach organizmów prokariotycznych, jak i eukariotycznych. W tych pierwszych służą jako powszechna strategia organizacji genów w genomie, co jest związane z niewielkimi rozmiarami tych organizmów. U drugich uczestniczą w wielu procesach komórkowych, takich jak imprinting genomowy, inaktywacja chromosomu X, splicing alternatywny, wyciszanie genów i metylacja, redagowanie RNA oraz translacja. Jednakże w przypadku genomów eukariotycznych obecność w komórce podwójnej nici mRNA, jaka powstaje w wyniku transkrypcji genów tworzących nakładającą się parę, powinna być rozpoznawana przez mechanizmy obronne danego organizmu jako ingerencja wirusów, które w ten sposób integrują swój materiał genetyczny z genomem gospodarza. Pomimo tego zjawisko to nie ulegało eliminacji i obserwujemy liczne przypadki nałożeń genów, które zostały zachowane. Można zatem przypuszczać, że pary nakładających się genów pełnią ważną rolę jako jednostki i stąd ich układ nie został rozerwany w toku ewolucji.

W opublikowanych pracach w zależności od przyjętych metod i wykorzystanego rodzaju danych obserwuje się z reguły kilku lub kilkunastoprocentową frakcję genów uczestniczących w nakładaniu w stosunku do całkowitej liczby genów danego organizmu. Dla przykładu: człowiek: od 9 do 22,7%, mysz: od 7,4 do 13,67%, kurczak: od 4,8 do 11,07% [7, 63, 64].

Można zadać pytanie, czy omawiane geny i tworzone przez nie pary są ewolucyjnie stare i powstały u wspólnych przodków organizmów? Czy też może są to geny młode ewolucyjnie, które ciągle powstają i są przystosowaniem danego organizmu do jego trybu życia, cechą charakterystyczną dla danej linii ewolucyjnej?

Istnieje kilka hipotez tłumaczących pochodzenie zjawiska nakładania się genów. Keese i Gibbs [26] zasugerowali, że nakładające się geny powstały jako rezultat *overprintingu* – procesu tworzenia nowych genów z wcześniej istniejących sekwencji nukleotydowych. Zjawisko to miało miejsce po rozdzieleniu się linii ewolucyjnych ssaków i ptaków, a nakładające się geny są przykładem młodych genów kodujących białka, związanych ze specjalizacją danego organizmu do jego trybu życia. Shintani [27] stwierdził, że nakładanie powstało podczas tranzycji z ssakokształtnych gadów do ssaków. Odbywało się to albo na zasadzie rearanżacji genów, której towarzyszyła utrata części regionu 3’ UTR, wraz z sygnałem poliadenylacji albo też na drodze rearanżacji genów, które wcześniej się nie

nakładały. W drugim przypadku jeden z genów w parze utracił swój oryginalny sygnał poliadenylacji, jednakże był w stanie wykorzystać taki sygnał z nici niekodującej drugiego genu. Dahary [30] umieszcza powstanie zjawiska nakładania się genów dużo wcześniej. Swoją hipotezę popiera obserwacją dużej liczby ludzkich nakładających się genów, które zakonserwowane są u rozdymki tygrysiej (*T. rubripes*). Twierdzi, że w większości przypadków nałożeń pomiędzy dwoma genami kodującymi białka zjawisko to ograniczone jest do ich regionów niepodlegających translacji (UTR). Dodatkowo w wielu przypadkach w nakładanie zaangażowana jest alternatywna poliadenylacja, co stwarza kilka wariantów transkryptu, które różnią się w swojej długości końca 3'. Jednakże, jak pokazują badania, nawet pomiędzy blisko spokrewnionymi gatunkami nałożenia genów nie są zakonserwowane [65]. Veeramachaneni i współautorzy w swojej pracy pokazują, że na 255 przypadków, w których oba geny z pary nakładającej się u człowieka miały swoje ortologi u myszy tylko 95 par było wspólnych dla obu organizmów. Co więcej, znacząca część genów z tych par wykazuje odmienne sposoby nakładania się w obu genomach. Podobny brak konserwacji wykazał w swojej pracy Dan i współautorzy [66]. Analizowane na przykładzie różnych genomów nałożenie pomiędzy genami *MINK* i *CHRNE* pojawiło się w toku ewolucji przynajmniej trzy razy i to w sposób niezależny. Oznacza to, że wiele nałożeń może być młodych ewolucyjnie oraz wzór powiązania pomiędzy genami nie musi być zakonserwowany nawet pomiędzy blisko spokrewnionymi gatunkami.

Celem niniejszej pracy była analiza genomu muszki owocowej (*Drosophila melanogaster*) pod kątem występowania genów nakładających się oraz przetestowanie omówionych powyżej hipotez dotyczących ewolucji tego zjawiska.

W przeanalizowanym na potrzeby niniejszej pracy genomie *D. melanogaster* w stosunku do 21243 transkryptów liczba znalezionych genów uczestniczących w nakładaniu wynosi 3504, co stanowi 16,49%. Jest więc to frakcja zbliżona do tej obserwowanej w badaniach innych organizmów. Podobny wzór widzimy w liczbie par nakładających się. Znaleziona liczba 2001 par u muszki owocowej nie odbiega znacznie od tej obserwowanej u człowieka – 1766, czy myszy – 2053 [7].

Na podstawie wyników etapu drugiego, mającego na celu poszukiwanie ortologów powyższych genów u 11 gatunków muszek z rodzaju *Drosophila* można wnioskować o ich dużej konserwacji. Spośród 3504 genów zidentyfikowanych w pierwszym etapie średnio 81% (2831) z nich ma swoje ortologi u analizowanych muszek. To samo tyczy się konserwacji par. Średnio 68% (1364) z nich wykazuje konserwację, a w przybliżeniu 96%

genów tworzących pary zlokalizowanych jest na tym samym chromosomie. W rezultatach trzeciego etapu analiz, w którym badano 6 organizmów modelowych obserwujemy niski poziom występowania ortologicznych genów i tworzonych przez nie par. Z 3504 genów uczestniczących w nakładaniu u *D. melanogaster* średnio 47% (907) wykazuje konserwację, a z 2001 par zakonserwowanych jest średnio 29% (311). W przybliżeniu tylko 26% genów z zakonserwowanej pary zajmuje ten sam chromosom.

Powyższe rezultaty wskazują więc na to, że pośród genów nakładających się istnieje wiele, które uformowały się stosunkowo niedawno w ewolucji. Znaczna większość zakonserwowanych genów, jak i par przez nie tworzonych znaleziona została wśród bardzo blisko spokrewnionych gatunków. Obserwacje te sprawiają, że bardziej wiarygodna wydaje się być hipoteza wyprowadzona przez Keese'a i Gibbsa [26], w której autorzy mówią o wytwarzaniu genów nakładających się *de novo* na drodze *overprintingu*. Stąd też mechanizm ten może być w głównej mierze odpowiedzialny za powstawanie zjawiska nakładania się wskutek generowania nowych genów związanych z trybem życia oraz warunkami danego organizmu, w którym są one znajdowane.

Wśród ortologicznych par obserwujemy średnio 17,5% (238 na 1364) konserwację nałożeń w genomach 11 *Drosophil* i nieco ponad 2% (7 na 311) w 6 przeanalizowanych organizmach modelowych. Liczba zidentyfikowanych ortologicznych nałożeń w analizowanym rodzaju *Drosophila* prawdopodobnie byłaby wyższa, gdyby nie słaba adnotacja regionów UTR w ich niedawno opublikowanych genomach. Przejawia się to w stosunkowo wysokiej liczbie zakonserwowanych zagnieżdżonych nałożeń (*nested*), a bardzo niewielkiej ilości przypadków nakładania pozostałych dwóch typów (*head-to-head* i *tail-to-tail)*. U 11 gatunków *Drosophila* obserwujemy niską, około 28% frakcję zakonserwowanych nałożeń typu *nested*, co w świetle tego przykładu przemawia za stwierdzeniem, że zjawisko nakładania jest specyficzne dla danej linii rozwojowej i powstało niedawno w ewolucji.

Jedyna znaleziona para genów (*CG6878* i *CG7011*), która zakonserwowana jest zarówno u *D. melanogaster* jak i pozostałych 11 muszek, u komara, pszczoły i danio pręgowanego może stanowić przykład starego ewolucyjnie nałożenia genów. W świetle przeprowadzonych analiz jest więc jedynym dowodem potwierdzającym hipotezę autorstwa Dahary'ego [30]. Nałożenie to zostało utracone u wyższych kręgowców – *G. gallus*, *M. musculus*, *H. sapiens* – być może na drodze rozrywającej translokacji albo też w związku ze zjawiskiem efektu dawki genu (*dosage effect*). Para genów mogła zostać początkowo

zduplikowana jako całość, a następnie jeden gen z pary utracony. W ten sposób geny ortologiczne dla genów *D. melanogaster* występują w genomach myszy, czy człowieka, ale nie odtwarzają nakładania. Zlokalizowane są na tych samych chromosomach w niewielkiej odległości od siebie, rozdzielone przez kilka innych genów. Jest to dowód na to, że w toku ewolucji nałożenia pomiędzy genami mogą być nie tylko tworzone, ale również tracone.

Wiele par genów, które nakładają się u *D. melanogaster* ma swoje ortologi w innych gatunkach, jednakże tworzące je geny nie tylko się nie nakładają, ale także zlokalizowane są na różnych chromosomach. W przeważającej większości ma to miejsce w przypadku analizowanych 6 organizmów modelowych – obserwujemy niezgodności chromosomów w ilości od 298 dla *A. gambiae* do 470 dla *A. mellifera* w stosunku do pierwotnie zidentyfikowanych 2001 par u *D. melanogaster*. U 11 muszek z rodzaju *Drosophila* wartości te wahają się od 2 u *D. erecta* do 104 u *D. willistoni*. Geny takie mogły pierwotnie tworzyć nakładającą się parę, lecz w toku ewolucji prawdopodobnie podlegały translokacjom lub duplikacjom z następującym po nich usuwaniem jednego genu z pary, co spowodowało utratę nałożenia. Potwierdza to analiza 13 zakonserwowanych nałożeń u pszczoły (*A. mellifera*). Spośród nich, przykładowo, aż 9 nie występuje wśród nałożeń zidentyfikowanych u *D. ananasae*, czy *D. erecta*. Z drugiej jednak strony takie ortologiczne geny, które zajmują różne chromosomy mogły w toku ewolucji utworzyć nowe nakładające się pary na drodze translokacji. W wynikach pracy obserwujemy przypadki, gdzie ortologiczne geny pszczoły z różnych chromosomów, nie nakładają się u *D. ananassae* (chociaż leżą już na tym samym chromosomie), ale w przypadku *D. melanogaster* mamy do czynienia z współdzieleniem tego samego *locus* genomowego (np. geny *CG10984* i *CG10973* w konfiguracji *tail-to-tail*). Przykłady te potwierdzają zatem hipotezę zaproponowaną przez Shintani'ego [27] o powstawaniu par genów nakładających się na drodze translokacji.

W toku analiz dokonano sprawdzenia jaka liczba ortologicznych genów nie odtwarza nakładania u analizowanych organizmów, ponieważ brak jest drugiego genu z pary. Zjawisko takie może w pewnym stopniu wskazywać na to, że kiedyś geny te brały udział w zjawisku nakładania, przez co byłyby przykładem starych ewolucyjnie nałożeń. W toku ewolucji drugi gen z pary został utracony, a razem z nim nałożenie, w którym uczestniczył. Takiego samego sprawdzenia dokonano dla przypadków, gdzie oba geny z pary pierwotnie zidentyfikowanej u *D. melanogaster* nie mają swoich ortologów u analizowanych

organizmów. Tutaj również można wnioskować, że dane nakładania powstały u wspólnego przodka muszki i porównywanego gatunku, a w toku ewolucji zostały u nich utracone i spotykane są wyłącznie u *D. melanogaster*. Hipotezę tę potwierdzają wyniki niniejszych badań. Wybierając do porównań zakonserwowane pary z pszczoły oraz zidentyfikowane geny ortologiczne kilku z 11 muszek obserwujemy oba omówione powyżej przypadki braku konserwacji nałożeń. W przypadku *D. ananassae* mamy 58 przypadków, gdzie jeden gen z pary nie posiada swojego ortologa i 12, w których brak ortologów dla obu genów z nakładającej się pary zidentyfikowanej u *D. melanogaster*. Podobnie wygląda to w przypadku *D. erecta*, gdzie liczby te wynoszą odpowiednio: 58 i 9, *D. grimshawi*: 62 i 7, *D. mojavensis*: 74 i 10 oraz *D. persimilis*: 101 i 11. Wyniki przeprowadzonych badań pokazują zatem, że zgodnie z hipotezą narodzin i śmierci w ewolucji genów (*birth and death evolution*) [67] nałożenia, jak i same geny nakładające się mogą być nie tylko tworzone w ewolucji, ale także podlegać utracie [68].

Pomimo faktu, że wielkoskalowe badania nakładających się genów przeprowadzane były już od dawna nadal nie rozumiemy w pełni jak takie nałożenia wyewoluowały i jakie jest prawdopodobne znaczenie dzielenia tego samego *locus* genomowego przez dwa geny. Rezultaty opublikowanych do tej pory prac, jak i wyniki niniejszej analizy genów nakładających się u *D. melanogaster* i 11 innych gatunków z rodzaju *Drosophila* wskazują, że nie ma jednego uniwersalnego modelu mogącego wytłumaczyć ewolucję tego fenomenu. Każda z istniejących hipotez w pewnym stopniu może zostać odniesiona do uzyskanych rezultatów. Obserwujemy zarówno przypadki tworzenia nowych nałożeń specyficznych gatunkowo, stare nakładające się geny zakonserwowane w toku ewolucji i wspólne dla odległych sobie grup oraz liczne przykłady na to, że pierwotne nakładania genów były tracone u organizmów.

Praktycznie każde pojawienie się w genomie nowego genu, czy egzonu może spowodować powstanie pary nakładających się genów. W świetle zaprezentowanych wyników można wnioskować, że spośród wszystkich mechanizmów tłumaczących ewolucję omawianego zjawiska w największym stopniu odpowiedzialny za ten proces może być *overprinting*, dający początek nowym genom i ich nowym wariantom z wykorzystaniem wcześniej istniejących sekwencji nukleotydowych.

Niniejsza praca z pewnością może stanowić pewien wgląd w to, jak w toku ewolucji powstawały geny nakładające się, dostarcza w głównej mierze dowody na to, że nałożenia genów są raczej gatunkowo specyficzne oraz pokazuje, że nakładania mogą być wytwarzane, jak i również tracone.

6. STRESZCZENIE

Nakładające się geny definiuje się jako parę różnych genów, których regiony genomowe pokrywają się w pewnym stopniu (współdzielą to same *locus* genomowe). Zjawisko to, wykryte pierwotnie u bakteriofaga ΦX174, przez długi czas uważane było jako specyficzne dla genomów wirusowych i bakteryjnych. W organizmach tych z uwagi na ich małe rozmiary mechanizm nakładania się genów może służyć między innymi jako strategia organizacji genomu. Ukończenie sekwencjonowania genomów coraz to większej liczby organizmów eukariotycznych pozwoliło stwierdzić, że fenomen ten nie jest, jak wcześniej uważano, rzeczą dla nich rzadką. Ciągle narastająca ilość nowych dowodów sugeruje, że nakładające się geny mogą regulować kluczowe procesy ekspresji genów u *Eukariota*, wliczając w to między innymi imprinting genomowy, interferencję RNA, regulację translacji i redagowanie RNA. Jednakże pomimo dużej liczby wykrytych przypadków takich genów ich pochodzenie i ewolucja nadal pozostają niejasne.

Istnieją trzy podstawowe mechanizmy, które tłumaczą pochodzenie omawianego zjawiska. Hipoteza zasugerowana przez Keese'a i Gibbsa [26] mówi, że nakładające się geny utworzone zostały w procesie zwanym *overprintingiem*. Polega on na tworzeniu nowych genów z istniejących już wcześniej sekwencji nukleotydowych. Dlatego też jeden z genów w nakładającej się parze jest reprezentantem ewolucyjnie i filogenetycznie młodych genów kodujących białka. Co więcej, według owej hipotezy ich funkcją jest adaptacja do obecnych warunków życia danego organizmu, w który znajdowane są konkretne pary nakładających się genów. Shintani [27] w swojej hipotezie zakłada, że nakładanie powstało przy rozdzieleniu się linii ewolucyjnych ssakokształtnych gadów od ssaków. Do tworzenia nałożeń dochodziło albo na drodze rearanżacji genów wraz z utratą części regionu 3' UTR z sygnałem poliadenylacji albo też podczas rearanżacji genów, których sekwencje wcześniej się nie nakładały. Dahary [30] umiejscawia wytworzenie nakładania pomiędzy genami dużo wcześniej, a swoją hipotezę popiera obserwacją dużej liczby ludzkich nakładających się genów, które posiadają swoje ortologiczne pary u rozdymki tygrysiej (*T. rubripes*). Twierdzi, że w większości przypadków nałożeń pomiędzy dwoma genami kodującymi białka zjawisko to ograniczone jest do ich regionów niepodlegających translacji (UTR), a dodatkowo w wielu przypadkach w nakładanie zaangażowana jest alternatywna poliadenylacja.

Badania przeprowadzone na potrzeby niniejszej pracy skupiały się w głównej mierze na blisko spokrewnionych gatunkach. Do analiz wybrano 12 gatunków z rodzaju *Drosophila* a jako zestaw referencyjny wykorzystano zbiór genów nakładających się jednego z nich - *D. melanogaster*. Sprawdzano zarówno konserwację nakładających się par genów, jak i występowanie ortologicznych genów, które są członkami poszczególnych par. Analizy porównawcze wykonano na trzech poziomach reprezentujących różne dystanse ewolucyjne: dla wszystkich przedstawicieli rodzaju *Drosophila*, innych insektów (komar i pszczoła) oraz kręgowców (człowiek, mysz, kurczak i danio pręgowany).

Wykryto 3504 unikatowe geny nakładające się u *Drosophila melanogaster* tworzące 2001 niepowtarzających się par. Stanowi to około 16,5% wszystkich transkryptów znajdowanych w genomie muszki owocowej. Rozkład liczby zidentyfikowanych ortologów (zakonserwowanych genów) nie odzwierciedla układu analizowanych 12 gatunków na ich drzewie filogenetycznym. To samo tyczy się ilości wykrytych ortologicznych par genów. Jednakże na podstawie analiz można wnioskować o ich dużej konserwacji. Spośród 3504 genów nakładających się u *D. melanogaster* średnio 81% z nich ma swoje ortologi u 11 analizowanych muszek. To samo tyczy się par tworzonych przez ortologi. Średnio 68% z nich wykazuje konserwację. Rezultaty analizy 6 organizmów modelowych prezentują niski poziom występowania ortologicznych genów i tworzonych przez nie par. Średnio 47% genów i 29% par wykazuje konserwację. Analiza porównawcza zestawu ortologicznych genów zidentyfikowanych u 11 muszek oraz pszczoły, jako grupy zewnętrznej na drzewie filogenetycznym pozwoliła stwierdzić, że w toku ewolucji z pierwotnych par nakładających się jeden lub oba geny uległy zanikowi, a razem z nimi tworzone przez nie nałożenie. Zidentyfikowano również przypadki bardzo starych ewolucyjnie par nakładających się, które wspólne są dla 12 gatunków muszek, pszczoły, komara i danio pręgowanego. Geny ortologiczne dla genów uczestniczących w tworzeniu tych par występują także u człowieka, myszy i kurczaka, jednakże rozdzielone są od siebie kilkoma genami i nie odtwarzają nakładania.

Jak pokazują rezultaty badań w obrębie nakładających się genów istnieją takie, które uformowały się stosunkowo niedawno w ewolucji i nawet w obrębie tego samego rodzaju jak *Drosophila* geny te i ich pary nie są zakonserwowane. Obserwujemy także przypadki starych ewolucyjnie par nakładających się genów oraz mamy dowody na to, że pierwotne nakładania zostały utracone w wielu organizmach, co stanowi dowód dla hipotezy narodzin i śmierci w ewolucji genów [67, 68].

Nadal nie poznano w pełni mechanizmu ewolucji nałożeń genów oraz nie wiadomo jakie jest prawdopodobne znaczenie dzielenia tego samego *locus* genomowego przez dwa geny. Rezultaty tej i innych prac wskazują, że nie ma jednego uniwersalnego modelu mogącego wytłumaczyć powstanie tego fenomenu.

7. SUMMARY

Overlapping genes can be defined as a pair of different genes, which genomic regions cover to some extent (they share the same genomic *locus*). This phenomenon originally detected in bacteriophage ΦX174 was for a long time believed as unique only to viral and microbial genomes. Due to small body sizes of such organisms the mechanism of gene overlap may be a strategy for genome organization. Completion of genome sequencing of many new eukaryotic organisms established that this phenomenon is not rare for them as it had been previously believed. A still increasing number of new evidences suggests that overlapping genes can regulate key processes of gene expression in *Eukariota*, including genomic imprinting, RNA interference, translational regulation and RNA editing. Despite the large number of these genes, their origin and evolution still remain unclear.

Basically, there are three main hypotheses explaining the origination of gene overlap phenomenon. Keese and Gibbs [26] say that overlapping genes are created in an overprinting process. New genes are generated from previously existing nucleotide sequences. That is why one of the genes from overlapping pair is a representative of evolutionary and phylogenetically young protein coding genes. Their function, according to the hypothesis, is adaptation to present life style of a given organism in which particular pairs of overlapping genes are found. Shintani [27] implies that overlapping genes phenomenon arose after the divergence of therapsid reptiles from mammals. Overlaps were created during genes rearrangements accompanied by loss of part of 3' UTR containing polyadenylation signal or alternatively by rearrangements of genes which previously did not overlap. Dahary [30] places the creation of overlapping genes much earlier, and supports his hypothesis by observation of a large number of human overlapping genes having their orthologous pairs in torafugu. He claims that in most cases of overlaps between protein coding genes they only share the regions that are not translated (UTR). Additionally, in many cases overlaps employ alternative polyadenylation.

The study discussed here focused mainly on closely related species. There were 12 species from *Drosophila* genus chosen as a basic set for testing and overlapping genes set from one of them – *D. melanogaster* – served as a reference set. Both conservation of overlapping genes pairs and single genes being a member of particular pair were examined. Comparative analyses were done in three levels representing different evolutionary distances:

for all representatives of *Drosophila* genus, other insects (mosquito and bee), and vertebrates (human, mouse, chicken, zebrafish).

Originally there were 3504 unique genes found to overlap in 2001 unique genes pairs in *Drosophila melanogaster*. The number of identified genes constitutes 16,5% of fruit fly transcripts. Proportions of the conserved genes do not reflect the arrangement of 12 species on their phylogenetic tree. This also relates to the amount of orthologous genes pairs found. However, on the basis of the analyses we can imply that these genes are strongly conserved. Out of 3504 overlapping genes found in *D. melanogaster* approximately 81% have their orthologs in 11 analyzed flies. Similar pattern is observed in pairs created by these orthologs. Approximately 68% of them demonstrate conservation. Results of study of 6 model organisms present low level of orthologous genes and their pairs. On average 47% of genes and 29% of pairs are conserved. Comparative analysis of orthologous genes set identified within 11 flies and a bee as a phylogenetic tree outgroup allowed to claim that in the course of evolution one or both genes out of original ancestral overlapping pairs could be lost. There were also cases of evolutionarily very old overlapping gene pairs identified which are common to all 12 *Drosophila* species, bee, mosquito and zebrafish. The orthologs of these genes are also found in human, mouse and chicken genomes but they are separated from each other by a few other genes and do not overlap.

All the results show that many overlapping genes have been formed relatively recently and even within the same genus like *Drosophila* these genes are not conserved. There are also cases of evolutionarily old overlapping genes pairs and many observations prove that some ancient overlaps were lost in many organisms confirming the gene birth and death hypothesis [67, 68].

We still do not fully know the mechanism of evolution of gene overlap and possible meaning of sharing the same genomic *locus* between two genes. Results presented here and also in other similar studies points out that there is no single universal model explaining the origination of this phenomenon.

8. BIBLIOGRAFIA

1. Shine, I., Wrobel, S.: *Thomas Hunt Morgan: Pioneer of Genetics*. Lexington: University of Kentucky Press, 1976.

2. Sanger F. et al.: *Nucleotide sequence of bacteriophage ΦX174 DNA*. Nature 1977, 265(5596):687-95.

3. Barrell B.G., Air G.M., Hutchison III C.A.: *Overlapping genes in bacteriophage ΦX174*. Nature 1976, 264:34-41.

4. Williams T., Fried M.: *A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3' ends*. Nature 1986, 322(6076):275-9.

5. Spencer C.A., Gietz R.D., Hodgetts R.B.: *Overlapping transcription units in the dopa decarboxylase region of Drosophila*. Nature 1986, 322(6076):279-81.

6. Boi S., Solda' G., Tenchini M.L.: *Shedding light on the dark side of the genome: overlapping genes in higher eukaryotes*. Current Genomics 2004, 5:509-524.

7. Makałowska I., Lin C.F., Makałowski W.: *Overlapping genes in vertebrate genomes*. Comput. Biol. Chem. 2005, 29(1):1-12.

8. The ENCODE Project Consortium: *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project*. Nature 2007, 447(7146):799-816.

9. Carninci P. et al.: *The transcriptional landscape of the mammalian genome*. Science 2005, 309(5740):1559-1563.

10. Cheng J. et al.: *Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution*. Science 2005, 308(5725):1149-1154.

11. Engstrom P.G. et al.: *Complex Loci in human and mouse genomes*. PLoS Genet. 2006, 2(4):e47.

12. Kapranov P. et al.: *Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays*. Genome Res. 2005, 15(7):987-997.

13. Katayama S. et al.: *Antisense transcription in the mammalian transcriptome*. Science 2005, 309(5740):1564-1566.

14. Lapidot M., Pilpel Y.: *Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms*. EMBO Rep. 2006, 7(12):1216-1222.

15. Li A.W., Murphy P.R.: *Expression of alternatively spliced FGF-2 antisense RNA transcripts in the central nervous system: regulation of FGF-2 mRNA translation*. Mol. Cell. Endocrinol. 2000, 170(1-2):233-242.

16. Munroe S.H., Lazar M.A.: *Inhibition of c-erbA mRNA splicing by a naturally occurring antisense RNA*. J. Biol. Chem. 1991, 266(33):22083-22086

17. Peters N.T. et al.: *RNA editing and regulation of Drosophila 4f-rnp expression by sas-10 antisense readthrough mRNA transcripts*. Rna 2003, 9(6):698-710.

18. Sleutels F., Zwart R., Barlow D.P.: *The non-coding Air RNA is required for silencing autosomal imprinted genes*. Nature 2002, 415(6873):810-813.

19. Tufarelli C. et al.: *Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease*. Nat. Genet. 2003, 34(2):157-165.

20. Britten R.J., Davidson E.H.: *Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty*. Q. Rev. Biol. 1971, 46:111-138.

21. Langridge J.: *Molecular Genetics and Comparative Evolution*. Research Studies Press, Taunton, Somerset, England 1991.

22. Kavaler J., Davis M.M., Chien Y.: *Localization of a T-cell receptor diversity-region element.* Nature 1984, 310(5976):421-3.

23. Biebricher C.K., Eigen M., Luce R.: *Template-free RNA synthesis by Q beta replicase.* Nature 1986, 321(6065):89-91.

24. Grassé P.P.: *Evolution of Living Organisms*. Academic Press, New York 1977, p. 297.

25. Evans R.M.: *The steroid and thyroid hormone receptor superfamily.* Science 1988, 240(4854):889-95.

26. Keese P.K., Gibbs A.: *Origins of genes: "big bang" or continuous creation?* Proc. Natl. Acad. Sci. U.S.A. 1992, 89:9489–9493.

27. Shintani S. et al.: *Origin of gene overlap: the case of TCP1 and ACAT2*. Genetics 1999, 152:743–754.

28. Nekrutenko A. et al.: *Oscillating evolution of a mammalian locus with overlapping reading frames: An XLas/ALEX relay*. PLoS Genet 2005, 1(2): e18.

29. http://en.wikipedia.org/wiki/T-complex_1

30. Dahary D., Elroy-Stein O., Sorek R.: *Naturally occurring antisense: transcriptional leakage or real overlap?* Genome Res. 2005, 15:364–368.

31. Iseli C. et al.: *Long-range heterogeneity at the 3 ends of human mRNAs*. Genome Res. 2002, 12:1068–1074.

32. Connelly S., Manley J.L.: *A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II*. Genes & Dev. 1988, 2:440–452.

33. Batt D.B., Luo Y., Carmichael G.G.: *Polyadenylation and transcription termination in gene constructs containing multiple tandem polyadenylation signals*. Nucleic Acids Res. 1994, 22:2811–2816.

34. http://pl.wikipedia.org/wiki/Muszka_owocowa

35. http://en.wikipedia.org/wiki/Drosophila_melanogaster

36. Adams M.D. et al.: *The genome sequence of Drosophila melanogaster*. Science 2000, 287:2185–2195.

37. Markow T.A., O'Grady P.M.: *Drosophila biology in the genomic age.* Genetics 2007, 177:1269-1276.

38. Powell J.R.: *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. Oxford Univ. Press 1997.

39. Drosophila 12 Genomes Consortium: *Evolution of genes and genomes on the Drosophila phylogeny*, Nature 2007, 450:203-218.

40. Richards S. et al.: *Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution*. Genome Res. 2005, 15:1–18.

41. Henikoff S. et al.: *Gene within a gene: nested Drosophila genes encode unrelated proteins on opposite DNA strands.* Cell 1986, 44(1):33-42.

42. Schulz R.A., Butler B.A.: *Overlapping genes of Drosophila melanogaster: Organization of the z600-gonadal-Eip28/29 gene cluster*. Genes & Dev. 1989, 3:232–242.

43. Pápai G.: *Study of two overlapping genes in Drosophila melanogaster.* Acta Biol. Szegediensis 2002, 46(1-2):43.

44. Kim D.S. *et al*., *EVOG: a database for evolutionary analysis of overlapping genes*. Nucleic Acids Res. 2009, 37:698-702.

45. Jiang L.W., Lin K.L., Lu C.L.: *OGtree: a tool for creating genome trees of prokaryotes based on overlapping genes.* Nucleic Acids Res. 2008, 36(suppl_2):W475-W480.

46. Yingqin L. et al.: *BPhyOG: An interactive server for genome-wide inference of bacterial phylogenies based on overlapping genes*. BMC Bioinformatics 2007, 8:266.

47. http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html

48. Giardine B. et al.: *Galaxy: a platform for interactive large-scale genome analysis.* Genome Res. 2005, 15(10):1451-5.

49. Karolchik D. et al.: *The UCSC Table Browser data retrieval tool*. Nucleic Acids Res. 2004, 32:D493–D496.

50. Yang Z., Nielsen R.: *Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models*. Mol. Biol. Evol. 2000, 17: 32–43.

51. Li W.H.: *Molecular evolution*. Sinauer Associates 1997.

52. http://pl.wikipedia.org/wiki/CGI

53. http://genome.ucsc.edu/FAQ/FAQformat#format1

54. Altschul S.: Basic Local Alignment Search Tool. J. Mol. Bio. 1990, 215(3):403-410.

55. Higgs P.G., Attwood T.K.: *Bioinformatyka i ewolucja molekularna*. PWN 2008, 213-215.

56. http://flyrnai.org/RNAi_orthology.html

57. http://en.wikipedia.org/wiki/Python_(programming_language)

58. http://pl.wikipedia.org/wiki/Python

59. Gruszecki P.: *Wprowadzenie do Pythona - język oprogramowania inny niż wszystkie*. http://www.internetmaker.pl/artykul/3423,1.html, 2008.

60. Nishikawa M. et al.: *Identification and characterization of endoplasmic reticulum-associated protein, ERp43*. Gene 2007, 386(1-2):42-51.

61. Chung Y.M. et al.: *Replicative senescence induced by Romo1-derived reactive oxygen species.* J Biol Chem. 2008, 283(48):33763-71.

62. Na A.R. et al.: *A critical role for Romo1-derived ROS in cell proliferation.* Biochem. Biophys. Res. Commun. 2008, 369(2):672-8.

63. Sanna C.R., Li W.H., Zhang L.: *Overlapping genes in the human and mouse genomes.* BMC Genomics 2008, 9:169.

64. Sun M. et al.: *Evidence for variation in abundance of antisense transcripts between multicellular animals but no relationship between antisense transcriptionand organismic complexity.* Genome Res. 2006, 16(7):922-33.

65. Veeramachaneni V. et al.: *Mammalian overlapping genes: the comparative perspective.* Genome Res. 2004, 14(2):280-6.

66. Dan I. et al.: *Overlapping of MINK and CHRNE gene loci in the course of mammalian evolution*. Nucleic Acids Res. 2002, 30(13):2906–2910.

67. Nei M., Rooney A.P.: *Concerted and birth-and-death evolution of multigene families*. Annu. Rev. Genet. 2005, 39:121-52.

68. Makałowska I., Lin C.F, Hernandez K.: *Birth and death of gene overlaps in vertebrates.* BMC Evol Biol. 2007, 7:193.

9. SPIS ILUSTRACJI

Ryc. 1. Ogólny schemat sposobu nakładania się genów, skategoryzowany w trzy główne typy. 8

Ryc. 2. Zrzut ekranowy interfejsu internetowego GALAXY. 20

Ryc. 3. Przykład danych wyjściowych z GALAXY zawierających powtórzenia nałożeń spowodowane wariantami splicingowymi tego samego genu. 26

Ryc. 4. Schemat analizy i obróbki danych w drugim etapie pracy zaprezentowany na przykładzie pary nakładających się genów z *D. melanogaster* w orientacji *tail-to-tail – CG10018* i *CG2929*.. 28

Ryc. 5. Schemat postępowania w metodzie *reciprocal* BLAST na przykładzie *D. melanogaster* i *M. musculus*. 30

Ryc. 6. Schemat identyfikacji ortologów dla par genów nakładających się z *D. melanogaster* oraz weryfikacji, czy odtwarzają one dane nałożenie. 31

Ryc. 7. Gen *CG15221* i trzy geny w nim zagnieżdżone: *CG34322, CG34321, CG2457*. 32

Ryc. 8. Typy analizowanych przypadków konserwacji par nakładających się genów. 34

Ryc. 9. Reprezentacja wyników analizy nakładających się genów wśród 12 gatunków z rodzaju *Drosophila* na jego drzewie filogenetycznym. 35

Ryc. 10. Reprezentacja wyników analizy nakładających się genów u 6 organizmów modelowych w formie drzewa rodowego badanych organizmów. 38

Ryc. 11. Geny *CG6878* i *CG7011* nakładające się u *D. melanogaster* oraz ich ortologi z *D. rerio*, *G. gallus* i *H. sapiens*. 39

10. SPIS TABEL

Tab. 1. Wyniki pierwszego etapu analiz nakładających się genów w genomie *D. melanogaster*. 32

Tab. 2. Wyniki analizy typu nakładania się dla genów *D. melanogaster* w pierwszym etapie analiz. 32

Tab. 3. Wynik analiz genów nakładających się dla 11 gatunków muszek z rodzaju *Drosophila*. 33

Tab. 4. Wynik analizy braku konserwacji genów ortologicznych oraz par nakładających się genów. 34

Tab. 5. Wyniki analizy sposobu nakładania się genów u 11 gatunków muszek z rodzaju *Drosophila*..... 35

Tab. 6. Wyniki analizy ortologów oraz par nakładających się genów u 6 modelowych organizmów. 36

Tab. 7. Wyniki analizy braku konserwacji genów ortologicznych oraz par nakładających się genów u 6 organizmów modelowych. 37

Tab. 8. Wynik analiz sposobu nakładania się genów dla 6 organizmów modelowych. 37