Streszczenie rozprawy doktorskiej pod tytułem „Modelowanie ewolucji genomów”*

Damian Wójtowicz

1 Wprowadzenie

Znajomość sekwencji genomowej dla coraz większej liczby organizmów przybliża nas do zrozumienia organizacji i ewolucji genomów. Podstawowym jej mechanizmem jest duplikacja genów [10]. Proces ten prowadzi do powstania redundancji koniecznej aby umożliwić jednej z kopii genu wytworzenie nowej funkcji korzystnej dla organizmu. Dwa genu, które powstały w wyniku zajścia zdarzenia duplikacji nazywamy *paralogami*. Geny te możemy pogrupować w maksymalne rodziny genów wzajemnie paralogicznych, które nazywamy *rodzinami paralogów*.

W ostatnich latach intensywnie badano ewolucję rodzin genów, szczególnie w genomice porównawczej [8]. Nie jest to zaskakujące biorąc pod uwagę fakt, że ponad połowa genów posiada wykrywalny gen paralogiczny w tym samym genomie [13], a porównywanie rodzin paralogów w genomach różnych gatunków pozwala wnioskować o ich ewolucji [1]. Genomy posiadają rodziny paralogów o różnych rozmiarach i te rodziny, jak i ich rozmiary, mogą się zmieniać w czasie. W wyniku duplikacji genów rodziny się powiększają, a straty genów powodują zmniejszanie się rodzin. Nowe geny dla danego gatunku (nieparalogiczne do istniejących), a stąd i nowe rodziny paralogów, mogą się pojawiać w wyniku horyzontalnego transferu materiału genetycznego lub odzyskania genów z części niekodującej DNA. Taką możliwość pojawienia się genu nazywamy *innowacją*. Punktowe zmiany sekwencji DNA (mutacje), a także różnego rodzaju rearanżacje genomu, mogą doprowadzić do tak znacznych zmian w sekwencji genu, że nie jest on już dłużej podobny do swego przodka i dlatego nie jest już rozpoznawany jako paralog[1]. Zdarzenie takie nazywamy *(zakumulowaną) zmianą genu*. Ono również prowadzi do powstania nowych rodzin genów w genomie. Pomimo ciągłych zmian zachodzących w rodzinach paralogów, rozkład ich rozmiarów, który będziemy również nazywać *rozkładem paralogów*, wydaje się być niezmienny [13, 5, 4, 19].

*Nazwy i kolejność rozdziałów w niniejszym streszczeniu zostały zmienione w stosunku do rozprawy.

[1]Paralogi są rozpoznawane głównie na podstawie podobieństwa sekwencyjnego, a nie prześledzenia historii ewolucji, o której nie mamy dokładnych informacji.

W rozprawie przedstawiamy dyskretny model ewolucji genomu w duchu prac Kimury [7], tj. przy zupełnym braku presji selekcyjnej. Zdajemy sobie sprawę, że taki model nie może być w pełni realistyczny, ale wierzymy jednak, że stanowi on istotny wkład w dyskusję nad problemem modelowania ewolucji genomów. Nasz model opisuje dynamikę genomu na poziomie genów i opiera się czterech wspomnianych już podstawowych zdarzeniach ewolucyjnych: duplikacji, stracie, zmianie i innowacji genów. Model ten nazywamy *modelem DLCI* (*ang. duplication, loss, change and innovation* model). W rozprawie skupiamy się na matematycznej analizie tego modelu z punktu widzenia rozkładu rozmiarów rodzin genów paralogicznych. Jest to pierwsza taka analiza modelu opartego na tych czterech zdarzeniach ewolucyjnych.

Problem analizowania rozkładów paralogów jest ważny, ponieważ określone kształty tych rozkładów są ściśle związane z modelami ewolucji. Badając te rozkłady możemy poznawać parametry ewolucji, np. prędkości zdarzeń ewolucyjnych czy określać które zdarzenia ewolucyjne są istotne, a które nie mają dużego wpływu na ewolucję. Aby osiągnąć ten cel, musimy zbadać powiązania między rozkładami paralogów a modelami ewolucji genomów. Zatem, ważne jest aby przeanalizować teoretycznie możliwie jak największą klasę modeli. Pozwoli nam to określić pewne parametry ewolucji istotne dla tego procesu.

2 Wcześniejsze badania i motywacje

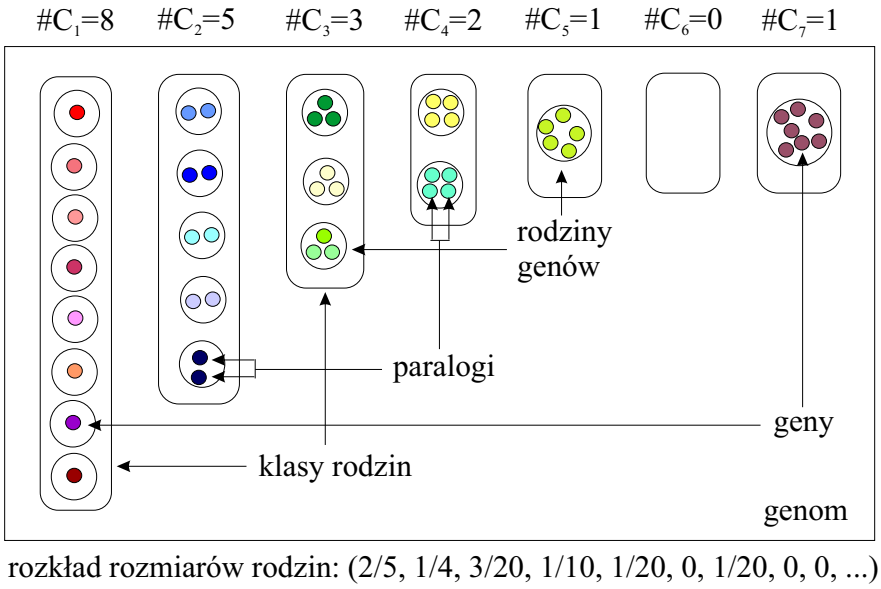
Motywacje tej pracy są związane z badaniami przeprowadzonymi w późnych latach dziewięćdziesiątych XX w., które analizowały rozkład paralogów w kilkunastu genomach mikrobów. Słonimski *et al.* [13] oraz niezależnie Huynen i van Nimwegen [4] policzyli i-genowe rodziny paralogów ($i = 1, 2, 3, \ldots$) w genomach, które były wówczas zsekwencjonowane i wyciągnęli dwa różne wnioski dotyczące kształtu zaobserwowanego rozkładu. Pierwsi stwierdzili, że rozkład ten jest rozkładem logarytmicznym ($P(i) = C\frac{\theta^i}{i}$, gdzie $0 < \theta < 1$, a $C = (-\ln(1-\theta))^{-1}$), a drudzy, że rozkładem potęgowym ($\Pr(i) = Di^{-\gamma}$, gdzie $\gamma > 1$, a $D = (\sum_{j \geqslant 1} j^{-\gamma})^{-1}$). Podobne badania przeprowadzili jeszcze Jordan *et al.* [5] oraz Yanai *et al.* [19], ale również nie były one jednoznaczne.

Powyższe wątpliwości co do kształtu rozkładu paralogów w genomach doprowadziły nas do wniosku, że sama analiza danych biologicznych jest niewystarczająca i należy zaproponować model ewolucji rodzin genów paralogicznych w genomie oraz przeanalizować go pod kątem rozkładu ich rozmiarów.

W tym samym czasie dwie inne grupy badawcze postawiły dokładnie taki sam problem [6, 11]. Jednak ich modele bazują tylko na trzech wspomnianych we wstępie zdarzeniach ewolucyjnych: w jednym brak zmiany genów, a w drugim innowacji genów. Niniejsza praca jest pierwszą analizą podejścia opartego na wszystkich czterech zdarzeniach.

3 Model DLCI

Obiektami podstawowymi w naszym modelu są *geny*, które traktujemy jako niezależnie ewoluujące i niepodzielne jednostki *genomu*, który jest zwykłym zbiorem genów. Aby zachować historię ewolucji rodzin paralogów zakładamy, że każdy gen posiada swój własny kolor. Intuicja stojąca za kolorami jest taka, że geny mają ten sam kolor wtedy, i tylko wtedy, gdy są *paralogami*. Co więcej, monochromatyczne rodziny genów odpowiadają *rodzinom paralogów*. Zatem genom dzieli się w sposób naturalny na rodziny genów ze względu na kolory. Dla $i \geqslant 1$, niech $\mathfrak{C}_i$ będzie kolekcją wszystkich i-genowych rodzin paralogów. Zbiór $\mathfrak{C}_i$ będziemy nazywać *i-tą klasą rodzin*, a jego rozmiar oznaczać przez $\#\mathfrak{C}_i$. W pracy tej badamy *rozkład paralogów* w genomie (rozkład rozmiarów rodzin genów paralogicznych), który jest zdefiniowany jako $(\pi_i)_{i\geqslant 1}$, gdzie $\pi_i = \frac{\#\mathfrak{C}_1}{\sum_{k\geqslant 1}\#\mathfrak{C}_k}$ jest prawdopodobieństwem zaobserwowania rodziny rozmiaru i w genomie. Na rysunku 1 prezentujemy zdefiniowane tu pojęcia na przykładowym genomie.



Rysunek 1: Przykładowy genom i ilustracja przedstawionych pojęć.

Teraz wspomniane we wstępie elementarne zdarzenia ewolucyjne możemy zdefiniować następująco:
- *duplikacja genu* – zdarzenie w wyniku którego w genomie powstaje identyczna kopia duplikowanego genu, czyli powstaje nowy gen i dziedziczy kolor po genie, który został poddany zdarzeniu duplikacji. Duplikacja genu pochodzącego z rodziny klasy $\mathfrak{C}_i$ przenosi tę rodzinę do klasy $\mathfrak{C}_{i+1}$ $(i \geqslant 1)$.
- *usunięcie genu* – zdarzenie, które prowadzi do usunięcia genu z puli genów

genomu. Dla $i > 1$, usunięcie genu z rodziny należącej do klasy $\mathfrak{C}_i$, przenosi tę rodzinę do klasy $\mathfrak{C}_{i-1}$; usunięcie genu z rodziny jednoelementowej całkowicie eliminuje także tę rodzinę z genomu.
- *zmiana genu* – zdarzenie, które prowadzi do zmiany koloru genu na kolor nieobecny w genomie. Pojawia się nowa jednoelementowa rodzina składająca się tylko z tego genu. Jest on jednocześnie usuwany z rodziny do której wcześniej należał.
- *innowacja genu* – zdarzenie, które wprowadza nowe geny do genomu. Zakładamy, że kolory wprowadzonych genów są parami różne. Zatem każdy gen wprowadza (zakłada) nową jednoelementową rodzinę. Rozmiar klasy $\mathfrak{C}_1$ zwiększa się o liczbę wprowadzonych genów.

Model DLCI ewolucji genomu jest dyskretnym łańcuchem Markowa. Stanami są nieskończone ciągi $(s_i)_{i\geqslant 1}$ liczb naturalnych, z których tylko skończenie wiele jest niezerowych. Stan $(s_i)_{i\geqslant 1}$ reprezentuje genom $\mathfrak{G}$, w którym dla każdego $i \geqslant 1$, liczba i-genowych rodzin paralogów wynosi s_i, tzn. $\#\mathfrak{C}_i = s_i$. Stanem początkowym w modelu DLCI jest genom składający się z $K > 0$ różnokolorowych genów, czyli stan $(K, 0, 0, \ldots)$.

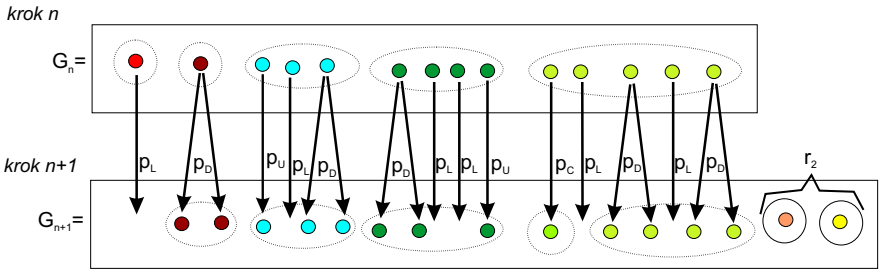
W modelu DLCI, proces ewolucji składa się z dwóch niezależnych podprocesów: *podprocesu wewnętrznego* i *podprocesu zewnętrznego*. Zakładamy, że w każdym kroku procesu ewolucji najpierw zachodzi jeden krok podprocesu wewnętrznego, a następnie jeden krok podprocesu zewnętrznego. Opiszemy oba te podprocesy osobno.

Podproces wewnętrzny (proces duplikacji, straty i zmiany genu): Podproces ten opisuje wewnętrzne zmiany zachodzące w genomie, tj. migrację rodzin pomiędzy klasami rodzin (lub ich zanikanie) oraz formowanie się nowych rodzin poprzez zdarzenie zmiany genu. Jest on parametryzowany trzema dodatnimi liczbami rzeczywistymi: p_D, p_L i p_C, spełniającymi warunek: $p_D + p_L + p_C < 1$, które są prawdopodobieństwami odpowiednich zdarzeń ewolucyjnych. W jednym kroku podprocesu, każdy gen aktualnego genomu jest niezależnie od pozostałych genów poddany jednemu z poniższych zdarzeń:

- duplikacji z prawdopodobieństwem $p_D > 0$,
- usunięciu z prawdopodobieństwem $p_L > 0$,
- zmianie z prawdopodobieństwem $p_C > 0$,
- lub pozostaje bez zmian z prawdopodobieństwem $p_U = 1 - pd + p_L + p_C) > 0$.

Podproces zewnętrzny (proces innowacji): Zakładamy, że mamy zewnętrzne źródło genów (*czarna skrzynka*), które w każdym kroku podprocesu umożliwia wprowadzenie do genomu $\mathfrak{G}$ pewnej losowej liczby genów. Liczba wprowadzonych genów jest opisana przez pewien rozkład prawdopodobieństwa $(r_i)_{i\geqslant 0}$ o wartości oczekiwanej $\tau = \sum_{i=0}^{\infty} i r_i > 0$. Oznacza to, że w jednym kroku ewolucji proces innowacji wprowadza do genomu i genów z prawdopodobieństwem r_i, a ich średnia liczba wynosi τ. Zakładamy również, że wprowadzone geny mają całkiem nowe kolory, tj. kolory nie występujące w genomie $\mathfrak{G}$, a co więcej kolory wprowadzonych genów są parami różne.

Jeden krok procesu ewolucji w modelu DLCI dla przykładowego genomu został przedstawiony na rysunku 2. Wprowadzenie i analizę modelu DLCI można także znaleźć w naszych pracach [16, 18].



Rysunek 2: Przykładowy krok ewolucji w modelu DLCI: $\mathfrak{G}_n \rightsquigarrow \mathfrak{G}_{n+1}$, czyli $(2,0,1,1,1,0,0,\ldots) \rightsquigarrow (3,1,2,1,0,0,\ldots)$.

Analiza rozkładu paralogów w modelu DLCI dotyczy zachowania w przypadku średnim, więc zajmujemy się wartościami oczekiwanymi rozmiarów rodzin. Wyprowadzamy zależność rekurencyjną dla tych wartości oczekiwanych, a następnie pokazujemy istnienie asymptotycznego rozkładu paralogów, przy liczbie kroków ewolucji dążącej do nieskończoności. Podajemy również równanie funkcyjne jednoznacznie charakteryzujące funkcję tworzącą tego rozkładu. Następnie pokazujemy, że jeśli prawdopodobieństwa rozważanych zdarzeń ewolucyjnych proporcjonalnie dążą do zera, to rozkład prawdopodobieństwa zaobserwowania rodziny określonego rozmiaru ma rozkład bliski rozkładowi logarytmicznemu. Dokładniej, otrzymujemy następujące twierdzenie dla modelu DLCI

Twierdzenie 1. *Niech p_D, p_L i p_C będą odpowiednio niezerowymi prawdopodobieństwami duplikacji, usunięcia i zakumulowanej zmiany genu, a $\tau > 0$ średnią liczbą genów dostarczanych do genomu przez proces innowacji. Wówczas, jeśli $p_L > p_D$, to dla dostatecznie małych wartości prawdopodobieństw p_D, p_L i p_C oraz dostatecznie dużej liczby kroków procesu ewolucji prawdopodobieństwo zaobserwowania rodziny rozmiaru $i \geqslant 1$ w genomie jest bliskie*

$C \cdot \theta^i / i$ oraz nie zależy od wartości τ, gdzie $\theta = \frac{p_D}{p_L + p_C}$, a C jest stałą normującą rozkład logarytmiczny. Ponadto, rozmiar genomu jest bliski $\tau/(p_L - p_D)$. Wszystkie powyższe własności nie zależą od początkowego rozmiaru genomu, pod warunkiem że na początku wszystkie genu są różnokolorowe (nieparalogiczne).

4 Inne modele

W rozprawie przedstawiamy również prostsze modele, oparte na mniejszej liczbie zdarzeń ewolucyjnych. Traktujemy je jednak głównie jako krok pośredni podczas analizy modelu DLCI. Szczególnie, najprostszy model DL (tylko duplikacje i straty genów), dla którego pokazujemy, że rozkład paralogów jest rozkładem geometrycznym, co jednak nie ma dużego odzwierciedlenia w faktycznie obserwowanych rozkładach paralogów w rzeczywistych genomach. Natomiast model DLC jest modelem, w którym nie ma zdarzenia innowacji i opiera się tylko na podprocesie wewnętrznym modelu DLCI. Może to być model dla organizmów z rzadko występującym transferem horyzontalnym i szybkim tempem mutacji w sekwencji genów. Okazuje się, że wprowadzenie zmiany genów do modelu DL prowadzi już do logarytmicznych rozkładów paralogów. Kroki wykonywane podczas analizy modeli DL, DLC i DLCI są analogiczne, a samą analizę przeprowadzamy od modelu najprostszego. Pozwala nam to wykorzystywać wyniki uzyskane dla prostszych modeli podczas analizy modelu DLCI. Przypadki szczególne modelu DLCI zostały wprowadzone i przeanalizowane w naszych wcześniejszych publikacjach [14] (model DL), [15] (model DLC), [12] (ciągłoczasowa wersja modelu DLC, nieopisywana w rozprawie, w której rozważamy tylko modele dyskretne) oraz [17] (model DLI oparty na duplikacjach, stratach i zmianach genów, który także został pominięty w rozprawie ze względu na analogiczną analizę do przedstawionej w pracy analizy modelu DLCI).

W rozprawie rozważamy również ogólniejszy model ewolucji oparty na niezależnej ewolucji grup rodzin paralogów. Zachowanie każdej z grup jest opisywane przez model DLCI z indywidualnymi prawdopodobieństwami ewolucyjnymi. Zakładamy, że w genomie mamy $M > 0$ grup rodzin, a każda rodzina należy do jednej, i tylko jednej, grupy rodzin. Zatem model ten jest rozłączną sumą M modeli DLCI i nazywamy go modelem M-DLCI.

Motywacją dla wprowadzenia modelu M-DLCI była obserwacja, że geny i rodziny genów ewoluują z różnymi prędkościami ewolucyjnymi [9, 2, 3]. Geny odpowiedzialne za ważne procesy życiowe organizmu prawdopodobnie nie wykazują skłonności do szybkich zmian z powodu silnej presji selekcyjnej, np. histony u wyższych organizmów. Z drugiej strony kinaza tyrozynowa zmienia się dość szybko [9]. Zatem, naturalnym wydaje się założenie, że istnieją grupy rodzin, które ewoluują w różnym tempie, tzn. z różnymi prawdopodobieństwami zdarzeń ewolucyjnych.

Dla modelu M-DLCI pokazaliśmy poniższy, dość naturalny, wynik:

Twierdzenie 2. *Niech w każdej grupie rodzin prawdopodobieństwo duplikacji genu będzie mniejsze niż prawdopodobieństwo straty. Wtedy dla dostatecznie małych wartości prawdopodobieństw zdarzeń ewolucyjnych oraz dostatecznie dużej liczby kroków procesu ewolucji rozkład paralogów jest bliski kombinacji liniowej rozkładów w każdej z grup z wagami równymi frakcji rodzin w odpowiedniej grupie wśród wszystkich rodzin w genomie.*

5 Wyniki eksperymentalne

Pod koniec rozprawy przedstawiamy wyniki weryfikacji naszych modeli. Rozkłady rozmiarów rodzin genów w modelach DLC, DLCI i M-DLCI, dla wartości parametrów biologicznie uzasadnionych, są rozkładami logarytmicznymi lub ich kombinacjami. Prezentujemy wyniki optymalnego dopasowania tych rozkładów do danych rzeczywistych. Obserwowane rozkłady rozmiarów rodzin pochodzą z 17 genomów bakteryjnych i 5 genomów drożdżowych. Analizujemy dopasowanie zarówno dla małych rodzin genów, jak i dla wszystkich rodzin genów z rozważanych genomów. Dla porównania, przedstawiamy również analogiczne wyniki dopasowania konkurencyjnego rozkładu do tych samych danych, rozkładu potęgowego postulowanego w modelach Kareva *et al.* [6] oraz Reeda i Hughesa [11]. Badamy także wpływ metody rozpoznawania rodzin genów paralogicznych na rozkład ich rozmiarów.

Pokazaliśmy, zakładając istnienie dwóch grup rodzin genów z indywidualnymi parametrami ewolucyjnymi, że liniowa kombinacja rozkładów logarytmicznych może właściwie opisywać obserwowane rozkłady. Co więcej, umożliwia ona podział rodzin genów na małe rodziny szybko ewoluujące i duże rodziny silnie konserwowane. Jednakże, zdajemy sobie sprawę, że czasami jest trudno znaleźć statystycznie istotną różnicę między naszym a konkurencyjnymi rozkładami, gdyż liczba rodzin w genomach jest zbyt mała. Zatem, dalsze badania w tej tematyce są nadal potrzebne. Ważne, że nawet proste modele ujawniają pewne potencjalnie ciekawe własności procesu ewolucji i wyznaczają interesujące kierunki dalszych badań.

Literatura

[1] M. Csűrös, I. Miklós. A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. A. Apostolico *et al.*, redaktor, *RECOMB*, wolumen 3909 serii *Lecture Notes In Bioinformatics*, strony 206–220, 2006.

[2] Josefa Gonzalez, Jose Maria Ranz, Alfredo Ruiz. Chromosomal Elements Evolve at Different Rates in the Drosophila Genome. *Genetics*, 161(3):1137–1154, 2002.

[3] Matthew W. Hahn, Tijl De Bie, Jason E. Stajich, Chi Nguyen, Nello Cristianini. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research*, 15:1153–1160, 2005.

[4] M.A. Huynen, E. van Nimwegen. The frequency distribution of gene family size in complete genomes. *Molecular Biology Evolution*, 15(5):583–589, 1998.

[5] K. Jordan, K.S. Makarova, J.L. Spouge, Y.I. Wolf, E.V. Koonin. Lineage-specific gene expansions in bacterial and archeal genomes. *Genome Research*, 11:555–565, 2001.

[6] G.P. Karev, Y.I. Wolf, A.Y. Rzhetsky, F.S. Berezovskaya, E.V. Koonin. Birth and death of protein domains: A simple model of evolution explains power law behavior. *BMC Evolutionary Biology*, 2:18, 2002.

[7] M. Kimura. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, 1983.

[8] E.V. Koonin, M. Galperin. *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics.* Kluwer Academic Publishers, 2002.

[9] H. Luz, M. Vingron. Family specific rates of protein evolution. *Bioinformatics*, 22(10):1166–1171, 2006.

[10] S. Ohno. *Evolution by Gene Duplication.* Springer Verlag, Berlin, 1970.

[11] W.J. Reed, B.D. Hughes. A model explaining the size distribution of gene and protein families. *Mathematical Biosciences*, 189(1):97–102, 2004.

[12] R. Rudnicki, J. Tiuryn, D. Wójtowicz. A model for the evolution of paralog families in genomes. *Journal of Mathematical Biology*, 53(5):759–770, 2006.

[13] P.P. Slonimski, M.O. Mosse, P. Golik, A. Henaût, Y. Diaz, J.L. Risler, J.P. Comet, J.C. Aude, A. Wozniak, E. Glemet, J.J. Codani. The first laws of genomics. *Microbial and Comparative Genomics*, 3(46), 1998.

[14] J. Tiuryn, R. Rudnicki, D. Wójtowicz. A case of genome evolution: from continous to discrete time model. *Proc. Mathematical Foundations of Computer Science MFCS'04*, wolumen 3153, strony 1–24, 2004.

[15] J. Tiuryn, D. Wójtowicz, R. Rudnicki. A discrete model of evolution of small paralog families. *Mathematical Models and Methods in Applied Sciences*, (zaakceptowany), 2007.

[16] D. Wójtowicz, J. Tiuryn. On genome evolution with accumulated change and innovation. *Proceedings of RECOMB Comparative Genomics Satellite Workshop*, wolumen 4205, strony 40–50, 2006.

[17] D. Wójtowicz, J. Tiuryn. On genome evolution with innovation. *Proceedings of Mathematical Foundations of Computer Science*, wolumen 4162, strony 801–811, 2006.

[18] D. Wójtowicz, J. Tiuryn. Evolution of gene families based on gene duplication, loss, accumulated change and innovation. *(manuskrypt zgłoszony do czasopisma)*, 2007.

[19] I. Yanai, C.J. Camacho, C. DeLisi. Predictions of gene family distributions in microbial genomes: Evolution by gene duplication and modification. *Physical Review Letters*, 85(12):2641–2644, 2000.