Małgorzata Dudkiewicz

Modelowanie presji mutacyjnej i selekcyjnej w genomie prokariotycznym

Praca doktorska wykonana
pod kierunkiem
prof. Stanisława Cebrata
w Zakładzie Genomiki
Instytutu Genetyki i Mikrobiologii
Uniwersytetu Wrocławskiego

Wrocław 2004

Spis treści:

Z uwagi na złożoność problemów będących punktem wyjścia, a zarazem podstawą teoretyczną poniższej pracy, uznano za zasadne zamieszczenie krótkiego streszczenia zawierającego najważniejsze przesłanki, na których oparto zarówno samą konstrukcję algorytmu symulacyjnego, jak i schemat przeprowadzonego doświadczenia. Zawarty w Streszczeniu zasób informacji jest wystarczający, jeżeli chodzi o zrozumienie zasadniczej treści pracy, natomiast rozwinięcie głównych punktów oraz podstawy teoretyczne i opisy badań eksperymentalnych prowadzących do sformułowania jej podstawowych założeń zamieszczono we Wprowadzeniu (str. 5-36). Objaśnienia i opisy metod i terminów technicznych przeniesiono do załączonego Słowniczka (str.90-92) i Aneksu I (str.78-81).

STRESZCZENIE

Podstawą konstrukcji opisywanego modelu ewolucji genomu prokariotycznego jest kierunkowa presja mutacyjna, wprowadzająca wyraźne różnice składu między nicią wiodącą i opóźniającą cząsteczki DNA. *(rozdz. 1.1.)*. Asymetryczna presja mutacyjna działająca na jedną z nici DNA (nić wiodącą) została opisana za pomocą macierzy substytucji nukleotydowych, stworzonej na podstawie danych eksperymentalnych *(rozdz. 1.5.)*. Presję działającą na nić do niej komplementarną opisuje tablica lustrzana. Sekwencje kodujące leżące na takiej asymetrycznej nici mają własną asymetrię wprowadzaną przez presję mutacyjną związaną z transkrypcją i siły selekcji. Zmiana położenia takiego niesymetrycznego genu względem kierunku replikacji wiąże się z odwróceniem kierunku presji mutacyjnej, co nie pozostaje bez wpływu na częstość mutacji. Kierunkowe substytucje

wprowadzane przez tablicę właściwą dla „macierzystej" nici mają szansę ulec odwróceniu pod wpływem tablicy lustrzanej *(rozdz.4.4).* Efekt mutacyjny, czyli potencjalna eliminacja genu przez wprowadzoną substytucję, zależy od tolerancji genu na zmianę specyficznych parametrów odpowiedzialnych za funkcje kodowanego białka.

Degeneracja kodu genetycznego pozwala na zachowanie pewnej dowolności w składzie kodonowym, a więc i nukleotydowym genu, ponieważ część mutacji ma charakter selekcyjnie neutralny, dotyczy to szczególnie tzw. mutacji cichych, nie zmieniających sensu kodowanego aminokwasu. Mutacje zmieniające sens kodonu mogą także zachowywać neutralny charakter, jeżeli nie mają wpływu na funkcję białka. Substytucja jest wtedy zachowywana, a sekwencja adaptuje się częściowo do kierunku działania presji mutacyjnej. W miarę upływu czasu gen jest jednak coraz bardziej narażony na „wybicie" z obszaru tolerancji przez każdą kolejną mutację. Aby temu zapobiec należałoby odwrócić kierunek działającej presji, czyli zmienić położenie genu względem kierunku ruchu widełek replikacyjnych. W poniższej pracy podjęto próbę stworzenia i opisania modelu symulacyjnego strategii ewolucyjnej opartej na inwersji sekwencji kodujących. Wyniki wskazują, że inwersje mogą znacząco zmniejszyć koszty ewolucji genomu prokariotycznego.

„Każda rzecz, w którą można uwierzyć, jest jakimś obrazem prawdy...”

William Blake: „Zaślubiny Nieba i Piekła”

(w przekładzie W. Juszczaka)

1. WPROWADZENIE

Analizując procesy ewolucyjne, stajemy przed zasadniczym problemem: w jaki sposób badać doświadczalnie zjawisko, którego nie sposób ani powtórzyć, ani odtworzyć w laboratorium, gdyż jest jedynym znanym dotąd tego typu eksperymentem, rozpoczętym prawie 4 mld lat temu i trwającym do dziś... W jaki sposób badać mechanizmy ewolucyjne? Można sięgać do archiwum natury, czyli nagromadzonych przez miliardy lat skamieniałości i na podstawie porównań z żyjącymi organizmami dochodzić ich filogenezy, lub pokusić się o poznanie elementarnych zasad rządzących ewolucją i próbować stworzyć ogólny model zjawiska. Coraz popularniejszą metodą modelowania naturalnych, stochastycznych procesów,

do których należy zaliczyć zjawiska ewolucyjne, są symulacje komputerowe.

Idea symulacji opiera się na założeniu, że losowe procesy można odtworzyć, a przynajmniej naśladować, stosując algorytmy oparte na wykorzystaniu generatorów liczb losowych. Czas biologiczny, którego miarą może być czas życia jednego pokolenia, czyli mówiąc w języku DNA, cykl replikacyjny, znajduje wtedy odzwierciedlenie w liczbie kroków symulacji (*MCS - Monte Carlo Steps*). Przyjmując takie założenie i stosując czysto losowe zasady, otrzymalibyśmy jednak raczej szereg światów alternatywnych niż model ziemskiej ewolucji. Procesy ewolucyjne są w znacznym stopniu ukierunkowane, każdy następny krok jest obciążony jakąś historią, a w miarę upływu czasu rosną ograniczenia. Rozwój świata żywego jest skanalizowany przez wczesne, uniwersalne „osiągnięcia" ewolucji, takie jak kod genetyczny czy polimerazy kwasów nukleinowych. Uwzględniając takie zjawiska w modelu symulacyjnym, można dojść do ciekawych wniosków co do mechanizmów narastania złożoności.

Asymetria w cząsteczce DNA

DNA jest heteropolimerem, zwiniętym w formę podwójnej helisy. Dwie nici tworzące jedną cząsteczkę DNA (nić Watsona i nić Cricka) są komplementarne i antyrównoległe Końce pojedynczej nici nie są równocenne, wyróżniamy koniec 5', czyli wolną grupę hydroksylową przy piątym węglu dezoksyrybozy i koniec 3', grupę OH przy trzecim węglu pierścienia cukrowego. Nici podwójnej helisy biegną w przeciwnych kierunkach (antyrównolegle: od 5'do 3' i od 3' do 5'), jedną nić nazywamy nicią Cricka, a drugą nicią Watsona. Zgodnie z zasadą Chargaffa *(Chargaff 1948)* ilość adeniny w cząsteczce DNA odpowiada ilości tyminy, a ilość guaniny jest równa zawartości cytozyny. Na jej podstawie sformułowano regułę łączenia się zasad w pary, zwaną BPR *(Base Pairing Rule*), która była podstawą konstrukcji modelu cząsteczki kwasu dezoksyrybonukleinowego w 1953 r., w którym naprzeciw tyminy znajduje się zawsze połączona z nią podwójnym wiązaniem wodorowym adenina, a naprzeciwko cytozyny – połączona wiązaniem potrójnym – guanina. Z sekwencji jednej nici można więc wydedukować sekwencję nici komplementarnej. Tak skomplikowana cząsteczka jest swoistym, niepowtarzalnym układem. Można jednak potraktować ją w czysto fizykochemiczny sposób i spróbować opisać jej dynamikę za pomocą układów równań różniczkowych pozwalających przewidzieć stan równowagi. Z zasady BPR wynika jej logiczna konsekwencja, zwana PR-1 *(interstrand type-1 parity rule) (Sueoka*

1995). Wiąże ona częstość substytucji (czyli zmian nukleotydowych w sekwencji) z zasadą komplementarności. PR-1 można wyrazić wzorem:

$$m(i \rightarrow j) = m(\hat{i} \rightarrow \hat{j}) \qquad (1)$$

gdzie m(i→j) jest częstością przejść **i** –tego nukleotydu w **j**-ty na nici Cricka, a m(î →ĵ) prawdopodobieństwem mutacji nukleotydu komplementarnego do **i** w komplementarny do **j** na nici Watsona. Innymi słowy, zmiana C→T na jednej nici odpowiada przejściu G→A na drugiej. Zasadę tę wykorzystał Lobry *(Lobry 1995)* do wyprowadzenia równania opisującego asymptotyczne zachowanie się ewoluującej cząsteczki DNA i tzw. reguły PR-2 *(type-2 parity rule)*, według której w stanie równowagi powinna być zachowana wewnątrzniciowa równość stężeń molowych A=T i G=C. Lobry rozwiązał układ równań różniczkowych opisujących zmianę frakcji nukleotydu X w czasie ewolucji cząsteczki DNA, wyrażony wzorem:

$$dX/dt = MX \qquad (2)$$

gdzie dX to zmiana frakcji nukleotydu X w czasie, a M to macierz częstości poszczególnych przejść uproszczona do sześciu wartości zgodnie z zasadą PR-1. Zakładając warunek równowagi (dx/dt=0), otrzymujemy rozwiązanie numeryczne, z którego wynika, że frakcje nukleotydów komplementarnych A i T oraz G i C w obrębie jednej nici DNA powinny być sobie równe. Czy taką symetrię obserwujemy rzeczywiście w przyrodzie, czy też metody opisujące stany równowagowe nie są najlepszym sposobem przybliżania zjawisk biologicznych? Wszystko wskazuje na to, że życie toczy się jednak z dala od stanu równowagi…

1.2.1. Definicja asymetrii

W latach dziewięćdziesiątych w licznych zsekwencjonowanych genomach eubakteryjnych i wirusowych zaobserwowano lokalne odchylenia od reguły PR-2 *(Lobry 1996, Blattner i współpr. 1997, Mrazek, Karlin 1998, Mc Lean i współpr. 1998, Mackiewicz i współpr. 1999, Kowalczuk i współpr. 2001a).* Okazało się, że badając skład nukleotydowy pojedynczej nici DNA bakteryjnego można wyodrębnić w niej dwie połowy, w obrębie pierwszej symetria A=T i G=C jest odchylona na korzyść guaniny i tyminy, a w obrębie drugiej na korzyść adeniny i cytozyny. Tak wyraźnych odchyleń nie udało się jednak zaobserwować ani u *Archaebakteria,* ani u *Eukaryota*. Skąd biorą się przeciwstawne tendencje w składzie cząsteczki DNA?

1.3. Mechanizmy generujące asymetrię

Istnienie asymetrii w cząsteczce DNA jest zgodne z teorią informacji. Cząsteczka kodująca powinna być asymetryczna, świadczy to o nierównowagowym charakterze kształtujących ją procesów. Całkowita symetria występuje tylko w stanie równowagi termodynamicznej, w którym zmiany, jeżeli zachodzą, są bezkierunkowe i nie wpływają na stan układu, a entropia osiąga maksimum. Ewoluujący świat żywy jest daleki od stanu równowagi. Według definicji Schrödingera, życie to negentropia, czyli stale podtrzymywana, oparta na informacji organizacja, uporządkowanie.

Gdyby zdegradować jakąkolwiek istniejącą cząsteczkę DNA do pojedynczych nukleotydów, a następnie użyć tych samych nukleotydów do stworzenia nowej sekwencji, do której byłyby one włączane w sposób całkowicie losowy, otrzymana cząsteczka byłaby symetryczna, ale nic by nie kodowała.

Mając poparcie ze strony teorii informacji, genomicy zaczęli poszukiwać mechanizmów biologicznych odpowiedzialnych za asymetrię obserwowaną między nicią wiodącą a opóźniającą w nadziei, że może to rzucić nieco światła na prawa rządzące ewolucją na poziomie molekularnym *(Francino, Ochman 1997, Mrazek, Karlin 1998, Frank, Lobry 1999, Tiller, Collins 2000a, Kowalczuk i współpr. 2001a)*.

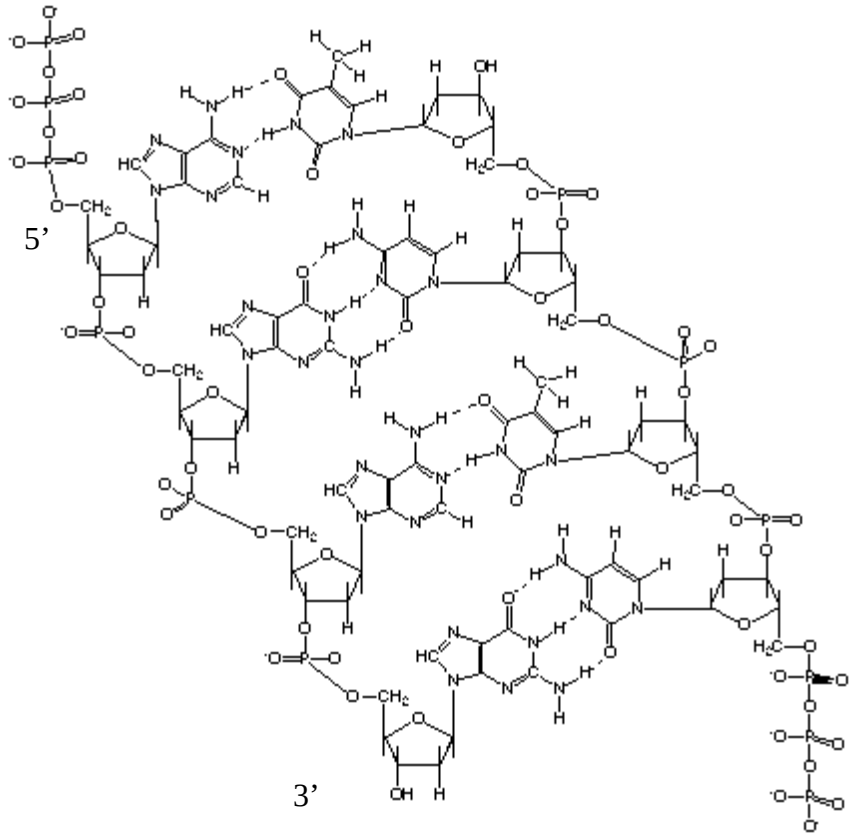
Skład nukleotydowy sekwencji DNA jest kształtowany przez dwie główne siły: presję mutacyjną i presję selekcyjną. Presja mutacyjna jest związana głównie z procesem replikacji i transkrypcji, presja selekcyjna wpływa na nielosowe rozmieszczenie genów i sekwencji sygnałowych na chromosomie oraz wymusza pewne trendy w użyciu i składzie nukleotydowym kodonów związane z konstrukcją kodu genetycznego i sekwencją aminokwasową białek.

1.3.1. Asymetria a mechanizm replikacji

DNA jest cząsteczką zbudowaną z dwóch antyrównoległych nici, każda z nici ma kierunek wyznaczany przez kierunek syntezy i replikacji kwasu dezoksyrybonukleinowego.

Substratami reakcji elongacji są trójfosforany dezoksyrybonukleotydów (dNTP). Przyłączanie kolejnego nukleotydu może zachodzić tylko w jednym kierunku – od końca 5’ do końca 3’. W wyniku reakcji jest uwalniany pirofosforan oraz powstaje wiązanie fosfodiestrowe między grupą hydroksylową przy piątym węglu dezoksyrybozy, a grupą OH przy trzecim węglu pierścienia cukrowego poprzedniego nukleotydu:

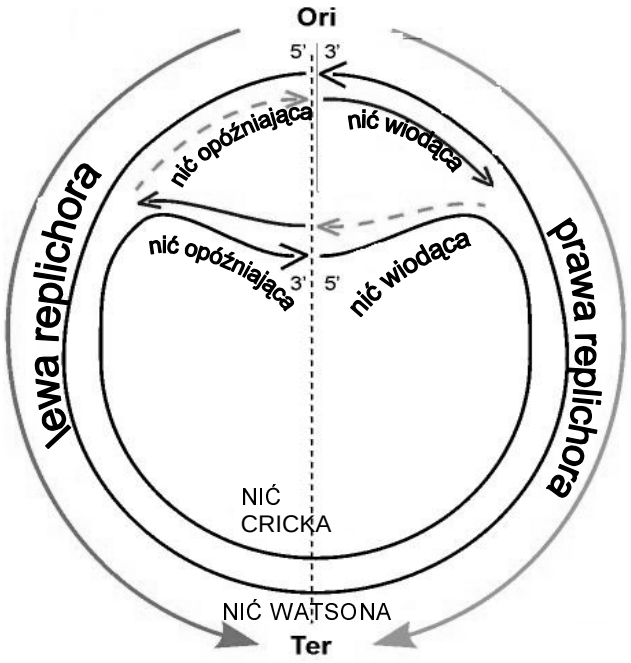
$$(DNA)_{n\ reszt}\ 3'\text{-}OH + 5'\text{-}\ dNTP \rightleftharpoons (DNA)_{n+1} + PP_i$$



Ryc.1. Wiązanie fosfodiestrowe, koniec 3’ i 5’ w dsDNA

Taki przebieg reakcji został utrwalony na wczesnym etapie ewolucji, jego specyfika nie mogła pozostać bez wpływu na strukturę i kształt syntetyzowanej cząsteczki.

Ponieważ w większości genomów bakteryjnych jest tylko jedno miejsce inicjacji replikacji *(ORI – origin of replication),* a reakcja replikacji biegnie w obie strony, tylko jedna połowa tej samej nici może być syntetyzowana w sposób ciągły, zgodnie z kierunkiem przesuwania się kompleksu enzymatycznego. Po drugiej stronie ORI synteza tej samej nici musi odbywać się etapami. W miarę rozplatania helisy na odsłoniętej nici syntetyzowane są startery RNA, od których zaczyna się wydłużanie łańcucha w kierunku 5’-3’, a więc ku ORI, przeciwnie do kierunku przesuwania się widełek replikacyjnych. Po prawej stronie ORI replikacja nici Watsona przebiega w sposób ciągły - kompleks polimerazy III nie oddysocjowuje od matrycy, proces zachodzi szybciej, dlatego przyjęto nazywać tę połowę nici nicią wiodącą. Reakcja po lewej stronie ORI przebiega przez tzw. *„fragmenty Okazaki”*.



Ryc.2. Schemat procesu replikacji chromosomu bakteryjnego: podział cząsteczki DNA na dwie replichory (linia podziału przebiega wzdłuż osi ORI-TER). Replikacja przebiega równocześnie w dwóch kierunkach i na dwóch niciach (na matrycy nici Watsona i nici Cricka) *(www.smorfland.microb.uni.wroc.pl*)

Podjednostka polimerazy III syntetyzuje tylko krótkie odcinki nici komplementarnej, które łączy potem ligaza, część nici Watsona po lewej stronie ORI *(Ryc.2.)* jest replikowana w sposób opóźniający. Jednocześnie na komplementarnej do nici Watsona nici Cricka sytuacja jest odwrotna. Część nici położona na prawo od ORI będzie syntetyzowana w sposób opóźniający, a część leżąca z lewej strony miejsca inicjacji replikacji będzie syntetyzowana w sposób wiodący. Cząsteczka DNA dzieli się więc na dwie replichory: prawą i lewą, granicą między nimi jest oś łącząca punkty ORI i TER. Każda replichora składa się z dwóch komplementarnych nici, z których jedna jest replikowana w sposób wiodący, a druga – opóźniający. Oczywiste jest, że presja mutacyjna związana z replikacją na nici wiodącej będzie różna od presji charakterystycznej dla nici opóźniającej. Skąd mogą wynikać te różnice?

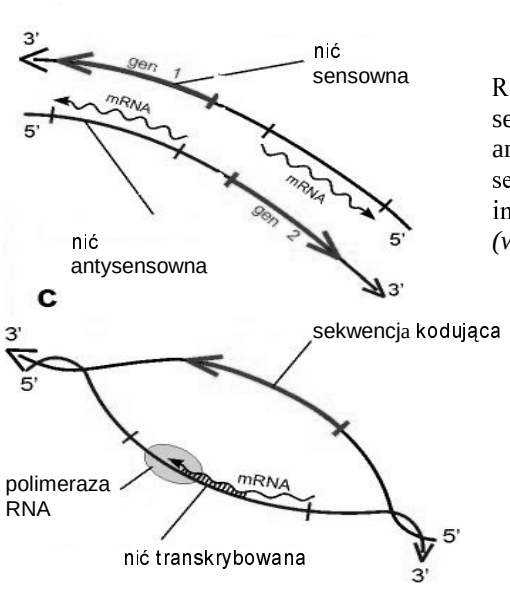
Skuteczność mechanizmów naprawczych w dużej mierze zależy od tego, czy ekscynuleazy mają dostęp do nowo syntetyzowanej nici *(Fijałkowska, Schapper 1996),* czy też jest on utrudniony przez kompleks replikacyjny. Nić opóźniająca powinna więc być replikowana w sposób wierniejszy. Z drugiej jednak strony, aby zachować stałe tempo replikacji, podjednostka polimerazy III syntetyzująca nić opóźniającą powinna katalizować szybsze włączanie nukleotydów do wydłużanego łańcucha, co może nie pozostawać bez wpływu na sprawność tzw. naprawy *„proofreading”* związanej z aktywnością

egzonukleolityczną podjednostki E *(Radman 1998).* Badania eksperymentalne nie dały jak dotąd jednoznacznej odpowiedzi na pytanie, która z nici jest replikowana w sposób wierniejszy *(Iwaki i współpr. 1996, Fijałkowska i współpr. 1998)*, co jest być może związane z faktem, że przeprowadzono je na szczepach pozbawionych mechanizmów naprawczych *(Roberts i współpr. 1994, Thomas i współpr. 1996).*

Inną próbą wyjaśnienia zjawiska asymetrii jest teoria dezaminacji cytozyny *(Frank, Lobry 1999).* Jak wynika z danych doświadczalnych *(Frederico i współpr. 1998),* ssDNA *(single stranded DNA)* jest 140 razy wrażliwszy na mutacje spowodowane dezaminacją cytozyny niż forma dwuniciowa. Tranzycje typu C→T są najczęściej obserwowanym rodzajem substytucji w cząsteczce DNA (cytozyna różni się od uracylu tylko jednym podstawnikiem, właśnie grupą aminową przy pierścieniu węglowym, a metylocytozyna po dezaminacji przechodzi bezpośrednio w tyminę). Nić wiodąca jest przez cały czas trwania replikacji narażona na tranzycje C→T, co musi znacząco zwiększać frakcję tyminy. Drugą najczęściej zachodzącą substytucją jest przejście A→G (w wyniku dezaminacji adeniny powstaje hipoksantyna, która łatwiej przechodzi w formę tautomeryczną tworzącą trzy wiązania wodorowe i łączy się z cytozyną, zamiast z tyminą). Może to tłumaczyć przewagę T i G na nici wiodącej oraz odpowiednio A i C na nici opóźniającej.

1.3.2. Asymetria a procesy transkrypcyjne

Na efekty związane z mechanizmem replikacji nakładają się skutki presji mutacyjnej związanej z transkrypcją. Przepisywanie informacji genetycznej z DNA na mRNA jest także



Ryc.3. Schemat procesu transkrypcji: nić sensowna (jej sekwencja jest identyczna do syntetyzowanego mRNA) i antysensowna (o sekwencji komplementarnej do nici sensownej). Jak widać nić sensowna podczas przepisywania informacji pozostaje odsłonięta i znajduje się w formie ssDNA *(www.smorfland.microb.uni.wroc.pl)*

procesem kierunkowym i odbywa się od końca 5' do 3'cząsteczki mRNA. Matrycą dla transkryptazy jest zawsze nić komplementarna do nici kodującej, zwana antysensowną *(Ryc.3.).* Podczas transkrypcji jedna z nici, nietranskrybowana (sensowna) jest odsłonięta i narażona na częstsze mutacje. Druga nić, na której jako na matrycy tworzone jest mRNA, jest zabezpieczona przez kompleks transkrypcyjny i dodatkowo chroniona przez enzymy naprawcze (szczególnie, jeżeli chodzi o usuwanie dimerów tymidynowych) *(Hanawalt 1991).* Jeżeli nic sensowna genu leży na nici replikowanej w sposób wiodący, to transkrypcja genu odbywa się w kierunku zgodnym z kierunkiem przesuwania się widełek replikacyjnych, w przeciwnym wypadku kompleks transkrypcyjny musi przesuwać się „pod prąd" procesu replikacji. Ponieważ w komórkach prokariotycznych transkrypcja i replikacja nie są od siebie ani czasowo, ani przestrzennie rozgraniczone, geny mogą preferować położenie na nici wiodącej, gdzie nie ma ryzyka kolizji kompleksów enzymatycznych. Takie preferencje istotnie zostały zaobserwowane u licznych gatunków bakterii *(Brewer 1988, Fraser i współpr. 1995).* Dokładniejsze analizy dotyczyły tempa ekspresji genów zlokalizowanych na nici wiodącej i opóźniającej. Geny o wysokiej ekspresji, w przypadku nierównomiernego rozkładu na chromosomie, powinny wprowadzać dodatkową asymetrię globalną w składzie nukleotydowym, ponieważ są kodowane przez specyficzne „szybkie" kodony. Większość aminokwasów jest kodowanych przez więcej niż jeden kodon, w komórce występuje często więcej niż jeden rodzaj akceptorowego tRNA dla danego aminokwasu. Stężenia molowe tych cząsteczek w komórce mogą być różne. Im więcej akceptorowego tRNA dla danego kodonu, tym szybsze tempo translacji białka kodowanego przez gen zbudowany z takich „szybkich" kodonów. Dla każdego genu można obliczyć tzw. CAI (Codon Adaptation Index) *(Gouy, Gautier 1982, Sharp, Li 1987),* który jest miarą szybkości jego translacji. Do tej pory jednak rola trendów związanych z białkami o wysokiej ekspresji w ogólnej asymetrii genomów bakteryjnych nie została jednoznacznie określona.

1.3.3. Asymetria a sekwencje sygnałowe

Jedną z przyczyn asymetrii składu nukleotydowego nici opóźniającej i wiodącej może być także nielosowe rozłożenie pewnych krótkich sekwencji o określonym składzie, jak np. sekwencje Chi (5' GCTGGTGG 3'), czyli gorące miejsca rekombinacji, które u *E.coli* są położone w większości na nici wiodącej *(Blattner i współpr. 1997).* Ich wkład w ogólną asymetrię nie jest jednak istotny, a już na pewno nie jest jej głównym źródłem *(Tillier, Collins 2000a).*

1.3.4 Trendy związane z kodowaniem białek

Przyczyną powstawania trendów w składzie nukleotydowym może być sam fakt kodowania białek. Konstrukcja kodu genetycznego wymusza występowanie specyficznych kombinacji nukleotydów w I i II pozycji kodonów *(Wong, Cedergren 1986, Zhang 1991, Cebrat i współpr. 1997, 1998, McLean i współpr. 1998).* Pierwsze pozycje kodonów wykazują przewagę adeniny i guaniny, podczas gdy drugie są bogatsze w adeninę i cytozynę. Te trendy tak silnie wyróżniają sekwencje kodujące, że na ich podstawie można je odróżniać od niekodujących sekwencji międzygenowych. Jest kilka powodów, dla których selekcja może preferować takie układy nukleotydów w kodonach. Pierwszy to przypuszczalne faworyzowanie kodonów z purynami w I pozycjach, ponieważ puryny są mniej wrażliwe na mutacje niż pirymidyny *(Hutchinson 1996),* poza tym guanina w I pozycji znacznie zwiększa siłę wiązania mRNA z kompleksem rybosomalnym, co wpływa na wierność translacji i ułatwia zachowanie prawidłowej ramki odczytu. Ponadto występowanie kodonów GAN i GNN świadczy o częstym występowaniu aminokwasów o charakterze kwaśnym (kwas asparaginowy i glutaminowy) i małych aminokwasów obojętnych (alanina, glicyna, walina) w łańcuchach polipeptydowych.

Podsumowując, należałoby uporządkować wymienione wyżej mechanizmy według ich istotności. Bezsprzecznie największy udział w tworzeniu asymetrii ma presja związana z replikacją. Nierównomierny rozkład genów może mieć dość istotny wpływ, ale tylko na asymetrię tych genomów, gdzie obserwuje się znaczącą dysproporcję

Phe Leu Ile Met Val Ser Pro Thr Ala Tyr His Gln Asn Lys Asp Glu Cys Trp Arg Gly Ochre Opal Amber																							
F	**L**	**I**	**M**	**V**	**S**	**P**	**T**	**A**	**Y**	**H**	**Q**	**N**	**K**	**D**	**E**	**C**	**W**	**R**	**G**	**-**	**-**	**-**
UUU	UAA	AUU	AUG	GUU	UCU	CCU	ACU	GCU	UAU	CAU	CAA	AAU	AAA	GAU	GAA	UGU	UGG	CGC	GGU	UAA	UGA	UAG
UUC	UUG	AUC		GUC	UCC	CCC	ACC	GCC	UAC	CAC	CAG	AAC	AAG	GAC	GAG	UGC		CGU	GGC			
	CUU	AUA		GUA	UCA	CCA	ACA	GCA										CGA	GGA			
	CUC			GUG	UCG	CCG	ACG	GCG										CGG	GGG			
	CUA				AGU													AGA				
	CUG				AGC													AGG				

Ryc.4. Kod genetyczny – niektóre aminokwasy są kodowane przez 6 różnych kodonów (Leu, Ser, Arg), niektóre przez dwa (np. Phe, His, Gln) lub tylko przez jeden (Trp, Met). Trzecie pozycje często nie mają wpływu na rodzaj kodowanego aminokwasu, stąd mutacje w trzeciej pozycji mają często charakter neutralny. O tym, że pewna tolerancja w oddziaływaniu kodon - antykodon jest dopuszczalna, świadczy występowanie w antykodonach tRNA specyficznej zasady, inozyny, która może oddziaływać z trzema różnymi zasadami (U,C,A), zawsze jednak występuje w pierwszej od końca 5' pozycji antykodonu, a więc może się łączyć tylko z trzecią pozycją kodonu.

między ilością genów na nici opóźniającej i wiodącej *(Borrelia burgdorferi, Mycoplasma genitalium, Mycoplasma pneumoniae, Bacillus subtilis)*. Wpływ rozkładu genów o wysokiej ekspresji i sekwencji sygnałowych jest prawdopodobnie marginalny. Presja transkrypcyjna jest natomiast czynnikiem istotnym, choć na pewno słabszym niż presja replikacyjna *(Frank, Lobry 1999).*

1.4 . Konsekwencje asymetrii dla genomu bakteryjnego

Presja mutacyjna związana z replikacją nie pozostaje bez wpływu na ewolucję i rearanżację genomów. Najczęściej stosowaną metodą badania tego oddziaływania jest porównywanie ortologów z blisko spokrewnionych genomów. Porównania takie, na poziomie składu nukleotydowego, kodonowego i aminokwasowego, przeprowadzili m.in. Lafay i współpr. (1999) oraz Tillier i Collins (2000b). Z obserwacji (objęto nimi łącznie cztery genomy: *B. burgdorfei i T. pallidum oraz Ch. trachomatis i Ch. pneumoniae*) wynikało, że ortologi leżące na różnych niciach wykazywały o wiele większą dywergencję, niż ortologi leżące w obu porównywanych genomach na tej samej nici

Analizę na szerszą skalę przeprowadzili Szczepanik i współpracownicy *(Szczepanik i współpr. 2001, Mackiewicz i współpr. 2003).* Celem badań było wykrycie różnic w tempie dywergencji między genami leżącymi na nici wiodącej, opóźniającej i takimi, które w trakcie swojej ewolucji zmieniły nić. Analizie poddano 12 645 par ortologów wybranych z 11 genomów bakteryjnych wykazujących znaczącą asymetrię. Niemal we wszystkich przypadkach stwierdzono, że dywergencja mierzona między ortologami z nici opóźniającej jest istotnie większa niż odległość między tymi samymi genomami liczona na podstawie różnic między ortologami wybranymi z nici wiodącej. Można to tłumaczyć bądź silniejszą presją mutacyjną działającą na geny leżące na nici opóźniającej, co sprawia, że ewoluują one szybciej, bądź słabsza presją selekcyjną, która nie utrzymuje tak restrykcyjnie, jak na nici wiodącej składu aminokwasowego genów *(Szczepanik i współpr. 2001)*. Ortologi, które na pewnym etapie ewolucji zmieniły nić, wykazywały jeszcze wyższe tempo dywergencji niż geny z nici opóźniającej, co wskazuje na to, że geny po inwersji gromadzą więcej mutacji, niż te, które po ewolucji na jednej nici zdążyły przystosować swój skład do danej presji.

Ostateczne wnioski z analizy przeprowadzonej na podstawie uzupełnionych baz danych można streścić następująco *(Mackiewicz i współpr. 2003)*: różna presja mutacyjna związana z replikacją nici w sposób wiodący i opóźniający powoduje, że przemieszczanie genów między replichorami podlega pewnym ograniczeniom, im asymetria jest wyraźniejsza, tym bariery

rearanżacyjne są ostrzejsze. Przeniesienie genu z jednej nici na drugą powoduje wzrost częstości mutacji, im częściej gen zmienia nić, tym większa jest obserwowana dywergencja. Ostateczny wynik pomiaru odległości filogenetycznej metodami genomicznymi zależy więc w znacznym stopniu od relacji między presją mutacyjną i selekcyjną działającymi na dany gen, które mogą zmieniać się w czasie. Zróżnicowanie tempa dywergencji genów powinno być uwzględnione przy tworzeniu drzew filogenetycznych (błędy wynikające z zaniedbania tych różnic mogą dotyczyć nie tylko długości gałęzi, ale i topologii drzew filogenetycznych).

Asymetria presji mutacyjnej wpływa także na możliwości rearanżacji genomów. Z porównania ortologów wynika, że wyraźnie preferowane są rearanżacje symetryczne względem osi ORI-TER *(Mackiewicz i współpr. 2001b)*, ponieważ nie powodują one ani zmiany nici, ani zmiany odległości od miejsca początku replikacji. Geny leżące blisko ORI, nazywane proksymalnymi, są to najczęściej geny o wysokiej ekspresji, replikowane w pierwszej kolejności, co zapewnia wysoki poziom stężenia ich produktów w komórce. Geny położone wokół TER, zwane dystalnymi, kodują białka, których stężenie w komórce powinno być utrzymane na niskim poziomie. Odległość od ORI ma więc duże znaczenie selekcyjne.

Zmiany położenia genów, czyli rearanżacje chromosomu, są jednym z ważnych mechanizmów ewolucji. Być może ucieczka spod szkodliwej presji mutacyjnej jest jednym ze sposobów obrony przed zbyt intensywną eliminacją i stratami ewolucyjnymi, zabezpieczającym niesioną przez geny informację i nie dopuszczającą do znaczącej redukcji liczebności populacji.

Asymetria może być podstawą i punktem wyjścia do stworzenia interesującego modelu zachowania się genomu prokariotycznego. Symulowanie ewolucji składu nukleotydowego sekwencji, a także zjawisk takich, jak eliminacja genów mogłoby dać odpowiedź na wiele pytań o korelacje i sprzężenia rządzące tymi zjawiskami w naturze. Na czym powinien opierać się taki model? Przede wszystkim powinien posiadać określony rozkład prawdopodobieństwa dla poszczególnych zmian mutacyjnych oraz algorytm określający siłę i warunki presji selekcyjnej.

1.5. Konstrukcja tablicy przejść mutacyjnych

Obserwowany przez nas skład sekwencji, ich właściwości i asymetria są efektem procesu ewolucji - rezultatem działania wypadkowych sił mutacji i selekcji. Konstrukcja spójnego algorytmu, który pozwalałby na obserwację skutków działania czystej siły selekcyjnej, wymaga rozdzielenia presji mutacyjnej i selekcyjnej. W jaki sposób odtworzyć

macierz przejść mutacyjnych na podstawie znanej sekwencji nukleotydowej? Porównując spokrewnione sekwencje można stworzyć tablice substytucji (aminokwasowych lub nukleotydowych), ale będą one odbiciem zarówno presji związanej z mutacją, jak i selekcją (tablice PAM – *Dayhoff i współpr. 1978)*. Czystą presję mutacyjną próbowano określić na podstawie funkcji matematycznych stanu równowagi opartych na założeniach teorii Markova, według której częstości występowania kolejnych zasad w sekwencji nukleotydowej (P_G P_C P_T P_A) w chwili t spełniają równanie:

$$P_\alpha (t+1) = P_\alpha (t) (1- \Sigma W_{\alpha\beta}) + P_\beta(t)\Sigma W_{\beta\alpha} \quad (\beta \neq \alpha) \qquad (3)$$

gdzie α i β to kolejne zasady ATGC, a $W_{\alpha\beta}$ reprezentuje elementy macierzy przejść nukleotydowych. Jak wynika ze wzoru *(3)* frakcja nukleotydu α w sekwencji w chwili t+1 zależy od częstości występowania tego nukleotydu w chwili t i od sumy częstości przejść $\alpha \rightarrow \beta$ (czyli substytucji przez pozostałe 3 nukleotydy) oraz od sumy iloczynu częstości przejść pozostałych nukleotydów w α ($W_{\beta\alpha}$) i częstości tych nukleotydów (P_β). Równanie to jest słuszne przy założeniu, że zdarzenia mutacyjne tworzą tzw. łańcuch Markova, czyli prawdopodobieństwa kolejnych zdarzeń zależą tylko od stanu układu w kroku poprzednim. W warunkach stanu stacjonarnego częstości nukleotydów nie ulegają zmianie w skutek zachodzących mutacji. Jeżeli w danym miejscu α przejdzie w β, to w innym α zastąpi β, sumaryczna zawartość nukleotydu α pozostanie więc niezmieniona. Można to zapisać matematycznie:

$$P_\alpha (t+1) - P_\alpha (t) = 0 \qquad 4),$$

uwzględniając zależność (3):

$$P_\alpha (t) (1- \Sigma W_{\alpha\beta}) + \Sigma W_{\beta\alpha} P_\beta(t) = P_\alpha (t), \qquad (5)$$

stąd:

$$P_\alpha (t) \Sigma W_{\alpha\beta} = \Sigma W_{\beta\alpha} P_\beta(t). \qquad (6)$$

Uzyskaliśmy w ten sposób układ 4 równań z 12 niewiadomymi - częstościami przejść $W_{\alpha\beta}$. Mamy więc nieskończenie wiele możliwych rozwiązań przy zadanych frakcjach $P_A(t)$, $P_G(t)$, $P_T(t)$ i $P_C(t)$. Skoro analityczne wyliczenie macierzy przejść ze składu nukleotydowego okazało się niemożliwe, próbowano stosować uproszczenia i przybliżenia tablicy substytucji (głównie w celu estymacji K, czyli standardowej odległości między sekwencjami, liczonej w ilości podstawień na miejsce). Istnieją trzy matematyczne modele estymacji K, każdy z nich opiera się na uproszczeniu dwunastoparametrowej macierzy przejść *(Tab.1.)*. Jednoparamertrowy model Jukesa i Cantora zakłada równe częstości przejść dla wszystkich

nukleotydów *(Tab.2.) (Jukes, Cantor 1969).* Dwuparametrowy model Kimury wprowadza zróżnicowane częstości dla tranzycji i transwersji *(Tab.3.) (Kimura 1980).* Wspomniane modele zakładają stałość częstości substytucji w czasie, trzeci model, Gu i Li *(1998),* wprowadza zależność częstości przejść od czasu, jest więc bardziej biologiczny, ale pozwala tylko na estymację, a nie dokładne wyliczenie K.

Pierwsze próby doświadczalnego konstruowania tablic przejść pojawiły się w latach osiemdziesiątych. Wu i Maeda próbowali je odtworzyć porównując sekwencje międzygenowe między homologami kilku niezbyt odległych filogenetycznie gatunków *(Wu, Maeda 1987).*

Tab.1. Ogólna postać macierzy przejść nukleotydowych. Każdy z jej elementów jest częstością przejść i-tego nukleotydu w j-ty, elementy diagonalne spełniają równanie $r_{ii} = -\Sigma_{i \neq j} r_{ij}$, tak, że suma wszystkich elementów w każdym wierszu wynosi 0.

	A	**T**	**G**	**C**
A	-(a+b+c)	a	b	c
T	d	-(d+e+f)	e	f
G	g	h	-(g+h+i)	i
C	j	k	l	-(j+k+l)

Tab.2. Jednoparametrowa tablica przejść nukleotydowych *(Jukes, Cantor 1969)*

	A	**T**	**C**	**G**
A	-3u	u	u	u
T	u	-3u	u	u
C	u	u	-3u	u
G	u	u	u	-3u

Tab.3. Dwuparametrowa tablica Kimury (s-tranzycja, v-transwersja) *(Kimura 1980)*

	A	**T**	**C**	**G**
A	-(2v+s)	v	v	s
T	v	-(2v+s)	s	v
C	v	s	-(2v+s)	v
G	s	v	v	-(2v+s)

Metoda opierała się na założeniu, że sekwencje międzygenowe nie podlegają selekcji, a więc mogą akumulować wszystkie substytucje, które zaszły w czasie ich ewolucji. Kowalczuk i współpracownicy *(2001b)* udoskonalili tę metodę, aby opracować tablicę mutacyjną, specyficzną dla genomu *B. burgdorferi*. Zamiast dowolnych sekwencji międzygenowych analizowano tylko pseudogeny, czyli sekwencje pochodzące ze zduplikowanych w obrębie jednego genomu genów uwolnionych spod presji selekcyjnej. Porównując je z ich genami macierzystymi określono częstości wszystkich dwunastu przejść mutacyjnych *(Kowalczuk i współpr. 2001a).* Metoda polegała na wybraniu wszystkich odcinków międzygenowych u *B. burgdorferi* i przetłumaczeniu ich na sekwencje aminokwasowe we wszystkich sześciu możliwych fazach. Przetłumaczonych na aminokwasy sekwencji międzygenowych użyto do przeszukania bazy *B. burgdorferi* programem FASTA *(Pearson i Lipman 1988).* Sekwencje ORFów i homologicznych do nich sekwencji międzygenowych dopasowano programem CLUSTAL X *(Jeanmougin i współpr. 1988).* Statystycznie istotną liczbę par homologów uzyskano tylko dla nici wiodącej, dla niej też stworzono tablicę substytucji opierając się na metodzie Gojoboriego i współpr. *(1982)* oraz Francino i Ochmana *(2000)* i modyfikując ją przez wprowadzenie poprawki na wielokrotne substytucje i rewersje dla każdego nukleotydu osobno, zamiast uniwersalnej poprawki Kimury *(1980).*

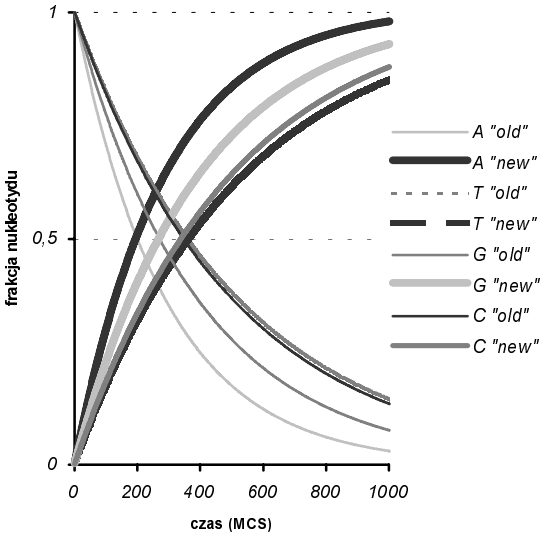
Tab.4. Asymetryczna tablica presji mutacyjnej *B. burgdorferi* dla nici wiodącej (a) i będąca jej lustrzanym odbiciem tablica przejść nukleotydowych dla nici opóźniającej (b) (nukleotyd w kolumnie przechodzi w nukleotyd w wierszu) *(Kowalczuk i współpr. 2001b)*

(a)	A	T	G	C	***(b)***	**A**	**T**	**G**	**C**
A	*0,807*	*0.103*	*0.067*	*0.023*	**A**	***0,865***	***0.065***	***0.035***	***0.035***
T	*0.065*	*0,865*	*0.035*	*0.035*	**T**	***0.103***	***0,807***	***0.023***	***0.067***
G	*0.164*	*0.116*	*0,705*	*0.015*	**G**	***0.261***	***0.070***	***0,622***	***0.047***
C	*0.070*	*0.261*	*0.047*	*0,622*	**C**	***0.116***	***0.164***	***0.015***	***0,705***

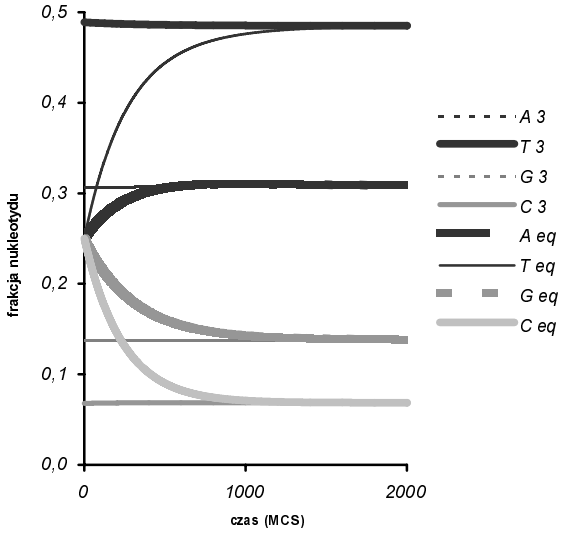
Otrzymana tablica empiryczna była następnie testowana przy pomocy symulacji komputerowych. Zgodnie z założeniami analitycznej definicji stanu równowagi *(wzór 3),*

czysto mutacyjna macierz substytucji powinna doprowadzić sekwencję do stałego w czasie składu równowagowego. W trakcie symulacji wybierano losowo nukleotyd (ze stałym prawdopodobieństwem p_{mut}), który następnie ulegał substytucji zgodnie z prawdopodobieństwami przejść określonymi w tablicy (możliwych przejść, łącznie z tożsamościowymi, było szesnaście, więc nie każdy nukleotyd wybrany do mutacji ulegał rzeczywiście substytucji). Niezależnie od tego, czy w pliku wejściowym dla programu była sekwencja losowa, czy sekwencje odpowiadające składem pierwszym, drugim czy trzecim pozycjom w kodonach, po odpowiednio dużej liczbie kroków Monte Carlo skład wszystkich sekwencji ustalał się na poziomie właściwym dla trzecich pozycji, skąd wniosek, że te ostatnie znajdują się w genomie *B. burgdorferi* w równowadze z presją mutacyjną. Uzyskana tablica utrzymuje nie tylko skład, ale i asymetrię naturalnych sekwencji będących w równowadze z presją mutacyjną (w sekwencji *B. burgdorferi* okazały się nimi sekwencje trzecich pozycji w ORFach z nici wiodącej). Takiego wyniku nie udało się osiągnąć za pomocą żadnego z typowych modeli parametrycznych *(Danchin, Rocha 2001)*

Symulacje oparte na skonstruowanej przez Kowalczuk i współpracowników tablicy substytucji *(Tab. 4.)* pozwoliły na obserwację dynamiki podstawień nukleotydów powodowanych czystą presją mutacyjną. Możliwość rejestrowania każdego podstawienia (w stosunku do poprzedniego kroku) stworzyła sposobność oszacowania częstości rewersji i podmian wielokrotnych, które nie są widoczne przy porównywaniu dwóch odległych od



Ryc. 5. Dynamika wymiany nukleotydów w sekwencji poddawanej działaniu czystej presji mutacyjnej – oryginalne nukleotydy (N „old') są podstawiane przez nowo wprowadzane (N „new"), punkty przecięcia krzywych wyznaczają długość okresów półtrwania poszczególnych nukleotydów.



Ryc. 6. Ewolucja składu nukleotydowego sekwencji o składzie $f_A=f_G=f_C=f_T=0,25$ pod wpływem tablicy przejść mutacyjnych; skład sekwencji dąży do składu trzecich pozycji w kodonach

siebie ewolucyjnie sekwencji.

Symulacja wykorzystująca parametry *Tab. 4.* pozwoliła też na obserwację dynamiki wymiany poszczególnych rodzajów nukleotydów w sekwencji, czyli czasu ich połowicznego zaniku, Modyfikacja algorytmu symulacyjnego pozwoliła na oszacowanie trwałości nukleotydów definiowanej analogicznie do okresu połowicznego rozpadu pierwiastków promieniotwórczych. Okres półtrwania nukleotydu to czas (mierzony w MCS) potrzebny na podstawienie połowy całkowitej ilości danego nukleotydu w sekwencji oryginalnej. Dla eksperymentalnej tablicy substytucji okres półrozpadu poszczególnych nukleotydów okazał się silnie liniowo skorelowany z frakcjami tych nukleotydów w sekwencji *B. burgdorferi* doprowadzonej do stanu równowagi z presją mutacyjną. Zależność tę można wyrazić wzorem:

$$F_a \sim p_{mut}\, t_a + \text{const} \qquad (7)$$

gdzie F_a to frakcja danego nukleotydu, p_{mut} –częstość mutacji, t_a –okres połowicznego zaniku nukleotydu *a*. Ze wzoru *(7)* wynika, że frakcja danego nukleotydu w sekwencji, która znajduje się w stanie równowagi z presją mutacyjną zależy tylko od częstości z jaką wpadają mutacje i długości okresu półrozpadu danego nukleotydu. Prawo to dotyczy jednak sytuacji szczególnej, w której zakładamy osiągnięcie stanu termodymanicznej równowagi, a nie bierzemy pod uwagę sił związanych z selekcją i dążeniem do zachowania niesionej przez sekwencję informacji.

Założenie, że trzecie pozycje w kodonie niosą istotną informację o kierunku i charakterze presji mutacyjnej u *Prokaryota* było testowane także przez innych badaczy. Analizy oparte o sekwencje dostępne w bazach danych zostały przeprowadzone m.in. przez zespół *Danchin i Rocha* (2001). Celem badań, którymi objęto dwadzieścia osiem genomów prokariotycznych dostępnych w *GenBanku*, było określanie asymetrii i poszukiwanie trendów związanych z replikacją metodą liniowej analizy dyskryminacyjnej. Zaobserwowaną asymetrię uznawano za znaczącą, jeżeli procent genów sklasyfikowanych prawidłowo wyłącznie na postawie ich składu nukleotydowego jako geny z nici wiodącej lub opóźniającej był większy niż 65-70%. Podstawą analizy były dwa parametry:

$$GCskew = (G - C) / (G+C) \qquad (8)$$

$$ATskew = (T - A) / (T+A) \qquad (9)$$

Parametry te, często stosowane w różnego typu analizach genomicznych, pozwalają na liczbowe oszacowanie asymetrii genomu. Aby porównać wielkości tych wskaźników dla genów leżących na różnych niciach, obliczono $\Delta GCskew$ i $\Delta ATskew$, zdefiniowane jako

różnice między średnimi wynikami dla genów z nici wiodącej i genów z nici opóźniającej. Największą asymetrię, opisaną wyżej wymienionym parametrem, wykazały genomy: *B. burgdorferi, Chlamydia trachomatis, Treponema pallidum i Chlamydia pneumaoniae*, znaczące wartości asymetrii składu nukleotydowego otrzymano w przypadku 21 spośród 28 badanych chromosomów bakteryjnych *(Danchin, Rocha 2001)*. Między wynikami analizy dyskryminacyjnej a wartościami Δ*GCskew* otrzymano wysoką korelację (współczynnik korelacji Spermana wyniósł 0,77), co świadczyłoby o decydującej roli asymetrii GC w generowaniu różnic między nićmi. Następnym etapem badań była deterministyczna analiza zmian składu nukleotydowego sekwencji w czasie z uwzględnieniem teorii dezaminacji cytozyny, zgodnie z którą presja mutacyjna związana z replikacją indukuje przejście C w T z częstością *z* w ciągu czasu *t*. Częstość ta jest stała i charakterystyczna dla danej nici. Analizując skład genu należy więc wziąć pod uwagę zmiany w czasie. Danchin i Rocha zmodyfikowali swój model wprowadzając równania różniczkowe opisujące dynamikę zmian składu sekwencji:

$$dN_A/dt = 0;\ dN_C/dt = -zN_C;\quad dN_G/dt = 0;\ dN_T/dt = zN_C; \qquad (10)$$

Zgodnie z układem równań *(8)* w chwili t skład sekwencji będzie następujący: $N_A = N_{A,0}$, $N_C = e^{-zt}N_{C,0}$, $N_G = N_{G,0}$, $N_T = N_{T,0} + (1- e^{-zt})N_{C,0}$, co oczywiście nie pozostaje bez wpływu na *GCskew* i *ATskew*, które trzeba wyliczyć na podstawie wzoru dynamicznego:

$$GC'skew = (N_{G,0} - e^{-zt}N_{C,0})/ N_{G} + e^{-zt}N_{C,0}$$
$$AT'skew = (N_{T,0} + (1- e^{-zt})N_{C,0} - N_{A,0})/ (N_{T,0} + (1- e^{-zt})N_{C,0} + N_{A,0}) \qquad (11)$$

Otrzymane równania opisują w sposób deterministyczny ewolucję składu nukleotydowego sekwencji. Badacze zastosowali równania *(11)* do obliczenia przewidywanych wartości ΔGC' i ΔAT' dla poszczególnych genomów, co jednocześnie pozwoliło na przetestowanie teorii dezaminacji. Założono, że największą zgodność wyników powinna zapewnić analiza trzecich pozycji w kodonach, które powinny znajdować się w równowadze z presją mutacyjną. Rzeczywiście dla 21 genomów stwierdzono znaczącą korelację między wartościami *zt* wyliczonymi na podstawie równań *(11)* i składu nukleotydowego trzecich pozycji, co potwierdza rolę dezaminacji cytozyny w generowaniu asymetrii oraz hipotezę o równowagowym składzie trzecich pozycji. Pewne niedoszacowania (szczególnie w przypadku *GCskew*) świadczą jednak o istotnym wpływie innych czynników generujących asymetrię

sekwencji, których analityczna transformacja mogłaby być o wiele bardziej skomplikowana niż dynamiczny model uwzględniający jedynie wpływ dezaminacji cytozyny. Użycie modelu symulacyjnego wydaje się więc być w tym przypadku uzasadnione. Otrzymane wyniki potwierdzają założenia o równowagowym składzie trzecich pozycji w kodonie, a jednocześnie zachęcają do użycia metod niedeterministycznych, które mogą być jedyną dostępną na aktualnym etapie badań drogą rozwiązania problemu.

1.6. Tablice PAM a procesy ewolucyjne

Znalezienie odpowiedzi na pytanie o mechanizm zmian ewolucyjnych na poziomie sekwencji nukleotydowej sprowadza się do poszukiwania wzorca, czy też matrycy przejść nukleotydowych, które doprowadziły do powstania istniejących sekwencji.

Jedną z metod określenia takiej macierzy przejść jest analiza homologicznych białek. Ponieważ informacja, którą można uzyskać z takich porównań dotyczy poziomu aminokwasowego, wyniki są zbierane w tablicy przejść aminokwasowych. Pierwsza taka tablica została opublikowana pod koniec lat sześćdziesiątych *(Dayhoff i współpr. 1969)* i była aktualizowana w miarę gromadzenia danych o nowopoznanych białkach *(Dayhoff i współpr. 1972; Schwartz, Dayhoff 1978; Risler i współpr. 1988)*. Konstrukcja macierzy Dayhoff opiera się na zaobserwowanych między homologicznymi sekwencjami różnicach w składzie aminokwasowym. Oczywiście opierając się na istniejących sekwencjach białkowych możemy obserwować tylko mutacje, które zostały zaakceptowane przez naturalną selekcję, gubimy natomiast informację o mutacjach eliminowanych przez presję selekcyjną oraz o wielokrotnych substytucjach. Zaktualizowaną wersję macierzy Dayhoff oparto na 1572 zmianach aminokwasowych zaobserwowanych w 71 grupach białek. W oparciu o analizowane sekwencje odtworzono 71 drzew filogenetycznych, przy czym nieznane sekwencje wspólnych przodków dla każdego odtworzonego drzewa zostały losowo wygenerowane. Dane zebrano w formie symetrycznej macierzy zmian aminokwasowych (przy zliczaniu zaobserwowanych substytucji nie uwzględniano kierunku przejścia, zakładając równe prawdopodobieństwa zmiany na przykład Glu $\rightarrow$ Ser, co Ser $\rightarrow$ Glu). Tak skonstruowana tablica nie zawierała jednak informacji o aminokwasach, które nie uległy substytucji, a więc tych, które są stabilizowane przez selekcję. Aby wprowadzić dodatkową informację do macierzy obliczono mutabilność dla każdego aminokwasu, według wzoru:

$$M_i = a_{i \rightarrow x} / F_i \qquad (12)$$

gdzie $a_{i \rightarrow x}$ to suma wszystkich zaobserwowanych podmian aminokwasu *i*, a F_i to całkowita liczba wystąpień *i*-tego aminokwasu we wszystkich analizowanych sekwencjach.

Porównanie relatywnych mutabilności wszystkich 20 aminokwasów budujących białka prowadzi do ciekawych wniosków. Najbardziej mutabilne aminokwasy to Asn, Ser, Asp i Glu, najstabilniejsze to Trp i Cys, czyli aminokwasy o charakterystycznym kształcie i właściwościach chemicznych, mające istotny wpływ na strukturę budowanego białka.

Aby zawrzeć informację o poszczególnych rodzajach przejść i o mutabilności poszczególnych aminokwasów, Dayhoff opracowała tzw. *„mutation probability matrix”*, czyli tablicę prawdopodobieństwa przejść aminokwasowych *(Schwartz i Dayhoff 1978).* Element tej macierzy, M_{ij}, określa prawdopodobieństwo, że aminokwas w kolumnie *j* zostanie zastąpiony przez aminokwas w wierszu *i* w ciągu określonego w jednostkach ewolucyjnych czasu (PAM-Point of Accepted Mutation lub Percent of Accepted Mutation) i jest obliczany wg wzoru:

$$M_{ij} = \lambda\, m_j\, A_{ij} / \Sigma_i A_{ij} \qquad (13)$$

gdzie A_{ij} to liczba wszystkich zaobserwowanych podmian aminokwasu *j* przez aminokwas *i* , m_j to relatywna mutabilność *j*-tego aminokwasu a λ to stała proporcjonalności. Elementy diagonalne macierzy są wyliczone na podstawie równania:

$$M_{jj} = 1 - \lambda\, m_j \qquad (14)$$

W ten sposób suma elementów każdego wiersza została znormalizowana do jeden - prawdopodobieństwa zajścia zdarzenia pewnego. Dystans ewolucyjny 1 PAM, dla którego skonstruowano tablicę odpowiada jednej zaakceptowanej substytucji na 100 aminokwasów. Aby uzyskać tablice przejść dla większych odległości filogenetycznych, podnosi się macierz PAM1 do odpowiedniej potęgi (czyli mnoży się macierz przez samą siebie odpowiednią ilość razy). W ten sposób otrzymano np. PAM75 i PAM250, które odpowiadają 75 i 250 podstawieniom na 100 aminokwasów. Średnia wartość 2,5 substytucji na aminokwas jest możliwa, gdy weźmiemy pod uwagę podstawienia wielokrotne. Oczywiście sekwencje odległe o 75 PAM nie różnią się w 75% pozycji. Procentowe różnice są znacznie mniejsze, gdyż nie wszystkie rejony białka ewoluują w jednakowym tempie. Niektóre aminokwasy pozostają niezmienne, podczas gdy inne ulegają wielokrotnym podstawieniom.

Tablic Dayhoff używano do symulacji losów konkretnych sekwencji po zadanym czasie ewolucji. Algorytm obejmował dwa etapy, każdy z nich wymagał użycia liczby losowej. Najpierw wybierano aminokwas do mutacji, przy czym prawdopodobieństwo wyboru było

zależne od mutabilności danego aminokwasu. Następnie losowano rodzaj substytucji, jakiej miał on podlegać, zgodnie z prawdopodobieństwami przejść z macierzy Dayhoff. Otrzymywano w ten sposób prawdopodobny obraz danej sekwencji po zadanym czasie ewolucji *(Dayhoff i współpr. 1978).*

Macierze PAM są tablicami mutacyjno-selekcyjnymi, opartymi o dane eksperymentalne, są więc ważnym punktem odniesienia dla dwuetapowych symulacji opartych na tablicach czysto mutacyjnych prowadzonych przez nasz zespół. Otrzymanie wysokiego współczynnika korelacji między tablicami PAM a tablicami, które można uzyskać z porównania sekwencji wygenerowanych przez program symulacyjny, byłoby cennym potwierdzeniem poprawności i zgodności opracowanego algorytmu z zachodzącymi w przyrodzie procesami.

```
C  12
S   0  2
T  -2  1  3
P  -3  1  0  6
A  -2  1  1  1  2
G  -3  1  0 -1  1  5
N  -4  1  0 -1  0  0  2
D  -5  0  0 -1  0  1  2  4
E  -5  0  0 -1  0  0  1  3  4
Q  -5 -1 -1  0  0 -1  1  2  2  4
H  -3 -1 -1  0 -1 -2  2  1  1  3  6
R  -4  0 -1  0 -2 -3  0 -1 -1  1  2  6
K  -5  0  0 -1 -1 -2  1  0  0  1  0  3  5
M  -5 -2 -1 -2 -1 -3 -2 -3 -2 -1 -2  0  0  6
I  -2 -1  0 -2 -1 -3 -2 -2 -2 -2 -2 -2 -2  2  5
L  -6 -3 -2 -3 -2 -4 -3 -4 -3 -2 -2 -3 -3  4  2  6
V  -2 -1  0 -1  0 -1 -2 -2 -2 -2 -2 -2 -2  2  4  2  4
F  -4 -3 -3 -5 -4 -5 -3 -6 -5 -5 -2 -4 -5  0  1  2 -1  9
Y   0 -3 -3 -5 -3 -5 -2 -4 -4 -4  0 -4 -4 -2 -1 -1 -2  7 10
W  -8 -2 -5 -6 -6 -7 -4 -7 -7 -5 -3  2 -3 -4 -5 -2 -6  0  0 17
   C  S  T  P  A  G  N  D  E  Q  H  R  K  M  I  L  V  F  Y  W
```

Ryc. 7. Tablica wartości log odds dla PAM250 – rezultat matematycznych przekształceń danych z tablicy PAM1. Elementy tablicy wyrażają stosunek rzeczywiście zachodzących substytucji do ich losowych częstości: wartość –10 oznacza, że obserwowana częstość podstawienia danej pary aminokwasów stanowi 1/10 ich spodziewanej liczby, wynik +2 oznacza, że dane substytucja zachodzi 1,6 raza częściej niż należało się spodziewać.

1.7. Zagadnienie kodu genetycznego - historia odkrycia i zmiany koncepcji

Kod genetyczny, podstawowy szyfr natury, został złamany dość szybko. Przyporządkowanie kolejnym kodonom odpowiednich aminokwasów miało charakter bezdyskusyjny (doświadczenia *Nirenberga i Matthei*) i zajęło około pięciu lat (1961-66), samo rozszyfrowanie kodu było jednak poprzedzone latami teoretycznych spekulacji i prób dopasowania koncepcji funkcjonowania kodu do powiększających się zasobów danych eksperymentalnych.

Za prekursora koncepcji cząsteczki chemicznej jako nośnika informacji należy uznać Schrödingera, który już w 1943 roku stwierdził, że materialna podstawa dziedziczenia musi mieć postać aperiodycznego kryształu, makromolekularnej struktury, w której pozycje są stałe, ale zajmujące je podjednostki mogą zachować pewną różnorodność, co dawałoby takiej strukturze możliwość gromadzenia informacji. Początki badań nad genezą kodu sięgają 1954 roku, kiedy kosmolog George Gamov przedstawił tezę o powstaniu kodu dzięki bezpośrednim oddziaływaniom DNA-białko. Bezpośrednio po odkryciu struktury DNA starano się powiązać białka z kwasem nukleinowym na drodze bezpośredniego przestrzennego dopasowania cząsteczek *(Woese 1967; Ycas 1969; Osawa 1995)*. Jedną z propozycji była koncepcja „diamentowego kodu" autorstwa G. Gamova (aminokwas miał dopasowywać się do odpowiedniego wgłębienia tworzonego przez parę zasad komplementarnych i dwa nukleotydy z bezpośredniego sąsiedztwa), która prawidłowo ograniczała pulę rozpoznawanych aminokwasów do dwudziestu, ale opierała się na błędnym założeniu, że synteza białka odbywa się bezpośrednio na matrycy DNA jądrowego. Założenie to zostało szybko podważone, a Gamov został zmuszony do zmodyfikowania swojej teorii. Zaproponowany wkrótce po odkryciu mRNA „kod trójkątny" był już kodem trójkowym - zawężał oddziaływanie kwas nukleinowy – aminokwas do trzech sąsiadujących nukleotydów mRNA. Dopuszczał jednak zachodzenie na siebie tripletów sekwencji i kładł nacisk nie tyle na sekwencję kodonów, co na ich kompozycję. Istotnym osiągnięciem była jednak sama koncepcja kodu trójkowego. Obserwacje Brennera z 1957 roku (analiza częstości występowania różnych dwupeptydów w białkach) doprowadziły wkrótce do odrzucenia założeń o zachodzeniu kodonów na rzecz kodu niezachodzącego. Kolejnym krokiem do współczesnej koncepcji kodu genetycznego było odejście od hipotezy o bezpośrednim oddziaływaniu cząsteczek aminokwasów i RNA i powstanie hipotezy cząsteczki adaptorowej *(Crick 1957)*, a wkrótce potem koncepcji „bezprzecinkowości" kodu, która zakładała, że żaden kodon poza ciągłą, bezpośrednią ramką odczytu nie mógł być interpretowany jako

znaczący. Tak jak w przypadku kodu diamentowego i trójkątnego, liczba możliwych do zakodowania aminokwasów i tym razem wynosiła dwadzieścia. Ostatnią odkrytą cechą kodu, być może najbardziej znaczącą dla ewolucji, była jego degeneracja. Obserwowane w kodzie prawidłowości i zbieżności są na tyle zadziwiające, że doprowadziły stopniowo, w połączeniu z rozwojem informatyki i sztucznych systemów selekcji, do odrzucenia hipotezy *„frozen accident”* Cricka *(1968).* Jednocześnie zaobserwowane odstępstwa od uniwersalnego kodu (dotyczące zwłaszcza organelli i małych genomów) dostarczyły argumentów świadczących za tezą, że kod genetyczny nie jest bezwzględnie konserwatywny, jak do tej pory zakładano i w pewnych okolicznościach może ewoluować. Porządek kodu i jego zastanawiające właściwości zdające się minimalizować zarówno możliwość błędu, jak i skutki pojedynczej mutacji, są pochodną tylko jednej jego cechy – degeneracji wynikającej z nadmiarowości *(Sonneborn 1965; Woese 1965; Zuckerkandl i Pauling 1965).* Wszystkie aminokwasy, z wyjątkiem metioniny i tryptofanu, są kodowane przez więcej niż jeden kodon, co więcej, kodony jednego aminokwasu leżą koło siebie w tablicy kodu, tworząc układy podwójnie (tzw. *semiboksy*) lub poczwórnie zdegenerowane (tzw. *boksy*) *(Ryc.9).* W przypadkach, kiedy aminokwas ma dwa kodony, różnią się one tylko nukleotydem w trzeciej pozycji, i to w ten sposób, że tranzycja nie zmienia sensu kodonu. Oczywiście w przypadku kompletnych boksów jakakolwiek substytucja w trzeciej pozycji kodonu nie zmienia jego znaczenia.

Degeneracja kodu zdaje się być sterowana zawartością GC w kodonach. Ma to bardzo proste chemiczne wytłumaczenie: wiązania wodorowe między guaniną a cytozyną są potrójne i znacznie mocniejsze niż wiązania adenina – tymina. Jeżeli dublet, czyli dwa pierwsze nukleotydy w kodonie, jest parą GC lub CG, tworzy zawsze kompletne boksy *(Ryc.9.),* natomiast jeżeli zasadami w dublecie są A lub T, powstają boksy podzielone, co łatwo wyjaśnić zdając sobie sprawę z faktu, że kodony GCN wiążą się na tyle silnie do swoich antykodonów, że trzeci nukleotyd tripletu nie odgrywa już właściwie żadnej roli, natomiast wiązanie par AT są na tyle słabe, że potrzebują wzmocnienia przez trzeci nukleotyd *(Lagerkvist 1978; Lagerkvist 1980).* Inaczej sformułowaną wersją tej prawidłowości jest tak zwana „zasada Jayaram’a”, według której, jeżeli dany dublet tworzy boks, jego konjugat tworzy boksy podzielone (konjugat danej zasady jest zasadą o przeciwnych do niej właściwościach, np. konjugat tworzącej potrójne wiązania puryny jest tworzącą podwójne wiązanie pirymidyną itd.). *(Jayaram 1997).*

Pochodzenie tych prawidłowości jest związane najprawdopodobniej z początkami kodu genetycznego, być może sięga jeszcze okresu sprzed wykształcenia współczesnych akceptorowych cząsteczek tRNA. Znaczenie dwóch pierwszych pozycji i częściowa lub

zupełna degeneracja trzeciej zdają się być reliktem z czasów bezpośrednich oddziaływań między aminokwasami a RNA, które na drodze ewolucji zostały wzmocnione i przejęte przez specyficzne właściwości stereochemiczne tRNA *(Szathmáry 1991).*

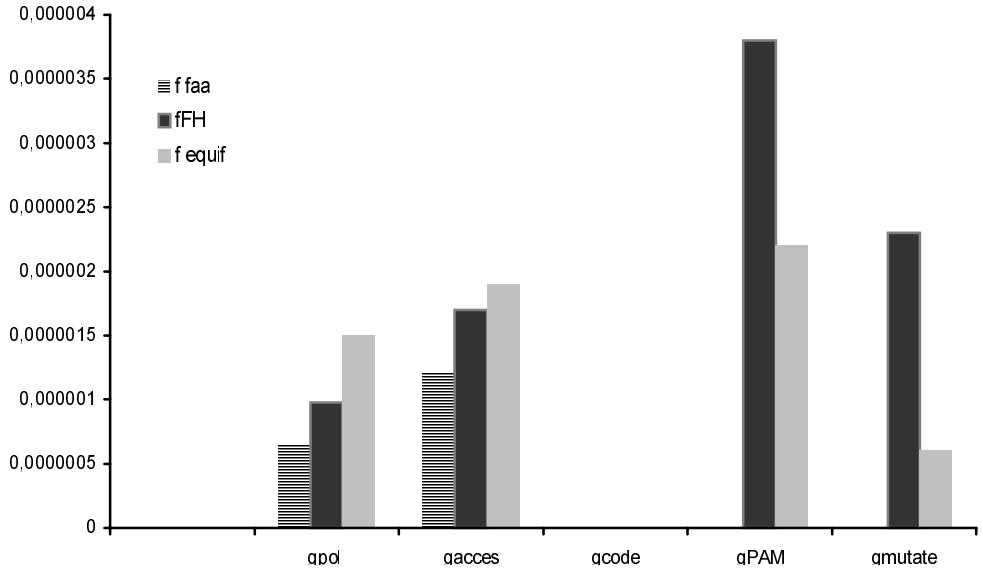
Hydrofobowość aminokwasów zmienia się w obrębie kodu bardzo regularnie. Pięć najbardziej hydrofobowych aminokwasów (fenyloalanina, izoleucyna, metionina, walina i leucyna) ma tyminę w drugiej pozycji kodonu, trzy najbardziej do siebie zbliżone pod względem hydrofobowości (leucyna, izoleucyna i walina) różnią się tylko nukleotydem w pierwszej pozycji kodonu. Sześć najsilniej hydrofilowych aminokwasów (histydyna, lizyna, asparagina, asparaginian, glutamina, glutaminian) ma adeninę w drugiej pozycji *(Woese 1965; Volkenstein 1966; Woese i współpr. 1966)*. Rezultatem tego jest ogólna prawidłowość, że aminokwasy o komplementarnych antykodonach mają zwykle przeciwne wartości hydrofobowości *(Volkenstein 1966; Blalock i Smith 1984).* Aminokwasy kodowane przez triplety z cytozyną w drugiej pozycji przyjmują najczęściej pośrednie wartości hydrofobowości między tymi, które zawierają adeninę i tyminę w tych pozycjach. Ponadto aminokwasy dzielące dublet cechują zwykle bardzo zbliżone właściwości polarne, z wyjątkiem pary cysteina – tryptofan *(Woese i współpr. 1966)*. Inne analizy *(Sjöström i Wold 1985)* wykazały, że także punkt izoelektryczny poszczególnych aminokwasów zmienia się w obrębie kodu z dużą regularnością.

Kodony aminokwasów o podobnych właściwościach chemicznych są ze sobą silnie związane. Kwaśne aminokwasy: asparaginian i glutaminian dzielą ten sam dublet, a ich amidowe pochodne, choć nie mają wspólnego dubletu, różnią się tylko pojedynczymi nukleotydami w pierwszej pozycji kodonu.

Trzy zasadowe aminokwasy (lizyna, arginina i histydyna) różnią się także pojedynczymi nukleotydami w kodonie. Podobnie niewielkie różnice obserwujemy między kodonami w grupie aminokwasów aromatycznych (fenyloalanina, tyrozyna i tryptofan) i zawierających grupy hydrofilowe (seryna, treonina i tyrozyna).

Analogicznych reguł uporządkowania można się dopatrzyć w kodzie wiele, po części dlatego, że jest małym, silnie powiązanym zbiorem, część zależności jest więc po prostu losowa i możliwa do uzyskania także w losowych kodach. Dlatego istotniejsze z punktu widzenie naukowego są niedawne odkrycia odstępstw od uniwersalnego kodu oraz rozwój badań nad wczesnymi systemami replikacyjnymi, jak również opracowanie systemu SELEX i możliwości syntetyzowania aptamerów (krótkich sekwencji RNA o zadanych właściwościach enzymatycznych) *(Knight 2001).* Niebagatelne znaczenie ma także dostępność danych pochodzących z symulacji komputerowych. Przykładem zastosowania podejścia

symulacyjnego w połączeniu z metodami analityczno-statystycznymi jest praca z 2001 roku, której autorzy *(Gilis i współpr. 2001)* usiłują określić poziom dostosowania oryginalnego kodu względem kodów wygenerowanych losowo. Podstawą metody jest przypisanie dowolnemu zestawowi kodonów i aminokwasów funkcji Φ określającej tzw. *„fitness"* kodu mierzoną jako średnia z funkcji opisujących różnicę między aminokwasem a i a' dla wszystkich kodonów i wszystkich możliwych błędów. Autorzy uwzględniali w swoich analizach ilość substytucji, jakie muszą zajść, aby dany aminokwas przeszedł w inny, jak również rodzaj tych substytucji (tranzycje i transwersje), a nawet częstości, z jakimi dane aminokwasy występują w żywych komórkach. Różnicę między aminokwasami obliczano kilkoma sposobami: jako funkcję g^{hydro} (różnice hydrofobowości liczone dwoma sposobami), g^{mutate} (macierz M[a, a'] uwzględniająca zmianę energii swobodnej tworzenia cząsteczki białka w wypadku zmiany jednego aminokwasu w jego sekwencji) i g^{PAM} (funkcja odwołująca się do wartości z tablic $PAM_{74\text{-}100}$). Po obliczeniu Φ dla kodu naturalnego i 10^9 losowo wygenerowanych kodów otrzymano dość jednoznaczne wyniki *(Gilis i współpr. 2001)*: dla funkcji g^{hydro} frakcja kodów, które okazały się lepiej przystosowane od rzeczywistego wyniosła od 0,5 do 1 na 10^6. Dla funkcji g^{mutate} (uwzględniającej szereg właściwości stereochemicznych białka, takich jak kąty torsyjne i specyficzne właściwości cysteiny, proliny i glicyny dla struktury trzeciorzędowej) na 10^9 losowych kodów tylko dwa uzyskały wynik lepszy niż kod rzeczywisty. Dla g^{PAM} wyniki były zbliżone do tych otrzymanych dla g^{mutate}. Na koniec wzięto pod uwagę strukturę kodu, tworząc funkcję g^{code} (a, a') uwzględniającą minimalną liczbę zasad, które muszą zmutować, aby aminokwas *a* przeszedł w *a'*. Wśród 10^9 losowych kodów nie znalazł się ani jeden, który uzyskałby wyższe Φ obliczane w oparciu o funkcję g^{code} niż kod uniwersalny *(Ryc.8.)*.



Ryc. 8.
Frakcje kodów losowych, które uzyskały *„fitness”* lepsze (wyższą wartość funkcji Φ) od kodu uniwersalnego *(Gilis i współpr. 2001).*
Funkcja Φ była obliczana trzema sposobami jako Φ^{FH} $(1/64\ \Sigma^{64}_{c=1}\ \Sigma^{64}_{c'=1}\ p(c/c')g(a(c),\ a(c')))$, Φ^{faa} $(\Sigma^{64}_{c=1}p(a(c)/n(c))\Sigma^{64}_{c'=1}g(a(c),a(c')))$, z uwzględnieniem częstości aminokwasów komórce i jako Φ^{equif} $(1/20\ \Sigma^{64}_{c=1}\ 1/n(c)\ \Sigma^{64}_{c'=1}\ g(a(c),\ a(c')))$, przy założeniu jednakowych częstości występowania dla wszystkich aminokwasów.

Podsumowując, kod genetyczny okazał się o wiele bardziej odporny na pojedyncze mutacje niż jego losowo wygenerowane odpowiedniki, co znaczy, że bardzo skutecznie minimalizuje koszty pojedynczych substytucji. Ostatnie odkrycia potwierdzają silny związek statystyczny między przynajmniej kilkoma kodonami a sekwencjami będącymi miejscami wiązania odpowiadających im aminokwasów do rybozymów (najsilniejsze powinowactwo stwierdzono dla argininy) *(Knight 1998)*, co sugeruje, że w pierwszych etapach ewolucji kodu dużą rolę mogły odegrać bezpośrednie oddziaływania RNA-białko. Czas ustabilizowania się i optymalizacji kodu nadal nasuwa szereg wątpliwości, ogólnie uznaje się jednak, że przyporządkowanie kodonów odpowiednim aminokwasom nastąpiło dość wcześnie, prawdopodobnie jeszcze w świecie RNA. Współczesne warianty kodu powstały przypuszczalnie na drodze niemal neutralnych mutacji standardowego kodu. Nie ma ostatecznych dowodów, że modyfikacje kodu odkryte w małych genomach są po prostu adaptacją do zredukowanej puli tRNA. Ale jest wiele przesłanek przemawiających za tezą, że do zmiany kodu dochodziło drogą ewolucji przez utratę kodonu. W niewielkich genomach istnieje pewne prawdopodobieństwo, że niektóre kodony mogą zupełnie zniknąć z sekwencji w skutek działania kierunkowej presji mutacyjnej. Jeżeli pojawią się znowu (np. po zmianie presji mutacyjnej), może się zdarzyć, że aparat translacyjny przypisze im inny sens.

	T		C		A		G	
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
	TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys
	TTA	Leu	TCA	Ser	TAA	Amber	TGA	Opal
	TTG	Leu	TCG	Ser	TAG	Ochre	TGG	Trp
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg
	CTC	Leu	CCC	Pro	CAC	His	CGC	Arg
	CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
	CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
	ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
	ATA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
	ATG	Met	ACG	Thr	AAG	Lys	AGG	Arg
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
	GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly
	GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly
	GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly

kodony stop

kodony pojedyncze

kodon start

aminokwasy kodowane przez 6 kodonów

semiboxy

boxy czterokodonowe

AT dublety zasad o podwójnym wiązaniu

AC dublety mieszane z pirymidyną w drugiej pozycji

TG dublety mieszane z puryną drugiej pozycji

GC dublety zasad o wiązaniu potrójny

Ryc.9. Podział kodu genetycznego na boxy i semiboxy z zaznaczeniem roli dubletów w tworzeniu grup kodonów. Kodony o dubletach złożonych z GC tworzą zawsze kompletne boxy, dublety AT tworzą boxy podzielone.

Innymi przyczynami odstępstw od standardowego kodu są mutacje w akceptorowym tRNA lub białkach rozpoznających kodony STOP oraz białkach pośredniczących w procesach inicjacji i terminacji translacji. Wysiłki badaczy zajmujących się obecnie zagadnieniami związanymi z kodem genetycznym są skierowane na dwa wyraźne cele: część naukowców dąży do wykazania, że pozornie „zamrożony" kod jest jednak zdolny do ewolucji, część usiłuje udowodnić, że istniejący kod jest optymalny i najlepszy z możliwych. O ile do celu pierwszej grupy nie można mieć żadnych metodologicznych zastrzeżeń, o tyle podstawy naukowe konstruowania dowodów na ewolucyjną doskonałość uniwersalnego kodu budzą wątpliwości. Trudno udowodnić, że jedyny istniejący egzemplarz jest optymalnym przedstawicielem swojego rodzaju bez obiektywnego materiału porównawczego.

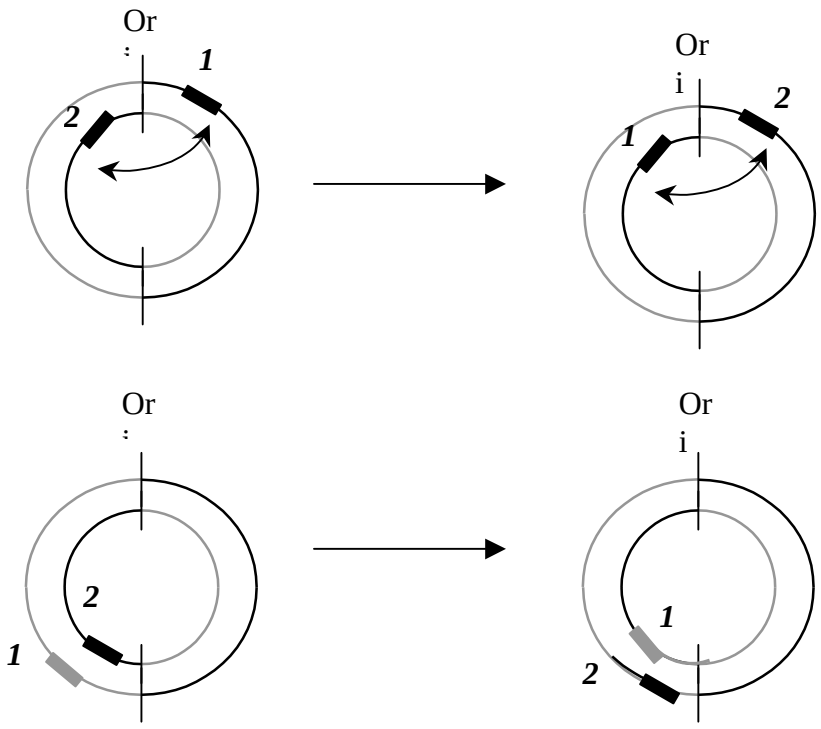
1.8. Mechanizmy ewolucji strukturalnej genomu prokariotycznego

Genomy prokariotyczne, w przeciwieństwie do eukariotycznych, cechuje bardzo wysoka gęstość kodowania, około 90% genomu stanowią sekwencje kodujące, zorganizowane często w operony. Drugą ważną cechą genomów bakteryjnych jest asymetria składu nukleotydowego wprowadzana przez trendy związane z replikacją, która różnicuje warunki panujące na nici wiodącej i opóźniającej. Wydaje się więc, że presja na zachowanie struktury genomu bakteryjnego powinna być bardzo silna. Obserwując położenie homologicznych genów i operonów w genomach różnych gatunków w obrębie linii ewolucyjnych bakterii należy jednak stwierdzić, że jest ono bardzo zmienne *(Mushegian i Koonn 1996; Casjens 1998; Ochman i współpr. 2000; Mackiewicz i współpr. 2003)*. Struktura genomu bakteryjnego nie jest konserwatywna, porównując ortologi z daleko spokrewnionych genomów można zauważyć silną tendencję do zmiany ich położenia na nici *(Szczepanik i współpr. 2001; Mackiewicz i współpr. 2001; Rocha 2003)*. Takie przeniesienia genu powodują wzrost tempa mutacji i zwiększenie dywergencji między ortologami *(Mackiewicz i współpr. 2003)*.

Z jednej strony mamy więc do czynienia z licznymi czynnikami ograniczającymi zmienność struktury chromosomu, z drugiej z częstymi zjawiskami jego rearanżacji. Głównym mechanizmem takich rearanżacji jest intrachromosomalna rekombinacja homologiczna, która może prowadzić do delecji, duplikacji, translokacji lub inwersji *(Smith 1988, Roth i współpr. 1996, Romero i Palacios 1997)*. Większość z wymienionych mutacji strukturalnych jest jednak niekorzystna dla genomu i bywa zwykle szybko eliminowana przez selekcję. Najczęściej akceptowanymi zmianami w strukturze genomu są inwersje, one też są

uznawane za główny mechanizm ewolucji strukturalnej genomu bakteryjnego *(Liu i Sanderson 1996; Roth i współpr. 1996; Hughes 2000).* Analizy par ortologów wybranych z kompletnie zsekwencjonowanych genomów potwierdzają tezę, że duże rearanżacje w skali genomu są skutkiem inwersji *(Eisen i współpr. 2000; Tillier i Collins 2000a; Zivanovic i współpr. 2002).* Wyjątkiem jest wynik analizy genomu *Mycoplasma pneumoniae* i *Mycoplasma genitalium*, która wykazała szereg translokacji przy braku dowodów na zajście inwersji *(Himmelreich i współpr. 1997).* Inwersje dużych fragmentów chromosomu zachodzą głównie dzięki rekombinacji homologicznej między odwróconymi powtórzeniami *(Achaz i współpr. 2003)*, czyli krótkimi, identycznymi sekwencjami o odwróconych polarnościach, które flankują najczęściej strukturę transpozonu. Najprostszymi bakteryjnymi transpozonami są tzw. sekwencje insercyjne *IS*, składające się z pary terminalnych odwróconych powtórzeń i sekwencji kodującej transpozazę. Bardziej złożone transpozony typu *Tn* składają się z kilku genów, niekoniecznie związanych z transpozycją i są flankowane przez terminalne powtórzenia o odwróconej lub tej samej polarności. Analizy wielu genomów wykazały silny związek statystyczny między ilością odwróconych powtórzeń w genomie, a liczbą zachodzących w trakcie jego ewolucji inwersji, przy braku takiej zależności dla powtórzeń o zgodnych polarnościach *(Achaz i współpr. 2003)*. Rozmieszczenie sekwencji *IS* na chromosomie nie jest czysto losowe, wydaje się podlegać selekcji, za czym przemawia fakt, że sekwencje *IS* nigdy nie leżą wewnątrz operonu. Sekwencje insercyjne dzielą chromosom na „ruchome" moduły, które mogą się przemieszczać w jego obrębie.

Odwróconych powtórzeń używa się w badaniach laboratoryjnych jako mechanizmu indukującego inwersje. Większość inwersji uzyskanych w laboratorium *(Roth i współpr. 1996)* powoduje jednak obniżenie tempa wzrostu kolonii lub śmierć komórek, co świadczy o silnej selekcji, jakiej podlegają te zjawiska w naturze. Inwersje obserwowane w istniejących genomach, czyli te zaakceptowane przez selekcję, jeżeli zachodzą między różnymi replichorami, są zwykle symetryczne względem miejsca inicjacji replikacji *(Tillier i Collins 2000a; Suyama i Bork 2001).* Taka symetria pozwala na zachowanie długości replichory i położenia genu względem ORI *(Ryc.10.) (Mackiewicz i współpr. 2003).* Jeżeli odwrócone powtórzenia leżą w obrębie tej samej replichory, inwersja prowadzi do zmiany nici. Gdyby inwersje zachodziły z całkowicie losową częstością, liczba genów, których sensy leżą na nici wiodącej byłaby mniej więcej równa liczbie genów na nici opóźniającej. Tak jednak nie jest, u wielu bakterii mamy do czynienia z nierównomiernym rozłożeniem genów na niciach.



Ryc.10. Schemat przebiegu rekombinacji intrachromosomalnej między odwróconymi powtórzeniami zlokalizowanymi na tej samej nici symetrycznie wobec ORI (a) i na różnych niciach tej samej replichory (b).

Większość sekwencji kodujących leży na nici wiodącej, dotyczy to zwłaszcza genomów, które wykazują wysoką asymetrię składu nukleotydowego *(Mackiewicz 2003).* Jest to spowodowane czynnikami selekcyjnymi, związanymi, jak wykazali niedawno *Rocha* i *Danchin (2003)*, raczej z istotnością transkryptu danego genu dla funkcjonowania komórki niż z presją na poziom jego ekspresji. Wyniki analizy porównawczej położenia genów odpowiedzialnych za podstawowe procesy metaboliczne komórki bakteryjnej pozwoliło na wyciągnięcie wniosku o decydującej roli funkcji genu dla organizacji genomu bakterii.

Położenie genu ma więc duże znaczenie selekcyjne, a istniejące struktury chromosomów są nieprzypadkowym tworem długiej ewolucji. Aby odpowiedzieć na pytanie o ewolucję strukturalną genomu, należy wziąć pod uwagę szereg czynników kształtujących środowisko genów, czyli warunki ewolucji w złożonym systemie, jakim jest chromosom bakteryjny.

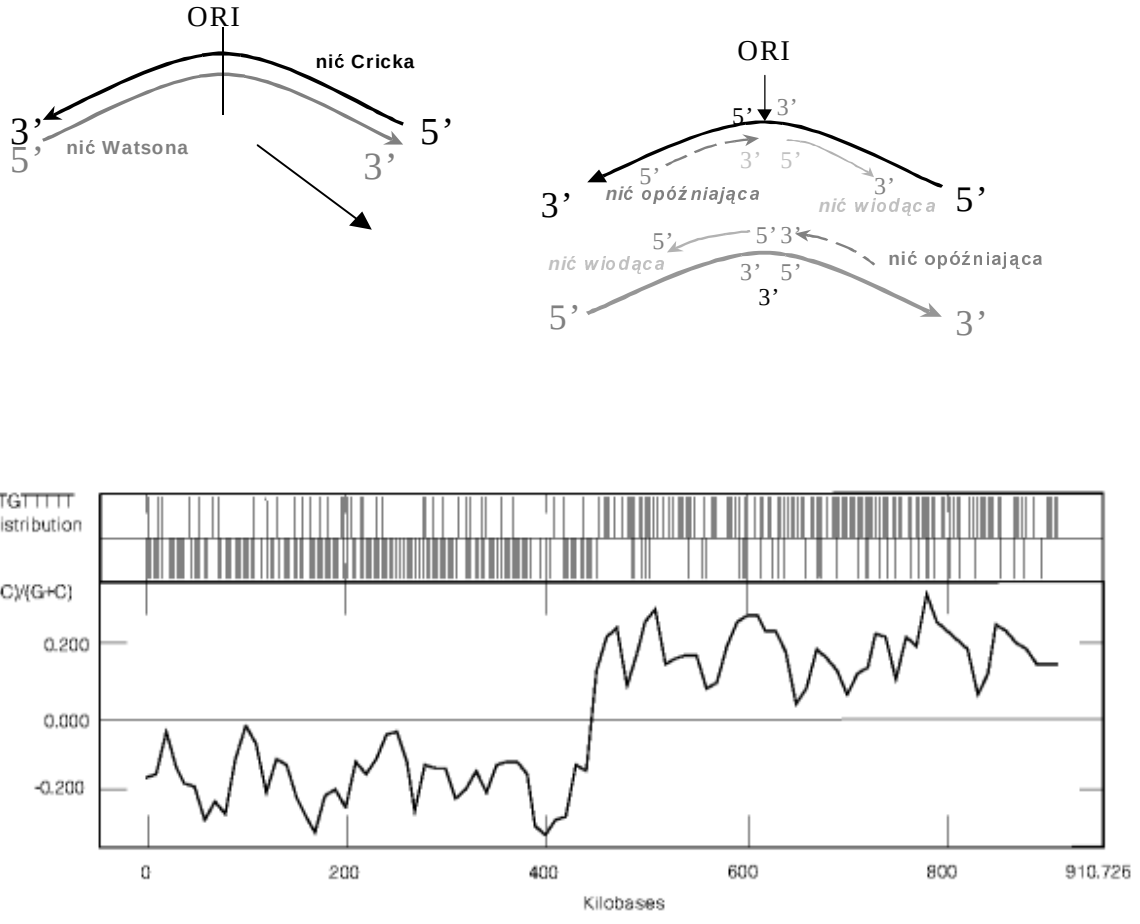
Właściwości genomu Borrelia burgorferi.

Genom *B. burgdorferi* B31 *(Fraser i współpr. 1997)*, czynnika etiologicznego tzw. *choroby z Lyme*, składa się z liniowej cząsteczki o długości 910 725 par zasad oraz 17 liniowych i kolistych plazmidów różnych rozmiarów o łącznej długości 533 000 par zasad. Na chromosomie znajduje się 850 genów kodujących głównie białka związane z replikacją, transkrypcją, translacją, transportem komórkowy. Chromosom nie zawiera natomiast genów odpowiedzialnych za funkcjonowanie szlaków metabolicznych syntezy związków organicznych, stąd przypuszczenie, że genom *B. burgdorferi* wyewoluował na drodze częściowej utraty funkcji od metabolicznie kompetentnego przodka.
Sekwencje kodujące białka stanowią 93% genomu, sekwencje kodujące RNA - 0.7%, sekwencje międzygenowe: 6.3%. Około 59% ORFów posiada zidentyfikowane homologi w bazach danych innych genomów, dla 29% nie znaleziono jeszcze homologów.

Średnia zawartość G+C w chromosomie *B. burgdorferi* wynosi ok. 28.6%, średnia długość ORFu wynosi około 330 kodonów. Analiza używalności kodonów wskazuje na przewagę tripletów bogatych w AU-. Najczęściej występującymi kodonami są: AAA (Lys, 8.1%), AAU (Asn, 5,9%), AUU (Ile, 5.9%), UUU (Phe, 5.7%), GAA (Glu, 5.0%), GAU (Asp, 4.2%). Średnia wartość punktu izoelektrycznego dla kodowanych białek wynosi 9.7, co wiąże się z wysoką częstością występowania kodonu dla zasadowej lizyny.

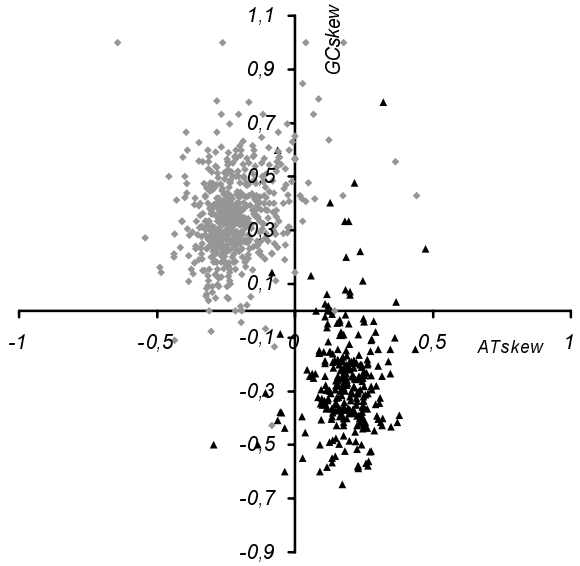
Mechanizm replikacji chromosomu i plazmidów *B. burgdorferi* został opisany w 1999 *(Picardeau i współpr. 1999)*. Jak wykazano metodą mapowania fizycznego miejsce inicjacji replikacji genomu *B. burgdorferi* znajduje się w środku liniowego chromosomu. 66% genów kodowanych na chromosomie *B. burgdorferi* jest transkrybowanych w kierunku od środka cząsteczki, co potwierdza wewnętrzne położenie ORI, przy założeniu, że optymalny kierunek transkrypcji jest zgodny kierunkiem replikacji *(Ryc.11)*. Za tym faktem przemawia także wyraźna asymetria genomu *B. burgdorferi* i widoczny przy analizie *GCskew* punkt zmiany trendu w środku genomu *(Ryc. 11)*. Przyjmując, że *ORI* znajduje się w środku chromosomu, na nici wiodącej leży 564 sekwencji kodujących, na nici opóźniającej prawie dwa razy mniej, tylko 286, co może wskazywać na preferencje co do położenia genów na chromosomie.

Ważną cechą genomu *B. burgdorferi* jest obecność licznych plazmidów i transpozonów oraz sekwencji o funkcjach transpozonowych. Wycofanie części kodowanej informacji genetycznej na plazmidy może wynikać z tendencji do minimalizacji rozmiarów nukleoidu, typowej dla organizmów pasożytniczych, zwłaszcza pasożytów wewnątrzkomórkowych. W genomie *B. burgdorferi* występuje łącznie 66 sekwencji o funkcjach transpozonowych (w tym 11 w chromosomie), co jest ciekawe biorąc pod uwagę, że kompletne transpozony są obecne wyłącznie na plazmidach, dokładnie na dziewięciu z nich (lp17, lp25, lp28(1-4), lp56, lp36 i lp38), stanowiąc ok. 0,74% całego genomu (*www.ncbi.nlm.nih.gov*). Wszystkie chromosomowe sekwencje traspozonowe są homologami sekwencji plazmidowych, co może świadczyć o licznych zjawiskach rearanżacji i wymianie informacji między genoforem a plazmidami.



Ryc.11. Model replikacji chromosomu *Borrelia burgdorferi*, zgodny z teorią umiejscawiającą ORI pośrodku chromosomu. Poniżej wykres *GC skew* i dystrybucji oktamerów *TTGTTTT* wzdłuż chromosomu *B. burgdorferi*. Punkt odwrócenia trendu odpowiada usytuowaniu ORI (*Fraser i współpr. 1997*).

Różnice w składzie trzecich pozycji kodonów genów z nici wiodącej i opóźniającej *B. burgdorferi* przedstawia *Ryc. 12.* Punkty na dwuwymiarowej płaszczyźnie, której osie wyznaczają wartości *AT i GC skew* dla trzecich pozycji w kodonach, reprezentują poszczególne geny z nici wiodącej (szare kwadraty) i opóźniającej (czarne trójkąty). Geny grupują się wokół centrów położonych w dwóch przeciwległych ćwiartkach układu współrzędnych. W ćwiartce II dominują sekwencje z nici wiodącej, w ćwiartce IV – z nici opóźniającej. Nieliczne punkty odbiegają od wyraźnie zaznaczonego schematu, zajmując położenia pośrednie, w ćwiartce I lub III. Są to sekwencje, które najprawdopodobniej uległy niedawnej inwersji. Na podstawie odległości punktów od odpowiednich centrów rozkładu można w przybliżeniu określić liczbę sekwencji, które zmieniły nić. Wyniki analizy przedstawiono w *Tab.5. (Mackiewicz i współpr. 2003).*



Ryc.12. Rozkład sekwencji kodujących *B. burgdorferi* w przestrzeni *ATskew/ GCskew* dla trzecich pozycji w kodonie. Szare punkty reprezentują geny z nici wiodącej, czarne – geny, których sensy leżą na nici opóźniającej.

Tab.5. Oszacowanie ilości inwersji genów ze zmianą nici, które zaszły w genomie *B. burgdorferi (Mackiewicz i współpr. 2003)*

liczba genów na nici wiodącej	liczba genów na nici opóźniającej	**geny, które zmieniły nić:**		N_{w-o} Inv_w/N_w	N_{o-w} Inv_o/N_o
		z wiodącej na opóźniającą Inv_w	z opóźniającej na wiodącą Inv_o		
564	286	24	27	0,043	0,094

Wymienione właściwości genomu *B. burgdorferi* predysponują ją do pełnienia funkcji organizmu modelowego w badaniach nad asymetrią DNA i jej wpływem na ewolucję molekularną.

2.CEL PRACY

Celem przeprowadzonych badań i analiz symulacyjnych było skonstruowanie modelu pozwalającego na symulację ewolucji genomu prokariotycznego z uwzględnieniem asymetrycznej, kierunkowej presji mutacyjnej i selekcyjnej.

3. MATERIAŁY I METODY

3.1. Sekwencje wykorzystane w konstrukcji modelu

Podstawą konstrukcji modelu symulacyjnego ewolucji genomu prokariotycznego była kompletna sekwencja *Borrelia burgdorferi B31 (Spirochaete) (Fraser i współpr. 1997)* zdeponowana na stronie National Center for Biotechnology Information *(Benson i współpr. 2000) (www.ncbi.nlm.nih. gov).* Opracowując model, wykorzystano informację zawartą w całym chromosomie, biorąc pod uwagę zarówno 850 sekwencji kodujących białka i rRNA, jak i sekwencje międzygenowe, uwzględniając położenie ORI i lokalizację poszczególnych genów na chromosomie.

Drugim genomem, z którego sekwencji skorzystano, budując model, był genom *Treponema pallidum*. Ze zbioru sekwencji aminokwasowych dostępnych na stronie *www.tigr.org* wybrano 433 ortologi odpowiednich sekwencji *B. burgdorferi.*

3.2. Moduł mutacyjny

Presję mutacyjną działającą na genom podlegający ewolucji określono korzystając z tablicy przejść nukleotydowych uzyskanej dla *B. burgdorferi* przez Kowalczuk i współpracowników *(2001b) (patrz. rozdz.1.5, Tab.4.).* Macierz substytucji otrzymano porównując ilość podstawień zaszłych w sekwencjach kodujących i homologicznych do nich pseudogenach, które z założenia nie podlegają presji selekcyjnej. Mutacje nagromadzone w sekwencjach pseudogenów powinny więc być odbiciem czystej presji mutacyjnej. Tak otrzymana macierz substytucji jest obrazem presji charakterystycznej dla genomu, na którego podstawie została skonstruowana.

3.3. Algorytm symulacyjny

Kompletną sekwencję *B. burgdorferi* poddawano symulacji według algorytmu typu Monte Carlo (*aneks*), którego działanie można podzielić na kilka zasadniczych etapów. Pierwszy obejmuje rozpoznawanie i wycinanie sekwencji kodujących z pliku wejściowego według podanych współrzędnych określających dokładne położenie genów oraz ORI. Program rozpoznaje jednocześnie nić, na której leży sens danego genu i grupuje osobno sekwencje odcinków kodujących położonych na nici wiodącej i opóźniającej. Takie grupy sekwencji przechodzą do etapu następnego, którym jest wprowadzanie mutacji. Etap ten obejmuje dwa nakładające się na siebie procesy: losowanie nukleotydu do mutacji (według zadanego prawdopodobieństwa p_{mut}) oraz podstawienie wybranego nukleotydu przez inny nukleotyd z prawdopodobieństwem określonym przez odpowiednią tablicę mutacji. Nie każdy wylosowany do mutacji nukleotyd ulega substytucji, jest to związane z niezerowymi wartościami diagonali macierzy mutacyjnej, a tym samym niezerowym prawdopodobieństwem przejść tożsamościowych.

Wprowadzone mutacje są zliczane i dostępne do dalszej analizy jako *mutacje wpadające,* czyli będące wynikiem działania czystej presji mutacyjnej.

Zmutowana sekwencja przechodzi do następnego etapu symulacji, w którym każdy gen przechodzi przez sito selekcyjne. Stosowany algorytm ma możliwość sprawdzania szeregu właściwości kodowanego przez gen białka, są to: sekwencja kodonowa, aminokwasowa, punkt izoelektryczny, hydrofobowość, generowanie startów i stopów W poniższej pracy skupiono się głównie nad testowaniem składu aminokwasowego, jako czynnika selekcyjnego w ewolucji genów, choć podjęto także próby analizy innych parametrów selekcyjnych.

Na etapie selekcji algorytm oblicza skład aminokwasowy każdego genu przed i po mutacji, a następnie oblicza parametr wyrażony wzorem:

$$\sum_{a=1}^{20} |f_a(0) - f_a(t)| \qquad (15)$$

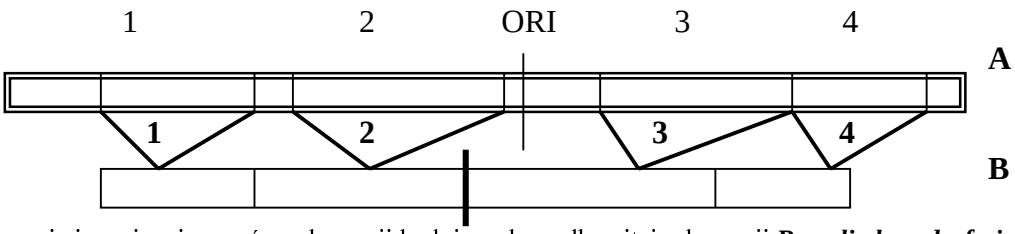
gdzie $f_a\ (0)$ to frakcja danego aminokwasu w oryginalnej sekwencji, a $f_a(t)$ frakcja tego samego aminokwasu w sekwencji poddanej mutacjom w kroku t .

Modelowanie ewolucji genomu według opisywanego algorytmu polega właściwie na symulacji koewolucji dwóch bliźniaczych sekwencji, które na zmianę podlegają mutacji i selekcji. Jeżeli wartość parametru *(15)* przekroczy ustaloną tolerancję T (sposób wyznaczenia zakresu tolerancji dla genów będzie omówiony w *rozdz. 3.4.*), gen jest zastępowany przez

swojego „ortologa”, który przeżył poprzedni krok symulacji w koewoluującej sekwencji. Zakładamy, że zmiana nie przekraczająca granic tolerancji nie powoduje istotnej selekcyjnie zmiany funkcji kodowanego białka. Po etapie selekcji algorytm zlicza zaakceptowane mutacje, jak również liczbę podmienionych genów. Ten ostatni parametr służy następnie do szacowania kosztów ewolucji. Każda eliminacja genu odpowiada w naturze śmierci jednego organizmu- podmieniony gen nie bierze już udziału w ewolucji, wygrywa i powiela się ten, który zajął jego miejsce. Procesy mutacji i selekcji zachodzą zadaną ilość razy, przy czym wzorcem selekcyjnym jest ciągle oryginalna sekwencja *B. burgdorferi.* Iteracja dotyczy obu współewoluujących sekwencji. W jednym kroku symulacji przez etapy mutacji i selekcji przechodzi najpierw genom *A*, a następnie genom *B*. W czasie, kiedy dana pula genów (A lub B) nie ulega ewolucji, stanowi „magazyn” sprawnych ortologów, z którego może korzystać bliźniacza sekwencja po przejściu przez etap selekcji.

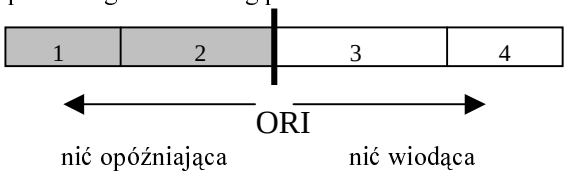
Program symulacyjny został napisany języku programowania C++ przez prof. M. R. Dudka. Język C++ umożliwia tworzenie kodu, który po odpowiedniej kompilacji, jest czytelny dla każdego komputera PC. Większość parametrów jest wprowadzana przez linię polecenia, ich zmiana nie wymaga bezpośredniej ingerencji w kod, co mogłoby być z czasem niebezpieczne dla integralności programu.

Język C++, udoskonalona wersja ANSI C, różni się od swojego poprzednika głównie wprowadzeniem możliwości programowania strukturalnego, bez czego tworzenie skomplikowanego oprogramowania byłoby znacznie trudniejsze. Główną zaletą tej metody jest możliwość tworzenia klas (odpowiedników struktur w ANSI C), czyli obiektów, stanowiących zbiór obiektów i stosowanych na nich metod (funkcji). Klasa jest jednostką kodu, na której można wykonać szereg podstawowych operacji i traktować jako całość. Dzięki zastosowaniu klas udało się zachować spójność struktury skomplikowanego programu.

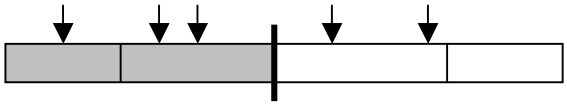


znajdowanie i wycinanie sensów sekwencji kodujących z całkowitej sekwencji ***Borrelia burgdorferi***
A – sekwencja oryginalna ***Borrelia burgdorferi***
B – wycięte sekwencje kodujące

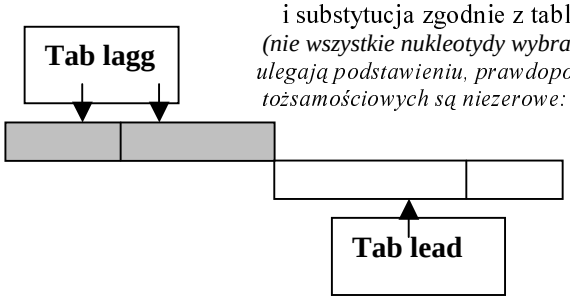
grupowanie genów według położenia na nici

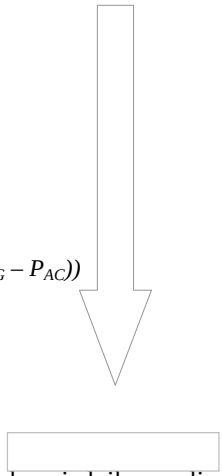


duplikacja sekwencji – tworzenie bliźniaczej puli genów



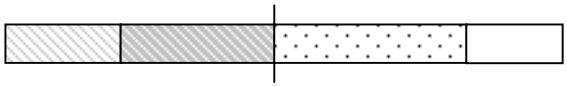
wybieranie nukleotydu do mutacji z prawdopodobieństwem pmut
i substytucja zgodnie z tablicą mutacyjną
(nie wszystkie nukleotydy wybrane do mutacji ulegają podstawieniu, prawdopodobieństwa przejść tożsamościowych są niezerowe: $P_{AA} = (1 - P_{AT} - P_{AG} - P_{AC})$)





obliczanie składu aminokwasowego sekwencji po mutacji
według wzoru $\Sigma|f_{a0} - f_{at}|$ i sprawdzanie warunku tolerancji

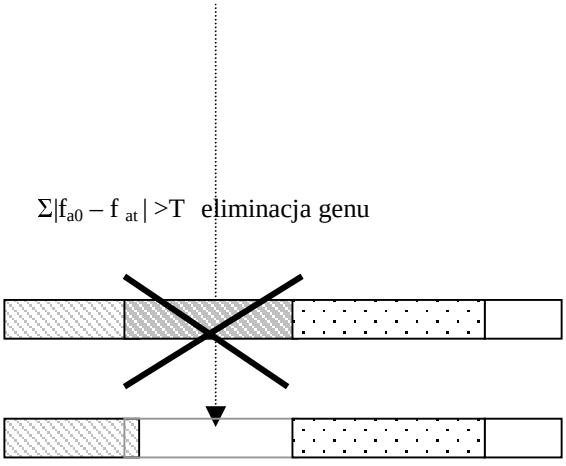
duplikacja odpowiednika wyeliminowanego genu



$\Sigma|f_{a0} - f_{at}| < T$

akceptacja wprowadzonych substytucji

$\Sigma|f_{a0} - f_{at}| > T$ eliminacja genu



podstawienie sekwencji jego ortologa z puli *B*

Ryc. 14. Schemat algorytmu symulacyjnego

3.4. Wyznaczanie parametru tolerancji

Parametrem określającym w symulacji siłę presji selekcyjnej jest przyjmowana wartość tolerancji, czyli akceptowany przez moduł selekcyjny zakres zmian w sekwencji. Arbitralne przyjęcie tolerancji na mutacje oddalałoby model od rzeczywistych zjawisk przyrodniczych. Najlepszym rozwiązaniem wydaje się wprowadzenie takiej wartości tolerancji T, która korespondowałaby z obserwowanym w przyrodzie poziomem zmienności składu aminokwasowego białka pełniącego daną funkcję. Aby oszacować tę wartość postanowiono obliczyć średnią różnicę składu aminokwasowego między ortologami pochodzącymi z dwóch spokrewnionych gatunków bakterii. W przypadku rozpatrywanego modelu wybrano ortologi z genomu *B. burgorferi i T. pallidum*, genomu najbliżej spokrewnionego z *B. burgdorferi* a jednocześnie kompletnie zsekwencjonowanego. W celu wyeliminowania z analizy ewentualnych paralogów i pseudogenów wybrano z bazy COG tylko ortologi obustronnie najbliższe, tak, aby żaden z genów nie posiadał bliższego ortologa niż jego odpowiednik z pary. Typowano więc tylko takie pary sekwencji, które odnajdywały siebie nawzajem po odwrotnym przeszukaniu bazy, a następnie sprawdzano, czy oba ortologi pełnią tę samą funkcję.

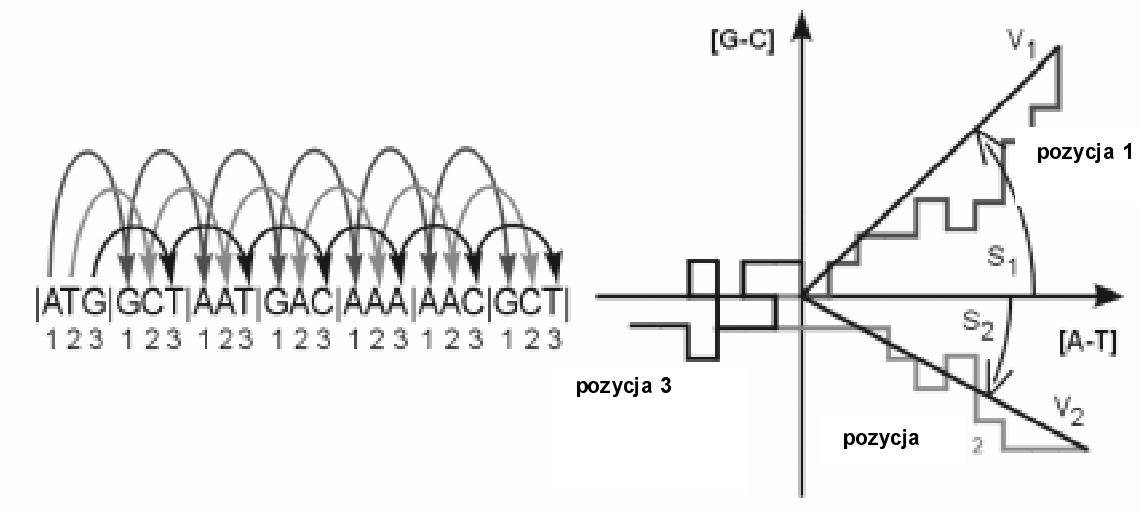
Dla każdej pary obliczono także parametr różnic składu aminokwasowego *(wzór 15.)*, którego średnią arytmetyczną zastosowano następnie do obliczeń. Dla 433 par ortologów obliczono średnią hydrofobowość produktów każdego z genów według wzoru:

$$\sum_{i=0}^{k} (H_{ai}) / k \qquad (16)$$

(gdzie H_a to hydrofobowość i-tego aminokwasu, a *k* - liczba wszystkich aminokwasów w danym białku) oraz średnią różnicę hydrofobowości dla wszystkich par ortologów, a także punkt izoelektryczny produktu każdego genu oraz średnią różnicę pI dla analizowanych par ortologów. Przy obliczaniu punktu izoelektrycznego dla 433 par ortologicznych białek skorzystano z dostępnej w sieci aplikacji *(http://us.expasy.org/tools/pi_tool.html)*. W oparciu o analizę korelacji badanych właściwości białek przyjęto średnią różnicę w składzie aminokwasowym białek *B. burgdorferi* i *T. pallidum* za parametr selekcyjny symulacji.

3.5. Metody obrazowania i analizy asymetrii

Jedną z metod graficznego przedstawiania asymetrii składu sekwencji nukleotydowych jest spacer w przestrzeni *A-T/G-C (Berthelsen i współpr. 1992)* Trendy występujące w składzie analizowanej sekwencji można obserwować w postaci wykresu w dwuwymiarowej przestrzeni *A-T, G-C*. Analiza polega na przesuwaniu się wirtualnego „spacerowicza" wzdłuż sekwencji nukleotydów o odpowiedni wektor *(Ryc.14)*. I tak, napotkanie adeniny powoduje przesunięcie spacerowicza do punktu wyznaczonego przez wektor o początku w punkcie, w którym zakończył się poprzedni krok spaceru i współrzędnych [1,0]. Napotkanie tyminy powoduje odpowiednio przesunięcie o wektor [-1,0], guaniny o wektor [0,1], a cytozyny - o wektor [0,-1]. Ten sposób obrazowania asymetrii pozwala na przedstawienie spaceru dla czterech różnych nukleotydów jednocześnie. Wykresy takie można robić zarówno dla całych chromosomów, jak i dla pojedynczych, czy sklejonych genów i przedstawiać oddzielnie spacer po trzech pozycjach w kodonie. „Spacerowicz" skacze wtedy tylko po pierwszych, drugich lub trzecich pozycjach w kodonie. Ten ostatni rodzaj spaceru typu Berthelsen nazywano, ze względu na charakterystyczny kształt, „pajączkiem". Pajączki wyraźne pokazują trendy związane z kodowaniem. Wykresy dla sekwencji niekodujących różnią się wyraźnie od spacerów wykonanych po sekwencjach kodujących, a nawet pseudogenach, które zachowują przez pewien czas trendy utrwalone w czasie, kiedy kodowały białko. Silnie wydłużone „nóżki" pajączka świadczą o właściwościach kodujących sekwencji. Krótkie i poplątane – o stochastycznym rozkładzie nukleotydów, czyli informacyjnym „szumie".



Ryc.14. Schemat konstrukcji pajączka dla dowolnej kodującej sekwencji nukleotydowej

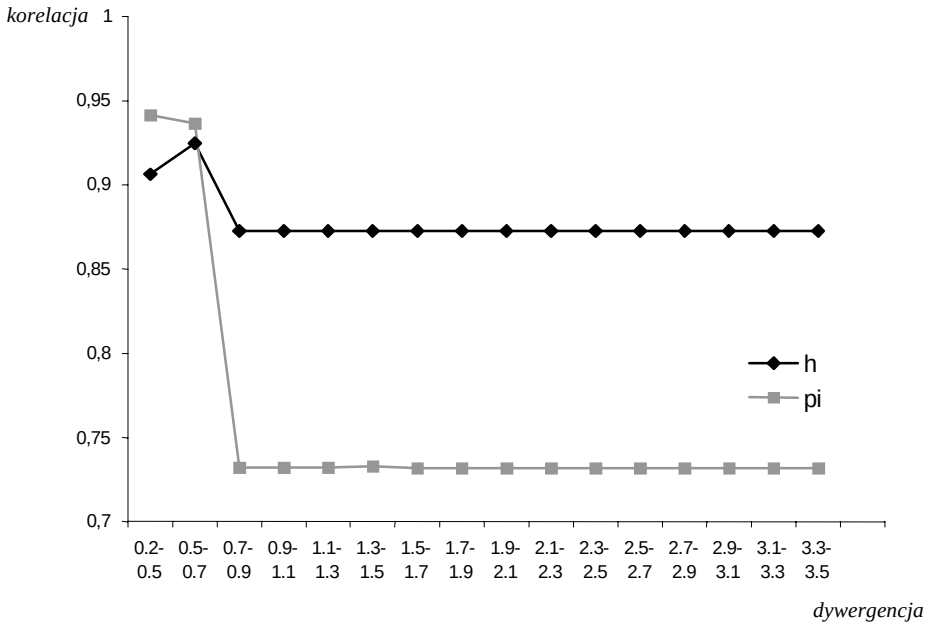
Pajączki dla wszystkich genów lub ORFów z danego genomu można przedstawić na jednym wykresie. Pojedynczy „pajączek” jest wtedy reprezentowany przez punkt, którego współrzędnymi są kąty nachylenia nóżek, czyli spacerów dla poszczególnych pozycji w kodonie, do osi A-T. Płaszczyzną takiego wykresu jest skończona projekcja powierzchni torusa.

4. WYNIKI I DYSKUSJA

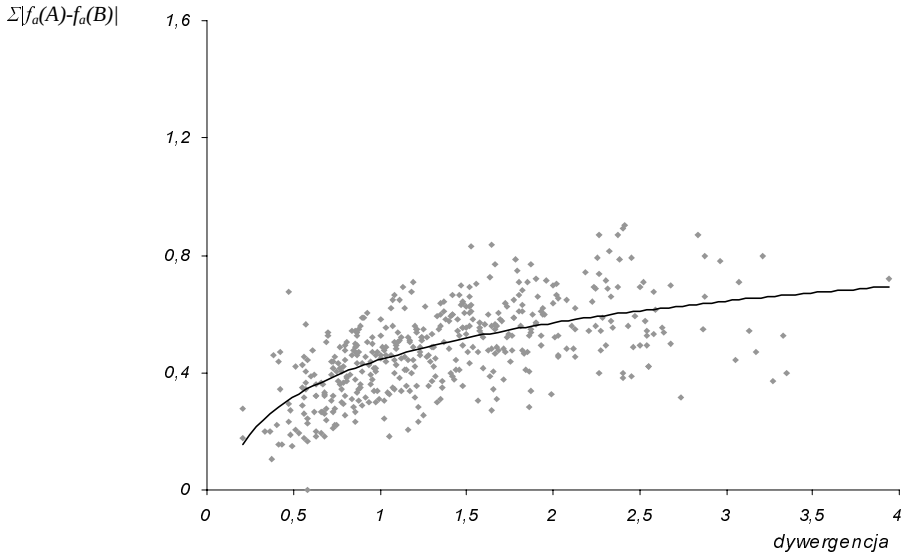
4.1. Szacowanie zakresu tolerancji

Aby zwiększyć wiarygodność homologii znalezionych między genami *B. burgdorferi* i *T. pallidum (rozdz.3.4, rozdz.3.5)* i powiązać je z właściwościami chemicznymi kodowanych białek, wykonano szereg dodatkowych analiz statystycznych, które pozwoliły na ostateczne wytypowanie zbioru ortologów, który służył za podstawę do oszacowania zakresu tolerancji. Obliczono współczynnik korelacji liniowej między wartościami punktów izoelektrycznych sekwencji ortologicznych z badanych genomów oraz współczynnik korelacji między wartościami ich hydrofobowości. Na osi x odłożono wartość hydrofobowości (lub pI) białka kodowanego przez sekwencję *B. burgdorferi*, a na osi y hydrofobowość (pI) homologicznego białka *T. pallidum*. Zbadano także zależność różnic poszczególnych parametrów od poziomu dywergencji między ortologami, oraz zależność współczynnika korelacji między wartościami hydrofobowości i pI od dywergencji między sekwencjami.

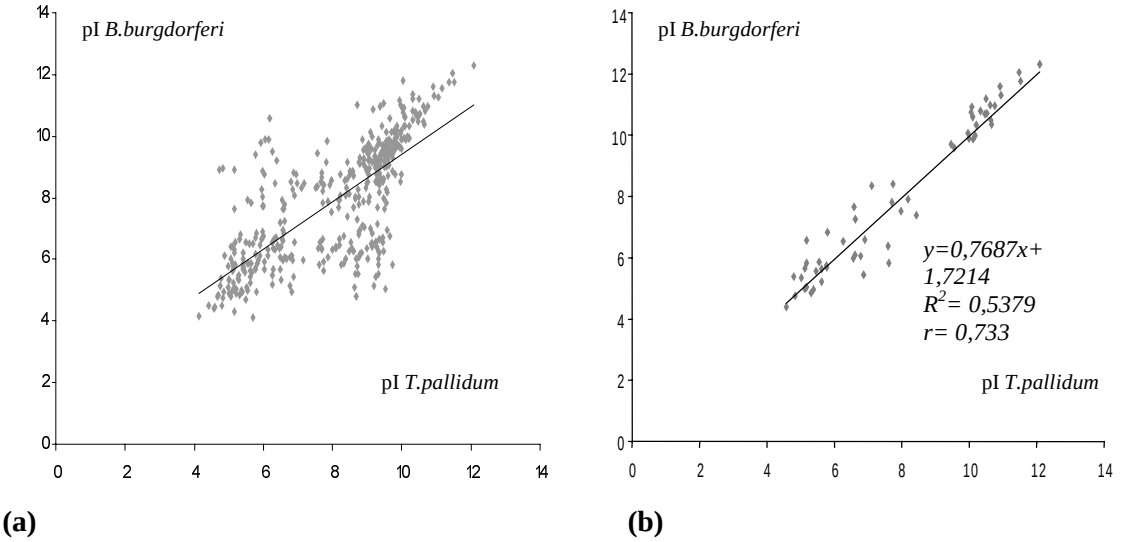
Okazało się, że korelacje między hydrofobowościami i pI par ortologów obniżają się znacząco, jeżeli dywergencja między badanymi parami ortologów przekroczy poziom 0,7 *(Ryc.15),* co ciekawsze po przekroczeniu tego poziomu dywergencji wartości globalnych różnic aminokwasowych między ortologami przestają rosnąć, osiągając w miarę stałe plateau *(Ryc.16.).* Jeżeli ograniczymy zbiór analizowanych danych do zbioru ortologów, których zakres dywergencji jest nie większy niż 0,7, współczynnik korelacji dla hydrofobowości wynosi ok. 0,9 *(Ryc.17.a,b*), a dla pI ok. 0,7 *(Ryc.18.a,b),* co świadczy o wysokim podobieństwie właściwości białek kodowanych przez te pary ortologów. Średnia różnica globalnego składu aminokwasowego wyliczona dla takiego podzbioru ortologów wynosi 0,33. Taką wartość parametru tolerancji na zmiany składu aminokwasowego zastosowano w dalszych symulacjach. Sekwencje użyte ostatecznie do wyznaczenia siły selekcji przeszły więc przez trójetapowy proces eliminacji – pierwszy i drugi etap obejmował sprawdzenie wiarygodności znalezionych homologii (wybranie ortologów obustronnie najbliższych oraz odrzucenie tych par, które nie kodowały białka o tej samej funkcji), trzeci eliminował pary o zbyt dużej dywergencji, dla których współczynnik korelacji między hydrofobowościami kodowanych białek wynosił poniżej 0,8, a współczynnik korelacji między ich punktami izoelektrycznymi nie przekraczał 0,7.



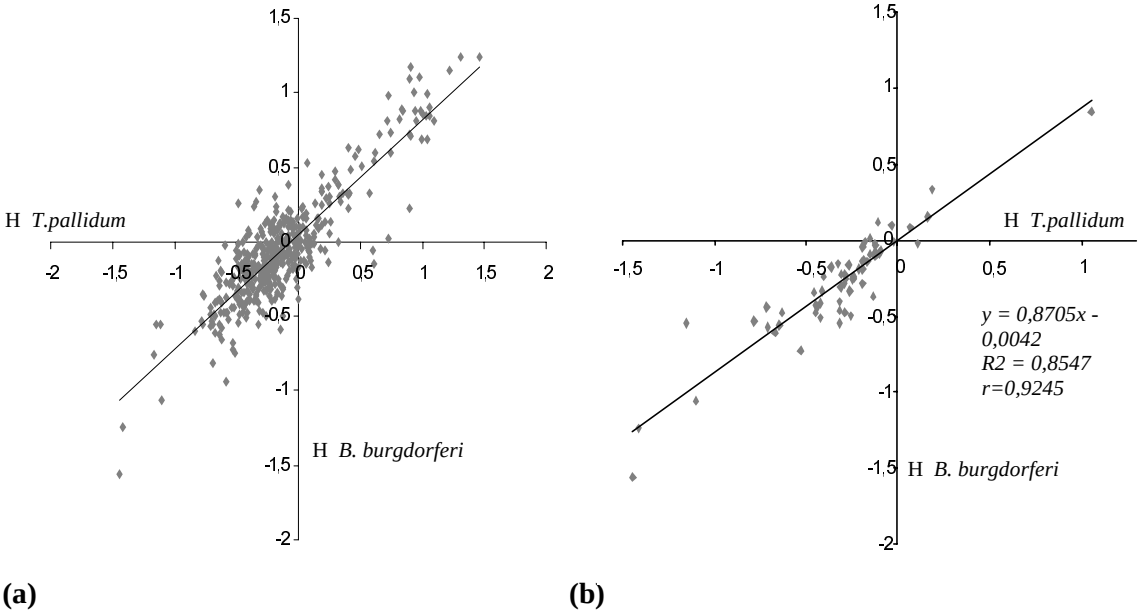
Ryc.15. Zależność korelacji między wartościami hydrofobowości (czarne punkty) i pI (szare punkty) sekwencji ortologów z poszczególnych par od poziomu dywergencji aminokwasowej (mierzonej ilością substytucji aminokwasowych na miejsce) między nimi. Pary ortologów podzielono na klasy według poziomu dywergencji (szerokość przedziału wynosi 0,2).



Ryc.16. Zależność różnic składu aminokwasowego między ortologami z pary, liczonych według wzoru *(15)*, od poziomu dywergencji między nimi. Na osi rzędnych odłożono wartości poziomu dywergencji między sekwencjami, na osi odciętych – wartości parametru $\Sigma|f_a(A)-f_a(B)|$



Ryc.17. Korelacje między wartościami pI (a) obliczanymi dla sekwencji ortologów wszystkich analizowanych par i (b) dla zbioru par sekwencji, których poziom dywergencji nie przekroczył 0,7. Na wykresie (b) zamieszczono wzór prostej regresji i wartość współczynnika korelacji liniowej dla przedstawionego zbioru punktów.

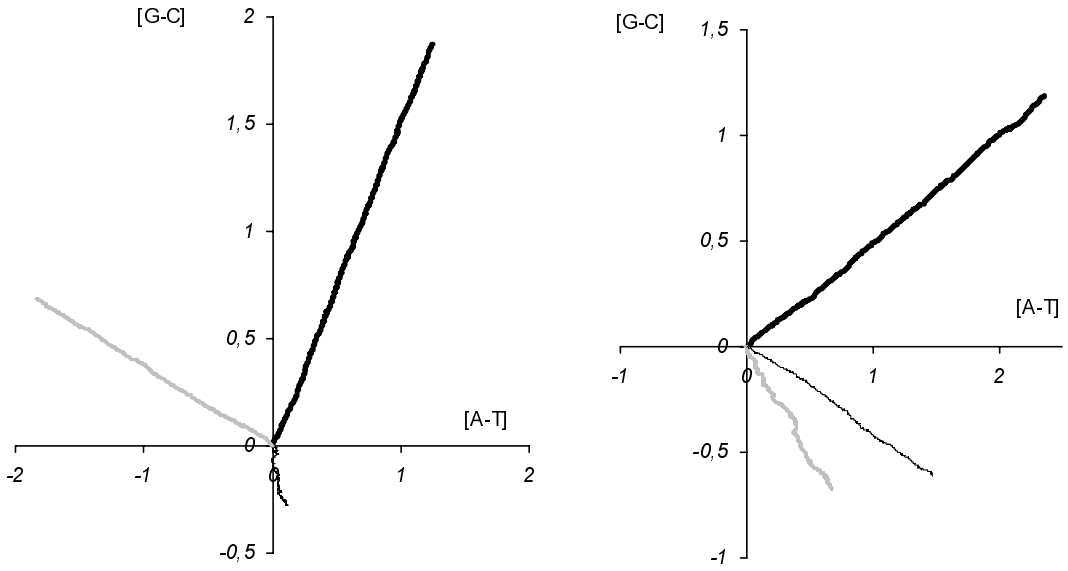


Ryc.18. Wykres przedstawiający korelacje między wartościami hydrofobowości w analizowanych pulach ortologów. Wykres (a) ukazuje korelację dla wszystkich par ortologów, wykres (b) dla grupy sekwencji, których dywergencja nie przekracza 0,7 substytucji aminokwasowej na miejsce. Na wykresie (b) zamieszczono wzór prostej regresji i wartość współczynnika korelacji dla przedstawionego zbioru punktów

4.2. Analiza presji selekcyjnej

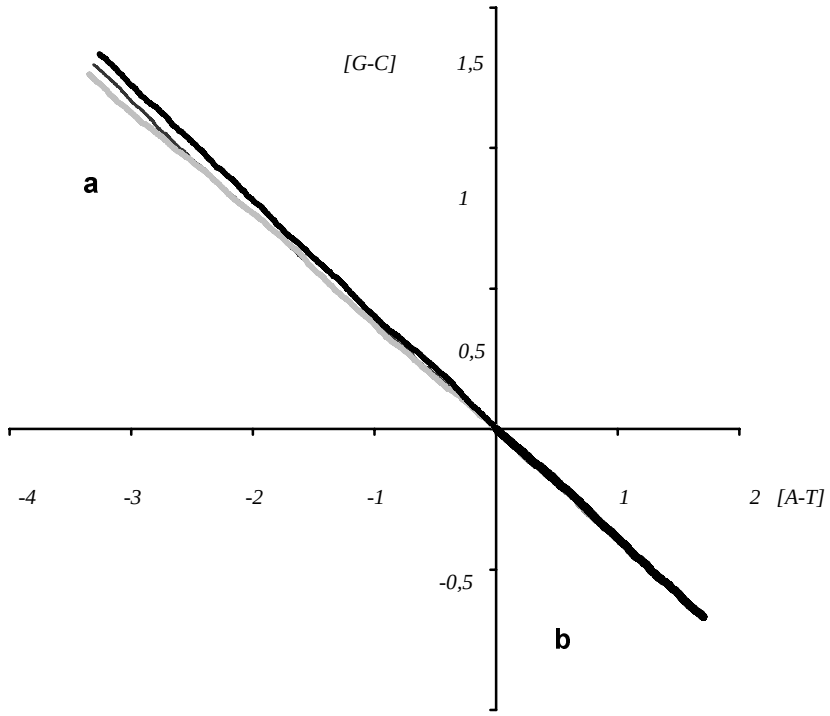
Rola presji selekcyjnej w procesie ewolucji sprowadza się do zachowania ilości i jakości niesionej informacji na poziomie, który pozwala na przeżycie i rozmnożenie się organizmu, czyli powielenie jej samej. Selekcja działa jednak najczęściej na poziomie funkcji białka, nie ciągu nukleotydów, co tworzy pewien zakres dopuszczalnej zmienności w kompozycji genów i pozwala na akceptację pewnej puli mutacji.

Aby zrozumieć znaczenie selekcji należy wyobrazić sobie jak wyglądałaby sekwencja poddana działaniu tylko i wyłącznie presji mutacyjnej. Na *Rycinie 20* przedstawiono spacer po sekwencjach kodujących *B. burgdorferi* doprowadzonych do równowagi z presją mutacyjną (po symulacji pod czystą presją mutacyjną przez 50000 MCS). Trendy związane z kodowaniem, widoczne wyraźnie na wykresie wykonanym metodą spaceru Berthelsen *(Rozdz.3.4.)* po sekwencji oryginalnej genów *B. burgdorferi (Ryc.19. a, b),* zanikają, pozostają tylko trendy związane z kierunkową presją mutacyjną. Informacja zawarta w sekwencji kodującej ulega zatarciu, a sama sekwencja osiąga stan równowagi, który jednak nie jest stanem całkowitej symetrii składu, jak należałoby się spodziewać, gdyby obowiązywała stochastyczna zasada PRII *(Rozdz.1.1.)*. Przeciwnie, stan równowagi także wykazuje pewne trendy, które wskazują na nielosowy charakter siły zwiększającej entropię systemu, jakim jest funkcjonująca w żywej komórce cząsteczka DNA. Spacer typu Berthelsen po poszczególnych pozycjach w kodonach sekwencji kodujących daje syntetyczny obraz dwóch rodzajów trendów – związanych z mutacją i wprowadzanych przez asymetryczną presję mutacyjną oraz tych będących bezpośrednim skutkiem działania sił selekcji. Spacer typu Berthelsen polega na „przejściu" przez wszystkie pozycje nukleotydowe analizowanej sekwencji z zachowaniem śladu takiego przejścia w dwuwymiarowej przestrzeni A-T/G-C. Ślad taki jest obrazem lokalnych trendów w składzie sekwencji, rodzaj napotkanego nukleotydu wymusza bowiem kierunek ruchu na płaszczyźnie.



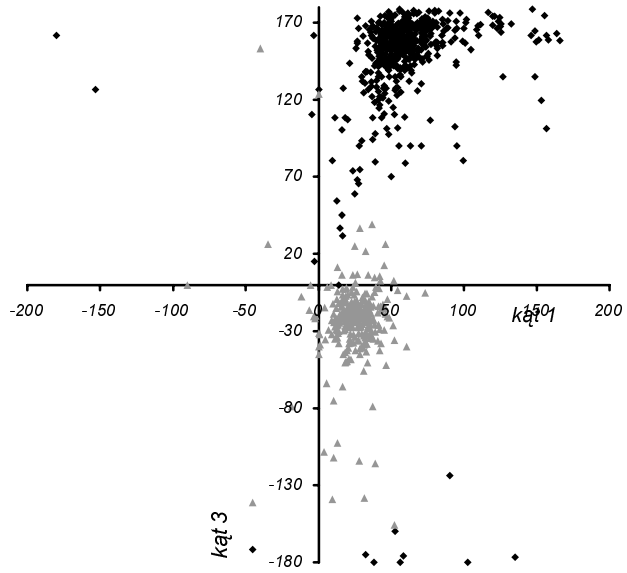
(a) **(b)**

Ryc. 19. Spacer typu Berthelsen po trzech pozycjach sekwencji kodujących nici wiodącej *(a)* i opóźniającej *(b) Borrelia burgdorferi.* Szara linia – trend dla trzecich pozycji w kodonie, czarna pogrubiona – pierwsze pozycje, czarna cienka linia – trend dla drugich pozycji.



Ryc. 20. Ten sam spacer wykonany dla sekwencji poddanych symulacji pod czystą presją mutacyjną przez 50000 kroków Monte Carlo. (a) wykres dla trzech pozycji w kodonie dla sekwencji z nici wiodącej, (b) - z nici opóźniającej (wykresy dla trzech pozycji w kodonie niemal zupełnie się pokrywają)

Brak trendów przy tym sposobie analizy obserwujemy w postaci kłębka statystycznego, powstającego wokół początku układu współrzędnych. Genom *B. burgdorferi* cechuje bardzo wyraźna asymetria składu nukleotydowego, przy czym trendy dla trzecich pozycji kodonów sekwencji kodujących z nici wiodącej i opóźniającej są do siebie wyraźnie przeciwstawne. Trzecie pozycje kodonów nici wiodącej są bogate w tyminę i guaninę, natomiast w kodonach leżących na nici opóźniającej przeważa adenina i cytozyna. Ta doskonała niemal komplementarność trendów zgadza się ze spotykanymi w literaturze *(Wolfe 1989)* wzmiankami o niewystępowaniu u *B. burgdorferi* presji selekcyjnej związanej z ewentualnym wpływem trzeciej pozycji kodonu na szybkość translacji. Może to być związane z ograniczoną wielkością genomu, a nawet dążeniem do jego minimalizacji. Jednym ze sposobów na zmniejszenie genomu może być ograniczenie ilości różnych rodzajów tRNA rozpoznających ten sam aminokwas, co jest możliwe dzięki regule wahadła oraz modyfikacjom składu tRNA (np. wprowadzaniu do antykodonów inozyny lub pseudourydyny, które zwiększają dowolność parowania). Przy minimalnej ilości różnych rodzajów tRNA, wpływ częstości występowania tRNA dla danego kodonu na tempo procesu translacji staje się znikomy, a tym samym znaczenie selekcyjne składu trzeciej pozycji kodonu jest do pominięcia. Można więc wnioskować, że skład trzecich pozycji kodonu u *B. burgdorferi* odzwierciedla jedynie wpływ presji mutacyjnej, a nie trendy związane z kodowaniem. Jednocześnie presja działająca na nić opóźniającą jest dokładnie komplementarna do presji określonej dla nici wiodącej, co usprawiedliwia wprowadzenie tablicy lustrzanej, jako macierzy opisującej presję mutacyjną na nici opóźniającej w symulacjach. Trendy dla trzecich pozycji w kodonie nie we wszystkich genomach są przeciwstawne, nie zawsze znaczenie selekcyjne trzecich pozycji jest do pominięcia, nie dla wszystkich genomów otrzymano też tak wyraźny obraz rozdziału genów na torusie, jak dla genomu *B. burgdorferi*. Jeżeli do utworzenia wykresu użyjemy kątów nachylenia spacerów po pierwszych i trzecich pozycjach sekwencji kodujących *B. burgdorferi,* geny z nici wiodącej i opóźniającej podzielą się na dwie odrębne grupy *(Ryc. 21).*



Ryc.21. Rozkład punktów reprezentujących wartości kątów nachylenia pierwszej i trzeciej nóżki „pajączka" do osi A-T dla poszczególnych genów *B. burgdorferi.* Szare trójkąty odpowiadają genom z nici wiodącej, czarne kwadraty- genom z nici opóźniającej

Aby wydobyć ze spaceru Berthelsen informację o czystej presji selekcyjnej, należy go zmodyfikować, uwzględniając wagę informacyjną każdego z nukleotydów. W tym celu należy nadać każdemu krokowi wartość odwrotnie proporcjonalną do częstości występowania danego nukleotydu w sekwencji doprowadzonej do stanu równowagi z presją mutacyjną (np. po napotkaniu adeniny przesuwamy się o wektor: $[1/N_{AR}, 0]$, gdzie N_{AR} oznacza frakcję adeniny w sekwencji równowagowej; po napotkaniu tyminy o wektor o współrzędnych $[-1/N_{TR},0]$, itd.). Im rzadziej dany nukleotyd występuje w sekwencji równowagowej, tym większa jest jego waga informacyjna, jeżeli pojawi się w sekwencji kodującej i odwrotnie, im częściej występuje losowo, tym statystycznie mniej niesie informacji.

Otrzymany wykres *(Ryc. 22.)* jest charakterystyczny dla genomu *B. burgdorferi,* potwierdza założenie o braku selekcji na trzecie pozycje w kodonie (nóżki dla trzeciej pozycji wykresu zarówno dla nici wiodącej, jak i opóźniającej, zwijają się wokół początku układu współrzędnych, nie wykazując żadnych trendów selekcyjnych) i silną selekcję na pierwsze i drugie pozycje w kodonie (silne trendy dla nici wiodącej *(Ryc.22.a)* i nieco słabiej zaznaczone dla nici opóźniającej *(Ryc.22.b).*

Innym sposobem uwidaczniania presji selekcyjnej jest konstrukcja tablic PAM *(Rozdz. 1.4)* dla sekwencji poddanej czystej presji mutacyjnej, a następnie porównanie tak otrzymanej tablicy z macierzami Dayhoff. Ewolucję sekwencji należy przerwać w momencie określonym przez oczekiwany poziom dywergencji (jeżeli porównujemy tablicę z macierzą PAM1

Dayhoff, symulujemy ewolucję do chwili, w której sekwencja macierzysta i potomna różnią się średnio jednym aminokwasem na sto). Następnie analizujemy otrzymaną sekwencję i tworzymy macierz zaobserwowanych podstawień aminokwasowych według wzoru *(13)*. W następnym kroku tworzymy pochodną macierz będącą wynikiem odjęcia uzyskanej tablicy symulacyjnej od macierzy PAM. Otrzymana tablica powinna opisywać presję selekcyjną. Konstruowane na podstawie porównań współczesnych sekwencji ortologicznych macierze przejść aminokwasowych zawierają informację zarówno o presji mutacyjnej jak i selekcyjnej działającej na sekwencje podczas ich ewolucji. Są one wynikiem uśrednienia danych zebranych z porównań wielu gatunków. Dlatego macierz selekcyjna otrzymana po odjęciu presji mutacyjnej wyznaczonej dla *B. burgdorferi* od ogólnej macierzy PAM nie została przyjęta za podstawę modułu selekcyjnego konstruowanego modelu.

Śledząc efekty działania selekcji wykonano szereg diagramów opisujących poszczególne właściwości białek kodowanych przez genom *B. burgdorferi*.

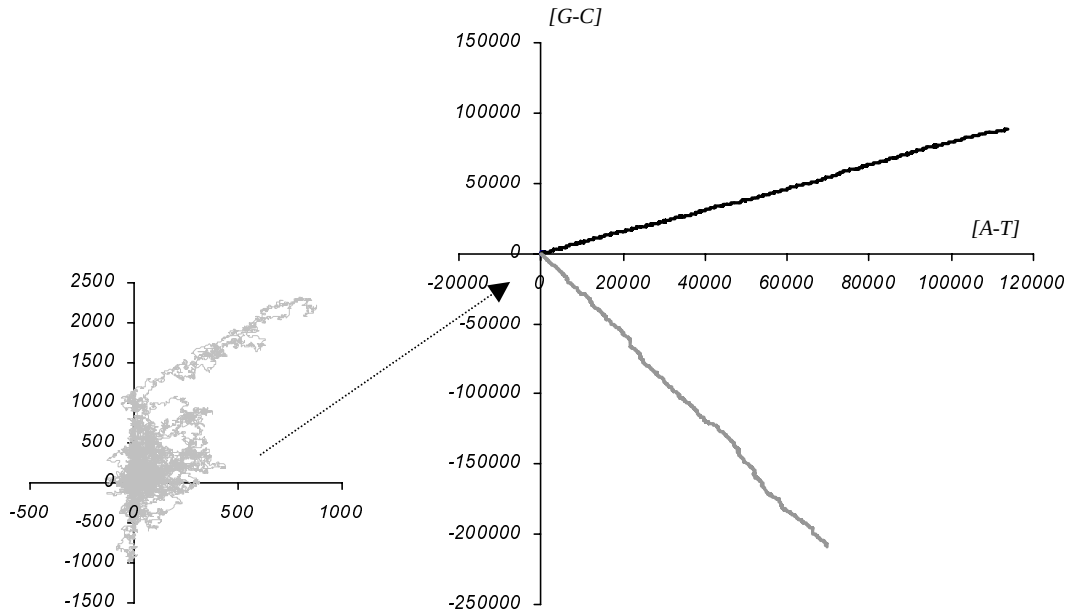
Pierwszy z prezentowanych histogramów *(Ryc.23.)* przedstawia rozkład hydrofobowości białek kodowanych przez oryginalną sekwencję *B. burgdorferi* (na osi x poszczególne klasy hydrofobowości, na osi y liczebność klas) i odpowiednich sekwencji w stanie równowagi z presją mutacyjną, osobno dla nici wiodącej i opóźniającej. Diagramy dla sekwencji w stanie równowagi są wyraźnie przesunięte w stosunku do diagramów wykonanych dla sekwencji oryginalnych. W przypadku genów z nici wiodącej dojście do stanu równowagi z presją mutacyjną doprowadziło do zwiększenia średniej hydrofobowości kodowanych białek, histogram jest wyraźnie przesunięty w prawo. Dla genów z nici opóźniającej przesunięcie jest mniej wyraźne, ale także zauważalne, średnia hydrofobowość kodowanych białek po osiągnięciu stanu równowagi uległa jednak obniżeniu. Sugeruje to, po pierwsze, różnice we właściwościach i podatności na mutacje tych dwóch grup genów, a po drugie silniejsze działanie presji selekcyjnej na geny z nici wiodącej. Przesunięcia histogramów charakteryzują presję selekcyjną działającą na geny z nici opóźniającej i wiodącej.

Podobne wykresy wykonano analizując zmiany punktu izoelektrycznego tych samych grup genów *(Ryc.24.)*. Również tutaj jako odwołania użyto rozkładu wartości pI produktów oryginalnych genów *B. burgdorferi* z puli nici wiodącej i opóźniającej. Histogram dla sekwencji wyjściowej ma dwa wyraźne maksima (w zakresie pH 4,5-6,5 oraz 8,5-10,0), których położenie ulega zmianie pod działaniem czystej presji mutacyjnej. Pozostaje jedno, wyraźne maksimum dla białek o zasadowym charakterze (o pI w zakresie pH od 8,5 do 9,5 dla nici wiodącej i 10,0-10,5 dla opóźniającej) i drugi, mniej wyraźny szczyt dla białek o charakterze kwaśnym, widoczny tylko dla genów nici wiodącej. Co ciekawe na wszystkich

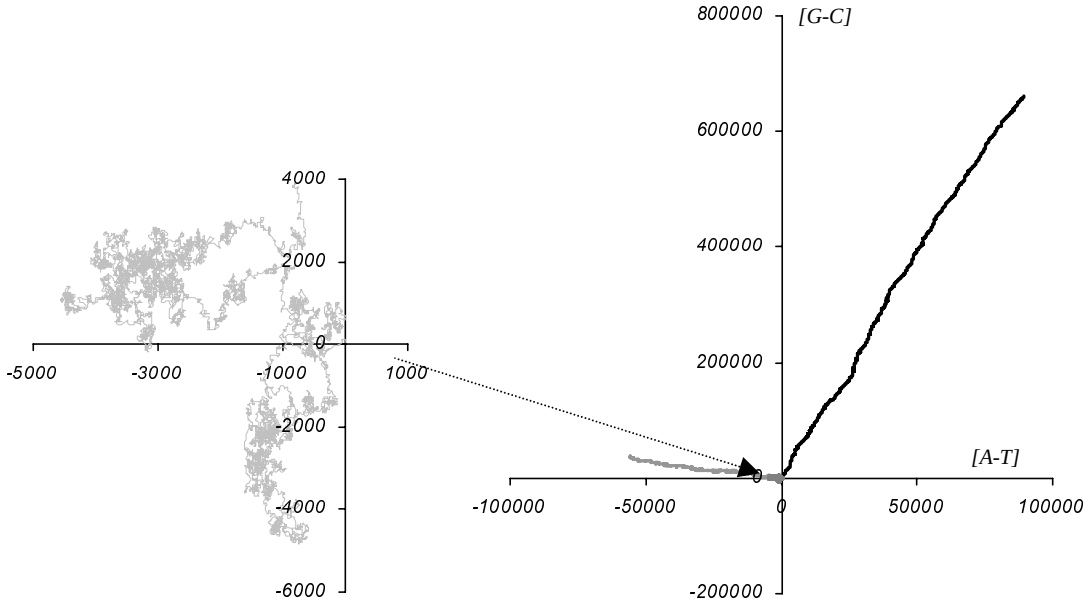
diagramach dla zakresu pH 7,0- 7,5 zaznacza się głębokie minimum lub obszar nieobsadzony, świadczące być może o jakimś stanie zakazanym, nie realizowanym przez białka najprawdopodobniej nie tylko ze względów selekcyjnych (w pH wnętrza komórki, które wynosi około 7,3, ich rozpuszczalność białek obojętnych byłaby silnie ograniczona), ale też z powodu użycia specyficznej grupy aminokwasów do ich budowy, których wybór narzuca konstrukcja kodu genetycznego. Za tą ostatnią tezą przemawia fakt, że diagramy dla sekwencji równowagowych, uzyskanych drogą symulacyjną, zawierają wspomniane minima.

Wymienione obserwacje zmuszają do zastanowienia się nad historią powstania i stabilizacji współczesnego kodu genetycznego, mogą być też podstawą i punktem wyjścia do dalszych analiz.

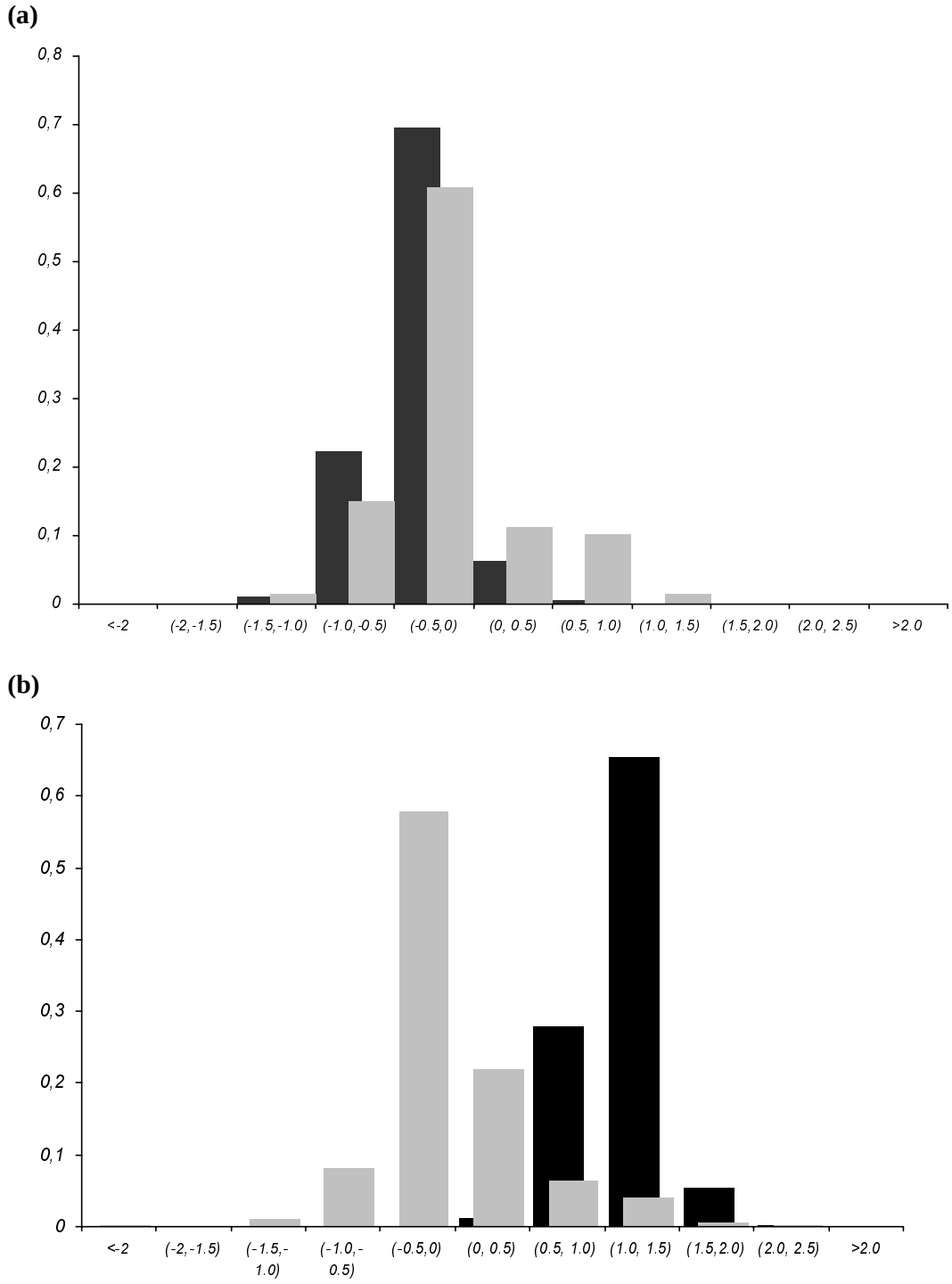
(a)



(b)

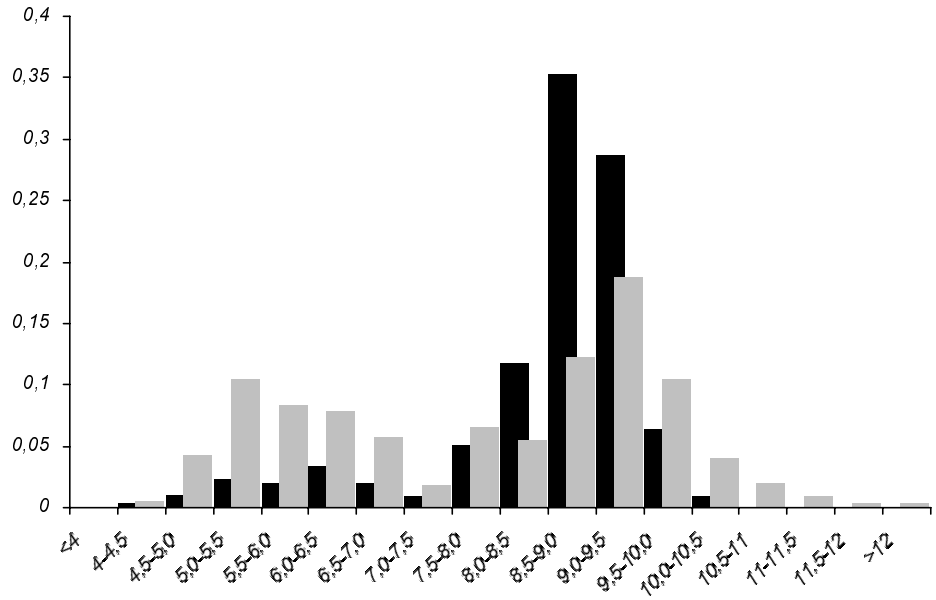


Ryc.22. Zmodyfikowane spacery typu Berthelsen po trzech pozycjach sekwencji kodujących *B. burgdorferi* z nici wiodącej (a) i opóźniającej (b). Czarne ciągłe linie przedstawiają trendy dla pierwszych pozycji w kodonie, szare linie- spacery po drugich pozycjach. Obok, w większej skali, przedstawiono spacery po trzecich pozycjach sekwencji kodujących.(kłębki statystyczne) Osie wyznaczają kierunki w dwuwymiarowej przestrzeni A-T/G-C.

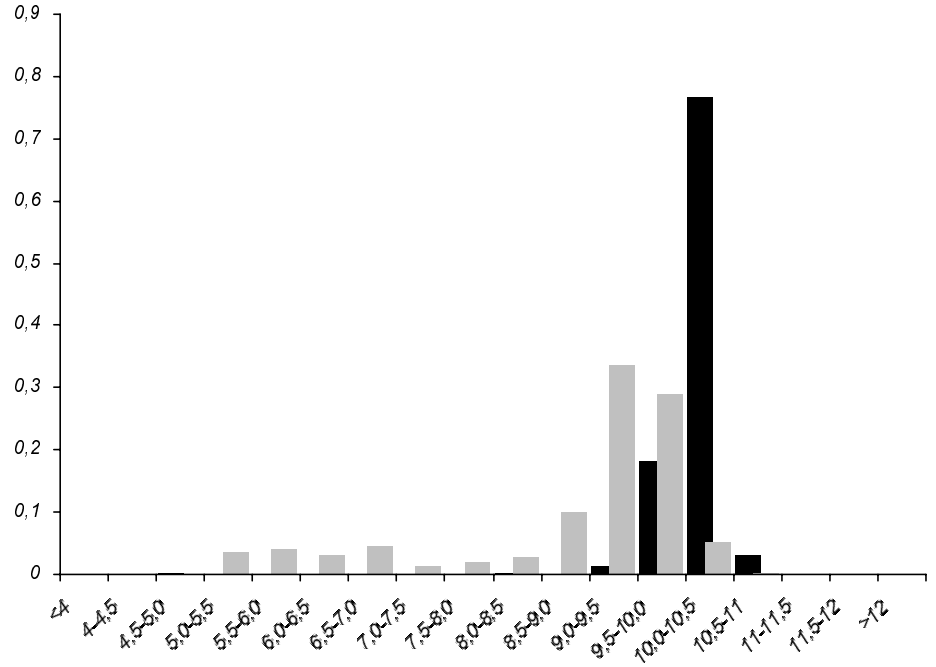


Ryc. 23. Histogramy rozkładu hydrofobowości dla sekwencji kodujących z nici wiodącej (a) i opóźniającej (b). Słupki szare reprezentują oryginalne grupy genów *B. burgdorferi*, czarne – odpowiednie sekwencje doprowadzone do stanu równowagi z presją mutacyjną. Na osi y odłożono znormalizowaną przez ilość genów na danej nici liczebność poszczególnych klas hydrofobowości.

(a)



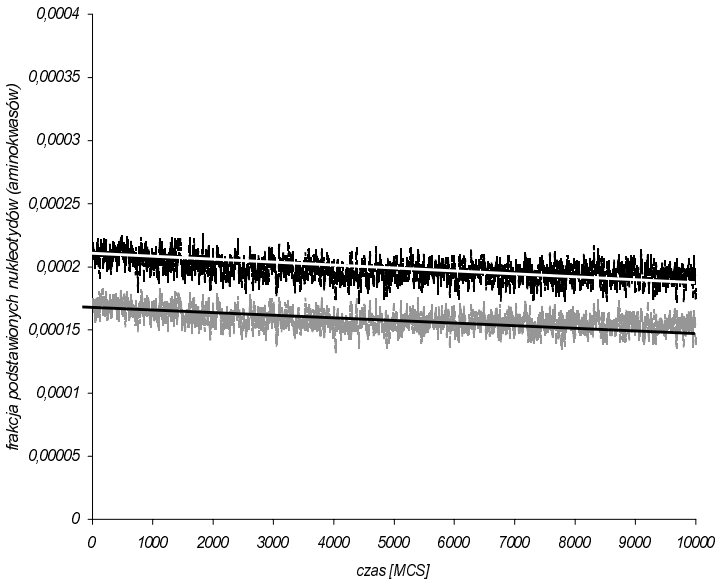
(b)



Ryc.24. Histogramy rozkładu pI dla sekwencji kodujących (a) z nici wiodącej i (b). opóźniającej Słupki szare reprezentują oryginalne grupy genów *B. burgdorferi*, czarne – odpowiednie sekwencje doprowadzone do stanu równowagi z presją mutacyjną. Na osi y odłożono znormalizowaną przez ilość genów na danej nici liczebność poszczególnych klas wartości pH przyjmowanych przez punkt izoelektryczny.

4.3. Miejsce kodu genetycznego w układzie presja mutacyjna – selekcja - genom

Na tle ostatnich badań kod genetyczny jawi się jako skomplikowany i wysoce uporządkowany system. Część naukowców usiłuje udowodnić, że jest on jednym z najdoskonalszych produktów ewolucji *(rozdz.1.7.),* trudno jednak rozstrzygać o tym w sytuacji, gdy mamy do czynienia z pojedynczym obiektem badań. Stan obecny, który poddaje się analizie naukowej, wygląda następująco: mamy jeden, zasadniczo uniwersalny wzorzec



Ryc.25. Liczba substytucji aminokwasowych (szara linia) i nukleotydowych (czarna linia) wpadających do *sekwencji B. burgdorferi* w czasie symulacji z czystą presją mutacyjną (bez selekcji). Na osi x czas symulacji w MCS.

kodu genetycznego i całe spektrum sekwencji nukleotydowych, dla których był on podstawą ewolucji i jednym z najistotniejszych jej ograniczeń. Być może słuszniejsze, z punktu widzenia metodologii naukowej, byłoby odwrócenie pytania o optymalizację kodu i zastanowienie się nad przystosowaniem do istniejącego wariantu kodu milionów systemów, na które składają się sekwencje nukleotydowe genomów oraz charakterystyczna dla nich presja mutacyjna i selekcyjna. Stawiając pytanie w ten sposób, dysponujemy obszernym materiałem porównawczym. Konstrukcja kodu dozwala na pewne przejścia aminokwasowe (jeżeli zachodzą na drodze pojedynczej substytucji a zmiana sensu kodonu nie wywołuje znaczącej zmiany właściwości kodowanego aminokwasu lub nie zmienia go wcale), do innych praktycznie nie dopuszcza (nawet jeżeli przejście aminokwasu *a* w *b* jest nieszkodliwe pod względem różnicy we właściwościach białka, może być niemożliwe do zrealizowania, ponieważ wymaga przejścia przez stadium *c*, zupełnie nie do zaakceptowania dla układu).
O takich ukrytych stanach zakazanych zapomina się często konstruując tablice przejść aminokwasowych, które uwzględniają wszystkie rodzaje podstawień.

Próbując zdefiniować powiązania między presją mutacyjną, selekcyjną i kodem genetycznym a sekwencją genomu, należy wyjść od prostej zależności wiążącej presję mutacyjną i frakcje poszczególnych nukleotydów w sekwencji doprowadzonej do stanu równowagi z presją mutacyjną *(wzór 7).* Oczywiście stan równowagi opisany równaniem *(7)* jest stanem, do którego dąży sekwencja pozostająca pod wpływem presji mutacyjnej, ale którego nigdy nie osiąga, ponieważ przeciwdziała temu presja selekcyjna. Rycina 25 przedstawia efekt *„płynięcia"* składu nukleotydowego sekwencji kodującej *B. burgdorferi* w kierunku stanu równowagi, a jednocześnie efekt degeneracji kodu genetycznego. Ilość podstawień aminokwasowych, do których doszło w sekwencji poddawanej czystej presji mutacyjnej przez 10 000 MCS jest wyraźnie niższa od liczby podstawień kodonowych. Każda zamiana nukleotydu prowadzi do zmiany kodonu, nie wszystkie substytucje w kodonie prowadzą jednak do zmiany kodowanego aminokwasu. Skład nukleotydowy sekwencji może więc dopasować się częściowo do presji mutacyjnej, ponieważ część podmian nie będzie miała znaczenia selekcyjnego, a jednocześnie zmniejszyć swoją mutabilność ponieważ w trakcie ewolucji kodony o krótszym okresie półtrwania będą stopniowo zastępowane przez kodony mniej mutabilne. Oczywiście zmiany takie zajdą tylko wtedy, jeżeli zaakceptuje je selekcja. O tym, jakie rodzaje podmian są akceptowalne decyduje sposób degeneracji kodu genetycznego.

Przyjmując hipotezę wczesnej stabilizacji kodu genetycznego, można przypuszczać, że to właśnie losowo utrwalona struktura kodu genetycznego odegrała decydującą rolę w koewolucji układu presja mutacyjna – selekcyjna – genom. Prawdopodobnie to specyfika degeneracji kodu genetycznego narzuciła ustalenie takich, a nie innych parametrów presji mutacyjnej, które współgrając z presją selekcyjną minimalizują szkodliwy wpływ substytucji zachodzących we współczesnych genomach *(Dudkiewicz i współpr. 2004a).* Za tezą o przystosowaniu presji mutacyjnej i selekcyjnej do istniejącego kodu świadczą dwa fakty: po pierwsze częstość występowania aminokwasów w białkach jest skorelowana z liczbą kodonów, jaką dany aminokwas dysponuje w tablicy kodu *(King i Jukes 1969),* po drugie najbardziej mutabilne kodony (np. TGG) odpowiadają aminokwasom najrzadziej występującym w białkach, a jednocześnie najsilniej pilnowanym przez selekcję. Aminokwasy szybko eliminowane przez mutację („uciekające" z sekwencji, na którą działa asymetryczna tablica substytucji *Tab.4.)*, takie jak tryptofan lub cysteina, są jednocześnie aminokwasami mającymi duże znaczenie dla struktury białka. Widać to wyraźnie na przykładzie tryptofanu, którego prawdopodobieństwo przeżycia pod presją mutacyjną jest bardzo niskie, jest go też stosunkowo niewiele w sekwencji, ale jeśli już występuje, jest utrzymywany przez siły

selekcji. Z kolei aminokwasy występujące często i kodowane przez 4 lub 6 kodonów (np. leucyna) mają niską mutabilność, jest ich dużo w sekwencji, ale nie podlegają tak wyraźnej presji selekcyjnej. Wyniki pozwalające na wyciągnięcie takich wniosków uzyskano drogą symulacji komputerowych *(Nowicka i współpr. 2003).*

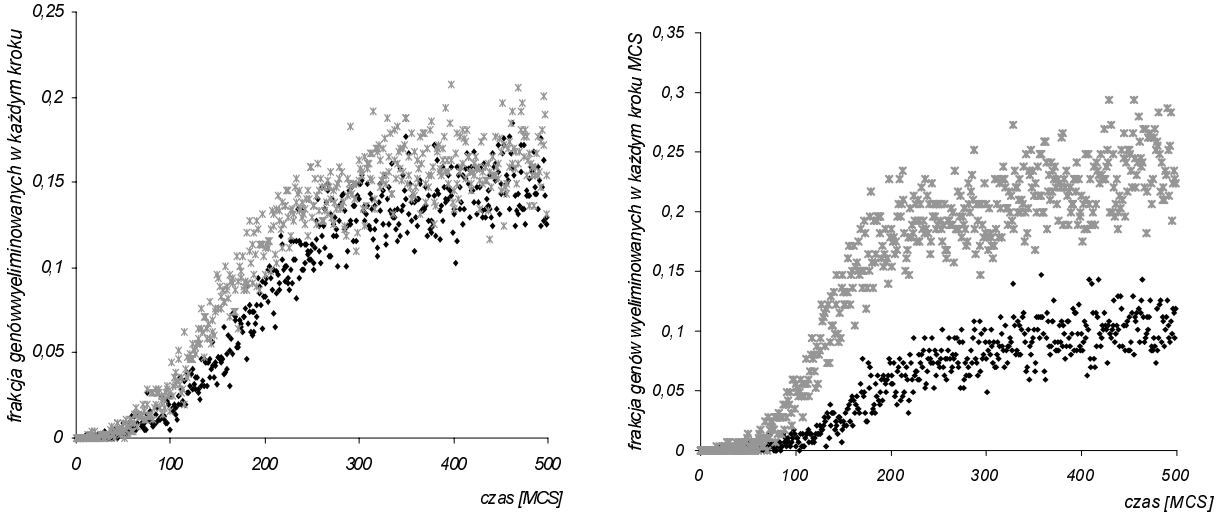
Opisywany model ewolucji genomu jest na tyle otwarty, że stwarza możliwości badania związków między presją mutacyjną, selekcyjną a samą sekwencją nukleotydową podlegającą tym procesom. Manipulując parametrami wejściowymi możemy próbować uzyskać odpowiedzi na wiele nierozstrzygniętych jeszcze pytań o sprzężenia między poszczególnymi czynnikami i produktami ewolucji. Taką próbę wyjaśnienia ukrytych zasad funkcjonowania podjęto dla systemu: genom *B. burgdorferi* / presja mutacyjna *B. burgdorferi* / uniwersalny kod genetyczny *(Dudkiewicz i współpr. 2004a)*. Celem doświadczenia było zbadanie wpływu naruszenia poszczególnych elementów tego układu na przeżywalność genów.

Założenia doświadczenia były bardzo proste. Za punkt odniesienia dla testowanych wariantów systemu uznano dynamikę eliminacji genów podczas symulacji trwającej 500 MCS (Monte Carlo Steps), przeprowadzonej na oryginalnych sekwencjach genów *B. burgdorferi* z nici wiodącej i opóźniającej pod presją mutacyjną określoną przez tablicę przejść stworzoną na podstawie sekwencji międzygenowych *(Tab.4.a.b) (Kowalczuk i współpr. 2001b)*. Parametrem selekcyjnym, ustalonym podczas przeprowadzania wszystkich symulacji będących podstawą doświadczenia, było zachowanie składu aminokwasowego genów z tolerancją 0,3. Opisane powyżej warunki symulacji będą nazywane w dalszej części rozważań standardowymi. Następną symulację przeprowadzono poddając oryginalną sekwencję genów *B. burgdorferi* presji tablicy symetrycznej, czyli takiej, w której wszystkie przejścia nukleotydowe zachodzą z jednakową częstością. Kolejna symulacja polegała na poddaniu losowo wygenerowanych sekwencji o długościach odpowiadających długościom genów *B. burgdorferi* oryginalnej presji mutacyjnej. Frakcje poszczególnych aminokwasów w całej sekwencji wygenerowanego genomu były równe. Porównanie otrzymanych wyników przedstawiono na *Ryc.26.* i *Ryc.27.* W obu przypadkach zmiana spowodowała zwiększenie liczby wyeliminowanych przez selekcję genów, a tym samym zwiększenie kosztów ewolucji (eliminacja jednego genu w praktyce oznacza śmierć noszącego go organizmu). Nasuwa to przypuszczenie, że presja mutacyjna i genom *B. burgdorferi* są do siebie dopasowane w sposób, który minimalizuje straty wywołane przy danych warunkach selekcji i funkcjonującym kodzie genetycznym. Ostatnia przeprowadzona symulacja miała odpowiedzieć na pytanie o wpływ sposobu degeneracji współczesnego kodu genetycznego na

przeżywalnośc genów *B. burgdorferi* w „standardowych" warunkach presji mutacyjnej i selekcyjnej. W tym celu zmodyfikowano wpisaną w program tablicę kodu w ten sposób, aby zachować stopień degeneracji, czyli liczbę i strukturę układów dwukrotnie i czterokrotnie zdegenerowanych kodonów, ale zmienić wzorzec przypisania do nich aminokwasów, zastępując jeden rodzaj aminokwasu przez inny o tym samym poziomie degeneracji I tak, walina kodowana przez cztery kodony mogła być podmieniona np. przez treoninę, dysponującą także pełnym boxem. W oparciu o tak takiej zmodyfikowaną tablicę kodu genetycznego wszystkie sekwencje aminokwasowe *B. burgdorferi* zostały ponownie przetłumaczone na sekwencje nukleotydowe, przy zachowaniu oryginalnego rozkładu używalności kodonów, co oznaczało, że dla każdej pozycji treoniny wylosowano nowe kodony (np. walinowe) zgodnie z rozkładem prawdopodobieństwa, z jakim pojawiały się one w oryginalnym genomie *B. burgdorferi*. Po takiej transformacji genomu kodony zmieniły swoje znaczenie, ale skład aminokwasowy sekwencji i poziom degeneracji kodu genetycznego pozostały niezmienione. Okazało się, że taka modyfikacja prowadzi do wzrostu liczby genów eliminowanych podczas symulacji *(Ryc.28)*, co świadczy o dużym wpływie nie tylko poziomu, ale i sposobu degeneracji kodu genetycznego na koszty ewolucji genomu.

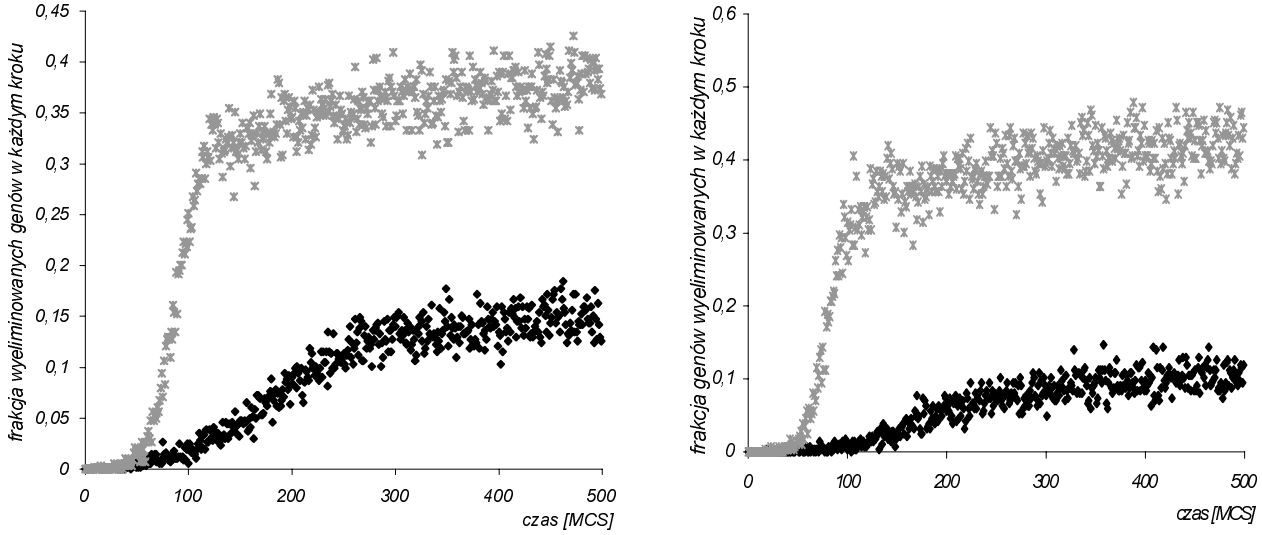
Czy układ złożony z istniejącej sekwencji i presji mutacyjnej jest w danych warunkach selekcyjnych idealnie dopasowany? Czy istnieje sekwencja jeszcze odporniejsza na panujące warunki presji mutacyjnej? Aby odpowiedzieć na to pytanie przeprowadzono symulację, podczas której poddawano sekwencję ORFów *B. burgdorferi* czystej presji mutacyjnej przez 50 000 kroków MC. Otrzymaną po zakończeniu symulacji sekwencję równowagową poddano następnie symulacji w warunkach standardowych. Okazało się, że doprowadzenie składu sekwencji do stanu równowagi z presją mutacyjną wyraźnie obniża liczbę genów eliminowanych przez selekcję *(Ryc.29)*, sekwencja kodująca *B. burgdorferi* mogłaby więc mieć skład jeszcze silniej minimalizujący liczbę potencjalnych substytucji aminokwasowych przy panującej presji mutacyjnej, nie osiąga jednak tego optimum. Prawdopodobnie można skonstruować kod genetyczny bardziej optymalny dla systemu presja mutacyjna / presja selekcyjna *B. burgdorferi*, trudno jednak przypuszczać, że byłby on tak samo zoptymalizowany w przypadku innych genomów.

Zarówno presja mutacyjna, jak i sekwencja *B. burgdorferi* mają charakter asymetryczny. Czy ta asymetria ma znaczenie dla struktury genomu i jego stabilności, a szczególnie dla kosztów ewolucji sekwencji kodujących?



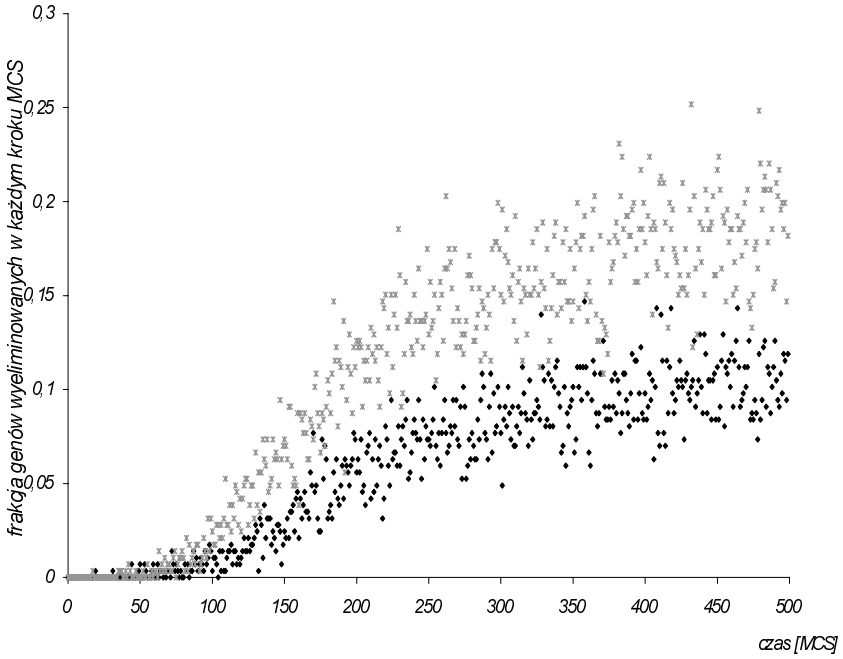
(a) **(b)**

Ryc.26. Porównanie liczby genów eliminowanych przez selekcję w warunkach standardowych (czarne punkty) i pod wpływem symetrycznej tablicy przejść (szare krzyżyki). (a) – wyniki dla nici wiodącej, (b) – dla nici opóźniającej

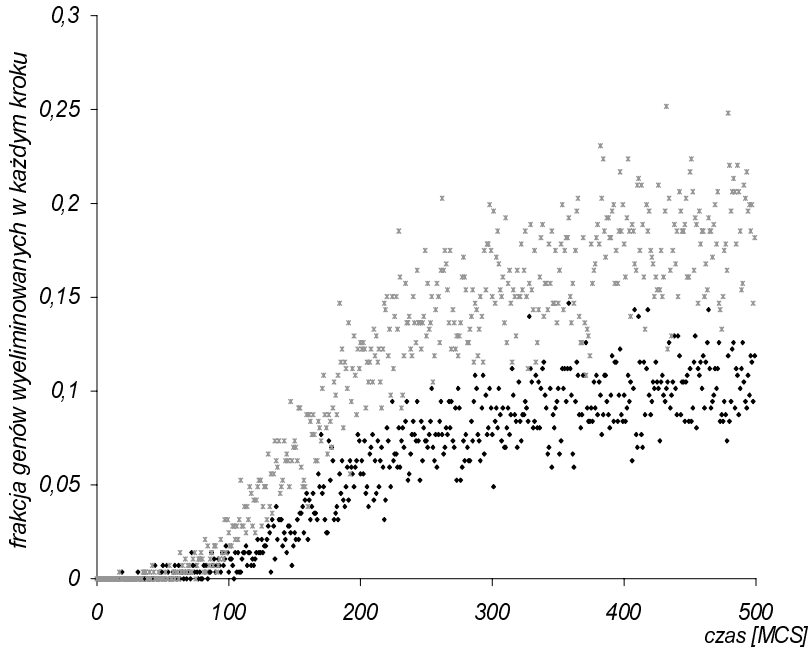


(a) **(b)**

Ryc.27. Eliminacja genów w przypadku symulacji standardowej (czarne punkty) i symulacji z sekwencją losową (szare krzyżyki) (a) – dla nici wiodącej (b) – dla opóźniającej

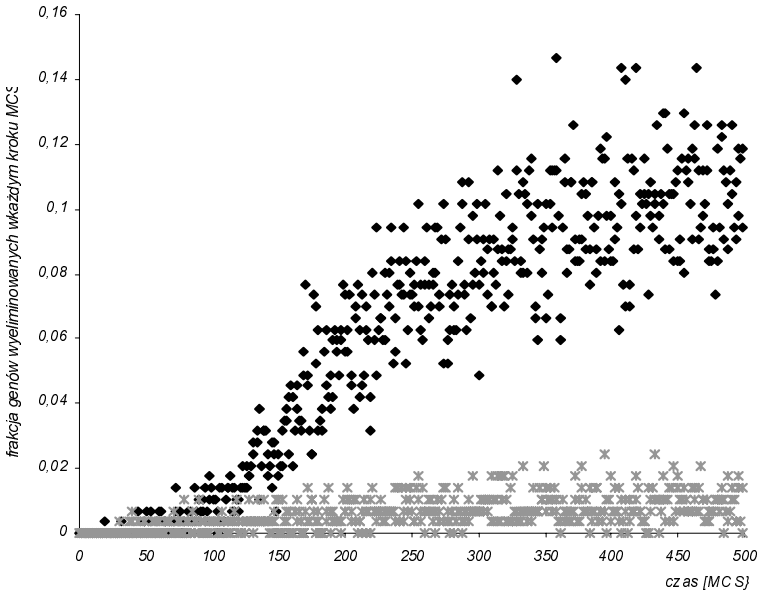


(a)

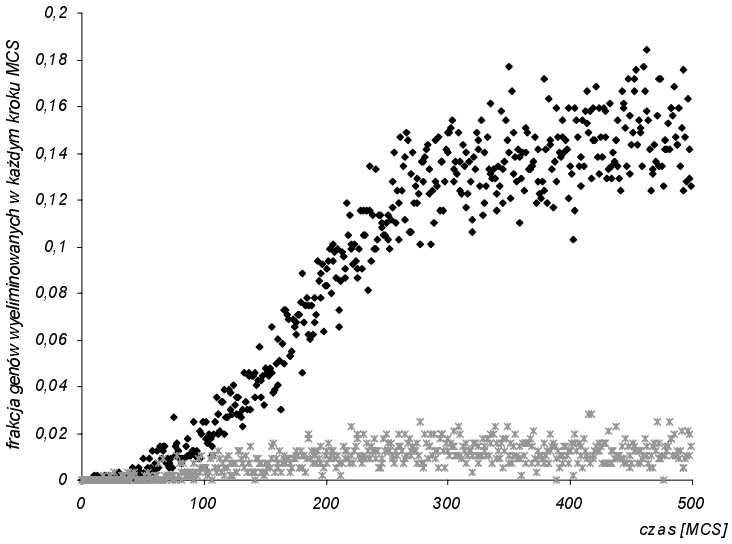


(b)

Ryc.28. Eliminacja genów podczas symulacji przeprowadzonej w warunkach standardowych (czarne kwadraty) i po zmianie sposobu degeneracji kodu genetycznego (szare krzyżyki). (a)- geny z nici wiodącej (b) – geny z nici opóźniającej



(a)



(b)

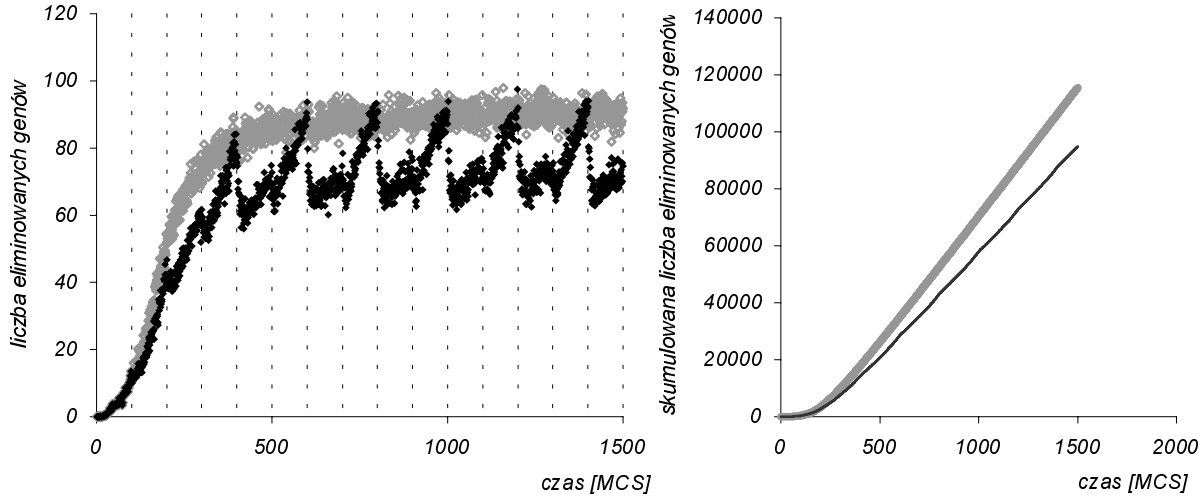
Ryc.29. Liczba genów wyeliminowanych przez selekcję po symulacji przeprowadzonej w warunkach standardowych dla sekwencji oryginalnej *B. burgdorferi* (czarne punkty) i dla sekwencji równowagowej (szare krzyżyki). (a) - sekwencje kodujące z nici wiodącej, (b) z nici opóźniającej

4.4. Analiza wpływu inwersji połączonych ze zmianą nici na przeżywalność genów

Inwersje genów są uznawane za główny mechanizm ewolucji strukturalnej genomu bakteryjnego, istnieje jednak szereg czynników ograniczających częstość takich rearanżacji *(rozdz. 1.8).* Obserwowana częstość inwersji jest więc pewnym, osiągniętym na drodze ewolucji konsensusem. Inwersje mogą prowadzić do zmiany położenia sekwencji kodującej z nici wiodącej na opóźniającą lub odwrotnie. Nasuwa się pytanie o skutki takiego przeniesienia dla przeżywalności genu. Zmiana nici łączy się z odwróceniem presji mutacyjnej działającej na gen. Skonstruowany model ewolucji pozwala na symulację warunków odpowiadających takiej sytuacji.

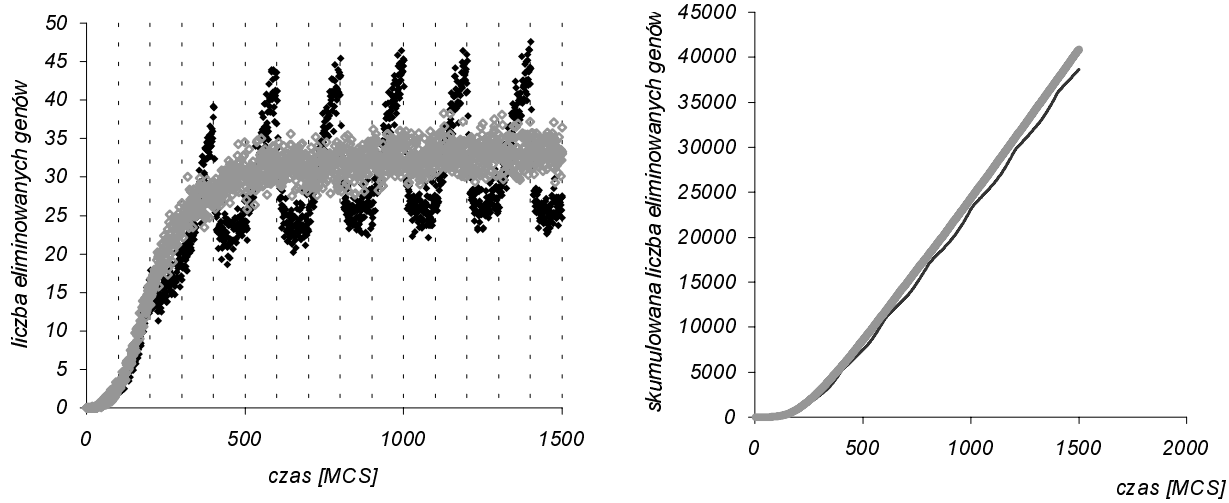
4.4.1. Porównanie zachowania się genów w warunkach stabilnej i zmiennej presji mutacyjnej.

Odpowiednikiem inwersji genu był etap algorytmu, w którym dochodziło do odwrócenia działającej tablicy substytucji. Jeżeli na przykład do mutacji został wytypowany gen z nici wiodącej, inwersję symulowano przez wprowadzenie substytucji zgodnie z prawdopodobieństwami przejść właściwymi dla nici opóźniającej. Odwrócenie tablicy można było powtarzać co krok lub co zadaną liczbę kroków (parametr wolny). Uzyskane wyniki zebrano na *(Ryc.30 i 31) (Dudkiewicz i współpr. 2003)*. Pierwszą symulację trwającą 1000 kroków Monte Carlo, przeprowadzono przy presji mutacyjnej 0,01, tolerancji na zmianę składu aminokwasowego równej 0,3, stosując stała presję mutacyjną właściwą dla danej nici. Następną symulację przeprowadzono w identycznych warunkach, zmieniając tablicę substytucji co 100 kroków. Każdy punkt na wykresie *(Ryc. 30, 31)* reprezentuje wartość średnią wyliczoną dla 10 symulacji przeprowadzonych w identycznych warunkach przy zmieniającej się liczbie inicjującej generator liczb losowych. Na osi rzędnych odłożono czas symulacji, na osi odciętych liczbę genów eliminowanych przez selekcję w każdym kroku. Linią przerywaną zaznaczono kolejne zmiany tablicy przejść. Szczególną uwagę zwraca zauważalne obniżenie liczby zabijanych genów w momencie zmiany nici, podczas gdy stałe odwrócenie presji wywołuje odwrotny skutek. Wyniki symulacji przeprowadzonej pod presją właściwą dla danej grupy genów i przy odwróconej presji przedstawia *(Ryc.32).*



(a) **(b)**

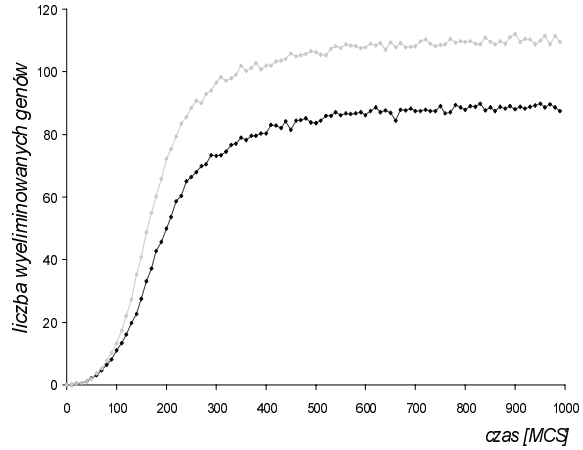
Ryc.30. Eliminacja genów z nici wiodącej podczas symulacji ze stałą presją mutacyjną (szare kwadraty i szara linia) i podczas symulacji , w czasie której presja mutacyjna zmieniała się na przeciwną co 100 MCS (czarne kwadraty i czarna linia). (a) – wartości bezwzględne (b) kumulacja danych z wykresu (a). Na osi x odłożono czas symulacji mierzony w krokach MC.



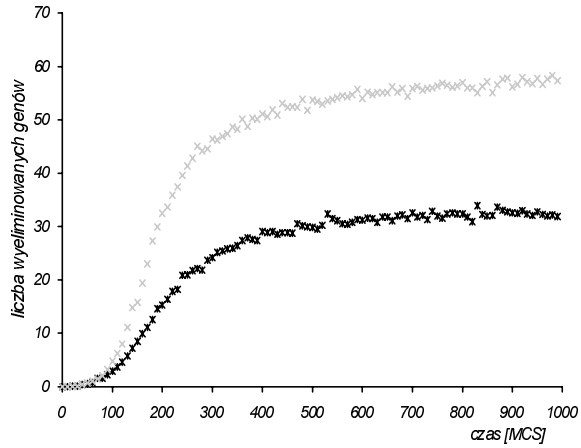
(a) **(b)**

Ryc.31. Eliminacja genów z nici opóźniającej podczas symulacji ze stałą presją mutacyjną (szare kwadraty i szara linia) i podczas symulacji, w czasie której presja mutacyjna zmieniała się na przeciwną co 100 MCS (czarne kwadraty i czarna linia). (a) – wartości bezwzględne (b) kumulacja danych z wykresu (a). Na osi x odłożono czas symulacji mierzony w krokach MC.

(a)



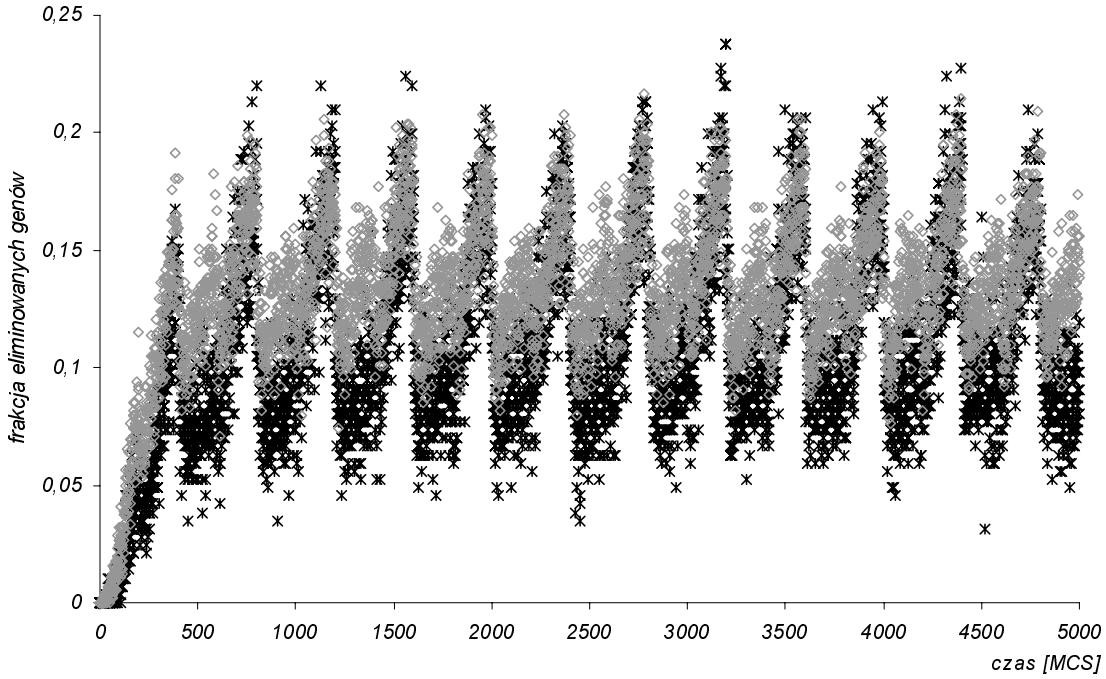
(b)



Ryc.32. Tempo eliminacji genów poddawanych presji tablicy właściwej dla nici, na której leżą (czarne punkty) i nici przeciwnej (szare punkty). (a) sekwencje z nici wiodącej, (b) – nić opóźniająca

Wybór presji mutacyjnej 0,01 był podyktowany względami natury technicznej, głównie długością czasu symulacji i pojemnością pamięci komputera. Prawdopodobieństwo mutacji równe 0,01 odpowiada wyborowi do mutacji co setnego nukleotydu. Tylko co czwarty nukleotyd z tej puli ulega rzeczywistemu podstawieniu, ponieważ około 75% przejść zachodzących zgodnie z tablicą substytucji to przejścia tożsamościowe. Szybkość ewolucji jest w przypadku tak dobranych warunków symulacji w przybliżeniu 10^6 razy większa niż w przyrodzie (spontaniczna częstość mutacji związanych z replikacją wynosi 10^{-9} w przeliczeniu na jedną parę nukleotydów i jeden cykl replikacyjny). Otrzymane wyniki należy przeskalować, przyjmując, że jeden krok Monte Carlo odpowiada w przybliżeniu 10^6 cykli replikacyjnych, a więc 100 kroków symulacyjnych to w naturze 10^8 pokoleń. Obserwowana

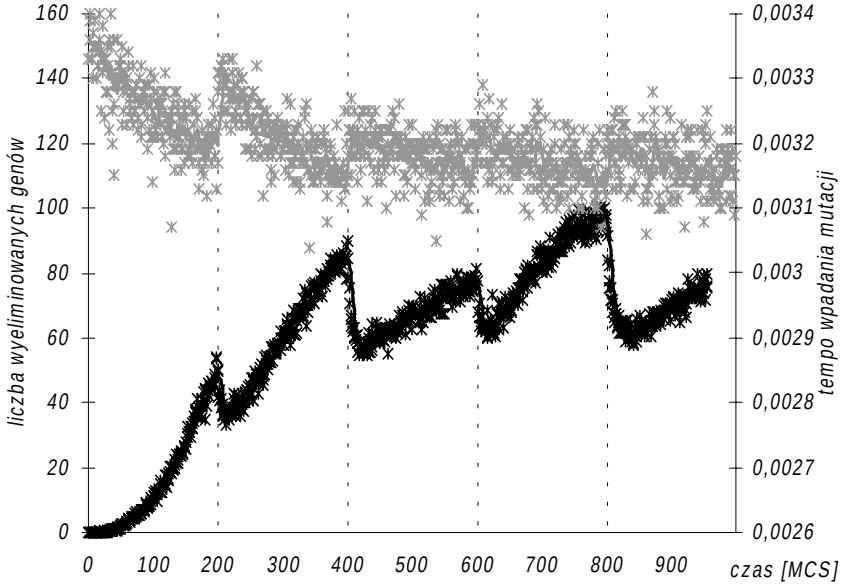
w naturze częstość zjawisk transpozycji w genomach bakteryjnych wynosi właśnie około 10^{-8}. Mechanizm transpozycji polega na rekombinacji, co prowadzi często do inwersji fragmentów DNA. Na Rycinach *30 i 31* przedstawiono wyniki symulacji standardowej i symulacji z presją mutacyjną zmieniającą się co sto kroków. Różnice w zabijaniu genów w czasie symulacji standardowej i symulacji ze zmienną presją mutacyjną, szczególnie w przypadku nici opóźniającej, nie wydają się duże. Trzeba jednak wziąć pod uwagę, że koszty ewolucji mierzymy całkowitą liczbą organizmów wyeliminowanych w czasie jej trwania, aby zinterpretować otrzymane wyniki należy skumulować je w czasie. Rezultaty takiej kumulacji ukazuje *Ryc.30.b* i *Ryc.31.b.* Aby odnieść wyniki do rzeczywistości należy także pamiętać, że każda podmiana pojedynczego genu odpowiada śmierci jednego organizmu. Obserwowane różnice mają więc w rzeczywistości duże znaczenie. Rycina *33* pokazuje dynamikę eliminacji genów w trakcie symulacji trwającej 5000 MCS, wystarczająco długo, aby sekwencja osiągnęła stan równowagi. Zmiana tablicy zachodziła co 200 kroków, co zwiększyło amplitudę wahań liczebności „ofiar ewolucji". Po około 1000 MCS liczba eliminowanych genów nie wzrasta, waha się tylko w pewnym zakresie, wyznaczanym przez częstość zmian presji mutacyjnej.



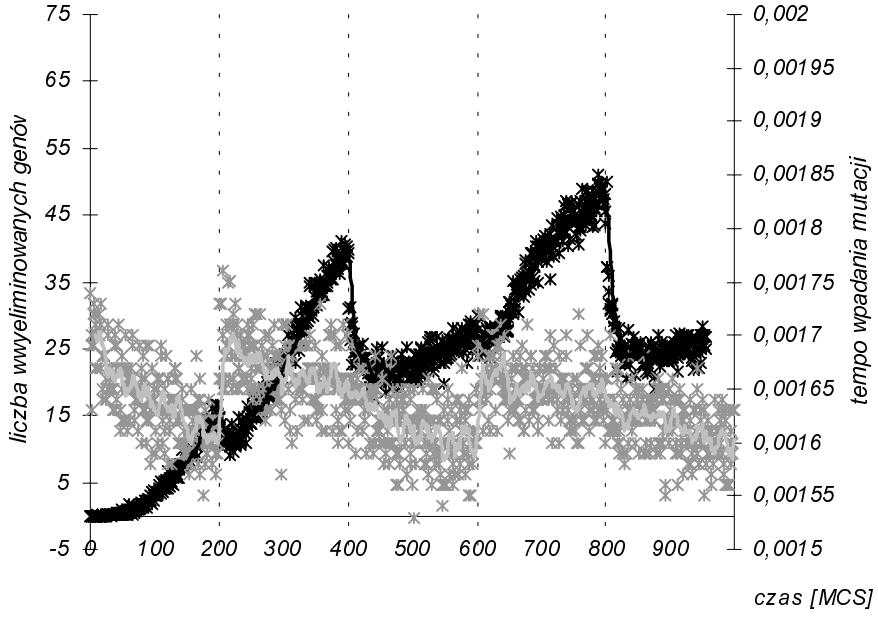
Ryc.33. Znormalizowana liczba genów eliminowanych z nici wiodącej (szare kwadraty) i opóźniającej (czarne krzyżyki) podczas symulacji trwającej 5000 kroków MC. Co 200 MCS zmieniano działającą tablicę mutacji na przeciwną.

Biorąc pod uwagę obserwowane w naturze zwiększenie dywergencji między ortologami, które w trakcie ewolucji zmieniły położenie na nici, postanowiono sprawdzić, ile substytucji zaszło w sekwencji ewoluującej pod zmienną presją mutacyjną. Na Rycinie *34.a.b* przedstawiono jednocześnie tempo eliminacji genów (oś y1) i tempo wpadania mutacji (oś y2), uzyskując odwrotną korelację. Model naśladuje więc rzeczywistość: po zmianie nici obserwujemy wzrost częstości mutacji.

(a)



(b)



Ryc.34. Dynamika eliminacji genów i częstości mutacji w czasie symulacji przeprowadzonej z odwróceniem presji mutacyjnej co 200 MCS. Na lewej osi odciętych odłożono bezwzględną liczbę wyeliminowanych przez selekcję w danym kroku genów, na osi prawej – tempo wpadania mutacji (czyli liczbę substytucji na aminokwas na 1MCS) (a) nić wiodąca, (b) – nić opóźniająca

Uzyskujemy też nową informację, której nie mogliśmy otrzymać z danych doświadczalnych – wzrost dywergencji nie oznacza wzrostu tempa eliminacji genów, przeciwnie, mutacje zachodzące po zmianie kierunku presji są częściej akceptowane przez selekcję, obserwujemy spadek liczby zabijanych genów i obniżenie kosztów ewolucji.
Opisywany efekt stwierdzono zarówno dla zbioru genów z nici opóźniającej, jak i wiodącej. Jednocześnie kształt wykresów i relacje między tempem eliminacji genów pod stałą i zmienną presją mutacyjną są nieco odmienne dla tych dwóch wyróżnionych zbiorów. Nasuwa to pytanie, czy poszczególne grupy genów różnią się reakcją na zmiany nici i jak dalece indywidualne jest ich zachowanie.

4.4.2. Analiza wpływu inwersji na wybrane geny.

Sekwencje kodujące białka rybosomalne, ze względu na ich znaczenie dla prawidłowego funkcjonowania komórki, są silnie konserwatywne, leżą prawie wyłącznie na nici wiodącej we wszystkich poznanych genomach bakteryjnych, postanowiono więc powtórzyć symulacje na wyodrębnionej grupie 51 genów rybosomalnych *B. burgdorferi*. Okazało się, że także z punktu widzenia wyników symulacji geny rybosomalne wykazują szereg specyficznych cech.
Rycina *35* przedstawia wyniki symulacji przeprowadzonej w warunkach standardowych na trzech grupach genów: genach z nici opóźniającej, genach rybosomalnych i genach z nici wiodącej bez rybosomalnych. Na wykres naniesiono średnie frakcje genów eliminowanych przez selekcję w każdym kroku symulacji, w każdej z trzech wyróżnionych grup. Zachowanie genów rybosomalnych nie odbiega wyraźnie od zachowania pozostałych genów z nici wiodącej. Różnice uwidaczniają się, jeżeli weźmiemy pod uwagę dywergencję tej grupy genów. Rycina *36* ukazuje dywergencję, czyli liczbę zmian aminokwasowych na miejsce, w analizowanych grupach genów. Geny rybosomalne wykazują najmniejsze tempo dywergencji, czyli zmieniają się najmniej w ciągu symulacji. Gromadzą więc najmniejszą liczbę mutacji, podczas gdy tempo ich eliminacji przez selekcję jest nieco niższe od wartości osiąganych przez pozostałe geny. Jaka jest przyczyna ich małej zmienności? Skoro wytłumaczeniem nie może być zwiększona eliminacja przez selekcję substytucji zmieniających skład aminokwasowy, przyczyna tkwi w mniejszym tempie wpadania mutacji do sekwencji rybosomalnych. Ilość substytucji, do których dojdzie w trakcie symulacji zależy od składu poddawanej presji mutacyjnej sekwencji wejściowej. Wynika stąd, że skład genów rybosomalnych jest dostosowany do presji mutacyjnej, tak, aby koszty ewolucji były jak

najmniejsze. Wyniki symulacji potwierdzają konserwatywny charakter sekwencji rybosomalnych. Ich konserwatyzm dotyczy jednak nie tylko sekwencji aminokwasowej, ale także położenia na nici, z analiz porównawczych genomów wynika, że geny rybosomalne praktycznie nie zmieniają nici w trakcie ewolucji. Zasadne jest więc sprawdzenie, jak zachowają się podczas symulacji przeprowadzonej ze zmianą tablicy substytucji. Ewolucję sekwencji rybosomalnych (51 genów) symulowano przez 1000 MCS w warunkach trzech różnych presji mutacyjnych: pod presją właściwą dla nici wiodącej, opóźniającej i pod presją zmieniającą się co krok. Wyniki przedstawiono na *Ryc.37*. Nie widać istotnych różnic w eliminacji genów, niezależnie od działającej presji. Geny rybosomalne we wszystkich przypadkach eliminowane były w podobnym tempie. Dlaczego więc nie obserwujemy inwersji tych genów, skoro wydają się one nie mieć wpływu na ich przeżywalność? Prawdopodobnie są za to odpowiedzialne mechanizmy selekcyjne zależne od położenia genu na chromosomie, szczególnie te związane z transkrypcją. Transkrypcja genu, którego sens leży na nici wiodącej odbywa się w kierunku zgodnym z ruchem widełek replikacyjnych. Prawdopodobnie następstwa spotkania się kompleksów replikacyjnego i transkrypcyjnego nie są aż tak poważne, jak w przypadku zderzenia czołowego, do którego może dojść, jeśli transkrypcja postępuje w kierunku przeciwnym do ruchu widełek replikacyjnych. Kolizja *„head to head”* kompleksów enzymatycznych prowadzi do dysocjacji białek i przerwania obu procesów.

Różnice w zachowaniu się genów podzielonych na trzy grupy skłaniają do zastanowienia się nad pytaniem o właściwości poszczególnych sekwencji kodujących. Nie wszystkie sekwencje preferują tę samą częstość zmiany tablicy, nie dla wszystkich koszt inwersji jest taki sam. Aby sprawdzić reakcję pojedynczych sekwencji na zmiany presji mutacyjnej wybrano cztery sekwencje o różnych długościach, trzy z nici wiodącej i jedną z nici opóźniającej. Do symulacji wybrano dwa geny z tej samej nici o różnych funkcjach i różnej istotności (gen adaptacyjny - acrB, warunkujący oporność na akryflawinę i gen niezbędny do przeżycia komórki – rpoB, podjednostka B polimerazy RNA) i dwie sekwencje o nieokreślonej funkcji (regiony przypuszczalnie kodujące: *BB0806 i BB0009*) leżące na przeciwnych niciach *(Dudkiewicz i współpr. 2004b)*. Wymienione sekwencje poddano symulacji w warunkach określonych przez osiem różnych układów zmian tablicy mutacyjnej. Wszystkie symulacje trwały po 500 MCS. Dwie pierwsze przeprowadzono bez zmiany tablicy (w warunkach standardowych i przy odwróconej presji), następne sześć przy zmianach tablicy zachodzących z określoną częstością. Wzięto pod uwagę następujące układy zmian tablicy:

1/1- zmiana tablicy następowała w każdym kroku, sekwencja połowę czasu symulacji „spędzała" na własnej nici, połowę na przeciwnej

1/2 – po jednym kroku symulacji pod wpływem tablicy właściwej dla genu następowały dwa kroki pod presją tablicy przeciwnej, sekwencja spędzała więc dwa razy więcej czasu na nici przeciwnej, niż na własnej

2/1 – sytuacja odwrotna, po dwóch krokach pod własną presją, następował jeden krok pod presją odwróconej tablicy, w tym wypadku sekwencja przebywała dwa razy dłużej na własnej nici

Podobnie należy rozumieć zapisy: 5/1, 10/1 i 1/20. W trakcie symulacji zmieniała się nie tylko częstość zmian tablicy, ale też czas pozostawania na danej nici. Jak widać na *Ryc. 38.* każdy z przedstawionych genów ma inne preferencje co do częstości inwersji i czasu pozostawania pod wpływem danej presji mutacyjnej.

Nasuwa się pytanie o przyczynę pozytywnego wpływu inwersji połączonych ze zmianą położenia na nici na przeżywalność genów. Czy taka reakcja na inwersję jest właściwością tylko sekwencji kodujących, czy też wykazują ją wszystkie sekwencje asymetryczne?

Aby odpowiedzieć na to pytanie, postanowiono sprawdzić, jak zareaguje na odwrócenie tablicy sekwencja genów doprowadzonych do stanu równowagi z presją mutacyjną, czyli sekwencja skrajnie asymetryczna, która straciła większość kodowanej informacji. Rycina *39* przedstawia wyniki dwóch symulacji, którym poddano taką równowagową sekwencję, pierwszą przeprowadzono w warunkach standardowych, drugą – przy zmiennej presji mutacyjnej (schemat zmian: 1/1). Inwersje, zamiast obniżyć, zwiększyły liczbę wyeliminowanych przez selekcję genów. Można stąd wyciągnąć wniosek, że oprócz asymetrii składu na przeżywalność genów w warunkach zmiennej presji mutacyjnej istotny wpływ mają ich właściwości związane z kodowaniem białek. Każdy gen zdaje się osiągać właściwy sobie optymalny skład, który może regulować przez zmiany położenia na nici. Ten optymalny skład jest oczywiście skutkiem długotrwałej selekcji.

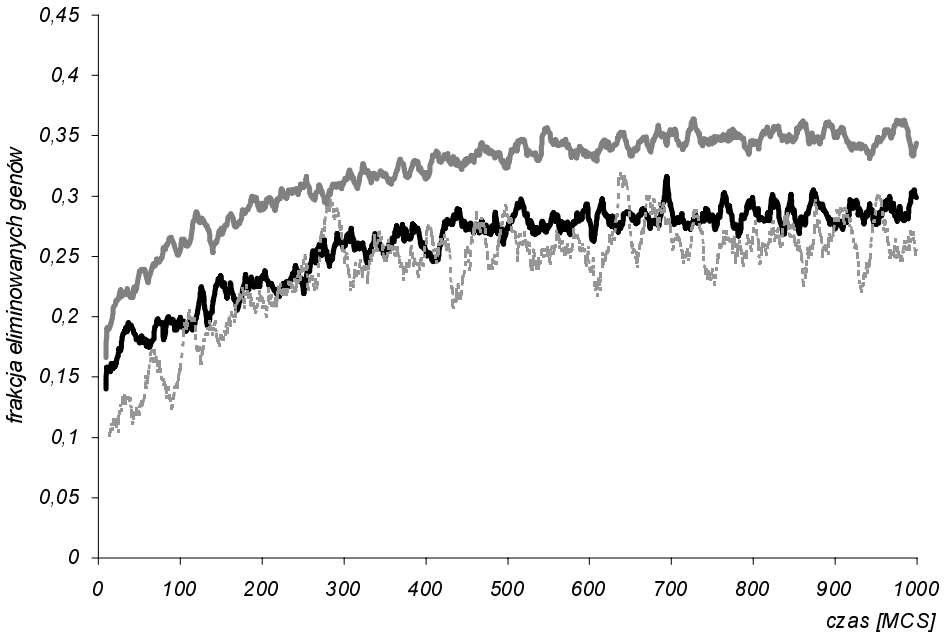
Otrzymane wyniki potwierdzają tezę, że genom bakteryjny jest wysoce zorganizowanym systemem, którego funkcjonalne elementy - sekwencje kodujące, przystosowują się w trakcie ewolucji do warunków presji mutacyjnej różnicujących loci chromosomu. Skład nukleotydowy sekwencji kodujących jest wypadkową działania kierunkowej presji mutacyjnej i sił selekcji. Istnienie na chromosomie dwóch stref poddanych przeciwstawnie działającym presjom mutacyjnym umożliwia genom odwrócenie trendów pojawiających się w ich składzie nukleotydowym pod wpływem kierunkowej presji. Konsekwencją długiego pozostawania na jednej nici jest pogłębiająca się asymetria składu sekwencji, która nie może

jednak osiągnąć maksimum wyznaczonego przez skład równowagowy, ze względu na bariery selekcyjne. Akceptowane w czasie ewolucji mutacje zmieniają stopniowo skład sekwencji w kierunku wyznaczonym przez kierunek działania presji mutacyjnej. Zgromadzone mutacje przesuwają skład sekwencji w stronę określonej granicy tolerancji na zmiany składu aminokwasowego. Gen znajdujący się na granicy tolerancji jest narażony na jej przekroczenie, każda następna substytucja może okazać się nieakceptowalna, co powoduje eliminację genu (w naturze oznacza to śmierć organizmu). Zmiana położenia na nici, a tym samym odwrócenie kierunku presji mutacyjnej nie dopuszcza do przekroczenia granicy tolerancji, a może doprowadzić do pojawienia się rewersji i supresji wewnątrzgenowych.
Inwersje połączone ze zmianą położenia na nici mogą być skutecznym mechanizmem obrony przed szkodliwym wpływem kierunkowej presji mutacyjnej. Jest to mechanizm charakterystyczny dla asymetrycznych genomów prokariotycznych. W genomach eukariotycznych jest kilka miejsc inicjacji replikacji, włączanych i działających w różnym czasie, niekiedy ze sporymi opóźnieniami. Dany gen może być w jednym cyklu replikowany w sposób wiodący, a w następnym – opóźniający, presją mutacyjna jest więc z definicji zmienna, nie obserwuje się związanej z replikacją asymetrii składu nukleotydowego. Między sąsiednimi miejscami inicjacji replikacji tworzy się jednak pewien gradient zmienności presji: im bliżej eukariotycznego ORI znajduje się sekwencja, tym większe ma szanse na stabilne warunki replikacji, im leży dalej, tym bardziej zmiennej presji mutacyjnej podlega. Czy sekwencje eukariotyczne mają możliwość regulacji kierunku presji, która na nie oddziałuje? Być może alternatywą dla inwersji w genomach eukariotycznych są sekwencje międzygenowe, które stanowią znaczny procent ich składu i mogą się przemieszczać, zmieniając jednocześnie odległość sekwencji kodujących od miejsc inicjacji replikacji.

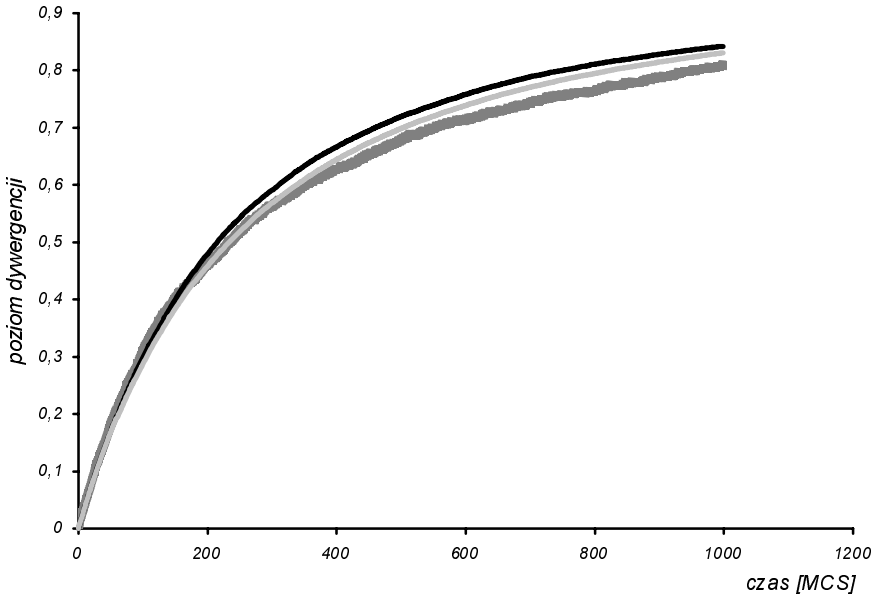
Stworzony model ewolucji sekwencji prokariotycznych pod zmienną presją mutacyjną pomógł odpowiedzieć na pytanie o przyczynę zwiększonej dywergencji obserwowanej między ortologami pochodzącymi z blisko spokrewnionych genomów, które zmieniły położenie na nici w trakcie ewolucji. Zwiększona dywergencja świadczy w tym przypadku o zwiększonym tempie mutacji, które nie jest jednak połączone ze zwiększoną eliminacją genów, ale ze wzrostem liczby mutacji zaakceptowanych.

Opisany model ewolucji opiera się na autentycznych, wyliczonych na podstawie rzeczywistych sekwencji tablicach przejść mutacyjnych, także parametr selekcji oszacowano, korzystając z autentycznych sekwencji. Tworząc algorytm starano się uwzględnić jak najwięcej przesłanek biologicznych, wykorzystując dostępne dane i zminimalizować liczbę parametrów wymagających arbitralnego ustalenia. O biologicznej wiarygodności tak

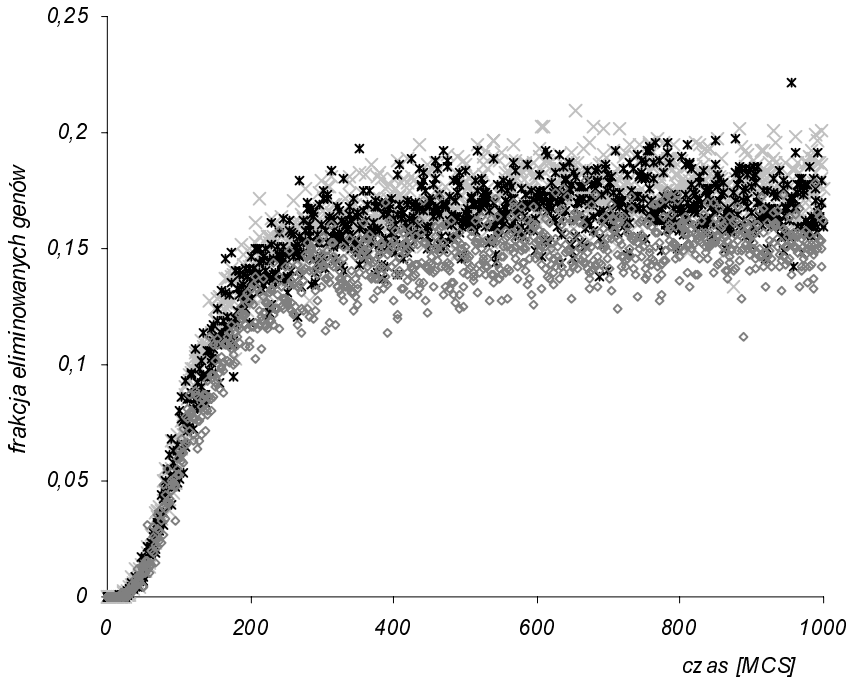
skonstruowanego modelu może świadczyć wysoki współczynnik korelacji (ok. 0,7) uzyskany dla tablic PAM1 obliczonych na podstawie sekwencji uzyskanych drogą symulacji i oryginalnych tablic PAM1. Na podstawie uzyskanych wyników można wnioskować, że asymetria obserwowana w składzie sekwencji prokariotycznych ma o wiele większe znaczenie dla ewolucji genomu niż przypuszczano. Asymetryczna presja mutacyjna generuje asymetrię sekwencji nukleotydowej. Razem tworzą one bilansujący się i ulegający optymalizacji układ, który umożliwia obniżenie kosztów ewolucji.



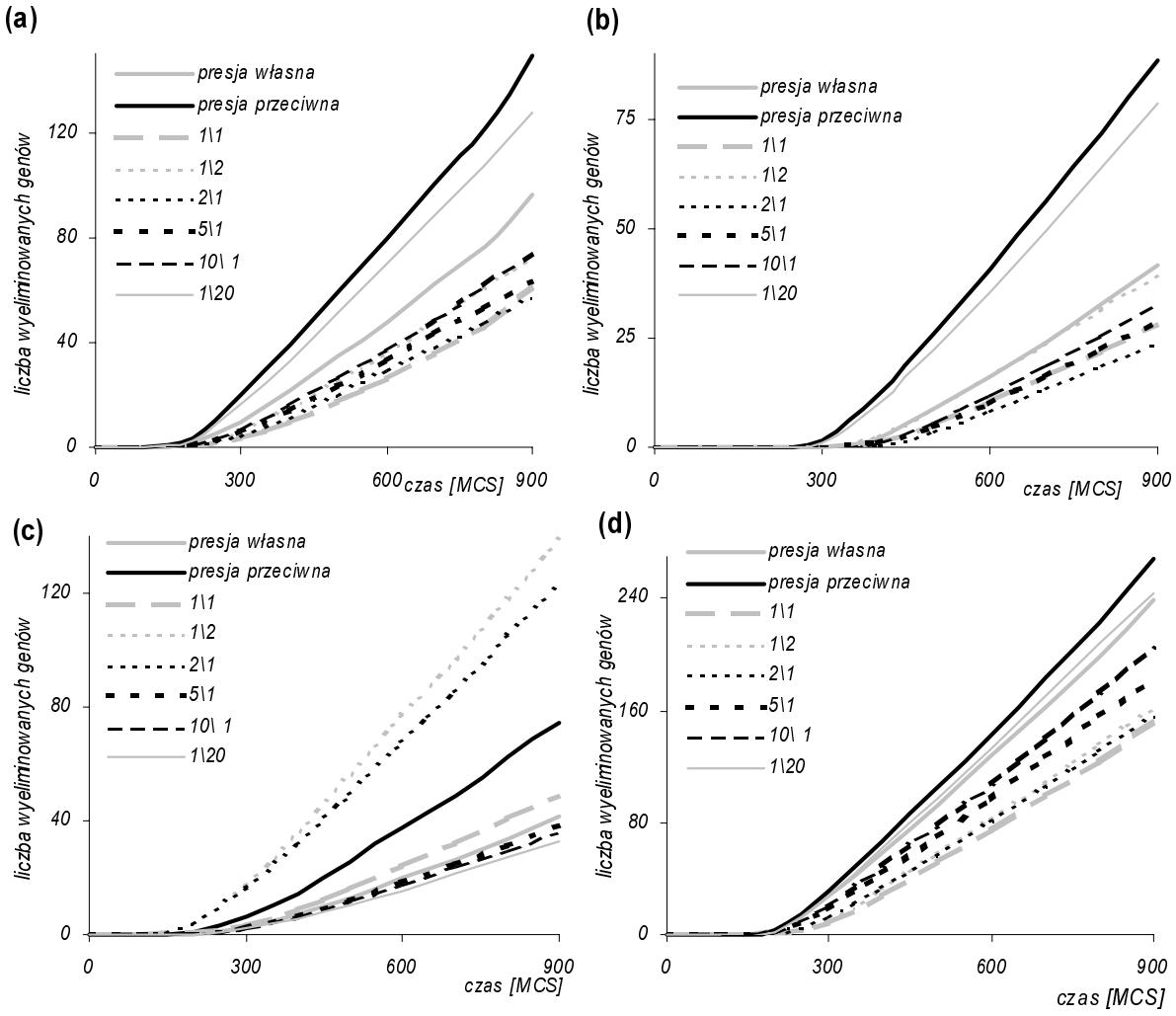
Ryc.35. Znormalizowane tempo eliminacji genów należących do trzech różnych grup podczas 1000 kroków symulacji przeprowadzanej w warunkach standardowych. Szara przerywana linia- sekwencje rybosomalne, czarna linia – sekwencje z nici opóźniającej, szara ciągła linia - sekwencje z nici wiodącej poza rybosomalnymi. Na osi y odłożono znormalizowaną przez liczebność danej grupy liczbę genów podmienionych w danym kroku MC.



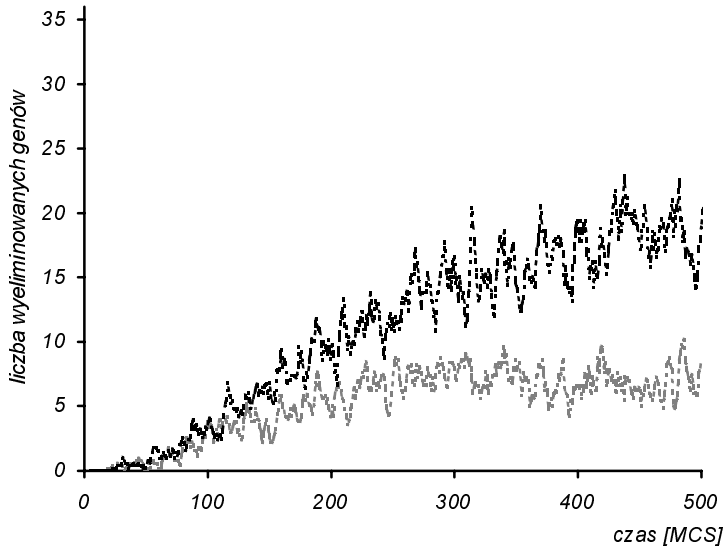
Ryc.36. Dywergencja mierzona ilością zmian aminokwasowych na miejsce w porównaniu do sekwencji wejściowej w grupie genów rybosomalnych (szara, pogrubiona linia), genów z nici opóźniającej (czarna linia) i wszystkich genów z nici wiodącej poza rybosomalnymi (szara, cienka linia). Na osi x – czas symulacji w MCS



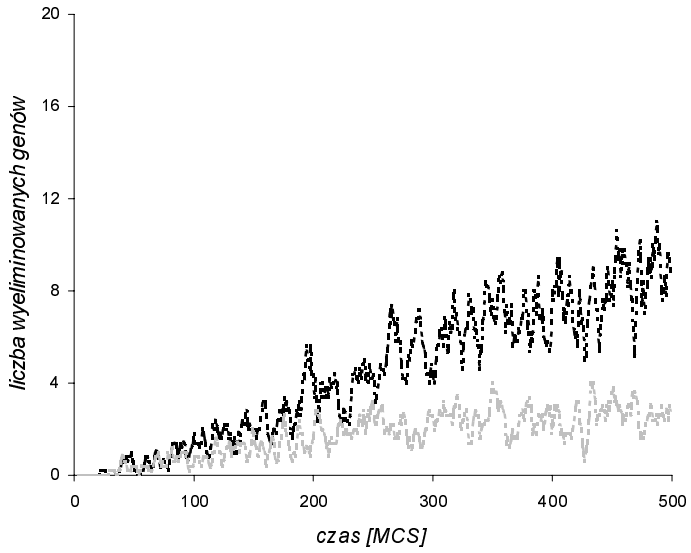
Ryc.37. Eliminacja genów rybosomalnych podczas symulacji przeprowadzonych w warunkach standardowych (czarne krzyżyki), pod presją mutacyjną właściwą dla nici opóźniającej (szare krzyżyki) i w warunkach zmiennej presji (zmiana tablicy następowała co krok - szare kwadraty).



Ryc.38. Eliminacja czterech różnych sekwencji w różnych warunkach presji mutacyjnej. Na osiach y odłożono zakumulowane liczby genów podmienionych w czasie ośmiu symulacji, których warunki opisano w legendzie *(patrz: tekst str. 70-71)* na osiach x – czas symulacji. (a) BB0806 (nić wiodąca) (b) (acrB) (nić wiodąca) (c) BB0009 (nić opóźniająca) (d) (rpoB) (nić wiodąca).



(a)



(b)

Ryc.39. Eliminacja sekwencji genów doprowadzonych do stanu równowagi z presją mutacyjną w warunkach symulacji standardowej (szara linia) i podczas symulacji ze zmienną presją mutacyjną (czarna linia). (a) nić wiodąca (b) nić opóźniająca

5. WNIOSKI

- symetryzacja presji mutacyjnej lub składu aminokwasowego sekwencji kodujących powoduje wzrost tempa eliminacji genów w porównaniu do oryginalnej, asymetrycznej tablicy przejść i oryginalnego składu
- presja mutacyjna i presja selekcyjna w danym genomie ulegają optymalizacji w warunkach kodowania narzuconych przez istniejący kod genetyczny
- stała zamiana presji mutacyjnej, pod której wpływem pozostaje gen na presję właściwą dla nici przeciwnej powoduje wyraźny wzrost liczby genów eliminowanych przez selekcję
- okresowe inwersje genów połączone ze zmianą nici wywołują obniżenie tempa ich eliminacji przez selekcję
- inwersje zachodzące w czasie ewolucji sekwencji ortologicznych powodują wzrost dywergencji między nimi, ponieważ wzrost częstości mutacji wpadających i obniżenie tempa eliminacji mutacji powoduje zwiększenie liczby akumulowanych substytucji.
- wpływ inwersji na geny rybosomalne jest niewielki, ich konserwatywne położenie należy tłumaczyć przyczynami innymi niż bezpośredni wpływ inwersji na tempo mutacji
- poszczególne geny różnią się preferowaną częstością inwersji i czasem pozostawania pod wpływem danej presji
- asymetryczna presja mutacyjna jest mechanizmem zmniejszającym koszty ewolucji genomu prokariotycznego

A N E K S

6.1. Zasady symulacji Monte Carlo

Wyniki komputerowych symulacji opartych na generowaniu liczb losowych zostały opublikowane po raz pierwszy w latach czterdziestych naszego wieku *(Metropolis i Ulam 1949).* Wykonanie instrukcji algorytmu programu symulacyjnego opartego na tej metodzie jest uzależnione od liczby r losowanej z zakresu od 0 do 1. Jeżeli liczba r jest mniejsza niż ustalone w programie prawdopodobieństwo p, pętla instrukcji jest wykonywana, jeżeli p < r program pozostaje w stanie wejściowym. Najprostszą metodą generowania liczb losowych jest mnożenie reszty ułamkowej liczby inicjującej generator przez pewną wartość *(Dudek 1992).* Głównym wymaganiem metody jest możliwość opisania symulowanego układu konkretną funkcją rozkładu prawdopodobieństwa, możemy wtedy naśladować wybrany system przez losowe typowanie zmiennych spełniających zadaną funkcję. *(http://csep1.phy.ornl.gov/mc/node1.html.)* Innymi słowy na wyjściu otrzymujemy szereg wyników, które będą obrazowały zachowanie się naszego układu w pewnym odcinku czasu Oczywiście takie strzelanie na oślep do rozkładu prawdopodobieństwa wymaga pocisków,

czyli generowanych losowo liczb z przedziału od zera do jedynki. Zasadę działania najprostszego generatora liczb losowych opisano poniżej. Metoda Monte Carlo w przeciwieństwie do konwencjonalnych metod numerycznych nie wymaga znajomości dokładnych równań różniczkowych opisujących układ, symulacja zjawisk stochastycznych wymaga tylko znajomości zasad statystyki.

Program do symulacji metodą Monte Carlo powinien posiadać kilka podstawowych elementów:

- po pierwsze – funkcję rozkładu prawdopodobieństwa dla modelowanego układu (fizyczny, biologiczny lub matematyczny system musi być opisany zbiorem funkcji *pdf*)
- po drugie – generator liczb losowych, źródło zbioru liczb losowych z przedziału [0,1]
- po trzecie określoną regułę typowania zmiennych z danego *pdf* (probability density function) *(sampling rule)*
- po czwarte wreszcie – algorytm opisujący sposób oznaczania i interpretacji otrzymywanych wyników.

Oczywiście program powinien mieć także określony sposób oceny błędu statystycznego i metodę redukcji wariancji.

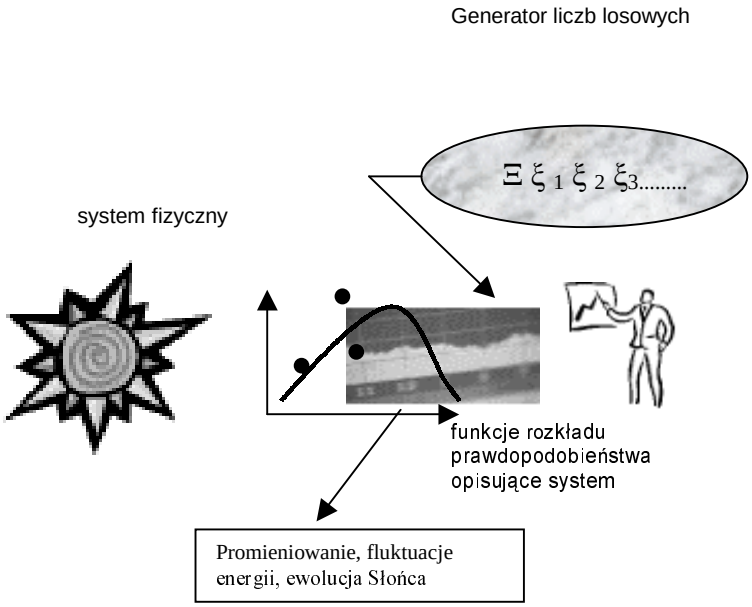
Rycina *40* przedstawia schemat modelowania zachowania Słońca jako układu fizycznego za pomocą symulacji typu MC Podstawowym wymaganiem metody jest znajomość krzywej rozkładu prawdopodobieństwa dla układu (może to być zbiór funkcji, jeżeli chcemy modelować różne aspekty zachowania się systemu, np. fluktuację wiatru słonecznego, zmiany wielkości promieniowania, ewolucję Słońca jako źródła energii itp.). Następnym warunkiem jest wygenerowanie zbioru liczb losowych z zakresu 0-1, które powinny być w tym przedziale rozłożone równomiernie. Taka liczba losowa „trafia" albo powyżej albo poniżej krzywej *pdf*.

Zależnie od tego program wykonuje określony algorytm i uzyskujemy pewną wartość wyjściową. Otrzymane wyniki muszą być odpowiednio oznaczone w programie, od tego zależy ich poprawna interpretacja.

Opisany moduł wprowadza losowy rozkład zdarzeń, który może być punktem wyjścia dowolnego typu symulacji. Zastosowanie symulacji w badaniach biologicznych ma niezbyt długą, lecz ciekawą historię. Podejmowane do tej pory próby wykorzystania metod informatycznych i symulacyjnych do lepszego poznania procesów ewolucyjnych można podzielić na trzy grupy. Do pierwszej zaliczamy symulacje prowadzące do powstania złożonych silikonowych światów, będących rodzajem alternatywnych ekosystemów. Prawa rządzące ich rozwojem wykazują niejednokrotnie zadziwiającą zbieżność ze znanym nam światem ożywionym *(Coveney, Highfield 1997)*. Drugą grupą są symulacje, które traktują

zasady ewolucji jako narzędzie pozwalające na otrzymanie pożądanego rozwiązania, czyli tzw. algorytmy genetyczne *(Holland, 1975).* Wreszcie trzecią grupą są symulacje wprowadzające pewne wstępne parametry zaczerpnięte z danych doświadczalnych i obserwacji przyrody, aby obserwować zachowanie się układów jak najbardziej zbliżonych do naturalnych i budować modele naśladujące ich rzeczywistą ewolucję *(Ausloos i współpr. 1997; Kowalczuk 2001b).*

Opisana w poniższej pracy symulacja i jej wyniki należą do grupy trzeciej. Metody komputerowe, możliwość generowania zdarzeń losowych są traktowane jako narzędzie pozwalające na lepsze poznanie zasad rządzących ewolucją genomów.



Ryc.40. Schemat algorytmu typu Monte Carlo – punkty nad i pod krzywą rozkładu prawdopodobieństwa obrazują efekty kolejnych losowań.

„Niczym nie da się tak wymodelować rzutów kością, jak tylko rzutami innej kostki”

Stanisław Lem: „Filozofia przypadku”

6.2. Rola symulacji komputerowych we współczesnej nauce

Dzięki rozwojowi technik komputerowych metodologia dzisiejszej nauki wzbogaciła się o wiele zupełnie nowych, a ułatwiających i przyspieszających pracę narzędzi, takich jak programy obliczeniowe, analityczne i symulacyjne. O ile o pożytku z tych zdobyczy postępu nie trzeba przekonywać matematyków czy fizyków, o tyle biologowie odnoszą się do nich ze sporą dozą nieufności.

„Jedną z ważniejszych metod odsłaniania zasad i zależności istotnych w przyrodzie jest konstruowanie metafor rzeczywistego świata, w których oszałamiające komplikacje mikroskopowe zostają poświęcone w celu zrozumienia ogólnego obrazu, mówiąc krótko chodzi o znalezienie sposobu, aby dojrzeć las, a nie tylko pojedyncze drzewa. Takie przenośnie mogą zilustrować, jak złożoność lasu wynika ze splątania drzew, krzewów i poszycia”- ten obszerny cytat z książki „Granice złożoności” P. Coveney’a i R. Highfielda *(1997)* streszcza powody konstruowania uproszczonych modeli złożonych zjawisk

przyrodniczych. Celem uproszczeń, według wielu tradycyjnych biologów zawsze zbyt daleko posuniętych, jest odrzucenie szczegółów, które zaciemniają obraz ogólnych praw rządzących zjawiskami. Prawidłowo skonstruowany i działający model nie może mieć zbyt wielu parametrów, gdyż większość z nich z konieczności byłaby wprowadzana arbitralnie. Modelowanie komputerowe polega raczej na ograniczaniu liczby parametrów, aby uniemożliwić manipulację wynikami, które powinny być konsekwencją wewnętrznych, nie narzuconych przez eksperymentatora, reguł i powiązań tworzących się w miarę wzrastania poziomu złożoności systemu. Takie podejście jest zgodne z rozwijającą się ostatnio teorią systemów i teorią złożoności, w świetle których życie jest jedną z postaci złożonego, podlegającego ewolucji systemu, rządzą nim więc te same zasady, które rządzą procesami narastania kryształów, ferromagnetyzmu, czy zachowaniem się gazów i mieszanin.

Wymienione w *rozdz. 6.1* rodzaje zastosowań symulacji komputerowych różnią się sposobem traktowania i wprowadzania parametru selekcji. W przypadku sztucznych światów generowanych *in silico* brak selekcji narzuconej z zewnątrz, „organizmy" walczą o przetrwanie konkurując o zasoby pamięci i czas procesora. W algorytmach genetycznych selekcja spełnia rolę utylitarną, zwycięża ten, kto najlepiej spełnia wyznaczone zadanie. Wreszcie w trzeciej grupie symulacji warunki selekcji narzuca przyroda (np. kryterium zachowania funkcji białka).

Mnogość silikonowych światów, które można obserwować za pośrednictwem internetu, jest ogromna, choć daleko jej jeszcze do bioróżnorodności. Warto przybliżyć chociaż niektóre z nich.

Jedną z wcześniejszych prób symulowania ewolucji w komputerze był projekt „Biomorfy" R. Dawkinsa. Zasady doświadczenia były bardzo proste, każdy „biomorf" był kodowany przez zespół dziewięciu genów określających kierunek, długość i głębokość rozgałęzień głównego pnia. Liczba możliwych kombinacji sięgała $5^* 10^{11}$. Początkowo celem było wygenerowanie kształtu drzewa, przy czym jedynym środkiem jego osiągnięcia było dychotomiczne rozgałęzianie wyjściowego pnia. W każdej iteracji powstawały dwie nowe gałęzie, do następnego kroku przechodził kształt wybrany przez obserwatora, którego oko pełniło rolę naturalnej selekcji. Ostateczny wynik przekroczył oczekiwania, Dawkins otrzymał zadziwiające spektrum fraktalnych bio-form: od „samolotów: przez „owady", „nietoperze", „pajaki" po różnego rodzaju „drzewa"*(Dawkins 1996).*

Znanym od dawna fenomenem samoorganizacji są tzw. automaty komórkowe (Cellular Automata - CA). Początki ich historii sięgają lat 40-tych – badań S. Ulama i idei kinematonu von Neumana. Swój rozwój i popularność zawdzięczają jednak znanej wśród intelektualistów

lat 70-tych grze J. H. Conway'a nazwanej po prostu *„Game of Life" (Coveney i Highfield 1997).* Reguły gry były bardzo proste. Podstawą była dwuwymiarowa siatka złożona z komórek o dwóch dopuszczalnych stanach: ON i OFF. Każda komórka miała ośmiu sąsiadów, od których zależał jej stan. Tym prostym układem rządziły proste zasady:

- jeżeli komórka jest wyłączona (nieożywiona) i ma trzech „żywych" sąsiadów, będzie włączona w następnym kroku
- jeżeli komórka „żyje" i ma dwóch lub trzech „żywych" sąsiadów, pozostanie włączona w kolejnym kroku
- jeżeli komórka nie ma żadnych lub ma tylko jednego włączonego sąsiada, nie przeżyje następnego kroku.

W ten sposób z układu komórek na siatce powstawał automat, który generował niepowtarzalne wzory odzwierciedlające trzy możliwe sytuacje: zanik ożywionych komórek , czyli wymarcie „życia", zachowanie stanu *„still life",* czyli zamrożenie automatu na pewnym etapie ewolucji oraz powstanie oscylatorów i samoreplikujących się wzorów. Dzięki automatom komórkowym po raz pierwszy zaobserwowano *in silico* zjawisko emergencji – całość okazała się czymś więcej, niż tylko sumą części składowych, między kolejnymi poziomami organizacji wytworzyła się luka jakościowa *(Gutowitz 1991).* Dowodem na to były zadziwiające struktury wytwarzane przez automaty komórkowe, czyli szybowce – propagujące się wzdłuż jednej prostej samoreplikujące się kształty oraz oscylatory – powtarzające się co pewien okres układy komórek. Automaty komórkowe można podzielić na cztery klasy, w zależności od stadium, które osiągają po pewnym czasie *(Wolfram 1986).* Ewolucja klasy I prowadzi do homogenizacji struktury automatu, w wyniku ewolucji CA klasy II powstają proste periodyczne struktury. Automaty klasy III po pewnym czasie osiągają stan chaotyczny, i w końcu na drodze ewolucji CA klasy IV powstają kompleksowe, niepowtarzalne struktury, bogactwem form przypominające życie. Automaty komórkowe znalazły zastosowanie w tak różnorodnych dziedzinach jak: modelowanie zachowania się gazów, zjawisk ferromagnetyzmu, procesów perkolacyjnych, przewidywanie propagacji pożarów lasów, rozwoju kompleksów urbanistycznych, modelowanie procesów krystalizacji a nawet generowanie obrazów i tworzenie grafiki komputerowej. Modyfikacją dwuwymiarowych CA są automaty jednowymiarowe – drugi wymiar płaszczyzny reprezentuje czas, można więc obserwować dynamikę takiego automatu w czasie, dostrzegając, że historia jego „życia" ma strukturę fraktalną. Inne modyfikacje dwuwymiarowych automatów polegają na manipulacji ilością znaczących komórek, ilością dopuszczalnych stanów lub wagą dla poszczególnych sąsiadów *(Langton 1989).*

W miarę rozwoju informatyki pojawiały się nowe systemy algorytmów naśladujące życie organiczne. Jednym z ciekawszych był prosty symulator M. Ray'a *Tierra*. *Tierra* to sztuczny system samoadaptacyjny, stworzony w języku programowania opartym o pewną liczbę wymienialnych (alternatywnych) instrukcji, analogicznych do aminokwasów w kodzie genetycznym. Każda instrukcja była pięciobitowym ciągiem. Zmiana jednego bitu powodowała tranzycję jednej instrukcji w inną tej samej grupy, a nie jej bezpośrednie zniszczenie *(Ray 1992, Adami 1995).* Ciągi kodów instrukcji spełniały rolę DNA, przy czym istniał pewien minimalny ciąg konieczny do przetrwania i powielenia się w pamięci komputera. Najciekawszym wynikiem otrzymanym po pewnym czasie swobodnej ewolucji było samorzutne powstanie populacji *pasożytów*, czyli algorytmów o ciągu instrukcji krótszym od minimalnego, które potrafiły wykorzystać do autoreplikacji fragmenty kodów programów – gospodarzy. Co jeszcze ciekawsze, następnym stadium ewolucji symulatora było pojawienie się populacji algorytmów odpornych na *pasożyty*, która w końcu wygrała ewolucyjny „wyścig zbrojeń" i opanowała CPU komputera.

Rozwinięciem jednowymiarowego schematu *Tierry* był system *Plateau* Maley'a. Dwuwymiarową płaszczyznę zamieszkiwały *Loopy,* czyli *pętlasy*, stworzenia składające się z koncentrycznych pętli instrukcji. Także i tu pojawiła się cała różnorodność form, w tym także *pasożyty.* Inną modyfikacją *Tierry* był model *C-zoo* Skippera – dwuwymiarową płaszczyznę rozpięto tym razem na torusie, przez który maszerował szereg „mrówek" szukających pożywienia. Każda mrówka była zbiorem czterech komórek zawierających trzydzieści dwie instrukcje. W miejscu gdzie spotykały się dwie „mrówki" dochodziło do konkurencji i walki o byt. Proste w założeniu modele zaowocowały nadzwyczaj bogatą menażerią zachowań i form. Bardziej skomplikowane próby organizacji sztucznego życia, jak na przykład projekt *E-den*, doprowadziły raczej do uproszczeń. *E-den* to zaprogramowana w pamięci komputera siatka zamieszkiwana przez stworzenia obrazowane przez ciągi cyfr od 0 do 9 (0 – próżnia, 1-9 materia: atomy ożywione i nieożywione). Każdy atom miał ładunek, energię i kolor, jeśli był ożywiony, także funkcję. Każdy organizm złożony z atomów miał swój genom, ciąg cyfr określający jego konstrukcję i funkcje. Sztuczne organizmy miały tu ogromne, nieporównywalne ze światem *Tierry* możliwości – mogły się poruszać, zmieniać kształt, rozmnażać się, pobierać pożywienie, przekazywać sygnały, konkurować i kooperować. Wyniki ewolucji tak złożonego systemu okazały się jednak nadzwyczaj skromne w porównaniu do tak ogromnego potencjału różnorodnych zachowań. Najbardziej wyrafinowane formy „życia" w *E-denie* miały postać kolonii organizmów, zbliżonych trochę do kolonii bakteryjnych, nie przejawiały jednak żadnych ciekawych zachowań. Symulacje

odegrały ważną rolę także w modelowaniu zachowań wyższych organizmów. Najprostsze symulatory naśladujące zachowania żywych stworzeń, tzw. *boidy* C. Reynoldsa, doskonale oddawały zjawiska grupowania i rozpraszania się organizmów. Podobne założenia leżały u podstaw modelu *A-Quarium*, w którym każda „ryba", modelowana przez osobny algorytm, starała się płynąć blisko sąsiada, jeżeli należał do tego samego gatunku, a uciekać, gdy reprezentował inny gatunek. Zachowanie całego *A-Quarium* było złożeniem zachowań poszczególnych „ryb". Bardziej skomplikowany model zachowania się ryb stworzył D. Terzopoulos. Każdą wirtualną rybę opisał przez indywidualny program komputerowy stanowiący część większego programu generującego prosty podmorski ekosystem. „Ryby", wyposażone w dopracowaną graficznie „cielesną powłokę", uczyły się pływać drogą drobnych, adaptacyjnych zmian w algorytmie. W miarę upływu czasu wzrastała kompleksowość ich zachowań: tworzyły ławice, robiły uniki, rozpraszały się, a nawet wytworzyły zachowania godowe. Mózgiem sztucznej ryby była sieć neuronowa, co pozwoliło im na naukę i stopniową komplikację zachowań.

Symulacje komputerowe umożliwiają nie tylko modelowanie behawioru zwierząt, ale także wzrostu roślin. Stworzeniu wirtualnych roślin służyły *L-systemy* Lindenmayera. Celem było opracowanie formalnego, matematycznego modelu wzrostu roślin, opartego na metodach zbliżonych do reguł gramatycznych, którymi posługują się lingwiści przy rozbiorze zdań. Organizm roślinny został przedstawiony w postaci łańcucha symboli oznaczających poszczególne moduły (liść, ogonek, pąk). Efekt wzrostu uzyskano przez manipulację tymi symbolami w trakcie stosowania odpowiedniego algorytmu. Nawet proste zbiory reguł doprowadziły do powstania „roślin", które wyglądały jak paprocie, bzy czy astry *(Prusinkiewicz i Lindenmayer 1990).*

Jednym z najnowszych i jednocześnie najbardziej skomplikowanych symulatorów jest projekt *Framestick*. Jest to rozbudowany system składający się z trójwymiarowej, mechanicznej symulacji sztucznego świata (moduł *Mechastick*) oraz organizmów posiadających określony genotyp i fenotyp, sterowanych przez sieci neuronowe i zaopatrzonych w pętle przetwarzania bodźców. Symulator dozwala na parametryzację większości operacji, właściwości środowiska, zasad symulacji i ewolucji systemu, umożliwia więc obserwację zarówno ewolucji otwartej, jak i ukierunkowanej. Dodatkową zaletą jest efektowny interfejs graficzny pozwalający na tworzenie krótkich filmów „z życia *Framestickó*w", zamieszczanych na stronach www. Wobec tak wielkiej liczby wprowadzanych z zewnątrz parametrów model jest raczej rodzajem efektownej gry w życie niż naukową symulacją. Podstawowy moduł budowy „*Framesticka*" to patykowate odnóże, które może ulec specjalizacji w kierunku odbioru

bodźców, zwiększonej wytrzymałość, szybkości itd. Moduły te budują stworzenia przybierające w trakcie ewolucji formy przypominające meduzy, skorpiony, węże, pseudoryby itd.

Na koniec warto wspomnieć o regułach rządzących algorytmami genetycznymi. Początki tej metody optymalizacyjnej sięgają lat 60-tych, kiedy J. Holland ogłosił jej podstawowe założenia *(Holland 1975).* Zgodnie z nimi pula genetyczna dużej populacji zawiera potencjalne rozwiązanie każdego problemu adaptacyjnego. Jest ono jednak początkowo nieaktywne, ponieważ potrzebne do jego utworzenie geny są rozproszone w populacji. Aby powstało rozwiązanie, optymalna kombinacja genów musi spotkać się w tym samym osobniku. Aby do tego doprowadzić, należy zakodować problem i jego rozwiązania w postaci ciągów binarnych, a następnie wygenerować losową populację, której pula genetyczna reprezentowałaby grupę możliwych rozwiązań problemu. Następnie należy przypisać „finess" każdemu wygenerowanemu osobnikowi. Jej wartość powinna być wprost proporcjonalna do odległości od optimum. Wybór osobników do reprodukcji powinien odbywać się zgodnie z ich procentowym udziałem w globalnym przystosowaniu populacji. W trakcie reprodukcji dochodzi do crossing-over i mutacji punktowych, które są następnie selekcjonowane ze względu na ich wpływ na „fitness" osobnika. Procesy mutowania, crossing- over i reprodukcji powtarzają się aż do osiągnięcia optimum, którym jest narzucona z góry, sparametryzowana funkcja, którą ma spełniać wygenerowany na drodze ewolucji program *(Goldberg 1995).*

Podsumowując można stwierdzić, że charakterystyczną cechą modeli i symulacji komputerowych jest zjawisko narastania złożoności w czasie stosowania jak najprostszych algorytmów. Prostota modelu pozwala na zadziałanie podstawowych praw statystyki. Komplikowanie modelu przez wprowadzanie szeregu parametrów mających go upodobnić do rzeczywistości odnosi raczej skutki przeciwne do zamierzonych.

Liczbę różnych parametrów, których współcześnie używa się do opisu białka, szacuje się na setki. Trudno rozstrzygnąć, które z nich są najważniejsze. Uwzględnienie wszystkich w symulacji, razem z arbitralnie wybranymi wagami, które należałoby im przypisać, jest praktycznie niemożliwe do zrealizowania. Dlatego w opisywanym modelu ograniczono się do tak ogólnego parametru, jak skład aminokwasowy białka, ale mimo daleko idącego uproszczenia, a właściwie dzięki niemu, otrzymano ciekawe wyniki.

Adresy stron internetowych poświęconych symulacjom komputerowym:

http://www.ics.uci.edu/~eppstein/ca/links.html

http://www.alife.org,

http://paradise.caltech.edu/~cook/Workshop/CAs/2DOutTot/Life/StillLife/StillLifeTheory.html

http://www.frams.poznan.pl

http://www.world-of-dawkins.com/Dawkins/ Work/Software

http://www.cpsc.ucalgary.ca/Research/bmv/software.html

http://www.isd.atr.co.jp/~ray/tierra

http://www.red3d.com/cwr/boids

http://www.vergenet.net/~conrad/boids

mrl.nyu.edu/~dt/alife.html - 3k

http://www.scs.carleton.ca/~csgs/resources/gaal.htm

http://www.solver.com

http://www.cs.ucl.ac.uk/staff/A.Steed/boids.html

SŁOWNICZEK

ATskew: parametr liczbowy określający asymetrię genomu, wyrażany wzorem
ATskew = (A-T)/(A+T)

GCskew: parametr liczbowy określający asymetrię genomu, wyrażany wzorem
GCskew = (G-C)/(G+C)

Asymetria składu nukleotydowego: odstępstwa od zasady *PR2 (*czyli *Parity Rule 2*), stochastycznej reguły równowagi między stężeniami molowymi komplementarnych nukleotydów w obrębie pojedynczej nici DNA) obserwowane np. po podzieleniu nici Watsona lub Cricka na dwie części wyznaczane przez położenie punktu ORI i TER

CAI: *Codon Adaptation Index*, wskaźnik obliczany dla każdego ORFu lub genu kodującego białko określający częstość występowania kodonów w odniesieniu do ich używalności w skali całego genomu, stosowany też do określania prawdopodobieństwa kodowania dla danego ORFu.

Generator liczb losowych: algorytm służący do generowania losowych ciągów liczb (między jego elementami nie powinno być powiązań, a kolejność powinna być przypadkowa)

Inwersja genu: zmiana położenia genu względem kierunku ruchu widełek replikacyjnych, połączona ze zmianą charakteru nici, na której leżał do tej pory gen, a więc z odwróceniem kierunku działania presji mutacyjnej

MCS: *Monte Carlo Step*, symulacyjna jednostka czasu, jedna iteracja, czyli jeden pełny cykl działania algorytmu symulacyjnego

Mutacja cicha: substytucja nukleotydowa w kodonie nie zmieniająca kodowanego aminokwasu, podstawienie synonimiczne

Mutacja neutralna: mutacja akceptowana przez selekcję, nie wywierająca znaczącego wpływu na przystosowanie organizmu

Nić Watsona: jedna z komplementarnych nici podwójnej helisy, jej sekwencja jest zwykle podawana w bazach danych

Nić Cricka: nić komplementarna do nici Watsona, do odtworzenia na podstawie sekwencji obecnych w bazach danych

Nić sensowna : nić odcinka kodującego kolinearna z sekwencją mRNA, nić nie transkrybowana

Nić antysensowna: nić komplementarna do *nici sensownej*, transkrybowana na mRNA

Nić wiodąca: odcinek nici DNA replikowany w sposób ciągły w kierunku zgodnym z kierunkiem przesuwania się widełek replikacyjnych

Nić opóźniająca: nić komplementarna do nici wiodącej, replikowana w kierunku przeciwnym do ruchu widełek replikacyjnych, za pośrednictwem tzw. fragmentów Okazaki

ORF: *Open Reading Frame,* czyli otwrta ramka odczytu, fragment sekwencji DNA rozpoczynający się kodonem start (zwykle ATG) i zakończony jednym z kodonów stop (*Opal* - TGA, *Amber* – TAG lub *Ochre* - TAA), potencjalny gen

ORI**:** *Origin of Replication,* miejsce inicjacji procesu replikacji, zwykle sekwencja o określonym dla danej grupy konsensusie rozpoznawanym przez kompleks polimerazy i towarzyszących jej białek

Presja mutacyjna: tendencja do wstawiania błędnych nukleotydów do nowo syntetyzowanej nici DNA, opisywana tablicą przejść, która określa prawdopodobieństwa 12 możliwych podstawień nukleotydowych.

Presja selekcyjna: eliminacja organizmu, w którym zaszła mutacja obniżająca jego ogólne przystosowanie, warunkowana dążeniem do zachowania funkcji kodowanego białka. W opisywanych symulacjach presja selekcyjna doprowadza do podstawienia genu, który uległ mutacji nie dopuszczalnej przez przyjęty poziom tolerancji.

PR1: *Parity Rule – 1*, zasada Chargaffa, wynikająca z komplementarności zasad azotowych równość A=T, G=C obowiązująca dla całej cząsteczki DNA.

PR2: *Parity Rule – 2*, stochastyczna konsekwencja *PR1*, zakładająca ,że równość A=T i G=C powinna obowiązywać także dla pojedynczej nici DNA.

Replichora: odcinek chromosomu między *ORI* a *TER*, chromosom stanowiący jeden replikon dzieli się na dwie replichory, lewą i prawą, u *Eubacteria* zazwyczaj prawie równej długości.

Sekwencja kodująca: odcinek DNA zawierający informację o białku, transkrybowany na mRNA, a następnie tłumaczony na sekwencję aminokwasową

Sekwencja międzygenowa: niekodujący odcinek DNA, leżący między sekwencjami kodującymi białka

Tablica przejść nukleotydowych: szesnastoelementowa macierz, normalizowana do 1, określająca prawdopodobieństwa zajścia wszystkich możliwych substytucji nukleotydowych, elementy diagonalne dla tzw. przejść tożsamościowych wyznaczają prawdopodobieństwo, że

dany nukleotyd pozostanie niezmieniony, tablica wyznacza charakter presji mutacyjnej działającej na dany genom

Tablica lustrzana: tablica będąca lustrzanym odbiciem tablicy przejść nukleotydowych wyznaczonej dla nici wiodącej, określająca presję działającą na nić komplementarną, czyli opóźniającą

Tablica PAM: *Percent of Accepted Mutation,* macierz przejść aminokwasowych stworzona na podstawie porównania spokrewnionych sekwencji, zawiera informację o akceptowanych przez selekcję substytucjach aminokwasowych, tablica łącząca informację o presji mutacyjnej i selekcyjnej działającej na analizowane genomy

TER: miejsce terminacji replikacji, obszar chromosomu, w którym spotykają się wędrujące w przeciwnych kierunkach widełki replikacyjne

Transwersja: substytucja polegająca na podstawieniu pirymidyny przez purynę lub puryny przez pirymidynę (np. G→C, A→T)

Tranzycja: podstawienie puryny przez inną purynę lub pirymidyny przez pirymidynę (np. A→G, C→T).

LITERATURA

Achaz G., Coissac E., Netter P., Roch E. P. C.(2003). Associations between inverted repeats and the structural evolution of bacterial genomes. Genetics.164: 1279-1289

Adami C. (1995). Learning and complexity in genetic auto-adaptive systems. Physica D 80:154-170

Ausloos M., Mróz I., Pękalski A., Vandewalle N. (1998). Lattice gas model of gradual evolution. Phisica A 273: 75-91

Berthelsen Ch.L., Glazier J.A.,Skolnick M.H. (1992) Global fractal dimension of human DNA sequences treated as pseudorandom walks. Phys. Rev. A. 45: 8902-8913

Blalock J. E., Smith E.M. (1984). Hydroparhic anti-complementarity of amino acids based on the genetic code. Biochem. Biophys. Res. Comm. 121: 203-207

Blattner F.R., Plunkett G. 3rd, Bloch C.A., Perna N., Burland V., Riley M.,Collado-Vides J., Glasner J.D., Rode C.K., Mayhew G.F.(1997). The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453-1462

Brewer B.J.(1988). When polymerase collide: Replication and the transcriptional organisation of the E. Coli chromosome. Cell 53: 679-686

Casjens S. (1998). The divers and dynamic structure of bacterial genomes. Annu. Rev. Genet. 32: 339-377

Cebrat S., Dudek M. R.(1998). The effect of DNA phase structure on DNA walks. Eur. Phys. J.3: 271-276

Cebrat S., Dudek M. R., Rogowska A. (1997). Asymmetry in nucleotide composition of senese and antisense strands as a parameter for discriminating open reading frames as protein sequences. J. Appl. Genet. 38: 1-9

Chargaff E. (1950). Chemical specifity of nucleic acids and mechanism of their enzymatic degradation. Experientia 6: 201-240

Coveney P., Highfield R. (1997). Granice złożoności. Poszukiwania porządku w chaotycznym świecie. Prószyński i s-ka. Warszawa

Crick F.H.C. (1957). The structure of nucleic acids and their role in protein synthesis. Biochem.Soc. Symp. 14: 25-26

Crick F.H.C. (1968). The origin of genetic code. J.mol. Biol. 38: 367-379

Dawkins R. (1996). Samolubny gen . Prószyński i s-ka. Warszawa

Dayhoff M. O., Eck R. V., Park C. M. (1972),. Atlas of protein sequence and structure.5:75-84 NBRF

Dayhoff M. O., Schwatrz R. M., Orcutt B. C. (1978). Atlas of protein structure and structure 5:345-352. National Biochemical Research Foundation, Washington DC

Dudek M.R. (1992). Liczby losowe i Monte Carlo. Mikroklan 2: 30-34

Dudkiewicz M., Mackiewicz P., Nowicka A., Kowalczuk M., Mackiewicz D., Polak N., Smolarczyk K., Dudek M. R., Cebrat S. (2003). Properities of the genetic code under directional, asymmetric mutational pressure. In: P. M. A. Sloot et al. (Eds.), Proc. ICCS 2003 Lect. Notes. Comput. Sc. 2657: 343-350

Dudkiewicz M., Mackiewicz P., Nowicka A., Kowalczuk M., Mackiewicz D., Polak N., Smolarczyk K., Banaszak J., Dudek M. R., Cebrat S. (2004a). Correspondance between mutation and selection pressure and the genetic code degeneracy In the gene evolution. Future Gener. Comput. Sc. *(in press)*

Dudkiewicz M., Mackiewicz P., Nowicka A., Kowalczuk M., Mackiewicz D., Polak N., Smolarczyk K., Banaszak J., Dudek M. R., Cebrat S. (2004b).Simulation of gene evolution under directional mutational pressure. Physica A 336: 63-73

Eisen J. A., Heidelberg J. F., Whitte O., Salzberg S. L. (2000). Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol. 1: 11.1-11.9

Fijałkowska I.J., Jonczyk P., Maliszewska – Tkaczyk M., Bialoskorska M., Schaaper R.M. (1998). Unequal fidelity of leading and lagging strand DNA replication on the Escherichia coli chromosome. Proc. Natl. Acad.Sci. USA 95: 10020-10025

Fijałkowska I.J., Schaaper R.M. (1996). Mutants in the Exo I motif of Escherichia coli dnaQ: defective proofreading and inviability due to error catastrophe. Proc. Natl. Acad. Sci USA 93:2856-2861

Fitch W.M. (1966). The relation between frequencies of amino acids and ordered trinucleotides. J. Mol. Bil. 16: 1-8

Francino M.P., Chao L., Riley M. A., Ochman H. (1996). Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science 272:107-109

Francino M.P., Ochman H. (1997). Strand asymmetries in DNA evolution. Trends Genet. 13: 240-245

Francino M.P., Ochman H.(2000). Strand symmetry around the beta-globin origin of replication in primates. Mol. Biol. Evol.17: 416-422

Frank A.C., Lobry J.R.(1999). Asymmetric substitution patterns: review of possible underlying mutational or selective mechanisms. Gene 238: 65-77

Fraser C.M., Casjens S., Huang W.M., Sutton G.G., Clayton R., Lathigra R., White O., Ketchum K.A. i współpr.(1997). Genomic sequence of Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390: 580-586

Fraser C.M., Gocayne J.D., White O., Adams m.D., Clayton R.A., Fleischmann R.D., Bult C.J., Kerlavage A.R., Sutton G.G., Kelley J.M. i współpr.(1995). The minimal gene complement of *Mycoplasma genitalium.* Science 270: 397-403

Frederico L.A., Kunkel T.A., Shaw B.R. (1990). A sensitive genetic assay for the detection of cytosine deamination of rare constants and the activation energy. Biochemistry 29: 2532-2537

Gilis D., Massar S., Cerf N.J., Rooman M. (2001). Optimality of the genetic code with respect to protein stability and amino acid frequencies. GenomeBiol. 2(11): 49.1-49.12

Gojobori T., Li W-H, Graur D. (1982). Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. 18: 360-369

Goldberg D. (1988). Genetic Algorithms. Addison –Weseley, Washington DC

Gouy M., Gautier C. (1982). Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10: 7055-7074

Gu X., Li W. H. (1998). Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. Proc.natl. Acad. Sci. USA 95: 5899-5905

Gutowitz H. (1991). Cellular Automata: theory and experiment. ISBN 0-262-57086-6

Hanawalt P.C. (1991). Heterogenity of DNA repair at the gene level. Mutat. Res. 247: 203-211

Himmelreich R., Plagens H., Hilbert H., Reiner B., Herrmann R. (1997). Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. Nucleic Acids. Res. 25: 701-712

Holland J. (1975). Adaptation in Natural and Artificial Systems. Ann Arbor (red).University of Michigan Press

Hughes D. (2000). Co-evolution of the *tuf* gene conversion with the gneration of chromosomal inversions. J. Mol. Biol. 297: 355-364

Hutchinson F. (1996). Mutagenesis. In: Neidhardt F.C. (red) *Escherichia coli* and *Salmonella*. Cellular and molecular biology. Asm.Press, Washington D.C.: 749-763

Iwaki T., Kawamura A., Ishino Y., Kohno K., Goshima N., Yara M., Furusawa M., Doi H., Imamoto F. (1996). Preferential replication – dependent mutagenesis in the lagging strand in the *Escherichia coli*. Mol. Gen. Genet. 251: 657-664

Jayaram B. (1997). Beyond the wobble: the rule of conjugates. J. Mol. Evol. 45: 704-705

Jeanmougin F., Thompson J.D., Gouy M., Higgins D.G., Gibson T.J. (1988). Multiple sequence alignement with Clustal X. Trends. Biochem. Sci. 23: 403-405

Jukes T. H., Cantor C. (1969). Mammalian Protein Metabolism, chapter Evolution of protein molecules. Academic Press, New York: 21-132.

Kimura M. (1968). Evolutrionary rate at the molecular level. Nature 217: 624-626

Kimura M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Mol. Evol. 16:111-120

King J.L., Jukes T.H. (1969). Non-Darwinian evolution. Science 164: 788-797

Knight R.D. (2001). The origin and evolution of the genetic code: statistical and experimental investigations. PhD Dissertation Princeton Univ.

Knight R.D., Landweber L.F. (1998). Rhyme or reason : RNA –arginine interactions and the genetic code. Chem.Biol. 5(9): R215-220

Kowalczuk M., Mackiewicz P., Mackiewicz D., Nowicka A., Dudkiewicz M., Dudek M.R., Cebrat S. (2001a). DNA asymmetry and the replicational mutational pressure. J. Appl. Genet. 42(4): 553-577

Kowalczuk M., Mackiewicz P., Mackiewicz D., Nowicka A., Dudkiewicz M., Dudek M.R., Cebrat S. (2001b). High correlation between the turnover of nucleotides under mutational pressure and DNA composition. BMC Evolutionary Biology 1: 13

Lafay B., Lloyd A.T., McLean M.J., Devine K.M., Sharp P.M., Wolfe K.H. (1999). Proteome composition and codon usage in spirochaetes: species–specific and DNA strand-specific mutational biases. Nucleic Acids Res. 27:1642-1649

Lagerkvist U. (1978). Two out of three. An alternative method for codon reading. Proc. Natl. Acad. Sci. USA 75: 1759-1762

Lagerkvist U. (1980). Codon missreading: a restriction operative in the evolution of the genetic code. Amwer. Scient. 68: 192-198

Langton C. (1989). Artificial Life. Proceedings of an interdisciplinary workshop on the synthesis and simulations of living systems. Addison-Wesley (red). Redwood City

Liu S. L., Sanderson K. E. (1996) Highly plastic chromosomal organisation in Salmonella typhi. Proc. Natl. Acad. Sci. USA. 93:10301-10308

Lobry J.R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13: 660-665

Lobry J.R.(1995). Properities of a general model of DNA evolution under no-strand bias conditions. J. Mol. Evol. 13: 660-665

Mackiewicz P., Gierlik A., Kowalczuk M., Dudek M.R., Cebrat S. (1999). Asymmetry of nucleotide composition of prokaryotic chromosomes. J. Appl. Genet. 40: 1-14

Mackiewicz P., Mackiewicz D., Gierlik A., Kowalczuk M., Nowicka A. (2001a). The differential killing of genes by inversions in prokaryotic genomes. J. Mol. Evol. 53: 615-621

Mackiewicz P., Szczepanik D., Kowalczuk M., Cebrat S. (2001b). Flip-flop around the origin and trerminus of replication in prokaryotic genom es. Gen. Biol. 2(12): interactions 1004.1-1004.4

Mackiewicz D., Mackiewicz P., Kowalczuk M., Dudkiewicz M., Dudek M. R., Cebrat S. (2003). Rearrangements between differently replicating DNA strands in asymmetric bacterial genomes. A. Mic. Pol. 52: 245-261

McLean M. J., Wolfe K. H., Devine K. M. (1998). Base composition skews, replication orientation and gene orientation in 12 prokaryote genom es. J. Mol. Evol. 47: 691-696

McLean M.J., Wolfe K.H., Devine K.M. (1998). Base composition skews replication orientation, and gene orientation in 12 prokaryote genom es. J.mol. Evol. 47: 691-696

Metropolis N., Ulam S.(1949). The Monte Carlo method. J. Amer.Stat. Assoc 44: 335-341

Mrazek J., Karlin S. (19980. Strand compositional asymmetry in bacterial and large virial genom es. Proc. Natl. Acad. Sci. USA 95: 3720-3725

Mushegian A.R., Koonin E.V. (1996). Gene order is not conserved in bacterial evolution, Trends Genet. 12: 289-290

Nowicka A., Mackiewicz P., Dudkiewicz M., Mackiewicz D., Kowalczuk M., Cebrat S., Dudek M.R. (2003). Correlation between mutation pressure, selection pressure. In: P. M. A. Sloot et al. (Eds.), Proc. ICCS 2003, Lect. Notes Comput. Sc. 2658: 650-657

Ochman H., Lawrence J.G., Groisman E. A. (2000). Lateral gene transfer and yhe nature of bacterial innovations. Nature 405:299-304

Osawa S. (1995). Evolution of the genetic code. Oxford, Oxford Univ. Press.

Pearson W.R., Lipman D.J.(1988). Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci USA 85: 2444-2448

Picardeau M., Lobry J.R., Hinnebusch B.J. (1999). Physical mapping of an origin of bidirectional replication at yhe center of the Borrelia burgdorferi linear chromosome. Mol. Microbiol. 32: 437-445

Prusinkiewicz P., Lindenmayer A. (1990). The algorithmic beauty of plants. Springer – Verlag. Berlin

Radman M. (1998). DNA replication: one strand may be more equal. Proc. Natl. Acad. Sci. USA 95: 9718-9719

Ray T.S. (1992). Artificial Life II. Proceedings Volume in the Santa Fe Institute in the Sciences of Complexity. Vol. 10: 317

Risler J. L., Delorne M. O., Delacroix H., Henaut A. (1988). Amino acid substitutions in structurally related proteins: a pattern recognition approach. J. Mol. Biol. 204: 1019-1029

Roberts J.D., Izuta S., Thomas D.C., Kunkel T.A. (1994). Mispair -, site, and strand specific error rates during simian virus 40 origin-dependent replication in vitro with excess deoxythimidine triophosphate. J. Biol. Chem. 269: 1711-1717

Rocha E. (2003). DNA repeats lead to the accelerated loss of gene order in bacteria. Trends Genet. 11(19):600-603

Rocha E. P. C., Danchine A. (2001). Ongoing evolution of strand composition in bacterial genomes. Mol. Biol. Evol. 18(9):1789-1799

Rocha E.P.C., Danchin A. (2003). Gene essentiality determines chromosome organisation in bacteria. Nucleic Acids Res**. 22(31)**:6570-6577

Romero D., Palacios R. (1997). Gene amplification and genomic plasticity in prokaryotes. Annu. Rev. Genet. 31: 91-111

Roth J. R., Benson N., Galitski T., Haack K., Lawrence G. (1996). Rearrangements of the bacterial chromosome: formation and application. *Escherichia coli* and *Salmonella* Cellular and Molecular Biology ASM Press, Washington, DC: 2256-2276

Schrödinger E. (1945). What Is Life ? Cambridge, Cambridge Univ. Press

Schwartz R. M., Dayhoff M. O. (1978). The point mutation process in proteins. Origin of Life: Proceedings of the Second ISOL Meeting (Tokyo Japan):457-469

Sharp P.M., Li W-H. (1987). The codon adaptation index: a measure of directional synonymous codon usage bias and its potential applications. Nucleic Acids Res. 15: 1281-1295

Shepherd J.C.(1981). Periodic correlation in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code. J.Mol. Evol. 17: 94-102

Sjöström M., Wold S. (1985). A multivariate study of the relationship between the genetic code and the physical-chemical properities of amino acids. J. Mol. Evol. 22: 272-277

Smith G. R.(1988). Homologous recombination in prokaryotes. Microbiol. Rev.52: 1-28

Sonneborn T.M. (1965). Degeneracy of the genetic code : extent, nature and genetic implications. In: Evolving Genes and Proteins. V. Bryson and H.J. Vogel. (red.) New York, Academic Press: 377-297

Sueoka N. (1995). Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J. Mol. Evol. 40: 318-325

Suyama M., Bork P. (2001). Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet. 17:10-13

Szathmàry E.(1991). Codon swapping as a possible evolutionary mechanism. J.Mol. Evol. 32: 178-182

Szczepanik D., Mackiewicz P., Kowalczuk M., Gierlik A., Nowicka A., Dudek M.R., Cebrat S. (2001). Evolution rates of genes on leading and lagging DNA strands. J.Mol. Evol. 9: 814-825

Thomas D.C., Svoboda D.L., Vos J.M., Kunkel T.A. (1996). Strand specifity of mutagenic bypass replication of DNA containing psoralen monoadducts in a human cell extract. Mol. Cell. Biol. 16: 2537-2544

Tillier E.R.M., Collins R.A. (2000a). Genome rearrangement by replication-directed translocation. Nat. Genet. 26: 195-197

Tillier E.R.M., Collins R.A. (2000b). The contributions of replication orienbtation, gene direction, and signal sequences to basr-composition asymmetries in bacterial genomes. J.Mol. Evol. 50: 249-257

Volkenstein M.V. (1966). The genetic coding of protein structure. Biochim. Biophys. Acta 119: 421-424

Woese C.R. (1965a). On the evolution of the genetic code. Proc. Natl. Acad. Sci. USA 54: 1546-1552

Woese C.R. (1965b). Order In the genetic code. Proc. Natl. Acad. Sci. USA 54" 71-75

Woese C.R. (1967). The genetic code. The molecular basis for genetic expression. Harper & Row (red.). New York

Woese C.R., Dugre W.C., Dugre S.A., Kondo M. I współpr. (1966). On the fundamental nature and evolution of the genetic code. Cold Spring Harb. Symp. Quant.Biol. 31: 723-736

Wolfe K.H., Sharp P.M., Li W.-H. (1989) Mutation rates differ among regions of the mammalian genom. Nature 337: 283-285

Wolfram S. (1986). Theory and applications of cellular automata. ISBN 9971-50-124-4

Wong J.T., Cedergren R. (1986). Natural selection versus primitive gene structure as determinant of codon usage. Eur. J. Biochem. 159: 175-180

Wong J.T.F. (1983). Membership mutation of the genetic code: los sof fitness by tryptophan. Proc. Natl. Acad. Sci. USA 80: 6303-6306

Wong J.T.F.(1976). The evolution a universal genetic code. Proc. Natl. Acad. Sci. USA 73: 2336-2340

Wu N. S., Maeda K.(1987). Inequality in mutation rates of the two strands of DNA. Nature 327: 169-170

Ycas M. (1969). The biological code. Amsterdam , North Holland Publishing Company

Zhang C.T., Zhang R. (1991). Analysis of distribution of base in codon in the coding sequences by diagrammatic technique. Nucleic Acids Res. 19: 6313-6317

Zivanovic Y., Lopez P., Phillippe H., Forterre P. (2002). Pyroccocus genome comparison evidences chromosome shuffling- driven evolution. Nucleic Acids Res. 30: 1902-1910

Zuckerkandl E., Pauling L. (1965). Evolutionary divergence and convergence in proteins.In: Evolving Genes and Proteins. V.Bryson & J.Vogel (red.). New York, Academic Press