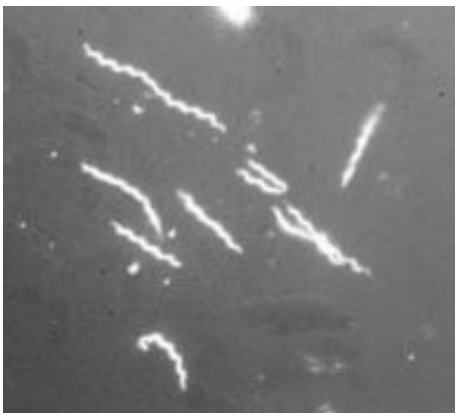
UNIWERSYTET WROCŁAWSKI
Wydział Nauk Przyrodniczych

Maria Kowalczuk

Presja mutacyjna i selekcyjna w genomie *Borrelia burgdorferi*

Mutational and selection pressures in the *Borrelia burgdorferi* genome



Praca doktorska
wykonana w Zakładzie Genetyki
Instytutu Mikrobiologii
pod kierunkiem
prof. dr hab. Stanisława Cebrata

Wrocław 2002

Składam serdeczne podziękowania
Profesorowi Stanisławowi Cebratowi
za życzliwą pomoc na każdym etapie powstawania tej pracy,
za stworzenie mi szansy rozwoju w SmORFlandzie,
i za wiarę w moje możliwości

Profesorowi Mirosławowi Dudkowi
za wszystkie generacje programów komputerowych,
które umożliwiły uzyskanie wyników tej pracy,
i za cierpliwość w uczeniu mnie linuxa

Pawłowi Mackiewiczowi,
Dorotce Mackiewicz, Agnieszce Gierlik, Oli Nowickiej, Gosi Dudkiewicz,
oraz wszystkim członkom zespołu SmORFlandu
za wspólną pracę w niepowtarzalnej atmosferze

Autoreferat

Borrelia burgdorferi

Borrelia jest gram-ujemnym krętkiem posiadającym 7-11 peryplazmatycznych rzęsek. Długość komórki waha się od 10 do 30 μm, szerokość od 0,2 do 0,5 μm (BARBOUR, HAYES 1986). *Borrelia burgdorferi* jest pasożytem wewnątrzkomórkowym kleszczy z rodzaju *Ixodes*, dla człowieka jest czynnikiem etiologicznym odkleszczowej choroby z Lyme (BURGDORFER i współpr. 1982). Analiza molekularna ujawniła, że *Borrelia burgdorferi* to grupa blisko spokrewnionych gatunków, określana obecnie mianem *Borrelia burgdorferi* sensu lato, do której zalicza się *B. afzelii* (BARANTON et al. 1992, CANICA et al. 1993), *B. garinii* (BARANTON et al. 1992), *B. japonica* (KAWABATA et al. 1993), *B. andersonii* (MARCONI et al. 1995), *B. tanukii* i *B. turdi* (FUKUNAGA et al. 1996), *B. valaisiana* (WANG et al. 1997), *B. lusitaniae* (LE FLECHE et al. 1997), *B. bissettii* (POSTIC et al. 1998), oraz oczywiście *B. burgdorferi* sensu stricto (BARANTON et al. 1992). Do tej pory całkowicie zsekwencjonowano tylko jeden szczep, *B. burgdorferi* sensu stricto B31, który jest przedmiotem niniejszej pracy.

Na genom *B. burgdorferi* B31 składa się liniowy chromosom o długości 910,725 par zasad (FRASER i współpr. 1997) oraz 12 liniowych i 9 kolistych plazmidów o łącznej długości 610,694 par zasad (CASJENS i współpr. 2000). 93% chromosomu stanowią 853 sekwencje kodujące, funkcje ok. 60% z nich zostały poznane. Miejsce inicjacji dwukierunkowej replikacji znajduje się pośrodku chromosomu, co potwierdza jego genetyczna organizacja (OLD i współpr. 1993) i GC skew (FRASER i współpr. 1997), natomiast eksperymentalnie potwierdzili to PICARDEAU i współpr. (1999). Taki sposób replikacji decyduje o tym, że połowa nici Watsona jest replikowana w sposób ciągły, jako nić wiodąca, a połowa jest syntetyzowana z fragmentów Okazaki jako nić opóźniająca. Komplementarne połowy nici Cricka są replikowane odpowiednio jako nić opóźniająca i wiodąca (patrz schemat organizacji genomu, str. 15). Pomimo, że obie nici są identycznej długości, na nici wiodącej znajduje się sensowna nić 564 ORFów, natomiast na opóźniającej – 286.

Asymetria DNA

Do analizy wybrano chromosom *B. burgdorferi*, ponieważ charakteryzuje go największa asymetria w składzie nukleotydowym DNA spośród dotychczas zsekwencjonowanych genomów bakteryjnych. Czym jest asymetria DNA? W podwójnej helisie DNA obowiązuje zasada komplementarności: liczba adenin w cząsteczce jest taka sama jak liczba tymin, natomiast liczba guanin odpowiada liczbie cytozyn, co jest spełnieniem reguł CHARGAFFa (1950). Gdyby cząsteczka była symetryczna, to te same reguły obowiązywałyby również w obrębie jednej nici. Jest to tak zwana druga reguła parowania (LOBRY 1995). Tak więc, gdyby wymieszać nukleotydy chromosomu *B. burgdorferi* i utworzyć z nich nową cząsteczkę w sposób losowy, nić wiodąca nie różniłaby się statystycznie od opóźniającej (Tabela 1, str. 19). Asymetria DNA definiowana jest jako odchylenie od równości [A]=[T] i [G]=[C] w pojedynczej nici. Jednym z mechanizmów generujących asymetrię w genomach jest presja mutacyjna związana z replikacją. Połówki nici Watsona prawdziwego chromosomu *B. burgdorferi* są replikowane w różny sposób i różnią się znacząco składem nukleotydowym (Tabela 1). Nić wiodąca jest bogata w tyminę i guaninę, natomiast opóźniająca – w adeninę i cytozynę.

Mechanizmy generujące asymetrię

Substytucja nukleotydu na jednej nici DNA prowadzi również do zmiany na nici komplementarnej. Jednak, aby zrozumieć pochodzenie i znaczenie asymetrii, należy określić gdzie i w jaki sposób zachodzą pierwotne zmiany, które prowadzą do różnych substytucji w różnych regionach chromosomu.

Mechanizmy, które mogą wprowadzać asymetrię do nici DNA, były wielokrotnie dyskutowane (prace przeglądowe: FRANCINO, OCHMAN 1997, MRAZEK, KARLIN 1998, FRANK, LOBRY 1999, KARLIN 1999, TILLIER, COLLINS 2000a, KOWALCZUK et al. 2001a). Skład nukleotydowy sekwencji jest kształtowany przez dwie różne, a czasami przeciwne siły: presję mutacyjną i selekcyjną. Presja mutacyjna to specyficzne substytucje wprowadzane do DNA podczas replikacji i transkrypcji. Presja selekcyjna natomiast jest związana głównie z funkcją kodowanych białek, a więc składem kodonowym genów i ich położeniem na chromosomie, oraz składem sekwencji

kontrolujących. Aby zobaczyć działanie wyłącznie presji mutacyjnej, należałoby znaleźć w genomie sekwencje nie podlegające selekcji.

Sekwencja DNA znajduje się w stanie równowagi z presją mutacyjną, gdy ogólny skład nukleotydowy odpowiada częstościom substytucji i nie podlega on dalszym zmianom ilościowym. Natomiast, kiedy sekwencja kodująca zaadaptuje się do panującej presji selekcyjnej, jej skład nie zmienia się i znajduje się ona w stanie równowagi dynamicznej, ale odchylonym od stanu sekwencji nie podlegających selekcji. Sparametryzowana odległość między daną sekwencją a sekwencją w stanie równowagi jest miarą presji selekcyjnej.

Obrazowanie i pomiar asymetrii - spacery DNA

Aby zobrazować asymetrię, daną sekwencję dzieli się zwykle na odcinki (zwane oknami), w których zlicza się badane nukleotydy lub różnice w ich liczbie i wynik przedstawia na wykresie w skali chromosomu. Rys. 1a (str. 37) przestawia wartości różnic [G]-[C] dla kolejnych odcinków nici Watsona. Wartości te są przeważnie ujemne dla części opóźniającej i przeważnie dodatnie dla części wiodącej. Obszar zmiany trendu pokrywa się z położeniem punktu inicjacji replikacji (ori), ale silne fluktuacje utrudniają analizę (MRAZEK, KARLIN 1998). Aby uzyskać bardziej czytelny obraz, można zwiększyć rozmiar okna (Rys. 1b) lub analizować sekwencję oknami zachodzącymi (Rys. 1c, LOBRY 1996a). Jednak, jeśli okna są zbyt duże, nie można precyzyjnie określić miejsca zmiany trendu (MCLEAN et al. 1998). Tych problemów można uniknąć, jeśli wartości różnic [G]-[C] zostaną skumulowane (Rys. 2a). Kumulatywne diagramy stosowali GRIGORIEV (1998), FREEMAN i współpr. (1998) oraz TILLIER i COLLINS (2000a). Najlepsze wyniki analizy daje jednak metoda spacerów DNA opracowana przez CEBRATA i DUDKA (1998) oraz rozwinięta przez MACKIEWICZA i współpr. (1999a,b). Spacerowicz analizuje sekwencję odcinkami lub nukleotyd po nukleotydzie. Aby uwypuklić lokalne trendy, przeskalowano wartości różnic [G]-[C] tak, aby wykres, (czyli spacer po sekwencji nici Watsona) kończył się w punkcie y=0 (Rys. 2b), oraz podzielono wartości przez długość analizowanej sekwencji, aby je znormalizować i umożliwić porównanie asymetrii sekwencji o różnej długości (Rys. 2c).

Transformacje spacerów – rozdzielenie różnych rodzajów asymetrii

Dodawanie i odejmowanie spacerów pozwala na rozdzielenie asymetrii wynikającej z działania różnych procesów. Procesy te można podzielić na dwie główne grupy: takie, które jednakowo oddziałują na nici Watsona i Cricka (asymetria ma taki sam znak na obu niciach), oraz takie, które działają przeciwnie (asymetria ma na jednej nici znak dodatni, na drugiej ujemny). Jeśli odejmie się wartości spacerów dla nici Watsona i Cricka, asymetria o takim samym znaku zniknie, a o znakach przeciwnych – zostanie uwypuklona. Ten typ asymetrii wprowadzany jest przez procesy związane z replikacją, i różnicuje on nić wiodącą i opóźniającą. Spacery DNA pozwalają na analizę zarówno różnic [G]-[C] i [A]-[T] (Rys. 3a), jak i asymetrii w rozkładzie poszczególnych nukleotydów (Rys. 3b). Dodanie tych samych spacerów spowoduje zanik asymetrii wynikającej z replikacji, i ujawnienie się asymetrii związanej z transkrypcją genów i kodowaniem białek, które nie zależą od sposobu replikacji nici. Ten typ asymetrii nie występuje w chromosomie *B. burgdorferi* (Rys. 4). Trendy widoczne na wykresach są statystycznie nieistotne.

Tego typu spacery można wykonać nie tylko dla sekwencji całego chromosomu, ale także dla sekwencji sklejonych ORFów, osobno dla każdej pozycji w kodonie, oraz dla sekwencji międzygenowych (leżących pomiędzy ORFami). Odejmowanie spacerów (Rys. 6-7) pokazuje obecność asymetrii związanej z replikacją w każdej pozycji w kodonie oraz w sekwencjach międzygenowych, natomiast dodawanie (Rys. 8-9) ujawnia brak wpływu transkrypcji i procesów związanych z kodowaniem na asymetrię. Asymetria obserwowana w genomie *B. burgdorferi* jest generowana wyłącznie przez procesy związane z replikacją (MACKIEWICZ i współpr. 1999c), a nie z transkrypcją, jak proponowali niektórzy autorzy (BELETSKII, BHAGWAT 1996, FRANCINO i współpr. 1996, FRANCINO, OCHMAN 1997, FREEMAN i współpr. 1998). Największą asymetrię obserwuje się w trzecich pozycjach w kodonie. Jest to zrozumiałe, ponieważ w większości przypadków tranzycja w tej pozycji nie powoduje zmiany sensu kodowanego aminokwasu, natomiast transwersja, która jest o wiele mniej prawdopodobna, zmienia sens kodonu jedynie w połowie przypadków (patrz tablica kodu genetycznego, str. 17). Asymetria puryn i pirymidyn w trzecich pozycjach kodonów czterokrotnie zdegenerowanych jest trzykrotnie większa niż w trzecich pozycjach kodonów dwukrotnie zdegenerowanych (Rys. 15). Ciekawe jest to, że asymetria w trzecich pozycjach jest większa nawet niż w sekwencjach międzygenowych. Prawdopodobnie

sekwencje międzygenowe łatwiej mogą przenosić się pomiędzy nicią wiodącą a opóźniającą niż geny. 96% genów białek rybosomalnych położonych jest na nici wiodącej, co wskazuje na nieprzypadkowe rozmieszczenie genów na chromosomie. Asymetrię stwierdza się także w pierwszych i drugich pozycjach. Jeśli różne typy substytucji zachodzą na niciach wiodących i opóźniających w pierwszych i drugich pozycjach w kodonie, musi prowadzić to do asymetrycznych podstawień aminokwasów w białkach i do asymetrii w położeniu kodonów różnych aminokwasów na chromosomie (Rys. 16).

Spacery typu „pajączek"

Innym rodzajem spaceru, przydatnym w analizie asymetrii, jest tak zwany „pajączek". W tej metodzie mierzy się jednocześnie asymetrię w rozkładzie wszystkich czterech nukleotydów. Sekwencja zostaje podzielona na niewielkie okna lub jest analizowana nukleotyd po nukleotydzie. Jeśli w badanym oknie przeważa guanina, „nóżka pajączka", czyli wykres kieruje się w górę, jeśli cytozyna – w dół, jeśli adenina – w prawo, jeśli tymina – w lewo. Wykresy dla pierwszej i drugiej pozycji w ORFach leżących na niciach wiodących i opóźniających wykazują wspólne cechy (Rys. 10). Pierwsze pozycje w kodonach bogate są w guaninę i adeninę, dlatego wykresy spacerów znajdują się w pierwszej ćwiartce układu współrzędnych. W drugich pozycjach przeważa adenina i cytozyna, dlatego wykresy położone są w czwartej ćwiartce. Te preferencje są uniwersalne i związane z kodowaniem białek (WONG, CEDERGREN 1986, ZHANG, ZHANG 1991, GUTIERREZ et al. 1996, CEBRAT et al. 1997b, 1998, MRAZEK, KARLIN 1998, MCLEAN et al. 1998, WANG 1998). Trzecie pozycje w kodonach wykazują przeciwne trendy na niciach wiodących i opóźniających. Trendy te są zgodne z preferencjami obserwowanymi w całej sekwencji i w sekwencjach międzygenowych: nić wiodąca jest bogata w guaninę i tyminę, opóźniająca – w adeninę i cytozynę. Sekwencja międzygenowa odczytana z nici wiodącej nie wykazuje struktury tripletowej, w każdej „pozycji" widać taki sam trend, typowy dla nici wiodącej (Rys. 11a). Jednak w genomie można także odnaleźć sekwencje międzygenowe pochodzące ze zduplikowanych genów, które zachowały jeszcze ślady swojej przeszłości. Sekwencja na Rys. 11b posiada strukturę tripletową i można określić fazę (ramkę odczytu), w której niegdyś kodowała ona białko, jednak wszystkie „nóżki" odchylone są już w kierunku typowym dla sekwencji międzygenowych.

„Pajączek” może także posłużyć do analizy sekwencji sklejonych ORFów z nici wiodącej i opóźniającej (Rys. 12). W celu normalizacji, wartości spacerów zostały podzielone przez długość odpowiednich sekwencji, aby porównać asymetrię sekwencji o różnej długości. Widać w nich takie same trendy jak w pojedynczych ORFach, tylko bardziej wyraźne. Sekwencje międzygenowe zostały odczytane dwukrotnie, z nici wiodącej i opóźniającej, aby pokazać ich przeciwne trendy. Ich wykresy są idealnie symetryczne. Należy zwrócić uwagę, że trzecie pozycje w kodonach ORFów z nici wiodącej, to inne sekwencje niż trzecie pozycje w kodonach ORFów z nici opóźniającej, a jednak one również wykazują lustrzaną asymetrię. Wskazuje to na silny wpływ presji mutacyjnej na ich skład nukleotydowy.

Długie „nóżki” świadczą o silnych trendach, czyli preferencjach w składzie nukleotydowym sekwencji, natomiast wykresy krótkie i poplątane świadczą o bardziej stochastycznym układzie nukleotydów w sekwencji. Parametry wykresów „pajączków”, takie jak długość wektora „nóżki” i kąt jego nachylenia do osi x, mogą być wykorzystane do dalszej analizy asymetrii (CEBRAT et al. 1997b, CEBRAT et al. 1998, KOWALCZUK et al. 1999a). Kąty nachylenia spaceru można przedstawić na płaszczyźnie o skończonej powierzchni, która jest w rzeczywistości powierzchnią torusa (Rys. 13). Na wykresie przedstawiającym kąty nachylenia spacerów po pierwszych pozycjach do kątów spacerów po trzecich pozycjach, ORFy z nici wiodącej i opóźniającej tworzą dwie odrębne, nie nachodzące na siebie grupy. Skład nukleotydowy ORFu zdradza jego położenie na nici, a nawet skład aminokwasowy białka pozwala określić gdzie leży kodujący je gen. Te cechy, oraz znaczna asymetria związana z replikacją, widoczna w każdej pozycji w kodonach a zwłaszcza w trzecich pozycjach, wskazują na to, że geny *B. burgdorferi* od dawna nie ulegały translokacjom na nić przeciwną i ich skład nukleotydowy dostosował się do położenia na chromosomie.

Określenie presji mutacyjnej – konstrukcja tablicy przejść mutacyjnych

Analiza spacerów DNA nie może ujawnić, jakie substytucje generują obserwowaną asymetrię, czyli jaka jest presja mutacyjna na sekwencję. Asymetryczny skład DNA może być uzyskany przez nieskończoną liczbę kombinacji częstości 12 typów substytucji. Co więcej, skład sekwencji zależy nie tylko od presji mutacyjnej, ale także od selekcji, która bezlitośnie eliminuje wszystkie niekorzystne substytucje. Jednak sekwencje międzygenowe pochodzące ze zduplikowanych genów powinny

akumulować wszystkie substytucje. Ich porównanie z macierzystymi genami ujawnia częstości wszystkich przejść mutacyjnych (KOWALCZUK i współpr. 2001a). Aby odnaleźć sekwencje wolne od selekcji, odcinki międzygenowe dłuższe niż 90 nukleotydów przetłumaczono na sekwencje aminokwasowe we wszystkich sześciu możliwych fazach (ramkach) odczytu. Kodony stop: amber i ochre zamieniono na tyrozynę, a opal – na tryptofan, gdyż tylko jedna substytucja w trzeciej pozycji tych kodonów wystarczy do powstania stopu. Następnie przeszukano białkową bazę danych *B. burgdorferi* programem FASTA (PEARSON, LIPMAN 1988). Ponieważ pseudogeny w przestrzeni międzygenowej kumulują wszystkie mutacje, zastosowano liberalne kryteria homologii, czyli wartość $E < 0,05$. Znaleziono około 30 homologów do sekwencji międzygenowych odczytanych z nici wiodącej. Sekwencje nukleotydowe homologicznych ORFów i sekwencji międzygenowych dopasowano przy pomocy programu CLUSTAL X (JEANMOUGIN et al. 1988). Suma długości pasujących sekwencji nukleotydowych wyniosła 3737 nukleotydów. Aby określić kierunek przejść mutacyjnych przyjęto, że wszystkie obserwowane różnice wynikają z substytucji zakumulowanych w sekwencjach międzygenowych. To założenie nie jest całkiem prawdziwe, ale jest niezbędne do konstrukcji tablicy, i było wykorzystywane z powodzeniem przez innych autorów (LI et al. 1984, YANG 1994). Tablicę substytucji skonstruowano według metody GOJOBORI'ego i współpr. (1982) oraz FRANCINO i OCHMANa (2000) (Tabela 2, str. 51). Ponieważ częstości substytucji są inne dla każdego nukleotydu, wprowadzono poprawki na wielokrotne substytucje i rewersje dla każdego nukleotydu oddzielnie zamiast jednej ogólnej poprawki KIMURY (1980). Jest to tablica empiryczna, w przeciwieństwie do tablic (modeli) parametrycznych, często wykorzystywanych w badaniach filogenetycznych (przeglądu najnowszej literatury na temat modeli empirycznych i parametrycznych dokonali LIÒ i GOLDMAN 1998, oraz WHELAN i współpr. 2001).

Testowanie tablicy

Uzyskana tablica przejść była następnie testowana za pomocą symulacji komputerowych. Pozwoliło to na zbadanie, jak zmienia się skład nukleotydowy sekwencji pod wpływem presji mutacyjnej opisanej przez tablicę, a także na zmierzenie ilości i rodzajów substytucji, które zaszły w sekwencji i które zostały zakumulowane. Można w ten sposób zaobserwować historię sekwencji. W czasie symulacji nukleotyd był wybierany do mutacji losowo, z prawdopodobieństwem p_{mut}, a następnie ulegał

substytucji zgodnie z prawdopodobieństwami przejść zawartymi w tablicy. Dzięki temu nie każdy nukleotyd wybrany do mutowania rzeczywiście ulegał podstawieniu. Takiej analizie poddano cztery sekwencje: pierwsze, drugie i trzecie pozycje w kodonach ORFów z nici wiodącej, oraz sekwencję o takiej samej długości, ale o równowagowym składzie, [A]=[T]=[G]=[C]. Rys. 18 pokazuje zmiany składu nukleotydowego w trakcie symulacji. Sekwencja pierwszych i drugich pozycji oraz sekwencja równowagowa stopniowo przybierają skład trzecich pozycji w kodonach, natomiast te pozostają bez zmian. Wynika z tego, że trzecie pozycje w kodonach znajdują się w stanie równowagi z presją mutacyjną, a wpływ selekcji nie jest widoczny.

Aby obliczyć ilość substytucji, które zaszły w trakcie symulacji, badaną sekwencję porównywano z sekwencją z poprzedniego kroku Monte Carlo (Rys. 19). Rodzaj i liczba podstawień zależy nie tylko od presji mutacyjnej opisanej tablicą, ale także od składu nukleotydowego analizowanej sekwencji. Innym badanym parametrem było tempo akumulowania substytucji. Obliczano je przez porównanie sekwencji po każdym kroku symulacji z sekwencją wyjściową (Rys. 20). Początkowo różnice między tymi sekwencjami narastają bardzo szybko, ponieważ przy braku selekcji każda substytucja zostaje zachowana w sekwencji. Z upływem czasu sekwencje coraz bardziej upodobniają się do siebie pod względem składu nukleotydowego, a wiele substytucji zachodzi w tych samych miejscach i zdarzają się również rewersje, natomiast liczba pozycji w których sekwencja po symulacji różni się od sekwencji wyjściowej utrzymuje się na stałym poziomie. Ten poziom jest inny dla każdego rodzaju substytucji i zależy również od składu sekwencji wyjściowej.

W badaniach filogenetycznych porównuje się sekwencje „po mutacjach", nie ma możliwości poznania sekwencji wyjściowych. Dlatego istnieje potrzeba wprowadzenia poprawki na wielokrotne substytucje i rewersje, aby obliczyć rzeczywistą liczbę substytucji, które zaszły od czasu ewolucyjnego rozejścia się sekwencji. Taką próbę stanowi poprawka Kimury (KIMURA 1980, 1983). Jak widać na Rys. 21, zastosowanie tej poprawki daje satysfakcjonujące wyniki tylko dla niewielkich odległości filogenetycznych, natomiast dla większych odległości liczba substytucji jest niedoszacowana. Znajomość tablicy przejść mutacyjnych umożliwia wprowadzenie precyzyjnych poprawek dla każdego rodzaju substytucji (KOWALCZUK i współpr. 2001b).

Właściwości tablicy substytucji

Tablica uzyskana w niniejszej pracy jest pierwszą, która tworzy DNA w stanie równowagi o składzie nukleotydowym obserwowanym w naturze. W przeciwieństwie do modeli parametrycznych, tablica ta zachowuje zarówno skład nukleotydowy DNA jak i asymetrię sekwencji będących w stanie równowagi z presją mutacyjną, tutaj – trzecich pozycji w kodonach ORFów z nici wiodącej.

Tempo podstawiania nukleotydów w sekwencji można porównać do procesu rozpadu pierwiastków radioaktywnych. Zanikanie nukleotydów w sekwencji i pojawianie się nowych pokazane jest na Rys. 22. Można obliczyć „półokres rozpadu" dla każdego nukleotydu, czyli czas podstawienia połowy nukleotydów danego rodzaju w sekwencji. Okazuje się, że dla rzeczywistej tablicy substytucji uzyskanej dla sekwencji w stanie równowagi z presją mutacyjną ten półokres substytucji jest liniowo skorelowany z frakcją danego nukleotydu w sekwencji (Rys. 23a). Im mniej jest danego nukleotydu, tym szybciej ulega on podstawieniu. Współczynnik korelacji jest rzędu 0,999. Takiej korelacji nie zaobserwowano dla tablicy przejść wygenerowanej przez komputer (Rys. 23b), mimo że utrzymuje ona asymetryczny skład DNA (KOWALCZUK i współpr. 1999b). Analiza tablic opublikowanych przez innych autorów ujawniła, że korelacji nie wykazują tablice dla sekwencji będących pod wpływem selekcji, natomiast tablice dla sekwencji wolnych od selekcji wykazują wysokie współczynniki korelacji (Tabela 3). Tablica znaleziona dla trzecich pozycji w czterokrotnie zdegenerowanych kodonach mitochondrialnego DNA *Drosophila melanogaster* (TAMURA 1992) stosuje się do tego prawa bardziej dokładnie niż tablica dla wszystkich trzecich pozycji w kodonach w tym genomie (takie same wyniki uzyskano dla tablic dla mtDNA naczelnych opublikowanych przez ADACHI i HASEGAWA, 1996). Takich różnic należałoby oczekiwać, gdyby niektóre substytucje w trzecich pozycjach, prowadzące do podstawień aminokwasowych, nie były neutralne. Można także zauważyć, że tablice uzyskane na podstawie analizy podstawień w różnych pseudogenach tego samego organizmu lub organizmów bardzo blisko spokrewnionych, dają różny skład DNA w stanie równowagi, co potwierdza tezę, że presja mutacyjna różni się w różnych regionach genomów eukariotycznych (FILIPSKI 1988, WOLFE et al. 1989, MATASSI et al. 1999).

Precyzyjne, niemal deterministyczne relacje między frakcją danego nukleotydu a tempem jego podstawiania umożliwiają ocenę, czy badana tablica została uzyskana dla sekwencji wolnych od selekcji. Umożliwiają także obliczenie odległości między badaną

sekwencją a sekwencją w stanie równowagi z presją mutacyjną. Ta odległość jest miarą presji selekcyjnej, która utrzymuje sekwencję w stanie odchylonym od równowagi. Dzięki temu można na przykład określić presję na każdą pozycję w kodonach sekwencji kodujących białka.

Podsumowanie i wnioski

- W genomie *Borrelia burgdorferi* położenie genu na nici wiodącej lub opóźniającej wpływa na jego skład nukleotydowy, co odbija się na składzie kodonowym i składzie aminokwasowym kodowanego białka.
- Na podstawie asymetrii pierwszych i trzecich pozycji w kodonach, ORFy można podzielić na dwie nie nachodzące na siebie grupy, leżące na różnych niciach DNA.
- Przez porównanie sekwencji międzygenowych pochodzących od genów z ich macierzystymi genami, obliczono częstości wszystkich rodzajów substytucji (BbTs) dla sekwencji wolnych od selekcji, pochodzących z nici wiodących.
- Empiryczna tablica substytucji (BbTs) podlega prawu liniowej korelacji pomiędzy czasem substytucji połowy nukleotydów danego typu a ich frakcją w sekwencji.
- Opierając się na tym prawie można obliczyć precyzyjne poprawki na wielokrotne substytucje i rewersje w badaniach filogenetycznych.
- Metody analizy opisane w niniejszej pracy umożliwiają oszacowanie relatywnego udziału presji mutacyjnej i selekcyjnej w obserwowanej asymetrii.
- Chromosom znajduje się w stanie równowagi dynamicznej z presją mutacyjną związaną z replikacją i z presją selekcyjną działającą na kodowaną przez niego informację.
- Trzecie pozycje w kodonach ORFów są w stanie równowagi z presją mutacyjną związaną z replikacją, bez widocznego wpływu presji selekcyjnej w czterokrotnie zdegenerowanych kodonach.

Abstract

Asymmetry in nucleotide composition of DNA is defined as a deviation from [A]=[T] and [G]=[C] parities within one DNA strand. The *Borrelia burgdorferi* B31 chromosome is the most asymmetric of all bacterial chromosomes sequenced to date. The asymmetry was analysed in the whole chromosome, and in coding and intergenic sequences separately. Using DNA walks allowed the author to visualise and to measure the asymmetry and also to separate the effects of different mutational and selection pressures that generate it. The mechanisms generating asymmetry include unequal mutation rates connected with replication and transcription, selection forces positioning genes and signal sequences nonrandomly in the genome, and protein coding constraints on coding sequences. Intergenic sequences as well as each position in the codon in protein coding sequences show the influence of replication-associated mutational pressure. Third codon positions in the *B. burgdorferi* chromosome were found to be at equilibrium with the mutational pressure, free from selection pressure.

To find which types of substitutions are responsible for the asymmetry observed in the chromosome, intergenic sequences that had originated from duplicated genes were compared to the original genes. Based on that comparison, the frequencies of each of the twelve possible substitutions were calculated for the leading strand. The obtained empirical table of substitutions was tested analytically and by computer simulations. Every DNA sequence under the influence of the mutational pressure described by the table mutated towards the nucleotide composition of the third positions in codons of ORFs from the leading strand. The table of substitutions represents pure mutational pressure. In the absence of selection, the time when a half of nucleotides of a given type are substituted by other nucleotides is linearly correlated with the fraction of the analysed nucleotide in the sequence. The higher substitution turnover of a nucleotide, the lower the fraction of this nucleotide in the DNA sequence. The correlation coefficient is of the order of 0.999. The same correlation was found for substitution matrices obtained by other authors for sequences from different genomes, which were free from selection pressure. The correlation was absent from computer-generated matrices, even though they kept the specific nucleotide composition of the third codon positions, and from tables of substitution rates found for sequences under strong selection. The precise, almost deterministic relations between the nucleotide fractions and their turnover rates enable estimating if a matrix of substitutions is influenced by

selection or not. Also, it enables counting the distance between the given sequence and the sequence in equilibrium with the mutational pressure. This distance is supposed to be a measure of selection pressure, which keeps the sequence at the steady state, far from equilibrium.

Author's original publications related to the subject of the thesis

I. MACKIEWICZ P., GIERLIK A., KOWALCZUK M., SZCZEPANIK D., DUDEK M.R., CEBRAT S. (1999). **Mechanisms generating correlation in nucleotide composition in *Borrelia burgdorferi* genome.** *Physica A*, **273**: 103-115.

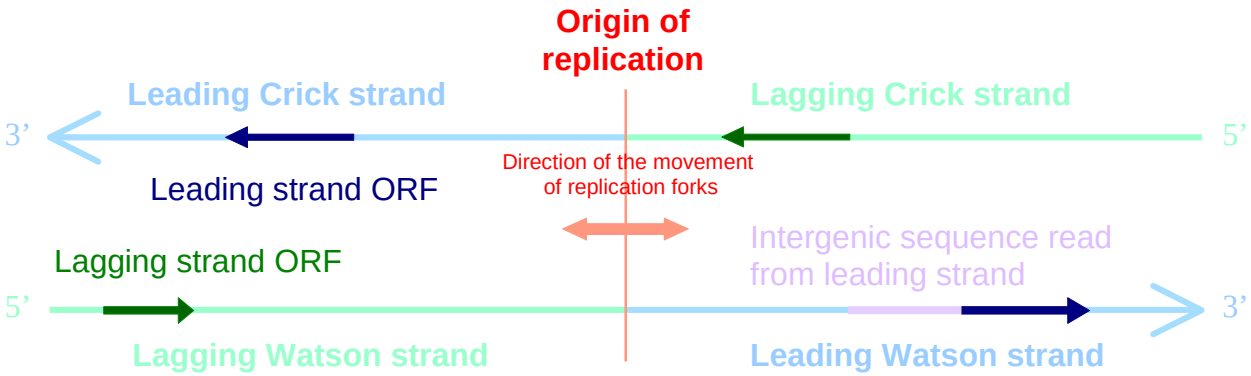
II. KOWALCZUK M., GIERLIK A., MACKIEWICZ P., CEBRAT S., DUDEK M.R. (1999). **Optimization of gene sequences under constant mutational pressure and selection.** *Physica A*, **273**: 116-131.

III. KOWALCZUK M., MACKIEWICZ P., SZCZEPANIK D., NOWICKA A., DUDKIEWICZ M., DUDEK M.R., CEBRAT S. (2001). **High correlation between the turnover of nucleotides under mutational pressure and the DNA composition.** BMC Evolutionary Biology, 1: 13.

IV. KOWALCZUK M., MACKIEWICZ P., MACKIEWICZ D., NOWICKA A., DUDKIEWICZ M., DUDEK M.R., CEBRAT S. (2001). **Multiple base substitution corrections in DNA sequence evolution.** International Journal of Modern Physics C, 12(7): 1043-1053.

V. KOWALCZUK M., MACKIEWICZ P., MACKIEWICZ D., NOWICKA A., DUDKIEWICZ M., DUDEK M.R., CEBRAT S. (2001). **DNA asymmetry and the replicational mutational pressure.** Journal of Applied Genetics, 42(4): 553-577.

Scheme of the B. burgdorferi chromosome



Abbreviations and definitions

W strand = Watson strand (usually the one located in data bases)

C strand = Crick strand (complementary to W strand)

Lagging DNA strand = synthesised from Okazaki fragments, in the opposite direction to the replication forks movement

Leading DNA strand = synthesised continuously, in the same direction as the replication forks movement

ORF = Open Reading Frame, a DNA sequence which begins with a start codon and ends with a stop codon, a potential protein coding sequence

An **ORF** is located on the **leading strand** when its sense strand is the leading strand, lagging strand ORFs respectively.

Sense strand = noncoding strand = non-template strand = nontranscribed strand of gene

Antisense strand = coding strand = template strand = transcribed strand of gene

CDS = (protein) coding sequence

Non-CDS = noncoding sequence

AT skew = (A-T)/(A+T)

GC skew = (G-C)/(G+C)

CAI = Codon Adaptation Index describing codon usage for optimal translation rate

MCS = Monte Carlo Steps (in computer simulations)

PR2 = Parity Rules Type 2, i.e. [A]=[T] and [G]=[C] are true for a single DNA strand

Mutational pressure = the rates of the twelve kinds of nucleotide substitutions

Selection pressure = the probability of elimination of a substitution

DNA in equilibrium = the sequence is free from selection pressure and the general composition of the evolving DNA sequence corresponds to the substitution frequencies (mutational pressure)

DNA in steady state = the sequence is in equilibrium with the selection pressure, far from the equilibrium with the mutational pressure

N = any base in DNA (A, T, G or C)

P = a purine (A or G)

Y = a pyrimidine (T or C)

BbTs = *Borrelia burgdorferi* table of substitutions

Amber, Ochre, Opal = stop translation codons

The table of the genetic code

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

Two-fold degenerated codons = the codons where a transition does not change the sense of the encoded amino acid, but a transversion does. In the above table they are located in the white boxes.

Four-fold degenerated codons = the codons where neither a transition nor a transversion change the sense of the encoded amino acid. They are located in the yellow boxes.

There are only two grey semi-boxes where a transition changes the sense of the codon.

1. Introduction

1.1. Definition of DNA asymmetry

The specific structure of the double-stranded DNA molecule implies many of its genetic and chemical features. One of the most important features is the complementarity of the two DNA strands, where the number of adenines is exactly the same as the number of thymines, while the number of guanines is exactly the same as the number of cytosines. These are the Chargaff's rules (CHARGAFF 1950) which helped WATSON and CRICK (1953) to describe the structure of the double helix and to find out that the Chargaff's rules are deterministic. If we assume that there are no mutational or selection pressures which influence the composition of the two DNA strands, the rules: [A]=[T] and [G]=[C] should be in force not only for double-stranded DNA but also for each of the two strands. These rules for a single DNA strand are stochastic instead of deterministic, and are called parity rules type 2, in short PR2 (LOBRY 1995). Deviation from PR2 means that the two DNA strands are under different mutational or selection pressures or both, which leads to asymmetric substitution patterns and DNA asymmetry.

1.2. Finding DNA asymmetry

PR2 are valid for random DNA molecules as well as for whole chromosomes. Let us count the number of each of the four types of nucleotides in the Watson strand of the *Borrelia burgdorferi* chromosome and construct a single-stranded "leading" and "lagging" DNA sequences by drawing the nucleotides randomly from the pool in which the frequency of each nucleotide is the same as in the *B. burgdorferi* chromosome. In Tab. 1 the composition of such a random DNA sequence (an example of one computer simulation) is shown. There are no significant differences in the nucleotide composition of these two sequences. It is not the case when one looks at the halves of the real Watson strand replicated as leading or lagging (see Table 1). The Chargaff's parity rules are valid but the PR2 are not: the differences between the numbers of complementary nucleotides in each strand are significant.

Strand type	Number of bases in strand				Strand length
	A	T	G	C	
Random leading Watson strand	160,481	162,812	64,209	64,946	452,448
Random lagging Watson strand	162,216	165,081	65,453	65,526	458,276
Real leading Watson strand	145,921	178,068	75,741	52,718	452,448
Real lagging Watson strand	177,186	149,128	53,911	78,051	458,276

Table 1. Numbers of nucleotides in the leading and lagging part of the Watson strand of the *Borrelia burgdorferi* chromosome. The random strand was obtained by drawing nucleotides randomly from a pool of all nucleotides from the Watson strand.

Deviations from PR2 and differences in composition between the leading and the lagging strands were observed in many eubacterial genomes (LOBRY 1996a, 1996b, BLATTNER et al. 1997, KUNST et al. 1997, FRASER et al. 1997, 1998, ANDERSON et al. 1998, FREEMAN et al. 1998, MRAZEK, KARLIN 1998, GRIGORIEV 1998, MCLEAN et al. 1998, SALZBERG et al. 1998, MACKIEWICZ et al. 1999a, 1999b, TILLIER, COLLINS 2000a, KOWALCZUK et al. 2001a) and are still detected in newly sequenced genomes. The asymmetry was also detected in many viruses (DANIELS et al. 1983, FILIPSKI 1990, MRAZEK, KARLIN 1998, GRIGORIEV 1999). Generally it was found that the leading strand is rich in guanine and thymine, and the lagging strand, in cytosine and adenine.

The asymmetry is observed even at the level of codons and amino acids (PERRIERE et al. 1996, MCINERNEY 1998, LAFAY et al. 1999, MACKIEWICZ et al. 1999b, ROCHA et al. 1999a, ROMERO et al. 2000). This kind of asymmetry generally does not exist in archaeal genomes (MRAZEK, KARLIN 1998, GRIGORIEV 1998, MCLEAN et al. 1998) and was found only in some of them (SALZBERG et al. 1998, LOPEZ et al. 1999, 2000). Analyses of eukaryotic genomes do not show the asymmetry on a large scale (MRAZEK, KARLIN 1998, GRIGORIEV 1998, GIERLIK et al. 2000), although some specific asymmetry in subtelomeric regions of yeast chromosomes was observed (GIERLIK et al. 2000, see also GRIGORIEV 1998).

The asymmetry is so strong that it can come to assistance in experimental searches for the origin and terminus of replication (e.g. QIN et al. 1999, PICARDEAU et al. 1999, 2000, ZAWILAK et al. 2001).

1.3. Mechanisms generating asymmetry

A substitution in one DNA strand is reflected by a change in the complementary strand. However, to understand the origin and meaning of asymmetry, it is important to find where, and how the primary changes occur which lead to different substitutions in different regions of the chromosome.

Mechanisms that could introduce asymmetry into DNA strands have been discussed many times (see for review: FRANCINO, OCHMAN 1997, MRAZEK, KARLIN 1998, FRANK, LOBRY 1999, KARLIN 1999, TILLIER, COLLINS 2000a, KOWALCZUK et al. 2001a). Nucleotide composition of a sequence is shaped by two different and sometimes opposite forces: mutational and selection pressures. Generally, they include mutational pressures on DNA during replication and transcription, selection forces positioning genes and signal sequences nonrandomly in the chromosome, and protein coding constraints on coding sequences.

1.3.1. Replication-associated mutational pressure

An important structural feature of the DNA molecule is that the two strands are antiparallel. Together with the properties of replication mechanisms, it has very significant genetic implications. DNA strands can be synthesised only in one direction: from the 5' to 3' end. Because the strands are antiparallel and replication forks move along the maternal double strand molecule, the two new strands have to be synthesised by different mechanisms and different replication-associated mutational pressures may influence their nucleotide composition. After many generations, parity rules type 2 should not be in force in such DNA molecules (see Table 1, the composition of the replichores of the *B. burgdorferi* chromosome).

Synthesis of one strand, called the leading strand, is continuous, while synthesis of the other strand, called the lagging strand, has short intermediates named Okazaki fragments (OKAZAKI et al. 1968). Differences between the synthesis of the strands have been reviewed by FRANK and LOBRY (1999). In *Escherichia coli* both DNA strands are synthesised by symmetric core enzymes (Pol III holoenzyme, BAKER, WICKNER 1992, MARIANS 1992, YUZHAKOV et al. 1996), and therefore base incorporation and proofreading should be the same for both strands. However, the enzyme complexes differ in processivity (tendency to remain on a single template, MARIANS 1992). The

leading strand complex needs to be more processive to remain on the template, while the lagging strand complex needs to dissociate more often, which facilitates excision of a mismatch by some cellular exonuclease (FIJALKOWSKA, SCHAAPER 1996). Thus, the lagging strand synthesis should be more faithful. Also, the strands may differ in stepwise progression speed and mismatch repair mechanisms (RADMAN 1998). The lagging strand polymerase should synthesise DNA faster to compensate for the time of its recycling, so more errors may be committed in the process, but on the other hand the discontinuous replication provides nicks in DNA, which are required by mismatch repair, so the lagging strand repair could be more efficient. Experimental analyses of the relative fidelity of the leading and lagging strand replication have given contradictory results (e.g. compare IWAKI et al. 1996 and FIJALKOWSKA et al. 1998). Generally, in experiments lagging strands seem more prone to mutations (e.g. TRINH, SINDEN 1991, BASIC-ZANINOVIC et al. 1992, VEAUTE, FUCHS 1993; ROBERTS et al. 1994, THOMAS et al. 1996). However, these results should be carefully considered, because the experiments were performed in specific conditions, e.g. the strains used in the studies were deficient in proofreading or mismatch repair.

A theory (named the cytosine deamination theory) that explains the influence of replication-associated mutational pressure on asymmetry was presented by FRANK and LOBRY (1999). During replication, stretches of the leading strand that are the template for the newly synthesised lagging strand are temporarily single-stranded. In this state the template is more exposed to damage and mutations (similarly to the sense strand during transcription). The most frequent mutation is deamination of cytosine and its homologue 5-methylcytosine to uracile (ECHOLS, GOODMAN 1991, LINDAHL 1993, KREUTZER, ESSIGMANN 1998). Uracile may be converted to thymine, which leads in consequence to C→T transition. It was found that cytosine deaminates 140 times faster in single-stranded DNA than in double-stranded (FREDERICO et al. 1990). This transition explains the excess of guanine and thymine in the leading strand, and adenine and cytosine in the lagging strand. The increase in the number of thymines is associated with a decrease in the number of cytosines. When the number of cytosines decreases, the percentage of guanines increases in the leading strand. Thus, in the leading strand the prevalence of thymine and guanine is observed. In the lagging strand, complementary to the leading one, prevalence of adenine and cytosine is observed. A similar result is given by a less common A→G transition which results from deamination of adenine to hypoxantine (LINDAHL 1993). Hypoxantine binds preferably

with cytosine, which leads to the increase of guanine in the leading strand. The decrease in the number of adenines results in the increase in the percentage of thymine, and in the opposite changes in the complementary, lagging strand – increase of adenine and cytosine. The deamination theory gives an especially convincing explanation of asymmetry in mitochondrial (TANAKA, OZAWA 1994, REYES et al. 1998) and viral genomes (GRIGORIEV 1998, 1999).

1.3.2. Transcription-associated mutational pressure

The genetic information stored in DNA can also be read only in one direction: from the 5' to 3' end. However, there are six possible reading frames and the transcribed strand of a gene may be located in the same or opposite direction to the replication fork movement (on the leading or lagging strand). Because genes are not distributed uniformly in the chromosome, and coding and non-coding strands are treated differently by transcription, PR2 may be violated.

A potential cause of asymmetry may also be deamination of methylated cytosines which leads to thymines. Some authors have claimed that this type of substitution differentiates sense and antisense strands of coding sequences, and that transcription mechanisms introduce the asymmetry into DNA strands (FRANCINO et al. 1996, FRANCINO, OCHMAN 1997, FREEMAN et al. 1998). During transcription a part of the nontranscribed DNA strand is exposed and more prone to deamination (BELETSKII, BHAGWAT 1996), while the other strand is protected by the enzymatic transcription complex and by transcription-coupled repair that preferentially repairs pyrimidine dimers (MELLON, HANAWALT 1989, HANAWALT 1991). Some experiments have proved that the frequency of mutations introduced into the non-transcribed DNA strand is higher than into the transcribed one (FRANCINO et al. 1996).

1.3.3. Unequal distribution of genes and oligomers on chromosome

Transcription-associated mutational pressure alone does not distinguish between the leading and lagging strand. However, if highly transcribed genes are preferably located on one strand, the bias between strands should be generated in coding sequences and in the intergenic regions that are partly transcribed.

There are preferences for transcribing DNA strands in the direction of replication, rather than in the inverse direction, possibly to avoid collisions between replication and transcription complexes (BREWER 1988). In *Mycoplasma genitalium*, *M. pneumoniae* and *Bacillus subtilis* over 75% of genes are located on the leading strands (FRASER et al. 1995, HIMMELREICH et al. 1996, KUNST et al. 1997). Hence, transcription-associated mutational pressure may contribute to the leading/lagging strand asymmetry. However, in the *E. coli* chromosome the bias is relatively low and only 54% of coding sequences are located on the leading strand (BLATTNER et al. 1997), so it cannot account for the asymmetry observed in that chromosome.

It has also been found that in prokaryotic chromosomes usually the majority of highly expressed genes are located on the leading strand. Those genes use a small subset of specific codons (GOUY, GAUTIER 1982, SHARP, LI 1987), which may contribute to the asymmetry observed. TILLIER and COLLINS (2000a) tried to estimate the contribution of transcription to asymmetry by analysing Codon Adaptation Index (CAI) values of genes. They found that genes with the highest and lowest CAI did not account for the correlation of base composition skew with replication orientation, and these skews were not completely explained by the selection for highly expressed genes on the leading strand. However, these analyses assumed that CAI value (which actually measures translation level/intensity) corresponded to transcription level/intensity, which may not be true. Only experimental research can provide evidence that genes on the leading strand are more intensively transcribed than genes on the lagging strand.

Leading- and lagging-strand-specific codon usage has been observed in *Borrelia burgdorferi*, *Treponema pallidum* and *Chlamydia trachomatis* (MCINERNEY 1998; LAFAY et al. 1999, ROCHA et al. 1999a, ROMERO et al. 2000). However, because the presumably highly expressed genes in those genomes do not differ in codon usage from other genes located on the same strand, some authors conclude that the two different patterns are the result of replication-associated mutational pressure and not selection, and codon usage is strand-specific and not correlated with the level of expression (MCINERNEY 1998; LAFAY et al. 1999).

Oligomers that are over-represented on one of the strands could contribute to DNA asymmetry. For example, Chi sequence 5’ GCTGGTGG 3’, which is a recombinational hot spot, is located preferentially on the leading strand of the *E. coli* chromosome (BLATTNER et al. 1997). SALZBERG et al. (1998) observed skewed distribution of some oligomers on leading and lagging strands. In *B. burgdorferi*, *T. pallidum*, *E. coli*, *B.*

subtilis and other genomes 7-, 8-, and 9-mers are statistically significantly skewed and are helpful in finding the origin and terminus of replication. The nucleotide composition of these oligomers is correlated with the most abundant codons in those genomes, although they do not occur preferentially within coding regions. Their asymmetry is much stronger than the inequality in the distribution of coding sequences. Their function is unknown but they are expected to play a role as biological signals in replication and transcription (ROCHA et al. 1998), so their distribution should be subject to selection. However, Chi-sites make up only 0.25 % of the *E. coli* chromosome and are not likely to be an important source for global base composition asymmetry (FRANK, LOBRY 1999), which is further supported by analyses of several genomes by TILLIER and COLLINS (2000a). Removal of all skewed octamer sequences from the *E. coli* and *Haemophilus influenzae* chromosomes gave reduced asymmetry but did not eliminate it. Apparently skewed oligomers are not the main source of asymmetry.

1.3.4. Protein coding constraints on coding sequences

Coding for proteins requires a specific nucleotide composition. It has been long known that coding strands of genes are rich in purines (e.g. SHEPHERD 1981, SMITHIES et al. 1981, KARLIN, BURGE 1995, FRANCINO et al. 1996, CEBRAT et al. 1997a, FREEMAN et al. 1998). DNA sequences which code for proteins have a triplet structure. Each position in the codon has specific preferences in nucleotide composition (WONG, CEDERGREN 1986, ZHANG, ZHANG 1991, GUTIERREZ et al. 1996, MRAZEK, KARLIN 1998, CEBRAT et al. 1997b, 1998, MCLEAN et al. 1998, WANG 1998), which suggests that it plays a unique role and remains under a specific selection pressure. Generally, the first codon positions of protein coding sequences are rich in adenine and guanine and the second are rich in adenine and cytosine. The asymmetry between coding and non-coding strands of genes is so strong that it can be used to successfully discriminate between coding and non-coding sequences (CEBRAT et al. 1997a, 1997b, 1998).

There are many mechanisms that contribute to the specific composition of genes. The common presence of purines in the sense strand is favoured by evolution because they are less prone to mutations than pyrimidines (especially dimers) (HUTCHINSON 1996). During transcription the sense strand is more exposed than the antisense strand which is preferably repaired by removal of pyrimidine dimers (MELLON, HANAWALT

1989, HANAWALT 1991). Therefore selection should increase the purine content of the coding strand (FREEMAN et al. 1998, FRANK, LOBRY 1999).

Furthermore, the base composition of the first and second positions in codons reflects the high usage of acidic amino acids coded by GAN (asparagine and glutamine), (KARLIN, MRAZEK 1996) and GNN (glycine, alanine and valine) (KARLIN et al. 1992). The second codon position determines the polarity of the encoded amino acid and its change may have a detrimental effect on the protein.

Periodical codon composition pattern $(GCU)_n$ plays a role in mRNA-rRNA interaction during translation in the ribosome (TRIFONOV 1987, 1992, LAGUNEZ-OTERO, TRIFONOV 1992). Guanines in the first codon positions interact with periodically distributed cytosines in rRNA and ensure the correct reading frame during translation.

Third codon positions are degenerated and most substitutions in them are silent. These substitutions are not necessarily neutral. They may change the rate of translation of the product (IKEMURA 1981, BENNETZEN, HALL 1982, SHARP, COWE 1991). However, selection on third positions in codons is the weakest and the effect of mutational pressure should be observed in them.

If coding sequences are in the same number on both leading and lagging strands, their compositional bias should be cancelled out. Otherwise, if they are not randomly distributed on chromosome, they can contribute to the global asymmetry (asymmetry of the whole chromosome).

1.3.5. Relative contribution of different factors to DNA asymmetry

Although there are many different and sometimes contradictory hypotheses and opinions about the influence of asymmetry, it is possible to draw some conclusions. The impact of uneven gene distribution on global asymmetry is different in various genomes. TILLIER, COLLINS (2000a) have assessed the relative contribution of gene orientation in many genomes to base composition asymmetry. In some genomes the influence of gene bias is opposite to that resulting from mutational pressure (MCLEAN et al. 1998, TILLIER, COLLINS 2000a). Highly expressed genes and signal sequences contribute to the bias only to a very small extent. The replication-associated mutational pressure is the most significant factor of the observed asymmetry. Some authors (CEBRAT et al. 1999, MACKIEWICZ et al. 1999a, 1999b, 1999c, TILLIER, COLLINS 2000a) have filtered by different methods the influence of replication from other mechanisms.

Although the degree of influence of transcription-associated mutational pressure still remains open, it seems that it is weaker than the influence of replication (FRANK, LOBRY 1999). The time of single-stranded state is shorter for the coding strand than for the lagging strand template. Deaminations in the sense strand cause only premutagenic lesions that have to wait for the next round of replication to become fixed and during this time can be repaired. The uracile resulting from deaminations of cytosine occurring in the lagging strand template is almost immediately paired with an incoming adenine in the synthesis of the lagging strand.

Furthermore, the influence of transcription on asymmetry may be consistent with the influence of replication-associated mutational pressure, because deamination of cytosine occurs both in the lagging strand template during replication, and during transcription in the coding strands, which are preferably located on the leading strand.

1.4. Rate of evolution of genes located on leading and lagging strands

1.4.1. Comparisons of orthologs from closely related genomes

Replication-associated mutational pressure is strong enough to influence gene evolution and rearrangements. A way to observe the influence of replication direction on gene evolution is to compare pairs of orthologs from closely related genomes. LAFAY et al. (1999) compared codon and amino acid usage between leading and lagging strand genes of *B. burgdorferi* and *T. pallidum*. Despite species-specific G+C content and chromosome structure and organisation, they found similar G-T versus A-C biases between the leading and lagging strands in these two species. The biases were found at the level of nucleotides, codons, and amino acids. The orthologs that have switched strands have adapted their codon and amino acid usage to their new strand and have the same codon usage as the genes of the new strand.

TILLIER and COLLINS (2000c), who compared orthologs from *Chlamydia trachomatis* and *Chlamydia pneumoniae*, also observed that the genes that switched the strand have acquired the skew of their current strand. Comparison of amino acid similarity and identity between the orthologs showed that the switched genes were on

average more diverged that the nonswitched ones. Changing the replication direction significantly changed the amino acid sequence and affected evolution of these sequences. Thus the substitutions resulting from mutational pressure are not neutral.

SZCZEPANIK et al. (2001) have measured differences in the rate of divergence between genes lying on the leading strand, lagging strand, and genes which changed their positions on chromosome during evolution. Analyses have been performed on 12,645 orthologs derived from 11 eubacterial genomes showing evident compositional asymmetry between leading and lagging strands. In almost all cases the distances between genomes measured by the divergence of orthologs from the lagging strand are statistically significantly larger than the distances counted on the basis of the leading strand orthologs. Apparently the orthologs situated on lagging strands diverge quicker than the orthologs situated on leading strands. This phenomenon can be explained either by a higher mutation rate on the lagging strand, or by stronger selection on the more conserved genes located on the leading strand (SZCZEPANIK et al. 2001). For closely related genomes the rate of divergence between the orthologs located on different strands is even greater than that of the lagging strand orthologs. The genes which have switched the strand recently are under a greater mutational pressure and diverge very quickly. The differences in the rate of divergence are significant enough to affect the structure of phylogenetic trees constructed on the basis of leading and lagging strand orthologs (SZCZEPANIK et al. 2001). Different mutational pressures on the two DNA strands group genes into slower and faster evolving groups. It may play an important role in adaptation to the quickly changing environment.

1.4.2. Rearrangements in genomes

MACKIEWICZ et al. (2001a) have found a method to determine which of two orthologs located on different strands has actually been relocated. Two pairs of highly asymmetric genomes were analysed, *C. trachomatis* vs. *C. pneumoniae*, and *B. burgdorferi* vs. *T. pallidum*. CG and AT skews were measured for each analysed gene as well as mean values and standard deviations for all leading and lagging strand genes in each genome. The gene whose GC and AT skews were more distant from the mean for its current strand was considered switched. The authors have found that genes have been relatively more often transferred from lagging to leading DNA strands than vice versa. That may be because the more conserved genes from the leading strand can

tolerate fewer substitutions which change their amino-acid composition and codon usage when affected by a higher mutational pressure after inversion. Highly expressed genes seem to be more sensitive to discrimination control through codon usage (i.e. IKEMURA 1981, GOUY, GAUTIER 1982, SHARP, LI 1987). Moreover, the possible collisions between transcription and replication complexes may be more deleterious for highly expressed genes switched from the leading to lagging strand (MCINERNEY 1998).

The most specific rearrangements occur around the origin of replication (SUYAMA, BORK 2001, EISEN et al. 2000, READ et al. 2000, TILLIER, COLLINS 2000b). In closely related genomes, many orthologs coding for the same function remain at the same distance and orientation to the origin or terminus of replication, but they can be positioned on either of the two replichores. This property gives a specific picture when the positions of genes in one genome are plotted against the positions of their homologs in a closely related genome. TILLIER and COLLINS (2000b) have argued that the structure of replication forks, which are hot-spots of recombination, is responsible for that picture. However, the strand and distance from the origin of replication may be as well conserved by selection (MACKIEWICZ et al. 2001b). Firstly, the distance from the origin of replication determines copy number of a gene in bacteria whose generation time is shorter than replication period. In those cells the newly replicated origins initiate the next round of replication before the end of the previous round. Thus, in the cell there are several copies of genes proximal to the origin. Highly and lowly expressed genes should be located in optimal distances from the origin (LIU, SANDERSON 1995, 1996). Secondly, transfer of a gene to the opposite strand increases mutational pressure on that gene, as mentioned above, and thus should be selected against. Thirdly, there is a trend to keep both replichores the same size (LIU, SANDERSON 1996), possibly because that ensures the shortest time of replication of the genome.

1.5. Effects of mutational and selection pressures

The biased substitutions occurring during replication and transcription are the mutational pressure on the sequence. To see the pure effect of the mutational pressure, one must find sequences which are free from selection. Sequences are in equilibrium with the mutational pressure when the general composition of the evolving DNA sequence corresponds to the substitution frequencies and it does not change any more. In equilibrium, the number of a given nucleotide *substituted* by other nucleotides is

balanced by the number of that nucleotide *substituting* the other nucleotides. The following four equations must be fulfilled:

$$N_{A\to G}+N_{A\to C}+N_{A\to T}=N_{G\to A}+N_{C\to A}+N_{T\to A}$$

$$N_{G\to A}+N_{G\to C}+N_{G\to T}=N_{A\to G}+N_{C\to G}+N_{T\to G}$$

$$N_{C\to A}+N_{C\to G}+N_{C\to T}=N_{A\to C}+N_{G\to C}+N_{T\to C}$$

$$N_{T\to A}+N_{T\to G}+N_{T\to C}=N_{A\to T}+N_{G\to T}+N_{C\to T}$$

where $N_{A\to G}=N_A*p(N_{A\to G})$, N_A is the number of adenines in the sequence, $p(N_{A\to G})$ is the probability of substitution of A by G, other symbols – respectively. After a long enough evolution time, the general composition of the evolving DNA sequence should not change any more, while the divergence between the original sequence and the evolved sequence should approximate the value which can be calculated directly from its nucleotide composition and is described by the equation:

$$D=1-(A_0*A_t+T_0*T_t+G_0*G_t+C_0*C_t)$$

where subscripts *0* ant *t* denote the fractions of nucleotides in the original sequence and the sequence evolved during the time *t*, respectively.

Most sequences in the genome are selected depending on where the mutation occurred and how it changed the fitness of the organism. Selection pressure positions genes in optimal regions of the genome and eliminates some mutations while it accepts others. When a sequence has adapted to the selection pressure, its general nucleotide composition stops changing, but the sequence is in the steady state, far from equilibrium. The distance between a given sequence and the sequence in equilibrium with the mutational pressure is the measure of selection pressure.

2. Aims of the Study

The present study was designed to investigate the relationship between mutational and selection pressures and location of the *Borrelia burgdorferi* genes on the chromosome. The objective of this study was to estimate the frequencies of substitutions which generate the asymmetry observed in the *B. burgdorferi* chromosome and to separate the effect of mutational pressure from the effect of selection.

The specific aims were to:

- analyse the asymmetry in nucleotide composition of the leading and lagging DNA strands,
- separate different factors contributing to the asymmetry,
- visualise and measure the asymmetry,
- find sequences which are in equilibrium with the mutational pressure,
- determine frequencies of the twelve kinds of substitutions which generate the asymmetry,
- study the influence of the asymmetry on gene evolution.

3. Materials and Methods

3.1. Borrelia burgdorferi species

Borrelia is a spiral-shaped, gram-negative bacterium with 7 to 11 periplasmic flagella. It varies from 10 to 30 μm in length and 0.2 to 0.5 μm in width (BARBOUR, HAYES 1986).

Borrelia burgdorferi was first isolated from the hard tick *Ixodes scapularis* by BURGDORFER et al. (1982) and recognized as the agent of Lyme disease. Further molecular analyses of different isolates have shown that *Borrelia burgdorferi* is in fact a group of related species, which are now referred to as "*Borrelia burgdorferi* sensu lato". This group includes *B. burgdorferi* sensu stricto and *B. garinii* (BARANTON et al. 1992), *B. afzelii* (BARANTON et al. 1992, CANICA et al. 1993), *B. japonica* (KAWABATA et al. 1993), *B. andersonii* (MARCONI et al. 1995), *B. tanukii* and *B. turdi* (FUKUNAGA et al. 1996), *B. valaisiana* (WANG et al. 1997), *B. lusitaniae* (LE FLECHE et al. 1997) and *B. bissettii* (POSTIC et al. 1998). Not all of them have been shown directly to cause human disease; the three major pathogenic species are *B. burgdorferi* sensu stricto, *B. garinii,* and *B. afzelii* (see WANG et al. 1999 for review).

3.2. Borrelia burgdorferi genome

Only one *Borrelia* genome has been sequenced so far, that of *Borrelia burgdorferi* sensu stricto strain B31. The genome contains a linear chromosome of 910,725 base pairs (bp), (FRASER et al. 1997), and 12 linear and 9 circular plasmids of combined length of 610,694 bp (CASJENS et al. 2000).

In this study only the linear chromosome was analysed. The chromosome contains 853 ORFs (852,486 bp total). 564 ORFs are located on the leading strand (560,553 bp) and 286 are located on the lagging strand (291,933 bp) Intergenic sequences are 102,009 bp in length. This number does not add up with the total length of ORFs, because some ORFs overlap. The chromosome is replicated bidirectionally from the middle, which was suggested by its genetic organization (OLD et al. 1993) and GC skew (FRASER et al. 1997), and confirmed experimentally by PICARDEAU et al. (1999). Chromosome sequence and information about coding regions of the *B. burgdorferi* chromosome were obtained from the National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/.

3.3. Analysis of DNA asymmetry

3.3.1. Walks along the chromosome

The idea of a walk is that a virtual walker moves along a DNA sequence and its movement is depicted on a two dimensional graph. The direction of the movement depends on the type of nucleotide visited. There are many kinds of walks. For example, a prevalence of adenine in the sequence can be analysed. The whole chromosome is divided into fragments (or windows) in which the number of adenines is counted. The walker goes up if there are more adenines in the fragment than the average for the chromosome, and down if there are fewer adenines than expected. The value of the walker jump for a given fragment is calculated from the equation:

$J = [N] - (\overline{F}_N \times L)$, where:

J = the value of the walker jump;

N - the number of adenines in the analysed fragment;

$\overline{F}_N$ - the frequency of the occurrence of adenine in the set of all analysed fragments;

L - length of the analysed fragment (window) in base pairs (bp). The lower limit of length is $L = 1$.

The values of the walker jumps are shown on the graph, where X-axis indicates the location of each fragment on chromosome, and Y-axis shows J, or the relative cumulative difference between the real number adenines and the expected number if adenine was uniformly distributed along the analysed sequence. Because the sequence is divided into hundreds of fragments, the points that represent them on the graph form a continuous-looking line. If the line goes up, it means that the chromosome region is rich in adenine, if it goes down it means that adenine is underrepresented in the region.

The following sequences have been analysed:

- whole Watson strand of the chromosome, divided into fragments of equal length (153 bp),
- subsequent intergenic sequences (located outside the ORFs in the database),
- three sequences: of the first, second, and third positions in codons in all ORFs spliced,
- in some walks sequences of ORFs transcribed from leading and lagging strands are analysed separately (which makes six sequences).

The above sequences have been analysed with different kinds of walks (CEBRAT, DUDEK, 1998, MACKIEWICZ et al. 1999b). Generally, the above equation has been used with different variables and modifications:

- walks on nucleotides, [A], [T], [G], and [C]. Here N is the number of the analysed nucleotide in the fragment, and F is the frequency of the occurrence of the nucleotide in the set of all analysed fragments,
- proportion of the number of nucleotides within coding sequences to whole chromosome (coding density). Here N is the number of nucleotides which in the analysed fragment of Watson or Crick strand are inside coding sequences, F is the number of nucleotides in all analysed coding sequences, and L is the length of the analysed chromosome fragment in bp,
- codon composition of genes; here N is the number of codons coding for a given amino acid in the analysed ORF sequence, and F is the frequency of the analysed group of synonymous codons in the whole set of coding sequences.
- walks on differences [A]-[T] and [G]-[C]. The walker jump is:

  $J=N-N_{av}$,

  where N is the value of the difference [A]-[T] or [G]-[C] in the analysed fragment, and N_{av} is the average value for the whole sequence. Often, this kind of walk gives a much clearer picture than analysing nucleotides separately,
- similarly, proportion of purines [P] to pyrimidines [Y] in the third positions in codons of coding sequences is analysed. The walks are done separately for the third positions in two-fold and four-fold degenerated codons. Here N is [P]-[Y] counted for third positions of two-fold or four-fold degenerated codons in the analysed sequence, and N_{av} is the mean value of [P]-[Y] in the set of coding sequences.

To compare walks on sequences of different length, J values were divided by the length of the whole analysed sequence (normalised).

3.3.2. Subtraction and Addition of DNA walks

Mechanisms that introduce asymmetry into DNA strands can be divided into two groups: the ones that treat the two complementary strands oppositely, and the ones that have the same influence on both strands. The first group of mechanisms are connected with replication of chromosome, the second group mainly with transcription. Transformations of DNA walks enable distinguishing between these effects (CEBRAT et al. 1999, MACKIEWICZ et. al. 1999a,b,c).

The asymmetries introduced by replication-associated mechanisms into the leading and lagging DNA strands are of reciprocal sign. Thus, subtracting DNA walks done on complementary strands enhances the picture of replication-associated asymmetry. In the analysis of coding sequences, J values of ORFs located on the Crick strand were multiplied by (-1) and cumulated with the J values calculated for ORFs located on the Watson strand, in the order in which they appeared on chromosome. In the analysis of noncoding sequences and the whole chromosome, the analysed fragments were read alternately from the Watson or Crick strand.

When DNA walks on sequences situated on W strand are added to DNA walks performed on sequences from C strand, the reciprocal values of replication-associated asymmetry compensate each other and disappear, leaving the effect of asymmetry introduced by other mechanisms. To add walks, J values of sequences located on C strand were cumulated with the values J calculated for sequences located on W strand, accordingly to their location on chromosome.

3.3.3. Spiders

In another version of DNA walk, the walker analyses not [A], [T], [G] or [C] separately, but all four nucleotides during one walk. The walker goes up if there is a prevalence of guanine, down for cytosine, right for adenine and left for thymine. Thus, in this kind of walk the graph does not show the position on chromosome, only trends in the sequence. The method was used in different variants by MIZRAJI, NINIO 1985, GATES 1986, BERTHELSEN et al. 1992, LOBRY 1996b, and CEBRAT et al. 1997b, 1998. Spider walks are used here to picture protein coding sequences. Each of the three codon positions in a gene is analysed separately. The whole graph is called a spider and the three walks are called spider legs. This kind of walk can be done for single ORFs as well as for the sequence of spliced ORFs from the whole chromosome, here for leading-

and lagging strand ORFs separately. The genomic walks have been normalised to compare asymmetry in sequences of different length. If there are no trends in the analysed sequence, the path of the walker resembles Brownian motion. If a prevalence of a nucleotide(s) is observed in the sequence, the walk produces a long line in the graph.

3.3.4. Angle distributions on torus surface

The length of and the angle of the vector between the beginning of and the end of a spider walk can be used as parameters for further analysis of the sequence (CEBRAT et al. 1997b, CEBRAT, DUDEK 1998). The angle of the vector is in fact arcus tangent (A-T)/(G-C), where A, T, G, and C are the numbers of the respective nucleotides in the analysed codon position of the analysed ORF. The arcus tangent is calculated to avoid dividing by zero and to normalize distributions. Values for the first versus second or the first versus third positions are presented on graphs. Because the values are angles, they are located on the finite surface of the torus. Each analysed sequence is a single point with co-ordinates of two values of asymmetry. In this way distributions of asymmetry of different groups of sequences can be analysed and compared.

3.4. Construction of the table of substitutions

It is impossible to find out which substitutions generate the observed asymmetry simply by measuring the asymmetry itself. The asymmetric DNA composition can be realised by an infinite number of combinations of frequencies of the twelve possible nucleotide substitutions. Furthermore, protein coding sequences are not only under the mutational pressure typical for their location, but also are subject to selection for function, which mercilessly eliminates all undesirable substitutions. However, the intergenic sequences which are remnants of duplications of genes should accumulate all mutations. Their comparison with the original genes should reveal the influence of mutational pressure. This approach was adopted by KOWALCZUK et al. (2001b). Intergenic sequences longer than 90 nucleotides (arbitrarily accepted value) were translated into amino acids in the six possible reading frames. The amber and ochre stop codons were arbitrarily translated for tyrosine and opal for tryptophan, because only one substitution in third positions of these codons is sufficient to generate a stop. The *B.*

burgdorferi protein database was searched with the FASTA program (PEARSON, LIPMAN 1988) for homologs to those “proteins”. Because pseudogenes in intergenic space supposedly accumulate all mutations, very liberal criteria of homology were adopted, namely E value < 0.05. About thirty such homologs to intergenic sequences read from the leading strand were found among *B. burgdorferi* ORFs. The nucleotide sequences of the pairs (3737 residues total) were aligned using CLUSTAL X program (JEANMOUGIN et al. 1988). All the observed differences between the ORFs and their homologs were assumed to result from substitutions in the intergenic sequences. In that way the frequencies of all substitutions were found, and the average number of substitutions per site was 0.46. A table of mutations for the leading strand was constructed according to GOJOBORI et al. (1982) and FRANCINO and OCHMAN (2000), (Tab. 2, page 51). Since the observed substitution rates were different for each of the four nucleotides, corrections for multiple substitutions and reversions were introduced for each nucleotide separately, instead of one general correction according to KIMURA (1980).

4. Results

4.1. DNA asymmetry

4.1.1. Whole chromosome sequence

Analysis of the whole *Borrelia burgdorferi* chromosome is shown in Fig. 1-3. Subsequent steps of analysis are shown on the example of a [G]-[C] walk along the Watson strand. Fig. 1a shows [G]-[C] values for consecutive 153-nucleotide long fragments of the chromosome. Around the origin of replication (vertical line) a major change of trend is detected. The lagging strand (the first half of the Watson strand) is richer in cytosine than guanine and the leading strand is richer in guanine. To clarify the picture, much larger windows (sequence fragments analysed) can be used (Fig. 1b), and they may also overlap (Fig. 1c).

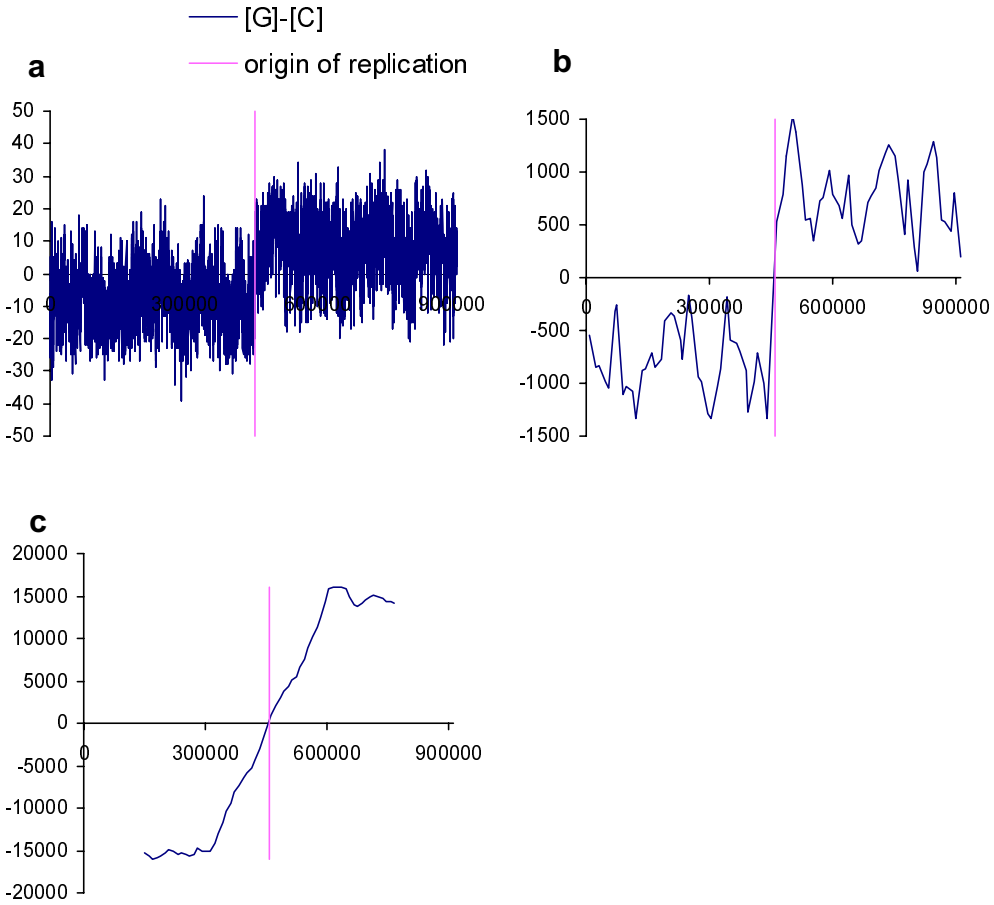


Figure 1. An analysis of [G]-[C] in the Watson strand of the *B. burgdorferi* chromosome. The number of [G]-[C] was analysed in fragments of different length, **a)** 153 bp, **b)** 15,300 bp, **c)** overlapping 300,000 bp fragments, the step size was 10,000 bp. Y-axis indicates location of fragments on chromosome.

The observed picture depends on an arbitrarily chosen size of the window. Cumulative diagrams give a much clearer picture of the change of the trend (Fig. 2a). Here, the values [G]-[C] from Fig. 1a have been cumulated. Extrema in the plot show the positions of the origin (minimum) and terminus (maximum) of replication, where the role of DNA strands changes from the leading to lagging or vice versa. The next step is to recalculate the walk to finish at y=0, which eliminates the trend of the whole sequence and makes the leading/lagging trends even clearer (Fig. 2b). The last step of the analysis is to normalise the values by the length of the walk to be able to compare the asymmetry between different sequences (Fig. 2c).

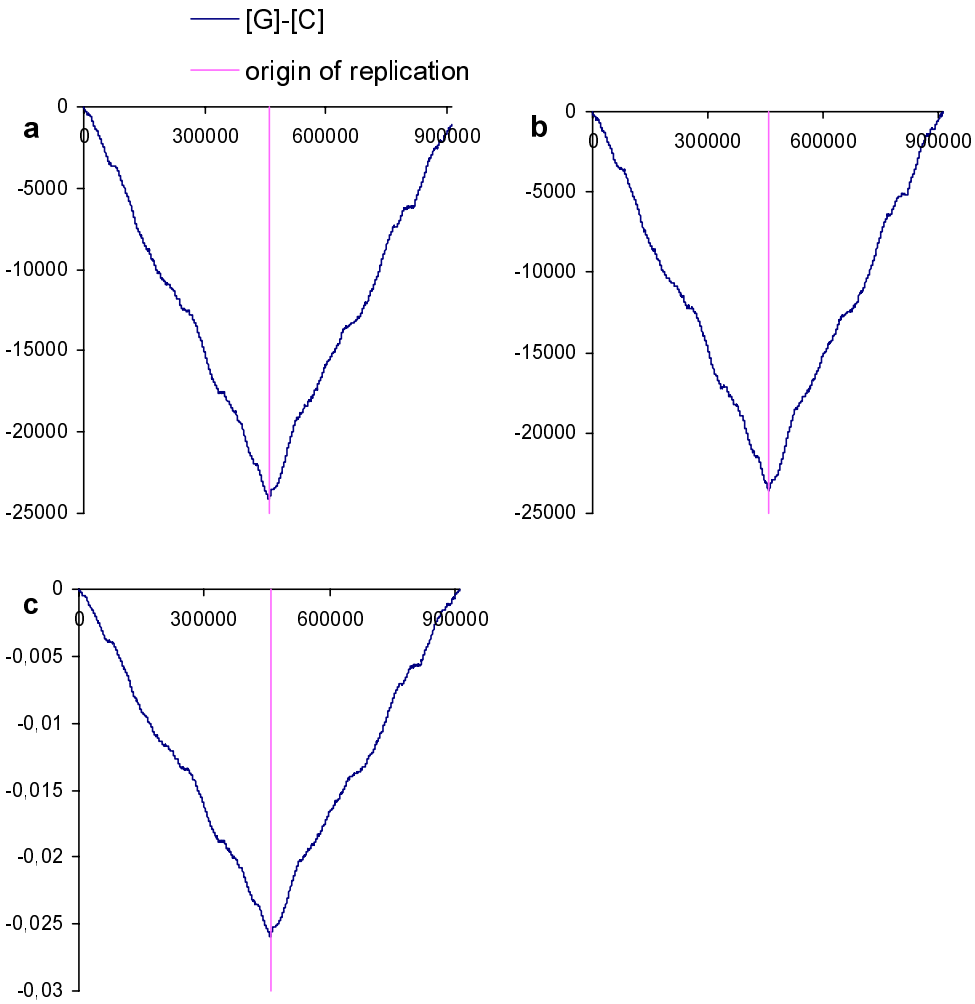


Figure 2. Cumulative diagrams for the Watson strand of the *B. burgdorferi* chromosome; **a)** cumulated values of [G]-[C] for consecutive 153 bp fragments of the chromosome, **b)** values from a) are detrended, according to the equation shown in the methods section, so the walk finishes at y=0, **c)** detrended values are normalized by the length of the sequence. Y-axis indicates location of fragments on chromosome.

Subtraction of walks on sequences read alternately from the Watson and Crick strands enhances the picture of the leading/lagging asymmetry because it eliminates the trends which are similar on both strands. In Fig. 3a such [G]-[C] and [A]-[T] walks are shown. Walks on particular nucleotides (Fig. 3b) reveal contribution of A, T, C and G to the chromosome asymmetry. It is clear that the lagging strand is rich in A and C, and the leading strand in G and T. In Fig. 4 a-b addition of walks is shown for comparison. The values of asymmetry are of one order of magnitude smaller.

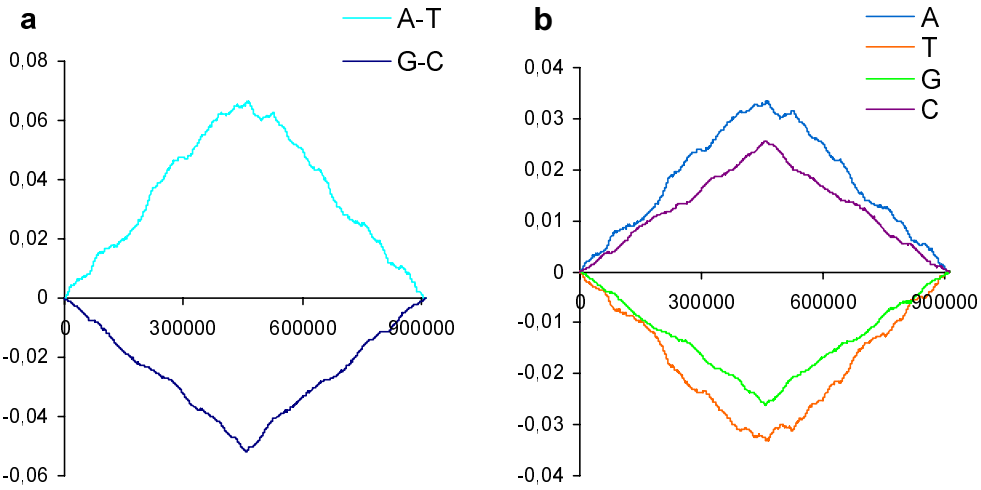


Figure 3. Subtracted DNA walks on the whole *B. burgdorferi* chromosome; **a)** walks on differences [A]-[T] and [G]-[C], **b)** walks on particular nucleotides. Y-axis indicates location on chromosome.

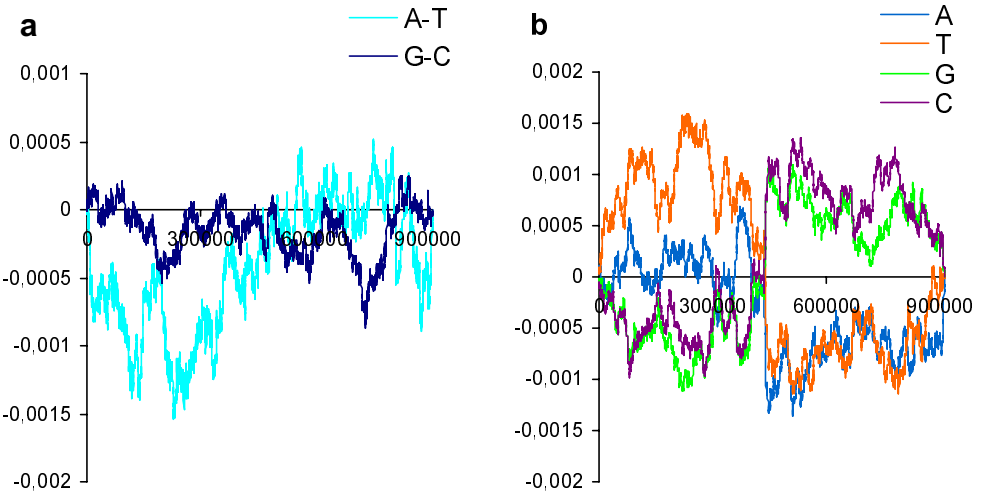


Figure. 4. Added DNA walks on the whole *B. burgdorferi* chromosome; **a)** walks on differences [A]-[T] and [G]-[C], **b)** walks on particular nucleotides. Y-axis indicates location on chromosome.

4.1.2. Coding and intergenic sequences

Cumulative walks on the first and second positions in codons of the W strand ORFs (Fig. 5a-b) do not show the leading/lagging trends. These trends are masked by nucleotide preferences specific for protein coding sequences in these codon positions, which are independent of the leading/lagging location of ORFs. That is why eliminating the coding trends is necessary in further analysis. In the first positions in codons there is a prevalence of A over T and of G over C, and in the second positions there is more A than T and more C than G. The trends observed in the third positions (Fig. 5c) are different for the leading and lagging halves of W strand, similarly to intergenic sequences (Fig. 5d).

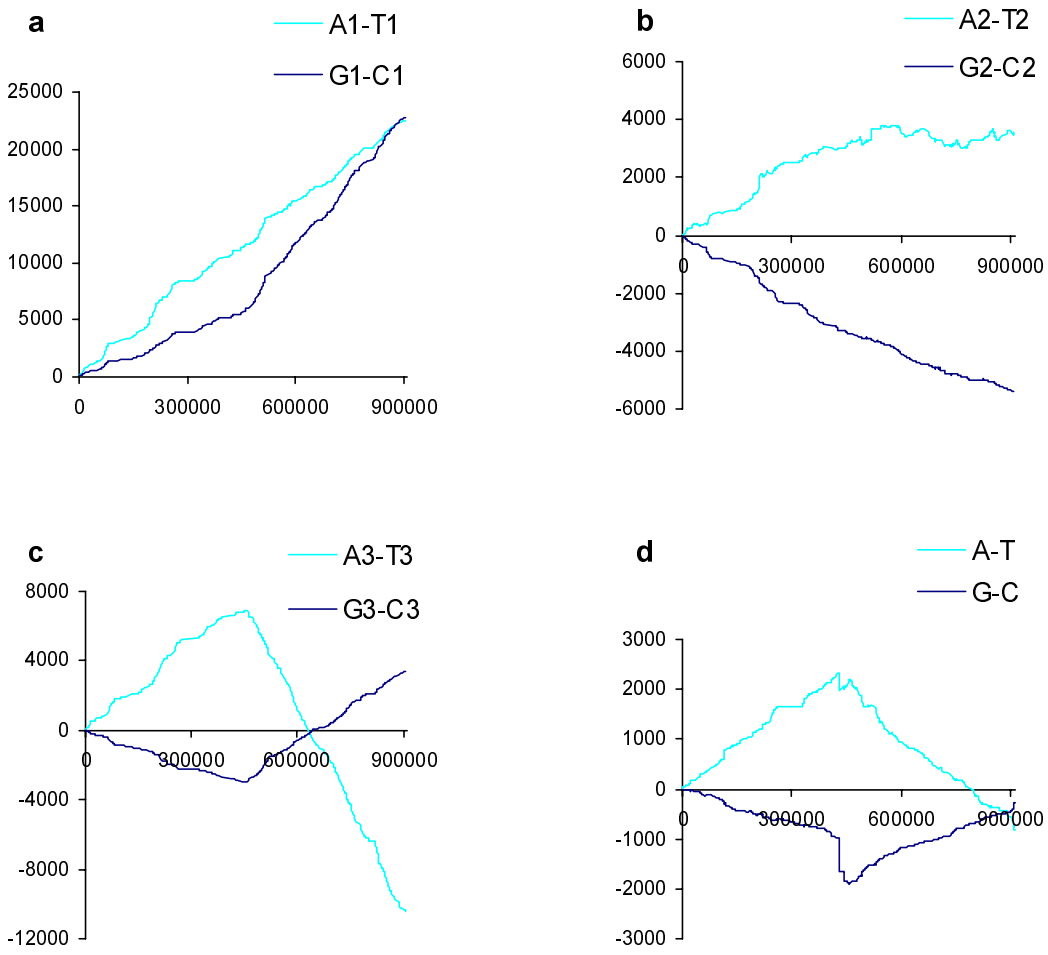


Figure 5. Cumulated values of differences [A]-[T] and [G]-[C] for consecutive protein coding sequences of the Watson strand of the *B. burgdorferi* chromosome; **a)** walks on first positions in codons, **b)** second positions in codons, **c)** third positions in codons. **d)** consecutive intergenic sequences from W strand. Y-axis indicates location of fragments on chromosome.

In the next step of analysis of ORFs, each jump of the walker is corrected to finish the whole walk at y=0 to bring out local trends (see chapter 3.3.1). In this way the strong coding trends are eliminated. The walks on ORFs from Crick strand have been subtracted from walks on ORFs from Watson strand, and also normalised by the length to enable comparing different sequences. Figure 6a-d shows walks on differences [A]-[T] and [G]-[C] done on each codon position of all ORFs and on intergenic sequences. Figure 7a-d shows the same kind of walks done for each of the four nucleotides separately.

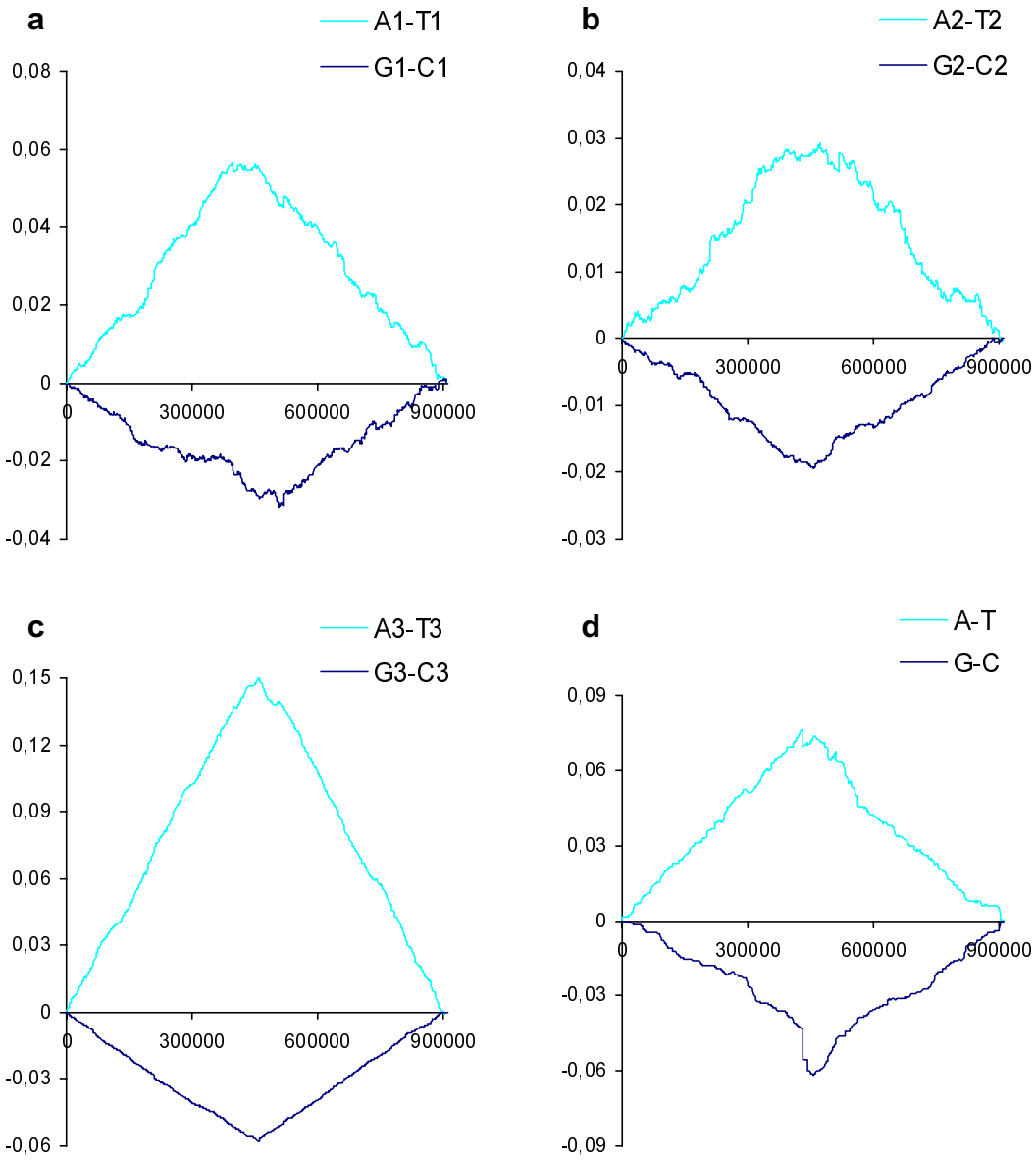


Figure 6. Subtracted DNA walks on the differences [A]-[T] and [G]-[C]; **a)** walks on the first positions in codons in all ORFs, **b)** second positions in codons, **c)** third positions in codons, **d)** intergenic sequences. Y-axis indicates location on chromosome.

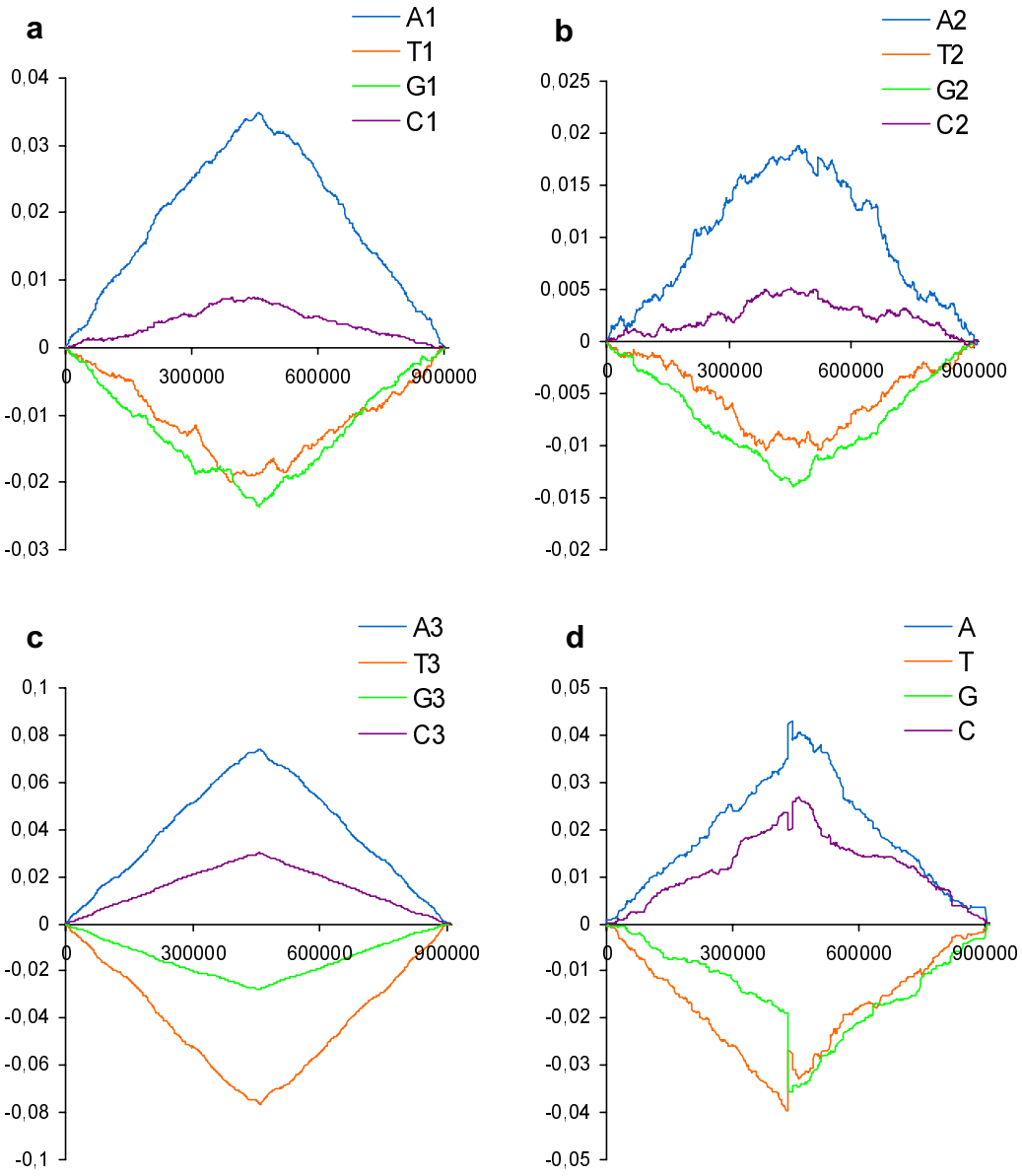


Figure 7. Subtracted DNA walks on the particular nucleotides [A], [T], [G], and [C]; **a)** walks on the first positions in codons in all ORFs, **b)** second positions in codons, **c)** third positions in codons, **d)** intergenic sequences. Y-axis indicates location on chromosome.

Subtraction of walks done on ORFs or sequences read alternately from Watson and Crick strands further intensifies the leading/lagging trends. Now the distinct asymmetry differentiating leading and lagging sequences is visible in each codon position.

The walk on intergenic sequences (Fig. 6d) looks similar to the one for the whole sequence (Fig. 3a), which also clearly shows that the asymmetry is a result of replication and not transcription-related processes.

Results of addition are presented on Fig. 8-9. The asymmetry observed here is a result of effects which have the same influence on both leading and lagging strands, like transcription and coding functions. That is why addition was not done for intergenic sequences. In coding sequences this asymmetry is negligible.

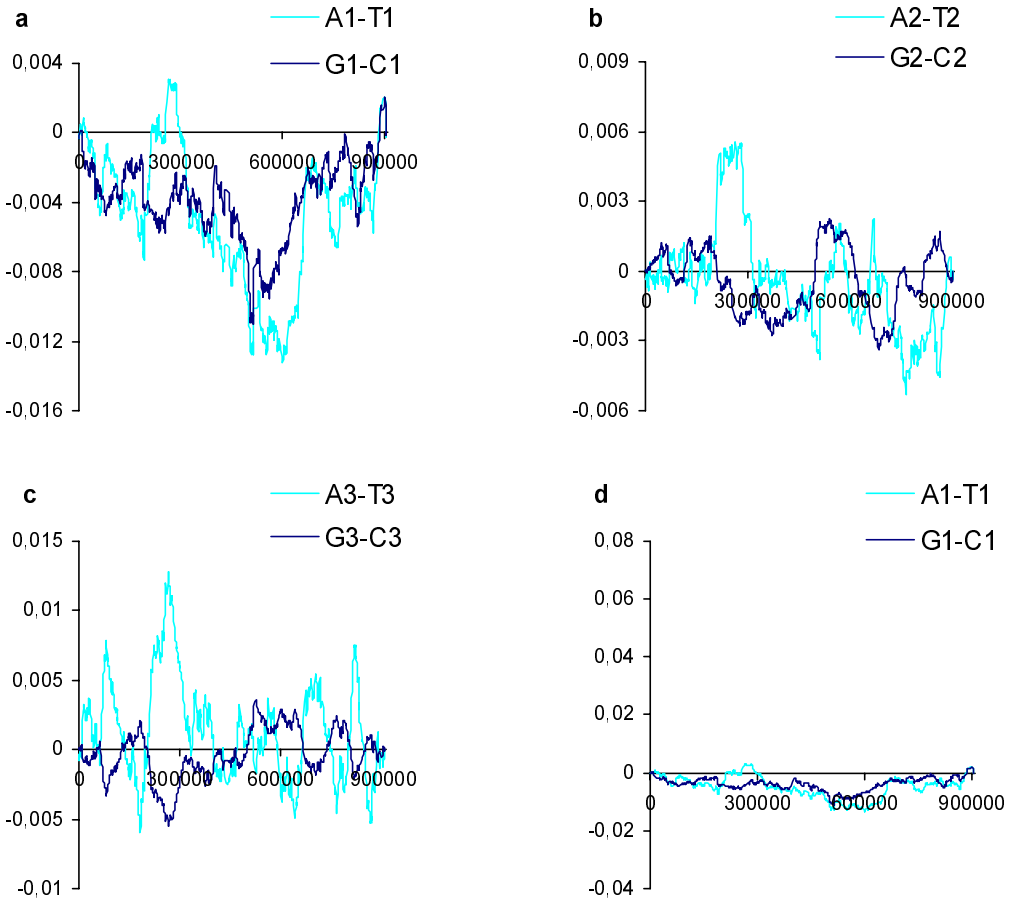


Figure 8. Added DNA walks on the differences [A]-[T] and [G]-[C]; **a)** walks on the first positions in codons, **b)** second positions in codons, **c)** third positions in codons, **d)** chart a) in the scale of Fig. 6a, i.e. subtraction of walks on first positions in codons. Y-axis indicates location on chromosome.

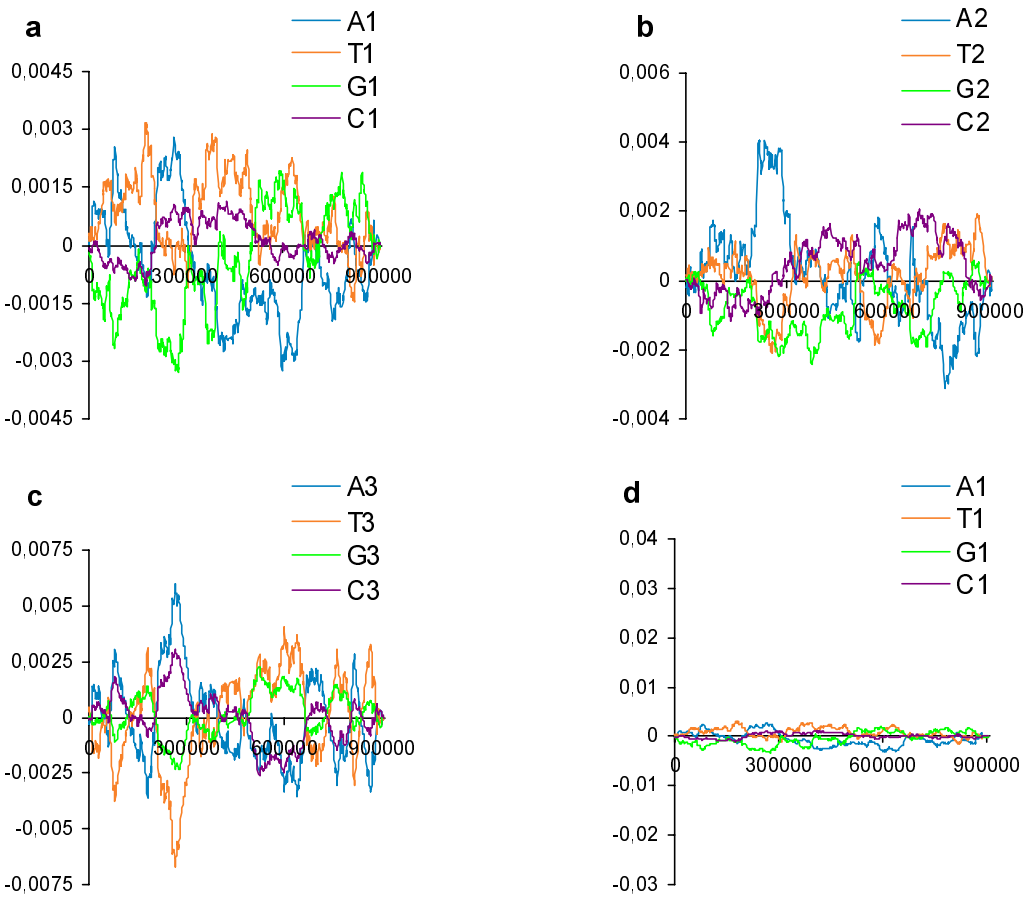


Figure 9. Added DNA walks on the particular nucleotides [A], [T], [G], and [C]; **a)** walks on the first positions in codons, **b)** second positions in codons, **c)** third positions in codons, **d)** chart a) in the scale of Fig. 7a, i.e. subtraction of walks on first positions in codons. Y-axis indicates location on chromosome.

4.1.3. Spider analysis of first, second and third positions in codons

Nucleotide preferences specific for protein coding sequences are best to be analysed with the so-called spiders. Figure 10a-b shows spider walks performed on two genes, located on leading and lagging strands, respectively, of the *B. burgdorferi* chromosome. The nucleotide preferences in each position in the codon are very apparent, because each spider leg goes in a different direction. The walks on first positions are similar for both strands, and show preference for guanine and adenine. Second positions show preference for C and A. Third codon positions of ORFs from leading and lagging strands show opposite trends, because selection pressure on them is the weakest. They show asymmetry introduced by replication-associated mutational pressure, which is of opposite sign on leading and lagging strands (MACKIEWICZ et al. 1999c).

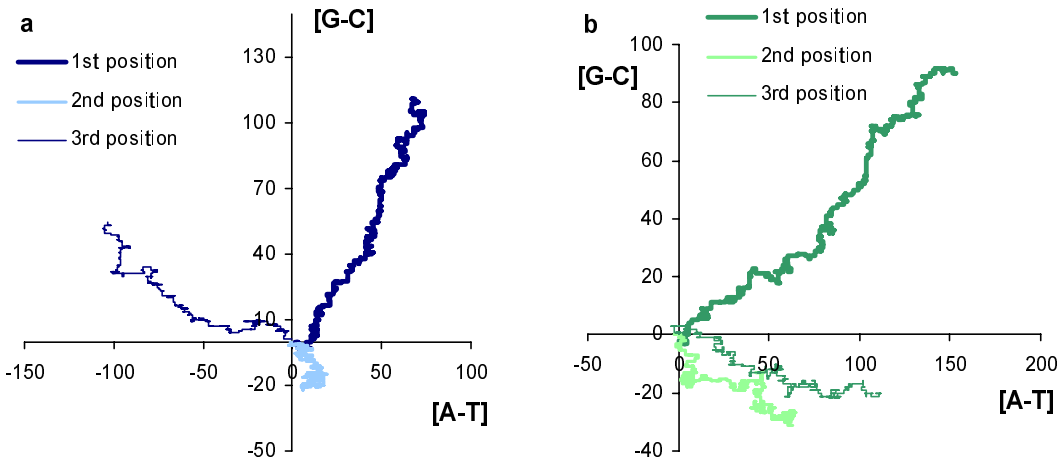


Figure 10. Spider walks on three positions in codons of two ORFs from the *B. burgdorferi* chromosome, **a)** a leading strand ORF BB0020, 556 codons long; **b)** a lagging strand ORF BB 0040, 598 codons long.

A spider done for an intergenic sequence read from the leading strand (Fig. 11a) shows no triplet structure of the sequence, but only abundance of G and T. Some intergenic sequences, however, have retained triplet structure and resemblance to protein coding ORFs (Fig. 11b). Note that all the legs are tilted towards the intergenic sequences trend. Such intergenic sequences are probably derived from coding sequences that were duplicated, lost their function, and began to accumulate nucleotide substitutions typical for their current strand.

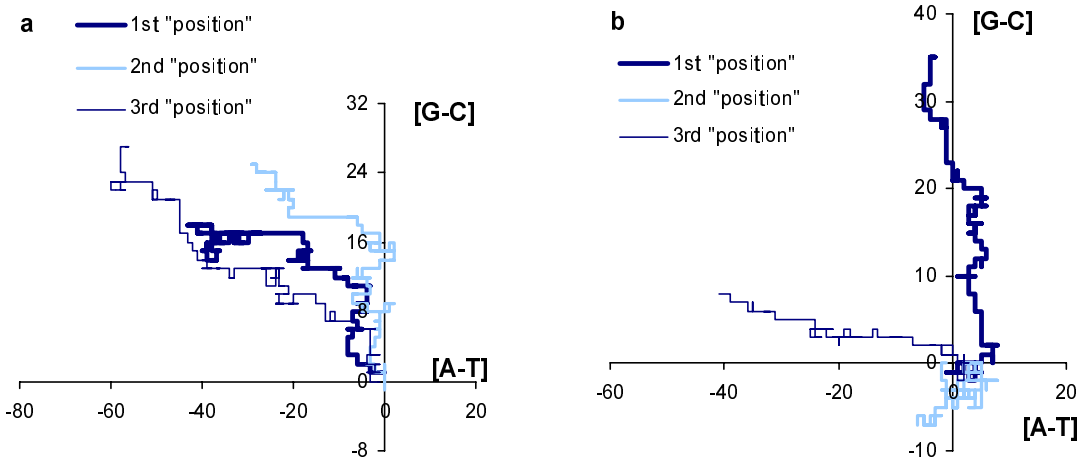


Figure 11. Spider walks on two intergenic sequences read from the leading strand, **a)** located between ORFs BB0472 and BB0473, 220 triplets long; **b)** located between ORFs BB0521 and BB0522, 139 triplets long.

The same trends, but considerably stronger, are visible on the level of the whole chromosome. Fig. 12 shows spider legs for the spliced sequences of all leading- and lagging strand ORFs as well as mirror walks on intergenic sequences read from the leading and lagging strands. The walks have been normalised by the length to compare different sequences.

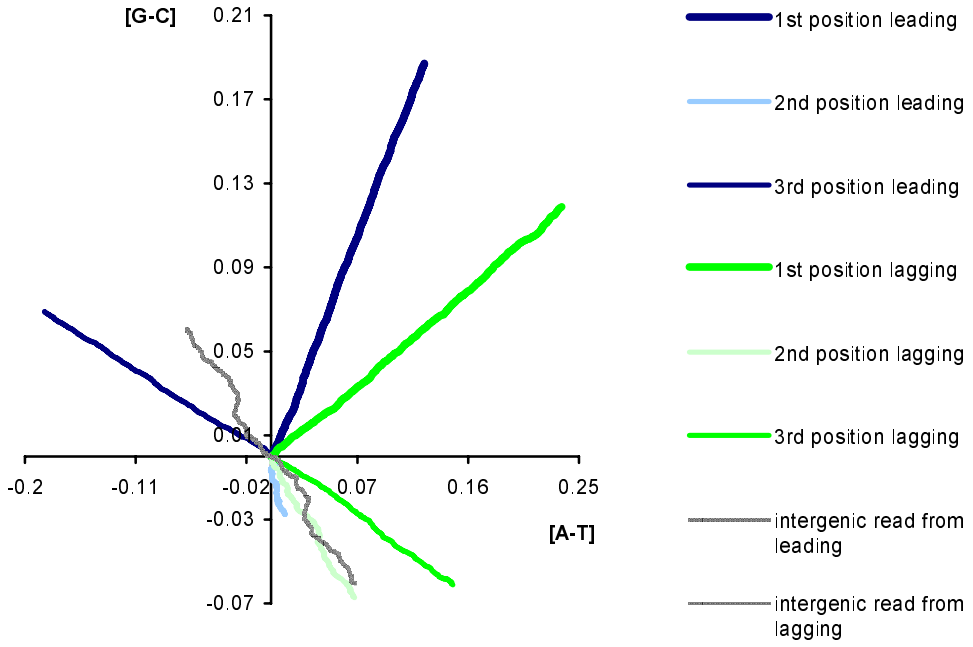


Figure 12. Spider walks on first, second and third positions in codons in spliced ORFs located on leading strands (blue) and lagging strands (green), and mirror walks on intergenic sequences, read from leading strands (blue) and from lagging strands (green).

4.1.4. Distributions of ORFs on torus surface

Similarities between ORFs can be also visualised as their distributions on the surface of the torus. Fig. 13a shows distribution of all ORFs, arctan([A-T]/[G-C]) for the first versus second positions in codons, and Fig. 13b shows first versus third positions. ORFs that group in specific regions of the graph have similar asymmetry. On the latter graph ORFs from the leading and lagging strands form two distinct sets. It is enough to measure asymmetry in nucleotide composition to be able to discriminate between ORFs from leading and lagging strands.

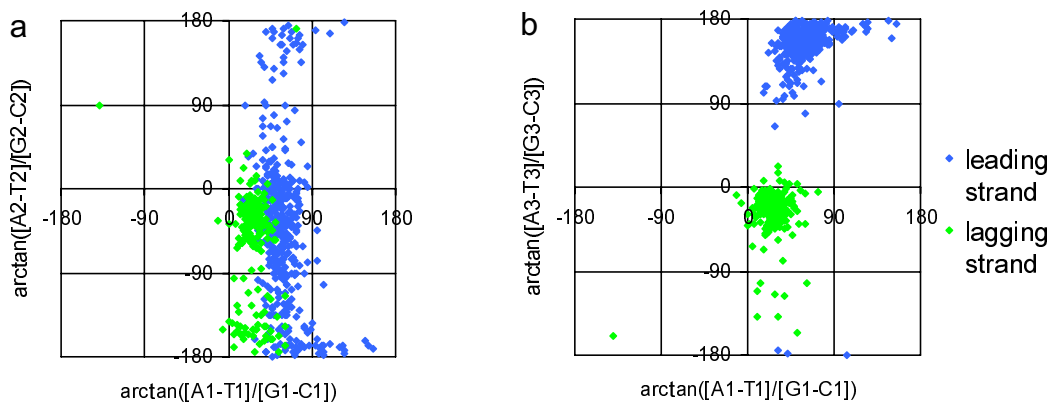


Figure 13. Distributions of ORFs on torus surface. Each ORF is represented by a point with co-ordinates of values of asymmetry (or angles of spider legs) of **a)** first versus second positions in codons; **b)** first versus third positions in codons

4.1.5. Coding density

DNA walks can be used to analyse distribution of ORFs on the chromosome. The walker analyses Watson and Crick strands separately. It moves up if the analysed nucleotide belongs to a coding sequence on the given strand, and down when the analysed nucleotide is between ORFs on the analysed strand. When walks done for Watson and Crick strand are subtracted, the resulting graph shows differences in the number of ORFs between the strands (Fig. 14a). In the first half of the chromosome there are more ORFs on the Crick strand, while in the second half – on the Watson strand. These halves of the chromosome are the leading strand, where the majority of ORFs are located in most bacterial genomes. When the walks are added, the graph shows differences between different regions of the chromosome (Fig. 14b).

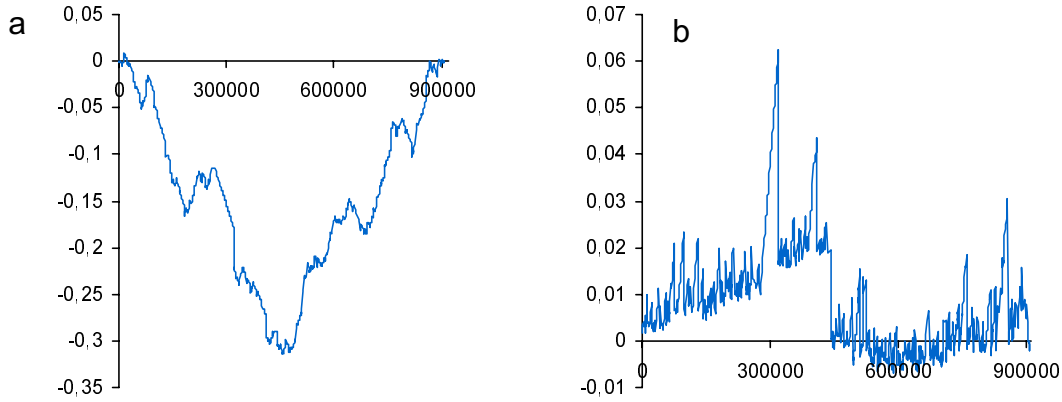


Figure 14. Analysis of coding density (proportion of the number of nucleotides within coding sequences to whole chromosome); **a)** subtraction of walks on Watson and Crick strands, **b)** addition of walks.

4.1.6. Codon composition of genes

As has been shown above, the leading – lagging asymmetry is present in each position in the codon, and is strongest in the third position. Transitions in this position do not change the sense of the encoded amino acid (with two exceptions, see the table of the genetic code, page 17), however transversions in two-fold degenerated codons do. [P]-[Y] walks on the third positions of two-fold degenerated codons in ORFs from Watson strand were subtracted from similar walks done on ORFs from Crick strand. Respective walks were done for four-fold degenerated codons. The resulting graph shows the difference in the occurrence of purines and pyrimidines, thus the difference in the number of accepted transversions in third positions in codons (Fig.15).

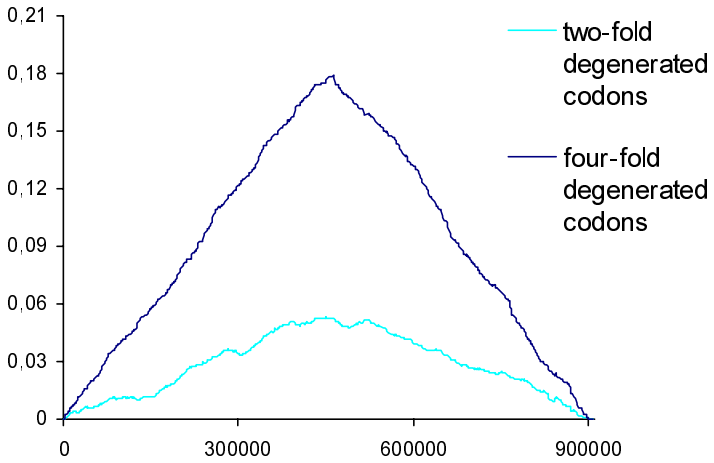
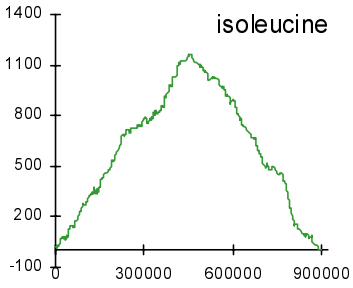
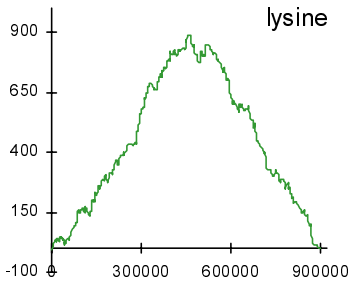
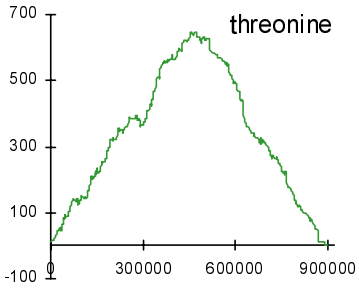


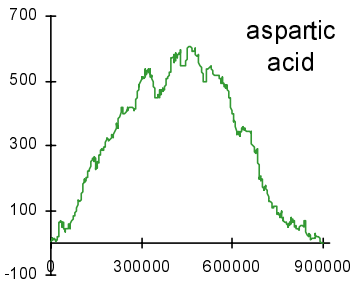
Figure 15. Walks on differences of purines and pyrimidines [P]-[Y] on the third positions in codons, done of two-fold and four-fold degenerated codons separately.

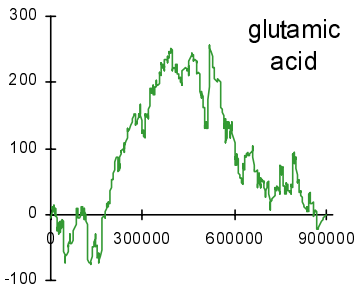
The leading/lagging asymmetry is seen also on the level of amino acid composition of proteins. Subtracted walks on groups of synonymous codons (for the same amino acid) show asymmetric distribution of most of them (Fig. 16). Generally, G and T-rich codons prevail on the leading, and A and C-rich ones on the lagging strand.

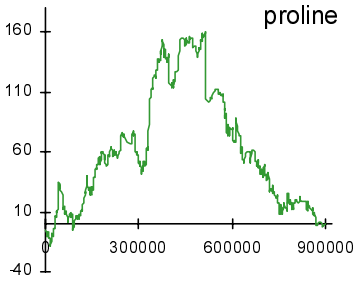

isoleucine
1400
1100
800
500
200
-100
0
300000
600000
900000

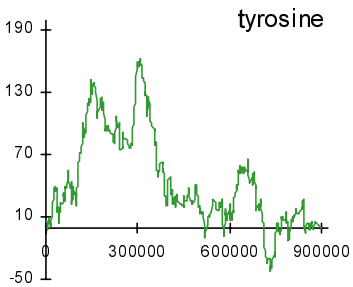

lysine
900
650
400
150
-100
0
300000
600000
900000

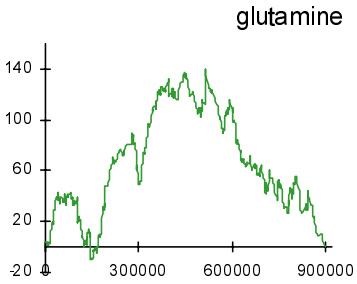

threonine
700
500
300
100
-100
0
300000
600000
900000

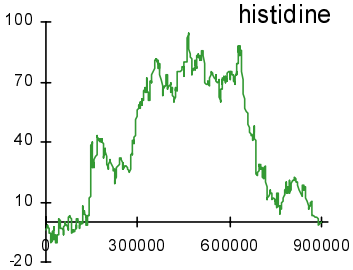

aspartic
acid
700
500
300
100
-100
0
300000
600000
900000

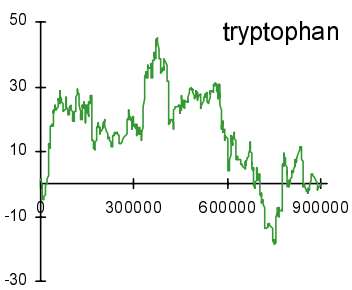

glutamic
acid
300
200
100
0
-100
0
300000
600000
900000


proline
160
110
60
10
-40
0
300000
600000
900000


tyrosine
190
130
70
10
-50
0
300000
600000
900000


glutamine
140
100
60
20
-20
0
300000
600000
900000


histidine
100
70
40
10
-20
0
300000
600000
900000


tryptophan
50
30
10
-10
-30
0
300000
600000
900000

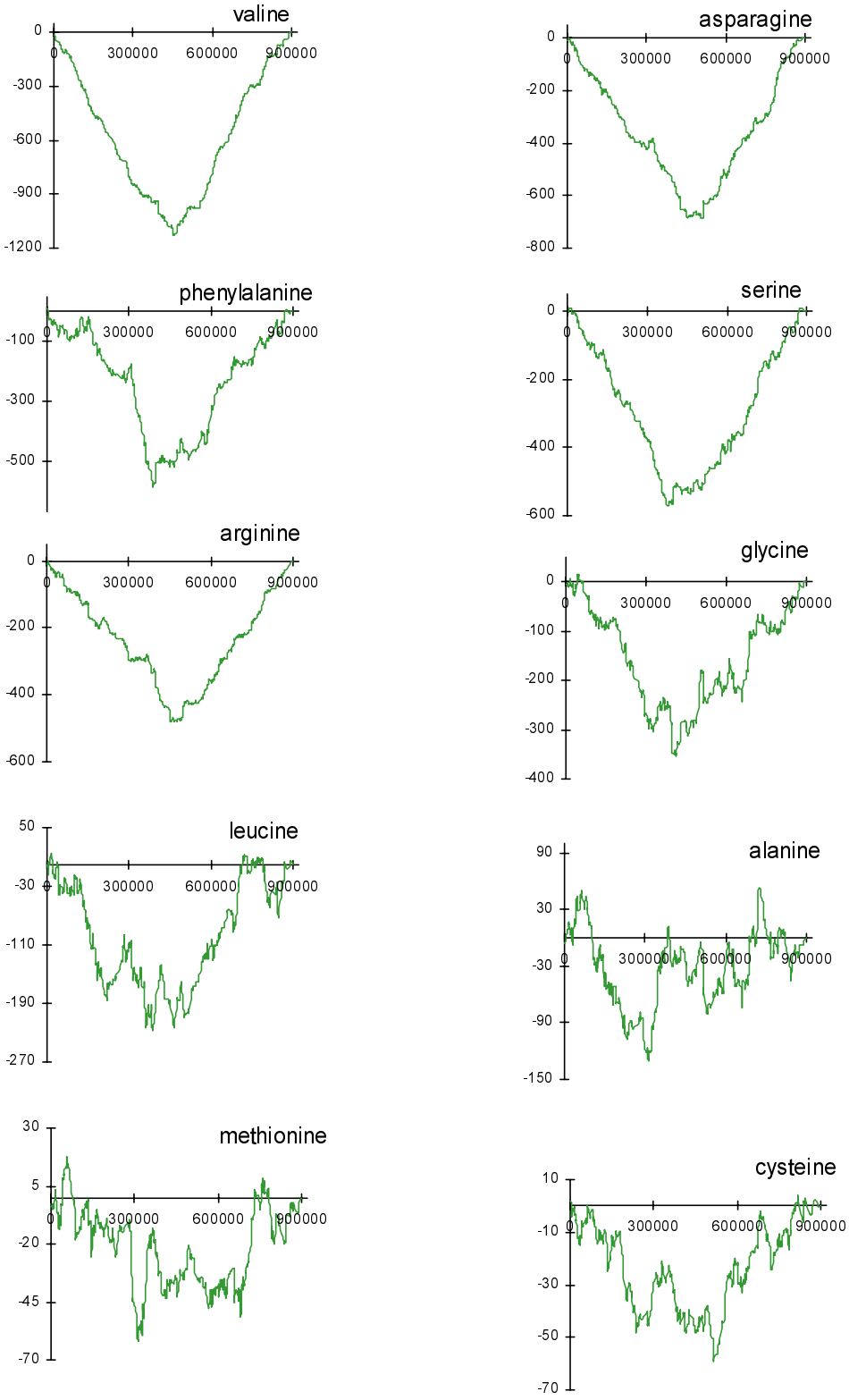


Figure 16. Subtracted walks on synonymous codons of the twenty amino acids. Each chart shows distribution of codons coding for a given amino acid. G and T-rich codons are used more often on the leading strand, and C and A-rich ones on the lagging strand. Thus, most amino acids are preferentially encoded in specific regions of chromosome.

4.2. Analysis of the table of substitutions

4.2.1. *Borrelia burgdorferi* table of substitutions (BbTs)

The most interesting result of the present work is the empirical model, or table of frequencies of the twelve kinds substitutions. The frequencies comprise the mutational pressure which is exerted on the leading strand of the chromosome during replication.

From: \ To:	A	T	G	C
A	-	0.103	0.067	0.023
T	0.065	-	0.035	0.035
G	0.164	0.116	-	0.015
C	0.070	0.261	0.047	-

Table 2. *Borrelia burgdorferi* table of substitutions (BbTs). Frequencies of substitutions in the leading strand of the *B. burgdorferi* chromosome. All frequencies sum up to 1.

4.2.2. Analytical studies

The obtained table of frequencies of the twelve kinds of substitutions (BbTs, Tab. 2) can be quickly tested analytically. The question is how nucleotide composition of a sequence changes under the influence of the table. First, equimolar composition (A:T:G:C = 0.25:0.25:0.25:0.25) was gradually altered accordingly to the frequencies of substitutions in the BbTs. The changes of nucleotide composition in time are shown in bold lines in Fig. 17. Also, nucleotide composition of the third positions in codons of the leading strand genes of *B. burgdorferi* was put under the mutational pressure of the table. The results are shown in fine lines in Fig. 17. After a number of steps, the originally equimolar sequence has the composition of the sequence of the third codon positions, while the composition of the sequence of the third positions remains unchanged.

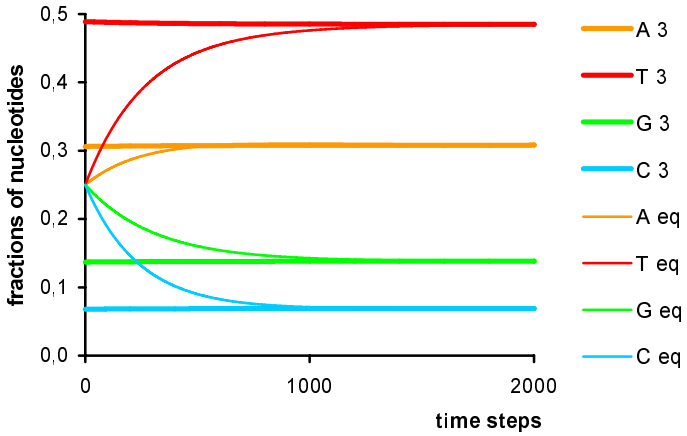


Figure 17. Analytical studies of the influence of the table of the 12 substitutions (BbTs) on fractions of nucleotides in a sequence. A3, T3, G3 and C3 (bold lines) are the fractions of respective nucleotides in third positions in codons of the *B. burgdorferi* leading strand ORFs. A eq, T eq, G eq and C eq (fine lines) are fractions of nucleotides in equimolar sequence (initially A=T=G=C). In time, frequencies of nucleotides in the equimolar sequence change into frequencies typical for the third positions in codons, while the third positions remain unchanged.

4.2.3. Computer simulations

The table of substitutions was further tested in computer simulations, using software written by M. R. Dudek. The sequence was analysed nucleotide by nucleotide, and nucleotides were chosen for mutation with a probability p_{mut}. After a nucleotide was chosen, it was substituted with the probability dictated by the BbTs. Thus, not every chosen nucleotide was substituted. By this method, it is possible to count each substitution that took place during the consecutive Monte Carlo Steps (MCS) of the evolution of the sequence, as well as the differences accumulated in each step between the mutated and the original sequences. Two sequences were analysed: a computer-generated random DNA sequence with equimolar nucleotide composition, and the sequence of spliced ORFs from the leading strand. These sequences were put under the mutational pressure of BbTs. Changes in the nucleotide composition of first, second and third codon positions of the ORF sequence during the computer simulations are shown bold lines in Fig. 18a-c, and the equimolar sequence in Fig. 18d. Composition of third positions in codons of the leading strand ORFs is shown in fine lines for comparison.

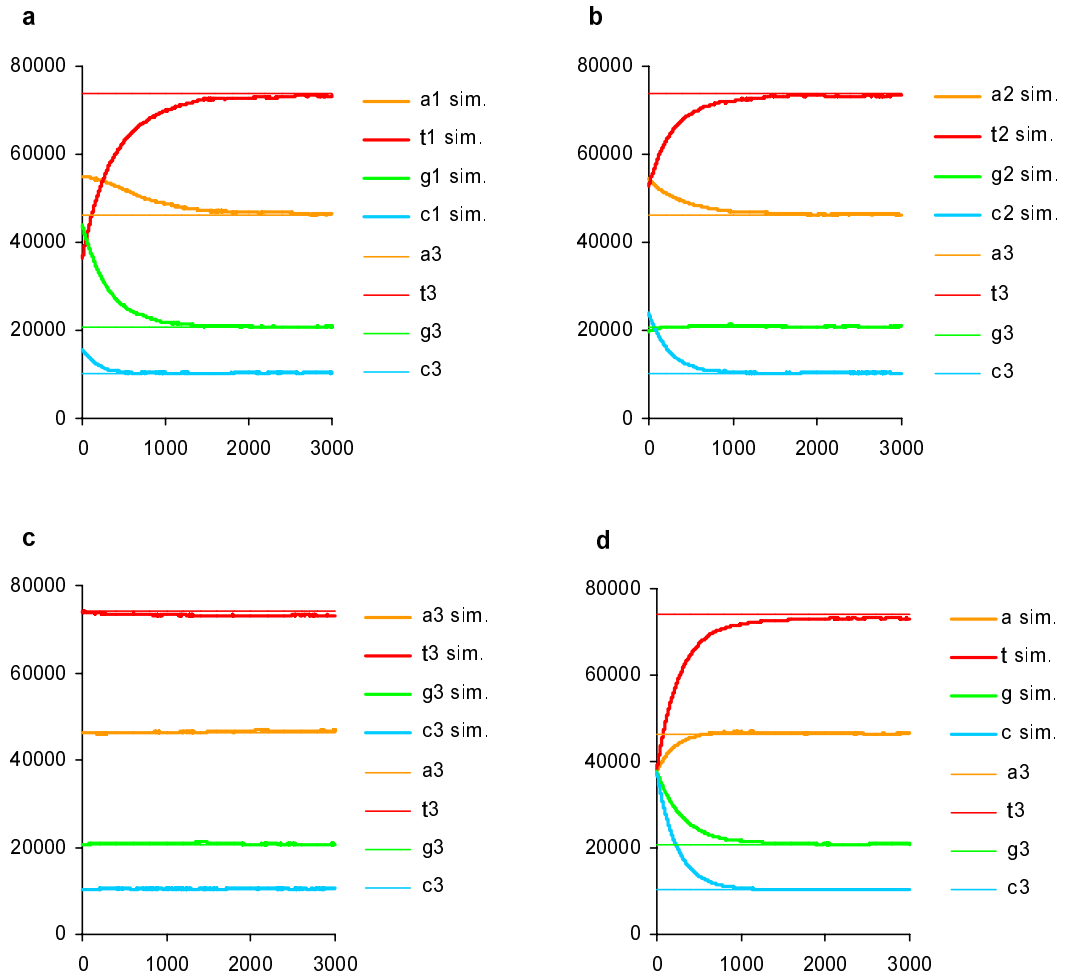


Figure 18. Results of simulations of evolution of DNA sequences under the influence of BbTs (bold lines); **a)**, **b)**, **c)** numbers of A, T, G, and C in first, second and third codon positions of spliced leading strand ORFs from the *B. burgdorferi* genome, **d)** evolution of equimolar sequence of the length of one third of the spliced ORFs sequence. Fine lines show the nucleotide composition of the third positions in codons in the leading strand ORFs. Y-axis shows the fraction of a given type of nucleotide during 3000 generations (X-axis).

The results are similar to the analytical analyses. After a sufficient number of MCS, the sequences are in equilibrium with the mutational pressure and they reach the nucleotide composition dictated by the pressure, which is the composition of the third positions in codons of the leading strand ORFs.

However, in computer simulations the whole sequence is analysed and not only fractions of nucleotides. Thus, the exact number of substitutions can be determined and frequencies of all types of substitutions accepted in the sequence can be calculated. Fig. 19a-d shows the numbers of all types of substitutions that occurred during the simulation.

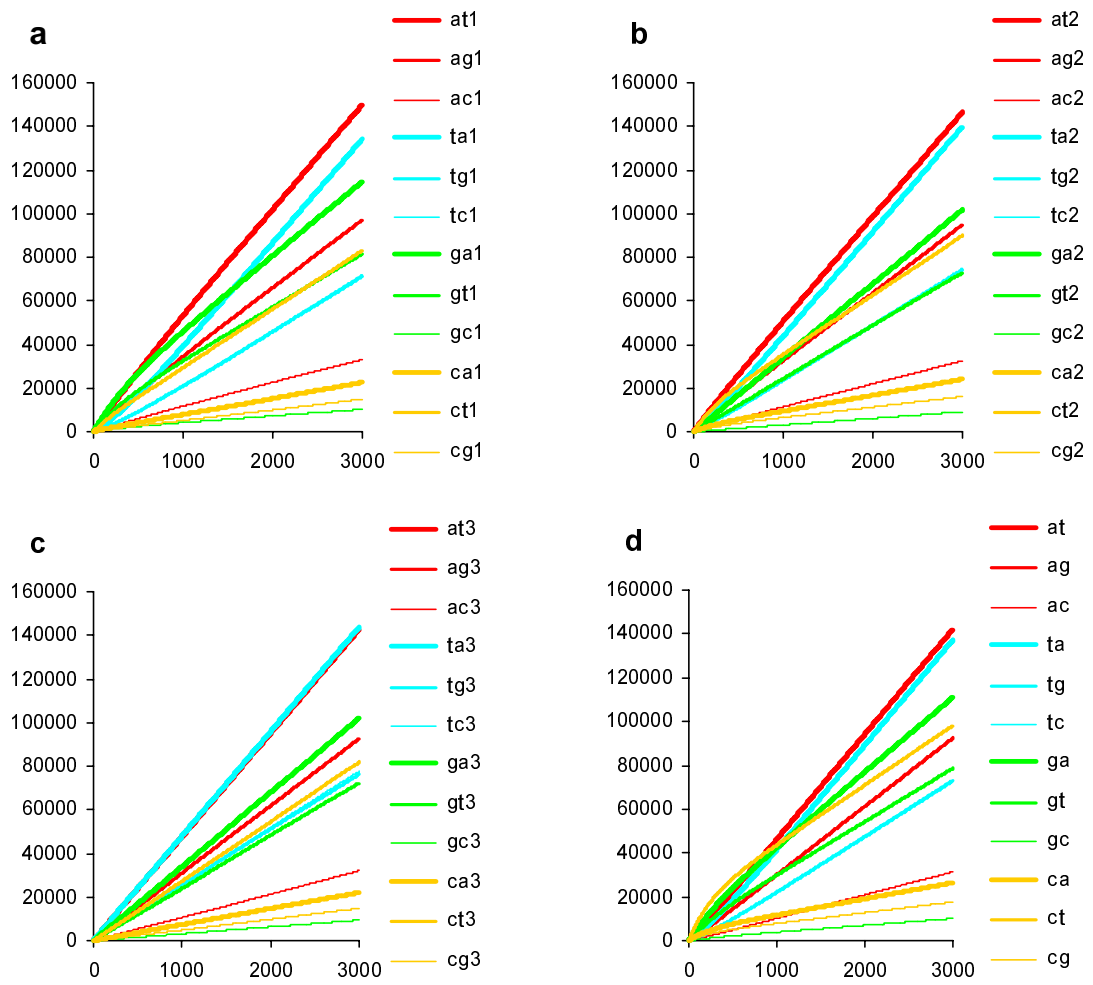


Figure 19. Numbers of all types of substitutions that occurred during 3000 Monte Carlo Steps (MCS) of the simulation. In each step the mutated sequence was compared to the sequence from the previous step. Y-axis shows the cumulated number of a given type of substitution during 3000 generations (X-axis). The numbers were calculated separately for **a)** first, **b)** second, **c)** third positions in codons in the leading strand ORFs in the *B. burgdorferi* genome, **d)** equimolar sequence.

The above figure was obtained by comparing the mutated sequence to the sequence from the previous evolution step. In this way each substitution was counted. The number of substitutions depends not only on the frequencies dictated by BbTs but also on the nucleotide composition of the mutated sequence. Fig. 20a-d was obtained by comparing the mutated sequence to the original sequence (before the simulations). In this way only the accumulated mutations were counted, excluding multiple substitutions and reversions. The rate of accumulation is different for each type of substitution and each position in the codon. When the sequence reaches equilibrium with the mutational pressure, the number of different sites between it and the original is constant; thus it seems to stop accumulating substitutions, although rate of mutation does not change.

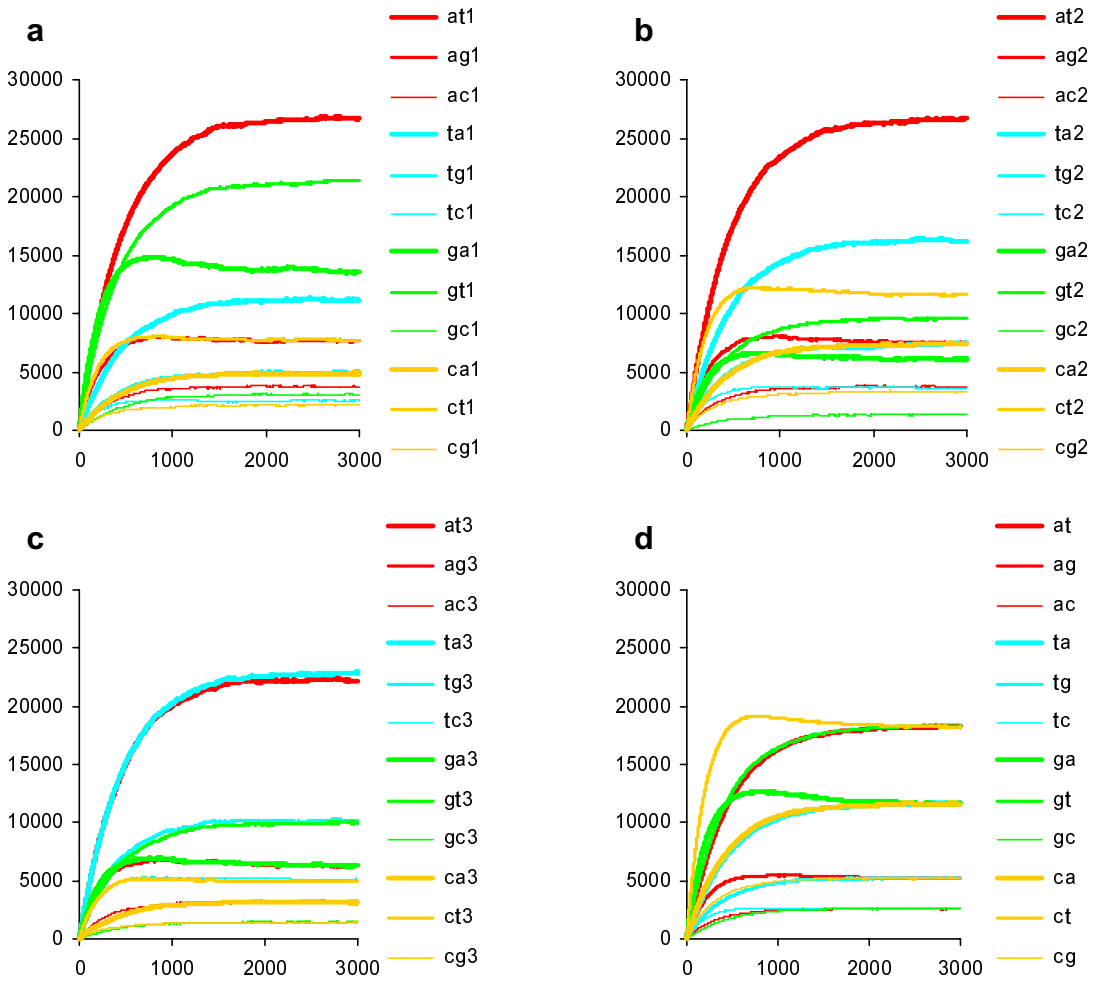


Figure 20. Accumulated substitutions. In each step, the mutated sequence was compared to the original one and thus the number of substitutions was counted. Y-axis shows the cumulated number of a given type of substitution during 3000 Monte Carlo Steps (MCS) of simulation generations (X-axis). The numbers were calculated separately for **a)** first, **b)** second, **c)** third positions in codons in the leading strand ORFs, **d)** equimolar sequence.

Fig. 21 shows changes in evolution time of the total number of mutations, accepted mutations and accumulated substitutions corrected with Kimura's formula for multiple substitutions and reversions:

K = -Ln(1-D-(D*D)/5)

where D is the observed distance and K is the corrected distance (KIMURA 1983).

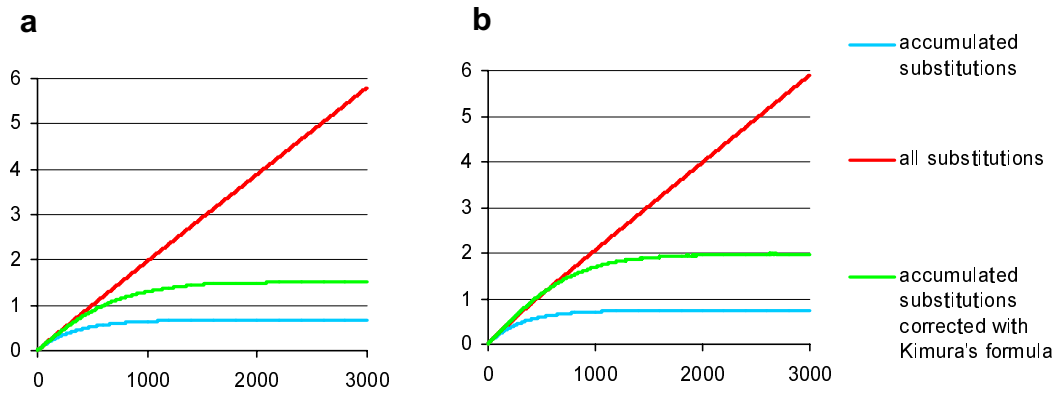


Figure 21. Changes in the number of substitutions during 3000 MCS **a)** evolution of the sequence of the leading strand ORFs, **b)** equimolar sequence of the same length.

4.2.4. Properties of BbTs

During evolution, substituted nucleotides disappear in the similar manner as radioactive isotopes. From the table of substitutions, half-times of substitution τ_A, τ_G, τ_T, τ_C can be calculated for each of the four nucleotides *A*, *G*, *T*, and *C*, respectively. This time is determined by the sum of probabilities of substitutions of a given nucleotide by the other three nucleotides, for example for adenine:

$$\tau_A = ln2/(p_{mut}*(p(A\rightarrow G)+p(A\rightarrow T)+p(A\rightarrow C))$$

where p_{mut} is a parameter which denotes the overall rate of mutation and does not influence the ratios between τs for different nucleotides. In the equilibrium, the fraction of a nucleotide which has been substituted is exactly the same as the fraction of this very nucleotide substituting the other ones. Thus, after the half time of substitutions the ratio between the "old" nucleotides and "new" nucleotides is 1:1 and this is a general property of any table of substitutions (Fig. 22a and b).

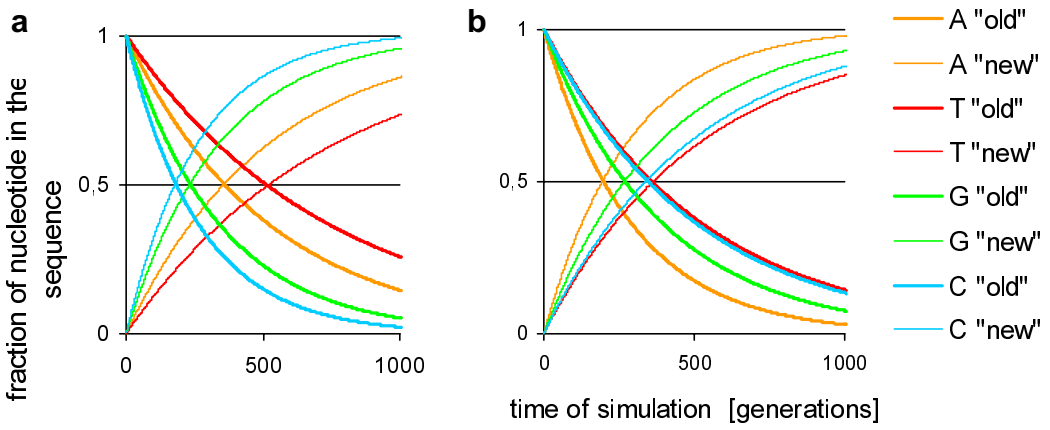


Figure 22. Half-time of substitutions. Bold lines show decreasing fractions of the original nucleotides, fine lines show increasing fractions of newly appearing nucleotides during 1000 MCS **a)** under the influence of BbTs **b)** under the influence of an artificial, computer-generated table of substitution which gives the same nucleotide composition as BbTs.

However, the table of substitutions obtained for *B. burgdorferi* has another interesting feature: the time when a half of nucleotides of a given type are substituted by other nucleotides is linearly correlated with the fraction of the analysed type of nucleotide in the sequence. The higher substitution turnover of a nucleotide, the lower the fraction of this nucleotide in the DNA sequence (Fig. 23a). It seems to be a property of the pure mutational pressure. An artificial, computer-generated table of substitutions by KOWALCZUK et al. (1999c) imposed asymmetry on DNA sequence but there was no correlation between fractions of nucleotides in that sequence and the rate of their substitution (Fig. 23b).

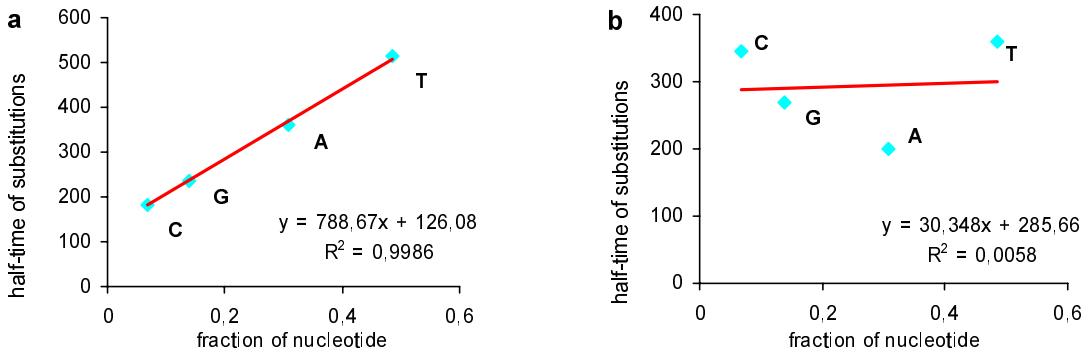


Figure 23. Correlation between the fraction of a nucleotide (X-axis) and its half-time of substitution (Y-axis) under the influence of **a)** BbTs, **b)** artificial BbTs.

Also, a table of substitution rates for sequences under strong selection (ZHANG 1999) gave no correlation (Tab. 3). However, correlation was found for substitution matrices which were obtained for sequences free from selection pressure, for example LI et al. 1984, YANG 1994, and FUKASAWA et al. 1982 (Tab. 3).

substitution	BbTS real	BbTS artifi-cial	pseud 1	pseud 2	pseud m	pseud h	mt dros 4d	mt dros 3p	mt DNA 1, 2p
A→T	0,103	0,222	0,047	0,024	0,017	0,087	0,124	0,019	0,069
A→G	0,067	0,111	0,050	0,138	0,068	0,149	0,013	0,148	0,082
A→C	0,023	0,016	0,094	0,031	0,068	0,000	0,029	0,005	0,094
T→A	0,066	0,189	0,044	0,024	0,057	0,025	0,004	0,021	0,054
T→G	0,035	0,002	0,082	0,030	0,114	0,013	0,075	0,005	0,016
T→C	0,035	0,002	0,033	0,126	0,095	0,025	0,379	0,086	0,182
G→A	0,164	0,023	0,210	0,214	0,166	0,177	0,056	0,449	0,118
G→T	0,116	0,178	0,072	0,047	0,030	0,076	0,000	0,041	0,030
G→C	0,015	0,056	0,053	0,051	0,060	0,050	0,000	0,000	0,031
C→A	0,070	0,186	0,065	0,052	0,076	0,052	0,288	0,000	0,087
C→T	0,261	0,010	0,210	0,211	0,172	0,312	0,000	0,227	0,216
C→G	0,047	0,004	0,042	0,054	0,076	0,035	0,124	0,000	0,020
DNA composition									
A	30,8	30,8	31,3	30,8	41,0	17,0	27,4	29,8	24,7
T	48,5	48,5	37,2	30,8	17,1	65,6	50,1	43,6	31,5
G	13,8	13,8	15,8	19,9	23,9	11,7	8,3	9,4	17,2
C	6,9	6,9	15,8	18,5	18,1	5,6	14,3	17,2	26,6
Half-time of substitution									
τ_A	361	199	364	360	450	294	412	404	283
τ_T	513	359	437	385	262	1103	637	622	274
τ_G	236	269	207	223	270	229	159	141	388
τ_C	183	346	219	219	214	174	240	306	214
Correlation	**0,999**	**0,072**	**0,998**	**0,991**	**0,967**	**0,998**	**0,997**	**0,988**	**-0,764**

Table 3. Examples of tables of substitutions, DNA composition in equilibrium with the mutational pressure and half times of nucleotide substitutions.

Explanations: BbTs real – table of substitutions in the *B. burgdorferi* genome on the leading strand estimated as described in the Materials and Methods section; BbTs artificial – one of the computer-generated tables which produce the DNA asymmetry and composition like BbTs real; pseud 1 –data for mammal pseudogene sequences (LI et al. 1984); pseud 2 – data for the psi-eta-globin pseudogenes of primates (YANG 1994); pseud m and pseud h – data for LDH-A pseudogenes of mouse and human respectively (FUKASAWA et al. 1986); mtdros 4d and mtdros 3p – data for cytochrome b and NADH dehydrogenase subunit 1 genes of *Drosophila* mtDNA for four-fold degenerate sites and the third codon positions (TAMURA 1992); mtDNA 1, 2p - data for the first and the second codon positions for vertebrate mitochondrial genes (ZHANG 1999) Notes: The last column represents data for substitutions in mitochondrial sequences under strong selection pressure; p_{mut} parameter for τ counting equals 0,01. For more explanations see text.

5. Discussion

5.1. Methods of measuring and showing DNA asymmetry

DNA asymmetry, or deviations from PR2, are usually analysed in terms of excess of the number of guanines relative to cytosines or adenines relative to thymines. The bias is measured by GC and AT skews, (G-C)/(G+C) and (A-T)/(A+T), respectively. The method of analysing GC and AT skews with a sliding window (LOBRY 1996a) is helpful to detect replication origin in some prokaryotic chromosomes but the results are often difficult to interpret. If a small-sized window is used, strong fluctuations obscure the asymmetry (MRAZEK, KARLIN 1998), if the window is large, the trends in nucleotide composition are diminished (MCLEAN et al. 1998), cf. Fig.1. Cumulative skew diagrams or plots of numerically integrated skew of GRIGORIEV (1998) and TILLIER and COLLINS (2000a) eliminate fluctuations and give a much clearer picture (cf. Fig. 2). FREEMAN et al. (1998) performed cumulative diagrams of purine (A+G versus T+C) and keto (G+T versus A+C) excess, which indicated the origin and terminus of replication, and regions of integration of foreign DNA in the eubacterial genomes analysed.

ROCHA et al. (1999a,b) applied a statistical linear discriminant function to assess strand asymmetry at the level of nucleotides, codons and amino acids.

A method which differentiates the influence of replication processes and transcriptional/translational forces on genomic sequences was proposed by TILLIER and COLLINS (2000a). Their ANOVA analyses quantify and measure statistical significance of individual effects of replication and gene direction on GC and AT skews. The skews were measured in each codon position of CDS and in non-CDS separately. They found that the effect of replication orientation is independent of the effects of transcriptional or translational processes and in fact can be of the opposite sign. They also found that AT and GC skews in non-CDS are similar in size and sign to the skews seen with replication orientation at the third positions in codons.

Much more information can be derived from DNA walks (CEBRAT, DUDEK 1998). Walks on differences of [A]-[T] and [G]-[C] give similar results to cumulative diagrams (Fig. 3a) but walks on particular nucleotides reveal their participation in asymmetry (Fig. 3b). Subtracting and adding DNA walks allows to separate the effect of replication-associated processes from the effect introduced by transcription and coding functions (MACKIEWICZ et al. 1999a,b). Subtraction of walks magnifies the trends in nucleotide substitutions which are reciprocal on Watson and Crick strands, as are the

ones connected with replication (Fig. 3), while addition of walks diminishes them and brings out the trends which are of the same sign on W and C strands, resulting from transcription and coding functions (or shows the lack of them, as in the case of the *B. burgdorferi* chromosome, Fig. 4).

The replication-induced asymmetry can be detected in protein coding sequences in each position in the codon (Fig. 6-7), and as a result, this mutational pressure is reflected in amino acid composition of proteins (Fig. 16). The walks on ORF sequences are normalised and presented in the scale of the chromosome, i.e. x-axis shows location of the analysed ORFs on the chromosome, so the observed asymmetry does not result from unequal numbers of ORFs on the leading and lagging strands. Analysis of coding density (subtraction of walks, Fig. 14a) shows that the majority of ORFs are located on the leading strand, which may add to GC and AT skews observed by some other authors. When one looks at both W and C strands at the same time (addition of walks, Fig. 14b), ORFs are distributed more evenly on the chromosome.

Furthermore, "spider" DNA walks can be used to distinguish between coding and non-coding sequences and to indicate the strand and the phase in which DNA is coding (CEBRAT, DUDEK 1998). Most coding ORFs have very strong and specific trends in nucleotide composition of each position in the codon, which can be seen in individual ORFs (Fig. 10) and even more clearly in spliced sequences of all ORFs from leading and lagging strands (Fig. 12). The parameters of the "spider" walks, like their angles and lengths of vectors, have been successfully used to discriminate protein coding from non-coding sequences and to estimate the total numbers of protein coding genes in genomes (CEBRAT et al. 1997b, CEBRAT et al. 1998, KOWALCZUK et al. 1999b). For some genomes, like *B. burgdorferi*, it is possible to determine by nucleotide composition of a gene, which strand, leading or lagging, it is located on (Fig. 13), and even amino acid composition of a protein reveals the strand the corresponding gene is located on. GIERLIK et al. (2000) have used DNA walks to analyse eukaryotic genomes and have been the first to find replication-associated asymmetry at the ends of chromosomes. DNA walks have proven to be the best method to visualise and compare sequence asymmetry.

5.2. The steady state of the B. burgdorferi chromosome

The *B. burgdorferi* chromosome is the most asymmetric of all sequenced so far (see http://smorfland.microb.uni.wroc.pl/bacasym.html for comparison). Subtraction of walks on the chromosome sequence shows extremely strong asymmetry, but addition of walks shows lack of it. The observed asymmetry differentiates sequences from leading and lagging strands (Fig. 3), while addition of walks differentiates ORFs proximal and distal to the origin of terminus of replication in some genomes, e.g. *Bacillus subtilis* (MACKIEWICZ et al. 1999b). However, this kind of asymmetry is absent in the *B. burgdorferi* chromosome (Fig. 4).

Cumulative walks on first and second positions in codons in ORFs show trends specific for the coding functions of these sequences, and independent of their location on chromosome (Fig. 5a-b). Third positions have trends similar to intergenic sequences (Fig. 5c-d). The same conclusions come from the genomic spider walks (Fig.12). Apparently, the most degenerated positions in codons have most adapted to their location on chromosome. However, detrended DNA walks reveal the leading/lagging asymmetry also in the first and second positions in codons (Fig. 6a-b). Selection on these positions is not strong enough to eliminate all asymmetric substitutions that occur during replication. Walks on particular nucleotides show that distribution of each type of nucleotide contributes to the observed asymmetry, although not to the same extent (Fig. 7). Addition of walks does not show any clear trends connected with transcription (Fig. 8a-c and 9a-c), and presented in the scale of subtracted walks shows no asymmetry at all (Fig. 8d and 9d). The asymmetry between leading and lagging strands of the chromosome is a result of substitutions introduced during replication (MACKIEWICZ et al. 1999c), and not transcription, as some authors have argued (BELETSKII, BHAGWAT 1996, FRANCINO et al. 1996, FRANCINO, OCHMAN 1997, FREEMAN et al. 1998).

The asymmetry is greatest in the third position in the codon, which is under weakest selection pressure. In half of the boxes of the table of the genetic code (page 17) any substitution in the third position does not change the sense of the encoded amino acid. Fig. 15 shows that in those positions (in four-fold degenerated codons) asymmetry is the strongest. However, the asymmetry is also present in the third positions of two-fold degenerated codons, which means that some asymmetric transversions that change the encoded amino acid, and thus should be selected against, are nevertheless fixed. The nucleotide composition of the third positions in codons follows precisely the sign of the

asymmetry of intergenic sequences (Fig. 5cd, 6cd, and 7cd) and the third positions of ORFs situated on the leading and lagging strands have precisely mirror asymmetry (Fig. 10, 12). Apparently, replication-associated mutational pressure is the main force that generates the observed asymmetry. Interestingly, the asymmetry is greater in third codon positions than in intergenic sequences. This paradox could be explained assuming that the highly degenerated third positions have accumulated more neutral or near neutral mutations introduced by the replication-associated processes because they stay at their positions for longer than intergenic sequences (Mackiewicz et al. 1999c). There are constraints on inversions of coding sequences but no constraints on inversions of intergenic sequences. Thus, some newly inverted intergenic sequences could complement the asymmetry of the "new host" strand. The third codon positions stay for longer under the influence of the mutational pressure typical for one strand, and because they are not under strong selection, their composition is close to equilibrium.

When asymmetry in the first and third positions in codons is taken into account, genes form two distinct, easily recognisable groups (Fig. 13b), which testifies for a particular conservation of location of genes in the genome (MACKIEWICZ et al. 1999c). 96% of genes coding for ribosomal proteins are located on the leading strand, which suggests that location and rearrangements of genes are under very strict constraints. In fact, the structure of the genome has been conserved for such a long time that asymmetric substitutions had time to accumulate in each position in the codon and influence the amino acid composition of proteins. The conservation may result either from lack of recombination between the strands or from the differential killing of genes relocated to the opposite strand by the replication-associated mutational pressure (MACKIEWICZ 2001a). It seems that genes are "fitted" to their location, where the mutational pressure is optimal, and the genome is in the steady state.

5.3. The B. burgdorferi table of substitutions

Analyses of long-range correlations in DNA have revealed that in the intergenic sequences a very strong triplet signal can be detected (VOSS 1992, GIERLIK et al. 1999). This signal can be created by fragments of coding sequences transferred into intergenic space by recombination mechanisms. Apparently, some intergenic sequences have derived from coding sequences and could freely accumulate mutations with frequencies determined by the replication-associated mutational pressure (see Fig. 11b for an

example). If the time of divergence has not been very long, homology between the intergenic sequences and their original protein coding sequences can be found. In this way the table of substitutions was constructed for the *B. burgdorferi* leading strand ORFs (BbTs, Table 2, see Materials and Methods section for details). An assumption was made that mutations have accumulated only in the intergenic sequences and not in the coding sequences, which is not exactly true, but which enabled constructing the table. This assumption could only lower the real mutational rate without influencing the ratios between the specific substitutions if selection kills mutants evenly with the same probability independently of the kind of substitution (the position in codon does not influence the results). This is a risky assumption but it could give a good approximation of the mutational pressure exerted on intergenic sequences. Some other authors who have constructed matrices of substitutions using the mutations accumulated in pseudogene sequences have made the same assumption (LI et al. 1984, YANG 1994).

The *B. burgdorferi* chromosome was chosen for analysis because there are many premises indicating that it is in the steady state (see above). Third positions in codons have been found in equilibrium with replication-associated mutational pressure by analytical analysis (Fig. 17) and computer simulations (Fig. 18), which show that the composition of these positions does not change under the influence of mutational pressure.

After aligning sequences under study and determining site-by-site homologies and differences between them, one needs to build a mathematical model of the evolution of the sequences in time. There are two kinds of models, or matrices of substitutions: empirical ones, using properties calculated through comparisons of observed sequences, and parametrical ones, using chemical or biological properties of DNA and amino acids (see LIÒ, GOLDMAN 1998, and WHELAN et al. 2001 for review and discussion). The table of substitution rates described in the present work is a phenomenological, empirical one. It is the first table that creates DNA in equilibrium, and of nucleotide composition observed in nature. Contrary to any parametrical model, it retains both DNA sequence composition and the strand asymmetry of the reference sequences, here the third positions in codons of the *B. burgdorferi* leading strand ORFs (Fig. 17, 18). Computer simulations of evolution of the leading strand ORFs under the influence of the table and no selective constraints allow for more than estimation of changes in nucleotide composition. They enable watching the history of the sequence and counting each substitution. The frequencies of substitutions are given in the table, but the number

of substitutions that actually occur also depends on the composition of the analysed sequence. That is why the numbers of substitutions are slightly different for each position in the codon, and for the equimolar sequence (Fig. 19). Much greater differences are observed in the number of substitutions that accumulated in the analysed sequence (Fig. 20). When a DNA sequence is put under a mutational pressure, but free from selection pressure, the number of different sites between this sequence and the original sequence increases. However, substitutions may occur in the same sites and after some time the number of different sites between sequences is constant, although the rate of mutations does not change. The relation between the numbers of accumulated substitutions and all substitutions is shown in Fig. 21. Usually when calculating evolutionary distances, one can compare sequences of living organisms, but ancestral sequences remain unknown. Thus only the number of substitutions accumulated between these sequences is known. To estimate evolutionary distance, it is necessary to find the number of substitutions that actually occurred, or to correct the observed number of substitutions for multiple substitutions and reversions. A way to do that was proposed by KIMURA (1983). Fig. 21 shows that Kimura's correction is accurate only for short evolutionary distances. Ideally, the correction should allow for calculating the actual number of substitutions from the number of accumulated substitutions, so it should be close to the graph showing all substitutions. However, Kimura's correction does not take into account different rates of substitution of each nucleotide, and works only for short evolutionary distances. From Fig. 19 it is clear that the corrections should be different for each type of substitution (KOWALCZUK et al. 2001c). Knowing the table of substitutions, one can count the corrections that should be introduced into experimentally found differences in nucleotide sequences to find the real numbers of substitutions during the divergence time. It is possible to count the corrections very precisely and for a wide degree of homology. Furthermore, it should be possible to separate the effects of mutational and selection pressures.

The empirical table of substitutions allows for calculating half times of substitution for each of the four nucleotides (Fig. 22). What is more, there is a linear evolution law that correlates the fractions of the four nucleotides in the sequence with the rates of their substitution (Fig. 23a). The law holds only for real matrices obtained for DNA in the equilibrium, under only mutational pressure, free from selection (KOWALCZUK et al. 2001b). Computer-generated tables and the ones obtained for sequences under selection pressure do not share that property (see Fig. 23b and Table 3). The matrix found for the

third positions in the four-fold degenerated codons in *Drosophila* mitochondrial DNA (TAMURA 1992) fulfils this law more precisely than for all third positions in codons in that organelle's genome (the same results were obtained for matrices of primates' mtDNA published by ADACHI and HASEGAWA, 1996). These differences could be expected if some mutations in the third positions, leading to amino acid substitutions are not neutral. One can also notice that matrices found by analysis of substitutions into different pseudogenes in the same organism or in very closely related organisms give a different DNA composition in equilibrium, which supports the thesis that the mutational pressure varies for different regions of the same eukaryotic genome (FILIPSKI 1988, WOLFE et al. 1989, MATASSI et al. 1999).

Precise, almost deterministic relations between the nucleotide fractions and their turnover rates enable estimating if the matrix of substitutions is influenced by selection or not. Also, it enables counting the distance between the given sequence and the sequence in equilibrium with this mutational pressure. This distance is supposed to be a measure of selection pressure, which keeps the sequence at the steady state, far from equilibrium. Using the mutational pressure matrix one can estimate the pressure on each position in codons in protein coding sequences.

6. Summary and Conclusions

- In *Borrelia burgdorferi*, location of a gene on the leading or lagging strand of the chromosome influences its nucleotide composition, which is reflected by its codon composition and amino acid composition of the encoded protein.
- By the asymmetry in nucleotide composition of first and third positions in codons ORFs can be divided into two non-overlapping groups located on different DNA strands.
- By comparing gene-derived intergenic sequences to their homologous genes, the frequency of each kind of substitution (BbTs) was found for leading strand sequences free from selection.
- The empirical matrix of substitutions (BbTs) obeys the linear law for the correlation of the half-time of substitution of a type of nucleotide with its fraction in the sequence.
- Basing on this law, it is possible to calculate precise corrections for multiple substitutions and reversions in phylogenetic studies.
- Methods of analysis described in this work enable assessment of relative contribution of mutation and selection forces to the observed asymmetry.
- The chromosome is in steady state with replication-associated mutational pressure and with selection pressure on the encoded information.
- The third positions in codons in the leading strand ORFs are in equilibrium with replication-associated mutational pressure, and the influence of selection is not visible in four-fold degenerated codons.

Future perspectives:

- If these methods are applied to other genomes, they could allow to estimate the differences in mutational pressure between genomes and then to estimate the role of selection on different sequences. It will enable measuring phylogenetic distances more precisely, and constructing more accurate phylogenetic trees.
- Further analysis of the table of substitutions should allow estimating the role of selection and the susceptibility of each position in codons of the coding sequences to mutational pressure. Further studies should also indicate if there is any correlation in elimination of substitutions from the third positions by truncated selection.

7. References

1. ADACHI J., HASEGAWA M. (1996). Tempo and mode of synonymous substitutions in mitochondrial DNA of primates. Mol. Biol. Evol. **13:** 200-208.

2. ANDERSSON S.G., ZOMORODIPOUR A., ANDERSSON J.O., SICHERITZ-PONTEN T., ALSMARK U.C., PODOWSKI R.M., NASLUND A.K., ERIKSSON A.S., WINKLER H.H., KURLAND C.G. (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature **396:** 133-140

3. BAKER T.A., WICKNER S.H. (1992). Genetics and enzymology of DNA replication in *Escherichia coli*. Annu. Rev. Genet. **26:** 447-477.

4. BARANTON G., POSTIC D., SAINT GIRONS I., BOERLIN P., PIFFARETTI J.C., ASSOUS M., GRIMONT P.A. (1992). Delineation of *Borrelia burgdorferi* sensu stricto, *Borrelia garinii* sp. Nov., and group VS461 associated with Lyme borreliosis. Int. J. Syst. Bacteriol. **42:** 378-383.

5. BASIC-ZANINOVIC T., PALOMBO F., BIGNAMI M, DOGLIOTTI E. (1992). Fidelity of replication of the leading and the lagging DNA strands opposite N-methyl-N-nitrosourea-induced DNA damage in human cells. Nucleic Acids Res. **20:** 6543-6548.

6. BARBOUR A.G., HAYES S.F. (1986). Biology of *Borrelia* species. Microbiol. Rev. **50:** 381-400.

7. BELETSKII A., BHAGWAT A.S. (1996). Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **93:** 13919-13924.

8. BENNETZEN J.H., HALL B.D. (1982). Codon selection in yeast. J. Biol. Chem. **257:** 3026 - 3031.

9. 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.

10. BLATTNER F.R., PLUNKETT G. 3rd, BLOCH C.A., PERNA N.T., BURLAND V., RILEY M., COLLADO-VIDES J., GLASNER J.D., RODE C.K., MAYHEW G.F., et al. (1997). The complete genome sequence of *Escherichia coli* K-12. Science **277:** 1453-1462.

11. BREWER B.J. (1988). When polymerases collide: Replication and the transcriptional organisation of the *E. coli* chromosome. Cell **53:** 679 – 686.

12. BURGDORFER W., BARBOUR A.G., HAYES S.F., BENACH J.L., GRUNWALDT E., DAVIS J.P. (1982). Lyme disease – a tick-borne spirochetosis? Science **216:** 1317-1319.

13. CANICA M.M., NATO F., DU MERLE L., MAZIE J.C., BARANTON G., POSTIC D. (1993). Monoclonal antibodies for identification of *Borrelia afzelii* sp. nov. associated with late cutaneous manifestations of Lyme borreliosis. Scand. J. Infect. Dis. **25:** 441-448.

14. CASJENS S., PALMER N., VAN VUGT R., HUANG W.M., STEVENSON B., ROSA P., LATHIGRA R., SUTTON G., PETERSON J., DODSON R.J., HAFT D., HICKEY E., GWINN M., WHITE O., FRASER C.M. (2000). A bacterial genome in flux: the

twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of Lyme disease spirochete *Borrelia burgdorferi*. Mol. Microbiol. **35(3):** 490-516.

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

16. CEBRAT S., DUDEK M.R., ROGOWSKA A. (1997a). Asymmetry in nucleotide composition of sense and antisense strands as a parameter for discriminating open reading frames as protein coding sequences. J.Appl. Genet. **38:** 1-9.

17. CEBRAT S., DUDEK M.R., MACKIEWICZ P., KOWALCZUK M., FITA M. (1997b). Asymmetry of coding versus non-coding strands in coding sequences of different genomes. Microb. & Comp. Genom. **2:** 259 - 268.

18. CEBRAT S., DUDEK M.R., MACKIEWICZ P. (1998). Sequence asymmetry as a parameter indicating coding sequences in *Saccharomyces cerevisiae* genome. Theory Bioscienc. **117:** 78-89.

19. CEBRAT S., DUDEK M.R., GIERLIK A., KOWALCZUK M., MACKIEWICZ P. (1999). Effect of replication on the third base of codons. Phys. A **265:** 78-94.

20. CHARGAFF E. (1950). Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia **6:** 201-240.

21. DANIELS D.L., SANGER F. COULSON A.R. (1983). Features of bacteriophage lambda: analysis of the complete nucleotide sequence. Cold Spring Harbor Symp. on Quantum Biology Vol. **47:** 1009-1024.

22. DESCHAVANNE P., FILIPSKI J. (1995). Correlation of GC content with replication timing and repair mechanisms in weakly expressed *E. coli* genes. Nucleic Acids Res. **23:** 1350-1353.

23. ECHOLS H., GOODMAN M.F. (1991). Fidelity mechanisms in DNA replication. Annu. Rev. Biochem. **60:** 477-511.

24. EISEN, J.A. HEIDELBERG JF, WHITE O, SALZBERG SL. (2000). Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Gen. Biol. **1:** 0011.1-0011.9.

25. FIJALKOWSKA 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.

26. FIJALKOWSKA I.J., JONCZYK P., MALISZEWSKA-TKACZYK M., BIALOSKORSKA M., SCHAAPER R.M. (1998). Unequal fidelity of leading strand and lagging strand DNA replication on the *Escherichia coli* chromosome. Proc. Natl. Acad. Sci. USA **95:** 10020-10025.

27. FILIPSKI J. (1990). Evolution of DNA sequences. Contributions of mutational bias and selection to the origin of chromosomal compartments. In: Advances in mutagenesis research 2 (ed. G. Obe), pp. 1-54. Springer Verlag, Berlin.

28. FILIPSKI J. (1998). Why the rate of silent codon substitutions is variable within a vertebrate's genome. J. Theor. Biol. **134:** 159-164.

29. FRANCINO M.P., CHAO L., RILEY M.A., OCHMAN H. (1996). Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science **272:** 107-109.

30. FRANCINO M.P., OCHMAN H. (1997). Strand asymmetries in DNA evolution. Trends Genet. **13:** 240-245.

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

32. FRANK A.C., LOBRY J.R. (1999). Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene **238:** 65-77.

33. 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. et al. (1995). The minimal gene complement of *Mycoplasma genitalium.* Science **270:** 397-403.

34. FRASER C.M., CASJENS S., HUANG W.M., SUTTON G.G., CLAYTON R., LATHIGRA R., WHITE O., KETCHUM K.A., DODSON R., HICKEY E.K. et al. (1997). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi.* Nature **390:** 580-586

35. FRASER C.M., NORRIS S.J., WEINSTOCK G.M., WHITE O., SUTTON G.G., DODSON R., GWINN M., HICKEY E.K., CLAYTON R., KETCHUM K.A. et al. (1998). Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science **281:** 375-388

36. FREDERICO L.A., KUNKEL T.A., SHAW B.R. (1990). A sensitive genetic assay for the detection of cytosine deamination: determination of rare constants and the activation energy. Biochemistry **29:** 2532-2537.

37. FREEMAN J.M., PLASTERER T.N., SMITH T.F., MOHR S.C. (1998). Patterns of genome organisation in bacteria. Science **279:** 1827.

38. FUKASAWA K.M., LI W.-H., YAGI K., LUO C.-C., LI S.S.-L. (1986). Molecular evolution of mammalian lactate dehydrogenase-A genes and pseudogenes: Association of a mouse processed pseudogene with a B1 repetitive sequence. Mol. Biol. Evol. **3:** 330-342.

39. FUKUNAGA M., HAMASE A., OKADA K., NAKAO M. (1996). *Borrelia tanukii* sp. nov. and *Borrelia turdae* sp. nov. found from ixoid ticks in Japan: rapid species identification by 16S rRNA gene-targeted PCR analysis. Microbiol. Immunol. **40:** 877-881.

40. GATES M.A. (1986). A simple way to look at DNA. J. Theor. Biol. **119:** 281-300.

41. GIERLIK A., MACKIEWICZ P., KOWALCZUK M., DUDEK M.R., CEBRAT S. (1999). Some hints on Open Reading Frame statistics – how ORF length depends on selection. Int. J. Modern Phys. C. **10:** 635-643.

42. GIERLIK A., KOWALCZUK M., MACKIEWICZ P., DUDEK M.R., CEBRAT S. (2000). Is there replication-associated mutational pressure in the *Saccharomyces cerevisiae* genome?. J. Theor. Biol. **202:** 305-314.

43. GOJOBORI T., LI W-H, GRAUR D. (1982). Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. **18:** 360-369.

44. GOUY M., GAUTIER C. (1982). Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. **10:** 7055-7074.

45. GRIGORIEV A. (1998). Analysing genomes with cumulative skew diagrams. Nucleic Acids Res. **26:** 2286-2290.

46. GRIGORIEV A. (1999). Strand-specific compositional asymmetries in double-stranded DNA viruses. Vir. Res. **60:** 1-19.

47. GUTIERREZ G., MARQUEZ L., MARTIN A. (1996). Preference for guanosine at first codon position in highly expressed *Escherichia coli* genes. A relationship with translation efficiency. Nucleic Acids Res. **24:** 2525-2528.

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

49. HIMMELREICH R., HILBERT H., PLAGENS H., PIRKL E., LI B.C., HERRMANN R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res. **24:** 4420-4449.

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

51. IKEMURA T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein sequence: a proposal for a synonymous codon choice that is optimal for the *E. coli* translation system. J. Mol. Biol. **151:** 389-409.

52. IWAKI T., KAWAMURA A., ISHINO Y., KOHNO K., KANO Y., GOSHIMA N., YARA M., FURUSAWA M., DOI H., IMAMOTO F. (1996). Preferential replication-dependent mutagenesis in the lagging DNA strand in *Escherichia coli*. Mol. Gen. Genet. **251:** 657-664.

53. JEANMOUGIN F., THOMPSON J.D., GOUY M., HIGGINS D.G., GIBSON T.J. (1988). Multiple sequence alignment with Clustal X. Trends. Biochem. Sci. **23:** 403-405.

54. KARLIN S. (1999). Bacterial DNA strand compositional asymmetry. Trends Microb. **8:** 305-308.

55. KARLIN S., BURGE C. (1995). Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. **11:** 283-290.

56. KARLIN S., MRAZEK J. (1996). What drives codon choices in human genome? J. Mol. Biol. **262:** 459-472.

57. KARLIN S., BLAISDELL B.E. I BUCHER P. (1992). Quantile distributions of amino acid usage in protein classes. Protein Eng. **5:** 729-738.

58. KAWABATA H., MASUZAWA T., YANAGIHARA Y. (1993). Genomic analysis of *Borrelia japonica* sp. nov. isolated from *Ixodes ovatus* in Japan. Microbiol. Immunol. **37:** 843-848.

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

60. KIMURA M. (1983). The neutral theory of Molecular Evolution. Camb. Univ. Press., p. 75.

61. KOWALCZUK M., MACKIEWICZ P., GIERLIK A., DUDEK M.R., CEBRAT S. (1999a). Total number of coding open reading frames in the yeast genome. Yeast **15:** 1031-1034.

62. KOWALCZUK M., GIERLIK A., MACKIEWICZ P., CEBRAT S., DUDEK M.R. (1999b). Optimization of gene sequences under constant mutational pressure and selection. Phys. A **273:** 116-131.

63. 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.

64. 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

65. KOWALCZUK M., MACKIEWICZ P., SZCZEPANIK D., NOWICKA A., DUDKIEWICZ M., DUDEK M.R., CEBRAT S. (2001c). Multiple base substitution corrections in DNA sequence evolution. Int. J. Modern Phys. C **12(7):** 1043-1053.

66. KREUTZER D.A., ESSIGMANN J.M. (1998). Oxidized, deaminated cytosines are a source of C→T transitions *in vivo*. Proc. Natl. Acad. Sci. USA **95:** 3578-3582.

67. KUNST F., OGASAWARA N., MOSZER I., ALBERTINI A.M., ALLONI G., AZEVEDO V., BERTERO M.G., BESSIERES P., BOLOTIN A., BORCHERT S., et al. (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature **390:** 249-256.

68. 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.

69. LAGUNEZ-OTERO J., TRIFONOV E.N. (1992). mRNA periodical infrastructure complementary to the proof-reading site in the ribosome. J. Biomol. Struct. Dyn. **10:** 455-464.

70. LE FLECHE A., POSTIC D., GIRARDET K., PETER O., BARANTON G. (1997). Characterization of *Borrelia lusitaniae* sp. nov. by 16S ribosomal DNA sequence analysis. Int. J. Syst. Bacteriol. **47:** 921-925.

71. LI W-H, WU C.I., LUO C-C. (1984). Nonrandomness of point mutation as reflected in nucleotide substitution in pseudogenes and its evolutionary implications. J. Mol. Evol. **12:** 58-71.

72. LINDAHL T. (1993). Instability and decay of the primary structure of DNA. Nature **362:** 709-715.

73. LIÒ P., GOLDMAN N. (1998). Models of molecular evolution and phylogeny. Gen. Res. **8:** 1233-1244.

74. LIU S.L., SANDERSON K.E. (1995). Rearrangements in the genome of the bacterium *Salmonella typhi*. Proc. Natl. Acad. Sci. USA **92:** 1018-1022.

75. LIU S.L., SANDERSON K.E. (1996). Highly plastic chromosomal organization in *Salmonella typhi*. Proc. Natl. Acad. Sci. USA **93:** 10303-10308.

76. LOBRY J.R. (1995). Properties of a general model of DNA evolution under no-strand bias conditions. J. Mol. Evol. **40:** 326-330. Erratum **41:** 680.

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

78. LOBRY J.R. (1996b). A simple vectorial representation of DNA sequence for the detection of replication origins in bacteria. Biochimie **78:** 323-326.

79. LOPEZ P., PHILIPPE H., MYLLYKALLIO H., FORTERRE P. (1999). Identification of putative chromosomal origins of replication in Archaea. Mol. Microbiol. **32:** 881-891.

80. LOPEZ P., FORTERRE P., LE GUYADER H., PHILIPPE H. (2000). Origin of replication of *Thermotoga maritima*. Trends. Genet. **16:** 59-60.

81. MACKIEWICZ P., GIERLIK A., KOWALCZUK M., DUDEK M.R., CEBRAT S. (1999a). Asymmetry of nucleotide composition of prokaryotic chromosomes. J .Appl. Genet. **40:** 1-14.

82. MACKIEWICZ P., GIERLIK A., KOWALCZUK M., DUDEK M.R, CEBRAT S. (1999b). How does replication-associated mutational pressure influence amino acid composition of proteins? Gen. Res. **9:** 409-416.

83. MACKIEWICZ P., GIERLIK A., KOWALCZUK M., SZCZEPANIK D., DUDEK M.R., CEBRAT S. (1999c). Mechanisms generating long-range correlation in nucleotide composition of the *Borrelia burgdorferi* genome. Phys. A **273:** 103-115.

84. MACKIEWICZ P., SZCZEPANIK D., GIERLIK A., KOWALCZUK M., NOWICKA A., DUDKIEWICZ M., DUDEK M.R., CEBRAT S. (2001a). The differential killing of genes by inversions in prokaryotic genomes. J. Mol. Evol. **53(6):** 615-621.

85. MACKIEWICZ P., SZCZEPANIK D., KOWALCZUK M., CEBRAT S. (2001b). Flip-flop around the origin and terminus of replication in prokaryotic genomes. Gen. Biol. **2(12):** interactions 1004.1-1004.4.

86. MARCONI R.T., LIVERIS D., SCHWARTZ I. (1995). Identification of novel insertion elements, restriction fragment length polymorphism patterns, and discontinuous 23S rRNA in Lyme disease spirochetes: phylogenetic analyses of rRNA genes and their intergenic spacers in *Borrelia japonica* sp. nov. and genomic group 21038 (*Borrelia andersonii* sp. nov.) isolates. J. Clin. Microbiol. **33:** 2527-2434.

87. MARIANS K.J. (1992). Prokaryotic DNA replication. Annu. Rev. Biochem. **61:** 673-719.

88. MATASSI G., SHARP P.M., GAUTIER C. (1999). Chromosomal location effects on gene sequence evolution in mammals. Curr. Biol. **9:** 786-791.

89. MCINERNEY J.O. (1998). Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc. Natl. Acad. Sci. USA **95:** 106698-10703.

90. MCLEAN M.J., Wolfe K.H., DEVINE K.M. (1998). Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J. Mol. Evol. **47:** 691-696.

91. MELLON I., HANAWALT P.C. (1989). Induction of the *Escherichia coli* lactose operon selectively increases repair of its transcribed DNA strand. Nature **342:** 95-98.

92. MIZRAJI E. I NINIO J. (1985). Graphical coding of Nucleic Acids sequences. Biochimie **67:** 445-448.

93. MRAZEK J., KARLIN S. (1998). Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl. Acad. Sci. USA **95:** 3720-3725.

94. OKAZAKI R., OKAZAKI T., SAKABE K., SUGIMOTO K., SUGINO A. (1968). Mechanism of DNA chain growth. I. Possible discontinuity and unusual secondary structure of newly synthesised chains. Proc. Natl. Acad. Sci. USA **59(2):** 598-605.

95. OLD I.G., MARGARITA D., SAINT GIRONS I. (1993). Unique genetic arrangement in the dnaA region of the *Borrelia burgdorferi* linear chromosome; nucleotide sequence of the dnaA gene. FEMS Microbiol. Lett. **111:** 109-114.

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

97. PERRIERE G., LOBRY J.R., THIOULOUSE J. (1996). Correspondence discriminant analysis: a multivariate method for comparing classes of protein and Nucleic Acids sequences. Comput. Appl. Biosci. **12:** 519-524.

98. PICARDEAU M., LOBRY J.R., HINNEBUSCH B.J. (1999). Physical mapping of an origin of bidirectional replication at the centre of the *Borrelia burgdorferi* linear chromosome. Mol. Microbiol. **32:** 437-445.

99. PICARDEAU M., LOBRY J.R., HINNEBUSCH B.J. (2000). Analyzing DNA strand compositional asymmetry to identify candidate replication origins of *Borrelia burgdorferi* linear and circular plasmids. Gen. Res. **10:** 1594-1604.

100. POSTIC D., RAS N.M., LANE R.S., HENDSON M., BARANTON G. (1998). Expanded diversity among California *Borrelia* isolates and description of *Borrelia bissettii* sp. nov. (formerly *Borrelia* group DN127). J. Clin. Microbiol. **36:** 3497-3504.

101. QIN M.H., MADIRAJU M.V., RAJAGOPALAN M. (1999). Characterization of the functional replication origin of *Mycobacterium tuberculosis*. Gene **233:** 121-130.

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

103. READ T.D., BRUNHAM R.C., SHEN C., GILL S.R., HEIDELBERG J.F., WHITE O., HICKEY E.K., PETERSON J., UTTERBACK T., BERRY K. et al. (2000). Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucleic Acids Res. **28:** 1397-1406.

104. REYES A., GISSI C., PESOLE G., SACCONE C. (1998). Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol. Biol. Evol. **15:** 957-966.

105. 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 deoxythymidine triphosphate. J. Biol. Chem. **269:** 1711-1717.

106. ROCHA E.P.C., VIARI A., DANCHIN A. (1998). Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons. Nucleic Acids Res. **26:** 2971-2980.

107. ROCHA E.P.C., DANCHIN A., VIARI A. (1999a). Universal replication biases in bacteria. Mol. Microbiol. **32:** 11-16.

108. ROCHA E.P.C., DANCHIN A., VIARI A. (1999b). Bacterial DNA strand compositional asymmetry: Response. Trends. Microb. **7:** 308.

109. ROMERO H., ZAVALA A., MUSTO H. (2000). Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. Nucl. Acids Res. **28:** 2084-2090.

110. SALZBERG S.L., SALZBERG A.J., KERLAVAGE A.R. TOMB J-F. (1998). Skewed oligomers and origins of replication. Gene **217:** 57-67.

111. SHARP P.M., COWE E. (1991), Synonymous codon usage in *Saccharomyces cerevisiae*. Yeast **7:** 657-678.

112. 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.

113. SHEPHERD J.C. (1981). Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc. Natl. Acad. Sci. USA **78:** 1596-1600.

114. SMITHIES O., ENGELS W.R., DEVEREUX J.R., SLIGHTOM J.L. I SHEN S.H. (1981). Base substitutions, length differences, and DNA strand asymmetries in the human G_γ and A_γ fetal globin gene region. Cell. **26:** 345-353.

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

116. 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. **52:** 426-433.

117. TAMURA K. (1992). The rate and pattern of nucleotide substitution in *Drosophila* mitochondrial DNA. Mol. Biol. Evol. **9:** 814-825.

118. TANAKA M., OZAWA T. (1994). Strand asymmetry in human mitochondrial DNA mutations. Genomics **22:** 327-335.

119. THOMAS D.C., SVOBODA D.L., VOS J.M., KUNKEL T.A. (1996). Strand specificity of mutagenic bypass replication of DNA containing psoralen monoadducts in a human cell extract. Mol. Cell. Biol. **16:** 2537-2544.

120. TILLIER E.R.M., COLLINS R.A. (2000a). The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. J. Mol. Evol. **50:** 249-257.

121. TILLIER E.R.M., COLLINS R.A. (2000b). Genome rearrangement by replication-directed translocation. Nat. Genet. **26:** 195-197.

122. TILLIER E.R.M., COLLINS R.A. (2000c). Replication orientation affects the rate and direction of bacterial gene evolution. J. Mol. Evol. **51:** 459-463.

123. TRIFONOV E.N. (1987). Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S rRNA nucleotide sequences. J. Mol. Biol. **194:** 643-652.

124. TRIFONOV E.N. (1992). Recognition of correct reading frame by the ribosome. Biochimie **74:** 357-362.

125. TRINH T.Q., SINDEN R.R. (1991). Preferential DNA secondary structure mutagenesis in the lagging strand of replication in *E. coli*. Nature **352:** 544-547.

126. VEAUTE X., FUCHS R.P.P. (1993). Greater susceptibility to mutations in lagging strand of DNA replication in *Escherichia coli* than in leading strand. Science **261:** 598-600.

127. VOSS R. (1992). Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys. Rev. Lett. **68:** 3805-3808.

128. WANG G., VAN DAM A.P., LE FLECHE A., POSTIC D., PETER O., BARANTON G., DE BOER R., SPANJAARD L., DANKERT J. (1997). Genetic and phenotypic analysis of *Borrelia valaisiana* sp. nov. (*Borrelia* genomic groups VS116 and M19). Int. J. Syst. Bacteriol. **47:** 926-932.

129. WANG G., VAN DAM A.P., SCHWARTZ I., DANKERT J. (1999). Molecular typing of *Borrelia burgdorferi* sensu lato: taxonomic, epidemiological and clinical implications. Clin. Microbiol. Rev. **12(4):** 633-653.

130. WANG J. (1998). The base contents of A, C, G, or U for three codon positions and the total coding sequences show positive correlation. J. Biomol. Struct. Dyn. **16:** 51-57.

131. WATSON J.D., CRICK F.C.H. (1953). A structure for deoxyribose nucleic acid. Nature **171:** 737-738.

132. WHELAN S., LIÒ P., GOLDMAN N. (2001). Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends genet. **17(5):** 262-272.

133. WOLFE K.H., SHARP P.M., LI W.-H. (1989). Mutation rates differ among regions of the mammalian genome. Nature **337:** 283-285.

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

135. YANG Z. (1994). Estimating the pattern of nucleotide substitution. J. Mol. Evol. **39:** 105-111.

136. YUZHAKOV A., TURNER J., O'DONNEL M. (1996). Replisome assembly reveals the basis for asymmetric function in leading and lagging strand replication. Cell **86:** 877-886.

137. ZAWILAK A., CEBRAT S., MACKIEWICZ P., KRÓL-HULEWICZ A., JAKIMOWICZ D., MESSER W., GOSCINAK G., ZAKRZEWSKA-CZERWINSKA J. (2001). Identification of a putative chromosomal replication origin from *Helicobacter pylori* and its interaction with the initiator protein DnaA. Nucleic Acids Res. **29:** 2251-2259.

138. ZHANG J. (1999). Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. Mol. Biol. Evol. **16:** 868-875.

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

CONTENTS