



Nowe spojrzenie na filogenezę z punktu widzenia genomiki

Maciej Sobczyński, Paweł Mackiewicz, Dorota Mackiewicz,
Kamila Smolarczyk, Stanisław Cebrat

Zakład Genomiki, Instytut Genetyki i Mikrobiologii,
Uniwersytet Wrocławski, Wrocław

New insights into phylogeny from genomic point of view

Summary

Availability of fully sequenced genomes contributes to the development of new science named phylogenomics which opens new possibilities of phylogenetic analyses and study of genome evolution based on the whole information coded in genomic DNA. The advantages and disadvantages of the new methods are described. Despite many phenomena such as lineage-specific gene loss, gene duplication and horizontal gene transfer disturbing phylogenetic analyses, the new methods are able to extract some phylogenetic signals in the analysed genomes and construct reliable phylogenetic trees. The genome-based studies support not only the three-domain concept of *Tree of Life* but they identify previously undetected major clades of prokaryotes as well.

Key words:

phylogenomics, genome, phylogeny, archaea, bacteria, tree of life.

Adres do korespondencji

Maciej Sobczyński,
Zakład Genomiki,
Instytut Genetyki
i Mikrobiologii,
Uniwersytet Wrocławski,
ul. Przybyszewskiego 63/77,
51-148 Wrocław;
e-mail:
pamac@microb.uni.wroc.pl

1. Molekularne analizy filogenetyczne

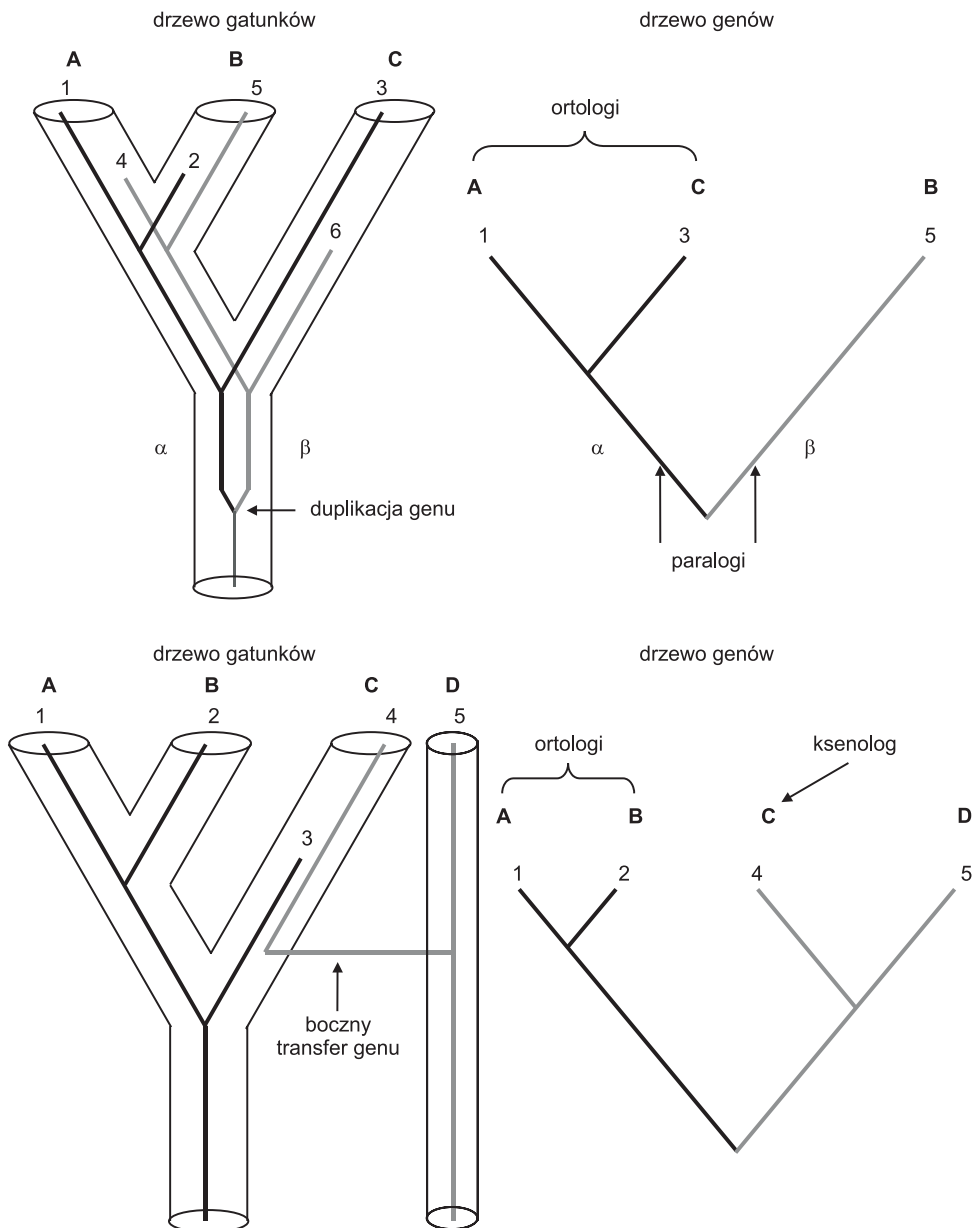
Rekonstrukcja genealogicznego drzewa wszystkich organizmów żyjących na Ziemi jest marzeniem biologów od czasów Karola Darwina. Nauką badającą i opisującą ewolucyjne związki między organizmami jest filogenetyka. W jej klasycznym podejściu związki te odkrywa się przez porównywanie cech morfologicznych badanych organizmów. Dzisiaj dodatkowym źródłem danych do analiz filogenetycznych są sekwencje kwasów nukleinowych i białek. Dzięki znaczącemu postępowi w sekwencjonowaniu

waniu, biolodzy otrzymują coraz większe ilości danych molekularnych pozwalających na szersze podejście do zagadnień systematyki i taksonomii. Ustalenie pokrewieństw filogenetycznych między organizmami pozwoli na zrozumienie ich ewolucji oraz opracowanie ich naturalnej klasyfikacji porządkującej gatunki w sposób umożliwiający ich rozpoznanie i opisanie. Naturalna klasyfikacja opiera się na podobieństwach wynikających ze wspólnego pochodzenia.

Schemat filogenezy przedstawiany jest najczęściej w postaci dichotomicznego grafu zwanego drzewem. Drzewo filogenetyczne złożone jest z węzłów i gałęzi, gdzie tylko jedna gałąź może łączyć dwa węzły. Węzeł przedstawia jednostkę taksonomiczną, natomiast gałąź to relacja pomiędzy dwoma jednostkami taksonomicznymi. Długość gałęzi jest odzwierciedleniem różnic między tymi jednostkami. W przypadku drzew konstruowanych na podstawie analizy sekwencji DNA długość gałęzi odpowiada liczbie substytucji nukleotydów. Badania molekularne dostarczają danych, które mają wiele zalet w porównaniu z innymi tradycyjnie badanymi cechami organizmu. Podstawową ich zaletą jest to, że dotyczą bezpośrednio struktury materiału genetycznego. Dzięki temu nie analizuje się cech plastycznych fenotypowo zmieniających się zależnie od czynników niegenetycznych. Ponadto sekwencje DNA, RNA i białek umożliwiają porównywanie gatunków odległych ewolucyjnie i słabo zróżnicowanych morfologicznie. Ilość danych o tych sekwencjach przyrasta w postępie wykładniczym. Efektywna ich analiza jest możliwa dzięki rozwojowi filogenetycznych programów komputerowych opartych na zdefiniowanych modelach ewolucyjnych (1,2).

W badaniach ewolucji genów i białek używa się sekwencji homologicznych, czyli o wspólnym pochodzeniu i pełniących podobną funkcję. Ważny jest jednak dobry wybór danych do analiz porównawczych i pewność, że interesujące nas sekwencje są ortologiczne, to znaczy pochodzą od wspólnego przodka i zostały rozdzielone w czasie rozchodzenia się linii rodowej analizowanych organizmów. Trudności w badaniach filogenetycznych następują geny występujące w tym samym organizmie, a powstałe w wyniku duplikacji. Geny te, zwane paralogicznymi, dają błędny obraz filogenezy, gdyż mogły powstać zanim linie rodowe analizowanych gatunków się rozeszły (rys. 1). Drzewa filogenetyczne mogą zawierać błędy także z powodu zjawiska horyzontalnego przenoszenia materiału genetycznego między organizmami (rys. 1). Geny w ten sposób nabyte zwane są ksenologicznymi.

Standardowa analiza filogenetyczna składa się z kilku etapów. Pierwszym krokiem jest wybór sekwencji DNA lub białek i ich porównanie. Drugim etapem jest zbudowanie matematycznego modelu ewolucji analizowanych sekwencji, który wyjaśni różnice między nimi. Polega on na określeniu mechanizmu podstawień (substytucji) nukleotydów lub aminokwasów. Trzecim krokiem jest konstrukcja drzewa najlepiej wyjaśniającego relacje filogenetyczne ukryte w sekwencjach i jego interpretacja.



Rys. 1. Niezgodność między drzewem gatunków i drzewem genów spowodowana obecnością paralogów i ksenologów w analizowanym zbiorze danych. Po prawej stronie przedstawiono rzeczywiste powiązania między trzema gatunkami (A, B, C) oraz losy genów (1-6) przedstawione w postaci linii. Brak kontynuacji linii oznacza, że dany gen zanikł u danego gatunku. Po lewej stronie przedstawiono drzewa skonstruowane na podstawie genów występujących u powstałych gatunków. Na górnym panelu widać, że z powodu uwzględnienia w analizie paralogów (a nie ortologów), gatunek A jest grupowany z C zamiast z B. Na dolnym panelu, z powodu zajęcia bocznego transferu genu, gatunek C jest łączony z niespokrewnionym gatunkiem D zamiast z gatunkami A i B.

2. Czy drzewo życia jest uniwersalne?

W przeszłości badania filogenetyczne opierały się na analizie cech morfologicznych, biochemicznych i fizjologicznych (3,4). Świat żywych organizmów był dzielony na *Prokaryota* i *Eukaryota*, głównie na podstawie ultrastruktury komórki. W ósmej edycji *Bergey's Manual of Determinative Bacteriology* z 1974 r. podzielono *Prokaryota* na dwa działy: sinice (*Cyanobacteria*) i resztę (5). Wśród „nie-sinic” wyróżniono 19 grup bakterii. Chociaż podziały takie były użyteczne, bez odpowiedzi pozostawało pytanie o rzeczywiste powiązania ewolucyjne, które odpowiadałyby tej klasyfikacji. Na takie analizy nie pozwalały jednak słabe zróżnicowanie morfologiczne i prosta organizacja komórek bakterii. Nowym podejściem w tego typu badaniach stało się zastosowanie do analiz filogenetycznych sekwencji małej podjednostki rybosomalnego RNA, 16S rRNA (6-8). Sekwencje rRNA okazały się szczególnie użyteczne w analizach filogenetycznych wszystkich organizmów. Sekwencje te występują zarówno u *Prokaryota*, jak i *Eukaryota*, podlegają stosunkowo powolnym zmianom w czasie i z tego powodu nadają się do porównań gatunków odległych ewolucyjnie, pochodzących nawet z różnych królestw. Na podstawie analiz rRNA wykazano, że świat organizmów należy podzielić nie na dwie, lecz na trzy wyraźnie oddzielne grupy. Podczas gdy organizmy eukariotyczne (nazwane *Eucarya*) zachowały swoją tożsamość, wśród organizmów prokariotycznych wyróżniono bakterie właściwe, czyli *Eubacteria* (obecnie *Bacteria* – bakterie) oraz na nową grupę – archebakterie (obecnie *Archaea* – archeony), obejmującą metanogeny i organizmy żyjące w ekstremalnych warunkach środowiskowych: termofile i halofile. Co ciekawe, *Archaea* okazały się równorzędną grupą, jednakowo odległą od dwóch pozostałych. Specyficzność *Archaea* potwierdzono również w badaniach morfologicznych i biochemicznych dotyczące składu lipidów w błonie komórkowej i szczegółów budowy ściany komórkowej. W analizach opartych na dużej liczbie nowych sekwencji rRNA (9) oraz białkach, jak: polimerazy RNA, ATPazy, czynniki elongacji translacji i syntetazy aminoacylo-tRNA potwierdzono podział świata żywego na trzy domeny (10).

Dostępność nowych sekwencji raczej skomplikowała analizy i koncepcje filogenetyczne zarówno na poziomie niższych jednostek taksonomicznych, jak i na szczeblu domen i królestw. Okazało się, że różne sekwencje dawały różne drzewa, a drzewo genów nie powinno być utożsamiane z drzewem gatunków (rys. 1). Analizując różne sekwencje białkowe uzyskuje się różne powiązania między bakteriami, archeonami i *Eukaryota* (10,11). *Bacteria* łączą się z *Archaea* na podstawie analiz ALADH, HSP 70, FtsZ, syntetazy cytrynianowej, dehydrogenazy glutaminianowej II, syntetazy glutaminy I i gyrazy B. Natomiast *Bacteria* są bliższe *Eucarya* według GAPDH i dehydrogenazy jabłczanowej, a *Archaea* są grupowane z *Eucarya* na bazie syntetazy argininobursztynianowej, różne syntetazy aminoacylo-tRNA, podjednostki ATPazy, czynniki elongacji, podjednostki polimerazy RNA, DNA-zależną polimerazę B, syntetazę karbamoilofosforanu i podjednostkę rozpoznającą sygnały (SRP). Większość analiz, opartych na genach paralogicznych, zduplikowanych jeszcze

przed podziałem na trzy domeny i umożliwiających ukorzenie drzewa, postuluje wcześniejsze oddzielenie się linii bakterii, od linii prowadzącej do archeonów i *Eukaryota* (12-17). Pojawiły się również koncepcje, w których sugeruje się, że *Eukaryota* powstały przez fuzję przedstawicieli *Archaea* i bakterii gramujemnych (18-21), co potwierdzono w analizach sekwencji syntetazy acetylo-CoA i fotolizy (11). Ponadto w analizach wykazano, że geny *Archaea* zaangażowane w procesy replikacji, transkrypcji i translacji są bliższe genom eukariotycznym, natomiast geny kodujące białka strukturalne i związane z metabolizmem wskazują na związki z bakteriami (21-23).

Niezgodność topologii drzew uzyskanych na podstawie różnych sekwencji ma wiele przyczyn. Wyłączając błędy metodologiczne związane ze stosowanymi algorytmami, błędne drzewa filogenetyczne mogą wynikać z następujących przyczyn:

- zbyt małej liczby miejsc w sekwencjach niosących sygnał filogenetyczny; dotyczy to szczególnie sekwencji krótkich i niewystarczająco zróżnicowanych w obrębie analizowanych organizmów;
- zbyt dużej dywergencji analizowanych sekwencji związanej z dużą liczbą podstawień wielokrotnych, tj. takich, które zaszły w tych samych miejscach sekwencji;
- zmiennego tempa ewolucji sekwencji w różnych liniach filogenetycznych;
- zróżnicowanego tempa podstawień w różnych regionach sekwencji;
- podobnego składu (np. % G+C) niespokrewnionych sekwencji, które są łączone razem na drzewie;
- obecności w analizowanym zbiorze sekwencji paralogicznych oraz pochodzących z bocznego transferu.

Problemem w analizach filogenetycznych są duplikacje sekwencji oraz utrata genów w danej linii filogenetycznej. To prowadzi do porównań sekwencji, które praktycznie nie są ortologami i do niepełnej reprezentacji ortologicznych genów u wszystkich gatunków, które chcemy analizować. Rzeczywiste powiązania filogenetyczne są szczególnie zaburzane przez horyzontalny transfer genów (HGT, *Horizontal Gene Transfer*) powszechny u organizmów prokariotycznych (24-29). Nabyty niespokrewniony gen może zastąpić jego funkcjonalny odpowiednik w genomie, co zwane jest, w zależności od historii sekwencji, nieortologicznym lub ksenologicznym zastępowaniem genów (30). W niektórych przypadkach nawet kilkadziesiąt procent genów zostało nabytych w wyniku HGT (28). W przytoczonych przykładach pokazano, jak istotne znaczenie ma zjawisko horyzontalnego przekazywania informacji genetycznej na treść i wiarygodność hipotez filogenetycznych. Ukazano też możliwość ilościowych i jakościowych zmian, jakie mogą zajść w genomach w wyniku tego procesu.

Wspomnianą mozaikowość przedstawicieli *Archaea* można tłumaczyć właśnie za pomocą HGT i nabyciem wielu genów od bakterii. Horyzontalny transfer genów wykazano u *Methanobacterium thermoautotrophicum*, u którego znaleziono regiony DNA o niższej (o ok. 10%) zawartości G+C w porównaniu z całym genomem tego archeona (31). Także nietypowa używalność kodonów w analizowanych genach sugeruje,

że zostały one nabyte w wyniku bocznego transferu. Co więcej, 490 genów (z 1869) wykazuje wysokie podobieństwo do genów bakteryjnych, podczas gdy tylko 352 wykazuje podobieństwo większe niż 50% do genów innego archeona – *Methanococcus jannaschi*. Geny, których produkty są zaangażowane w procesy transkrypcji, translacji i replikacji u *Archeoglobus fulgidus* są bardzo podobne do swoich odpowiedników u *M. jannaschi*. Natomiast geny niektórych szlaków metabolicznych tych archeonów różnią się znacznie sekwencją. Biosynteza tryptofanu u *A. fulgidus* przebiega niemal identycznie jak u bakterii *Bacillus subtilis*, mimo że gatunki te dzieli znaczna odległość na drzewie rRNA (32). Stwierdzono, że *Pyrococcus furiosus* posiada szlaki transportu i metabolizmu maltozy i maltodekstryn nieobecne u blisko spokrewnionego *P. horikoshii*. W związku z tym, że jeden z tych szlaków jest bardzo podobny do szlaku występującego u *E. coli*, można założyć, iż *P. furiosus* nabył te geny właśnie od *E. coli* (33).

Transfer mógł następować również od termofilnych *Archaea* do hipertermofilnych bakterii. Około 25% genów *Thermotoga maritima* oraz 16% *Aquifex aeolicus*, gatunków należących do hipertermofili, wykazuje największe podobieństwo właśnie do białek archeonów (34,35). Horyzontalne przekazywanie DNA występuje także w obrębie bakterii oraz między bakteriami i *Eukaryota*. Dla *E. coli* wykazano, że 675 genów ma wysoki stopień podobieństwa do swoich odpowiedników u sinicy *Synechocystis*, 231 genów – do archeona *M. jannaschi* i 254 genów – do drożdży *Saccharomyces cerevisiae* (36). Uważa się, że 755 z 4288 genów *E. coli* zostało nabytych horyzontalnie w 234 zdarzeniach, jakie zaszły w czasie 100 mln lat ewolucji tego gatunku (37). Gramodatnia laseczka *Clostridium acetobutylicum* jest zdolna do sporulacji i kiełkowania, jednak nie znaleziono u niej odpowiednich ortologów odpowiedzialnych za te procesy u blisko spokrewnionej laseczki *B. subtilis*. Sugeruje to, że *C. acetobutylicum* i *B. subtilis* nabyły geny sporulacji i kiełkowania z różnych źródeł (38). Sekwencje insercyjne (IS, *Insertion Sequences*) w genomie *Synechocystis* wykazują cechy sekwencji *E. coli*. Interesujące jest także to, że niektóre z genów tej sinicy mają swoje homologii jedynie w niektórych genomach eukariotycznych, w tym w genomie ludzkim (39). Uważa się również, że prawie 1/3 genów krętków mogła zostać nabyta przez HGT (28,40).

Powszechność HGT rodzi pytanie, czy graf filogenetyczny powinien w tych okolicznościach mieć postać drzewa, czy może raczej formę sieci, nawet wychodzącej poza świat prokariotyczny. HGT czyni koncepcję oryginalnego drzewa ewolucyjnego niespójną. Wydaje się, że scharakteryzowanie genów, które nie podlegają temu procesowi pomoże ominąć wynikające z niego trudności dla filogenetyki molekularnej, chociaż znalezienie takich „markerów filogenetycznych” może być trudne. W szczegółowych analizach 28 rodzin białek u *Prokaryota* wykazano, że po wyeliminowaniu ze zbioru genów nabytych w wyniku bocznego transferu, w pozostałych sekwencjach nie pozostał istotny sygnał filogenetyczny (41). Podobne wyniki uzyskano dla konserwatywnych białek archeonów. Mimo że wszystkie one wykazywały większą konserwatywność w obrębie *Archaea* niż poza tą domeną, nie udało się określić dla nich istotnych powiązań filogenetycznych (42). Wydawałoby się, że do-

brymi kandydatami na geny nie podlegające HGT powinny być geny zaangażowane w fundamentalne procesy w komórce, jak replikacja DNA, transkrypcja i translacja. Jednakże znaleziono dowody na transfer sekwencji kodujących polimerazy RNA, czynniki elongacji, syntetazy aminoacylo-tRNA, a nawet białka rybosomalne (S14) i rRNA (24,43-46). Podważono nawet niektóre powiązania oparte na drzewie rRNA związane z pozycją bakterii hipertermofilnych umieszczanych na nim u podstawy pnia bakterii (47-49). Ponadto rRNA posiada stosunkowo mało informatywnych miejsc w swojej sekwencji, w której około 1/3 bierze udział w tworzeniu struktur drugorzędowych (słupków i pętli), dlatego zmiany w tych miejscach nie są niezależne.

3. Analizy filogenetyczne w erze poligenomicznej

Chociaż obecność wymienionych problemów może bardzo utrudniać znalezienie prawidłowych powiązań filogenetycznych, to, jak się wydaje, analizy oparte na pełnych zbiorach genów mogą pomóc w rozwiązaniu przynajmniej niektórych z nich. Jest to możliwe dzięki dostępności coraz większej liczby kompletnie zsekwencjonowanych genomów, co umożliwi wiele nowych analiz filogenetycznych, w tym tworzenie drzew genomowych. Jest to przedmiotem nowej dziedziny wiedzy – filogenetyki genomów lub filogenomiki. Opracowano kilka metod tworzenia drzew genomowych posiadających jednak zarówno wady, jak i zalety (50).

3.1. Metody oparte na składzie genów

W metodach tych zakłada się, że blisko spokrewnione gatunki powinny posiadać dużą frakcję wspólnych genów występujących u obu gatunków. Natomiast gatunki bardziej odległe powinny utracić część genów odziedziczonych po wspólnym przodku, co powinno odbić się w mniejszej frakcji wspólnych genów. Jeżeli utrata genów jest proporcjonalna do czasu oddzielania się gatunków, powinno się uzyskać dobre związki filogenetyczne na bazie frakcji wspólnych genów. Jednakże, genomy prokariotyczne są bardzo plastyczne i bardzo łatwo tracą wiele genów (np. pasożyty wewnątrzkomórkowe) lub nabywają je w wyniku bocznego transferu. Dlatego drzewa skonstruowane w ten sposób są raczej drzewami fenetycznymi niż filogenetycznymi i odzwierciedlają ogólne podobieństwo i styl życia genomów, a nie związki ewolucyjne. Stworzono wiele drzew opartych na tej metodzie różniących się wyborem analizowanych genów i metodą konstrukcji drzewa (51-60). Drzewa te dobrze odzwierciedlały powiązania na poziomie trzech domen i blisko spokrewnionych gatunków, jednak błędnie odzwierciedlały związki na średnich odległościach filogenetycznych. Na przykład rozdzielały gatunki wolnożyjące i pasożytnicze należące do proteobakterii (57), co wiąże się z silną redukcją genomów tych ostatnich. Lepsze wyróżnienie głównych linii filogenetycznych uzyskano, gdy wybierano do analiz tyl-

ko rodziny białkowe (53) i dzielono odległości między dwoma genomami przez liczbę genów w mniejszym genomie (52,60).

3.2. Metody oparte na ułożeniu genów

W metodach tych przyjmuje się, że rearanżacje zachodzą równomiernie w genomach i losowo zmieniają ułożenie genów. Im genomy są bardziej odległe, tym mniej genów wykazuje podobne położenie. Operonowa organizacja genomu wpływa jednak z jednej strony na ograniczenie tempa rearanżacji, a z drugiej, sprawia ona, że geny biorące udział w tych samych procesach komórkowych i wspólnie ekspresjonowane mają tendencję do występowania razem, co można wykorzystać w tych analizach (61,62). Całe operony są jednak podatne na HGT, co zaburza powiązania filogenetyczne na uzyskiwanych drzewach. Ograniczenia tej metody wynikają również ze zbyt częstych rearanżacji zachodzących w genomach *Prokaryota* (63-68). Nawet stosunkowo blisko spokrewnione genomy nie wykazują konserwatywnego ułożenia genów. Podobnie jak we wcześniej opisanej metodzie, skonstruowane drzewa dobrze odzwierciedlają związki filogenetyczne na poziomie domen i blisko spokrewnionych organizmów, a słabo na średnich odległościach filogenetycznych. Topologia tych drzew jest ponadto zaburzona przez boczny transfer genów (57,60).

3.3. Metody oparte na odległościach ewolucyjnych między ortologami

W metodach tych wykorzystywane są odległości ewolucyjne genomów (np. średnie lub mediany) mierzone procentem identyczności lub stopniem podobieństwa między parami wielu ortologów. Zakłada się w nich, że odległości te korelują z dywergencją porównywanych gatunków (57,58,69). Drzewa wyraźnie rozdzielają *Archaea* od *Bacteria*, wyróżniają wiele znanych linii bakteryjnych i proponują nowe (patrz niżej).

3.4. Metody oparte na połączonych przyrównaniach wielu ortologów

Metody te oparte są na tradycyjnych molekularnych analizach filogenetycznych, zliczających zmiany w analizowanych sekwencjach, jednakże wykorzystują one dużą liczbę sekwencji. Dzięki temu unika się wielu problemów związanych z pojedynczymi sekwencjami. Jeżeli nie występował intensywny horyzontalny transfer genów lub geny takie mają niewielki udział w analizowanym zbiorze sekwencji, to drzewa oparte na tej metodzie powinny wiernie odzwierciedlać związki filogenetyczne. Ograniczenia związane są jednak z niepełną reprezentacją grup ortologów we wszystkich analizowanych genomach i niebezpieczeństwem wyboru do analizy pa-

ralogów w przypadku, gdy właściwy ortolog zanikł. Ponadto jednoczesna analiza sekwencji wymaga zastosowania takiego samego modelu podstawień dla wszystkich sekwencji, co nie zawsze jest słuszne. W tych analizach z reguły wybierano sekwencje kodujące białka związane z procesami replikacji, transkrypcji i translacji, np. rybosomalne (41,57,70-74). Uzyskane wyniki są podobne lub lepsze od tych opartych na metodzie odległości ewolucyjnych. Metoda ta istotnie wyróżnia, poza trzema domenami, wiele głównych linii bakterii i archeonów.

3.5. Metody oparte na wielu drzewach

W metodach tych obchodzi się problem stosowania jednakowego modelu ewolucji dla wszystkich sekwencji i ograniczonej liczby homologicznych genów, które muszą występować w analizowanych gatunkach. Konstruowane są osobne drzewa dla poszczególnych grup homologów, a następnie tworzy się dla nich jedno drzewo konsensusowe. W tym podejściu pojawia się jednak problem obecności paralogów w zbiorze danych oraz problem związany z porównywaniem i pogodzeniem wielu drzew opisujących często różne zbiory organizmów. Zastosowano wiele różnych podejść, aby obejść te ograniczenia (57,75-78). Wyniki tej metody zbliżone są do poprzedniej, jednakże w niektórych przypadkach lepiej identyfikowane są poszczególne grupy.

3.6. Metody oparte na indelach w sekwencjach

W przeciwieństwie do większości metod analizujących substytucje w sekwencjach, w tych metodach analizowana jest obecność lub brak charakterystycznych insercji lub delecji (zwanych łącznie indelami) w sekwencjach wspólnych dla danych linii filogenetycznych (79). Dzięki temu unika się problemu wielokrotnych podstawień. Na dużych dystansach ewolucyjnych indele mogą dawać lepsze sygnały filogenetyczne ze względu na to, że zdarzają się o wiele rzadziej niż substytucje. Istnieje jednak możliwość ich silnego związku z presją selekcyjną powodującą podobne zmiany w niezależnych liniach. Wspólne indele mogą być również związane z sekwencjami nabytymi w wyniku HGT. Analizy oparte na sygnaturach indeli dały wyniki sprzeczne z podziałem świata żywego na trzy domeny, dając alternatywny pogląd na ewolucję życia. Wykazano, że *Archaea* są polifiletyczne i bardziej powiązane z bakteriami gramdodatnimi. Ponadto na ich podstawie sugeruje się, że *Prokaryota* powinny zostać podzielone na dwie grupy nazwane ze względu na liczbę błon otaczających komórkę: monodermata, obejmującą wszystkie archeony i bakterie gramdodatnie oraz didermata, obejmujące bakterie gramujemne. Natomiast *Eukaryota* miałyby powstać przez fuzję termoacidofilnych archeonów i bakterii gramujemnych.

3.7. Metody oparte na składzie sekwencji i genomów

Reprezentują one alternatywne podejście do analiz filogenetycznych, a przedmiotem tych analiz jest ogólny skład sekwencji, tj. częstość występowania dwunukleotydów (80), tetranukleotydów (81), aminokwasów (82) i oligopeptydów (83). Mimo znacznej degeneracji informacji wiele uzyskiwanych drzew wykazuje dużą zgodność z drzewem opartym na rRNA. Ponadto w analizie składu aminokwasowego proteomu pokazuje się, że koresponduje on z wymaganiami środowiskowymi organizmu, np. temperaturą (82). Powiązania organizmów uzyskiwane tymi metodami mogą jednak nie odzwierciedlać rzeczywistych powiązań rodowych i grupować genomy, które niezależnie uzyskały podobny skład, np. w wyniku podobnego stylu i warunków życia.

4. Powiązania filogenetyczne w obrębie organizmów prokariotycznych

Wśród bakterii wyróżniono początkowo 12 grup systematycznych (84). Jednakże w ostatnich latach zaproponowano wiele dodatkowych grup bakterii (85). Niektóre z tych grup zawierają tylko jeden lub kilka gatunków (np. *Thermomicrobia*, *Fibrobacteries*), a inne (*Actinobacteria*, *Firmicutes*) liczą setki gatunków. Z powodu braku jasnych i obiektywnych reguł, nie jest pewne, czy wyodrębnianie konkretnych jednostek systematycznych jest uzasadnione. Na podstawie analiz 16S rRNA, zaproponowano podział *Proteobacteria* na pięć podgrup nazywanych α , β , γ , δ i ϵ (84). Niektóre z tych podgrup (α , β i γ) zawierają tysiące gatunków i są znacznie bogatsze niż większość wyższych jednostek systematycznych domeny *Bacteria*. Brak wspomnianych kryteriów filogenetycznych sprawia, że nie jest jasne, czy te bardzo liczne zbiory gatunków mają otrzymać status podgrupy, podczas gdy te znacznie uboższe, zawierające niejednokrotnie pojedyncze gatunki, powinny być wyodrębniane jako jednostki wyższego rzędu. Związki filogenetyczne między poszczególnymi grupami bakterii i archeonów zależą od stosowanych metod i wyboru sekwencji do analiz.

Wiele linii filogenetycznych *Prokaryota* wyróżnionych na podstawie rRNA, pojedynczych białek lub metod klasycznych jest identyfikowanych metodami genomowymi. Przykładowo, na podstawie przeprowadzonych analiz wielu genów aparatu translacyjnego wykazano, że *Proteobacteria*, gramodatnie bogate w G+C (*Actinobacteria*), gramodatnie ubogie w G+C (*Firmicutes*), krętki (*Spirochaetes*) i sinice są wyraźnymi grupami monofiletycznymi (73). Monofiletyczność tych grup została wykazana także na podstawie metody superdrzewa (75,78). Dobrze wyróżniane są metodami genomowymi podgrupy w obrębie proteobakterii (57,58,78), co jest zgodne z analizami opartymi na rRNA i pojedynczych białkach (86).

W analizach filogenetycznych zarówno genomowych, jak i opartych na rRNA oraz czynnika translacji Tu (87) wykazano, że tradycyjnie wyróżniana grupa bakterii gramodatnich jest polifiletyczna, co sugeruje niezależną utratę błony zewnętrznej

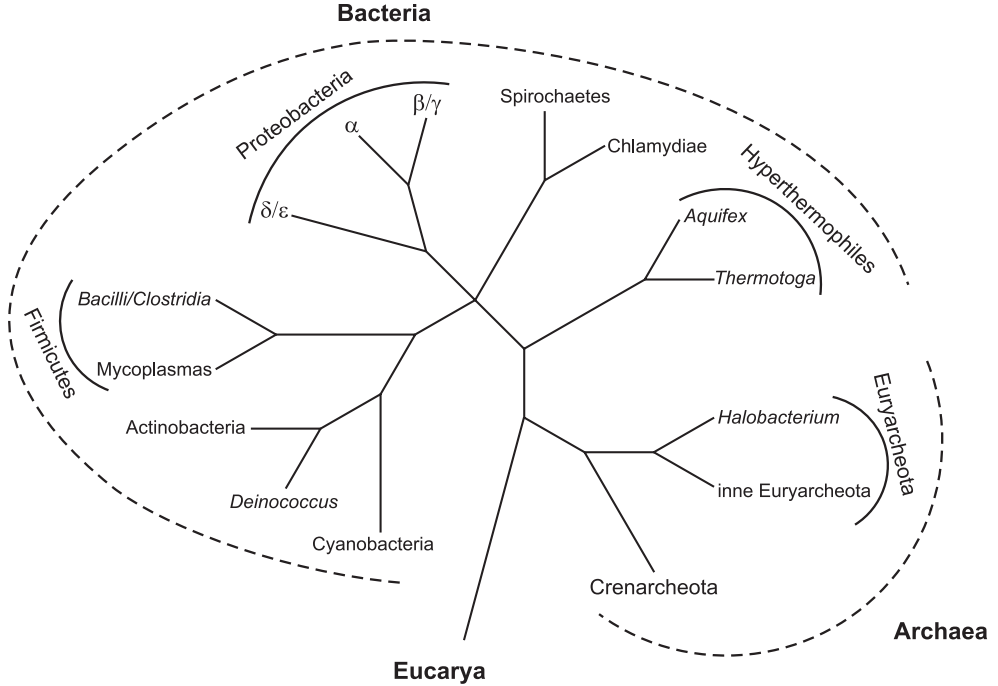
w dwóch liniach filogenetycznych bakterii bogatych i ubogich w G+C. Wiele metod genomowych wyróżnia jako osobną grupę bakterie gramdodatnie ubogie w G+C, natomiast gramdodatnie bogate w G+C (*Actinobacteria*) są łączone w jeden kład z *Deinococcus* i *Cyanobacteria* (50,57,73,75,78,83). W innych metodach (71) łączy się *Deinococcus* z sinicą *Synechocystis*, czyli podobnie jak na drzewie opartym na rRNA (73), a *Actinobacteria* z zielonymi bakteriami siarkowymi.

Na bazie metod genomowych (57,58,73,75,78) i czynnika translacji Tu (88) mykoplazmy łączone są z bakteriami gramdodatnimi ubogimi w G+C. Natomiast na niektórych drzewach rRNA (73) i białek rybosomalnych (70) klady te leżą daleko od siebie. Wiele metod genomowych opartych na średnich odległościach i przyrównaniach wielu sekwencji łączy razem krętki i chlamydie (57,58,71,73,83) w przeciwieństwie do drzewa rRNA (73) i metody superdrzewa (75,78).

Na podstawie analiz rybosomalnego RNA bakterie hipertermofilne *Aquifex aeolicus* i *Thermotoga maritima* umiejscowiono u podstawy drzewa filogenetycznego bakterii, co zasugerowało hipotezę głoszącą, że przodek wszystkich bakterii na Ziemi był termofilem (9,84). W wielu metodach genomowych również umieszcza się te bakterie u podstawy pnia prowadzącego do *Bacteria* (57,58,71,73,78). Jednakże, gatunki te charakteryzują się specyficznym składem nukleotydowym i aminokwasowym, dlatego ich grupowanie może być artefaktem (47-49) zwłaszcza, że ich kład nie wykazuje dużej istotności statystycznej (78). Niewykluczone zatem, że ostatni wspólny przodek bakterii nie był termofilem, a *Thermotoga* i *Aquifex* wtórnie przystosowały się do życia w wysokich temperaturach (47,89). Z uwagi na wspólne środowisko życia tych bakterii, ich podobieństwo może być spowodowane horyzontalną wymianą informacji genetycznej lub ewolucją równoległą. Uwzględnienie w analizach dodatkowych przedstawicieli tej grupy może wyjaśnić te wątpliwości.

Na podstawie analiz genomowych zasugerowano, że tradycyjny podział archeonów na *Euryarchaeota* i *Crenarchaeota* może być niepoprawny, gdyż *Euryarchaeota*, jak się wydaje, są grupą parafiletyczną, a *Crenarchaeota* powstały w obrębie *Euryarchaeota* (57,58). Jednakże testy Kashino-Hasegawy na istotność statystyczną topologii drzew nie uprawniają do odrzucenia alternatywnej hipotezy, że *Euryarchaeota* i *Crenarchaeota* są grupami siostrzanymi. Być może parafiletyczność *Euryarchaeota* jest artefaktem, którego źródłem mogła być szybka ewolucja *Thermoplasma* oraz *Halobacterium*, umieszczonej zwykle u podstawy pnia *Archaea* (57). W przeciwieństwie do tych metod *Euryarchaeota* i *Crenarchaeota* są jednak traktowane jako grupy siostrzane w innych analizach genomowych (74,75,78). W większości przypadków metody genomowe, z wyjątkiem (74), są zgodne co do bliskiego pokrewieństwa *Pyrococcus* i metanogenów (57,58,71,75,78).

Na podstawie wyników uzyskanych przy zastosowaniu opisanych metod można skonstruować przynajmniej tymczasowe drzewo konsensusowe *Prokaryota* obejmujące kompletnie zsekwencjonowane genomy (rys. 2). Dla porównania, na rysunku 3 przedstawiono drzewo oparte na analizie 16S rRNA skonstruowane dla bakterii, których genomy zostały kompletnie zsekwencjonowane.



Rys. 2. Tymczasowe drzewo konsensusowe *Prokaryota* skonstruowane na bazie analiz kompletnie zsekwencjonowanych genomów.

5. Zakończenie

Dostępność kompletnie zsekwencjonowanych genomów mikroorganizmów otwiera nowe możliwości badania ich ewolucji. Pokazuje jak duże znaczenie w ich ewolucji i w określaniu powiązań filogenetycznych ma utrata genów i horyzontalny transfer informacji genetycznej. Jednak pomimo tych zjawisk zacierających powiązania rodowe, można z analizowanych genomów wydobyć sygnały filogenetyczne i określić przynajmniej niektóre powiązania filogenetyczne. Rozwój dodatkowych metod badawczych i wzrost liczby kompletnie zsekwencjonowanych genomów ułatwi w przyszłości dokładniejsze odtworzenie drzewa życia (*Tree of Life*).

Rys. 3. Drzewo filogenetyczne sporządzone na podstawie 16S rRNA skonstruowane dla 112 kompletnie zsekwencjonowanych bakterii. Przedstawiciela *Archaea* – *Archaeoglobus fulgidus* (ar f) wybrano do ukorzenia drzewa. Drzewo skonstruowano metodą ME za pomocą programu MEGA 2.1 (90) przy założeniu modelu substytucji Tamura-Nei (91). Przy węzłach głównych grup filogenetycznych zaznaczono wartości bootstrap większe niż 60 (istotności węzłów). Objasnienia skrótów nazw rodzajowych: ag – *Agrobacterium*, aq – *Aquifex*, ba – *Bacillus*, bc – *Bacteroides*, bl – *Bdellovibrio*, bi – *Bifidobacterium*, bl – *Blochmannia*, bo – *Bordetella*, be – *Borrelia*, bd – *Bradyrhizobium*, br – *Brucella*, bu – *Buchnera*, ca – *Campylobacter*, ca – *Caulobacter*, ch – *Chlamydia*, cb – *Chlorobium*, cr – *Chromobacterium*, cl – *Clostridium*, co – *Corynebacterium*, cx – *Coxiella*, de – *Deinococcus*, en – *Enterococcus*, e – *Escherichia*, fu – *Fusobacterium*, ge – *Geobacter*, gl – *Gloeobacter*, ha – *Haemophilus*, he – *Helicobacter*, lb – *Lactobacillus*, lc – *Lactococcus*, le – *Leptospira*, li – *Listeria*, me – *Mesorhizobium*, mb – *Mycobacterium*, mp – *Mycoplasma*, n – *Neisseria*, ni – *Nitrosomonas*, no – *Nostoc*, oc – *Oceanobacillus*, pa – *Pasteurella*, ph – *Photorhabdus*, pt – *Phytoplasma*, pi – *Pirellula*, po – *Porphyromonas*, pr – *Prochlorococcus*, ps – *Pseudomonas*, ra – *Ralstonia*, rh – *Rhodospseudomonas*, ri – *Rickettsia*, sa – *Salmonella*, sh – *Shewanella*, sg – *Shigella*, si – *Sinorhizobium*, st – *Staphylococcus*, sc – *Streptococcus*, sm – *Streptomyces*, so – *Synechococcus*, sy – *Synechocystis*, ta – *Thermoanaerobacter*, ts – *Thermosynechococcus*, th – *Thermotoga*, tr – *Treponema*, to – *Tropheryma*, ur – *Ureaplasma*, vi – *Vibrio*, wi – *Wigglesworthia*, wl – *Wolbachia*, wo – *Wolinella*, xa – *Xanthomonas*, xy – *Xylella*, ye – *Yersinia*.



Literatura

1. Lio P., Goldman N., (1998), *Genome Res.*, 8, 1233-1244.
2. Whelan S., Goldman N., (2001), *Trends Genet.*, 14, 262-272.
3. Stanier R. Y., van Niel C. B., (1962), *Arch. Mikrobiol.*, 42, 17-35.
4. Murray R. G. E., (1986), in: *Bergey's Manual of Systematic Bacteriology*, Eds. Sneath P. H. A., Mair N. S., Sharpe M. E., Holt J. G., 31-34, Williams & Wilkins, Baltimore.
5. Buchanan R. E., Gibbons N. E., (1974), *Bergey's Manual of Determinative Bacteriology*, 8th ed., Williams & Wilkins, Baltimore.
6. Woese C. R., Fox G. E., (1977), *Proc. Natl. Acad. Sci. USA*, 74, 5088-5090.
7. Fox G. E., Stackebrandt E., Hespell R. B., Gibson J., Maniloff J., Dyer T. A., Wolfe R. S., Balch W. E., Tanner R. S., Magrum L. J., et al., (1980), *Science*, 209, 457-463.
8. Woese C. R., (1990), *Science*, 247, 789.
9. Olsen G. J., Woese C. R., Overbeek R., (1994), *J. Bacteriol.*, 176, 1-6.
10. Forterre P., (1997), *ASM News*, 63, 89-92.
11. Brown J. R., Doolittle W. F., (1997), *Microbiol. Mol. Biol. Rev.*, 61, 456-502.
12. Gogarten J. P., Kibak H., Dittrich P., Taiz L., Bowman E. J., Bowman B. J., Manolson M. F., Poole R. J., Date T., Oshima T., et al., (1989), *Proc. Natl. Acad. Sci. USA*, 86, 6661-6665.
13. Iwabe N., Kuma K., Hasegawa M., Osawa S., Miyata T., (1989), *Proc. Natl. Acad. Sci. USA*, 86, 9355-9359.
14. Brown J. R., Doolittle W. F., (1995), *Proc. Natl. Acad. Sci. USA*, 92, 2441-2445.
15. Baldauf S. L., Palmer J. D., Doolittle W. F., (1996), *Proc. Natl. Acad. Sci. USA*, 93, 7749-7754.
16. Lawson F. S., Charlebois R. L., Dillon J. A., (1996), *Mol. Biol. Evol.*, 13, 970-977.
17. Gribaldo S., Cammarano P., (1998), *J. Mol. Evol.*, 47, 508-516.
18. Zillig W., Klenk H. P., Palm P., Leffers H., Puhler G., Garrett R. A., (1989), *Endocytobiosis Cell Res.*, 6, 1-25.
19. Gupta R. S., Golding G. B., (1993) *J. Mol. Evol.*, 37, 573-582.
20. Lake J. A., Rivera M. C., (1994), *Proc. Natl. Acad. Sci. USA*, 91, 2880-2881.
21. Golding G. B., Gupta R. S., (1995), *Mol. Biol. Evol.*, 12, 1-6.
22. Gupta R. S., Golding G. B., (1996), *Trends Biochem. Sci.*, 21, 166-171.

23. Koonin E. V., Mushegian A. R., Galperin M. Y., Walker D. R., (1997), *Mol. Microbiol.*, 25, 619-637.
24. Doolittle W. F., (1999), *Trends Cell. Biol.*, 9, M5-M8.
25. Doolittle W. F., (1999), *Science*, 284, 2124-2129.
26. Eisen J. A., (2000), *Curr. Opin. Genet. Dev.*, 10, 606-611.
27. Garcia-Vallve S., Romeu A., Palau J., (2000), *Genome Res.*, 10, 1719-1725.
28. Koonin E. V., Makarova K. S., Aravind L., (2001), *Annu. Rev. Microbiol.*, 55, 709-742.
29. Ragan M. A., (2001), *Curr. Opin. Genet. Dev.*, 11, 620-626.
30. Koonin E. V., Mushegian A. R., Bork P., (1996), *Trends Genet.*, 12, 334-336.
31. Smith D. R., Doucette-Stamm L. A., Deloughery C., Lee H., Dubois J., (1997), *J. Bacteriol.*, 179, 7135-7155.
32. Klenk H. P., Clayton R. A., Tomb J. F., White O., Nelson K. E., (1997), *Nature*, 390, 364-370.
33. DiRuggiero J., Dunn D., Maeder D. L., Holley-Shanks R., Chatard J., (2000), *Mol. Microbiol.*, 38, 684-693.
34. Aravind L., Tatusov R. L., Wolf Y. I., Walker D. R., Koonin E. V., (1998), *Trends Genet.*, 14, 442-444.
35. Nelson K. E., Clayton R. A., Gill S. R., Gwinn M. L., Dodson R. J., Haft D. H., Hickey E. K., Peterson J. D., Nelson W. C., Ketchum K. A., et al., (1999), *Nature*, 399, 323-329.
36. 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), *Science*, 277, 1453-1474.
37. Lawrence J. G., Ochman H., (1998), *Proc. Natl. Acad. Sci. USA*, 95, 9413-9417.
38. Nolling J., Breton G., Omelchenko M. V., Makarova K. S., Zeng Q., (2001), *J. Bacteriol.*, 183, 4823-4838.
39. Kaneko T., Sato S., Kotani H., Tanaka A., Asamizu E., Nakamura Y., Miyajima N., Hirose M., Sugiyama M., Sasamoto S., et al., (1996), *DNA Res.*, 3, 109-136.
40. Subramanian G., Koonin E. V., Aravind L., (2000), *Infect. Immun.*, 68, 1633-1648.
41. Teichmann S. A., Mitchison G., (1999), *J. Mol. Evol.*, 49, 98-107.
42. Nesbo C. L., Boucher Y., Doolittle W. F., (2001), *J. Mol. Evol.*, 53, 340-350.
43. Asai T., Zaporozhets D., Squires C., Squires C. L., (1999), *Proc. Natl. Acad. Sci. USA*, 96, 1971-1976.
44. Ueda K., Seki T., Kudo T., Yoshida T., Kataoka M., (1999), *J. Bacteriol.*, 181, 78-82.
45. Yap W. H., Zhang Z., Wang Y., (1999), *J. Bacteriol.*, 181, 5201-5209.
46. Brochier C., Philippe H., Moreira D., (2000), *Trends Genet.*, 16, 529-533.
47. Forterre P., (1996), *Cell*, 85, 789-792.
48. Philippe H., Laurent J., (1998), *Curr. Opin. Genet. Dev.*, 8, 616-623.
49. Galtier N., Tourasse N., Gouy M., (1999), *Science*, 283, 220-221.
50. Wolf Y. I., Rogozin I. B., Grishin N. V., Koonin E. V., (2002), *Trends Genet.*, 18, 472-479.
51. Huynen M. A., Snel B., Bork P., (1999), *Science*, 286, 1443a.
52. Snel B., Bork P., Huynen M. A., (1999), *Nat. Genet.*, 21, 108-110.
53. Fitz-Gibbon S. T., House C. H., (1999), *Nucleic Acids Res.*, 27, 4218-4222.
54. Tekaia F., Lazcano A., Dujon B., (1999), *Genome Res.*, 9, 550-557.
55. Lin J., Gerstein M., (2000), *Genome Res.*, 10, 808-818.
56. Natale D. A., Shankavaram U. T., Galperin M. Y., Wolf Y. I., Aravind L., Koonin E. V., (2000), *Genome Biol.*, 1, RESEARCH0009.
57. Wolf Y. I., Rogozin I. B., Grishin N. V., Tatusov R. L., Koonin E. V., (2001), *BMC Evol. Biol.*, 1, 8.
58. Clarke G. D., Beiko R. G., Ragan M. A., Charlebois R. L., (2002), *J. Bacteriol.*, 184, 2072-2080.
59. House C. H., Fitz-Gibbon S. T., (2002), *J. Mol. Evol.*, 54, 539-547.
60. Korbel J. O., Snel B., Huynen M. A., Bork P., (2002), *Trends Genet.*, 18, 158-162.
61. Lathe W. C. 3rd, Snel B., Bork P., (2000), *Trends Biochem. Sci.*, 25, 474-479.
62. Wolf Y. I., Rogozin I. B., Kondrashov A. S., Koonin E. V., (2001), *Genome Res.*, 11, 356-372.
63. Mushegian A. R., Koonin E. V., (1996), *Trends Genet.*, 12, 289-290.
64. Kolsto A. B., (1997), *Mol. Microbiol.*, 24, 241-248.
65. Watanabe H., Mori H., Itoh T., Gojobori T., (1997), *J. Mol. Evol.*, 44(Suppl. 1), S57-S64.
66. Bellgard M. I., Itoh T., Watanabe H., Imanishi T., Gojobori T., (1999), *Ann. N. Y. Acad. Sci.*, 18, 293-300.
67. Itoh T., Takemoto K., Mori H., Gojobori T., (1999), *Mol. Biol. Evol.*, 16, 332-346.

68. Hughes D., (2000), *Genome Biol.*, 1, REVIEWS0006.
69. Grishin N. V., Wolf Y. I., Koonin E. V., (2000), *Genome Res.*, 10, 991-1000.
70. Hansmann S., Martin W., (2000), *Int. J. Syst. Evol. Microbiol.*, 50, 1655-1663.
71. Brown J. R., Douady C. J., Italia M. J., Marshall W. E., Stanhope M. J., (2001), *Nat. Genet.*, 28, 281-285.
72. Brochier C., Philippe H., (2002), *Nature*, 417, 244.
73. Brochier C., Bapteste E., Moreira D., Philippe H., (2002), *Trends Genet.*, 18, 1-5.
74. Matte-Tailliez O., Brochier C., Forterre P., Philippe H., (2002), *Mol. Biol. Evol.*, 19, 631-639.
75. Daubin V., Gouy M., Perriere G., (2001), *Genome Inform. Ser. Workshop Genome Inform.*, 12, 155-164.
76. Sicheritz-Ponten T., Andersson S. G., (2001), *Nucleic Acids Res.*, 29, 545-552.
77. Zhaxybayeva O., Gogarten J. P., (2002), *BMC Genomics*, 3, 4.
78. Calteau A., Daubin V., Perriere G., (2004), *Lecture Notes in Computer Science*, 3039, 700-708.
79. Gupta R. S., (1998), *Microbiol. Mol. Biol. Rev.*, 62, 1435-1491.
80. Nakashima H., Ota M., Nishikawa K., Ooi T., (1998), *DNA Res.*, 5, 251-259.
81. Pride D. T., Meinersmann R. J., Wassenaar T. M., Blaser M. J., (2003), *Genome Res.*, 13, 145-158.
82. Tekaiia F., Yeramian E., Dujon B., (2002), *Gene*, 297, 51-60.
83. Qi J., Wang B., Hao B. I., (2004), *J. Mol. Evol.*, 58, 1-11.
84. Woese C. R., (1987), *Microbiol. Rev.*, 51, 221-271.
85. Ludwig W., Klenk H.-P., (2001), in: *Bergey's Manual of Systematic Bacteriology*, Eds. Boone D. R., Castenholz R. W., 49-65, Springer-Verlag, Berlin.
86. Gupta R. S., (2000), *FEMS Microbiol. Rev.*, 24, 367-402.
87. Ludwig W., Strunk O., Klugbauer S., Klugbauer N., Weizenegger M., Neumaier J., Bachleitner M., Schleifer K. H., (1998), *Electrophoresis*, 19, 554-568.
88. Kamla V., Henrich B., Hadding U., (1996), *Gene*, 171, 83-87.
89. Miller S. L., Lazcano A., (1995), *J. Mol. Evol.*, 41, 689-692.
90. Kumar S., Tamura K., Nei M., (1993), *MEGA: Molecular Evolutionary Genetics Analysis*, Pennsylvania State University, University Park, PA.
91. Tamura K., Nei M., (1993), *Mol. Biol. Evol.*, 10, 512-526.