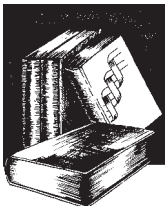


PRACE PRZEGLĄDOWE

Genomika – dziedzina wiedzy XXI wieku

Paweł Mackiewicz[1], Jolanta Zakrzewska-Czerwińska[2], Stanisław Cebrat[1]

[1]Zakład Genomiki, Instytut Genetyki i Mikrobiologii, Uniwersytet Wrocławski, Wrocław
[2]Zakład Mikrobiologii, Instytut Immunologii i Terapii Doświadczalnej im. L. Hirszfelda, Polska Akademia Nauk, Wrocław

Genomics – science of the 21st century

Summary

Genomics is a new field of biology. Its fast development is caused mainly by quick progress in large-scale genome sequencing and in computer technology. In spite of a huge number of sequenced microbial genomes available in databases, their taxonomical diversity is biased and reflects the interests of researches and facility of microorganisms' isolation and culture in laboratory conditions. More than 80% of genome sequencing projects are focused on the members of Proteobacteria, Firmicutes and Actinobacteria. Environmental genome shotgun sequencing reveals that microbial diversity is much greater than we expected. Particular levels of genomic analysis, the problems and subjects of genomics are specified and described here.

Key words:
genomics, bioinformatics, microbial genomes, microbial diversity.

Adres do korespondencji

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

biotechnologia
3 (70) 7–21 2005

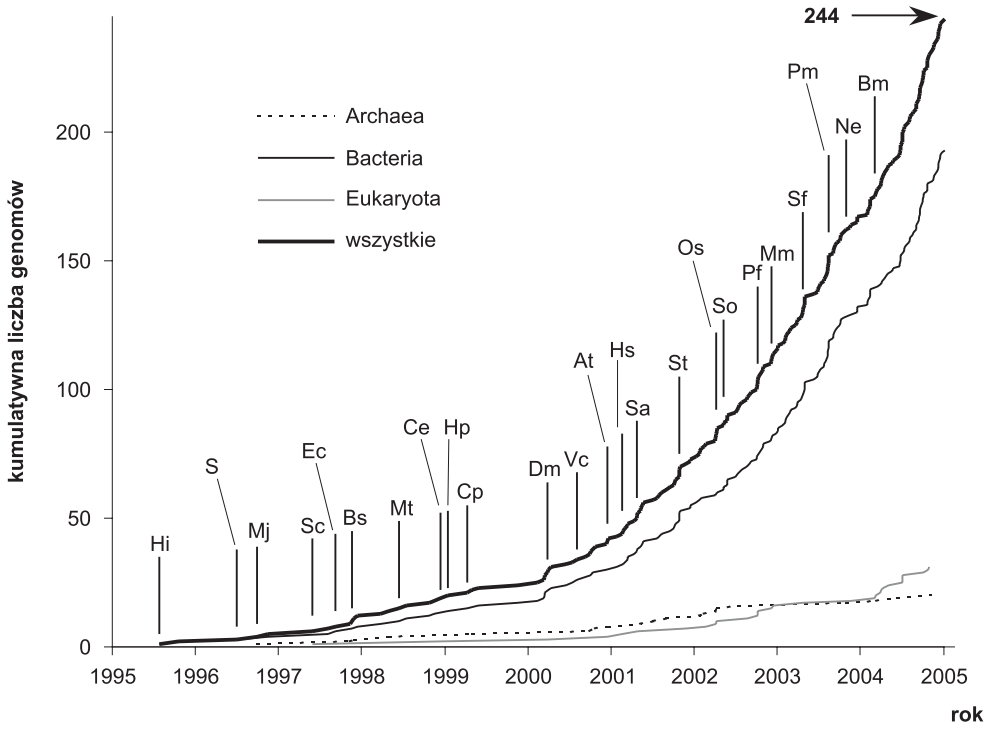
1. Postępy w sekwencjonowaniu genomów

Genomika, czyli nauka o genomach jest stosunkowo nową, ale dynamicznie rozwijającą się dziedziną biologii. Do powstania genomiki przyczynił się intensywny rozwój technik biologii molekularnej, który umożliwił podjęcie i zrealizowanie projektów sekwencjonowania całych genomów. Pierwszym zsekwencjonowanym genomem był genom bakteriofaga MS2, zbudowany z RNA o długości 3569 nukleotydów (1). Przełomem stało się

wprowadzenie w 1977 r. technik sekwencjonowania DNA przez Sangera i wsp. (2) oraz Maxama i Gilberta (3). Szczególnie popularna stała się metoda z użyciem dideoksynukleotydów Sangera zwana metodą terminacji łańcucha. Pozwoliła ona na poznanie sekwencji całego genomu faga ΦX174 o długości 5,4 tys. nukleotydów, opublikowanej w 1977 r. (4). Kolejnymi zsekwencjonowanymi genomami był genom mitochondrialny człowieka o długości 16,6 tys. pz (5) oraz faga λ o długości 48,5 tys. pz (6) – bardzo popularnego modelowego obiektu wielu badań molekularnych i genetycznych.

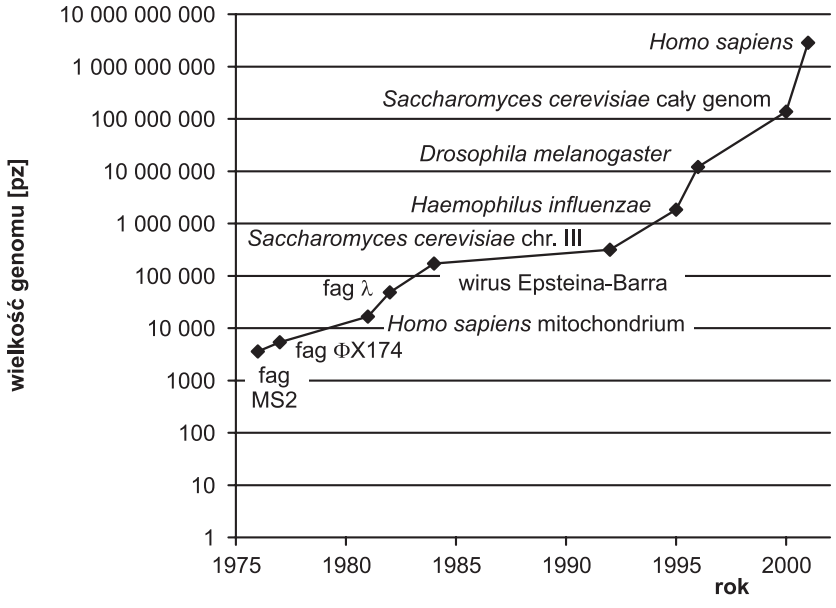
W latach osiemdziesiątych XX w. sekwencjonowanie małych genomów stało się już stosunkowo proste i mało kosztowne, co doprowadziło do opublikowania sekwencji genomów wielu wirusów i organelli komórkowych. Jednak analiza sekwencji dużych genomów wciąż była poza zasięgiem ówczesnych możliwości. Dlatego za ważne wydarzenie uznano poznanie pełnej sekwencji (315 tys. pz) – chromosomu III drożdży *Saccharomyces cerevisiae* (7). Sekwencja całego genomu drożdży o długości ponad 12 milionów pz została opublikowana na początku 1996 r. (8). Krokiem milowym w genomice stało się wprowadzenie nowych technik sekwencjonowania dużych genomów, tak zwaną metodą *shotgun* („strzału na ślepo") polegającej na sekwencjonowaniu dużej liczby sekwencji generowanych przez losowe fragmentowanie genomu, które następnie są składane komputerowo (9). To właśnie wprowadzenie metod obliczeniowych składających setki tysięcy losowo uzyskanych sekwencji DNA (początkowo o długości 300-500 pz, a obecnie do 1500 pz) w dłuższe fragmenty zmniejszyło znacznie koszty i skróciło czas sekwencjonowania, eliminując tradycyjne metody polegające na żmudnym i czasochłonnym mapowaniu oraz składaniu kolejno ułożonych kosmidów lub subklonów (10). Dzięki metodzie *shotgun*, jeszcze przed ogłoszeniem kompletnej sekwencji genomu drożdży, opublikowano sekwencję genomu bakterii *Haemophilus influenzae* – 1,8 mln pz (11), a tuż po nim genomu *Mycoplasma genitalium* – 0,6 mln pz (12).

Od tego czasu można obserwować w przybliżeniu wykładniczy wzrost liczby kompletnie zsekwencjonowanych genomów i intensywny rozwój genomiki (rys. 1). Na początku 2005 r. liczba zsekwencjonowanych genomów wynosiła 244 (wg bazy danych GOLD, www.genomesonline.org; 13, 14), w tym z królestwa *Archaea* – 20, *Bacteria* – 193, *Eukaryota* – 31. Znaczny udział, jak widać, stanowią genomy *Prokaryota*. Licząc od 1999 r. liczba poznawanych genomów podwaja się średnio co 15 miesięcy, a od 2000 r. co miesiąc publikowane są średnio sekwencje czterech genomów. Według bazy danych GOLD na początku 2005 r. rozpoczętych było 1000 projektów sekwencjonowania różnych genomów (w tym: *Archaea* – 27, *Bacteria* – 509, *Eukaryota* – 464). Zakładając, że dotychczasowe tempo przyrostu liczby zsekwencjonowanych genomów prokariotycznych będzie się utrzymywać, to do 2030 r. poznamy ponad 5400 genomów. Dla porównania liczba znanych gatunków *Prokaryota* wynosi obecnie 5536 (według DSMZ Bacterial Nomenclature Up-to-date, www.dsmz.de/bactnom/bactname.htm).



Rys. 1. Skumulowana liczba kompletnie zsekwencjonowanych genomów z podziałem na trzy królestwa (według danych z bazy GOLD). Pionowymi liniami zaznaczono czas opublikowania sekwencji niektórych organizmów istotnych z punktu widzenia: poznawczego, biotechnologicznego lub medycznego Hi – *Haemophilus influenzae* KW20 (pierwszy zsekwencjonowany organizm komórkowy, patogen), S – *Synechocystis* sp. PCC6803 (sinica), Mj – *Methanococcus jannaschii* DSM 2661 (archeon), Sc – *Saccharomyces cerevisiae* S288C (*Eukaryota*, drożdże, organizm modelowy, znaczenie biotechnologiczne), Ec – *Escherichia coli* K12 (organizm modelowy, fakultatywny patogen), Bs – *Bacillus subtilis* 168 (organizm modelowy), Mt – *Mycobacterium tuberculosis* H37Rv (patogen), Ce – *Caenorhabditis elegans* (*Eukaryota*, nicień, organizm modelowy), Hp – *Helicobacter pylori* J99 (patogen), Cp – *Chlamydophila pneumoniae* CWL029 (patogen), Dm – *Drosophila melanogaster* (muszka owocowa, organizm modelowy), Vc – *Vibrio cholerae* N16961 (patogen), At – *Arabidopsis thaliana* (roślina, rzodkiewnik pospolity organizm modelowy), Hs – *Homo sapiens*, Sa – *Staphylococcus aureus* N315 (MRSA) (patogen), St – *Salmonella typhi* CT18 (patogen), Os – *Oryza sativa japonica* (ryż, znaczenie gospodarcze), So – *Streptomyces coelicolor* A3(2) (wytwarzanie antybiotyków), Pf – *Plasmodium falciparum* 3D7 (pierwotniak, zarodziec sierpowaty, patogen), Mm – *Mus musculus* (ssak, mysz, organizm modelowy), Sf – *Shigella flexneri* 2a 2457T (patogen), Pm – *Prochlorococcus marinus* CCMP1375 (SS120) (sinica, znaczenie ekologiczne), Ne – *Nanoarchaeum equitans* Kin4-M (przedstawiciel nowej grupy *Archaea*), Bm – *Bombyx mori* p50T (jedwabnik morwowy, znaczenie przemysłowe).

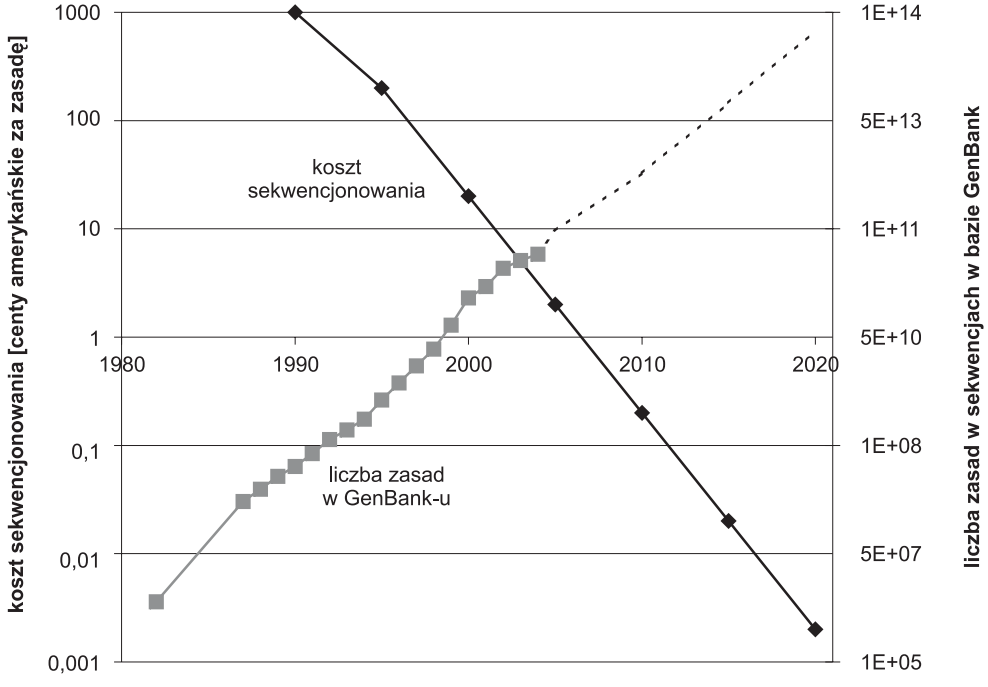
W wykładniczy sposób rośnie również wielkość sekwencjonowanych chromosomów i genomów (rys. 2). Zsekwencjonowane dotychczas chromosomy organizmów prokariotycznych charakteryzują się dużym zróżnicowaniem wielkości: *Archaea* od 0,5 mln pz (*Nanoarchaeum equitans*) do 5,8 mln pz (*Methanosarcina acetivorans)*, *Bacteria* od 0,58 mln pz (*Mycoplasma genitalium*) do 9,1 mln pz (*Bradyrhizobium japonicum*).



Rys. 2. Wzrost wielkości kolejno sekwencjonowanych chromosomów lub genomów. Oś Y przedstawiono w skali logarytmicznej.

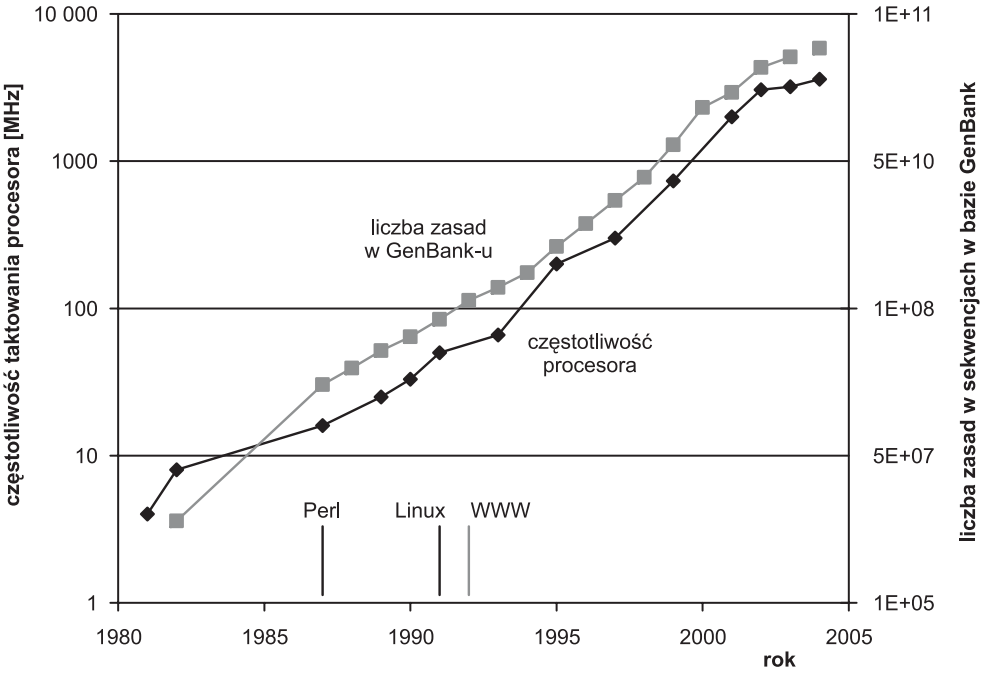
Wśród *Eukaryota* najmniejszym kompletnie zsekwencjonowanym genomem jest genom pasożytniczego grzyba *Encephalitozoon cuniculi* o wielkości 2,5 mln pz, a największym – genom człowieka o wielkości 3,1 mld pz. Największym znanym genomem eukariotycznym czekającym na zsekwencjonowanie jest genom ameby *Amoeba dubia* o wielkości aż 670 mld pz.

Ogromnym przyśpieszeniem sekwencjonowania, obniżenia kosztów i zwiększenia dokładności odczytów było wprowadzenie elektroforezy kapilarnej i znakowanie nukleotydów fluorochromami, co pozwoliło na zautomatyzowanie całego procesu. Koszty sekwencjonowania w przeliczeniu na zasadę zmniejszają się dwukrotnie co 18 miesięcy, co daje 10-krotny spadek kosztów co 5 lat (15). W 1995 r. sekwencjonowanie kosztowało 100-300 centów amerykańskich za zasadę, a w 2000 r. już tylko 10-30 centów amerykańskich. Zakładając, że w roku 2020 koszty te będą wynosić 0,001-0,003 centów amerykańskich za zasadę, a na sekwencjonowanie będzie się przeznaczać rocznie 1 miliard USD, to za 15 lat będzie się uzyskiwać sekwencje odpowiadające prawie 17 tysiącom genomów człowieka (5×10^{13} par zasad na rok). Sugeruje to, że tempo przyrostu danych sekwencyjnych będzie jeszcze bardziej rosnąć. Jest to wariant optymistyczny, ponieważ w końcu i tak dojdzie się do granic możliwości stosowanych technologii, wynikających po prostu z ograniczeń praw przyrody.



Rys. 3. Zależność między kosztem sekwencjonowania a liczbą zasad w sekwencjach deponowanych w bazie GenBank (www.ncbi.nlm.nih.gov/Entrez). Linią przerywaną zaznaczono przewidywany wzrost liczby zasad w przyszłości według (15). Obie osie Y przedstawiono w skali logarytmicznej.

Rola komputerów sprowadza się nie tylko do składania zsekwencjonowanych fragmentów oraz gromadzenia danych w postaci skomputeryzowanej, ale również do analiz sekwencji, np. rozpoznawania sekwencji kodujących, poszukiwania sekwencji podobnych, porównywania genomów, czy przewidywania struktur białek. Nieocenioną rolę odgrywa także internet, który umożliwia szybki dostęp do gromadzonych danych oraz ich przesyłanie między badaczami, centrami sekwencjonującymi genomy oraz bazami danych. Widać wyraźny związek między liczbą gromadzonych sekwencji a rozwojem technologii komputerowych – mierzonych szybkością procesorów lub pojemnością twardych dysków, opisywanego najczęściej prawem Moore'a mówiącego, że wydajność komputerów ulega podwojeniu co około 18 miesięcy (rys. 4). W podobnym tempie podwaja się liczba danych w GenBank-u (co 14 miesięcy). Znaczna część analiz genomicznych jest przeprowadzana za pomocą różnorodnych metod obliczeniowych, zaawansowanych algorytmów i skomputeryzowanego sprzętu, dlatego genomika jest ściśle powiązana z bioinformatyką – również intensywnie rozwijającą się dziedziną interdyscyplinarną łączącą biologię z naukami i technikami informatycznymi oraz obliczeniowymi.



Rys. 4. Związek między liczbą gromadzonych sekwencji w bazie GenBank a rozwojem technologii komputerowych mierzonych częstotliwością procesora. Zaznaczono również ważne daty dla bioinformatyki: wprowadzenie języka programowania Perla i systemu Linux oraz początek powszechnego działania sieci www.

2. Zróżnicowanie filogenetyczne sekwencjonowanych genomów prokariotycznych

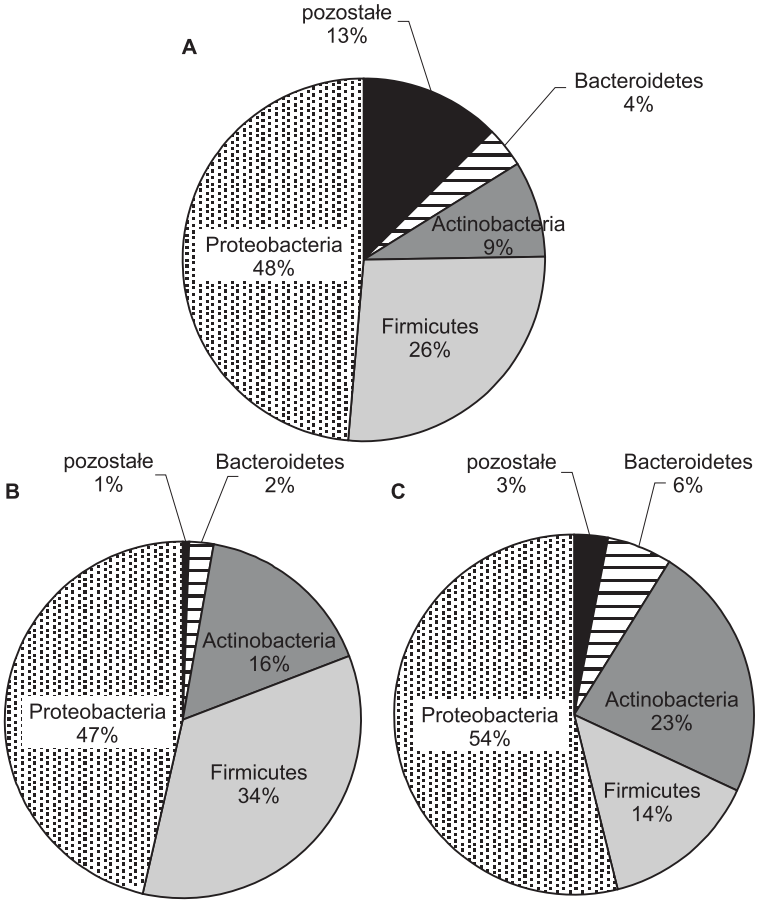
Najwięcej poznanych genomów należy do organizmów prokariotycznych, co pozwala przyjrzeć się ich zróżnicowaniu filogenetycznemu. Jednak pomimo dużej liczby genomów już zsekwencjonowanych lub będących w trakcie sekwencjonowania (w sumie 751), nie reprezentują one równomiernie większych grup filogenetycznych (tab.). Najsłabiej są reprezentowane grupy królestwa *Archaea* (47 projektów ukończonych i nie ukończonych). Projektów dotyczących genomów z królestwa *Bacteria* jest aż 704. Wśród nich dominują trzy grupy bakterii: Proteobacteria, stanowiących prawie połowę wszystkich poznawanych genomów, Firmicutes – ponad 1/4 projektów i Actinobacteria – prawie 10% (rys. 5A). Wśród projektów proteobakterii dominują gamma-proteobakterie – 25% wszystkich projektów. Kolejnymi grupami wybieranymi do analiz genomowych są bakterie z grupy Bacteroidetes/Chlorobi, sinice, chlamydie i krętki.

Tabela

Zróżnicowanie taksonomiczne projektów, ukończonych i nie ukończonych, związanych z sekwencjonowaniem genomów prokariotycznych

Główne grupy filogenetyczne	Liczba projektów	
ARCHAEA	**47**	**(100%)**
Euryarchaea	35	(74,5%)
Crenarchaea	11	(23,4%)
Nanoarchaeota	1	(2,1%)
Korarchaeota	0	(0%)
BACTERIA	**704**	**(100%)**
Actinobacteria (promieniowce)	60	(8,5%)
Aquificae	4	(0,6%)
Grupa Bacteroidetes/Chlorobi	26	(3,7%)
Bacteroidetes	15	(2,1%)
Chlorobi	11	(1,6%)
Grupa Chlamydiae/Verrucomicrobia	19	(2,7%)
Chlamydiae	17	(2,4%)
Verrucomicrobia	2	(0,3%)
Chloroflexi (bakterie zielone niesiarkowe)	3	(0,4%)
Chrysiogenetes	1	(0,1%)
Cyanobacteria (sinice)	26	(3,7%)
Deferribacteres	0	(0%)
Deinococcus-Thermus	5	(0,7%)
Dictyoglomi	1	(0,1%)
Grupa Fibrobacteres/Acidobacteria	5	(0,7%)
Acidobacteria	4	(0,6%)
Fibrobacter	1	(0,1%)
Firmicutes (bakterie gramdodatnie)	187	(26,6%)
Fusobacteria	3	(0,4%)
Gemmatimonadetes	0	(0%)
Nitrospirae	2	(0,3%)
Planctomycetes	5	(0,7%)
Proteobacteria (bakterie purpurowe)	343	(48,7%)
Alpha-Proteobacteria	81	(11,5%)
Beta-Proteobacteria	46	(6,5%)
Gamma-Proteobacteria	176	(25%)
Delta-Proteobacteria	23	(3,3%)
Epsilon-Proteobacteria	17	(2,4%)
Spirochaetes (krętki)	10	(1,4%)
Thermodesulfobacteria	1	(0,1%)
Thermotogae	3	(0,4%)

Liczby projektów pochodzą z bazy danych GOLD (www.genomesonline.org) z początku 2005 r., a podział filogenetyczny z bazy NCBI (www.ncbi.nlm.nih.gov/Taxonomy).



Rys. 5. Zróżnicowanie filogenetyczne mikroorganizmów w obrębie: (**A**) 704 sekwencjonowanych genomów prokariotycznych (baza GOLD, www.genomesonline.org); (**B**) 177 szczepów izolowanych z próbek środowiskowych, weterynaryjnych i klinicznych (18); (**C**) 3767 kultur prokariotów pochodzących z Australijskiej Kolekcji Mikroorganizmów (www.biosci.uq.edu.au/micro/culture/culture.htm).

Selektywny wybór genomów do sekwencjonowania wynika z trudności izolowania niektórych mikroorganizmów ze środowiska oraz ich dalszego hodowania w klasycznych warunkach laboratoryjnych. Dotyczy to szczególnie mikroorganizmów, w tym wielu przedstawicieli *Archaea*, żyjących w skrajnych warunkach środowiskowych: halofili, termofili i acidofili. Większość znanych i analizowanych genomów należy do mikroorganizmów charakteryzujących się szybkim wzrostem na standardowych, sztucznych podłożach w warunkach tlenowych i przy średnich temperaturach. Ocenia się, że te organizmy stanowią mniej niż 1% całego świata mikroorganizmów, czyli do odkrycia i zbadania pozostaje aż 99% pozostałych (16). Rzeczywiście, przedstawiony rozkład sekwencjonowanych genomów odpowiada proporcjom

szczepów w kolekcjach mikoroorganizmów w poszczególnych grupach taksonomicznych (17) – rysunek 5. W przeprowadzonych badaniach taksonomicznych – 177 środowiskowych, weterynaryjnych i klinicznych izolatów wykazano (18), że z wyjątkiem jednego, wszystkie należały tylko do czterech grup bakterii: Proteobacteria (82 izolaty), Firmicutes (61 izolatów), Actinobacteria (29 izolatów) i Bacteroidetes (4 izolaty), które są również podobnie reprezentowane w sekwencjonowanych genomach. Również w Australijskiej Kolekcji Mikroorganizmów, 97% szczepów należy właśnie do tych czterech grup (www.biosci.uq.edu.au/micro/culture/culture.htm).

Obserwowany rozkład taksonomiczny sekwencjonowanych genomów podyktowany jest również dotychczasowymi zainteresowaniami badaczy i wyborem do sekwencjonowania genomów tych bakterii, które są organizmami modelowymi lub mają duże znaczenie w biotechnologii, rolnictwie, przemyśle, ekologii i medycynie (19). Osiem najintensywniej badanych rodzajów należy do trzech grup bakterii: Proteobacteria (*Escherichia*, *Helicobacter*, *Pseudomonas*, *Salmonella*), Firmicutes (*Bacillus*, *Streptococcus*, *Staphylococcus*) i Actinobacteria (*Mycobacterium*) – (20). Według danych z bazy GOLD prawie wszystkie z 751 prokariotycznych projektów sekwencjonowania dotyczy patogenów (52%) lub organizmów wykorzystywanych w biotechnologii (47%), a zaledwie 1% jest związanych z badaniami podstawowymi w ramach projektu *Tree of Life,* dotyczącego badania zróżnicowania mikroorganizmów w naszej biosferze i odkrywania nie znanych jeszcze gatunków.

O tym jak mało reprezentatywna jest nasza wiedza o puli mikroorganizmów całej biosfery i jak wiele zostało do zbadania, mogą świadczyć badania przeprowadzone przez Ventera i wsp. (21), którzy zastosowali sprawdzoną już metodę *shotgun* do losowego sekwencjonowania genomów należących do populacji mikroorganizmów „wychwyconych" przez filtry z około 1500 litrów wody, pochodzącej z Morza Sargassowego. Było to przedsięwzięcie na dotychczas nie znaną skalę. Zsekwencjonowano w sumie 1,045 miliarda par zasad i zidentyfikowano 1,2 miliona genów o łącznej długości 700 mln pz. Z tego prawie 70 tysięcy okazało się nowymi genami (w tym 782 geny kodujące fotoreceptory podobne do rodopsyny). Oszacowano, że w pobranych próbkach znajdowało się co najmniej 1800 gatunków genomowych. Przyjmując, że podobieństwo sekwencji rRNA mniejsze niż 97% upoważnia do wyróżnienia nowego gatunku, zidentyfikowano 148 potencjalnie nowych gatunków bakteryjnych. Na podstawie genów 16S rRNA, RecA, czynników elongacji Tu i G, HSP70 oraz RNA-zależnej polimerazy B (RpoB) oszacowano, że w próbkach najbardziej reprezentowane były geny należące do proteobakterii, z przewagą grup alfa oraz gamma, następnie sinic, Firmicutes, Actinobacteria oraz Bacteroidetes/Chlorobi. Na pierwszy rzut oka taki rozkład taksonomiczny bardzo przypomina wcześniej opisane proporcje. Jednak, jak sami autorzy stwierdzają, taki wynik jest skutkiem specyficzności reakcji PCR służącej do powielania próbek sekwencji, która prowadzi do preferencyjnego powielania genów występujących w większej liczbie kopii zarówno w danym genomie, jak i w całej populacji (z powodu dominacji danego gatunku w próbce). Gamma-proteobakterie rzeczywiście mają przynajmniej 5 kopii operonu

rRNA, a nadreprezentowane klony stanowiące 53% tych sekwencji należały tylko do dwóch powszechnych gatunków proteobakterii *Shewanella* i *Burkholderia*. Wiele mikroorganizmów nie zostało zidentyfikowanych z powodu ich małej reprezentacji w próbkach oraz wielkości porów w stosowanych filtrach (0,1-3 μm). Szacunki wskazują, że 80% mikroorganizmów (około 47 700 „gatunków") to rzadko występujące organizmy znajdujące się poniżej progu detekcji w tego typu badaniach. Venter, inspirowany podróżą Karola Darwina dookoła świata, zamierza w czasie podobnej podróży pobierać próbki oceanu co 200 mil. Może te badania pozwolą nam przynajmniej w części poznać jak duże jest zróżnicowanie świata mikroorganizmów.

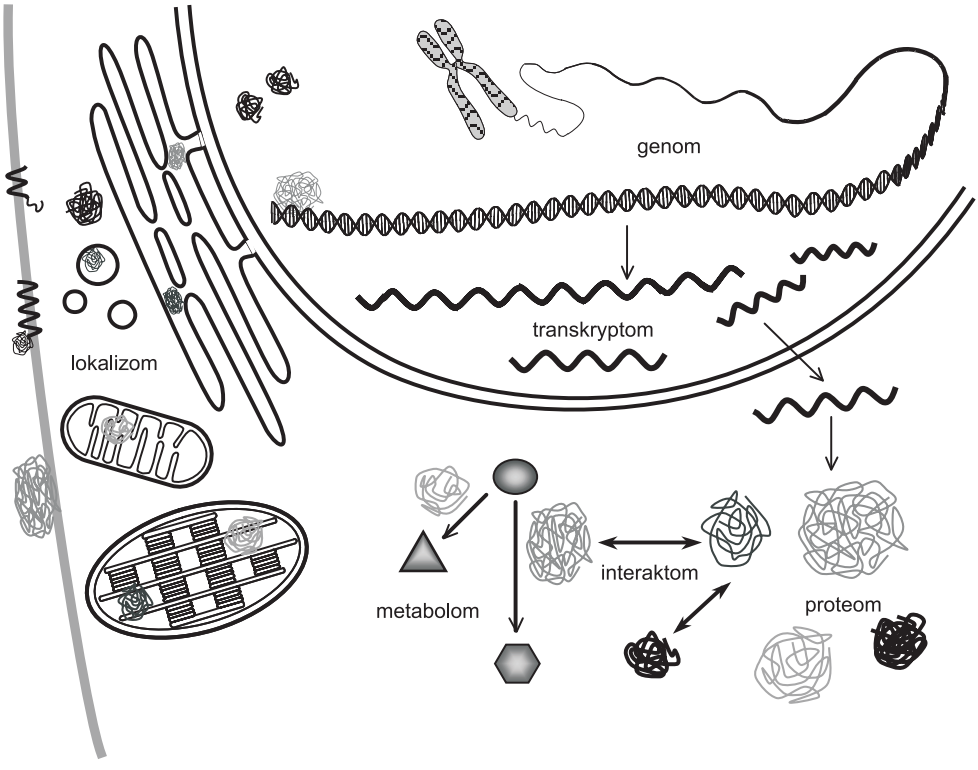
Warto dodać, że szczególnie słabo reprezentowane są bardzo zróżnicowane grupy mikroorganizmów eukariotycznych znajdujących się u podstawy pnia *Eukaryota,* mimo że liczba projektów sekwencjonowania genomów tych grup (w sumie 125) jest porównywalna z liczbą projektów innych grup (Fungi – 122, Viridiplantae – 85, Metazoa – 165).

3. Poziomy analiz i problematyka badań genomiki

Po etapie sekwencjonowania, uzyskane sekwencje są deponowane w postaci elektronicznej w bazach danych pierwotnych (archiwalnych), najczęściej w GenBank (Stany Zjednoczone), EMBL (European Molecular Biology Laboratory Nucleotide Sequence Database, Wielka Brytania) lub DDBJ (DNA Data Bank of Japan, Mishima, Japonia), które tworzą jedno konsorcjum – The International Sequence Database Collaboration i wymieniają się na bieżąco danymi. Natomiast baza PDB (The Protein Data Bank) gromadzi głównie informacje z badań eksperymentalnych dotyczące struktur przestrzennych białek. Z tych danych korzystają tzw. bazy wtórne (pochodne), np. SWISS-PROT/TrEMBL (przy EMBL) czy PIR (The Protein Information Resource), które w różny sposób przetwarzają informacje o sekwencjach i strukturach. Osobną kategorię stanowią wyspecjalizowane bazy poświęcone określonym genomom i innym zagadnieniom genomicznym oraz bioinformatycznym. Wyczerpujące informacje o bazach danych można znaleźć w artykułach w każdym pierwszym numerze roku czasopisma "Nucleic Acids Research", na stronie którego znajduje się ich przejrzysta klasyfikacja oraz łącza do ich stron internetowych (http://nar.oupjournals.org). W numerze z 2005 r. zebrano i opisano 719 baz.

Zgromadzone sekwencje są przedmiotem dalszych analiz, których celem jest zdobycie jak najwięcej informacji o danym genomie i funkcjonowaniu komórki gospodarza. Poziomy analiz informacji genomu odpowiadają etapom ekspresji tej informacji w komórkach (rys. 6):

– **Genom** – wszystkie sekwencje DNA zawarte w organizmie (lub RNA w przypadku niektórych wirusów). Jego bezpośrednia analiza dotyczy głównie rozpoznawania sekwencji kodujących, sekwencji regulatorowych i sekwencji powtórzonych oraz określania ogólnej organizacji, np. zróżnicowania składu nukleotydowego



Rys. 6. Poziomy analiz organizmu (komórki), którego genom został zsekwencjonowany, odpowiadające organizacji informacji biologicznej i jej ekspresji.

w regionach chromosomu, rozmieszczenia genów na chromosomie, organizacji genów w operony.

– **Transkryptom** – wszystkie sekwencje RNA syntetyzowane (transkrybowane) w organizmie. Analiza skupia się na regulacji ekspresji genów w różnorodnych warunkach i/lub tkankach. Badania są przeprowadzane za pomocą mikromatryc oligonukleotydowych i cDNA, popularnie zwanych chipami DNA. Duże nadzieje w zrozumieniu funkcjonowania komórki wiąże się ze stosunkowo niedawnym odkryciem zjawiska zwanego interferencją RNA (RNAi – RNA *interference*) i rolą różnych niskocząsteczkowych RNA w regulacji ekspresji genów, organizacji materiału genetycznego i ochronie przed pasożytami.

– **Proteom** – wszystkie białka wytwarzane w organizmie. Analizy dotyczą identyfikowania konserwatywnych regionów i motywów w sekwencjach, przewidywania struktur drugorzędowych oraz przestrzennych. Białka i ich struktury są klasyfikowane w różne grupy, np. rodziny i nadrodziny. Zidentyfikowanym białkom przypisywana jest kategoria funkcjonalna i określana jest ich rola w komórce.

– **Lokalizom** – opisuje subkomórkowe położenie białek w komórce. Analizy komputerowe dotyczą poszukiwania swoistych motywów w sekwencjach aminokwasowych oraz peptydów sygnałowych i tranzytowych kierujących sekwencje do odpowiednich przedziałów komórki.

– **Interaktom** – dotyczy zależności i interakcji między makrocząsteczkami w komórce. Obecnie najintensywniej są badane oddziaływania między białkami. Są one przedstawiane za pomocą sieci zależności.

– **Metabolom** – opisuje wszystkie szlaki metaboliczne, łącznie z metabolitami i procesami zachodzącymi w organizmie. Celem dotychczasowych badań jest określenie, jakie szlaki metaboliczne funkcjonują w danym organizmie, głównie na podstawie analiz porównawczych między genomami. Znajomość występowania lub braku danego szlaku może mieć duże znaczenie praktyczne w biotechnologii i medycynie.

Genomika jest dziedziną nową, intensywnie rozwijającą się, dlatego stosowana terminologia nie jest jeszcze ustalona i często tym samym terminom różni autorzy przypisują odmienne znaczenia. Genomika w swoich analizach bardzo często posługuje się skomplikowanymi algorytmami i technikami obliczeniowymi ze względu na złożoność badanych problemów i dlatego często określa się ją terminem **genomika obliczeniowa** (22). Natomiast termin **genomika funkcjonalna** często stosuje się w celu określenia badań eksperymentalnych związanych z analizą genomu, przeprowadzanych w skali całego genomu, np. inaktywacji genów w celu zidentyfikowania efektów fenotypowych, analizy interakcji między białkami w systemach dwuhybrydowych, lokalizowania białek w komórce za pomocą różnych znaczników, analizy ekspresji genów za pomocą chipów DNA, izolowania i charakterystyki elektroforetycznej i strukturalnej białek oraz identyfikowania kompleksów białek za pomocą spektrometrii masowej. Granica między tymi dziedzinami jest jednak płynna. Genomika obliczeniowa stara się również interpretować wyniki eksperymentalne za pomocą technik obliczeniowych. Koonin i Galperin (23) proponują również inne użycie terminu genomika funkcjonalna, jako odpowiednika genomiki strukturalnej. W tym znaczeniu termin ten dotyczyłby badań genomiki obliczeniowej proponujących potencjalne cele (geny) do dalszych badań eksperymentalnych, w celu dokładniejszego określenia ich funkcji w komórce. Takimi celami miałyby być szczególnie geny o nieznanych funkcjach, które są konserwatywne i występują w wielu genomach, a zatem prawdopodobnie są istotne dla funkcjonowania komórki. Metody komputerowe są tańsze i szybsze niż analizy eksperymentalne, dlatego długo będą jednym z głównych źródeł informacji o genomach, które następnie powinny być weryfikowane w badaniach doświadczalnych.

Dostęp do wielu sekwencji genomów i różnorodnych informacji spowodował rozwój nowych, bardziej wyspecjalizowanych działów genomiki. Poza dziedzinami, których przedmiotem są poszczególne poziomy analiz (patrz wyżej), jak genomika (*sensu stricte*), transkryptomika, proteomika, interaktomika, metabolomika itp., można wyróżnić jeszcze takie dziedziny jak:

– **Genomika porównawcza** – porównuje genomy i poszczególne sekwencje za pomocą algorytmów FASTA lub BLAST w celu znalezienia charakterystycznych regionów w sekwencjach – motywów lub domen oraz rozpoznania sekwencji homologicznych (wywodzących się od wspólnego przodka). Ułatwia dokonanie właściwych adnotacji analizowanych sekwencji przez przeniesienie przypisanej funkcji lub innych informacji z jednej sekwencji na inną – homologiczną, na bazie ich podobieństwa. Porównywanie sekwencji z wielu genomów ułatwia ponadto wyznaczenie właściwych granic genu i określenie jego struktury, szczególnie dotyczy to genów podzielonych na eksony i introny oraz umożliwia zidentyfikowanie nieznanych regionów regulatorowych, motywów i domen w sekwencjach. Analizy porównawcze ułatwiają klasyfikowanie białek i ich struktur w różne grupy – rodziny, nadrodziny, ortologi, paralogi itp. W skali genomowej badania porównawcze identyfikują regiony poddane rearanżacjom, duplikacjom i delecjom.

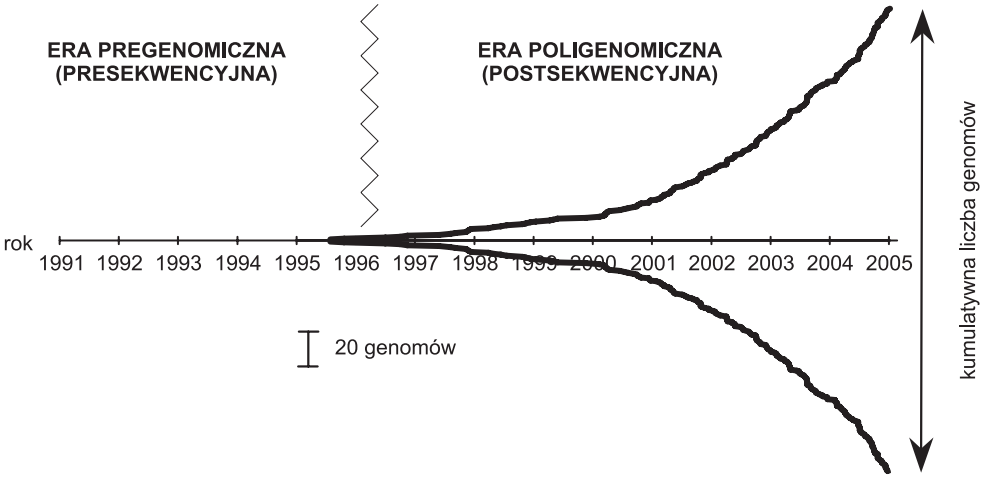
– **Genomika ewolucyjna** – opisuje wszystkie wymienione aspekty w kontekście procesów ewolucyjnych, np. duplikacje i utraty genów, fuzje genów, zmiany ułożenia genów, transfer boczny oraz analizuje drzewa filogenetyczne organizmów uwzględniając informacje pochodzące z całych genomów (**filogenomika**).

– **Genomika strukturalna** – stara się określić struktury przestrzenne wszystkich białek.

– **Farmakogenomika –** zajmuje się inżynierią białek i projektowaniem nowych leków na podstawie informacji płynących z analiz genomowych.

Wyniki badań genomicznych są coraz częściej wykorzystywane w poszukiwaniu czynników wirulencji, nowych szczepionek, związków bakteriobójczych, genów i ich produktów będących celem działania nowych leków. Pomagają w zrozumieniu mechanizmów wirulencji i patogenezy, a wobec tego przyczyniają się do lepszego diagnozowania i leczenia wielu chorób infekcyjnych. Poza tym lepsza znajomość genomów wielu mikroorganizmów umożliwia skuteczniejsze ich wykorzystanie w biotechnologii, przemyśle, rolnictwie i ochronie środowiska.

Jednym z pierwszych przykładów zastosowania genomiki do poszukiwania nowych szczepionek są badania przeprowadzone na genomie bakterii *Neisseria meningitidis* serotyp B szczep MC58 (24). Na podstawie analiz komputerowych kompletnie zsekwencjonowanego genomu tej bakterii zidentyfikowano 570 przypuszczalnych białek sekrecyjnych lub powierzchniowych. Kodujące je geny sklonowano w komórkach *Escherichia coli*, w których ekspresji uległo 61% analizowanych genów. Oczyszczone rekombinowane białka z *E. coli* użyto do immunizacji myszy. Z surowicy immunizowanych myszy wybrano następnie siedem przeciwciał, które wykazywały aktywność bakteriobójczą i zdolność do wiązania się z powierzchnią komórek meningokoków. Na podstawie dalszych analiz wybrano dwa białka, które charakteryzowały się dużą konserwatywnością w obrębie wielu izolatów i serotypów *N. meningitidis*. Dzięki takim analizom można w ciągu kilku miesięcy, uwzględniając w tym już sekwencjonowanie i analizę genomu, wyselekcjonować właściwe antygeny do produkcji skutecznej szczepionki.



Rys. 7. Podział historii biologii molekularnej na erę pregenomiczną i poligenomiczną. Podział ten jest wynikiem wzrostu liczby zsekwencjonowanych genomów oraz danych wynikających z ich analiz biologicznych, które są coraz częściej przeprowadzane w kontekście całego genomu.

Dostępność sekwencji wielu genomów oraz danych wynikających z ich analiz spowodowały, że badania biologiczne coraz częściej uwzględniają kontekst całego genomu (rys. 7). Dlatego często stosowany jest termin era pregenomiczna, lub presekwencyjna na określenie okresu, kiedy sekwencje genomów nie były jeszcze dostępne, a analizy przeprowadzano na bazie sekwencji pojedynczych genów. Gdy liczba kompletnych sekwencji genomów wystarczająco wzrosła weszliśmy w erę poligenomiczną. Stosowany jest również termin era postgenomiczna, ale jest on raczej niewłaściwy, gdyż przedrostek *post-* sugeruje, że mamy do czynienia z badaniami nie dotyczącymi już genomu. Bardziej zasadny jest natomiast termin era postsekwencyjna oznaczający czas po etapie sekwencjonowania genomu, czyli czas jego analizowania. W początkowych etapach rozwoju genomiki i bioinformatyki, tworzono głównie pierwotne bazy danych gromadzące szybko przyrastające dane sekwencyjne, a badania były skierowane na identyfikowanie i rozumienie funkcji poszczególnych genów i białek (25). Później zaczęto stosować na dużą skalę genomiczne i proteomiczne badania eksperymentalne. Obecnie intensywnie rozwijają się również bazy wtórne – pochodne, przetwarzające zgromadzone dane w celu uzyskania nowych informacji, a dominują badania mające na celu zrozumienie funkcji na poziomie molekularnym, komórkowym oraz na poziomie organizmu. W przyszłości będą powstawać komputerowe reprezentacje całych komórek i organizmów opisujące ich funkcjonowanie, co pozwoli lepiej zrozumieć podstawowe zasady rządzące złożonymi zjawiskami i układami biologicznymi. Bioinformatyka i genomika staną się bardziej fundamentalnymi dziedzinami łączącymi w sobie poza naukami biologicznymi i informatycznymi, jak jest obecnie, również matematykę, fizykę, chemię i medycynę.

Literatura

1. Fiers W., Contreras R., Duerinck F., Haegeman G., Iserentant D., Merregaert J., Min Jou W., Molemans F., Raeymaekers A., van den Berghe A., et al., (1976), Nature, 260, 500-507.
2. Sanger F., Nicklen S., Coulson A. R., (1977), Proc. Natl. Acad. Sci. USA, 74, 5463-5467.
3. Maxam A. M., Gilbert W., (1977), Proc. Natl. Acad. Sci. USA, 74, 560-564.
4. Sanger F., Air G. M., Barrell B. G., Brown N. L., Coulson A. R., Fiddes C. A., Hutchison III C. A., Slocombe P. M., Smith M., (1977), Nature, 265, 687-695.
5. Anderson S., Bankier A. T., Barrell B. G., de Bruijn M. H., Coulson A. R., Drouin J., Eperon I. C., Nierlich D. P., Roe B. A., Sanger F., et.al., (1981), Nature, 290, 457-465.
6. Sanger F., Coulson A. R., Hong G. F., Hill D. F., Petersen G. B., (1982), J. Mol. Biol., 162, 729-773.
7. Oliver S. G., van der Aart Q. J. M., Agostoni-Carbone M. L., Aigle M., Alberghina L., Alexandraki D., Antoine G., Anwar R., Ballesta J. P. G., Benit P., et al., (1992), Nature, 357, 38-46.
8. Goffeau A., Barrel B. G., Bussey H., Davis R. W., Dujon B., Feldmann H., Galibert F., Hoheisel J. D., Jacq C., Johnston M., et al., (1996), Science, 274, 546-567.
9. Venter J. C., Smith H. O., Hood L., (1996), Nature, 381, 364-366.
10. Sutton G. G., White O., Adams M. D., Kerlavage A. R., (1995), Genome Sci. Technol., 1, 9-19.
11. Fleischmann R. D., Adams M. D., White O., Clayton R. A., Kirkness E. F., Kerlavage A. R., Bult C. J., Tomb J., Dougherty B. A., Merrick J. M., et al., (1995), Science, 269, 496-512.
12. 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), Science, 270, 397-403.
13. Kyrpides N., (1999), Bioinformatics, 15, 773-774.
14. Bernal A., Ear U., Kyrpides N., (2001), Nucleic Acids Res., 29, 126-127.
15. Overbeek R., (2000), Genome Biol., 1, COMMENT2002.
16. Staley J. T., Konopka A., (1985), Ann. Rev. Microbiol., 39, 321-346.
17. Hugenholtz P., (2002), Genome Biol., 3, reviews 0003.1–0003.8.
18. Drancourt M., Bollet C., Carlioz A., Martelin R., Gayral J. P., Raoult D., (2000), J. Clin. Microbiol., 38, 3623-3630.
19. Janssen P., Audit B., Cases I., Darzentas N., Goldovsky L., Kunin V., Lopez-Bigas N., Peregrin-Alvarez J. M., Pereira-Leal J. B., Tsoka S., Ouzounis C. A., (2003), Genome Biol., 4, 402.
20. Galvez A., Maqueda M., Martinez-Bueno M., Valdivia E., (1998), ASM News, 64, 269-275.
21. Venter J. C., Remington K., Heidelberg J. F., Halpern A. L., Rusch D., Eisen J. A., Wu D., Paulsen I., Nelson K. E., Nelson W. et al., (2004), Science, 304, 66-74.
22. Koonin E. V., (2001), Curr. Biol. 11, R155-158.
23. Koonin E. V., Galperin M. Y., (2003), *Sequence – Evolution - Function. Computational Approaches in Comparative Genomics,* Kluwer Academic Publishers, Boston, Dordrecht, London.
24. Pizza M., Scarlato V., Masignani V., Giuliani M. M., Arico B., Comanducci M., Jennings G. T., Baldi L., Bartolini E., Capecchi B., et al., (2000), Science, 287, 1816-1820.
25. Kanehisa M., Bork P., (2003), Nature Genet. Suppl., 33, 305-310.