PRACE PRZEGLĄDOWE

Kurs szybkiego czytania DNA – nowoczesne techniki sekwencjonowania

Magdalena Kotowska[1], Jolanta Zakrzewska-Czerwińska[1,2]

[1]Instytut Immunologii i Terapii Doświadczalnej im. Ludwika Hirszfelda, Polska Akademia Nauk, Wrocław
[2]Wydział Biotechnologii, Uniwersytet Wrocławski, Wrocław

DNA speed reading course – high-throughput DNA sequencing technologies

Summary

Development of high-throughput DNA sequencing technologies that omit time consuming and labour intensive cloning steps have opened unprecedented possibilities in life sciences. Massive scale generation of raw sequences requires constant improvement of computational methods of data analysis. New disciplines of genomics, metagenomics and transcriptomics have emerged which revolutionize experimental approach to different fields of biology. Both basic studies, such as species evolution or microbial ecology, and applied sciences of biotechnology and medicine benefit greatly from the new tools available. In this article next-generation DNA sequencing technologies are reviewed. Information on data analysis and applications is also provided.

Key words:
next-generation DNA sequencing, pirosequencing, single molecule sequencing, genomics, metagenomics.

Adres do korespondencji

Magdalena Kotowska,
Instytut Immunologii
i Terapii Doświadczalnej
im. Ludwika Hirszfelda,
Polska Akademia Nauk,
ul. Rudolfa Weigla 12,
53-114 Wrocław;
e-mail:
szulc@iitd.pan.wroc.pl

biotechnologia
4 (91) 24–38 2010

1. Wstęp

W roku 1977 Sanger i wsp. (1) oraz Maxam i Gilbert (2) wprowadzili techniki sekwencjonowania DNA. Od tego momentu datuje się intensywny rozwój tych technik. Najczęściej stosowaną stała się enzymatyczna **metoda Sangera** (z użyciem dideoksynukleotydów) często nazywana metodą terminacji łańcucha. Po raz

pierwszy metodę tę zastosowano do poznania sekwencji DNA faga ΦX174 o długości 5,4 tys. nukleotydów (3). W późniejszych latach metoda Sangera ulegała licznym modyfikacjom poprzez np. zastosowanie rekombinowanych polimeraz DNA czy też znaczników fluoroforowych (zamiast izotopowych) (4-6). W 10 lat po opublikowaniu metody Sangera dwie firmy, Applied Biosystems i Amersham–Pharmacia (obecnie General Electric Healthcare) wprowadziły jako pierwsze techniki automatycznego sekwencjonowania (4-6). Początkowo produkty sekwencjonowania rozdzielano w żelach poliakrylamidowych, a w późniejszym okresie automatyczne sekwenatory zostały wyposażone w cienkie kapilary służące do rozdziału DNA. W połowie lat 90. ubiegłego wieku japońska firma Hitachi wprowadziła wysoko wydajne kapilarne sekwenatory DNA. Dalszy rozwój automatyzacji i miniaturyzacji sekwenatorów umożliwił równoczesne sekwencjonowanie kilkuset fragmentów DNA. Równolegle z rozwojem tych technik opracowywano nowe strategie tworzenia bibliotek genomów do ich sekwencjonowania. Krokiem milowym stało się wprowadzenie nowej techniki sekwencjonowania dużych genomów, tzw. metody *shotgun* („strzału na ślepo"), polegającej na sekwencjonowaniu dużej liczby losowo pofragmentowanych odcinków DNA, które następnie są składane komputerowo (7). Takie podejście wymagało z kolei wprowadzenia nowatorskich komputerowych metod obliczeniowych składających setki tysięcy losowo uzyskanych sekwencji DNA w dłuższe fragmenty. Strategia ta obniżyła znacznie koszty i skróciła czas sekwencjonowania, eliminując tradycyjne metody polegające na żmudnym i czasochłonnym mapowaniu oraz składaniu kolejno ułożonych kosmidów lub subklonów (8). W 1996 r. dzięki metodzie *shotgun* opublikowano po raz pierwszy sekwencję całego genomu bakteryjnego: *Haemophilus influenzae* – $1{,}8\times10^6$ pz (9). Metodę tę wykorzystano do ustalenia sekwencji genomu człowieka – $2{,}91\times10^9$ pz, którą opublikowano w 2001 r. (10).

Od końca XX w. notujemy gwałtowny przyrost liczby kompletnie zsekwencjonowanych genomów i równolegle intensywny rozwój stosunkowo młodej dziedziny z pogranicza biologii i informatyki – **genomiki**, analizy całych genomów. Sekwencje DNA gromadzone w niemal wykładniczym postępie są analizowane za pomocą coraz bardziej wyrafinowanych i zaawansowanych metod obliczeniowych, przy wykorzystaniu komputerów charakteryzujących się coraz szybszymi procesorami i pojemniejszymi dyskami twardymi.

2. Nowe generacje sekwencjonowania

Odkrycia dokonane w ostatnich kilku latach pozwoliły na wprowadzenie kolejnych, po metodzie Sangera, generacji sekwencjonowania DNA (11,12). Ich pojawienie się nie byłoby możliwe bez współdziałania naukowców reprezentujących różne dziedziny wiedzy: biologię, chemię, fizykę, matematykę i informatykę. Dzięki miniaturyzacji i automatyzacji stworzono wysoko przepustowe sekwenatory umożliwiające jednoczesne sekwencjonowanie nawet miliona fragmentów DNA – **sekwencjo-**

nowanie drugiej generacji. Najnowsze osiągnięcia nanotechnologii umożliwiają bezpośredni odczyt sekwencji z pojedynczej cząsteczki DNA (SMS, ang. *Single Molecule Sequencing*) bez konieczności jej amplifikacji – **sekwencjonowanie trzeciej generacji**. Przykładowe zestawienie technik sekwencjonowania obu generacji przedstawiono w tabeli.

Tabela

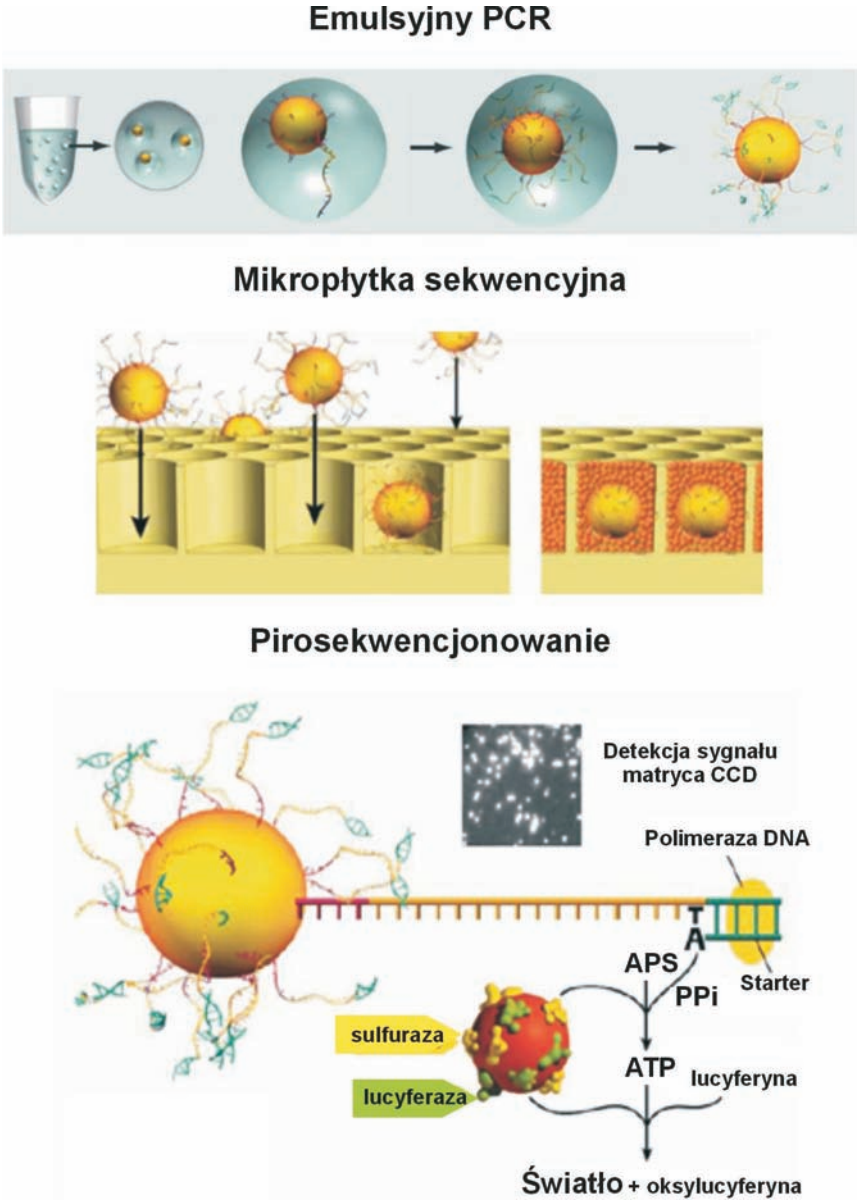
Przykładowe zestawienie technik sekwencjonowania

Metoda sekwencjonowania (firma), strona www		Maksymalna długość odczytu	Liczba pz na pojedynczy eksperyment
II generacji	'454' (454 Life Sciences) **www.454.com**	300-500 pz	0,05-1,0 Gpz
	Solexa (Illumina) **www.illumina.com**	35-50 pz	2,0 Gpz
	SOLiD (Applied Biosystems) **www.appliedbiosystems.com**	~35 pz	> 2,0 Gpz
III generacji	tSMS (Helicos) **www.helicosbio.com**	100-200 pz	1 Gpz
	Visigen (VisiGen Biotechnologies, Inc.) **www.visigenbio.com**	nieograniczona długość	bd
	SMRT (Pacific Biosciences) **www.pacificbiosciences.com**	100 000 pz	bd

bd – brak danych

2.1. Jak to się robi?

Pirosekwencjonowanie jest pierwszą metodą nowej generacji sekwencjonowania DNA, którą wdrożono przed kilkoma laty (rys. 1). W metodzie tej rejestruje się syntezę DNA w czasie rzeczywistym poprzez pomiar ilości pirofosforanu (PP_i) uwalnianego w momencie włączenia komplementarnej zasady do nowo syntezowanej nici DNA. Pomiaru uwalnianego PP_i dokonuje się za pomocą dwóch reakcji enzymatycznych (z udziałem sulfurazy i lucyferazy), w wyniku których powstaje strumień fotonów rejestrowany przez kamerę ze światłoczułą matrycą CCD (ang. *Charge Coupled Device*). Fragmenty DNA przeznaczone do sekwencjonowania poddaje się ligacji z adapterami (krótkimi fragmentami o znanej sekwencji) i amplifikuje się na podłożu stałym – kulkach opłaszczonych streptawidyną, które wiążą produkt amplifikacji (jeden ze starterów znakowany jest biotyną). Amplifikacji dokonuje się w emulsji wodno-olejowej: naczynie reakcyjne stanowi kropelka wody. Emulsję przygotowuje się w taki sposób by w jednej kropelce wody znajdowała się pojedyncza kulka, na której amplifikowany jest dany fragment DNA. Metoda ta jest często nazywana bąbelkowym lub emulsyjnym PCR-em, a dzięki niej można jednocześnie



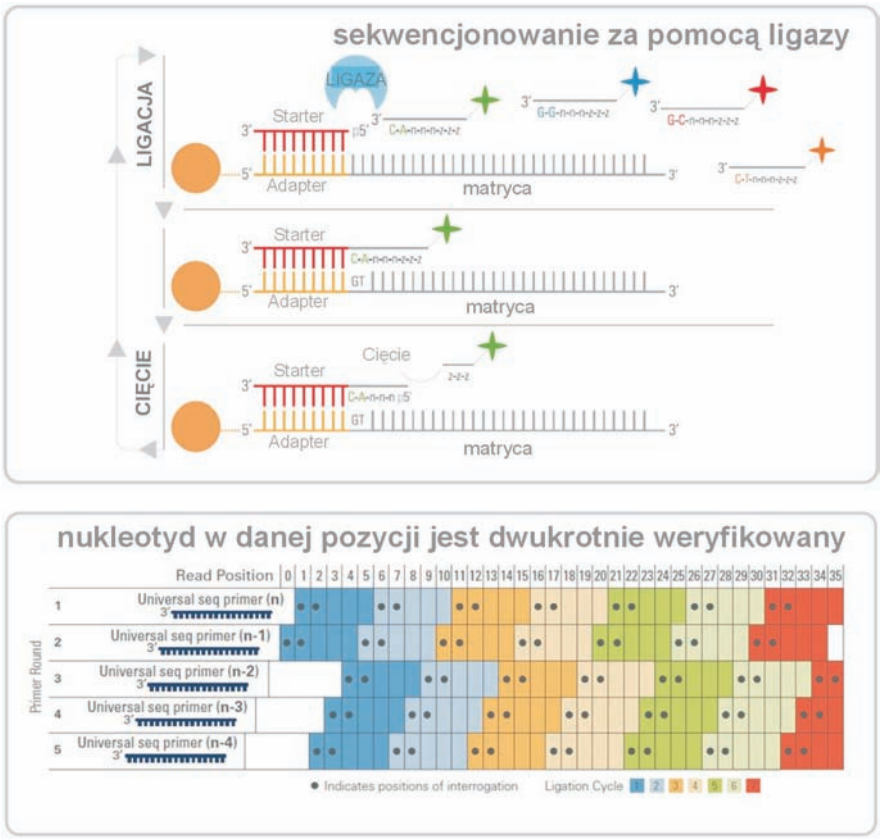
Rys. 1. Pirosekwencjonowanie (metoda '454'). Opracowano według 454 Life Sciences (www.454.com). Opis w tekście.

przeprowadzać miliony reakcji amplifikacji. Powielone na kulkach matryce DNA umieszcza się następnie w studzienkach (o średnicy 44 μm) mikropłytki (jeden zamplifikowany fragment DNA – jedna kulka – jedna studzienka), w których dokonuje się pirosekwencjonowania w czasie rzeczywistym. Każda studzienka znajduje się na

końcu pojedynczego włókna światłowodowego połączonego z kamerą CCD. Przykładowo jedna mikropłytka może zawierać ponad 1 500 000 studzienek, a zatem tyle reakcji sekwencjonowania można przeprowadzać jednocześnie. Wysoko przepustowe sekwenatory (**The 454 Genome Sequencer FLX**) wykorzystujące tę technologię są już komercyjnie dostępne; sprzedaje je firma **454 Life Sciences** należąca do szwajcarskiego koncernu **Roche AG**. Przy użyciu tej technologii zsekwencjonowano w rekordowym tempie 2 miesięcy cały genom Jamesa D. Watsona, a koszt tego przedsięwzięcia wyniósł około 1 miliona USD. Dla porównania, zakończony 7 lat temu projekt sekwencjonowania ludzkiego genomu (ang. *Human Genome Project*) wykorzystujący metodę Sangera kosztował 3 miliardy dolarów i trwał 13 lat przy zaangażowaniu międzynarodowego konsorcjum złożonego z 20 instytucji **(http://www.genome.gov).**

Kolejną metodę sekwencjonowania drugiej generacji, **Solexa** (**Genome Analyzer**), wprowadziła firma Illumina (San Diego, USA). W metodzie tej do pofragmentowanego DNA dołącza się krótkie dwuniciowe adaptery. Po denaturacji mieszaninę jednoniciowych fragmentów DNA oraz oligonukletydów (przy znaczącym nadmiarze tych ostatnich w stosunku do fragmentów DNA) unieruchamia się na powierzchni stałej (mikropłytce umieszczonej w komorze przepływowej). W następnym etapie dokonuje się amplifikacji fragmentów DNA za pomocą reakcji PCR typu „koci grzbiet" bez dodatku wolnych oligonukleotydów. Polega to na tym, że unieruchomiony z jednej strony jednoniciowy fragment DNA odszukuje w najbliższym sąsiedztwie komplementarny oligonukleotyd, również związany z podłożem, tworząc tzw. „koci grzbiet", a polimeraza w obecności nukleotydów dobudowuje komplementarną nić. Po denaturacji cykl się powtarza, aż do momentu powstania w okolicy danego fragmentu odpowiedniej do sekwencjonowania liczby kopii tego fragmentu (~1000 kopii). W ten sposób otrzymuje się na mikropłytce sektory, z których każdy reprezentuje statystycznie inny powielony fragment DNA – ang. „*polonies*" – *polymerase generated colonies*. Kolejnym etapem jest jednoczesne sekwencjonowanie DNA (poprzez syntezę) we wszystkich sektorach. DNA jest sekwencjonowany przy użyciu specjalnie zaprojektowanych nukleotydów zaopatrzonych w usuwalne znaczniki fluorescencyjne, które każdorazowo kończą syntezę DNA w danym cyklu – każdy z 4 nukleotydów znaczony jest innym fluoroforem. Matryca CCD rejestruje sygnały w poszczególnych sektorach pochodzące od nowo przyłączonych w danym cyklu komplementarnych nukleotydów, po czym fluorofory ze wszystkich sektorów zostają usunięte, tak by kolejne znakowane nukleotydy mogły zostać przyłączone w następnym cyklu sekwencjonowania.

W 2007 r. firma **Applied Biosystems** (Foster City, USA) wprowadziła nową metodę sekwencjonowania drugiej generacji **SOLiD** opartą na ligacji (rys. 2). W metodzie tej, podobnie do sekwencjonowania „454", badany materiał amplifikuje się za pomocą emulsyjnego PCR-u. Po reakcji PCR 3' koniec jednej z nici DNA jest modyfikowany, co umożliwia jej kowalencyjne przyłączenie do stałego podłoża (szklanej mikropłytki). W ten sposób, podobnie jak w metodzie Solexa, otrzymuje się mikro-



Rys. 2. Sekwencjonowie metodą SOLiD. Opracowano według Applied Biosystems (**http://www3.appliedbiosystems.com**). Opis w tekście.

płytkę z sektorami, z których każdy reprezentuje statystycznie inny powielony fragment DNA. W następnym etapie, w wyniku hybrydyzacji dołączany jest krótki oligonukleotyd komplementarny do jednego ze starterów (adapterów) użytego do emulsyjnego PCR-u. W kolejnym etapie dołączane są za pomocą ligacji krótkie znakowane fluorescencyjnie oligonukleotdy, w których znane są pierwsze dwa nukleotydy; przykładowo dinukleotyd CA odszukuje komplementarną sekwencję GT na matrycy jednoniciowego DNA. Następnie fluorescencyjny znacznik jest usuwany, a w kolejnych cyklach znajdowane są następne dinukleotydy tak, że w sumie otrzymujemy podwójne pokrycie każdej pozycji w analizowanej sekwencji.

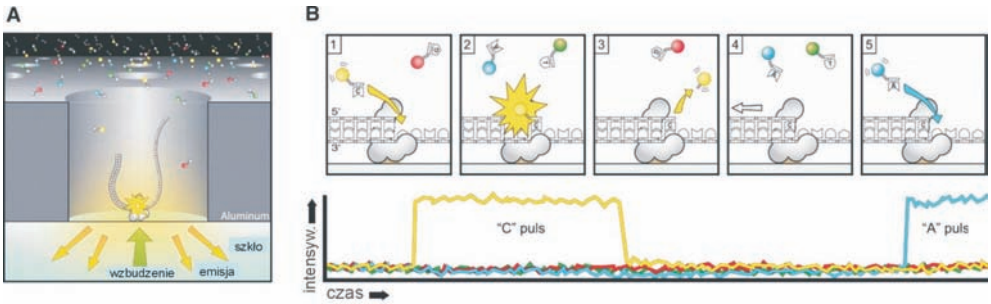
Pod koniec 2008 r. firma **Helicos** (Cambridge, USA, www.helicosbio.com) jako pierwsza wprowadziła sekwenator trzeciej generacji **tSMS** (ang. *true Single Molecule Sequencing*) umożliwiający bezpośredni odczyt sekwencji z matrycy bez konieczności jej amplifikacji. Zasada działania tSMS jest zbliżona do metody Solexa opracowanej przez firmę Illumina: po przyłączeniu komplementarnego do matrycy kolejnego

znakowanego fluorescencyjnie nukleotydu, następuje usunięcie fluoroforu i dobudowanie następnego znaczonego nukleotydu. W sekwenatorze tSMS zastosowano bardzo czuły detektor wykrywający pojedynczą cząsteczkę fluoroforu.

Ciekawą metodę sekwencjonowania trzeciej generacji, w której wykorzystano zjawisko FRET (ang. *Förster Resonance Energy Transfer*) zaproponowała firma **VisiGen Biotechnologies**, **Inc**. (Houston, USA). Jest to metoda syntezy DNA w czasie rzeczywistym naśladująca proces replikacji DNA. W metodzie tej wykorzystuje się polimerazę DNA z kowalencyjnie dołączonym fluoroforem – donorem oraz cztery modyfikowane nukleotydy, z których każdy ma przyłączony do reszty fosforanowej inny fluorofor – akceptor. Cząsteczka DNA przesuwająca się przez centrum aktywne modyfikowanej polimerazy DNA jest odczytywana bezpośrednio: w momencie przyłączania kolejnego komplementarnego do matrycy nukleotydu dochodzi do zbliżenia donora z akceptorem i przekazania energii, co jest monitorowane przez czuły system detekcyjny. Po usunięciu pirofosforanu z przyłączonym akceptorowym fluoroforem dochodzi do zaniku zjawiska FRET, a następnie zostaje przyłączony kolejny nukleotyd. Szybkość takiego odczytu może dochodzić do 300 zasad na sekundę dla pojedynczego nanometrowego czujnika.

Nukleotydy z fluoroforami przyłączonymi do fosforanu wykorzystuje również metoda sekwencjonowania **SMRT** (ang. *Single-Molecule Real-Time*) wprowadzona przez firmę **Pacific Biosciences** (PacBio, MenloPark, USA). Pojedyncze cząsteczki polimerazy są immobilizowane na dnie studzienek reakcyjnych **ZMWs** (ang. *Zero-Mode Waveguides*) – otworków o średnicy kilkudziesięciu nm wykonanych w warstwie metalu o grubości 100 nm na szklanej powierzchni (rys. 3). Oświetlając płytkę od spodu promieniem lasera można obserwować zawartość poszczególnych studzienek. Ponieważ średnica otworków jest mniejsza niż długość fali, rzeczywista objętość, do której dociera światło wynosi 20 zeptolitrów (20×10^{-21} litra). Rejestrowany jest sygnał pochodzący od każdego wbudowywanego nukleotydu. Po odłączeniu pirofosforanu z fluoroforem sygnał zanika. Pomimo że wszystkie nukleotydy są obecne w mieszaninie reakcyjnej w wysokim stężeniu, poziom tła jest bardzo niski, ponieważ czas reakcji jest o kilka rzędów dłuższy niż czas przypadkowego pojawiania się wolnych nukleotydów w polu obserwacji (13).

Ostatnio brytyjska firma **Oxford Nanopore Technologies** opracowała nowy typ sekwencjonowania trzeciej generacji – nanoporowe sekwencjonowanie. W przeciwieństwie do poprzednich metod sekwencjonowanie odbywa się nie poprzez syntezę, ale poprzez kontrolowane odcinanie nukleotydu po nukleotydzie z analizowanej nici DNA. Odłączane nukleotydy sukcesywnie przechodzą przez nanopory, w których mierzone jest natężenie prądu. Każdy z czterech nukleotydów powoduje inną zmianę natężenia prądu (rzędu pikoamperów) co jest rejestrowane przez czuły amperomierz.



Rys. 3. Sekwencjonowanie pojedynczej cząsteczki DNA w czasie rzeczywistym, SMRT (ang. *Single-Molecule Real-Time*). Opracowano według (13). **A**. Studzienka ZMW (ang. *Zero-Mode Waveguide*), w której dokonuje się detekcji pojedynczego komplementarnego nukleotydu przyłączanego w trakcie syntezy. **B**. Zasada działania techniki SMRT. Polimeraza dobudowuje pojedynczy komplementarny znakowany fluoroforem nukleotyd (**1**), czego konsekwencją jest rejestracja sygnału (**2**). Po odłączeniu $\mathbf{PP_i}$ z fluoroforem sygnał zanika (**3**), matryca się przesuwa (**4**) i dobudowywany jest następny nukleotyd (**5**).

2.2. Do czego jeszcze to się może przydać?

Wysoko wydajne techniki sekwencjonowania nowej generacji szybko znalazły szersze zastosowania aniżeli tylko sekwencjonowanie genomów. Możliwość jednoczesnego badania milionów krótkich sekwencji bez konieczności ich klonowania pozwoliła na analizę złożonych mieszanin kwasów nukleinowych. Techniki te przyczyniły się do rozwoju nowej dziedziny – **metagenomiki**, która zajmuje się globalną analizą materiału genetycznego (DNA/RNA) pozyskanego bezpośrednio ze środowiska (np. gleba, woda) bez konieczności hodowli organizmów zasiedlających analizowaną biocenozę.

Genomika otworzyła drogę do **transkryptomiki** – całościowego badania powstającego w komórkach RNA. Dominująca w dotychczasowych badaniach, oparta na hybrydyzacji, technika *microarray* wymaga znajomości sekwencji genomu badanego organizmu do właściwego zaprojektowania mikromatryc oligonukleotydowych (tzw. chipów DNA). Nowe techniki określane terminem RNA-seq pozwalają na sekwencjonowanie wszystkich powstających transkryptów, umożliwiają precyzyjne porównanie ilościowe i są wolne od ograniczeń związanych z niespecyficzną hybrydyzacją (14).

Również inne eksperymenty wymagające identyfikacji dużej liczby krótkich fragmentów DNA mogą być obecnie przeprowadzane w skali całego genomu dzięki zastosowaniu wysoko wydajnego sekwencjonowania.

Immunoprecypitacja chromatyny (ChIP, ang. *Chromatin Immunoprecipitation*) służy do wyszukiwania fragmentów DNA wiązanych przez badane czynniki transkrypcyjne i inne białka. Bezpośrednie sekwencjonowanie (ChIP-seq) ma znaczną przewagę nad stosowaniem mikromatryc do identyfikacji związanych fragmentów DNA (ekspery-

menty *ChIP-on-chip*), jest bowiem dokładniejsze i mniej pracochłonne (15). Inne przykłady technik, które można stosować na nie spotykaną dotąd skalę to identyfikacja regionów otwartej chromatyny wrażliwej na trawienie DNazą I (DNase-seq) (16), badanie zmienności liczby kopii fragmentów DNA (CNV-seq) (17) czy też badanie polimorfizmu pojedynczych nukleotydów (SNP) (18,19).

3. Analiza wyników sekwencjonowania

Nowoczesne urządzenia wykorzystujące techniki sekwencjonowania nowej generacji dostarczają olbrzymią liczbę danych, przekraczającą o kilka rzędów wielkości wyniki uzyskiwane starszymi metodami. Odczytywane sekwencje są jednak bezwartościowe bez odpowiednich narzędzi informatycznych do ich analizy. Olbrzymia liczba generowanych w coraz szybszym tempie danych wymaga zaangażowania potężnych komputerów i udoskonalonych programów. Obecnie etapem ograniczającym uzyskiwanie znaczących informacji staje się nie samo sekwencjonowanie genomów, lecz ich analiza (20). Aby w pełni wykorzystać potencjał, jaki niesie upowszechnienie nowych technologii badacze muszą poznać podstawowe zasady działania narzędzi analitycznych i zdawać sobie sprawę z ich ograniczeń.

3.1. Długa sekwencja z krótkich odczytów

Pierwszym etapem opracowania danych jest złożenie pojedynczych odczytów (ang. *reads*) w ciąg zasad odpowiadających długim odcinkom DNA (ang. *contigs*), najlepiej całym badanym chromosomom. Drugi etap to rozszyfrowanie „treści" zakodowanej w odczytanej sekwencji i wyciąganie wniosków wyjaśniających zjawiska biologiczne.

„Składanie" krótkich sekwencji w całość odbywa się za pomocą dwóch zasadniczych typów algorytmów: *alignment* i *assembly* (21). Pierwsze podejście (*alignment*) ma zastosowanie, wtedy gdy dysponujemy sekwencją genomu innego organizmu tego samego gatunku, lub też gatunku spokrewnionego. Pojedyncze odczyty są wówczas przypisywane do odpowiednich regionów genomu referencyjnego. Składanie sekwencji *de novo* (*assembly*) tradycyjnie opierało się na identyfikacji kolejnych częściowo nakładających się fragmentów. Metody te sprawdzały się dobrze przy odczytach o długości ok. 800 pz uzyskiwanych metodą Sangera, lecz okazały się niepraktyczne w odniesieniu do dużo większej liczby znacznie krótszych sekwencji 50-400 pz. Nowe algorytmy wykorzystują odmienne podejście – konstrukcję wykresu de Bruijna (22). Do ich zalet należy lepsza identyfikacja regionów powtórzonych, a także liniowa zależność czasu trwania obliczeń od liczby analizowanych sekwencji, w przeciwieństwie do starszych metod, gdzie czas trwania analizy wzrasta proporcjonalnie do kwadratu liczby sekwencji.

Konieczność porównywania miliardów krótkich sekwencji wymusza optymalizację programów pod względem szybkości działania i wymagań sprzętowych. Istotnym elementem, który musi zostać uwzględniony przy konstrukcji algorytmu jest także fakt, że wyniki pochodzące z poszczególnych typów sekwenatorów różnią się strukturą i rodzajami generowanych błędów (21). Wysoko wydajne technologie nowej generacji przestały być domeną nielicznych wielkich centrów sekwencjonowania. W wielu laboratoriach na całym świecie powstają nowe programy, których aktualną listę można znaleźć na stronach internetowych (**http://seqanswers.com/wiki/Software**, **http://en.wikipedia.org/wiki/List_of_sequence_alignment_software**, **http://en.wikipedia.org/wiki/Sequence_assembly**).

Analiza wyników sekwencjonowania DNA metagenomowego musi sprostać trudniejszemu wyzwaniu analizy pomieszanych sekwencji pochodzących z różnych organizmów, które są reprezentowane w nierównym stopniu. W tym przypadku bierze się pod uwagę kilka kryteriów, takich jak zawartość par GC, charakterystyczny rozkład dinukleotydów oraz podobieństwo do poznanych genomów (23,24). Możliwa jest nawet rekonstrukcja z próbek metagenomowych całych, lub prawie całych, genomów mikroorganizmów dominujących w danej populacji. Najbardziej znany program do analizy sekwencji metagenomów to MEGAN (25). Analizę porównawczą ułatwiają ogólnodostępne serwery MG-RAST (26) i SEED (27).

3.2. Rozszyfrowanie treści

Pierwszym krokiem do poznania informacji zakodowanej w DNA jest odszukanie genów kodujących białka. Choć procedury znajdowania sekwencji kodujących są powszechnie stosowane od wielu lat, właściwe opisanie (ang. *annotation*) nowo sekwencjonowanych genomów, szczególnie eukariotycznych, nie jest sprawą trywialną (28-30). Domniemane funkcje zidentyfikowanych białek wnioskowane są na podstawie porównania z dostępnymi bazami danych, które niestety nie są wolne od błędów (31). Podejmowane są międzynarodowe wysiłki zmierzające do udoskonalenia procedur opisywania genomów, np. projekt SEED (27). Aktualną listę dostępnych programów wraz z omówieniem zasad ich działania można znaleźć w pracy przeglądowej (32).

W ciągu ostatnich kilkunastu lat zaczęto odkrywać fascynujący świat niekodującego RNA (ncRNA, ang. *non-coding RNA*) (33,34). Do dobrze nam znanych z podręczników mRNA, tRNA i rRNA doszły nowe skróty: siRNA (ang. *small interfering RNA*), miRNA (ang. *microRNA*), lincRNA (ang. *large intervening non-coding* RNA). Wykazano, że większość wykrywanych transkryptów eukariotycznych nie odpowiada genom kodującym białka. Kwestia, które z nich mają konkretną funkcję, a które stanowią tylko „tło", pozostaje otwarta (34). Również organizmy prokariotyczne wykorzystują regulatorowe cząsteczki RNA (35). Dopiero zaczynamy poznawać ich rolę biologiczną. Wyszukiwanie w sekwencjach genomowych genów kodujących funk-

cjonalne RNA jest znacznie trudniejsze niż znajdowanie genów kodujących białka (36). Głównym kryterium analizy jest przewidywanie struktur drugorzędowych tworzonych przez cząsteczki RNA. Jeśli elementy o podobnej strukturze znajdują się w genomach różnych organizmów, może to świadczyć o ich istotnej roli. Metody komputerowe do wyszukiwania ncRNA są jeszcze bardzo niedoskonałe, wyniki uzyskiwane przez różne programy są rozbieżne, niemniej jednak dają wyobrażenie o skali zjawiska i stanowią dobry punkt wyjścia do badań eksperymentalnych (36).

4. Co z tego wynika?

W chwili pisania tego artykułu (początek 2010 r.) w bazie danych GOLD (**www.genomesonline.org/gold.cgi**) zgromadzono pełne sekwencje 1198 genomów (5 razy więcej niż 5 lat temu), przy ogólnej liczbie 6638 projektów trwających i zakończonych. W tej liczbie mieści się także 207 metagenomów z tak różnych środowisk jak norweskie fiordy, amerykańskie oczyszczalnie ścieków czy przewód pokarmowy termita z Kostaryki. Jest to olbrzymi materiał zarówno do analiz porównawczych jak i szczegółowych badań poszczególnych organizmów.

Wielka liczba nowych danych umożliwia rozszerzenie analiz filogenetycznych, pozwala lepiej zrozumieć ewolucję gatunków, oraz zjawiska związane z przekazywaniem genów (37-39). Szczególnie wiele nowych informacji dotyczy organizmów prokariotycznych. Zmienia się także nasze spojrzenie na ich rolę w ekosystemach.

4.1. Kopalnia złota dla biotechnologii

Badania metagenomów potwierdzają szacunki, według których poznane dotychczas gatunki stanowią zaledwie 1% całego świata mikroorganizmów (40). Tajemnicza „większość" kryje w sobie trudne do wyobrażenia bogactwo nieznanych enzymów i metabolitów – potencjalnych leków. Uzasadnione nadzieje na nowe odkrycia inspirują rozwój metod pozwalających sięgać do tych zasobów. Do takich metod należy tworzenie bibliotek metagenomowych w organizmach łatwych do hodowli i przeszukiwanie ich zarówno na poziomie sekwencji jak i w testach funkcjonalnych (40-42).

Od czasu wyprawy Craiga Ventera (24) na Morze Sargassowe gwałtownie wzrosło zainteresowanie mikroorganizmami morskimi. Najbardziej obiecujące potencjalne leki wytwarzane przez bakterie pochodzące z tego środowiska zostały opisane w pracy przeglądowej (43). Na chromosomie morskiego promieniowca *Salinispora tropica* wytwarzającego związek przeciwnowotworowy – salinosporamid A odkryto 17 zespołów genów biosyntezy metabolitów wtórnych z różnych klas chemicznych (44).

Podobnie w wyniku poznania sekwencji genomów innych mikroorganizmów produkujących antybiotyki, np. promieniowców z rodzaju *Streptomyces*, ujawniono,

że posiadają one geny biosyntezy kilkudziesięciu metabolitów wtórnych. Wielu z tych związków nie znano wcześniej, gdyż nie są one wytwarzane w standardowych warunkach hodowli (45). Znajomość sekwencji genomu umożliwia takie manipulacje genetyczne, np. w obrębie genów regulatorowych, które pozwalają na biosyntezę pożądanych produktów. Zaplanowane zmiany w obrębie genomu służą także do zwiększania wydajności szczepów przemysłowych (46) oraz do konstrukcji gospodarzy, w których zachodzić będzie ekspresja genów przeniesionych z innych organizmów (47).

Oprócz produkcji nowych leków, zainteresowanie biotechnologów koncentruje się na poszukiwaniu alternatywnych źródeł energii. Do ciekawych przykładów można zaliczyć odkrycie 782 genów nowych analogów rodopsyny – napędzanych światłem pomp protonowych (24) oraz próby wykorzystania mikroorganizmów do produkcji wodoru (48). Inną ważną dziedziną czerpiącą z informacji genomowych i metagenomowych jest usuwanie trudno degradowalnych zanieczyszczeń ze środowiska (49).

Badania tego typu nie ograniczają się wyłącznie do bakterii i archeonów. Ważną grupę z punktu widzenia zastosowań biotechnologicznych stanowią grzyby niższe. Za przykład może posłużyć kropidlak *Aspergillus flavus*, który powoduje zanieczyszczenie żywności niebezpiecznymi aflatoksynami. Analiza genomu tego organizmu wykazała obecność genów biosyntezy wielu innych metabolitów wtórnych o potencjalnym znaczeniu terapeutycznym. Enzymy *A. flavus* degradujące biopolimery – celulazy, ksylanazy, chitynazy – mogą być wykorzystane do produkcji biopaliw (50).

4.2. Nowe możliwości w medycynie

Zasadniczym celem sekwencjonowania genomów mikroorganizmów chorobotwórczych jest poszukiwanie skutecznych metod profilaktyki, diagnostyki i terapii. Analiza genomu pozwala na sprawne testowanie wielu epitopów i uzyskanie lepszych szczepionek (51,52). Projektowane są leki zdolne do selektywnego blokowania głównych ścieżek metabolicznych. Wprowadzane są molekularne metody identyfikacji patogenów w próbkach od pacjentów.

Ustalenie sekwencji genomu ludzkiego (10) jest punktem wyjścia do identyfikacji mutacji odpowiedzialnych za powstawanie chorób oraz różną podatność poszczególnych osób na działanie leków, chemicznych zanieczyszczeń środowiska lub infekcje. Stale zwiększa się oferta dostępnych testów do badania predyspozycji genetycznych, np. do wystąpienia niektórych nowotworów. Rozpoczyna się rozwój medycyny „spersonalizowanej", w której lekarz oceniając profil genetyczny pacjenta będzie mógł przewidzieć prawdopodobną reakcję organizmu na terapię i dobrać odpowiedni rodzaj i dawkę leku. Można oczekiwać postępu także w dziedzinie terapii genowej, projektowaniu leków oraz w stosowaniu białek ludzkich do celów terapeutycznych.

Tylko niektóre choroby są wynikiem mutacji pojedynczego genu; w większości przypadków zależność między genotypem i fenotypem jest znacznie bardziej skomplikowana. Dostępna sekwencja genomu człowieka oraz postęp technologiczny w dziedzinie sekwencjonowania bardzo przyspieszyły badania tej złożonej sieci powiązań rządzących funkcjonowaniem organizmu. Wyjaśnienie molekularnych podstaw zróżnicowania w obrębie naszego gatunku jest celem projektu sekwencjonowania genomów ok. 1200 osób z całego świata (ang. *1000 Genome Project*) (www.1000genomes.org).

Kolejnym niezwykle interesującym przedsięwzięciem jest rozpoczęty w listopadzie 2007 r. projekt sekwencjonowania DNA mikroorganizmów bytujących we wnętrzu i na powierzchni organizmu ludzkiego (ang. *Human Microbiome Project*) zwany niekiedy „drugim genomem człowieka" (http://nihroadmap.nih.gov/hmp/). Jego celem jest m.in. określenie w jaki sposób zmiany w składzie mikroflory organizmu wpływają na stan zdrowia.

5. Zakończenie

Tempo przyrostu danych sekwencyjnych, jak się wydaje, będzie rosnąć, aż do granic możliwości stosowanych technologii, wynikających po prostu z ograniczeń prawami przyrody. Nie wiadomo, czy uda się stworzyć metodami inżynierii genetycznej polimerazę DNA, która będzie w stanie szybciej przyłączać kolejne nukleotydy (powyżej 1000 nukleotydów na sekundę) niż polimerazy DNA w żywej komórce w trakcie replikacji DNA. Oprócz wydajności, istotnym czynnikiem przy poszukiwaniu nowych technologii sekwencjonowania DNA jest obniżenie kosztów. Obecnie koszt sekwencjonowania genomu ludzkiego ocenia się na 50-70 tysięcy dolarów. Szacuje się, że za parę lat koszt ten może obniżyć się do około tysiąca dolarów, a zatem sumy, o wydanie której pokusić się może wiele osób pragnących poznać sekwencję własnego genomu.

Praca finansowana ze środków Fundacji na rzecz Nauki Polskiej, projekt MISTRZ(JZC).

Literatura

1. Sanger F., Nicklen S., Coulson A. R., (1977), Proc. Natl. Acad. Sci. USA, 74, 5463-5467.
2. Maxam A. M., Gilbert W., (1977), Proc. Natl. Acad. Sci. USA, 74, 560-564.
3. Sanger F., Air G. M., Barrell B. G., Brown N. L., Coulson A. R., Fiddes C. A., Hutchison C. A., Slocombe P. M., Smith M., (1977), Nature, 265, 687-695.
4. Smith L. M., Sanders J. Z., Kaiser R. J., Hughes P., Dodd C., Connell C. R., Heiner C., Kent S. B., Hood L. E., (1986), Nature, 321, 674-679.
5. Ansorge W., Sproat B. S., Stegemann J., Schwager C., (1986), J. Biochem. Biophys. Methods, 13, 315-323.
6. Ansorge W., Sproat B., Stegemann J., Schwager C., Zenke M., (1987), Nucleic Acids Res., 15, 4593-4602.

7. Venter J. C., Smith H. O., Hood L., (1996), Nature, 381, 364-366.
8. Sutton G. G., White O., Adams M. D., Kerlavage A. R., (1995), Genome Sci. Technol., 1, 9-19.
9. 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.
10. Venter J. C., Adams M. D., Myers E. W., Li P. W., Mural R. J., Sutton G. G., Smith H. O., Yandell M., Evans C. A., Holt R. A., et al., (2001), Science, 291, 1304-1351.
11. Gupta P. K., (2008), Trends Biotechnol., 26, 602-611.
12. Ansorge W. J., (2009), N. Biotechnol., 25, 195-203.
13. Eid J., Fehr A., Gray J., Luong K., Lyle J., Otto G., Peluso P., Rank D., Baybayan P., Bettman B., et al., (2009), Science, 323, 133-138.
14. van Vliet A. H., (2009), FEMS Microbiol. Lett., 302, 1-7.
15. Gilchrist D. A., Fargo D. C., Adelman K., (2009), Methods, 48, 398-408.
16. Crawford G. E., Holt I. E., Whittle J., Webb B. D., Tai D., Davis S., Margulies E. H., Chen Y., Bernat J. A., Ginsburg D., et al., (2006), Genome Res., 16, 123-131.
17. Xie C., Tammi M. T., (2009), BMC Bioinformatics., 10, 80.
18. van Tassell C. P., Smith T. P., Matukumalli L. K., Taylor J. F., Schnabel R. D., Lawley C. T., Haudenschild C. D., Moore S. S., Warren W. C., Sonstegard T. S., (2008), Nat. Methods., 5, 247-252.
19. Kerstens H. H., Crooijmans R. P., Veenendaal A., Dibbits B. W., Chin-A-Woeng T. F., den Dunnen J. T., Groenen M. A., (2009), BMC Genomics., 10, 479.
20. McPherson J. D., (2009), Nat. Methods., 6(11 Suppl), S2-5.
21. Flicek P., Birney E., (2009), Nat. Methods., 6(11 Suppl), S6-S12.
22. Idury R. M., Waterman M. S., (1995), J. Comput. Biol., 2, 291-306.
23. Tyson G. W., Chapman J., Hugenholtz P., Allen E. E., Ram R. J., Richardson P.M., Solovyev V. V., Rubin E. M., Rokhsar D. S., Banfield J. F., (2004), 428, 37-43.
24. 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.
25. Huson D. H., Auch A. F., Qi J., Schuster S. C., (2007), Genome Res., 17, 377-386.
26. Meyer F., Paarmann D., D'Souza M., Olson R., Glass E. M., Kubal M., Paczian T., Rodriguez A., Stevens R., Wilke A., et al., (2008), BMC Bioinformatics, 9, 386.
27. Overbeek R., Begley T., Butler R. M., Choudhuri J. V., Chuang H. Y., Cohoon M., de Crécy-Lagard V., Diaz N., Disz T., Edwards R., et al., (2005), Nucleic Acids Res., 33, 5691-5702.
28. Korf I., (2004), BMC Bioinformatics., 5, 59.
29. Parra G., Bradnam K., Korf I., (2007), Bioinformatics., 23, 1061-1067.
30. Bączkowski K., Mackiewicz P., Kowalczuk M., Banaszak J., Cebrat S., (2005), Biotechnologia, 3, 22-44.
31. Schnoes A. M., Brown S. D., Dodevski I., Babbitt P. C., (2009), PLoS Comput. Biol., 5, e1000605.
32. Rentzsch R., Orengo C. A., (2009), Trends Biotechnol., 27, 210-219.
33. Eddy S. R., (2001), Nat. Rev. Genet., 2, 919-929.
34. Guttman M., Amit I., Garber M., French C., Lin M. F., Feldser D., Huarte M., Zuk O., Carey B. W., Cassady J. P., et al., (2009), Nature, 458, 223-227.
35. Güell M., van Noort V., Yus E., Chen W. H., Leigh-Bell J., Michalodimitrakis K., Yamada T., Arumugam M., Doerks T., Kühner S., et al., (2009), Science, 326, 1268-1271.
36. Gorodkin J., Hofacker I. L., Torarinsson E., Yao Z., Havgaard J. H., Ruzzo W. L., (2010), Trends Biotechnol., 28, 9-19.
37. Zhu Y., Pulukkunat D. K., Li Y., (2007), Nucleic Acids Res., 35, 2283-2294.
38. Pearson T., Okinaka R. T., Foster J. T., Keim P., (2009), Infect. Genet. Evol., 9, 1010-1019.
39. Sobczyński M., Mackiewicz P., Mackiewicz D., Smolarczyk K., Cebrat S., (2005), Biotechnologia, 3, 102-117.
40. Singh J., Behal A., Singla N., Joshi A., Birbian N., Singh S., Bali V., Batra N., (2009), Biotechnol. J., 4, 480-494.
41. Singh B. K., Trends Biotechnol., [Epub ahead of print], PubMed PMID: 20005589.
42. Hugenholtz P., Tyson G. W., (2008), Nature, 455, 481-483.
43. Williams P. G., (2009), Trends Biotechnol., 27, 45-52.

44. Udwary D. W., Zeigler L., Asolkar R. N., Singan V., Lapidus A., Fenical W., Jensen P. R., Moore B. S., (2007), Proc. Natl. Acad. Sci. USA, 104, 10376-10381.
45. Ikeda H., Ishikawa J., Hanamoto A., Shinose M., Kikuchi H., Shiba T., Sakaki Y., Hattori M., Omura S., (2003), Nat. Biotechnol., 21, 526-531.
46. Stratigopoulos G., Bate N., Cundliffe E., (2004), Mol. Microbiol., 54, 1326-1334.
47. Donadio S., Sosio M., Lancini G., (2002), Appl. Microbiol. Biotechnol., 60, 377-380.
48. Kalia V. C., Lal S., Ghai R., Mandal M., Chauhan A., (2003), Trends Biotechnol., 21, 152-156.
49. Paul D., Pandey G., Pandey J., Jain R. K., (2005), Trends Biotechnol., 23, 135-142.
50. Cleveland T. E., Yu J., Fedorova N., Bhatnagar D., Payne G. A., Nierman W. C., Bennett J. W., (2009), Trends Biotechnol., 27, 151-157.
51. Giuliani M. M., Adu-Bobie J., Comanducci M., Aricò B., Savino S., Santini L., Brunelli B., Bambini S., Biolchi A., Capecchi B., et al., (2006), Proc. Natl. Acad. Sci. USA, 103, 10834-10839.
52. Telford J. L., (2008), Cell Host Microbe, 3, 408-416.