



Wykorzystanie sekwencjonowania nowej generacji do poznania genomu ziemniaka

Robert Gromadka

Pracownia Sekwencjonowania DNA i Syntezy Oligonukleotydów,
Instytut Biochemii i Biofizyki, Polska Akademia Nauk, Warszawa

Application of next-generation sequencing method to decode potato genome

Summary

The last years have been a time of exponential progress of sequencing methods. The beginning of the 21st closed a thirty-year domination of sequencing by the Sanger's method. Next generation technology resulted in a mass sequencing of the successive genomes. The time necessary to discover full genomes, as occurred for potato, was dramatically reduced.

Key words:

Sanger's sequencing, pyrosequencing, Potato Genome Sequencing Consortium, Next generation sequencing, National Project Sequencing of Potato Genome.

Adres do korespondencji

Robert Gromadka,
Pracownia
Sekwencjonowania DNA
i Syntezy
Oligonukleotydów,
Instytut Biochemii
i Biofizyki PAN,
ul. Pawińskiego 5A,
02-106 Warszawa.

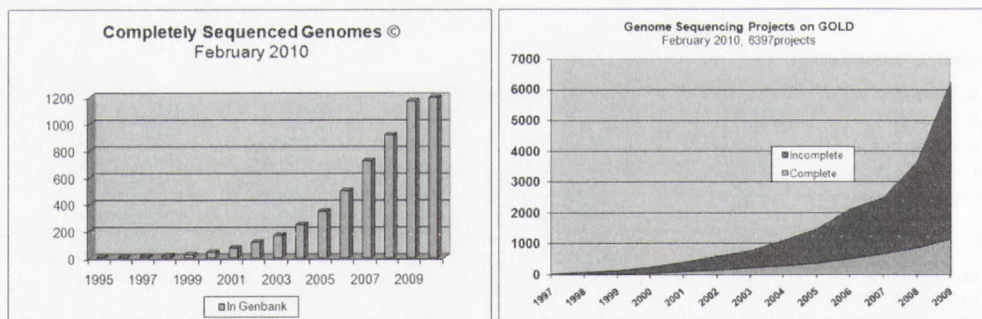
biotechnologia

4 (91) 69–75 2010

1. Wstęp

Kiedy w roku 1977 ukazała się praca Sangera i Alana (1,2), nastąpił pierwszy krok w kierunku uproszczenia i przyspieszenia techniki sekwencjonowania DNA (3). Metoda Sangera dominowała przez kolejne trzy dekady. Metoda ta oparta na syntezie *de novo* nici DNA z wykorzystaniem dideoksyrybonukleotydowych terminatorów podlegała ciągłym usprawnieniom. Początkowo stosowano izotopowo znakowane deoksyrybonukleotydy i sekwencjonowanie przeprowadzano w czterech niezależnych mieszaninach reakcyjnych. Do każdej mieszaniny dodawano niewielką

ilość jednego rodzaju dideoksyrybonukleotydu (ddATP, ddCTP, ddGTP lub ddTTP), o który zatrzymywały przebieg syntezy nowej nici. Po rozdiale elektroforetycznym na niezależnych ścieżkach w żelach poliakrylamidowych pozycję fragmentów DNA ustalano przez naświetlenie kliszy rentgenowskiej (autoradiografia). Obecnie bazuje się na znacznikach fluorescencyjnych dołączonych do dideoksyrybonukleotydu terminatorów. Do każdego rodzaju dołączany jest barwnik emitujący światło o innej długości fali po wzbudzeniu laserem. Pozwala to cały proces przeprowadzić w jednej probówce. Kolejna istotna zmiana nastąpiła wraz z pojawieniem się aparatury nowej generacji (NG, ang. *Next Generation*). Zapowiedzią jej był m.in. artykuł w „Nature” w roku 2005 o maszynym równoległym sekwencjonowaniu na płytce w dołkach o objętości pikolitrowej, opisujący nową technikę sekwencjonowania przez syntezę z wykorzystaniem aparatu firmy 454 (obecnie własność Roche) (4). W tym czasie na rynku pojawiły się sekwenatory Solex firmy Illumina i aparat Solid firmy Applied Biosystems. Pierwsze aparaty 454 generowały ponad 50 razy więcej sekwencji DNA niż powszechnie stosowany 96-kapilarny Applied Biosystems 3730XL i to za jedną szóstą kosztów. Pomimo tak obiecujących wyników naukowcy podchodzili do nich bardzo sceptycznie. Pewien wpływ na to mógł mieć fakt, że duże centra sekwencjonowania zainwestowały ogromne sumy w sekwencjonowanie techniką kapilarną. Przejście na korzystanie z aparatów NG widać wyraźnie po zmianach ilości projektów dotyczących sekwencjonowania całych genomów realizowanych w kolejnych latach (rys. 1). Obecnie, jak wypowiedział się Adam Lowe, dyrektor marketingowy Illuminy (5), użytkownicy ich systemu tygodniowo otrzymują tysiące miliardów nukleotydów sekwencji, co odpowiada 20-krotnej ilości sekwencji zgromadzonych w GenBanku w roku 2005.



Rys. 1. Liczba zsekwenowanych genomów w kolejnych latach (<http://www.genomesonline.org/GOLD> – Genomes OnLine Database v.3.0).

2. Polski Narodowy Projekt Sekwencjonowania Genomu Ziemniaka

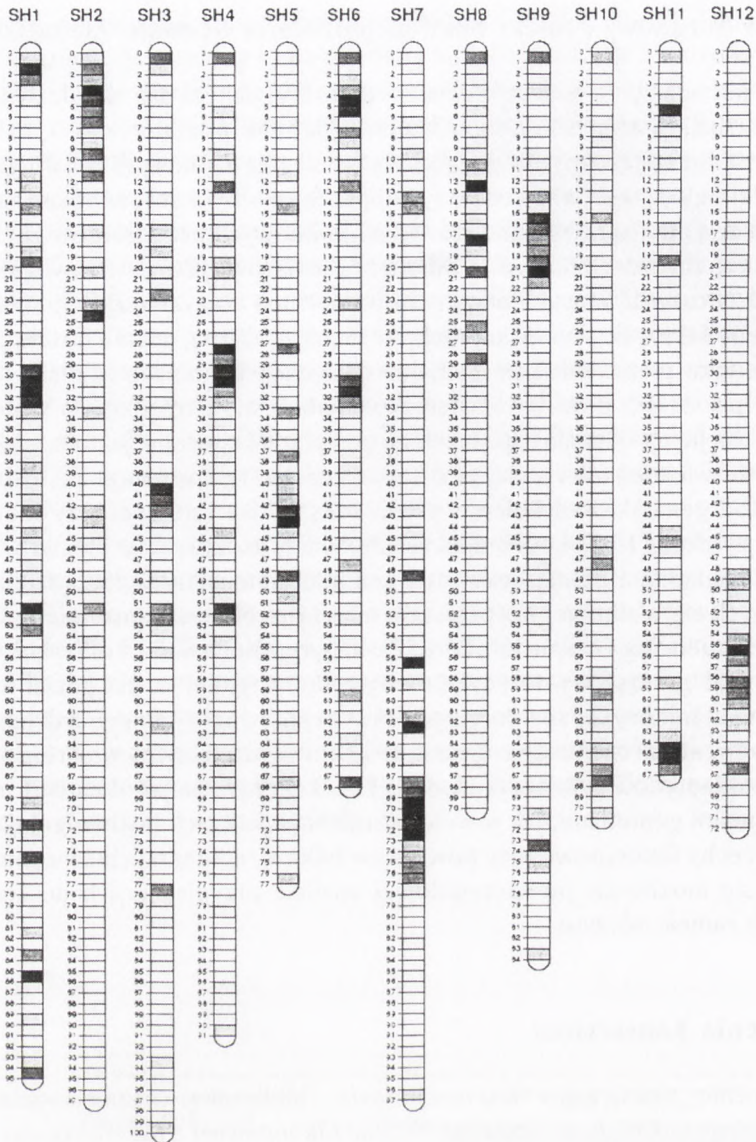
Ziemniak należy do rodziny *Solanace*, do której zalicza się również wiele innych gatunków roślin istotnych ekonomicznie, takich jak pomidor, tytoń, bakłażan czy pieprz. Jest on uprawiany na dużą skalę w Europie i Ameryce od drugiej połowy XIX w. i zajmuje trzecie miejsce po ryżu i pszenicy wśród źródeł pokarmu dla ludzi.

Polska znajduje się na 6. miejscu wśród światowych producentów ziemniaka (6). Na świecie uprawianych jest wiele odmian ziemniaka (7), które stanowią potencjalne źródło do tworzenia nowych odmian odpornych na choroby czy szkodniki (wirusy, zaraza ziemniaka, rak ziemniaka, mokra i sucha zgnilizna, czarna nóżka, parch zwykły, bakterioza pierścieniowa). Poznanie sekwencji genomu ziemniaka (około 840 milionów par zasad) powinno dostarczyć nowoczesne narzędzie do wprowadzania na rynek kolejnych odmian odpornych na pestycydy i choroby, jak również charakteryzujące się większą tolerancją na suszę czy niskie temperatury. Zastosowanie nowych metod genetyki molekularnej skróci czas wytwarzania nowych odmian, obecnie trwający około 10-12 lat, i jednocześnie obniży koszty prowadzonych badań.

W celu poznania pełnej sekwencji genomu ziemniaka i sporządzenia anotacji znajdujących się w nim genów powstało międzynarodowe Konsorcjum „The Potato Genome Sequencing Consortium” (PGSC). Instytut Biochemii i Biofizyki PAN w Warszawie został zaproszony do udziału w tym Konsorcjum. Polski udział w realizacji projektu miał między innymi dotyczyć sekwencjonowania jednego z dwunastu chromosomów genomu ziemniaka. Celem projektu było nie tylko proste poznanie sekwencji nukleotydowej. Miała ona posłużyć w kolejnym etapie do analiz porównawczych z innymi genomami, jak również ustalenia lokalizacji poszczególnych genów. Niektóre cechy fenotypowe były powiązane tylko z regionami chromosomów, teraz pojawiła się możliwość po szczegółowej analizie przypisania ich do konkretnych otwartych ramek odczytu.

3. Założenia Konsorcjum

Pierwotnie zamierzano sekwencjonować bibliotekę chromosomalnego DNA sklonowanego w BAC (ang. *Bacterial Artificial Chromosome*) z heterozygotycznego genotypu diploidalnego ziemniaka RH89-039-16 (RH). Koordynator projektu, *Centre for Biosystems Genomics* w Wageningen, Holandia, przygotował 78 000 klonów niosących wstawki genomowego DNA o średniej długości 120 Kb. Na podstawie analizy restrykcyjnej klony zostały połączone w grupy i przyporządkowane do ~7000 fizycznie zmapowanych kontigów. Równocześnie zostały one przekazane do firmy biotechnologicznej Keygene Inc., Wageningen Holandia (8). Opierając się na własnej technologii AFLP (9) (ang. *Amplified Fragment Length Polymorphism*) w Keygene zmapowano 30 000 klonów BAC na ultra-gęstej mapie zawierającej wzajemną pozycję 10 000 unikatowych markerów AFLP przypisanych do wszystkich 12 chromosomów



Rys. 2. Gęstość i położenie markerów AFLP na dwunastu chromosomach ziemniaka. Markery zaznaczono na ciemno.

ziemniaka. Mapę opracował holenderski zespół z Wageningen (*Plant Breeding Group*) i (10). Mapa genetyczna (rys. 2) wskazuje istniejące zagęszczenia markerów AFLP w różnych chromosomach ziemniaka.

Do każdego chromosomu zostało na tej podstawie przypisane od 50 do 150 BAC. Ich położenie zostało zweryfikowane przez hybrydizację fluorescencyjną

in situ DNA BAC z chromosomami. Miały stać się one miejscami startu do sekwencjonowania metodą BAC-to-BAC.

Kiedy powstawał projekt podstawową metodą sekwencjonowania były techniki oparte na metodzie Sangera z dideoksyrybonukleotydami terminującymi reakcje polimeryzacji nowo powstającej nici DNA. W zależności od firmy dostarczającej aparaty do sekwencjonowania stosowano różne rozwiązania techniczne. Najbardziej wydajny był system firmy Applied Biosystem oparty na czterech różnych barwnikach fluorescencyjnych i rozdziale elektroforetycznym w polimerze, odpowiedniku żelu poliakrylamidowego, umieszczonym w kapilarze. Dzięki automatyzacji procesu możliwe było równoczesne odczytywanie 96 matryc i automatyczne, po zakończeniu jednego odczytu, podstawianie kolejnej porcji reakcji do analizy. W ten sposób można było uzyskać ponad 1000 odczytów o średniej długości 600-800 pz w ciągu doby na jednym aparacie.

W trakcie trwania projektu powszechnie dostępne stały się sekwenatory nowej generacji, tzw. sekwenatory genomowe. W roku 2008 Instytut Biochemii i Biofizyki PAN zakupił jedno z takich urządzeń GS FLX (454) firmy Roche. Umożliwia ono równoczesny odczyt około 100 milionów fragmentów DNA o średniej długości 250 nukleotydów. Zastosowanie tego aparatu zmieniło radykalnie podejście do poznania pełnej sekwencji genomu ziemniaka. Sekwencjonowanie metodą Sangera stało się pomocniczym, używanym głównie do wyjaśnienia miejsc wątpliwych w sekwencji, zamykania przerw między kontigami czy rozwiązywaniem problemów związanych z powtórzeniami w sekwencji.

Z powodu powolnego postępu prac nad sekwencjami genotypu RH związanych z ograniczeniami spowodowanymi heterozygotycznością i dostępem do aparatury nowej generacji, Konsorcjum zdecydowało się na równoległe prowadzenie sekwencjonowania drugiego genotypu, homozygoty (monoploida) DM1-3 516R44.

4. Strategia sekwencjonowania metodą Sangera

DNA genomowe ziemniaka zostało pofragmentowane przez niepełne trawienie enzymami restrykcyjnymi *EcoRI* i *HinDIII*, a następnie fragmenty o długości 100-150 kbp sklonowano w BAC. Tak przygotowana biblioteka została udostępniona członkom Konsorcjum. Korzystając z oprogramowania przygotowanego przez holenderskiego koordynatora i udostępnionego członkom Konsorcjum, na podstawie mapy fizycznej wytypowano BACi przypisane do chromosomu VII. Miał to być wyjściowy materiał do sekwencjonowania metodą Sangera. Wyizolowane DNA BAC zostało ponownie pofragmentowane na odcinki długości 2-3 tysięcy par zasad i sklonowane w wektorze pBluScript. Taka strategia wymaga tworzenia kolejno dwóch bibliotek i selekcji klonów niosących DNA genomu ziemniaka. Z powodu niezupełnie losowego fragmentowania DNA przez niepełne trawienie duża część sekwencji w kolejnych BAC się powtarza. W niektórych przypadkach nawet do

20-30 kbp. Ponadto niektóre sekwencje z różnych powodów takich jak toksyczność czy odmienna zawartość par G/C, okazały się trudne do klonowania w bakteriach lub niestabilne. Tych sekwencji brakowało w trakcie składania konsensusu.

5. Strategia sekwencjonowania nowej generacji

W nowej technologii sekwencjonowania całych genomów pomija się wstępne etapy klonowania i tworzenia bibliotek w bakteriach. Dzięki temu nie ma wad związanych z nietolerancją bakterii na informację przenoszoną przez badany DNA czy skład nukleotydowy. Technologia ta ma też minusy wynikające z dużej ilości uzyskiwanych informacji, która musi zostać poddana analizie przy składaniu konsensusu. Pewnym problemem są też krótsze, pojedyncze odczyty, chociaż w marcu 2010 r. można już było odczytywać około 500 pz, co zaczyna być porównywalne do sekwencjonowania metodą Sangera. Pomimo że nowa technologia wymaga większej liczby odczytów, 30-krotne pokrycie w porównaniu z 8-10-krotnym pokryciem wymaganym przy metodzie Sangera, koszt uzyskania konsensusu jest znacząco niższy. W celu wykorzystania już istniejącej biblioteki BAC do sekwencjonowania została zastosowana technika tzw. MID (ang. *Multiplex IDentifier*). Standardowo DNA po fragmentowaniu do odcinków długości średnio 800 pz jest przygotowywany do przyłączenia na końcach dwuniciowych adapterów niezbędnych do amplifikacji materiału przed sekwencjonowaniem. Sekwencje adapterowe służą również jako miejsca przyłączenia startera do sekwencjonowania. W technice MID adaptery są dodatkowo przedłużone o specyficzne 10-nukleotydowe sekwencje, inne dla każdego BAC-a, które pozwolą później je rozróżnić i przypisać do poszczególnych klonów BAC przez program komputerowy (np. *sffile* firmy Roche). Dzięki tej metodzie jest możliwe równoległe sekwencjonowanie kilku BAC-ów jednocześnie w trakcie jednego przebiegu reakcji w genomowym sekwenatorze.

6. Sekwencjonowanie BAC dwoma metodami

W trakcie wdrażania nowego systemu sekwencjonowania przeprowadziliśmy dla porównania ponowne sekwencjonowanie jednego z wcześniej wybranych BAC, którego konsensus został ustalony metodą Sangera. Powtórne sekwencjonowanie na aparacie GS FLX (454) wygenerowało konsensus zgodny z wcześniej otrzymanym. Jednak różnica w czasie i nakładzie sił i środków była nieporównywalna na korzyść aparatu nowej generacji. Podczas gdy przygotowanie banku DNA klonu BAC RH200P15 do sekwencjonowania zajęło ponad dwa tygodnie i wymagało żmudnej selekcji klonów niosących wstawki DNA, a następnie zabezpieczenia ich w postaci zamrożonych hodowli na płytkach 96-dołkowych. W drugim przypadku wyizolowany DNA klonu BAC RH200P15 był wprost użyty do sekwencjonowania. W ciągu jed-

nego tygodnia zostało zakończone sekwencjonowanie klonu RH200P15 i można było analizować otrzymane sekwencje, które złożyły się w konsensus o długości 149532 pz.

7. Podsumowanie

Gwałtowny rozwój nowych technologii pozwala wierzyć, że genomika personalna to nie odległa przyszłość, ale właściwie już teraźniejszość. Sekwencjonowanie pierwszego genomu człowieka trwało ponad 10 lat i pochłonęło kwotę ponad 3 bilionów dolarów. Obniżenie kosztów i skrócenie czasu generowania sekwencji stało się punktem wyjściowym ogłoszenia w 2007 r. projektu „Projekt 1000 genomów” (*The 1000 Genome Project*), w którym mamy poznać genomy ludzi z różnych części świata. W styczniu 2010 r. firma Illumina podała, że mający wejść do użycia ich nowy sekwenator HiSeq2000 powinien zsekwencjonować równocześnie dwa genomy ludzkie za równowartość 10 tysięcy USD z 30-krotnym pokryciem (11). Duże zainteresowanie budzi również, mający w tym roku wejść do powszechnego użycia, nowy system sekwencjonowania w czasie rzeczywistym pojedynczej cząsteczki DNA firmy Pacific Biosciences (12).

Opracowanie powstało w ramach realizacji projektu badawczego finansowanego przez Ministerstwo Nauki i Szkolnictwa Wyższego: 47/PGS/2006/01.

Literatura

1. Sanger F., Air G. M., Barrell B. G., Brown N. L., Coulson A. R., Fiddes J. C., Hutchison C. A., Slocombe P. M., Smith M., (1977), *Nature*, 265, 687-695.
2. Sanger F., Nicklen S., Coulson A. R., (1977), *Proc. Natl. Acad. Sci. USA*, 74, 5463-5467.
3. Sanger F., Coulson A. R. J., (1975), *Mol. Biol.*, 94, 441-448.
4. Margulies M., et al., (2005), *Nature*, 437 (7057), 376-380.
5. Perkel J. M., (2009), *Science Magazine*, 324, (5924), 275-279.
6. FAO Crops statistics database <http://faostat.fao.org/>
7. Spooner D. M., Hijmans R. J., (2001), *American Journal of Potato Research*, 78, 237-2268.
8. <http://www.keygene.com/>
9. Vos P., Hogers R., Bleeker M., et al., (1995), *Nucleic Acids Res.*, 23 (21), 4407-4414.
10. Isidore E., et al., (2003), *Genetics*, 165, 2107-2116.
11. Press Release Source: Illumina, Inc. On Tuesday January 12, (2010), 2:50 pm EST.
12. <http://www.pacificbiosciences.com/>