KOSMOS
PROBLEMY NAUK BIOLOGICZNYCH

Tom 58 2009
Numer 1–2 (282–283)
Strony 17–28

Polskie Towarzystwo Przyrodników im. Kopernika

MARTA KASPRZAK[1,2], ALEKSANDRA ŚWIERCZ[1,2]

[1]Instytut Informatyki,
Politechnika Poznańska
Piotrowo 2, 60-965 Poznań
[2]Instytut Chemii Bioorganicznej PAN
Noskowskiego 12/14, 61-704 Poznań
E-mail: mkasprzak@cs.put.poznan.pl
aswiercz@cs.put.poznan.pl

SEKWENCJONOWANIE I ASEMBLACJA DNA — PODEJŚCIA, MODELE GRAFOWE, ALGORYTMY

WPROWADZENIE

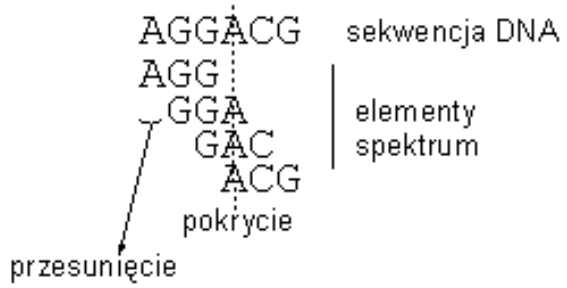
Kwas deoksyrybonukleinowy (DNA) jest cząsteczką kodującą informację genetyczną organizmu, złożoną z dwóch nici połączonych ze sobą za pomocą wiązań wodorowych (tzw. helisa DNA). Każda nić to łańcuch nukleotydów, których kolejność stanowi tę informację, zapisywanych symbolicznie jako 'A', 'C', 'G', 'T'. 'A' oznacza nukleotyd z zasadą azotową adeniną, 'C' z cytozyną, 'G' z guaniną, 'T' z tyminą. Nukleotyd A z jednej nici helisy łączy się za pomocą wiązania wodorowego z nukleotydem T położonym na przeciwko niego w drugiej nici, natomiast nukleotyd G łączy się z C. Ta zasada występowania w parach A z T oraz G z C nazywana jest zasadą komplementarności. Dzięki tej właściwości, znając fragment jednej nici, możemy odtworzyć komplementarną do niego sekwencję nukleotydów w drugiej nici. Krótki fragment jednoniciowego DNA nazywany jest oligonukleotydem.

Odczytywanie kolejności nukleotydów w sekwencji DNA (czyli rozpoznanie informacji genetycznej organizmu) odbywa się w kilku etapach. W pierwszym etapie, sekwencjonowaniu, odczytywane są sekwencje DNA o długości zazwyczaj kilkuset nukleotydów. W kolejnym etapie, asemblacji, sekwencje te są łączone w dłuższe odcinki, w wyniku czego otrzymujemy sekwencje o długości nawet do miliona nukleotydów. Dla krótkich genomów, np. wirusów lub bakterii, te dwa etapy często wystarczą, aby odczytać cały genom. Dla dłuższych, potrzebny jest trzeci etap, mapowanie, polegający na właściwym uszeregowaniu względem siebie zasemblowanych sekwencji. W tym celu stosuje się inne niż w poprzednich etapach podejścia biochemiczne, np. z użyciem enzymów restrykcyjnych (mapowanie restrykcyjne, patrz np. BŁAŻEWICZ i współaut. 2001).

Etap sekwencjonowania może być realizowany na kilka sposobów. Najbardziej popularną metodą do niedawna była elektroforeza żelowa (MAXAM i GILBERT 1977, SANGER i współaut. 1977), która generuje sekwencje o długości kilkuset nukleotydów. Metoda ta ogranicza się do etapu laboratoryjnego (eksperyment biochemiczny) i nie wymaga etapu algorytmicznego (przetwarzanie danych), co z jednej strony czyni ją prostą i przystępną, z drugiej strony jednak nieodporną na błędy eksperymentalne.

Bardziej zaawansowanym technologicznie i koncepcyjnie podejściem jest sekwencjonowanie przez hybrydyzację (SBH), w wyniku którego otrzymujemy sekwencje o długości do tysiąca nukleotydów, ale w tym przypadku potrzebne są metody algorytmiczne do przetwarzania danych eksperymentalnych (SOUTHERN 1988). Na mikromacierz DNA nanosi się bibliotekę oligonukleotydów, czyli zbiór krótkich, jednoniciowych fragmentów łańcucha DNA (np. wszystkie 4^l oligonukle-

otydy o zadanej długości *l*). Podczas eksperymentu hybrydyzacyjnego badana sekwencja DNA przykleja się do komplementarnych oligonukleotydów na mikromacierzy. W jego wyniku znajdowany jest zbiór *S* (spektrum) fragmentów, które zawarte są w oryginalnej sekwencji (szerszy opis eksperymentu z mikromacierzą DNA zawarty jest np. w (SOUTHERN i współaut. 1992)). Celem części obliczeniowej SBH jest rekonstrukcja oryginalnej sekwencji o znanej długości *n* z elementów spektrum. Rekonstrukcja polega na takim uszeregowaniu wszystkich elementów spektrum, aby kolejne oligonukleotydy w rozwiązaniu nakładały się na siebie z przesunięciem równym 1 (przy założeniu braku błędów eksperymentalnych, patrz Ryc. 1).



Ryc. 1. Rekonstrukcja sekwencji AGGACG ze spektrum *S*={ACG, AGG, GAC, GGA}.

Kolejne oligonukleotydy nakładają się na siebie z przesunięciem równym 1. Pokrycie każdej pozycji w sekwencji wyznacza liczba oligonukleotydów, które nakładają się w tym miejscu na sekwencję, np. pokrycie czwartego nukleotydu w sekwencji wynosi 3.

Znalezienie rozwiązania (sekwencji) jest problemem łatwym obliczeniowo, o ile w spektrum nie pojawią się błędy (PEVZNER 1989). Jednakże w rzeczywistym eksperymencie pojawiają się błędy i mogą być one dwojakiego rodzaju: błędy negatywne — jeśli brakuje jakiegoś oligonukleotydu w spektrum, pomimo że występuje on w badanej sekwencji — oraz błędy pozytywne, gdy w spektrum znajduje się nadmiarowy oligonukleotyd, który nie występuje w sekwencji. W przypadku, gdy w spektrum występują błędy, podczas rekonstrukcji badanej sekwencji zakłada się, że przesunięcie pomiędzy kolejnymi elementami spektrum może być większe niż 1 (co odpowiada błędom negatywnym), a niektóre oligonukleotydy nie znajdą się w ogóle w rozwiązaniu (błędy pozytywne). Problem ten jest wówczas obliczeniowo trudny (BŁAŻEWICZ i KASPRZAK 2003). Szczególnym rodzajem błędów negatywnych są powtórzenia fragmentów sekwencji DNA, które są co najmniej tak długie jak oligonukleotydy użyte w eksperymencie. Jeśli oligonukleotyd występuje kilkakrotnie w sekwencji, w spektrum pojawia się on tylko raz. Odtworzenie sekwencji oryginalnej w przypadku powtórzeń jest trudniejsze ze względu na wiele możliwości optymalnego (z obliczeniowego punktu widzenia) dopasowania fragmentów. W celu zminimalizowania liczby błędów pochodzących z eksperymentu zaproponowane zostały nowe podejścia, które w trakcie hybrydyzacji zamiast oligonukleotydów o równej długości wykorzystują oligonukleotydy, dla których temperatura zajścia idealnej hybrydyzacji jest taka sama, albo oligonukleotydy zawierające zasady uniwersalne mogące łączyć się z dowolną zasadą azotową. Algorytmy do rekonstrukcji oryginalnej sekwencji dla każdego z tych podejść (z oligonukleotydami o równej długości, o równej temperaturze oraz zawierającymi zasady uniwersalne) zostaną przedstawione w części „Sekwencjonowanie".

Od niedawna stosuje się wysoce zautomatyzowane podejścia do sekwencjonowania, które w stosunkowo krótkim czasie generują miliony sekwencji o długości kilkudziesięciu (technologie Solexa i SOLID) lub kilkuset nukleotydów (metoda 454). Każdy z tych systemów posiada inną strategię generowania wysokiej jakości danych. Metoda 454 firmy Roche (MARGULIES i współaut. 2005) polega na sekwencjonowaniu przez syntezę fragmentów DNA, które przyczepiane są do małych koralików, a następnie klonowane w emulsji wodno-oleistej. Kolejne nukleotydy są syntetyzowane do fragmentów DNA w wyniku szeregu reakcji chemicznych. W jednym kroku może się przykleić nawet kilka nukleotydów tego samego typu. Podczas sekwencjonowania metodą Solexa firmy Illumina (BENNETT 2004) sekwencja jest przyczepiana do powierzchni płytki. Do sekwencji kolejno syntetyzowane są specjalnie zaprojektowane, znakowane kolorem nukleotydy, które każdorazowo kończą syntezę DNA, a następnie (po odczytaniu koloru) za pomocą enzymów są odblokowywane. W metodzie SOLID firmy Applied Biosystems (FU i współaut. 2008), zamiast pojedynczych nukleotydów, przyczepiane są znakowane kolorem oligonukleotydy, w których znane są pierwsze dwa nukleotydy. W kolejnych cyklach znajdowane są następne dwu-nukleotydy tak, że w sumie otrzymujemy podwójne pokrycie każdej pozycji w sekwencji.

Drugi etap odczytywania sekwencji DNA, asemblacja, polega na połączeniu fragmentów DNA pochodzących z etapu sekwencjonowania w dłuższe odcinki. Spośród wielu algorytmów, które rozwiązują problem asemblacji, kilka z nich zostanie omówionych tutaj ze względu na interesujące podejścia grafowe. Są wśród nich zarówno algorytmy skonstruowane dla dłuższych fragmentów wejściowych (klasyczna asemblacja), jak i dla krótszych, pochodzących z nowych metod sekwencjonowania. Wszystkie one zostaną przedstawione w części „Asemblacja".

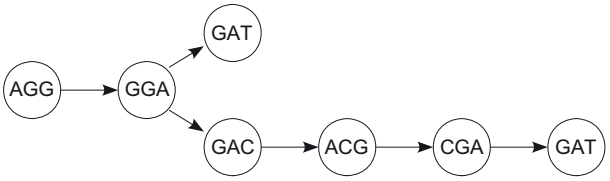
SEKWENCJONOWANIE

Pierwszy algorytm do rekonstrukcji oryginalnej sekwencji na podstawie wyników eksperymentu hybrydyzacyjnego z mikromacierzą DNA został zaproponowany w (BAINS i SMITH 1988). Autorzy założyli brak błędów w spektrum. Algorytm buduje drzewo, w którym wierzchołkami są elementy spektrum, natomiast łuki łączą wierzchołki, jeśli l-1 ostatnich liter poprzednika nakłada się na l-1 pierwsze litery następnika. Rozwiązaniem jest sekwencja odpowiadająca ścieżce w tym drzewie od korzenia do liścia zawierającej wszystkie elementy spektrum dokładnie jeden raz (Przykład 1). Jako korzeń drzewa wybierany jest ten element, który jest początkiem badanej sekwencji. Jeśli natomiast nie jest on znany, algorytm musi skonstruować $|S|$ drzew z każdym kolejnym elementem spektrum jako korzeniem.

LYSOV i współaut. (1988) zauważyli, że problem SBH bez błędów w spektrum można sprowadzić do znanego problemu poszukiwania ścieżki Hamiltona w pewnym grafie. Skierowany graf H konstruowany jest w następujący sposób. Każdy wierzchołek grafu odpowiada innemu elementowi spektrum. Łuk (u,v) łączy wierzchołek u z wierzchołkiem v jeśli l-1 ostatnich liter etykiety wierzchołka u nakłada się na l-1 pierwsze litery etykiety v. W takim grafie poszukiwana jest ścieżka przechodząca przez wszystkie wierzchołki dokładnie jeden raz (ścieżka Hamiltona) (Przykład 1).

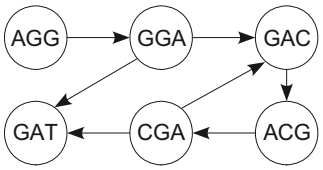
PRZYKŁAD 1

Załóżmy, że dla badanej sekwencji nukleotydów AGGACGAT eksperyment hybrydyzacji przebiegł bez błędów, w wyniku czego otrzymano spektrum: S = {ACG, AGG, CGA, GAC, GAT, GGA}. Długość badanej sekwencji n = 8, długość oligonukleotydów l = 3, natomiast $|S| = n–l+1$. Dla uproszczenia przykładu załóżmy, że znany jest pierwszy oligonukleotyd (AGG). Metoda Bainsa i Smitha tworzy drzewo dodając kolejne elementy o ile nie znajdują się już w bieżącej ścieżce (Ryc. 2). Tylko dolna ścieżka przechodzi przez wszystkie elementy spektrum. Odczytując kolejne etykiety można zrekonstruować sekwencję oryginalną.



Ryc. 2. Drzewo dla metody zaproponowanej przez BAINSA i SMITHA (1988).

Metoda LYSOVA i współaut. (1988) utworzy na podstawie tego samego spektrum graf, który zaprezentowany jest na Ryc. 3. W grafie istnieje dokładnie jedna ścieżka Hamiltona, która odpowiada badanej sekwencji.



Ryc. 3. Graf dla metody zaproponowanej przez LYSOVA i współaut. (1988).

Powyższe metody działają tylko dla idealnego, bezbłędnego spektrum, jednakże ich złożoność obliczeniowa jest wykładnicza (tzn. liczbę elementarnych operacji algorytmu można wyrazić funkcją wykładniczą, w której rozmiar instancji rozwiązywanego problemu jest wykładnikiem potęgi). Algorytm o wielomianowej złożoności czasowej (tzn. w którym liczba operacji jest wyrażona funkcją wielomianową) dla problemu SBH został zaproponowany przez PEVZNERA (1989), co dowodzi, że problem należy do klasy problemów łatwych obliczeniowo. Algorytm ten szuka ścieżki przechodzącej przez wszystkie łuki w grafie skierowanym dokładnie raz (tj.

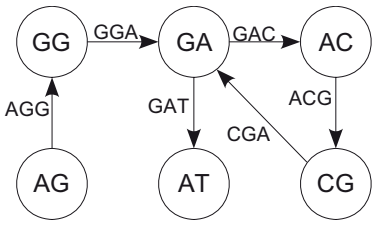
ścieżki Eulera). Tym razem elementy spektrum odpowiadają łukom, a każdy z łuków wychodzi z wierzchołka, który jest zaetykietowany *l*-1-literowym prefiksem etykiety łuku i wchodzi do wierzchołka zaetykietowanego *l*-1-literowym sufiksem (Przykład 2).

Transformacja grafu, w którym poszukiwana jest ścieżka Hamiltona w graf, w którym poszukiwana jest ścieżka Eulera, zmienia złożoność obliczeniową problemu. Klasa grafów etykietowalnych, dla których taka transformacja jest możliwa, została szeroko omówiona przez BŁAŻEWICZA i współaut. (1999c). Grafy budowane na podstawie spektrum w metodzie LYSOVA i współaut. (1988), nazwane grafami DNA, należą do klasy grafów etykietowalnych. Grafy skonstruowane na podstawie metody LYSOVA i współaut. (1988) są grafami liniowymi grafów PEVZNERA (1989). Dla takiej pary grafów poszukiwanie ścieżek Hamiltona i Eulera jest równoważne (patrz BŁAŻEWICZ i współaut. 1999c).

Kolejny algorytm zaprezentowany przez PEVZNERA (1989) dopuszczał błędy negatywne w spektrum. Złożoność czasowa algorytmu jest wielomianowa, jednakże nie w każdym przypadku algorytm potrafi znaleźć rozwiązanie, więc nie może być traktowany jako algorytm dokładny. W tej metodzie najpierw wyznaczana jest liczba brakujących oligonukleotydów, równa n-l+1-|S|, a następnie oligonukleotydy te są znajdowane poprzez transformację problemu sekwencjonowania do problemu poszukiwania przepływu w sieci zbudowanej na podstawie grafu dwudzielnego $K_{m,m}$. Graf dwudzielny zbudowany jest z wierzchołków grafu Pevznera, które posiadają różną liczbę łuków wchodzących i wychodzących. Jeśli różnica ta jest większa niż 1 dla pewnego wierzchołka, liczba jego wystąpień jest odpowiednio zwiększana. Po lewej stronie grafu dwudzielnego umieszczane są wierzchołki z większą liczbą łuków wchodzących, a po prawej z większą liczbą łuków wychodzących. Liczba wierzchołków z lewej strony równa jest liczbie wierzchołków z prawej strony i wynosi m. Z każdego wierzchołka z lewej strony wychodzi łuk do każdego wierzchołka z prawej strony. Koszt łuku jest równy najmniejszemu przesunięciu względem siebie etykiet wierzchołków minus 1. Zatem koszt jest równy 1, jeśli nałożenie etykiet jest równe *l*-2, natomiast jeśli etykiety w ogóle się na siebie nie nakładają koszt równy jest *l*-1 (koszt łuku oznacza ile wierzchołków/oligonukleotydów brakowałoby w grafie Pevznera, gdybyśmy chcieli dany łuk wykorzystać). Dodatkowo do grafu dwudzielnego dodawane jest źródło s, z którego wychodzą łuki do każdego wierzchołka z lewej strony, oraz ujście t, do którego dochodzą łuki od każdego wierzchołka z prawej strony. W tak skonstruowanej sieci, gdzie wszystkie łuki mają pojemność równą 1, poszukiwany jest przepływ o wartości m-1 i o minimalnym koszcie. Jeśli koszt okazałby się równy liczbie brakujących oligonukleotydów (czyli n-l+1-|S|), wówczas graf Pevznera zostałby uzupełniony o brakujące łuki (lub ścieżki) i można by w nim poszukiwać ścieżki Eulera (Przykład 2).

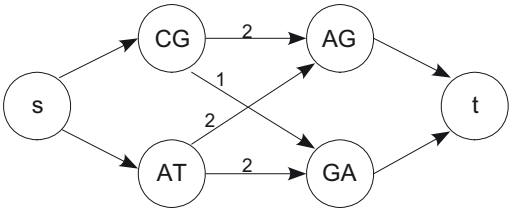
PRZYKŁAD 2

Rozważmy tę samą sekwencję jak w Przykładzie 1. Spektrum idealne S = {ACG, AGG, CGA, GAC, GAT, GGA}. Graf Pevznera został przedstawiony na Ryc. 4.



Ryc. 4. Graf dla metody PEVZNERA (1989).

Ścieżka Eulera w tym grafie odpowiada badanej sekwencji AGGACGAT. Załóżmy teraz, że podczas eksperymentu hybrydyzacji wystąpiły błędy i w spektrum brakuje elementu CGA (błąd negatywny). Aby utworzyć kompletny graf Pevznera należy wyznaczyć liczbę brakujących oligonukleotydów n-l+1-|S| = 1 (przepływu o takim koszcie będziemy szukać) a następnie skonstruować sieć z wierzchołków o różnej liczbie łuków wchodzących i wychodzących w grafie z Ryc. 4 pozbawionego łuku CGA (Ryc. 5).



Ryc. 5. Sieć dla metody PEVZNERA (1989).

Koszty na łukach są równe przesunięciom pomiędzy etykietami wierzchołków przy zało-

żeniu dokładnego nałożenia prefiksu lewego wierzchołka z sufiksem prawego wierzchołka. W takiej sieci szukamy przepływu o wartości m-1=1 i o koszcie równym 1. Istnieje tylko jeden taki przepływ zawierający wierzchołki (CG, GA), co odpowiada brakującemu oligonukleotydowi CGA. Po dodaniu tego łuku do grafu Pevznera możemy poszukiwać ścieżki Eulera.

Pierwszy algorytm dokładny, który dopuszcza występowanie błędów zarówno pozytywnych jak i negatywnych w spektrum i nie wymaga żadnej dodatkowej informacji o elementach spektrum został zaproponowany przez BŁAŻEWICZA i współaut. (1999a). Problem sekwencjonowania został sformułowany jako wariant problemu komiwojażera (ang. selective traveling salesman problem). Tworzony jest graf skierowany pełny, w którym wierzchołki są zaetykietowane oligonukleotydami a koszt łuku pomiędzy dwoma wierzchołkami jest równy minimalnemu przesunięciu odpowiednich etykiet przy założeniu ich dokładnego nałożenia. Z każdym wierzchołkiem skojarzony jest zysk o wartości 1. W takim grafie poszukiwana jest ścieżka o największym sumarycznym zysku i koszcie nie przekraczającym n-l, która jest równoważna sekwencji o długości nie większej niż n utworzonej z maksymalnej liczby oligonukleotydów ze spektrum (Przykład 3).

PRZYKŁAD 3

Załóżmy, że dla sekwencji AGGACGAT spektrum z błędami negatywnymi i pozytywnymi wygląda następująco: S = {ACG, AGG, CAC, GAC, GAT, GGA}. W spektrum, oprócz błędu negatywnego CGA, pojawił się jeden błąd pozytywny CAC. Graf pełny skonstruowany dla tej metody (BŁAŻEWICZ i współaut. 1999a) zawierałby łuki o kosztach przedstawionych Tabela 1.

Zysk za odwiedzenie każdego wierzchołka jest równy 1. Algorytm szuka ścieżki przechodzącej co najwyżej jeden raz przez każdy wierzchołek, o maksymalnym zysku i koszcie, nie większym niż n-l=5. Jako rezultat otrzymujemy dwie ścieżki, które przechodzą przez 5 wierzchołków i o koszcie równym 5. Odczytując etykiety wierzchołków otrzymamy rozwiązania: AGGACGAT oraz AGGACACG.

Tabela 1. Koszty łuków dla metody wg BŁAŻEWICZA i współaut. (1999a).

	ACG	AGG	CAC	GAC	GAT	GGA
ACG	–	3	3	2	2	2
AGG	3	–	3	2	2	1
CAC	1	3	–	3	3	3
GAC	1	3	2	–	3	3
GAT	3	3	3	3	–	3
GGA	2	2	3	1	1	–

Problem SBH w przypadku, gdy występują błędy, nawet błędy tylko jednego typu (pozytywne lub negatywne), jest problemem silnie NP-trudnym (tzn. prawdopodobnie nie może dla niego zostać skonstruowany algorytm o wielomianowej złożoności czasowej). Zatem algorytm dokładny dla większych instancji problemu może nie zakończyć działania w sensownym czasie. W efekcie większość zaproponowanych dla tego problemu algorytmów to heurystyki, które w znacznej mierze przyspieszają przeszukanie przestrzeni rozwiązań, jednocześnie nie gwarantując, że znalezione rozwiązanie jest optymalne.

KRUGLYAK (1998) zaproponował wielostopniowe podejście do SBH. Kolejno wykonywane są eksperymenty z oligonukleotydami o zwiększanej długości. Pierwszy eksperyment przeprowadzany jest dla kompletnej biblioteki oligonukleotydów o długości l. W kolejnych eksperymentach biblioteka oligonukleotydów składa się z połączonych ze sobą oligonukleotydów, które znalazły się w spektrum po poprzednim eksperymencie, przy założeniu jak największego możliwego nałożenia. Ustawienie progu dla minimalnej liczby nakładających się nukleotydów na wartość większą niż 0 (0 oznacza konkatenację pary oligonukleotydów) zmniejsza znacząco rozmiar biblioteki generowanej dla następnego eksperymentu. Zaletą tego podejścia jest redukcja błędów wynikających z powtórzeń ciągu nukleotydów w sekwencji oryginalnej, jednakże inne błędy negatywne oraz błędy pozytywne będą propagowane w kolejnych krokach tej metody.

Interaktywne podejście do SBH zaproponowali PHAN i SKIENA (2001). Zwykły eksperyment hybrydyzacji wzbogacony jest o dodatkowe eksperymenty, które mają na celu rozwianie wszelkich niejednoznaczności podczas rekonstrukcji oryginalnej sekwencji. Metoda ta konstruuje graf, taki sam jak w przypadku metody LYSOVA i współaut. (1988), gdzie oligonukleotydy są etykietami wierzchołków a łuki łączą wierzchołki, których etykiety przesunięte są względem siebie o jedną pozycję. Seria dodatkowych zapytań (eksperymentów biochemicznych) z oligonukleotydami o zwiększanej długości pozwala na wyeliminowanie niepotrzebnych wierz-

chołków (błędy pozytywne), lub dodanie brakujących (błędy negatywne). Kiedy wszystkie rozgałęzienia w grafie są rozstrzygnięte (graf staje się ścieżką prostą), algorytm kończy zapytania i jako wynik zwraca jedyną ścieżkę Hamiltona w grafie.

Metoda zaprezentowana przez BUI i YOUSSEFA (2004), to algorytm genetyczny. W początkowej fazie algorytmu oligonukleotydy, które nakładają się na siebie na *l*-1 pozycjach, łączone są na stałe w dłuższe fragmenty. Jeśli jest więcej niż jedna możliwość połączenia oligonukleotydów, wówczas takie oligonukleotydy nie są łączone. W dalszej części algorytm operuje na dłuższych fragmentach zamiast na oligonukleotydach. Każdy osobnik w algorytmie genetycznym reprezentowany jest jako permutacja wszystkich fragmentów. Liczba osobników w populacji pozostaje cały czas stała. Dopasowanie osobnika jest obliczane na podstawie liczby nakładających się pozycji pomiędzy sąsiednimi fragmentami oraz długości sekwencji. Rodzice następnego pokolenia w populacji wybierani są proporcjonalnie do ich dopasowania. Następnie rodzice przekazują cechy swojemu potomstwu przy użyciu operatorów krzyżowania oraz mutacji. Podczas krzyżowania wyznaczane są 3 miejsca podziału u rodziców a następnie ich potomstwo wybiera różne części od rodziców. Przy takim podziale niektóre fragmenty mogą zostać zduplikowane u potomstwa, więc mechanizm naprawczy usuwa zduplikowane fragmenty a dodaje fragmenty, które zostały pominięte. Dodatkowo mechanizm naprawczy ulepsza rozwiązanie poprzez lokalne przeszukanie sąsiedztwa w zbiorze rozwiązań. W nowo utworzonym potomstwie mutacja zachodzi z 10% prawdopodobieństwem, zamieniając miejscami losowo wybrane fragmenty. Potomstwo zastępuje najgorszych osobników w pokoleniu rodziców. Algorytm kończy działanie, gdy nie uda się polepszyć rozwiązania.

Według ZHANGA i współaut. (2003) algorytm zakłada ograniczenie na liczbę brakujących kolejnych oligonukleotydów w sekwencji. Autorzy zauważyli, że limit błędu Δ (liczba brakujących kolejnych oligonukleotydów) może być równy $1 \le \Delta < l$. Jeśli $\Delta \ge l$, wówczas brakowałoby informacji do rekonstrukcji sekwencji oryginalnej. Autorzy najpierw przetransformowali problem sekwencjonowania z błędami negatywnymi i pozytywnymi do problemu z błędami tylko pozytywnymi. Dla każdej pary elementów *u, v* ze spektrum *S*, dla której przesunięcie pomiędzy *u* a *v* jest równe *t*, $1 < t \le \Delta+1$, wyznaczana jest ścieżka złożona z *t*-1 oligonukleotydów wypełniających lukę między *u* i *v*. Wszystkie elementy ze ścieżki, których nie ma w spektrum, dodawane są do rozszerzonego spektrum *S'*. Jeśli Δ jest zbyt duże, spowoduje to dodanie zbyt dużej liczby elementów do *S'* (założono, że $\Delta \le 3$). Po tej wstępnej transformacji w *S'* mogą się pojawić tylko błędy pozytywne (przy założeniu, że limit Δ jest poprawny). Następnie algorytm tworzy macierz sąsiedztwa *A* o rozmiarze $|S'|\times|S'|$, w której '1' oznacza, że dwa elementy nakładają się na *l*-1 pozycjach, a '0' przeciwny przypadek. W dalszej kolejności obliczane są macierze A^k, w których na pozycji $a^k_{i,j}$ znajduje się liczba ścieżek o długości *k* pomiędzy elementem s_i oraz s_j. Celem algorytmu jest maksymalizacja *k*, czyli wydłużanie ścieżki, a tym samym zwiększanie liczby elementów ze spektrum, przez które przechodzi. Jednocześnie $k \le n$-1, gdyż długość rozwiązania nie może przekroczyć *n*. Jako rozwiązanie zwracane są wszystkie ścieżki o maksymalnej wartości *k*.

Algorytm zaproponowany przez BŁAŻEWICZA i współaut. (2004) jest połączeniem dwóch heurystyk — przeszukiwania tabu i *scatter*. Celem przeszukiwania tabu jest maksymalizacja liczby elementów ze spektrum wykorzystanych w rozwiązaniu, natomiast zadaniem drugiej heurystyki jest wprowadzenie różnorodności w przeszukiwaniu przestrzeni rozwiązań. Najlepsze rozwiązania znalezione w trakcie przeszukiwania tabu są zapamiętywane a następnie wykorzystywane przy tworzeniu kolejnego rozwiązania początkowego i cały proces przeszukiwania tabu rozpoczyna się od nowa. Algorytm kończy działanie po kilku powtórzeniach całego cyklu i zwraca najlepsze znalezione rozwiązanie.

Wszystkie zaprezentowane powyżej algorytmy działają dla spektrum, w którym oligonukleotydy są równej długości. Czasami jednak z powodu błędów w spektrum nie jest możliwa jednoznaczna rekonstrukcja sekwencji. Zaproponowane zostało nowe podejście (PREPARATA i współaut. 1999, PREPARATA i UPFAL 2001), które ma na celu wydłużenie oligonukleotydów przy jednoczesnym zachowaniu rozmiaru biblioteki. Wydłużone oligonukleotydy zawierają zasady uniwersalne, tj. takie cząsteczki chemiczne, które przyklejają się do każdej ze standardowych zasad azotowych (A, C, G, T). Taką uniwersalną zasadą mógłby być np. 5-nitroindol (LOAKES i BROWN

1994) lub 3-nitropyrrole (BERGSTROM i współaut. 1995).

Oligonukleotydy muszą być utworzone według specjalnego wzorca nazwanego GP(*s*,*r*) (ang. gapped probe). Dla ustalonych parametrów *r* i *s*, wzorzec wygląda następująco: $X^s(U^{s-1}X)^r$, gdzie X oznacza nukleotyd z jedną ze standardowych zasad (A, C, G, T), a U nukleotyd z zasadą uniwersalną. Dla przykładu, wzorzec GP(3,2) wygląda następująco: XXXUUXUUX, a oligonukleotyd pasujący do wzorca to np. GACUUCUUT. Algorytm zaproponowany przez PREPARATĘ i współaut. (1999) jest prostą heurystyką. Rozpoczyna poszukiwanie rozwiązania od znanego prefiksu sekwencji o długości $s(r+1)$. Z każdą iteracją próbuje wydłużyć sekwencję o 1 nukleotyd, poprzez dopasowanie ze spektrum elementów, których ($s(r+1)$-1)–prefiks nakłada się na ($s(r+1)$-1)–sufiks tworzonej sekwencji (uniwersalne zasady mogą się łączyć z dowolną inną). Jeżeli więcej niż jeden element pasuje jako rozwinięcie, tworzonych jest kilka równoległych ścieżek. Algorytm zakłada, że spektrum wejściowe zawiera tylko bardzo niewielki odsetek błędów eksperymentalnych.

Kolejny algorytm (HEATH i współaut. 2003) operuje na dwóch spektrach łączonych, które składają się ze zwykłych oligonukleotydów oraz oligonukleotydów odwróconych. Zwykłe oligonukleotydy są utworzone na podstawie wzorca używanego przez PREPARATĘ i współaut. (1999): $X^s(U^{s-1}X)^r$, natomiast odwrócone na podstawie wzorca $(XU^{s-1})^r X^s$. Użycie spektrów obu typów ułatwia rekonstrukcję sekwencji w przypadku błędnych elementów w spektrum.

HALPERIN i współaut. (2003) zaproponowali utworzenie mikromacierzy, która zamiast elementów tworzonych według deterministycznego wzorca będzie składała się z elementów, w których pozycje ze znanymi zasadami (A, C, G, T) będą wybierane losowo. Elementy są tworzone w następujący sposób. Długość każdego elementu jest równa $l = c \cdot k+1$, gdzie *c* jest zazwyczaj między 3 a 10, natomiast *k* jest liczbą znanych nukleotydów. Następnie generowane są zbiory A_i, gdzie i=1...β*k*, a β zależy od liczby błędów. Dla każdego zbioru A_i wybierany jest losowo zbiór *k* pozycji z {1, 2,..., *ck*} i tworzone są wszystkie możliwe 4^{k+1} oligonukleotydy ze znanymi zasadami na *k* wybranych oraz na ostatniej pozycji, a reszta zasad jest uniwersalna. Wszystkich elementów użytych w eksperymencie jest $\beta k 4^{k+1}$ i jest to suma zbiorów A_i. Zrekonstruowanie sekwencji jest możliwe z dużym prawdopodobieństwem, nawet dla spektrów z błędami negatywnymi i pozytywnymi.

Chociaż udało się już wyprodukować zasady uniwersalne w laboratorium (LOAKES i BROWN 1994, BERGSTROM i współaut. 1995), to jednak nadal są one jedynie rozważane teoretycznie i nie wiadomo, czy będą mogły skutecznie brać udział w eksperymencie hybrydyzacji.

PREPARATA i OLIVER (2004) ponownie zajęli się problemem sekwencjonowania przy użyciu zasad uniwersalnych, ale tym razem zamiast nieosiągalnych idealnych zasad uniwersalnych zastosowali zasady zdegenerowane, które są jednolitą mieszaniną czterech naturalnych zasad. Liczba zasad zdegenerowanych w oligonukleotydzie jest ograniczona, gdyż każda taka zasada pogarsza sygnał hybrydyzacji na mikromacierzy. W standardowym modelu hybrydyzacji zakłada się, że każda para zasad komplementarnych generuje sygnał hybrydyzacji o takiej samej sile. Jednakże siła sygnału zależy od zawartości zasad G/C i A/T (para G/C łączy się silniejszym wiązaniem niż para A/T, co przekłada się na siłę sygnału w trakcie eksperymentu) oraz od pozycji nukleotydu w oligonukleotydzie. Autorzy wywnioskowali, że pozycja zasady zdegenerowanej ma znaczący wpływ na siłę sygnału hybrydyzacji. Ustalona została zatem minimalna wartość energii wiązania oraz zbiór pozycji, na których może się pojawić zasada zdegenerowana. Wykazano, że rozrzut wartości energii wiążącej dupleks jest zbyt duży, co czyni zasady zdegenerowane nieuniwersalnymi. Jako praktyczną realizację zasad uniwersalnych autorzy zaproponowali dwa zbiory zasad częściowo zdegenerowanych, zbiory A/T (słabo-wiążące) oraz G/C (mocno-wiążące).

Rozważania na temat różnic w sile sygnałów hybrydyzacji, mające na celu zmniejszenie błędów hybrydyzacji, doprowadziły do zaproponowania nowego, izotermicznego podejścia do sekwencjonowania (BŁAŻEWICZ i współaut. 1999b). Podczas eksperymentu hybrydyzacji, zamiast oligonukleotydów o równej długości wykorzystywane są oligonukleotydy o równej temperaturze zajścia idealnej hybrydyzacji, co ma na celu ujednolicić warunki, w której oligonukleotydy będą hybrydyzowały do badanej sekwencji. W przybliżeniu zakłada się, że wiązanie pary G/C jest dwa razy silniejsze niż wiązanie pary A/T (WALLACE i współaut. 1981). Ten

model, chociaż przybliżony, pozwala zrekompensować niższą stabilność dupleksów (fragmentów dwuniciowego DNA) bogatych w pary A/T poprzez zwiększenie ich długości. Aby zatem wyznaczyć temperaturę dla oligonukleotydu zakłada się, że każdy nukleotyd G lub C zwiększa temperaturę oligonukleotydu o 4 stopnie, natomiast każdy nukleotyd A lub T o 2 stopnie. Wszystkie oligonukleotydy o tej samej temperaturze zajścia idealnej hybrydyzacji tworzą bibliotekę izotermiczną.

Użycie tylko jednej biblioteki izotermicznej nie jest wystarczające. Na przykład nie jest możliwe pokrycie fragmentu sekwencji DNA składającej się tylko z nukleotydów G i C za pomocą oligonukleotydów o temperaturze niepodzielnej przez 4. Biblioteka oligonukleotydów o temperaturze podzielnej przez 4 nie pokryje natomiast sekwencji, w której występuje pojedynczy nukleotyd A lub T otoczony nukleotydami G lub C. Z drugiej strony dwie biblioteki izotermiczne o temperaturach różniących się o 2 stopnie (temperatura nukleotydu A lub T) mogą pokryć każdą sekwencję DNA, co więcej, przesunięcie pomiędzy oligonukleotydami z tych bibliotek pokrywającymi sekwencję DNA będzie nie większe niż 1.

Problem sekwencjonowania przez hybrydyzację przy użyciu bibliotek izotermicznych jest problemem łatwym w przypadku, gdy w spektrum nie ma błędów. Jeśli natomiast w spektrum pojawią się błędy negatywne lub pozytywne, lub błędy obu rodzajów, wówczas problem jest problemem silnie NP-trudnym (trudnym obliczeniowo) (BŁAŻEWICZ i KASPRZAK 2006).

BŁAŻEWICZ i KASPRZAK (2006) zaproponowali algorytm (wielomianowy) dokładny dla przypadku spektrum idealnego (bez błędów). Na podstawie spektrum tworzony jest skierowany graf G, w którym po pewnych transformacjach poszukiwana jest ścieżka. Oligonukleotydy są etykietami wierzchołków, a łuki łączą wierzchołki, których etykiety są równej długości i nakładają się z przesunięciem o jedną literę (o ile nie spowoduje to błędu negatywnego). Jeśli oligonukleotyd o_i jest zawarty w o_j i dosunięty do lewej jego strony, wówczas wszystkie łuki wchodzące do o_j i wychodzące z o_i są usuwane, a dodany zostaje łuk z o_i do o_j. Z drugiej strony, jeśli o_i jest zawarty w o_j i dosunięty do prawej jego strony, wówczas wszystkie łuki wchodzące do o_i i wychodzące z o_j są usuwane, a dodany zostaje łuk z o_j do o_i. Wszystkie łuki wchodzące do pierwszego wierzchołka lub wychodzące z ostatniego wierzchołka są usuwane. Następnie łuki, które z pewnością nie zostaną wykorzystane do tworzenia ścieżki są usuwane, a niektóre łuki są tymczasowo wstawiane tak, że graf G staje się grafem liniowym. Graf G zostaje przetransformowany do swojego grafu oryginalnego H i odtąd oligonukleotydy są etykietami łuków w grafie H. Algorytm może poszukiwać ścieżki Eulera w grafie H pomijając przy tym połączenia odpowiadające tymczasowym łukom w grafie G.

BŁAŻEWICZ i FORMANOWICZ (2005) przedstawili metodę rozwiązującą problem SBH, która łączy w sobie podejście wielostopniowe z izotermicznym. Podejście wielostopniowe (KRUGLYAK 1998) omówione zostało wcześniej przy zastosowaniu bibliotek z oligonukleotydami o równej długości, jest ono jednak wrażliwe na błędy eksperymentalne. Połączenie podejścia wielostopniowego z izotermicznym ma na celu zmniejszenie liczby błędów, szczególnie negatywnych.

BŁAŻEWICZ i współaut. (2006) zaproponowali hybrydowy algorytm genetyczny. Algorytm rozpoczyna działanie od utworzenia pierwszego pokolenia osobników. Każdy osobnik to permutacja wszystkich oligonukleotydów ze spektrum, które razem po złożeniu tworzą sekwencję dłuższą niż n. Następnie z każdego osobnika wybierana jest podsekwencja o długości nie większej niż n, w której zawarta jest największa liczba elementów spektrum. Liczba tych elementów jest oceną każdego osobnika. Osobnicy wybierani są zgodnie z ich oceną jako pula rodziców. Im lepsza ocena osobnika, tym więcej razy może on zostać wybrany jako rodzic. Następnie z pary rodziców tworzony jest jeden nowy osobnik za pomocą krzyżowania. Najlepsze połączenia sąsiednich oligonukleotydów u rodziców dziedziczone są przez potomka. Wśród losowo wybranych rodziców i potomków zachodzi także mutacja, polegająca na zamianie kolejności najsłabiej nakładających się oligonukleotydów. Najlepsi osobnicy przetrwają w następnym pokoleniu. Liczba osobników w kolejnych pokoleniach jest stała. Algorytm kończy swoje działanie, jeśli przez kilka pokoleń nie powiększyła się liczba oligonukleotydów tworzących rozwiązanie i jako wynik zwracane jest najlepsze znalezione do tej pory rozwiązanie.

ASEMBLACJA

Asemblacja polega na połączeniu ze sobą w dłuższe odcinki (optymalnie w jedną spójną całość) fragmentów DNA (do 1000 nukleotydów) pochodzących z etapu sekwencjonowania (Ryc. 6). Fragmenty mogą pochodzić z obu nici helisy DNA, przy czym nie wiadomo, z której nici jest dany fragment. W efekcie nie wiadomo, czy do utworzenia rozwiązania należy dany fragment użyć czytany wprost, czy też czytany od końca i przetłumaczony na nukleotydy komplementarne (czyli jego odwrotnie komplementarny odpowiednik). We fragmentach mogą też pojawić się błędy eksperymentalne: losowe lub charakterystyczne dla metody sekwencjonowania, za pomocą której otrzymane były te fragmenty. Celem algorytmu rozwiązującego problem asemblacji jest odtworzenie badanej sekwencji, będącej często całym genomem bądź długim wycinkiem genomu, np. poprzez maksymalizację dopasowania fragmentów lub maksymalizację liczby użytych fragmentów. Podczas wyznaczania dopasowania do siebie fragmentów należy dopuścić pewien odsetek niezgodności (błędy we fragmentach) oraz rozważać dopasowanie także z odwrotnie komplementarnymi odpowiednikami wszystkich fragmentów. Ze względu na to, że badana sekwencja może nie być równo pokryta przez fragmenty, a w niektórych miejscach może nie być pokryta przez żaden fragment, często nie uda się odtworzyć jednej spójnej sekwencji, lecz tylko jej część podzieloną na krótsze odcinki. Wówczas potrzebna jest dodatkowa wiedza ekspertów oraz dodatkowy eksperyment biochemiczny, który pozwoli na rekonstrukcję całej sekwencji.

```
A C C T G C A _ C T _ C G
-------------------------
A C C T   C A C C T
    C T G C T   C T T C G
          A _ C T _ C
```

Ryc. 6. Na podstawie fragmentów wejściowych {ACCT, ACTC, CACCT, CGAAG, CTGCT} podczas asemblacji udało się odtworzyć sekwencję ACCTGCACTCG.

Podczas rekonstrukcji użyto fragmentów ze zbioru wejściowego lub fragmentów do nich odwrotnie komplementarnych (zamiast CGAAG wykorzystano CTTCG). Niektóre sekwencje zawierają błędy, dlatego sąsiadujące fragmenty nie zawsze się idealnie nakładają na siebie.

Pierwsza część algorytmów przedstawionych w tej sekcji może być teoretycznie zastosowana do rozwiązania asemblacji fragmentów pochodzących ze wszystkich metod sekwencjonowania, kolejne są dedykowane głównie do asemblacji krótkich fragmentów pochodzących z nowych metod sekwencjonowania (CHAISSON i współaut. 2004, ZERBINO i BIRNEY, 2008). Algorytm przedstawiony przez KECECIOGLU i MYERSA (1995) potrafi zasemblować (połączyć) fragmenty, w których mogą występować błędy eksperymentalne i które mogą pochodzić z obu nici DNA. Na początku algorytm dodaje do wejściowego zbioru fragmentów sekwencje do nich odwrotnie komplementarne. Następnie wszystkie pary fragmentów są porównywane, aby wyznaczyć ich przybliżone nałożenie o pewnej istotności statystycznej (wadze połączenia). Utworzony zostaje graf pełny, w którym wierzchołki odpowiadają fragmentom a łuki odpowiadają ich nałożeniom. Niektóre łuki są usuwane ze względu na niską wagę, przez co graf staje się rzadszy. Wstępna orientacja fragmentów jest wyznaczana przez algorytm heurystyczny, który znajduje w grafie drzewo rozpinające o maksymalnej wartości wag (tylko połowa wierzchołków jest brana pod uwagę — oryginalny fragment lub fragment do niego odwrotnie komplementarny). Rozwiązaniem jest ścieżka Hamiltona o największym poziomie istotności (sumarycznej wadze).

Metoda IDURYEGO i WATERMANA (1995) wykorzystuje metodę Pevznera do sekwencjonowania przez hybrydyzację z oligonukleotydami o równych długościach (PEVZNER 1989). Każdy fragment wejściowy rozdzielony zostaje na kolekcję n-l+1 oligonukleotydów o długości l, gdzie n jest długością fragmentu. Następnie na podstawie tej kolekcji tworzony jest graf Pevznera, w którym poszukiwana jest ścieżka Eulera. Algorytm działa najlepiej gdy wartość l jest duża, gdyż w ten sposób lepiej jest zachowana informacja o fragmentach wejściowych. Metoda działa jedynie dla danych bezbłędnych i pochodzących z jednej nici DNA. Autorzy podają jednak, jak przystosować tę metodę do rzeczywistych warunków. Do zbioru fragmentów wejściowych dodawane są sekwencje odwrotnie o nich komplementarne. W rezultacie algorytm zwraca dwa komplementarne do siebie rozwiązania. Ze względu na błędy w danych dopuszczane jest także nie wykorzystywanie

niektórych łuków (nadmiarowe fragmenty), lub wykorzystanie ich kilka razy (błędy negatywne pochodzące z powtórzeń). Wówczas stosuje się wariant problemu poszukiwania ścieżki Eulera w podobny sposób jak w problemie selektywnego komiwojażera.

JIANG i LI (1996) rozpatrywali problem asemblacji jako wariant problemu najkrótszego wspólnego superciągu (ang. shortest common superstring). Fragmenty muszą pochodzić z jednej nici DNA i każdy może zawierać nie więcej niż k błędów. Algorytm łączy w każdej iteracji dwa fragmenty w dłuższy ciąg. Wybierana jest taka para elementów, dla których stosunek pomiędzy długością ciągu i sumą długości fragmentów jest najmniejszy. Metoda dopuszcza co najwyżej k błędów przy połączeniu ciągów. W pierwszej fazie tworzone są krótkie ciągi z niewielkiej liczby fragmentów bardzo ściśle ze sobą połączonych. W następnej kolejności fragmenty te łączone są w dłuższe, a w ostatnim etapie otrzymywany jest jeden długi ciąg złożony ze wszystkich fragmentów.

PEVZNER i współaut. (2001) przedstawili jeszcze jedną metodę poszukującą ścieżki Eulera w grafie Pevznera. Metoda ta potrafi sobie poradzić zarówno z błędami w sekwencjach wejściowych, jak i z długimi powtórzeniami fragmentów w badanej sekwencji. W pierwszym kroku algorytm próbuje wyeliminować błędy w sekwencjach. W tym celu wejściowe fragmenty są dzielone na oligonukleotydy o długości l, gdzie l jest znacznie krótsze od długości fragmentów. Zliczana jest liczba wystąpień każdego oligonukleotydu i jeśli jest ona większa niż M (pewien ustalony próg) dla pewnego oligonukleotydu, wówczas oznaczany jest on jako „silny", w przeciwnym przypadku jako „słaby". Jeśli jest błąd na jakiejś pozycji we fragmencie, wówczas kilka kolejnych oligonukleotydów będzie słabych. Procedura naprawcza poprzez redukcję prawdopodobnych błędów zamienia słabe oligonukleotydy w mocne i zmniejsza w ten sposób liczność spektrum. Następnie z elementów spektrum utworzony zostaje graf podobny jak w podejściu Pevznera. Dla każdego fragmentu zapamiętywana jest reprezentująca go ścieżka w grafie. W końcowym etapie poszukiwana jest ścieżka Eulera w grafie, która zawiera w sobie wszystkie zapamiętane ścieżki.

CHAISSON i współaut. (2004) zaprezentowali metodę asemblacji krótkich fragmentów wejściowych (100–200 nukleotydów). Algorytm oparty jest na przedstawionym powyżej podejściu (PEVZNER i współaut. 2001). W pierwszej fazie naprawiane są błędy pochodzące z eksperymentu. Fragmenty dzielone są na oligonukleotydy o długości 15–20 i oznaczane są jako „silne", jeśli pojawią się więcej niż M razy, lub „słabe" w przeciwnym przypadku. Aby rozstrzygnąć błędy w sekwencjach zaproponowano algorytm programowania dynamicznego. Wraz ze wzrostem długości fragmentów wzrasta też niestety liczba błędów we fragmentach i przestrzeń rozwiązań staje się zbyt duża do przeszukania. Po fazie naprawy błędów z oligonukleotydów konstruowany jest graf, w którym poszukuje się ścieżki Eulera.

Metoda VELVET do asemblacji bardzo krótkich fragmentów wejściowych (ok. 35 nukleotydów) została zaproponowana w (ZERBINO i BIRNEY 2008). Działa ona efektywnie przy bardzo dużym pokryciu sekwencji zapewnianym przez technologie Solexa. Metoda przeprowadza kolejne operacje na grafie, który jest zbudowany podobnie u PEVZNERA i współaut. (2001). W pierwszej fazie algorytm eliminuje błędy w sekwencjach i łączy sekwencje, które nakładają się na siebie (idealnie). W drugim etapie aparat naprawczy wyszukuje powtórzeń, czyli fragmentów w grafie, które są współdzielone przez co najmniej dwie różne ścieżki i rozstrzyga konflikty poprzez przeszukiwanie lokalnych nałożeń sekwencji. Dodatkowo metoda może wykorzystywać informację o odległości pomiędzy parami sekwencji wejściowych, co umożliwia w kolejnym etapie sklejanie krótkich odcinków w dłuższe.

PODSUMOWANIE

Zaprezentowane metody stanowią jedynie wycinek niezmiernie bogatej literatury, która powstała dla sekwencjonowania i asemblacji DNA. Zostały one wybrane pod kątem zarówno znaczenia dla rozwoju tej gałęzi badawczej (historycznie najistotniejsze podejścia i algorytmy), jak i ich atrakcyjności (interesujące modele grafowe, efektowne rozwiązania). Znaczenie sekwencjonowania i asemblacji jako pierwszego etapu drogi do poznania i zrozumienia informacji genetycznej organizmów gwarantuje dalszy rozwój metod biochemicznych i coraz skuteczniejsze algorytmy.

DNA SEQUENCING AND ASSEMBLING — APPROACHES, GRAPH MODELS, AND ALGORITHMS

Summary

Reading genetic information of an organism, i.e. reading a sequence of nucleotides of a DNA fragment, can be done in two or three stages. In the first stage, the sequencing, one can obtain sequences up to a few hundreds of nucleotides. There are several approaches to carry out this stage. The historically oldest approach is gel electrophoresis, also called by the name of the author – the Sanger method. Another approach is sequencing by hybridization, which is technologically more sophisticated and it involves also algorithmic methods to process the experimental data (as opposed to the previous approach). The novel, fully automated approaches (owned by Roche, Illumina, Applied Biosystems) generate millions of short DNA sequences in short time. Next stage in reading a DNA sequence is the assembling: the output of the sequencing stage is assembled together into longer contigs of length up to even a few million nucleotides. The last stage, called the mapping or the finishing, consists in scheduling assembled sequences in the right order.

The methods presented in the paper are only a part of immensely rich literature, which is available for the DNA sequencing and assembling. They were chosen both from the point of view of their importance for the development of this research branch (historically most important approaches and algorithms) and for their attractiveness (interesting graph models). The meaning of the sequencing and the assembling as the first steps on the way of understanding genetic information of organisms, guarantees further development of associated biochemical and computational approaches..

LITERATURA

BAINS W., SMITH G. C., 1988. *A novel method for nucleic acid sequence determination*. J. Theoretical Biology 135, 303–307.

BENNETT S., 2004. *Solexa Ltd.* Pharmacogenomics 5, 433–438.

BERGSTROM D. E., ANDREWS P. C., NICHOLS R., ZHANG P., 1995. *3–Nitropyrrole Nucleoside.* US Patent No. 5,438,131. 08/01/95.

BŁAŻEWICZ J., KASPRZAK M., 2003. *Complexity of DNA sequencing by hybridization*. Theoretical Comput. Sci. 290, 1459–1473.

BŁAŻEWICZ J., KASPRZAK M., 2006. *Computational complexity of isothermic DNA sequencing by hybridization*. Discrete Appl. Math. 154, 718–729.

BŁAŻEWICZ J., FORMANOWICZ P., 2005. *Multistage isothermic sequencing by hybridization.* Comput. Biol. Chemistry 29, 69–77.

BŁAŻEWICZ J., FORMANOWICZ P., KASPRZAK M., MARKIEWICZ W. T., WĘGLARZ J., 1999a. *DNA sequencing with positive and negative errors*, J. Comput. Biol. 6, 113–123.

BŁAŻEWICZ J., FORMANOWICZ P., KASPRZAK M., MARKIEWICZ W. T., WĘGLARZ J., 1999b. *Method of sequencing of nucleic acids*. Polish Patent Application P335786.

BŁAŻEWICZ J., HERTZ A., KOBLER D., DE WERRA D., 1999c. *On some properties of DNA graphs* . Discrete Appl. Math. 98, 1–19.

BŁAŻEWICZ J., FORMANOWICZ P., KASPRZAK M., JAROSZEWSKI M., MARKIEWICZ W. T., 2001. *Construction of DNA restriction maps based on a simplified experiment.* Bioinformatics 17, 398–404.

BŁAŻEWICZ J., GLOVER F., KASPRZAK M., 2004. *DNA sequencing — tabu and scatter search combined*. INFORMS J. Comput. 16, 232–240.

BŁAŻEWICZ J., OĞUZ C., ŚWIERCZ A., WĘGLARZ J., 2006. *DNA sequencing by hybridization via Genetic Search*. Operations Res. 54, 1185–1192.

BUI T. N., YOUSSEF W. A, 2004. *An enhanced genetic algorithm for DNA sequencing by hybridization with positive and negative errors*. Lect. Notes Comput. Sci. 3103, 908–919.

CHAISSON M., PEVZNER P., TANG H., 2004. *Fragment assembly with short reads*. Bioinformatics 20, 2067–2074.

FU Y., PECKHAM H. E., MCLAUGHLIN S. F., RHODES M. D., MALEK J. A., MCKERNAN K. J., BLANCHARD P., 2008. *SOLID sequencing and Z-Base encoding*. [W:] *The Biology of Genomes Meeting*, Cold Spring Harbour Laboratory (http://www.appliedbiosystems.com).

HALPERIN E., HALPERIN S., HARTMAN T., SHAMIR R., 2003. *Handling long targets and errors in sequencing by hybridization*. J. Comput. Biol 10, 483–497.

HEATH S. A., PREPARATA F. P., YOUNG J., 2003. *Sequencing by hybridization by cooperating direct and reverse spectra*. J. Comput. Biol 10, 499–508.

IDURY R., WATERMAN M., 1995. *A new algorithm for DNA sequence assembly*. J. Comput. Biol 2, 291–306.

JIANG T., LI M., 1996. *DNA sequencing and string learning*. Math. Systems Theory 29, 387–405.

KECECIOGLU J. D., MYERS E. W., 1995. *Combinatorial algorithms for DNA sequence assembly*. Algorithmica 13,7–51.

KRUGLYAK S., 1998. *Multistage sequencing by hybridization.* J. Comput. Biol 5,165–171.

LOAKES D., BROWN D. M., 1994. *5-Nitroindole as an universal base analogue*. Nucl. Acids Res. 22, 4039–4043.

LYSOV Y. P., FLORENTIEV V. L., KHORLIN A. A, KHRAPKO K. R., SHIK V. V., MIRZABEKOV A. D., 1988. *Determination of the nucleotide sequence of DNA using hybridization of oligonucleotides. A new method*. Dokl. Akademii Nauk SSSR 303, 1508–1511.

MARGULIES M., EGHOLM M., ALTMAN W. E., ATTIYA S. i współaut., 2005. *Genome sequencing in microfabricated high density picolitre reactors*. Nature 437, 376–380.

MAXAM A. M., GILBERT W., 1977. *A new method for sequencing DNA*. Proc. Natl. Acad. Sci. USA 74, 560–564.

PEVZNER P. A., 1989. *l-tuple DNA sequencing: computer analysis*. J. Biomol. Struct. Dyn. 7, 63–73.

PEVZNER P., TANG H., WATERMAN M. S., 2001. *A new approach to fragment assembly in DNA sequencing*. Proc. 5th Ann. Inter. Conf. Res. Comput. Molecular Biology (RECOMB), ACM Press, Montreal , 256–267.

PHAN V. T., SKIENA S., 2001. *Dealing with errors in interactive sequencing by hybridization*. Bioinformatics 17, 862–870.

PREPARATA F. P., UPFAL E., 2001. *System and methods for sequencing by hybridization.* United States Patent Application US 2001/0004728. 21/07/2001.

PREPARATA F. P., OLIVER J. S., 2004. *DNA sequencing by hybridization using semi-degenerate bases.* J. Comput. Biol. 11, 753–765.

PREPARATA F. P., FRIEZE A. M., UPFAL E.,1999. *Optimal reconstruction of a sequence from its probes.* J. Comput. Biol. 7, 361–368.

SANGER F., NICKELEN S., COULSON A. R., 1977. *DNA sequencing with chain-terminating inhibitors.* Proc. Natl. Acad. Sci. USA 74, 5463–5467.

SOUTHERN E. M., 1988. *United Kingdom Patent Application GB8810400.*

SOUTHERN E. M, MASKOS U., ELDER J. K., 1992. *Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models.* Genomics 13, 1008–1017.

WALLACE R. B., JOHNSON M. J., HIROSE T., MIYAKE T., KAWASHIMA E. H., ITAKURA K., 1981. *The use of synthetic oligonucleotides as hybridization probes. II. Hybridization of oligonucleotides of mixed sequence to rabbit beta-globin DNA.* Nucleic Acids Res. 9, 879–894.

ZERBINO D. R., BIRNEY E., 2008. *Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.* Genome Res. 18, 821–829.

ZHANG J.-H., WU L.-Y., ZHANG X.-S., 2003. *Reconstruction of DNA sequencing by hybridization.* Bioinformatics 19, 14–21.