RZECZPOSPOLITA POLSKA

Urząd Patentowy Rzeczypospolitej Polskiej

(12) **OPIS PATENTOWY** (19) **PL** (11) **216761**

(13) **B1**

(21) Numer zgłoszenia: **381053**

(22) Data zgłoszenia: **15.11.2006**

(51) Int.Cl.
C12Q 1/68 (2006.01)
G06N 3/00 (2006.01)
G06N 3/12 (2006.01)

(54) **Sposób identyfikacji cząsteczek DNA opisanych wyrażeniem regularnym**

(43) Zgłoszenie ogłoszono:
26.05.2008 BUP 11/08

(45) O udzieleniu patentu ogłoszono:
30.05.2014 WUP 05/14

(73) Uprawniony z patentu:
POLITECHNIKA WARSZAWSKA, Warszawa, PL

(72) Twórca(y) wynalazku:
ROBERT NOWAK, Chełm, PL
ANDRZEJ PŁUCIENNICZAK, Warszawa, PL

(74) Pełnomocnik:
rzecz. pat. Joanna Bocheńska

PL 216761 B1

Opis wynalazku

Przedmiotem wynalazku jest sposób identyfikacji cząsteczek DNA opisanych wyrażeniem regularnym, mogący znaleźć zastosowanie do identyfikacji chorób genetycznych lub przy prowadzeniu eksperymentów biologicznych, do wykrywania pewnych innych zależności genetycznych u organizmów.

Aby identyfikować cząsteczki DNA można użyć procesu sekwencjonowania, ujawnionego w amerykańskim opisie patentowym US5683881. Proces ten pozwala odczytać sekwencję nici DNA. Znając sekwencję danej nici DNA można, przy pomocy komputerów elektronicznych oraz algorytmów wyszukiwania badać, czy należy ona do szukanej podklasy. Metoda taka nie jest optymalna, ponieważ proces sekwencjonowania jest bardzo kosztowny i czasochłonny.

Istnieje kilka prac dotyczących przetwarzania informacji przez molekuły, które wykorzystują inne metody. Jedną z pierwszych prac, która dostrzegła, możliwości zastosowań molekuł do przetwarzania informacji, jest publikacja WO 9707440 oraz L. M. Adleman, *Molecular computation of solutions to combinatorial problems*, Science, 266: 1021-1024, November 11, 1994. Pokazano tam metodę wykorzystującą wielką liczbę cząsteczek do rozwiązywania problemów kombinatorycznych, należących do klasy NP. W opisanym przez Adlemana eksperymencie pokazano możliwość rozwiązania, problemu ścieżki Hamiltona. (zwanego też problemem komiwojażera) przy pomocy cząsteczek DNA. W pracy tej nie porusza się zagadnień związanych z identyfikacją cząsteczek DNA.

Kilkukrotnie były podejmowane próby konstrukcji automatów skończonych za pomocą metod biologii molekularnej: publikacja R. Deaton, R. C. Murphy, J. A. Rose, M.H. Garzon, D. R. Franceschetti, and S. E. Stevens, Jr. *A DNA based implementation of an evolutionary search for good encodings for DNA computation*, JEEE International Conference on Evolutionary Computations, pages 267-271, 1997. Indiana University Purdue Universit'y, Indianapolis, Illinois, USA, publikacja WO 0077732, WO 03042395, publikacja US2005112614.

W pracy *A DNA based implementation of an evolutionary search for good encodings for DNA computation* pokazano metodę przeprowadzenia doświadczenia na DNA, aby symulować automat skończony. Dla każdego przejścia projektuje się odpowiednią cząsteczkę DNA. Rozpoznawanie, czy słowo wejściowe należy do danego języka regularnego polega na cyklicznym odczytywaniu kolejnych symboli ze słowa wejściowego przez osobę prowadzącą doświadczenie, dodawaniu cząsteczek reprezentujących przejścia dla tego symbolu do roztworu i wykonywaniu reakcji, w wyniku których w roztworze pojawiają się cząsteczki reprezentujące stany automatu. Powyższa koncepcja z trudem mogłaby znaleźć zastosowanie praktyczne, ponieważ nie pozwala umieścić wielu niezależnych automatów w tym samym miejscu, a także nie pozwala obrabiać wielu różnych słów wejściowych. Opisane rozwiązanie nie jest potwierdzone eksperymentem.

Sposób opisany w publikacji WO 0077732 (opublikowany również w K. Komiya, K. Sakamoto, H. Gouzo, S. Yokoyama, M. Anta, A. Nishikawa, aud M. Hagiya, *Successive state transitions with i/o interface by molceules*, volume 2054 of Lecture Notes in Computer Science, pages 17-26. Springer Verlag, Berlin, Heidelberg. New York, 2000), bazuje na metodzie self-assembly przedstawionej w US20055112614 oraz E. Winfree, F. Lin, L. A. Wenzler, and N. C. Seeman, *Design end self-assembly of two-dimensional DNA crystals*, Nature, 394(6693): 539-545, August 6 1998. Automat skończony oraz słowo wejściowe są zakodowane na tej samej, pojedynczej nici DNA, co pozwala na jednoczesną pracę bardzo wielu takich automatów. Przedstawiony sposób jest potwierdzony doświadczeniem i może mieć zastosowania praktyczne. Wadą tego rozwiązania jest brak gwarancji, że krok obliczeń będzie przejściem w automacie, co nie pozwala (w sposób deterministyczny) ustalić w jakiej liczbie kroków uzyska się odpowiedź, w szczególności liczba kroków może być bardzo duża.

Interesująca metoda identyfikacji cząsteczek opisana w WO 030423395 oraz Y. Benenson, T. Paz-Elizur, R. Adar, E. Keinan, Z. Livneh, and E. Shapiro, *Programmable and autonomous computing machinie made of biomolecules*, Nature, 414(1): 430-434, November 22, 2001, bazuje na pewnych szczególnych właściwościach niektórych enzymów restrykcyjnych, takich jak na FokI, które tną cząsteczkę w pewnej odległości od rozpoznawanej sekwencji. Poprawność metody jest potwierdzona doświadczeniem. Główną wadą opisanego tam rozwiązania jest ograniczenie na maksymalną liczbę stanów i przejść. Do ich kodowania można wykorzystać maksymalnie 4 nukleotydy, więc maksymalna liczba stanów i przejść wynosi 256, co eliminuje sensowność wykorzystania opisanego sposobu dla praktycznie występujących przypadków.

Sposób identyfikacji cząsteczek DNA opisanych wyrażeniem regularnym według wynalazku pozwala zbadać, czy nieznana cząsteczka posiada sekwencję, która jest opisana wyrażeniem regular-

nym. Bazuje ono na operacjach znanych z inżynierii genetycznej i nie wymaga poznawania sekwencji cząsteczki. Opisywane rozwiązanie jest nowym sposobem przetwarzania informacji za pomocą molekuł. Jest to praktyczna realizacja abstrakcyjnego urządzenia nazywanego automatem skończonym.

Cząsteczkę DNA można potraktować jako sekwencje symboli czyli słowo nad pewnym językiem. Wyrażenia regularne są pewną formą opisu zbioru słów (języka). Języki opisywane wyrażeniami regularnymi są nazywane językami regularnymi. Są to języki nieskończone, tzn. mogą zawierać nieskończoną liczbę słów.

Definicje słowa, języka, języka regularnego są zawarte w pozycji Hopcroft and Ullman, Wprowadzenie do teorii automatów, języków i obliczeń, PWN, 1994. Słowo spełnia wyrażenie regularne, jeżeli należy do języka opisanego przez to wyrażenie.

Przykłady zebrano w tabeli 1

Wyrażenie	Słowa spełniające to wyrażenie
a*	ε, a, aa, aaa, ...
(ab)*b	b, abb, ababb,
a(a|b)*b	aa, aab, aabb, ...„ ab, abb, abbb,
(a|b)(a|b)	aa, ab, ba, bb

T a b e l a 1: Przykłady wyrażeń regularnych, alfabet (zbiór symboli) jest dwuliterowy {a, b}.

Identyfikacja cząsteczek DNA, których sekwencja jest opisana wyrażeniem regularnym ma znaczenie przy prowadzeniu eksperymentów biologicznych, oraz przy wykrywaniu pewnych zależności genetycznych u organizmów, co może być przydatne np. do wykrywania chorób genetycznych. Załóżmy, że interesująca nas choroba genetyczna występuje, gdy w określonym miejscu występuje jedna z sekwencji DNA:

ATGGGGCCCC
ATATATGGGGCCCC
ATATATATATGGGGCCCC
ATATATATATATATGGGGCCCC
ATATATATATATATATATGGGGCCCC
…

T a b e l a 2: Sekwencje opisujące pewną chorobę genetyczną.

Sekwencje wymienione w tab. można krócej zapisać za pomocą wyrażenia regularnego:

AT(ATAT) * GGGGCCCC

albo:

a(aa) * b, gdzie { a = AT ; b = GGGGCCCC

Aby zbadać, czy nieznana sekwencja jest jedną z wymienionych w tabeli można:

1. Odczytać jego sekwencję.

2. Porównać tę sekwencję z wzorcami. Tworzy się cząsteczki, które odpowiadają każdej z sekwencji wymienionych w tabeli, a następnie (za pomocą reakcji biochemicznych) bada się identyczność sekwencji z wzorcami.

3. Zastosować sposób według wynalazku.

Metoda pierwsza (odczyt sekwencji nieznanego DNA) jest droga, czasochłonna, ponieważ wymaga wykorzystania sekwencjonowania. Druga z metod nie zawsze jest poprawna, ponieważ liczba

interesujących badacza sekwencji jest nieskończona, zaś można wyprodukować tylko skończoną liczbę wzorców. Natomiast samo badanie jest szybkie i tanie, można tutaj zastosować np. DNA chip.

Ponieważ prezentowane urządzenie wykorzystuje cząsteczki DNA, w opisie wynalazku pojawiają się terminy związane z biologią molekularną. Cząsteczka DNA składa się z dwóch oplatających się nici złożonych z nukleotydów. W naturze występują 4 różne nukleotydy DNA, które w skrócie oznacza się literami A, G, C, T. Nukleotyd A z jednej nici łączy się z nukleotydem T w drugiej nici, zaś G z C (zasada komplementarności). Nici DNA nie są symetryczne, jeden z końców oznacza się 5' zaś drugi 3'. Zwyczajowo sekwencje DNA pisze się od 5' końca. Hybrydyzacja jest to reakcja polegająca na połączeniu dwóch nici DNA o sekwencjach komplementarnych w jedną cząsteczkę DNA. Denaturacja jest procesem odwrotnym. Polimeryzacja jest procesem budowy nici komplementarnej do danej. Proces ten, oprócz nici bazowej wymaga dostarczenia nukleotydów oraz odcinka DNA, od którego zacznie się budowa nowej nici zwanego starterem. PCR jest procesem wykładniczego powielania cząsteczek DNA.

Symbol jest reprezentowany w prezentowanym rozwiązaniu przez sekwencję nukleotydów, na przykład: ATCCCA, GGTCCT. Symbole w niniejszym opisie oznaczane będą identyfikatorami będącymi małymi literami alfabetu lub cyframi. Sekwencja nukleotydów komplementarna do sekwencji dla danego symbolu oznaczana jest tutaj identyfikatorem ze znakiem negacji, więc jeśli symbol *a* jest reprezentowany przez sekwencję ATCCCA to a odpowiada sekwencji 3' - TAGGGT - 5', czyli TGGGAT.

Terminem alfabet Σ oznacza się zbiór symboli. Termin słowo określa skończony ciąg symboli z danego alfabetu. Tutaj słowo jest reprezentowane przez jednoniciową cząsteczkę DNA zawierającą sekwencje reprezentujące kolejne symbole. Przykładowo dla alfabetu Σ = {a, b}, gdzie symbol *a* jest reprezentowany przez ATCCCA, zaś *b* przez GGTCCT, słowo R = *abb* jest opisywanie przez nić o sekwencji ATCCCA GGTCCT GGTCCT. Długość słowa R, oznaczana |R| jest liczbą symboli w tym słowie, na przykład słowo *aab* ma długość 3. Słowo ε ma długość O. Słowa będą oznaczane wielkimi literami.

Terminem język określa się zbiór słów dla danego alfabetu. Wyrażenia regularne są formą opisu niektórych języków tzw. języków regularnych. Mówimy, że słowo spełnia wyrażenie regularne, jeżeli należy do języka opisanego przez to wyrażenie.

Dla każdego wyrażenia regularnego można skonstruować równoważną, to znaczy opisującą ten sam język, gramatykę prawostronnie liniową. Gramatyka prawostronnie liniowa jest czwórką uporządkowaną G = (V, T, P, q_0), gdzie V oznacza zbiór symboli nieterminalnych, T oznacza zbiór symboli terminalnych, P jest zbiorem produkcji, zaś $q_0 \in V$ jest stanem początkowym. Zbiory T oraz V są rozłączne. Produkcje $p \in P$ są typu $1 \rightarrow a2$ lub $1 \rightarrow a$, gdzie $1, 2 \in V$ oraz $a \in T$. W opisie wynalazku symbole nieterminalne oznaczane są cyframi, zaś symbole terminalne małymi literami.

Automat skończony dla danego wyrażenia regularnego jest urządzeniem, które w skończonej liczbie przejść potrafi odpowiedzieć, czy dane słowo spełnia wyrażenie regularne. Dla każdego wyrażenia regularnego można zbudować automat skończony rozpoznający to wyrażenie. Jeżeli wyrażenie regularne R ma długość |R|, to można, skonstruować automat niedeterministyczny o liczbie stanów *O*(|R|). Analiza, czy dane słowo X spełnia wyrażenie regularne przy użyciu tego podejścia wykonuje się w czasie *O**(|R|* |X|) gdzie |X| jest długością słowa. Możliwa jest także konstrukcja automatu deterministycznego; liczba jego stanów jest wykładniczo zależna od długości wyrażenia, regularnego $O(2^{|R|})$, zaś analiza, czy dane słowo X spełnia wyrażenie regularne przy użyciu tego podejścia wykonuje się w czasie *O*(|X|).

Sposób identyfikacji cząsteczek DNA opisanych wyrażeniem regularnym polega na tym, że definiuje się wyrażenie regularne, reprezentacje symboli, reprezentacje słów wejściowych, oraz sekwencję cząsteczki DNA reprezentującej słowo wyjściowe będące odpowiedzią pozytywną. Z badanych próbek materiału genetycznego izoluje się nici DNA w znany sposób. Do pojedynczej badanej nici DNA na końcu 3' dołącza się sekwencję nukleotydów nazywaną dalej end, a na końcu 5' sekwencję nukleotydów nazywaną dalej start. Tak przygotowana nić reprezentuje słowo wejściowe. Na podstawie określonego wcześniej wyrażenia regularnego tworzy się gramatykę prawostronnie liniową, a następnie tworzy się reguły redukcji odpowiadające tej gramatyce. Kolejnym krokiem jest synteza znanymi metodami inżynierii genetycznej zbioru cząsteczek zwanych dalej motorami produkcji, których sekwencje odpowiadają odpowiednim regułom redukcji: posiadają na 5' końcu sekwencję komplementarną do sekwencji end oznaczaną dalej $\overline{\text{end}}$ następnie sekwencję komplementarną do sekwencji reprezentującej lewą stronę redukcji, następnie sekwencję reprezentującą prawą stronę redukcji, zaś na

końcu 3' sekwencję end. Następnie badane nici DNA umieszcza się razem z motorami produkcji i wykonuje się wstępną denaturację, a następnie prowadzi się proces hybrydyzacji cząsteczek reprezentujących słowa wejściowe z motorami produkcji. Po zakończeniu hybrydyzacji w roztworze umieszcza się startery o sekwencji end, następnie prowadzi się reakcję polimeryzacji z dodatkiem polimerazy o dużym prawdopodobieństwie skoku, przy czym w roztworze znajdują się w nadmiarze wolne nukleotydy A, T, G i C. Następnie prowadzi się proces denaturacji w znany sposób po czym dodaje się startery o sekwencji start, a następnie ponownie prowadzi się proces polimeryzacji w celu uzyskania nici reprezentującej słowo wyjściowe. Jeżeli uzyskano nową cząsteczkę, to bada się, czy ma ona sekwencję cząsteczki reprezentującej odpowiedź pozytywną. Jeżeli tak, to badana cząsteczka DNA jest opisana wyrażeniem regularnym. Jeżeli nie, to cały cykl od momentu wstępnej denaturacji powtarza się, aż do uzyskania określonej cząsteczki DNA lub do momentu, gdy nie powstanie nowa cząsteczka wyjściowa. W tym wypadku badana cząsteczka DNA nie jest opisana danym wyrażeniem regularnym. Jeżeli znana jest liczba symboli *k*, które zawiera cząsteczka reprezentująca słowo wejściowe, to badanie można uprościć prowadząc cykle *k* razy, nie wykonując pośredniego badania powstawania kolejnych słów wyjściowych, zaś sprawdzanie czy uzyskano cząsteczkę reprezentującą odpowiedź pozytywną wykonuje się jednorazowo po tych cyklach.

Odmianą sposobu według wynalazku jest zmodyfikowany sposób identyfikacji cząsteczek DNA opisanych wyrażeniem regularnym, który zawiera dodatkowy krok wykładniczego powielania cząsteczek DNA.

Sposób identyfikacji cząsteczek DNA opisanych wyrażeniem regularnym polega na tym, że definiuje się wyrażenie regularne, reprezentacje symboli, reprezentacje słów wejściowych, oraz sekwencję cząsteczki DNA reprezentującej słowo wyjściowe będące odpowiedzią pozytywną. Z badanych próbek materiału genetycznego izoluje się nici DNA w znany sposób. Do pojedynczej badanej nici DNA na końcu 3’ dołącza się sekwencję nukleotydów nazywaną dalej end, a na końcu 5’ sekwencję nukleotydów nazywaną dalej start. Tak przygotowana nić reprezentuje słowo wejściowe. Na podstawie określonego wcześniej wyrażenia regularnego tworzy się gramatykę prawostronnie liniową, a następnie tworzy się reguły redukcji odpowiadające tej gramatyce. Kolejnym krokiem jest synteza, znanymi metodami inżynierii genetycznej, zbioru cząsteczek zwanych dalej motorami produkcji, których sekwencje odpowiadają odpowiednim regułom redukcji: posiadają na 5’ końcu sekwencję komplementarną do sekwencji end oznaczaną dalej $\overline{\text{end}}$, następnie sekwencję komplementarną do sekwencji reprezentującej lewą stronę redukcji, następnie sekwencję reprezentującą prawą stronę redukcji, sekwencję end, sekwencję r rozpoznawaną przez wybrany enzym restrykcyjny oraz sekwencję pomocniczą p. Następnie badane, nici DNA umieszcza się razem z motorami produkcji i wykonuje się wstępną denaturację, a następnie prowadzi się proces hybrydyzacji cząsteczek reprezentujących słowa wejściowe z motorami produkcji. Po zakończeniu hybrydyzacji w roztworze umieszcza się w nadmiarze startery o sekwencji $\overline{\text{prend}}$, następnie prowadzi się reakcję polimeryzacji z dodatkiem polimerazy o dużym prawdopodobieństwie skoku, przy czym w roztworze znajdują się w nadmiarze wolne nukleotydy A, T, G i C.

Następnie dodaje się w nadmiarze startery o sekwencji start i przeprowadza się procesem wykładniczego powielania cząsteczek DNA (PCR), który powiela łańcuch zawierający sekwencję reprezentującą słowo wejściowe. Następnie przeprowadza się reakcję cięcia wybranym enzymem restrykcyjnym, który przecina te cząsteczki w miejscu r. Uzyskuje się cząsteczki reprezentujące słowo wyjściowe. Jeżeli uzyskano nową cząsteczkę, to bada się, czy ma ona sekwencję cząsteczki reprezentującej odpowiedź pozytywną. Jeżeli tak, to badania cząsteczka DNA jest opisana wyrażeniem regularnym. Jeżeli nie, to cały cykl od momentu wstępnej denaturacji powtarza się, aż do uzyskania określonej cząsteczki DNA lub do momentu, gdy nie powstanie nowa cząsteczka wyjściowa. W tym wypadku badana cząsteczka DNA nie jest opisana danym wyrażeniem regularnym. Jeżeli znana jest liczba symboli *k*, które zawiera cząsteczka reprezentująca słowo wejściowe, to badanie można uprościć prowadząc cykle *k* razy, nie wykonując pośredniego badania powstawania kolejnych słów wyjściowych, zaś sprawdzanie czy uzyskano cząsteczkę reprezentującą odpowiedź pozytywną wykonuje się jednorazowo po tych cyklach.

Znajomość sekwencji badanego DNA nie jest w sposobie według wynalazku konieczna. Istotna jest natomiast odpowiedź na pytanie: czy badane DNA posiada sekwencję, która jest opisana wyrażeniem regularnym, np. czy badane DNA jest nośnikiem określonej choroby genetycznej? Sposób według wynalazku nie posiada wad metod znanych ze stanu techniki, które zostały wymienione wyżej. Proponowana metoda jest tania i nie wymaga sekwencjonowania, gwarantuje zawsze uzyskanie pra-

widłowej odpowiedzi, zaś liczba kroków jest liniowo zależna od długości badanej sekwencji. Opisywane rozwiązanie można traktować również jako nowy sposób przetwarzania informacji za pomocą molekuł. Jest to praktyczna realizacja abstrakcyjnego urządzenia nazywanego automatem skończonym.

Przedmiot wynalazku został przedstawiony na rysunku, na którym Fig. 1 przedstawia automat skończony rozpoznający słowa należące do języka: a(a|b) * b oraz reguły produkcji dla gramatyki prawostronnie liniowej odpowiadające temu językowi, Fig. 2 przedstawia produkcję molekularną A → B; gdzie X, A, B, C są dowolnymi słowami oraz |A| > 0, Fig. 3 Figura 3: Automat molekularny, który wykonuje k kroków, gdzie k jest długością słowa wejściowego, wykorzystywany, gdy słowa wejściowe mają taką samą długość a każdy z kroków jest procesem molekularnej produkcji, Fig. 4: Produkcja molekularna ze skokiem. Przykład dla produkcji *ab → c*; a) motor produkcji molekularnej; b) słowo wejściowe *xxab;* c) hybrydyzacja motora produkcji do słowa wejściowego; d) początek polimeryzacji, starterem jest sekwencja **end**; e) skok polimerazy; f) koniec polimeryzacji, utworzony łańcuch pośredni; g) h) i) budowa łańcucha reprezentującego słowo wyjściowe, j) cząsteczka reprezentująca słowo wyjściowe, Fig. 5: Struktura cząsteczek po hybrydyzacji. Fig. 6: Zdjęcie żelu, łańcuchy uzyskane po przeskoku polimerazy DNA. Kieszonki 1 oraz 2: reakcja trwająca 1 min; kieszonki 3 oraz 6: wzorzec (1444, 736, 587, 476, 458, 434, 298, 267, 257, 174, 102, 80, 30); kieszonki 4 oraz 5; reakcja trwająca 20 min. Fig. 7: Zdjęcie żelu pokazujące łańcuch cięty restryktazami. Kieszonka 1: *HinfI*] Kieszonka 2 oraz 6: wzorzec (1444, 736, 587, 476, 458, 434, 298, 267, 257, 174, 102, 80, 30); kieszonka 3: HpaII; kieszonka 4: RsaI; kieszonka 5: fragment nietrawiony. Fig. 8, Produkcja molekularna z PCR. Przykład dla produkcji *ab → c*; a) motor produkcji molekularnej; b) słowo wejściowe *xxab;* c) hybrydyzacja motora, produkcji do słowa wejściowego; d) polimeryzacja ze skokiem polimerazy, sekwencja $\overline{\mathbf{prend}}$ jest starterem; e) PCR ze starterami start oraz $\overline{\text{prend}}$ f) cięcie enzymem restrykcyjnymi w miejscu **r**; j) łańcuch kodujący słowo wyjściowe *xxc,* Fig. 9: Automat molekularny (motory produkcji) dla języka a(a|b)* b. Fig. 10: Reprezentacja słowa *abb* za pomocą łańcucha DNA, Fig. 11: Badanie przynależności słowa aab (do języka a(a|b) * b). Krok 1 (aktywna produkcja molekularna b → 1, Fig. 12: Badanie słowa *aab* (dla języka a(a|b) * b) Krok 2, Fig. 13: Badanie słowa *aab* dla języka a(alb) * b. Krok 3, dwie produkcje molekularne są aktywne Fig. 14: Rozpoznawanie słów opisanych wyrażeniom a(a|b) * b dla słowa wejściowego *bbb*, Fig. 15: Automat molekularny zmodyfikowany. Wykonuje się dopóki można zastosować jakąś produkcję a Fig. 16: Automat molekularny. Separacja polega na stosowaniu PCR wraz z rozcieńczaniem. Działa dla słów o ustalonej długości równej *k*.

Prezentowane idee zostały potwierdzone eksperymentami, przeprowadzonymi w laboratorium inżynierii genetycznej.

P r z y k ł a d 1. Przykład 1 Ilustruje sposób według pierwszej odmiany wynalazku. Przykładowy automat skończony pokazany na figurze 1 rozpoznaje słowa należące do języka a(a|b) * b. Gramatyka **G (N, T, P**, q_0), gdzie **N** = {0, 1}, **T** = {a, b}, q_0 = 0, zaś zbiór reguł produkcji P pokazano na Fig. 1 generuje słowa dla tego języka.

Produkcja molekularna

Produkcja molekularna jest nową koncepcją procesu operującego na cząsteczkach DNA. Umożliwia ona między innymi implementację automatów skończonych. Produkcja molekularna bada, czy wystąpiła określona sekwencja nukleotydów na końcu łańcucha i jeśli tak, to tworzy nowy łańcuch o początkowych nukleotydach identycznych z wejściowymi, zaś końcowe zamienia się na inne, z góry określone.

Jeżeli łańcuch DNA przedstawimy jako ciąg symboli, to produkcja molekularna A → B utworzy słowo XB, jeżeli słowem wejściowym będzie XB (A, B, X oznaczają słowa, przy czym A nie może być słowem pustym ε). Działanie takiej produkcji zostało pokazane na Fig. 2. Należy dodać, że słowo XA będzie obecne na wyjściu procesu, ponieważ produkcja molekularna nie separuje słów wejściowych od wyjściowych. Jeżeli słowo wejściowe nie zawiera odpowiedniej sekwencji symboli, to produkcja molekularna nie utworzy słowa wyjściowego (Fig. 2 pokazuje słowo XC, które nie jest zmieniane przez produkcję molekularną A → B).

Automat molekularny bazujący na produkcjach molekularnych

Prezentowane urządzenie, nazywane molekularnym automatem skończonym jest nową realizacją automatu skończonego na cząsteczkach DNA. Wykorzystuje ono omówioną wcześniej produkcję molekularną do realizacji zmiany stanu. Umożliwia ono badanie, czy cząsteczka DNA o nieznanej sekwencji reprezentuje słowo wejściowe, opisane wyrażeniem regularnym.

Badanie przynależności słowa do języka poprzez redukcje

Do konstrukcji automatu molekularnego wykorzystuje się następujące twierdzenie.

Twierdzenie; Słowo S = $w_1 w_2 ... w_n$ jest wygenerowane przez gramatykę prawostronnie liniową o symbolu początkowym q_0 i produkcjach o postaci A → *w*B lub A → w, jeżeli to słowo można zredukować do symbolu początkowego gramatyki q_0 za pomocą reguł redukcji o postaci *w*B → A lub *w* → A. Reguły redukcji są odwróconymi regułami produkcji rozpatrywanej gramatyki, stosuje się je zawsze do ostatnich symboli w słowie.

Dowód: Podczas generowania słowa przy użyciu gramatyki prawostronnie liniowej jest obecny co najwyżej jeden symbol nieterminalny. Znajduje się on zawsze na, końcu słowa. Na początku generowania jest to symbol startowy gramatyki, zaś w każdym kolejnym kroku używa się produkcji o postaci A → *w*B (co zachowuje prawdziwość tego warunku) lub A → w (co także zachowuje prawdziwość warunku, a przy tym kończy proces generacji). Generowanie słowa wygląda następująco: $q_0 \rightarrow w_1A_1 \rightarrow w_1w_2A_2 \rightarrow \ldots \rightarrow w_1w_2 ... w_{n-1}A_{n-1} \rightarrow w_1w_2 ... w_n$.

Słowo S = $w_1w_2 ... w_n$ może być bezpośrednio wygenerowane jedynie ze słów o postaci $w_1w_2, \ldots w_{n-1} A_{n-1}$ przy czym A_{n-1} jest takim symbolem nieterminalnym, że w zbiorze produkcji istnieje produkcja $A_{n-1}w_n$. Zbiór takich słów można uzyskać stosując dla słowa S = $w_1w_2 ...$ w reguły redukcji $w_n \rightarrow A_{n-1}$.

Słowo $w_1w_2 ... w_{n-2}w_{n-1}A_{n-1}$ jest generowane ze słów typu $w_1w_2 ... w_{n-2} A_{n-2}$ przy czym A_{n-2} jest takim symbolem nieterminalnym, że w zbiorze produkcji istnieje produkcja $A_{n-1} \rightarrow w_{n-1}A_{n-1}$. Jest to prawdą, ponieważ istnieje tylko jeden symbol nieterminalny w słowie i jest to symbol ostatni, a tylko takie reguły zapewniają prawdziwość tego postulatu.

Automat molekularny bazujący na redukcjach

Korzystając z powyższego twierdzenia tworzy się odpowiednie produkcje molekularne (które implementują przedstawione tutaj reguły redukcji), a następnie próbuje się uzyskać symbol początkowy gramatyki. Dla każdej reguły produkcji gramatyki prawostronnie liniowej opisującej dany język buduje się produkcję molekularną. Należy stosować następujące reguły:

- dla produkcji regularnych typu 1 → a2, należy zbudować produkcję molekularną a2 → 1;
- dla produkcji regularnych typu 1 → a, należy zbudować produkcję molekularną a → 1.

Automat molekularny wykorzystuje kolejne redukcje do badania, czy słowo wejściowe należy do danego języka. Produkcja molekularna jest użyta do realizacji redukcji. Na podstawie słowa wejściowego produkcje molekularne tworzą łańcuchy pośrednie. Krok obliczeń jest procesem molekularnej produkcji, w którym jest zamieniany ostatni symbol ze słowa wejściowego. Liczba kroków jest równa długości słowa wejściowego. Odpowiedź pozytywną uzyskuje się, gdy w roztworze pojawi się cząsteczka reprezentująca symbol początkowy gramatyki. Kroki obliczeń (produkcje molekularne) są rozdzielone. Polega to na odseparowaniu cząsteczek wejściowych od wyjściowych, na przykład przez umieszczenie ich w oddzielnych probówkach. Automat molekularny przedstawiono schematycznie na Fig. 3.

Przykład działania automatu molekularnego

Dla języka opisanego wyrażeniem regularnym a(a|b)*b, dla którego automat; niedeterministyczny oraz zbiór produkcji gramatyki prawostronnie liniowej został pokazany na fig. 1 redukcje są następujące:

- b → 1,
- b1 → 1,
- a1 → 1,
- a1 → 0.

Analiza, czy słowo wejściowe *abb* należy do tego języka wymaga 3 kroków (|*abb*| = 3). Redukcje są następujące: *abb* → *ab1* → *a1* → {0,1}. Należy zauważyć, że *a1* może zostać zredukowane do 1 (przez redukcję *a1* → 1) lub do 0 (przez redukcję *a1* → 0), dlatego po trzecim kroku algorytmu mamy różne słowa w roztworze. Ponieważ symbol początkowy gramatyki jest obecny, automat akceptuje słowo *abb*. Słowo *abb* należy do języka a(a|b)*b.

Analiza dla rozpoznawanego automatu i słowa, wejściowego *bbb* przebiega, następująco; *bbb* → *bb1* → *b1* → *1*. Tutaj symbol początkowy gramatyki nie jest obecny w roztworze, zatem słowo nie należy do badanego języka.

Realizacja produkcji molekularnej

Jedną z możliwych realizacji produkcji molekularnej jest proces nazywany produkcją molekularną ze skokiem, opisany alg. 1 i pokazany na Fig. 4. Jeżeli słowem wejściowym jest Xab (gdzie X jest dowolnym słowem, a, b, c, x są symbolami), to po zastosowaniu produkcji molekularnej *ab* → c otrzyma się słowo X*c*.

1. **Umieszczenie w roztworze zawierającym motory produkcji cząsteczek kodujących słowo wejściowe; denaturacja;**
2. **hybrydyzacja;**
3. **polimeryzacja z przeskokiem;**
4. **produkcja cząsteczki wyjściowej.**

Algorytm. 1: Produkcja molekularna ze skokiem.

Przetwarzane słowa są reprezentowane przez łańcuchy DNA, które posiadają ustaloną sekwencję nukleotydów oznaczaną start na końcu 5', oraz na końcu 3' sekwencję oznaczaną **end**. Pomiędzy nimi sekwencje nukleotydów odpowiadają kolejnym symbolom w słowie, Fig. 4b pokazuje łańcuch reprezentujący słowo *xxb*, zaś Fig. 4j - słowo *xxc.*

Do budowy produkcji molekularnej ze skokiem użyto jednoniciowej cząsteczki DNA, nazywanej dalej **motorem produkcji**. Motor produkcji posiada na 5' końcu sekwencję $\overline{\mathbf{end}}$, następnie sekwencję opisującą produkcję, zaś na 3' końcu sekwencję **end**. Sekwencja opisująca produkcję jest następująca: na 5' końcu występuje sekwencja komplementarna do sekwencji reprezentującej lewą stronę produkcji, następnie zaś sekwencja reprezentująca prawą stronę produkcji. Dla produkcji molekularnej *ab* → c budowa tej cząsteczki pokazana jest na Fig. 4a.

Produkcja molekularna składa się z kroków, opisanych przez alg. 1. Po umieszczeniu w roztworze motorów produkcji cząsteczek reprezentujących słowa wejściowe i wstępnej denaturacji, która likwiduje przypadkowe połączenia pomiędzy cząsteczkami, przeprowadzana jest hybrydyzacja. Temperatura roztworu zostaje ustalona w taki sposób, aby do łańcuchów reprezentujących słowa wejściowe dołączały się odpowiednie motory produkcji (tutaj motor produkcji *ab* → c czyli cząsteczka z Fig. 4a).

Jeżeli roztwór zawiera wiele różnych motorów produkcji, to tylko te z nich będą hybrydyzowały (będą „aktywne"), które mają odpowiednie sekwencje. Produkcje molekularne są od siebie niezależne pod warunkiem, że liczba cząsteczek reprezentujących słowo wejściowe jest dostatecznie duża. W tym rozdziale założono, że bada się pojedynczą produkcję molekularną i jedno słowo wejściowe.

Cząsteczka uzyskana w wyniku hybrydyzacji (krok 2, alg. 1) motoru produkcji *ab* → c do cząsteczki reprezentującej słowo wejściowe *xxab* jest pokazana na Fig. 4c

Tworzenie łańcucha pośredniego przez polimerazę DNA jest kolejnym krokiem algorytmu. W roztworze umieszcza się startery, cząsteczki o sekwencji **end**. Następnie prowadzi się reakcję polimeryzacji, w takich warunkach, aby polimeraza mogła używać najpierw jednej a następnie drugiej nici jako wzorca. Użyty enzym charakteryzuje się dużym prawdopodobieństwem skoku, to znaczy przejścia z jednej nici na drugą pokazanego na Fig. 4d i 4c. Powstaje struktura zbudowania z trzech nici DNA, schematycznie pokazana na Fig. 4f. Zjawisko skoku jest kluczowe dla działania przedstawionej realizacji molekularnej produkcji. Aby takie zjawisko zaistniało należy, oprócz zastosowania odpowiedniego enzymu, zapewnić sprzyjające warunki.

Łańcuch pośredni, jest to nić DNA zbudowana przez polimerazę. Ma on sekwencję komplementarną do sekwencji reprezentującej słowo wyjściowe. Dla rozpatrywanego przykładu, nić ta rozpoczyna się sekwencją $\overline{\mathbf{end}}$, następnie występują sekwencje $\overline{\mathbf{c}}$, $\overline{\mathbf{x}}$, $\overline{\mathbf{x}}$ zaś jest zakończony sekwencją $\overline{\mathbf{start}}$ (patrz Fig. 4f).

Ostatni krok algorytmu to budowa łańcucha reprezentującego słowo wyjściowe. Użyto reakcji polimeryzacji ze starterami o sekwencji **start**, nicią bazową był łańcuch pośredni (Fig. 4g i 4h). Otrzymaną dwuniciową cząsteczkę (fig. 4i) denaturuje się i do dalszych obliczeń przekazuje się tylko nić kodującą słowo, to znaczy zawierającą sekwencję **start** na początku i **end** na końcu. Jest to łańcuch pokazany na Fig. 4j.

Jeżeli słowo wejściowe nie będzie zawierało sekwencji *ab* na końcu 3' to motor produkcji *ab* → *c* nie będzie hybrydyzował, więc cząsteczka pokazana na Fig. 4c nie powstanie. Nie zajdzie skok polimerazy i w konsekwencji nie zostanie utworzona cząsteczka reprezentująca słowo wyjściowe.

Weryfikacja eksperymentalna produkcji molekularnej ze skokiem.

Przedstawiona koncepcja produkcji molekularnej ze skokiem (alg. 1) została zweryfikowana doświadczalnie. Eksperyment opisany poniżej potwierdził możliwość syntezy nici DNA przez polimerazę na podstawie dwóch nici bazowych. Można przyjąć, że przedstawiony schemat stanowi przykład działania produkcji molekularnej ze skokiem, a więc kluczowy proces w automacie molekularnym.

Doświadczenie składało się z trzech etapów:

1. hybrydyzacja;
2. polimeryzacja ze skokiem;
3. badanie cząsteczki powstałej po przeskoku.

Celem eksperymentu, oprócz potwierdzenia możliwości występowania skoku polimerazy, była próba oszacowania częstości takiego zjawiska.

Budowa cząsteczek

Słowo wejściowe było reprezentowane przez jednoniciową cząsteczkę DNA faga M13, oznaczanej m13mp18 (genBank: www.ncbi.nlm.nlh.gov; No X02513) o długości 7249 bp (producent: Amersham). Motorem produkcji był syntetyczny oligonukleotyd o sekwencji m13HAK. Pozostałe cząsteczki użyte w doświadczeniu to także syntetyczne oligonukleotydy oznaczane m13PKO oraz M13P. Służą one jako startery polimerazy. Sekwencje użytych cząsteczek zostały zebrane w tabeli 3.

Nazwa	Długość	Sekwencja
m13	7249	M13mp18 - patrz genBank
m13PKO	21	CTAGCACTACAACTCGGACTA
M13HAK	55	GAGGTCATTTTTGCGGATGGCTTAGAGCTTCCG - GTAGTCCGAGTTGTAGTGCTAG
m13P	22	CTATTAGTAGAATTGATGCCAC

T a b e l a 3: Sekwencje łańcuchów DNA używanych w doświadczeniu.

Hybrydyzacja

Pierwszym etapem doświadczenia była hybrydyzacja. Składniki roztworu zostały podane w tab. 4. Mieszaninę podgrzano do 95°C na 60 s, aby zlikwidować przypadkowe hybrydyzacje. Następnie temperaturę zmniejszono do 72°C

Nazwa	Stężenie	Ilość
m13	0.44 pM/µl	5 µl
M13HAK	5 pM/µl	1 µl
M13PKO	5 pM/µl	1 µl
Bufor (Amersham)	10 x	4 µl
H_2O		29 µl
RAZEM		40 µl

T a b e l a 4: Mieszanina użyta do hybrydyzacji i polimeryzacji.

na 60 s. W tej temperaturze zachodziła reakcja hybrydyzacji. W wyniku otrzymano strukturę przedstawioną schematycznie na fig. 5. Do cząsteczki **m13**, której fragment jest widoczny w górnej części rysunku, zostaje przyłączona cząsteczka **m13HAK**. Sekwencje tych cząsteczek są tylko w części komplementarne, dlatego hybrydyzacja jest tylko częściowa. Z kolei (10 m13HAK dołączana zostaje cząsteczka **m13PKO**, która będzie użyta jako starter dla polimerazy. W doświadczeniu dodano nadmiar cząsteczek **m13HAK** i **m13PKO**, aby uzyskać większe prawdopodobieństwo wystąpienia skoku polimerazy.

Polimeryzacja z przeskokiem

Kolejny etap eksperymentu to reakcja polimeryzacji z przeskokiem. Użyto 2 jednostek enzymu Taq DNA polimerazy (otrzymywanej w IBiA), które dodano do mieszaniny przedstawionej w tab. 4. Mieszaninę podzielono następnie na dwie części, jedną z nich utrzymywano w temperaturze 50°C przez 60 s, drugą przez 20 min.

Badanie wyników

Badanie wyników miało na celu udowodnienie możliwości zachodzenia zjawiska przeskoku polimerazy. Cząsteczka, która powstała w wyniku molekularnej produkcji powinna mieć ściśle określoną sekwencję. w wyniku polimeryzacji z przeskokiem, która jest najważniejszym etapem produkcji mole-

kularnej, powinno się uzyskać cząsteczki, które byłyby wykładniczo powielane przez PCR o starterach **m13P** i **m13PKO**. Cząsteczki te powinny mieć długość 242 bp.

W celu wykrycia takich cząsteczek mieszaninę uzyskaną w poprzednim etapie odbiałczono za pomocą fenolu. Do tak uzyskanego roztworu cząsteczek DNA dodano startery **m13P** i **m13PKO** oraz inne składniki, które zostały wymienione w tab. 5. Wykonano 26 i 29 cykli PCR dla tego roztworu: 94°C przez 15 s, 50°C przez 15 s, 72°C przez 30 s. Następnie cząsteczki poddano procesowi elektroforezy na żelu poliakrylamidowym 6% (59:1) w buforze TE. Żel barwiono bromkiem etydyny przez 10 min i sfotografowano w świetle UV. Elektroforetogram jest przedstawiony na fig. 6.

Nazwa	Stężenie	Ilość
Mieszanina DXA po odbiałczeniu	0.05 pM/µl	1 µl
m13PKO	5 pM/µl	1 µl
m13P	5 pM/µl	1 µl
DNTP		2 µl
Bufor	10 x	4 µl
Tag DNA polimeraza (dodawana później)	2 u/µl	1 µl
H_2O		30 µl
RAZEM		40 µl

T a b e l a 5: Mieszanina do PCR.

W pierwszej kieszonce (Fig. 6) są cząsteczki DNA uzyskane po PCR fragmentu przeskoku polimerazy trwającym 60 s po 29 cyklach. Kieszonka 2 zawiera ten sam fragment, po 26 cyklach PCR. Kieszonki 4 i 5 zawierają fragment, uzyskany przez przeskok trwający 20 min, po 29 i 26 cyklach PCR.

Dla reakcji przeskoku trwającej 60 s uzyskano prążki dłuższe niż spodziewane. Dla reakcji trwającej 20 min uzyskana cząsteczka ma właściwą długość równą 242 bp. Zauważono, że długość trwania reakcji przeskoku nie wpływa na liczbę uzyskanych cząsteczek.

Sprawdzenie

Aby dodatkowo upewnić się, że otrzymana cząsteczka jest tą właściwą przeprowadzono reakcję trawienia restryktazami. Sprawdzanie to dotyczy cząsteczki uzyskanej po przeskoku trwającym 20 min, gdyż ma ona właściwą długość.

Reakcja trawienia enzymami przeprowadzona została dla trzech restryktaz: Rsal, *Hinf*I Hpall (producent: Amersham). Składniki mieszaniny są pokazane w tab. 6.

Nazwa	Stężenie	Ilość
DNA		5 µl
Restryktaza	10 u/µl	1 µl
Bufor	10 x	4 µl
BSA		1 µl
Woda		2 µl
RAZEM		10 µl

T a b e l a 6: Skład mieszaniny do cięcia restryktazą uzyskanych fragmentów DNA.

Przy trawieniu restryktazą Rsal, która rozpoznaje sekwencję **T'GTAC** znajdującą się na 165 miejscu w łańcuchu, powinno się uzyskać dwa fragmenty o długościach 77 bp i 165 bp. Trawienie enzymem H*inf*I, rozpoznającym sekwencję **T'GANTC** występującą na 127 i 207 miejscu powinno dać 3 fragmenty o długościach: 35 bp, 80 bp oraz 127 bp. Trawienie Hpall, powinno dać dwa fragmenty o długościach 23 bp i 219 bp, ponieważ enzym ten rozpoznaje sekwencję **T'CCGG** znajdującą się na 219 miejscu w rozpatrywanym łańcuchu.

Reakcja przebiegała przez 1 godzinę w 37°C. Uzyskane DNA poddano elektroforezie w żelu poliakrylamidowym 6% (59:1), następnie, barwieniu bromkiem etydyny a w końcu wykonano fotografię przedstawioną na fig. 7.

Jak widać (Fig. 7) cząsteczka uzyskana po przeskoku jest cięta na fragmenty o długości równe spodziewanym, więc cząsteczka jest z dużym prawdopodobieństwem tą, której oczekiwano.

Prążek uzyskany po przeskoku dla reakcji trwającej 20 mm, po 26 cyklach (Fig. 6 kieszonka 5) ma jasność pomiędzy jasnością prążka we wzorcu dla długości 298 bp (około 0.05 pM DNA), a jasnością prążka dla długości 267 bp+257 bp (0.1 pM DNA). Szacunkowo w tym prążku znajduje się 0.07 pM DNA.

Podsumowanie wyników eksperymentu

Po zakończeniu reakcji przeskoku w probówce było 0.05 pM struktur DNA po hybrydyzacji (pokazanych na Fig. 5). Zakładając, że proces odbiałczania fenolem jest wydajny w 70%, po odbiałczaniu mamy 0.035 pM DNA. Ponieważ użyto 26 cykli PCR, zakładając, że cały czas powielanie było wykładnicze, można oszacować prawdopodobieństwo przeskoku. Wynosi ono $3.2*10^{-8}$ ($0.7/(0.35 *2^{-26})$).

P r z y k ł a d 2

Przykład 2 ilustruje drugą odmianę sposobu według wynalazku.

Produkcja molekularna z PCR jest zmodyfikowanym procesem produkcji molekularnej ze skokiem. Od pierwowzoru (alg. 1) różni się dodatkowym krokiem PCR, który powiela cząsteczki reprezentujące słowo wyjściowe. Przedstawiony poniżej proces może być stosowany wielokrotnie, gdy słowo wyjściowe jednej produkcji molekularnej jest wejściowym kolejnej, gdyż liczba cząsteczek reprezentujących słowo wyjściowe jest zbliżona do liczby cząsteczek reprezentujących słowo wejściowe (nie jest o kilka rzędów wielkości mniejsza, co było właściwością poprzedniej implementacji). Taka właściwość jest potrzebna przy wykrywaniu sekwencji DNA opisanych wyrażeniem regularnym, gdy liczba kroków jest większa od jednego.

Produkcję molekularną z PCR opisano algorytmem 2. Motor produkcji jest jednoniciową cząsteczką DNA, pokazaną schematycznie na Fig. 8a. Budowa tej cząsteczki jest podobna do budowy motoru produkcji użytego w poprzedniej realizacji (Fig. 4a). Wyliczając od 5' końca, cząsteczka ta ma sekwencje: **end**, kodującą produkcję (w sposób identyczny jak poprzednio), **end**, **r** - sekwencja rozpoznawana przez enzym restrykcyjny, **p** - jest to sekwencja pomocnicza (sztucznie wydłużono cząsteczkę).

1. Umieszczenie w roztworze zawierającym motory produkcji cząsteczek kodujących słowo wejściowe;
2. hybrydyzacja;
3. polimeryzacja z przeskokiem;
4. PCR;
5. cięcie enzymem restrykcyjnym.

Algorytm 2: Produkcja, molekularna z PCR.

Algorytm 2, dla pierwszych trzech kroków jest podobny do algorytmu 1 PCR ze starterami **start** oraz **prend** jest kolejnym, czwartym krokiem omawianego algorytmu (Fig. 8e). Powoduje on zwiększenie ilości łańcuchów pośrednich. Dodatkową korzyścią jest otrzymywanie cząsteczki wyjściowej, ponieważ PCR dostarcza obu komplementarnych nici.

Ostatnim etapem algorytmu jest cięcie enzymem restrykcyjnym, który przecina cząsteczki w miejscu **r** (Fig. 8f). Cząsteczki po tym kroku (Fig. 8g) reprezentują słowo wyjściowe. Dla rozpatrywanego przykładu jest to słowo *xxc*.

Produkcja molekularna z PCR umożliwia przeprowadzenie, procesu obliczeniowego tyko wtedy, gdy wiele cząsteczek reprezentujących słowo wejściowe jest obecnych w roztworze (przynajmniej 10^8, biorąc pod uwagę prawdopodobieństwo skoku oszacowane poprzednio). Jeżeli cząsteczek tych nie będzie wystarczająco dużo, to można nie uzyskać cząsteczki reprezentującej słowo wyjściowe.

Realizacja automatu niedeterministycznego

W poprzednim punkcie przedstawiono kilka realizacji nowego procesu na cząsteczkach DNA, pozwalającego przetwarzać informację, nazywanych produkcją molekularną. Badanie, czy cząsteczka DNA jest opisane wyrażeniem regularnym jest procesem wykorzystującym cyklicznie różne produkcje molekularne. Poniżej zostanie opisany sposób, który jest jedną z możliwych realizacji molekularnego automatu skończonego.

Zasada działania automatu molekularnego

Mając automat molekularny (czyli zbiór produkcji molekularnych) można badać, czy automat ten akceptuje słowo wejściowe. Słowo to jest analizowane przez kolejne produkcje molekularne, które tworzą łańcuchy pośrednie. Jeżeli uda się uzyskać łańcuch reprezentujący symbol startowy gramatyki, to analiza daje wynik pozytywny - łańcuch wejściowy jest opisywany przez wyrażenie regularne. w przeciwnym wypadku odpowiedź jest negatywna. Krok obliczeń jest procesem molekularnej produkcji. W pojedynczym kroku obliczeń jest analizowany jeden symbol (ostatni) ze słowa wejściowego, więc liczba kroków wynosi |R| (długość słowa wejściowego). Automat molekularny działający w przedstawiony sposób przedstawiono schematycznie na Fig. 3 oraz opisany alg. 3.

1. Umieszczenie cząsteczek repr. słowo S w roztworze zawierającym motory produkcji
2. powtarzaj |S| razy:
(a) produkcje molekularne;
(b) separacja;
3. badanie obecności cząsteczki wyjściowej.

Algorytm 3: Automat molekularny.

Można przyjąć ze automat molekularny bada wszystkie możliwe przejścia dla automatu niedeterministycznego jednocześnie. Stosuje się tutaj jednocześnie kilka produkcji molekularnych, które działają na tym samym słowie wejściowym.

Jak wynika z alg. 3, po procesie produkcji molekularnej stosuje się proces separacji mający na celu usunięcie z roztworu cząsteczek reprezentujących słowa wejściowe dla produkcji molekularnych, zaś pozostawienie cząsteczek reprezentujących słowa wyjściowe. Separacja może także usunąć zbędne cząsteczki, używane jako pomocnicze w procesie produkcji molekularnej.

Rozpoznawanie słów realizowane przez automat molekularny

Działanie automatu molekularnego zostanie rozpatrzone na przykładzie języka opisanego wyrażeniem regularnym a(a|b)*b. Automat niedeterministyczny, rozpoznający słowa należące do tego języka, został przedstawiony na Fig. 1. Automat molekularny dla tego języka pokazany na fig. 9 jest reprezentowany przez zbiór motorów produkcji. Dla rozpatrywanego języka produkcje molekularne odpowiadają redukcjom: a1→ 0, a1 → 1, b1 → 1 oraz b → 1.

Słowo wejściowe *abb* jest reprezentowane przez łańcuch DNA przedstawiony na fig. 10. Oprócz sekwencji reprezentujących kolejne symbole łańcuch ten zawiera dwie sekwencje pomocnicze oznaczone etykietami **start** oraz **end**.

Początek obliczeń to dodanie cząsteczek reprezentujących słowa wejściowe do probówki zawierającej cząsteczki reprezentujące automat molekularny. Następnie przeprowadza się kilkukrotnie procesy produkcji molekularnych rozdzielone procesem separacji.

Dla słowa *abb* odpowiedź uzyskuje się w 3 krokach (długość tego słowa wynosi 3), które zostały schematycznie pokazane na Fig. 11, 12, 13. Rysunki obrazują pierwszą fazę procesu produkcji molekularnej, którą jest przyłączanie motoru produkcji do łańcucha wejściowego (krok 2 alg. 1 lub krok 2 alg. 2).

Na Fig. 11 pokazano pierwszy krok obliczeń. Do łańcucha reprezentującego słowo wejściowe *aab* (Fig. 10) przyłącza się motor produkcji b → 1. Pozostałe motory (reprezentujące inne produkcje molekularne, fig. 9) nie mogą hybrydyzować, ponieważ nie mają wystarczającej liczby komplementarnych nukleotydów. Można powiedzieć, że produkcja, molekularna b → 1 jest aktywna. Sprawia ona, że uzyskuje się łańcuch DNA reprezentujący słowo *ab*1.

Po separacji (odizolowaniu cząsteczek reprezentujących słowo *abb*) przeprowadza się kolejny krok molekularnej produkcji. Tym razem produkcja b1 → 1 jest aktywna (do łańcucha *ab*1 przyłącza się motor produkcji molekularnej b1 → 1), co zostało pokazano na Fig. 12. Po zakończeniu molekularnej produkcji uzyskuje się słowo *a*1.

Do łańcucha reprezentującego słowo *a1* przyłączają się dwa różne motory produkcji:

a1 → 0 oraz a1 → 1, co pokazano na Fig. 13. W pojedynczym kroku będą analizowane wszystkie możliwe przejścia, zatem uzyskuje się dwa wyjściowe łańcuchy DNA: jeden reprezentuje słowo 0 zaś drugi słowo 1. W kroku tym pokazano możliwość aktywacji kilku produkcji molekularnych, co można interpretować jako równoległe badanie kilku przejść w automacie niedeterministycznym.

Ponieważ jest to trzeci, ostatni krok obliczeń, bada się, czy w roztworze wyjściowym znajduje się odpowiedni łańcuch DNA. W przedstawionym przypadku jest to łańcuch reprezentujący słowo 0, który znajduje się w roztworze, zatem odpowiedź jest pozytywna.

Dla tego automatu oraz słowa *bbb* analiza została pokazana na Fig. 14. W pierwszym kroku zostaje użyta produkcja b → 0, więc powstaje słowo *bb0*. W kolejnym produkcja b0 → 1 daje łańcuch reprezentujący słowo b1. W trzecim kroku żaden motor produkcji nie przyłączył się, zatem nie wykonuje się żadna produkcja molekularna. Krok ten jest krokiem pustym, nie powstaje nowy łańcuch. Ponieważ w roztworze nie znajduje się łańcuch reprezentujący słowo 0, odpowiedź jest negatywna.

W przedstawionej realizacji wykonuje się dokładnie tyle kroków algorytmu, ile symboli ma badane słowo. Dla słów, które nie spełniają gramatyki wejściowej często już wcześniej można stwierdzić, że wynik będzie negatywny (tak jak w przypadku pokazanym słowa *bbb* dla ostatnio opisanego automatu). Jeżeli zostanie wykazane, że nie powstały żadne nowe cząsteczki w wyniku stosowania produkcji molekularnych, można dać od razu odpowiedź negatywną. Automat działający według tego schematu przedstawiono na Fig. 15. Automat tego typu nie wymaga znajomości długości słowa wejściowego, co może być zaletą w niektórych przypadkach. Wadą natomiast jest konieczność wielokrotnego sprawdzania warunków zakończenia, co może być kosztowne. Proces separacji dla automatu molekularnego.

Analizowanie słów realizowane przez automat skończony przedstawiony w poprzednim punkcie wymaga aby kroki obliczeń były rozdzielone. Krokiem obliczeń jest proces produkcji molekularnej. Rozdzielenie polega na odseparowaniu cząsteczek wejściowych od wyjściowych, na przykład przez umieszczenie ich w oddzielnych probówkach. Poniechanie tego sprawia, że będą istniały przypadki, gdy oddziaływania pomiędzy tymi cząsteczkami zaburzy proces analizy.

Separacja może zostać zrealizowana na kilka różnych sposobów, np. poprzez rozpoznawanie długości. Dokładniejszym i mniej kosztownym rozwiązaniem jest użycie techniki PCR w połączeniu z rozcieńczaniem. Stosuje się PCR ze starterami dla cząsteczki wyjściowej, a następnie rozcieńcza się roztwór. Wykonując kilkukrotnie ten proces doprowadza się do tego, że w roztworze znajdują się tylko cząsteczki wyjściowe. Aby zastosować powyższą metodę, cząsteczka wyjściowa produkcji molekularnej powinna mieć inne końce niż cząsteczka wejściowa (co nie jest spełnione w implementacji automatu molekularnego przedstawionej poprzednio).

Automat molekularny, w którym użyto wspomnianej metody do realizacji separacji, opisano alg. 4 i przedstawiono na Fig. 16. W kolejnych krokach algorytmu stosuje się naprzemiennie separację wykorzystując dwa różne zestawy starterów. Wymaga to podwojenia liczby motorów produkcji molekularnej (każda produkcja molekularna, ma dwie odmiany, każda z odmian działająca dla danego zestawu starterów).

1. Umieszczenie cząsteczek kodujących słowo S w roztworze zawierającym motory produkcji;
2. Niech j=0;
3. powtarzaj |S| razy:
(a) produkcje molekularne;
(b) jeżeli j=O
PCR ze starterem1
Niech j=1;
W przeciwnym wypadku
PCR ze starterem 2;
Niech j=O;
(c) rozcieńczanie;
4. badanie obecności cząsteczki wyjściowej.

Algorytm 4 Automat, molekularny wykorzystujący PCR do separacji.

Reprezentacja słowa zostaje zmieniona. Słowo jest reprezentowane przez jednoniciową cząsteczkę DNA, zaczyna się sekwencją **start**, zaś zakończone jest sekwencją **end1** lub **end2**. Jeżeli sekwencje **end1** oraz **end2** są identyczne, to definicja jest, zgodna z poprzednią (nie ma wtedy możliwości użycia separacji poprzez PCR i rozcieńczanie).

Istnieje możliwość realizacji automatu w ten sposób, aby znajomość długości słowa wejściowego nic była potrzebna do prawidłowego jego działania. Wykorzystuje się tutaj wykrywanie obecności cząsteczki wyjściowej w roztworze, podobnie jak dla automatu z Fig. 15.

Zastrzeżenia patentowe

1. Sposób identyfikacji cząsteczek DNA opisanych wyrażeniem regularnym, **znamienny tym**, że definiuje się wyrażenie regularne, reprezentacje symboli odpowiadające sekwencjom DNA charakterystycznym dla danej zależności genetycznej oraz symboli pomocniczych, w tym używanych do definiowania motorów produkcji, reprezentacje słów wejściowych odpowiadające sekwencji DNA, oraz sekwencję cząsteczki DNA reprezentującej słowo wyjściowe będące odpowiedzią pozytywną, a z badanych próbek materiału genetycznego izoluje się nici DNA w znany sposób, po czym do pojedynczej danej nici DNA na końcu 3' dołącza się dowolną, odmienną od reprezentacji symboli sekwencję nukleotydów nazywaną dalej **end**, a na końcu 5' dowolną, odmienną od reprezentacji symboli sekwencję nukleotydów nazywaną dalej **start**, jednocześnie na podstawie określonego wcześniej wyrażenia regularnego syntezuje się, znanymi metodami inżynierii genetycznej, zbiór cząsteczek zwanych dalej motorami produkcji, które posiadają na 5' końcu sekwencję komplementarną do sekwencji **end** oznaczaną dalej $\overline{\mathbf{end}}$, następnie sekwencję komplementarną do sekwencji reprezentującej lewą stronę redukcji w rozumieniu teorii automatów skończonych, następnie sekwencję reprezentującą prawą stronę redukcji w rozumieniu teorii automatów skończonych, zaś na końcu 3' sekwencję **end**, po czym badane wstępnie przygotowane nici DNA umieszcza się razem z motorami produkcji i wykonuje się wstępną denaturację, a następnie prowadzi się proces hybrydyzacji cząsteczek reprezentujących słowa wejściowe z motorami produkcji, po czym po zakończeniu hybrydyzacji w roztworze umieszcza się startery o sekwencji **end**, następnie prowadzi się reakcję polimeryzacji z dodatkiem polimerazy o dużym prawdopodobieństwie skoku, w nadmiarze wolnych nukleotydów **A**, **C**, **G**, **T**, po czym prowadzi się proces denaturacji w znany sposób po czym dodaje się startery o sekwencji **start**, a następnie ponownie prowadzi się proces polimeryzacji i jeżeli uzyskano nową cząsteczkę, to bada się, czy ma ona analogiczną sekwencję do cząsteczki reprezentującej odpowiedź pozytywną i jeśli tak, to proces się kończy, a jeśli nie to cały cykl od momentu wstępnej denaturacji powtarza się, aż do uzyskania określonej cząsteczki DNA lub do momentu, gdy nie powstanie nowa cząsteczka wyjściowa.

2. Sposób według zastrz. 1, **znamienny tym**, że cykl od momentu wstępnej denaturacji prowadzi się *k* razy, gdzie *k* odpowiada znanej długości cząsteczki reprezentującej słowo wejściowe, zaś sprawdzanie czy uzyskano cząsteczkę reprezentującą odpowiedź pozytywną wykonuje się jednorazowo po *k* cyklach.

3. Sposób identyfikacji cząsteczek DNA opisanych wyrażeniem regularnym, **znamienny tym**, że definiuje się wyrażenie regularne, reprezentacje symboli odpowiadające sekwencjom DNA charakterystycznym dla danej zależności genetycznej oraz symboli pomocniczych, w tym używanych do definiowania motorów produkcji, reprezentacje słów wejściowych, oraz sekwencję cząsteczki DNA reprezentującej słowo wyjściowe będące odpowiedzią pozytywną, a z badanych próbek materiału genetycznego izoluje się nici DNA w znany sposób, po czym do pojedynczej badanej nici DNA na końcu 3' dołącza się dowolną, odmienną od reprezentacji symboli sekwencję nukleotydów nazywaną dalej **end** a na końcu 5' dowolną, odmienną od reprezentacji symboli sekwencję nukleotydów nazywaną dalej **start**, jednocześnie na podstawie określonego wcześniej wyrażenia syntezuje się, znanymi metodami inżynierii genetycznej, zbiór cząsteczek zwanych dalej motorami produkcji, które posiadają na 5' końcu sekwencję komplementarną do sekwencji **end** oznaczaną dalej $\overline{\mathbf{end}}$, następnie sekwencję komplementarną do sekwencji reprezentującej lewą stronę redukcji w rozumieniu teorii automatów skończonych, następnie sekwencję reprezentującą prawą stronę redukcji w rozumieniu teorii automatów skończonych, sekwencję **end**, sekwencję **r** rozpoznawaną przez wybrany enzym restrykcyjny oraz dowolną, inną niż zdefiniowane poprzednio sekwencję pomocniczą **p**, następnie badane nici DNA umieszcza się razem z motorami produkcji i wykonuje się wstępną denaturację, a następnie prowadzi się proces hybrydyzacji cząsteczek reprezentujących słowa wejściowe z motorami produkcji, a po zakończeniu hybrydyzacji w roztworze umieszcza się w nadmiarze startery o sekwencji **prend**, następnie prowadzi się reakcję polimeryzacji z dodatkiem polimerazy o dużym prawdopodobieństwie skoku, przy czym w roztworze znajdują się w nadmiarze wolne nukleotydy **A**, **T**, **G** i **C**, po czym dodaje się w nadmiarze startery o sekwencji **start** i przeprowadza się procesem wykładni-

czego powielania cząsteczek DNA (PCR), następnie przeprowadza się reakcję cięcia wybranym enzymem restrykcyjnym, który przecina te cząsteczki w miejscu **r**, i jeżeli uzyskano nową cząsteczkę, to bada się, czy ma ona analogiczną sekwencję do cząsteczki reprezentującej odpowiedź pozytywną i jeśli tak, to proces się kończy, a jeśli nic to cały cykl od momentu wstępnej denaturacji powtarza, się, aż do uzyskania określonej cząsteczki DNA lub do momentu, gdy nie powstanie nowa cząsteczka wyjściowa.

4. Sposób według zastrz. 3, **znamienny tym**, że cykl od momentu wstępnej denaturacji prowadzi się *k* razy, gdzie *k* odpowiada znanej długości cząsteczki reprezentującej słowo wejściowe, zaś sprawdzanie czy uzyskano cząsteczkę reprezentującą odpowiedź pozytywną wykonuje się jednorazowo po *k* cyklach.

Rysunki

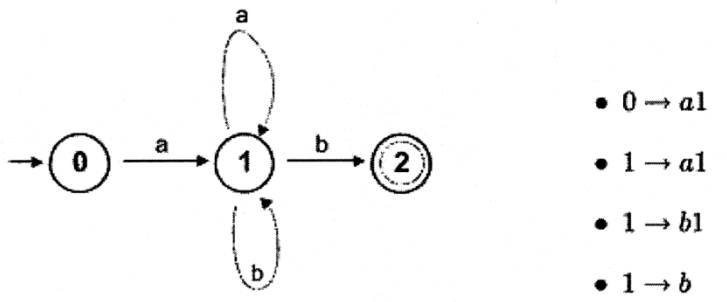

Fig. 1

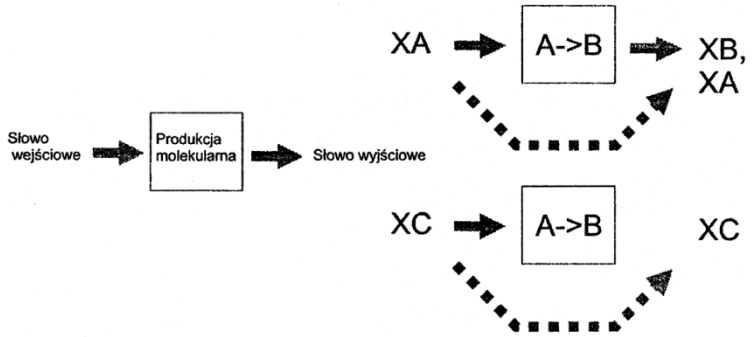

Fig. 2

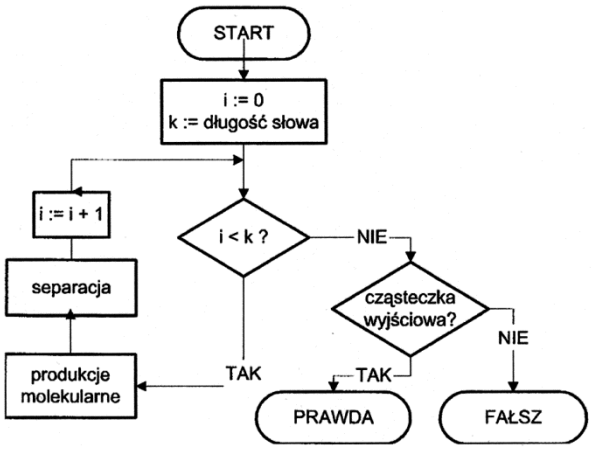

Fig. 3

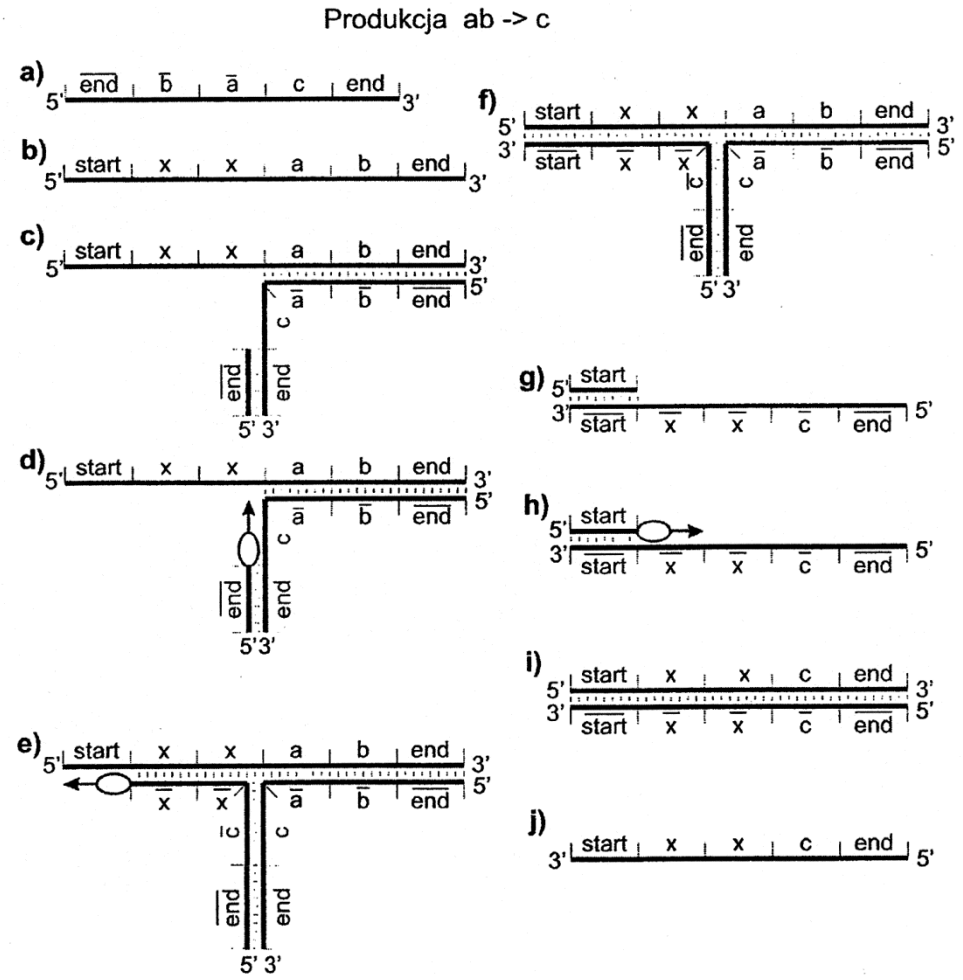
Produkcja ab -> c
a)
b)
c)
d)
e)
f)
g)
h)
i)
j)
start
end
5'
3'

Fig. 4

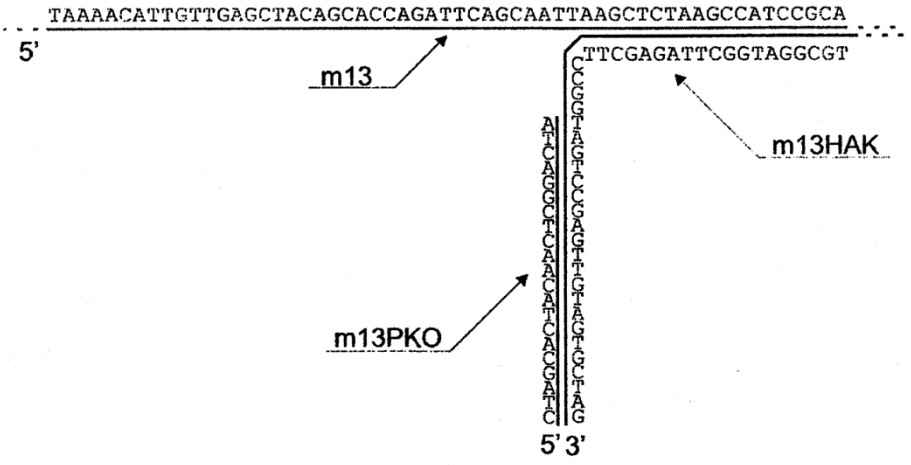
TAAAACATTGTTGAGCTACAGCACCAGATTCAGCAATTAAGCTCTAAGCCATCCGCA
5'
m13
TTCGAGATTCGGTAGGCGT
m13HAK
m13PKO
5' 3'

Fig. 5

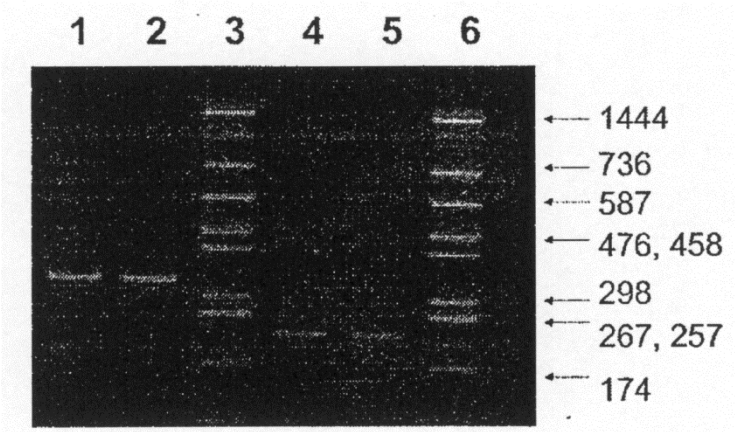
1
2
3
4
5
6
1444
736
587
476, 458
298
267, 257
174

Fig. 6

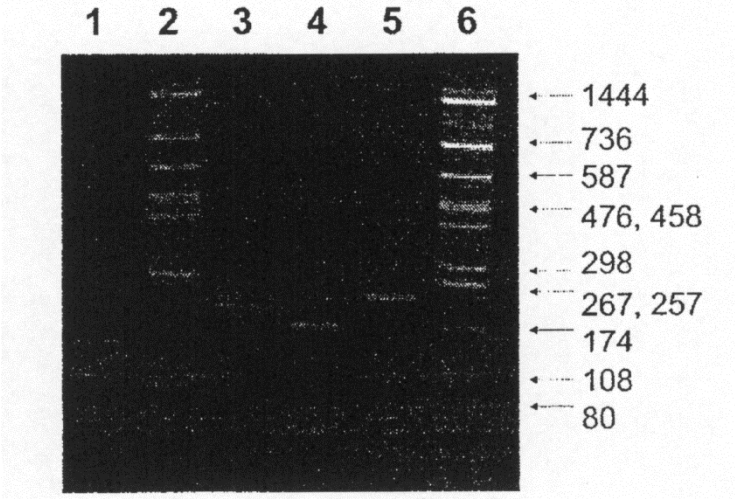
1
2
3
4
5
6
1444
736
587
476, 458
298
267, 257
174
108
80

Fig. 7

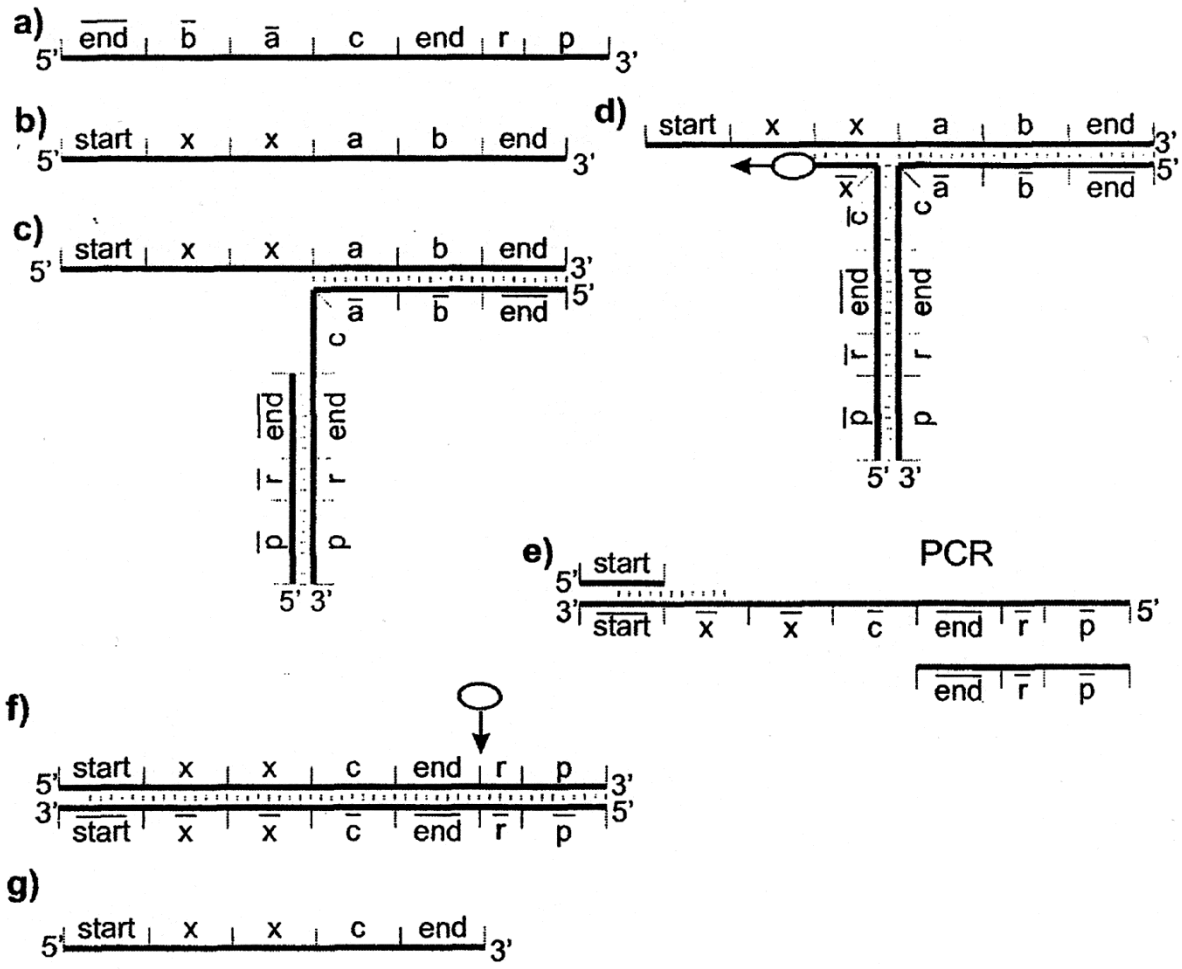
a)
5' end b̄ ā c end r p 3'
b)
5' start x x a b end 3'
c)
5' start x x a b end 3'
3' 5'
ā b̄ end
c
end r p
5' 3'
d)
start x x a b end 3'
5'
x̄ ā b̄ end
c
c
end r p
5' 3'
e)
PCR
5' start
3' start x̄ x̄ c̄ end r̄ p̄ 5'
end r̄ p̄
f)
5' start x x c end r p 3'
3' start x̄ x̄ c̄ end r̄ p̄ 5'
g)
5' start x x c end 3'

Fig. 8

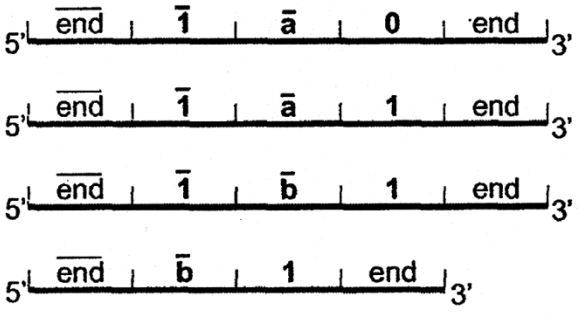
5' end 1̄ ā 0 end 3'
5' end 1̄ ā 1 end 3'
5' end 1̄ b̄ 1 end 3'
5' end b̄ 1 end 3'

Fig. 9

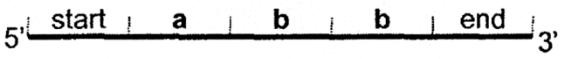

Fig. 10

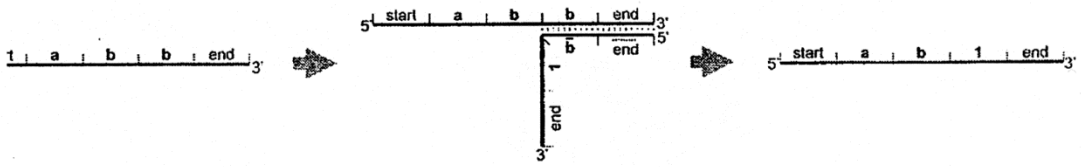

Fig. 11

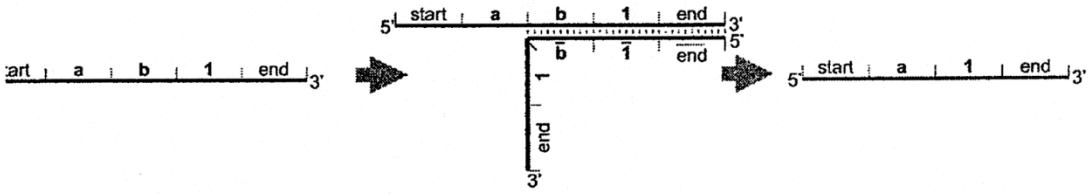

Fig. 12

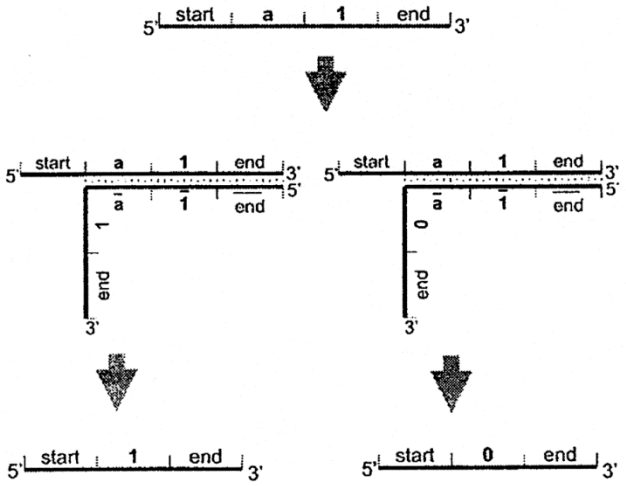

Fig. 13

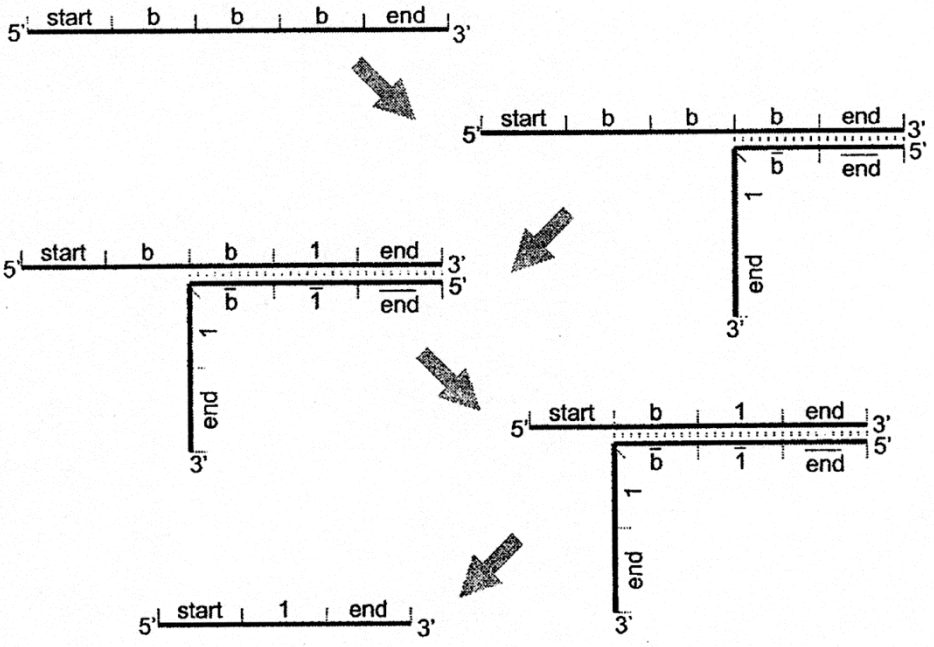
5' start b b b end 3'
5' start b b b end 3'
5'
b
end
1
end
3'
5' start b b 1 end 3'
5'
b
1
end
1
end
3'
5' start b 1 end 3'
5'
b
1
end
1
end
3'
5' start 1 end 3'

Fig. 14

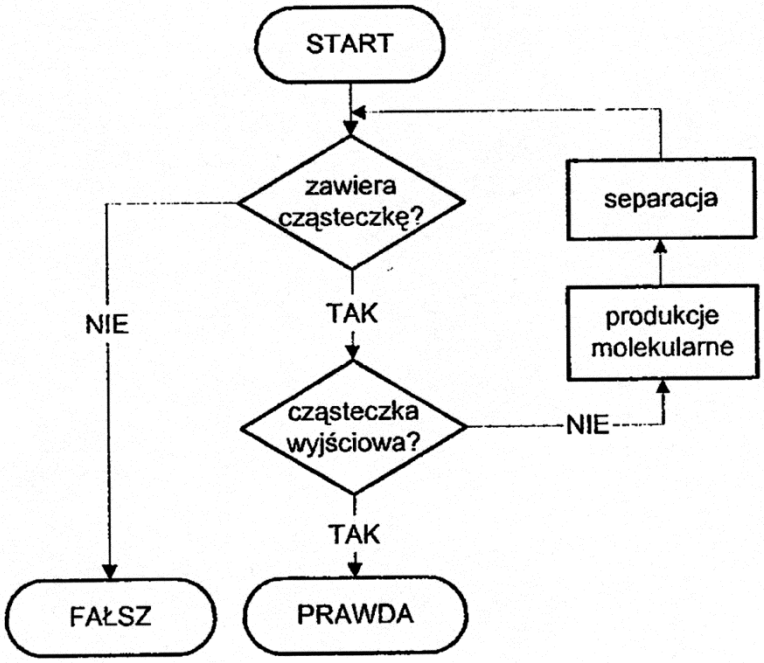
START
zawiera cząsteczkę?
NIE
TAK
cząsteczka wyjściowa?
NIE
TAK
separacja
produkcje molekularne
FAŁSZ
PRAWDA

Fig. 15

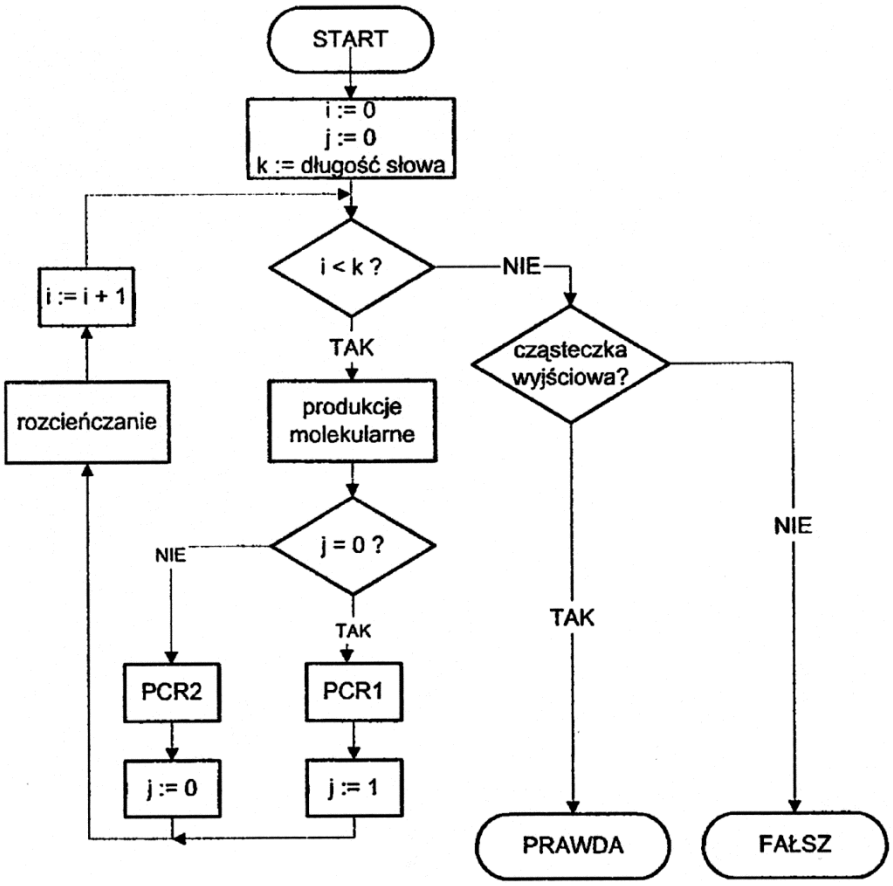

Fig. 16

Departament Wydawnictw UPRP
Cena 4,92 zł (w tym 23% VAT)