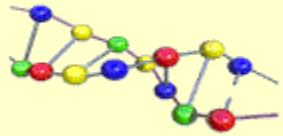
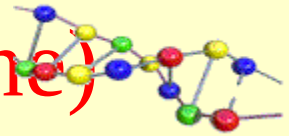
Obliczenia DNA



Jacek Mańdziuk

Obliczenia DNA (obliczenia molekularne)



Łączy w sobie elementy informatyki i biologii molekularnej. Należy do obszaru metod Computational Intelligence / Bioinformatyki.

Zasadnicza cecha: wykorzystanie materiału genetycznego do wykonywania obliczeń (rozwiązywania problemów matematyki dyskretnej i logiki).

Obliczenia wykonuje się w oparciu o tradycyjne algorytmy realizowane na (rzeczywistych) niciach molekularnych (DNA) stosując podstawowe operacje cząsteczkowe stosowane przy manipulacjach strukturami DNA.

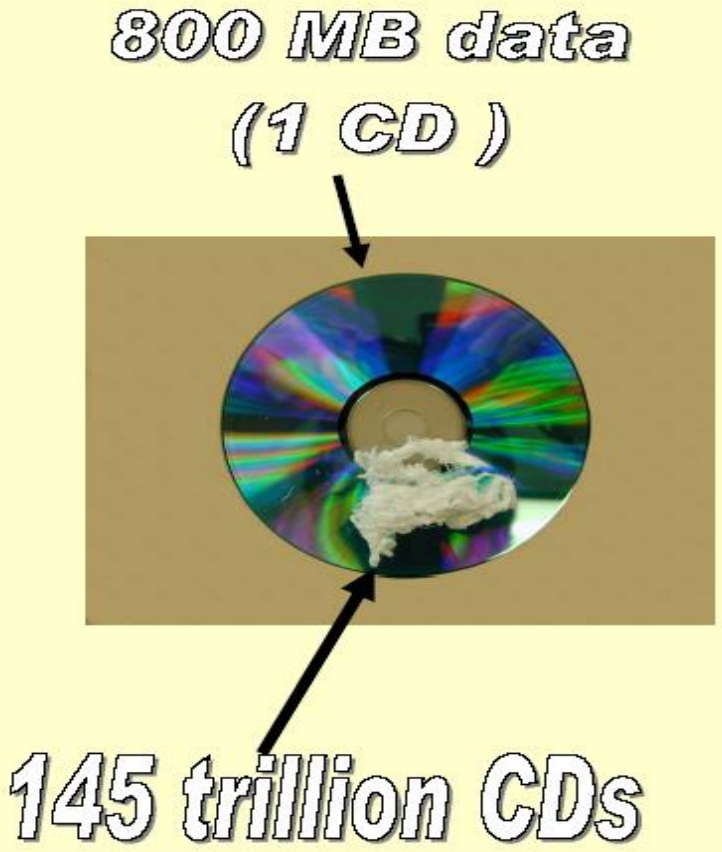
Pierwotnie wykorzystywano struktury DNA – obecnie także RNA.

Gęstość upakowania informacji DNA

1 gram materiału DNA na krązku CD o pojemności 800 MB.

W tym 1 gramie mieści się ok. 10^{15} MB danych.

Długość linii złożonej z krążków CD (dotykających się brzegami) niezbędnych do przechowania takiej ilości informacji pozwoliłaby na okrążenie Ziemi 375 razy !!! Czas ich przesłuchania wyniósłby 163,000 wieków !!!



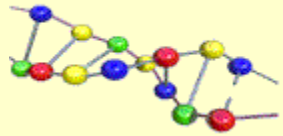
Astronomiczne ilości danych. Według Google (2015):

Co dwa dni wytwarzamy tyle danych, ile cała ludzkość wyprodukowała od początku swojego istnienia do roku 2001.

Zgromadziliśmy już 3,2 zetabajtów danych (1 zetabajt to miliard terabajtów = 10^{21} B)

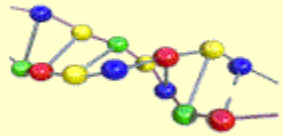
Objętość wszystkich zapisanych na świecie informacji podwaja się średnio co 18 miesięcy.

Gęstość upakowania informacji DNA



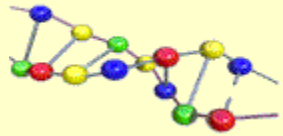
Zakładając, że linie DNA ułożone są w odstępie 0.35 nm jedna od drugiej, gęstość zapisu DNA wynosi przeszło milion Gb/ cal^2 – w porównaniu do ok. 740 Gb/cal^2 w typowym nowoczesnym dysku twardym.

W 2012 roku dwaj bioinżynierowie z Harvardu zmieścili 700 terabajtów na jednym gramie materiału genetycznego.
700TB odpowiada 14 000 dyskom Blu-ray o pojemności 50 GB.
Na tradycyjnych nośnikach taka ilość danych ważyłaby ok. 151 kg.

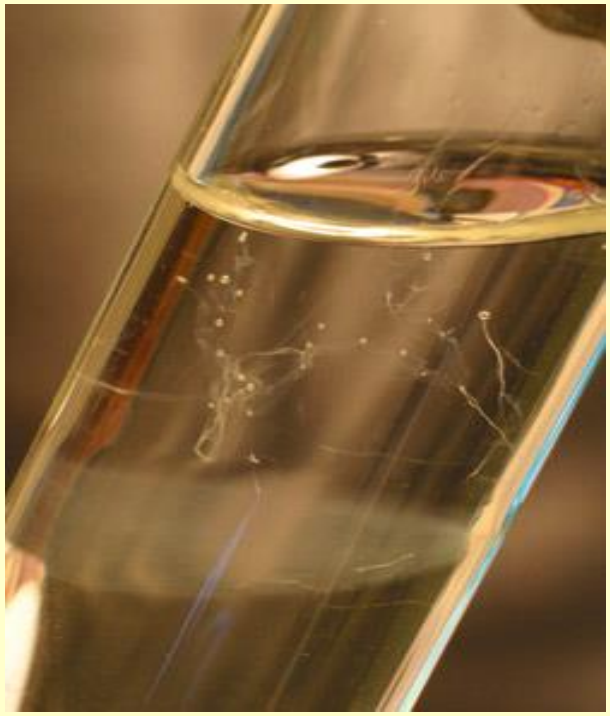


Niewielkie wymagania energetyczne - nici DNA zużywają bardzo małe ilości energii, którą czerpią bezpośrednio z otoczenia.

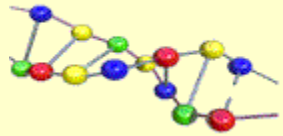
Operacje na ciągach DNA



Obliczenia DNA wykonywane są w próbówce z żelem, w której zawarte są (odpowiednio spreparowane) nici DNA.



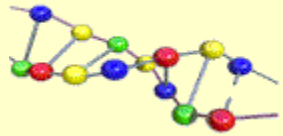
Operacje na ciągach DNA



Podstawowe kroki (operacje) algorytmów DNA, to (poprzez podgrzewanie, chłodzenie, … próbówek) **konkatenacja ciągów**, **wycinanie fragmentów ciągów** (przy pomocy enzymów), **wybieranie ciągów zawierających określony podciąg**, …

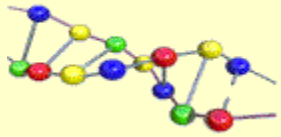
Operacje te odpowiadają operacjom AND,OR,NOT,NOR itd. klasycznego procesora.

Masowa równoległość przetwarzania



W tubie testowej zwykle umieszczone są setki miliardów nici DNA, na których wykonywane są operacje w sposób masowo-równoległy ➔ możliwe są rozwiązania problemów optymalizacyjnych metodami siłowymi!

Złożoność algorytmów DNA

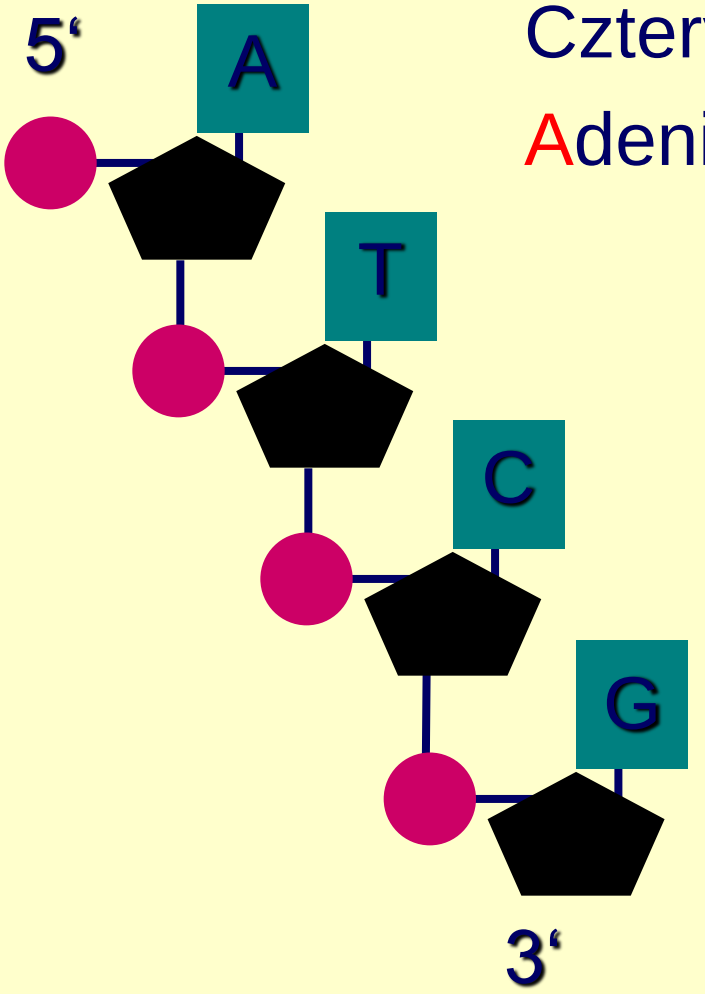


Złożoność obliczeniową algorytmów DNA mierzymy liczbą wymaganych operacji.

Złożoność pamięciową (zwykle) liczbą wykorzystanych próbówek.

Silną stroną obliczeń DNA jest masowa równoległość oraz relatywnie duża moc obliczeniowa pojedynczych elementów (nici DNA) w stosunku do ich rozmiarów.

Struktura DNA – pojedyncza nić



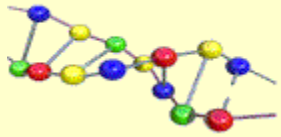
Cztery nukleotydy:

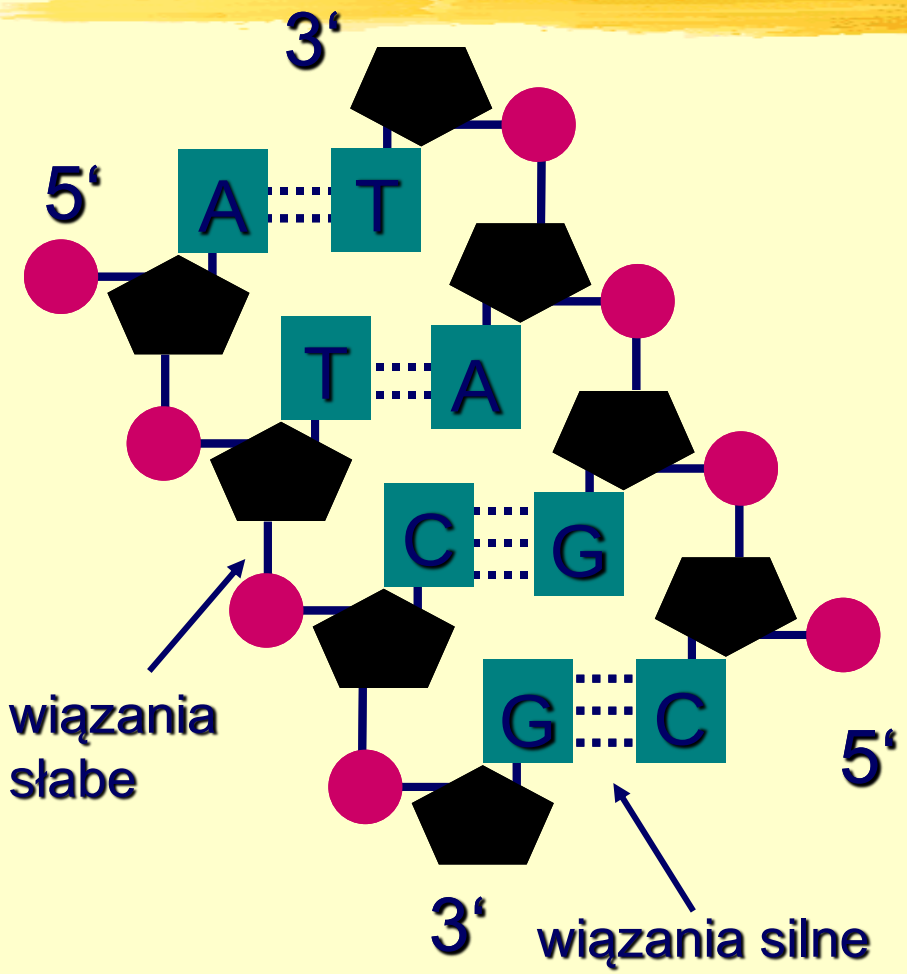
Adenina Tymina Guanina Cytozyna

- Rozróżnialne końce
  - 3’ i 5’
- Zapis sekwencji:

  5’-ATCG-3’

Struktura DNA – podwójna nić

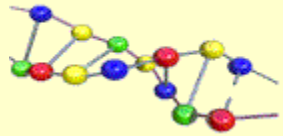




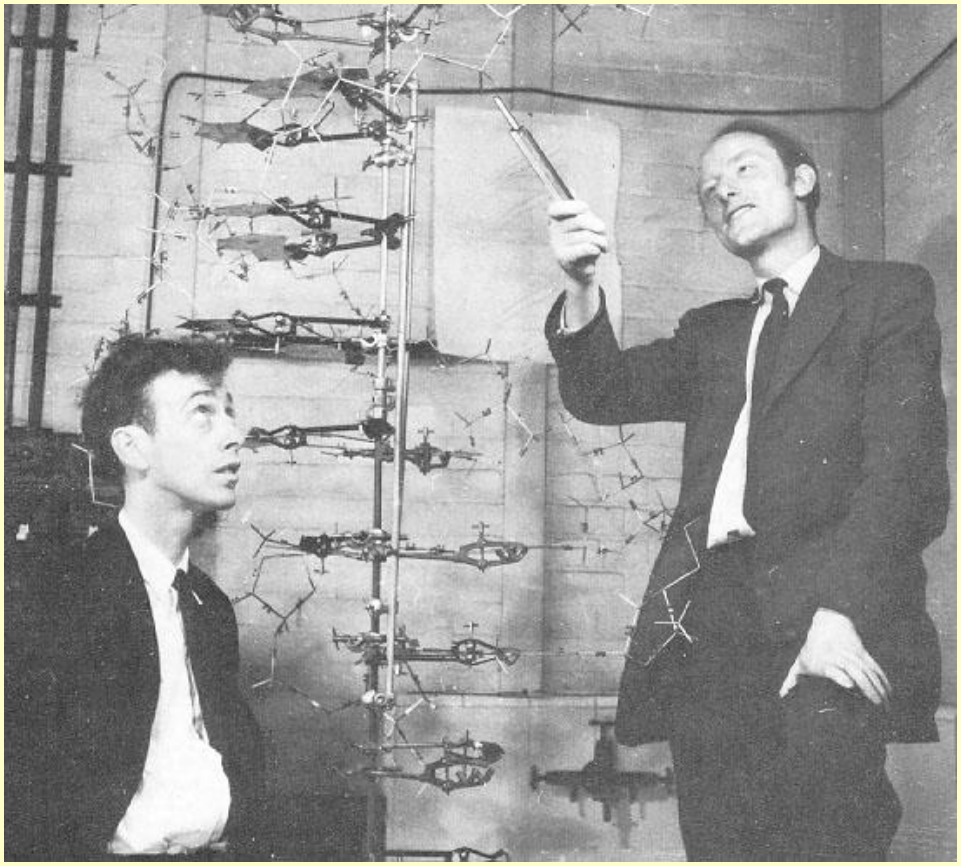
- A łączy się z T (powójnym wiązaniem słabym)
- C łączy się z G (wiązaniem potrójnym)

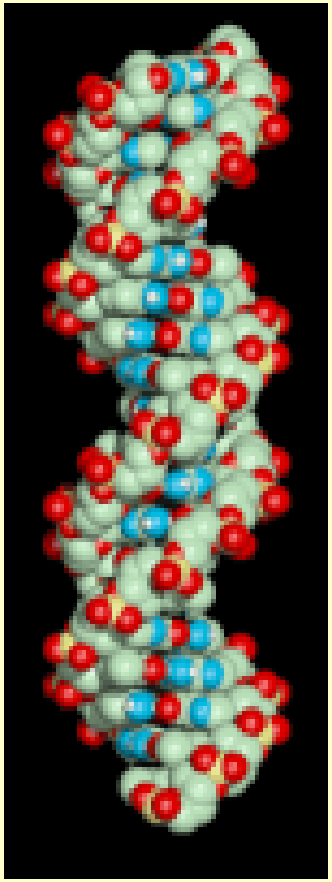
- Podwójna Helisa Watsona-Cricka
  - 5‘-ATCG-3‘
  - 5‘-CGAT-3‘

Podwójna Helisa



Odkrycie 1953, Nobel 1962

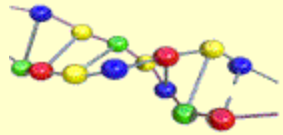




spektroskopia rentgenologiczna

+ Maurice Wilkins (Nobel - medycyna) ; Rosalind Franklin (pominięta).

Obliczenia molekularne - intuicja

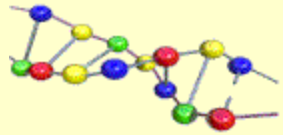


„Naturalne” podstawy ideowe:

- skomplikowana struktura żywych organizmów w ostatecznym rozrachunku jest wynikiem zastosowania skończonego zbioru operacji elementarnych (kopiowanie, podział, wstawianie, usuwanie, itd.) względem informacji początkowej (pierwotnej) zakodowanej w sekwencji DNA

- dowolne obliczenie, niezależnie od stopnia komplikacji, jest wynikiem odpowiedniego zastosowania elementarnych operacji arytmetycznych oraz logicznych

Obliczenia DNA (molekularne)



Pierwszy rzeczywisty eksperyment: Leonard Adleman (1994) wykorzystał narzędzia biologii molekularnej do rozwiązania egzystencjalnej wersji problemu TSP w niepełnym grafie skierowanym o wymiarze 7
z 14 krawędziami.

Oryginalny rysunek z pracy Adlemana

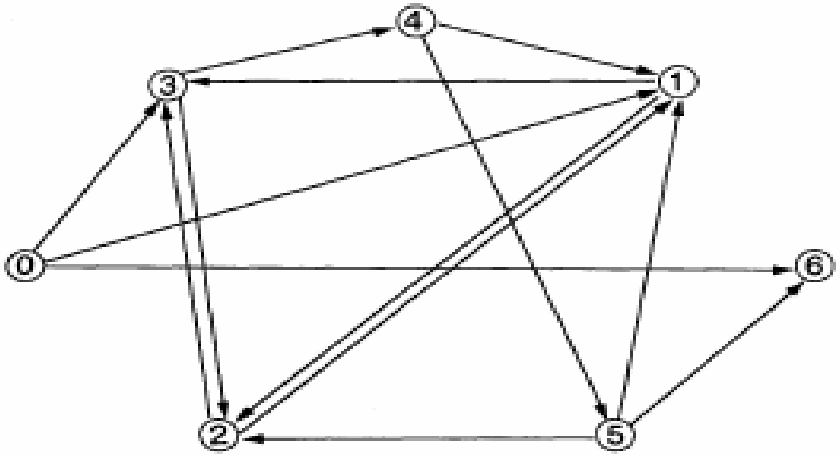
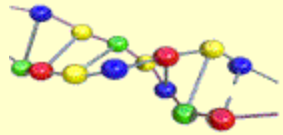


Fig. 1. Directed graph. When $v_{in} = 0$ and $v_{out} = 6$, a unique Hamiltonian path exists: 0→1, 1→2, 2→3, 3→4, 4→5, 5→6.

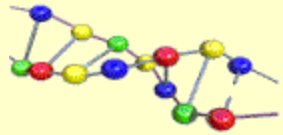
Obliczenia DNA (molekularne)



Eksperyment polegał na odpowiedniej manipulacji rzeczywistymi łańcuchami DNA umieszczonymi w tubie.

Zakodowanie specyfiki problemu w ciągach DNA, a następnie wykorzystanie technik biologii molekularnej (podgrzewanie, studzenie tuby, cięcie enzymami, łączenie poprzez dopisanie albo wstawianie ciągów, itd.) w celu wymuszenia odpowiedniej sekwencji działań - "programu DNA" zmierzających do realizacji zadania.

Problem komiwojażera - kodowanie

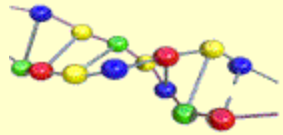


Kodowanie problemu

Każde miasto zostało zakodowane jako 20-znakowy ciąg DNA. Następnie dla każdej (zorientowanej) krawędzi grafu wygenerowano odpowiadające jej 20-znakowe ciągi złożone z sekwencji komplementarnych do drugiej połowy wierzch. pocz. (tej kraw.) oraz pierwszej połowy wierzch. końc. (tej kraw.).

Komplementarne ciągi DNA stanowią „łączniki” - ciągi „bazowe” DNA odpow. tym ciągom samorzutnie „doklejają się" do nich, a w konsekwencji generowane są ścieżki w grafie miast (ciągi bazowe i komplementarne są sklejane „na zakładkę”).

Problem komiwojażera - kodowanie



Reprezentacja wierzchołków (20-znakowe ciągi DNA):

Np.

S_2 = GTCACACTTCGGACTGACCT

S_4 = TGTGCTATGGGAACTCAGCG

S_5 = CACGTAAGACGGAGGAAAAA

5’ 20 znaków 3’

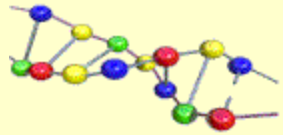
Ciągi komplementarne :

$\underline{S_2}$ = AGGTCAGTCCGAAGTGTGAC

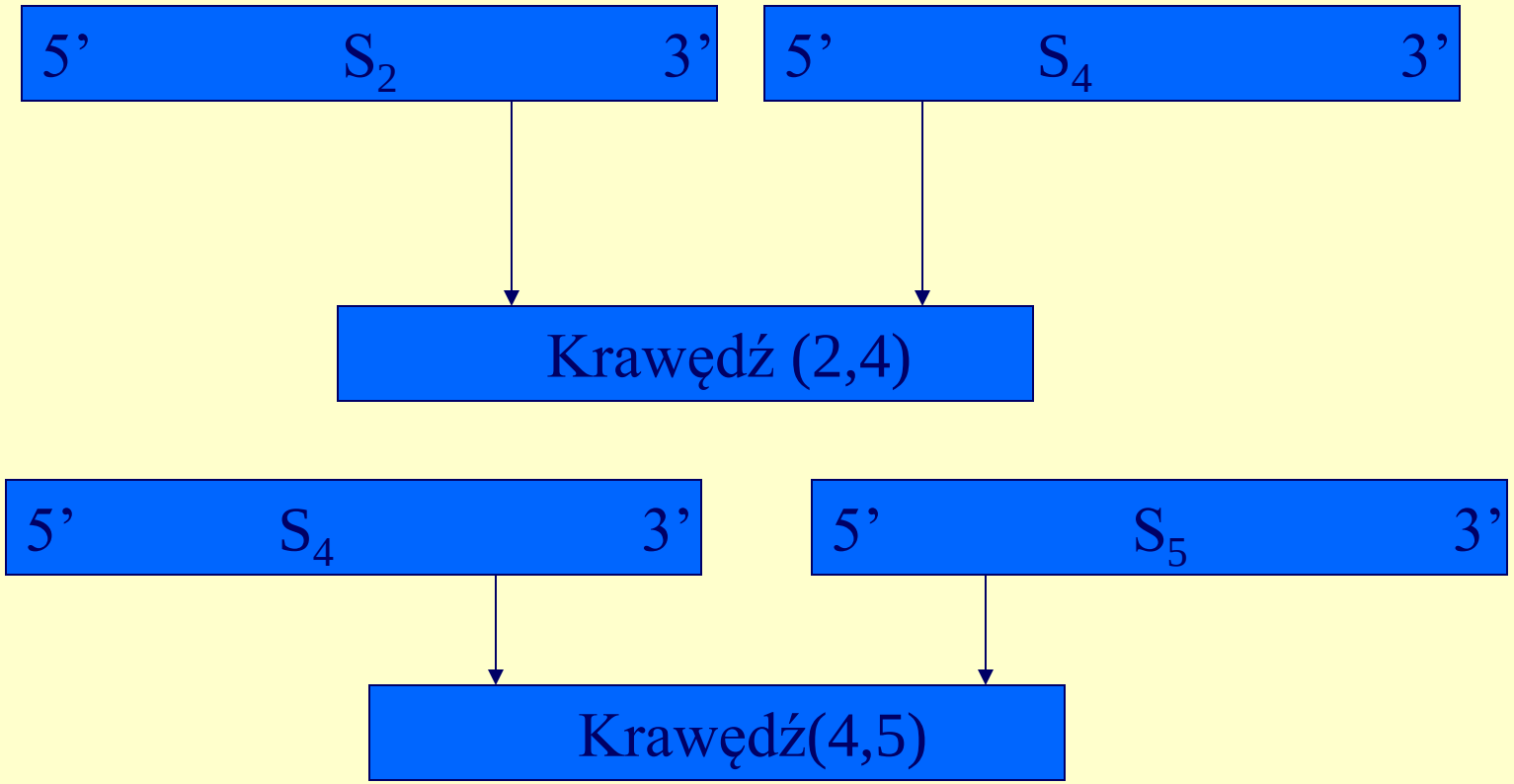
$\underline{S_4}$ = CGCTGAGTTCCCATAGCACA

$\underline{S_5}$ = TTTTTCCTCCGTCTTACGTG

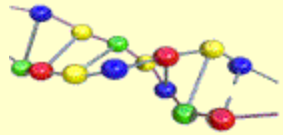
Problem komiwojażera - kodowanie



Reprezentacja krawędzi:



Problem komiwojażera - kodowanie



Reprezentacja krawędzi:

S_2 = GTCACACTTCGGACTGACCT

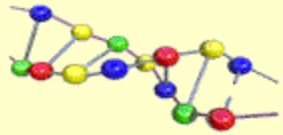
S_4 = TGTGCTATGGGAACTCAGCG

S_5 = CACGTAAGACGGAGGAAAAA

(2,4) = CCATAGCACAAGGTCAGTCC

(4,5) = GTCTTACGTGCGCTGAGTTC

Eksperyment L. Adlemana



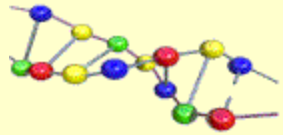
Algorytm niedeterministyczny:

Ustalony wierzchołek początkowy (vp)

[1] Wrzuć do probówki po szczypcie DNA (ok. 10^{14} nici) każdego z miast i każdej z możliwych krawędzi, dodaj wody i enzymów. Całość wymieszaj (mieszanie nici, konkatenacja, w obecności odpowiedniego enzymu (Ligazy))

[2] zatrzymaj jedynie te ścieżki, które zaczynają się w vp (w eksperymencie jest to wierzchołek nr 0)

Eksperyment L. Adlemana - c.d.

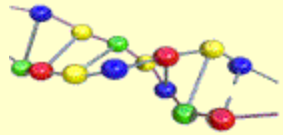


[3] jeżeli graf składa się z n wierzchołków, to zatrzymaj wyłącznie takie ścieżki, które przechodzą przez dokładnie n wierzchołków (ekstrakcja ciągów o dł. 140 za pomocą żelu agarowego)

[4] zatrzymaj wyłącznie te ścieżki, które prowadzą przynajmniej raz do każdego wierzchołka,

[5] jeżeli została chociaż jedna ścieżka odpowiedź brzmi "TAK", wpp. odpowiedź brzmi "NIE''.

Obliczenia molekularne - zalety



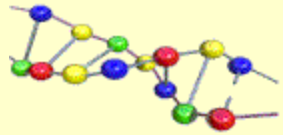
Zalety

nieograniczona pojemność – ilość DNA mieszcząca się w szklance ma moc obliczeniową dziesięciokrotnie większą od najszybszego istniejącego obecnie komputera równoległego.

obliczenia wykonywane w sposób masowo równoległy

bardzo niskie wymagania energetyczne

Obliczenia molekularne - wady



Wady

duże skomplikowanie techniczne eksperymentu - (wykonanie go zajęło prawie tydzień)

operacje [1]-[5] wykonywane na łańcuchach nie są bezbłędne. Wzrost wymiaru problemu → wzrost prawdopodobieństwa błędu

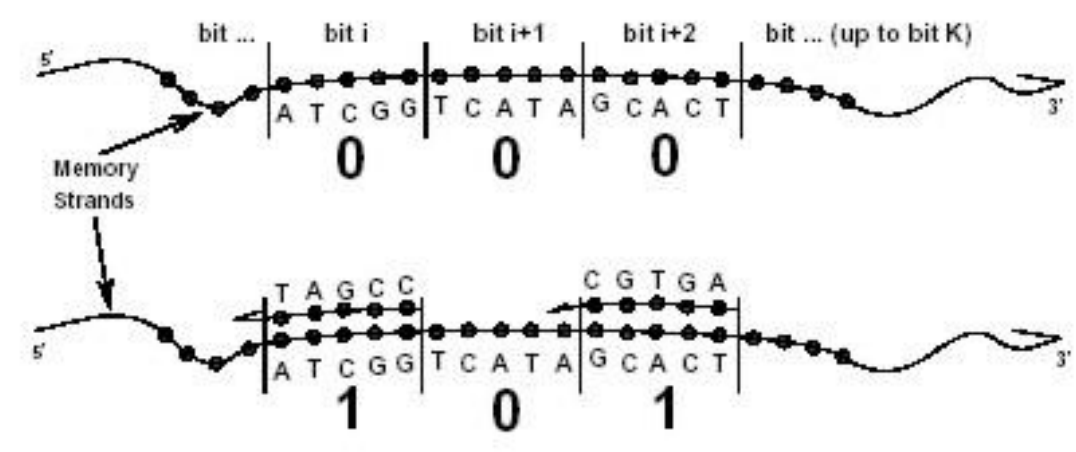
duże wymagania ilościowe odnośnie DNA (czas obliczeń skaluje się dobrze, ale wymagana ilość nici DNA rośnie ekspotencjalnie jak 2^N !!!);

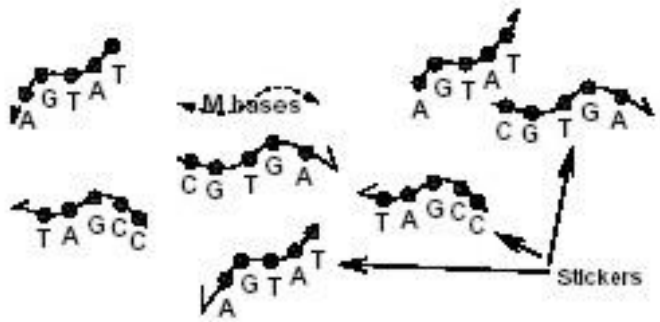
➔ zastosowanie wprost metodyAdlemana do problemu o wymiarze 200 wymagałoby ilości DNA o wadze przekraczającej wagę ziemi !!!.

”Komputer DNA” – model *sticker*

„Architektura” komputera DNA:

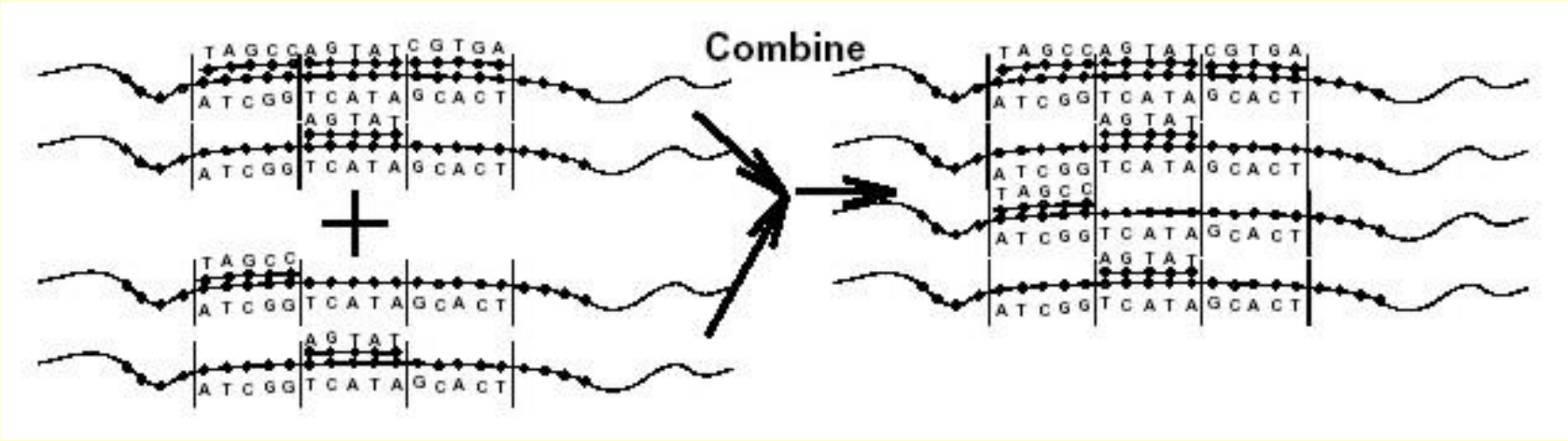
Nici pojedyncze – zera bitowe
Nici podwójne – jedynki bitowe





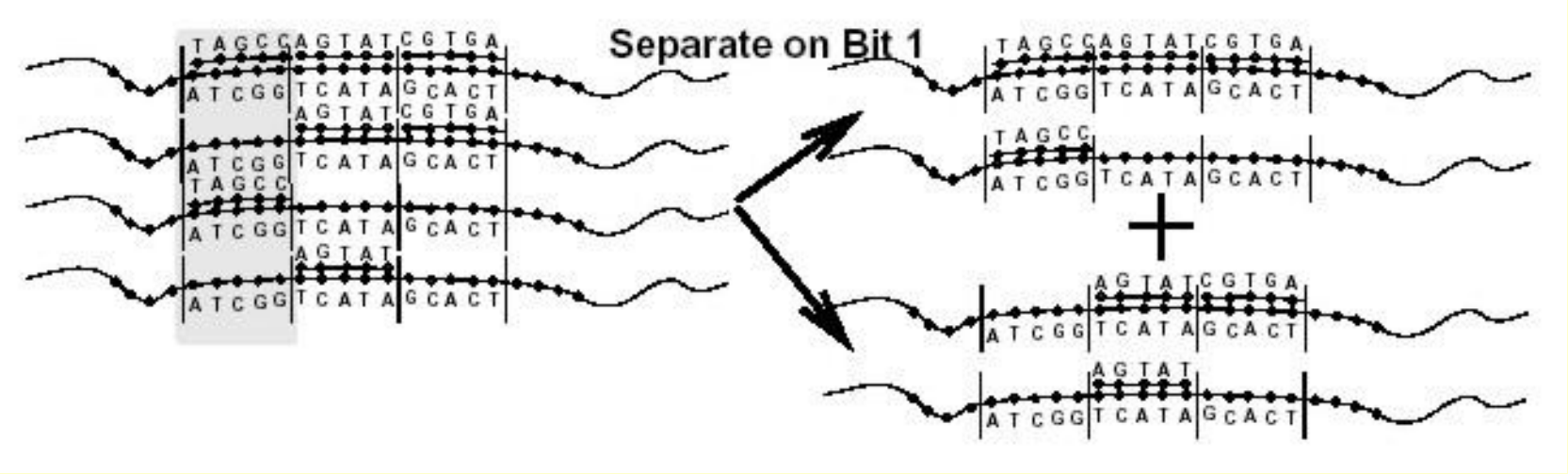
Operacja łączenia

Łączenie probówek zawierających różne łańcuchy:



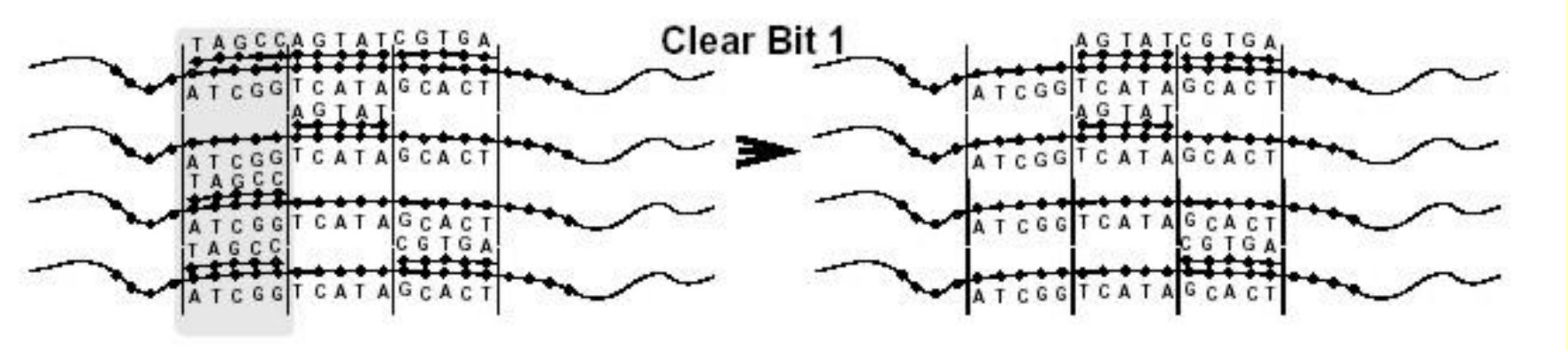
Separacja względem bitu

Rozdzielenie łańcuchów (na dwie klasy) względem wartości określonego bitu:



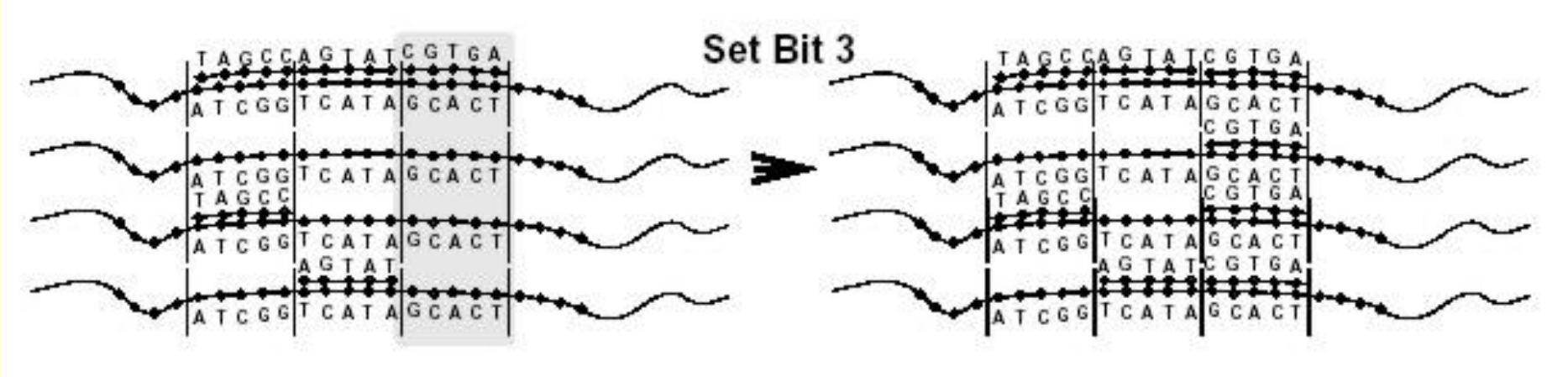
Zerowanie bitu

Wyzerowanie określonego bitu = usunięcie fragmentu ciągu komplementarnego:

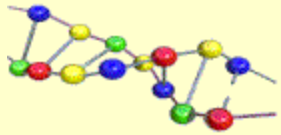


Ustawienie bitu

Ustawienie określonego bitu = dołączenie ciągu komplementarnego do danego bitu:

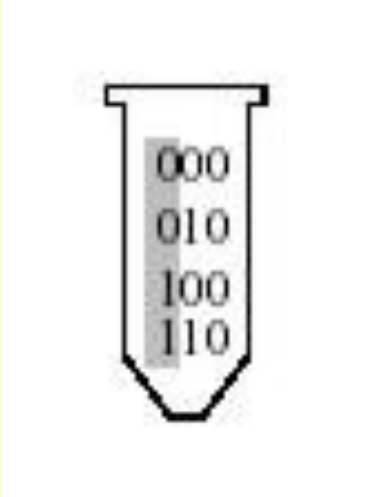


Problem XOR



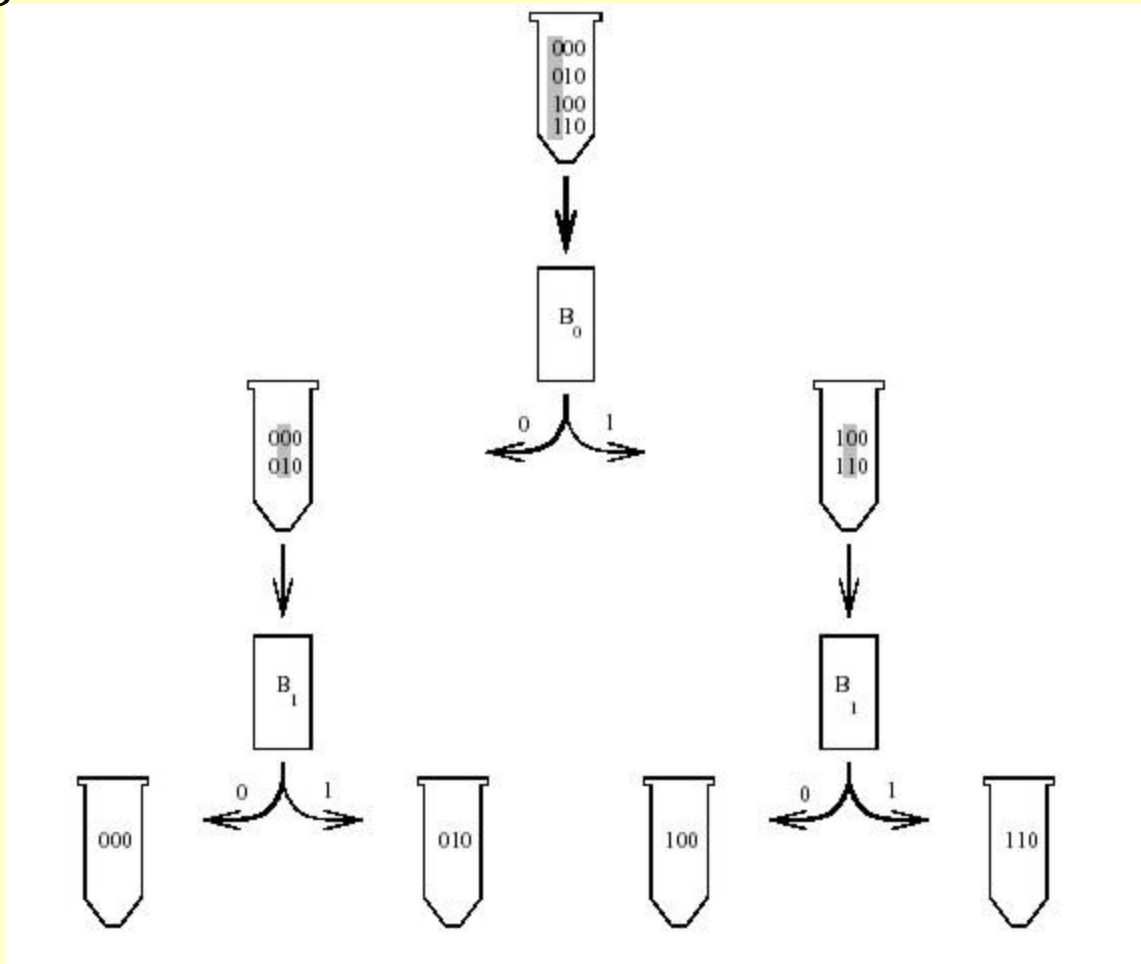
Wykorzystywane m. in. w rozwiązaniu molekularnym problemu 3-SAT i problemu łamania szyfru DES.

1. Wygeneruj wszystkie możliwe kombinacje dwóch bitów z wyzerowanym trzecim bitem (na którym będzie zapisany wynik)



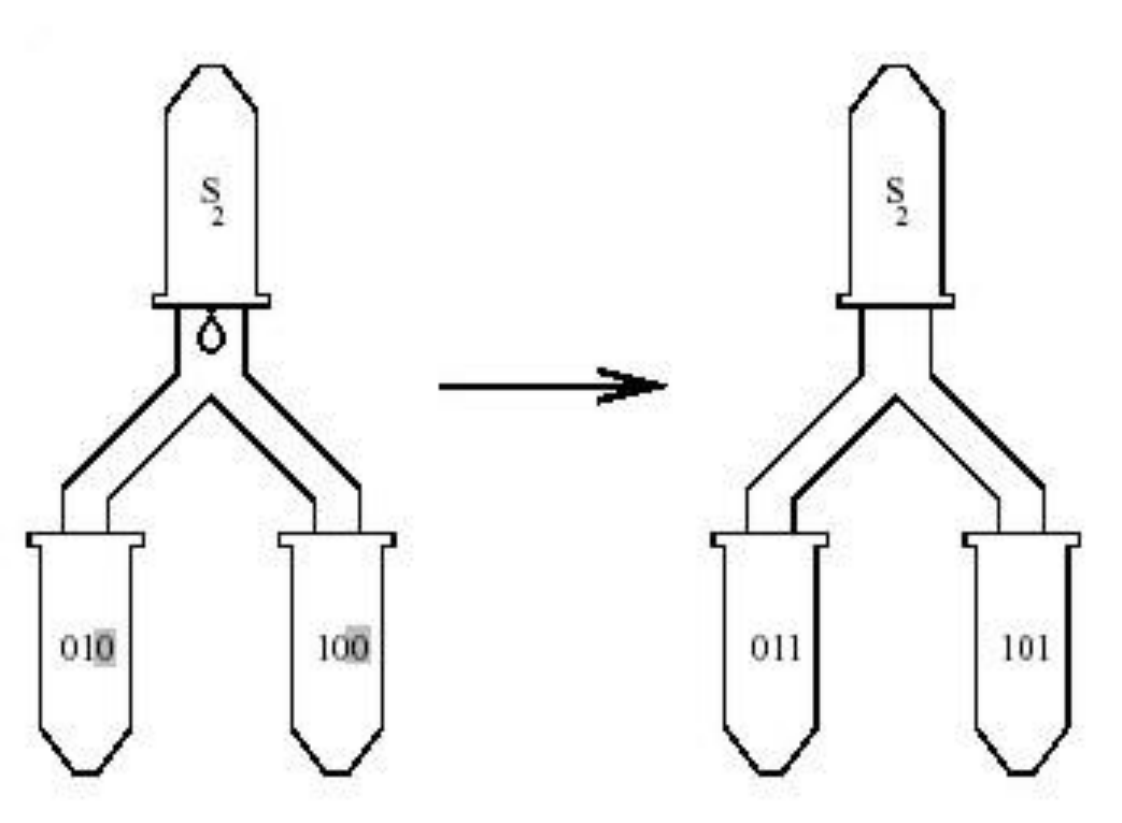
Problem XOR

2. Rozdziel probówkę kolejno ze względu na wartość pierwszego i drugiego bitu:

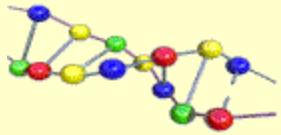


Problem XOR

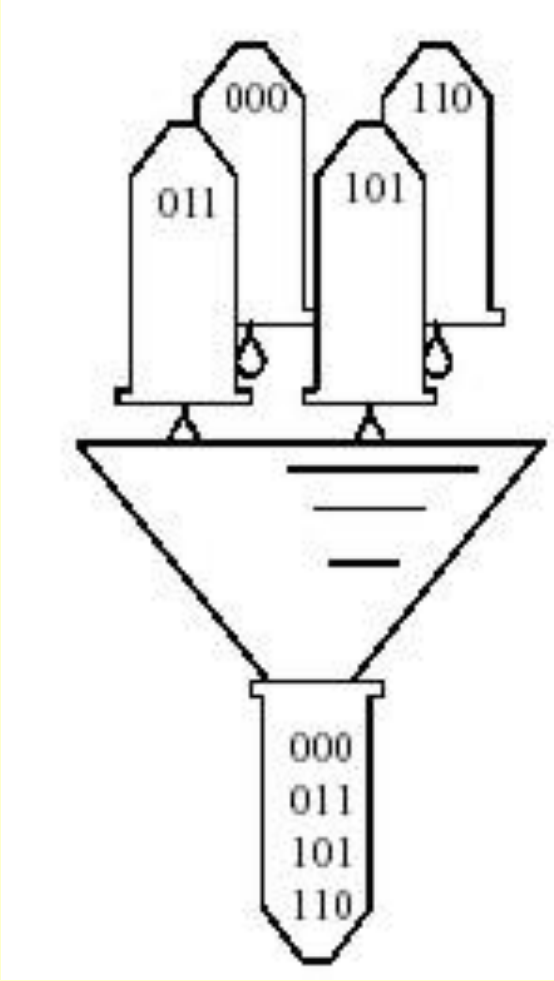
3. Ustaw bit wynikowy w dwóch probówkach, których wynikiem ma być 1 (probówki 2 i 3) ←→ dodaj ciąg komplementarny do ostatniego bitu



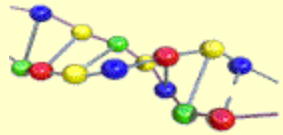
Problem XOR



4. Połącz probówki 1 i 4 z kroku 2 oraz 2 i 3 z kroku 3 (z ustawionym ostatnim bitem)



Obliczenia molekularne - wnioski

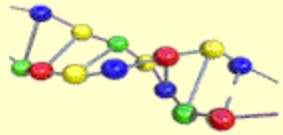


Wysoki stopień komplikacji technologicznej

Na obecnym etapie rozwoju obliczeń DNA komputery molekularne nie są jeszcze w stanie zagrozić tradycyjnym komputerom. Przede wszystkim z powodu skomplikowanego i pracochłonnego sposobu generowania i odczytywania rozwiązania.

DNA Computing nie zmniejsza złożoności obliczeniowej rozwiązań. Jego zaletą jest masowa równoległość.

Obliczenia molekularne - przyszłość

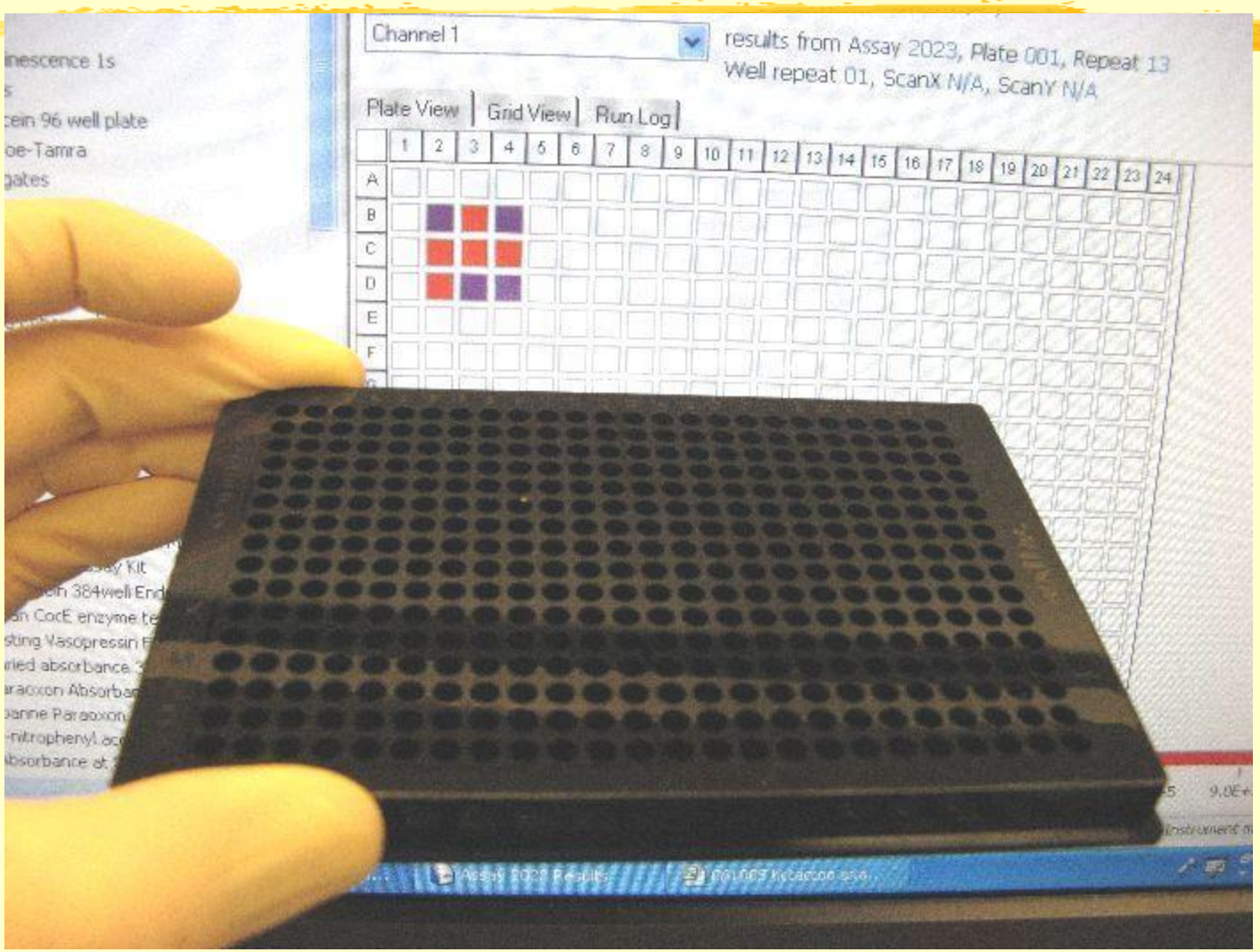


W miarę rozwoju technologii można się spodziewać bardziej wyrafinowanych algorytmów.

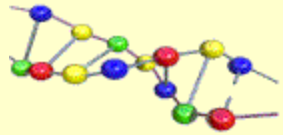
Jest duża szansa na znaczny postęp w metodach manipulacji ciągami DNA → powstają szybsze, mniej pracochłonne i dokładniejsze (pewniejsze) metody.

W takiej sytuacji, komputery DNA będą miały duże szanse na zastosowanie w dziedzinie kryptografii, programowania genetycznego, NLP, routingu, itp. z uwagi na szybkość obliczeń i ich masową równoległość.

MAYA II – komputer DNA (t-t-t)

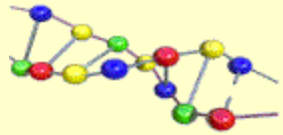


MAYA II – komputer DNA (t-t-t)



- Naukowcy z Columbia University oraz University of New Mexico
- Macierz 100 molekularnych YES / NO bramek logicznych
- Potrafi zagrać prawidłowo grę w KIK przeciwko człowiekowi i praktycznie gra optymalnie, ale ...
- MAYA II zawsze zaczyna i zawsze stawia pierwszy znak w środku
- Jeden ruch zajmuje ok. 30 minut
- Ruch człowieka wskazywany jest poprzez „dodanie” do systemu odpowiedniej nici DNS reprezentującej jedno z 8 pól zawnętrznych

Obliczenia molekularne – *quo vadis?*



I believe things like DNA computing will eventually lead the way to a "molecular revolution," which ultimately will have a very dramatic effect on the world. – L. Adleman