

wykład 9

Sekwencjonowanie DNA

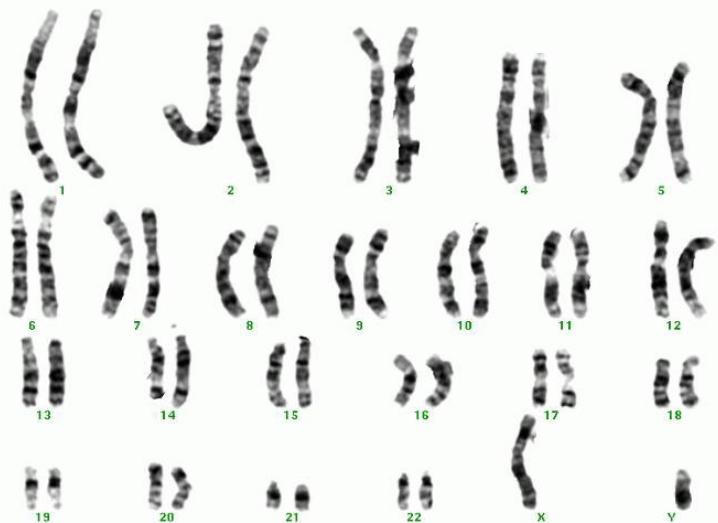
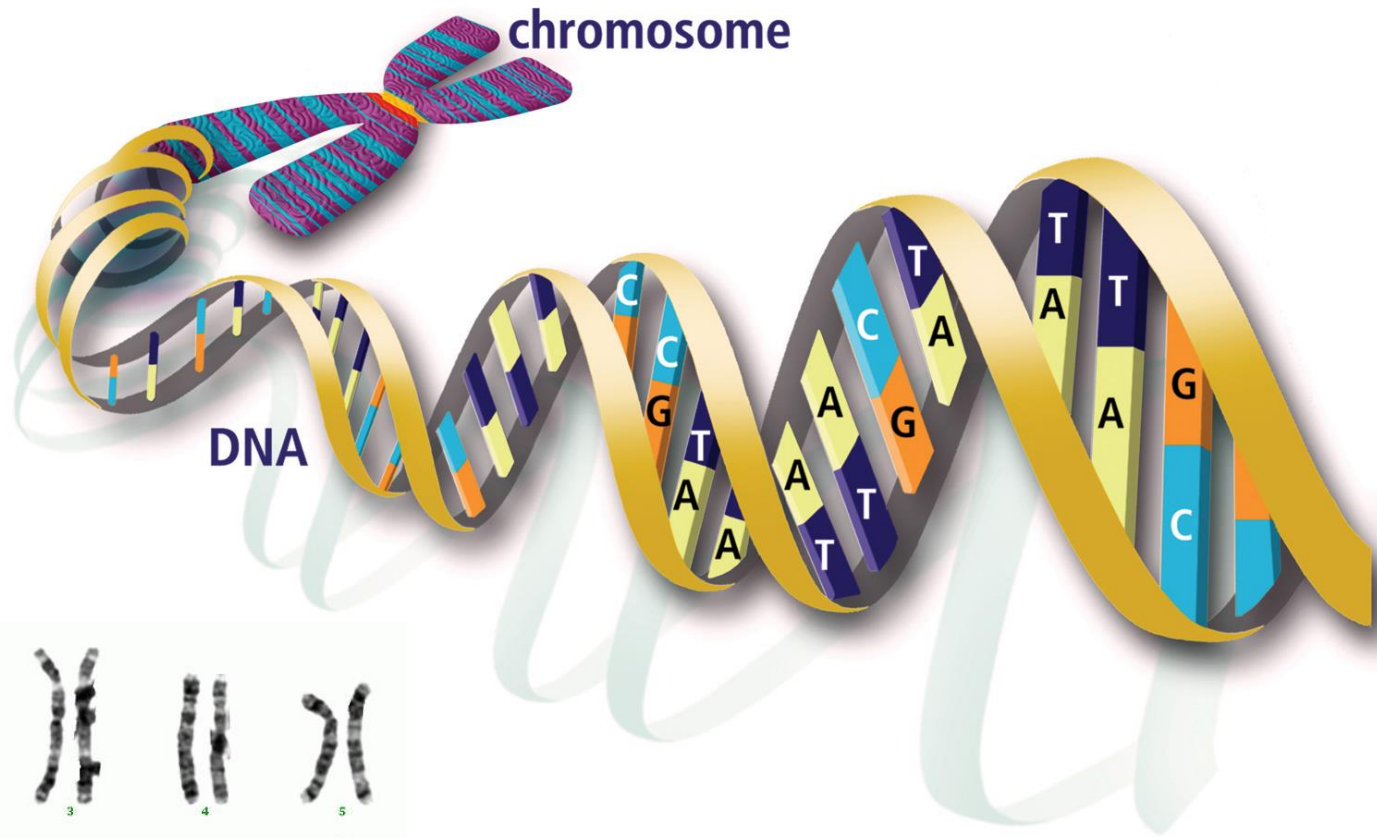
dr Jacek Śmietański

jacek.smietanski@ii.uj.edu.pl

<http://jaceksmietanski.net>

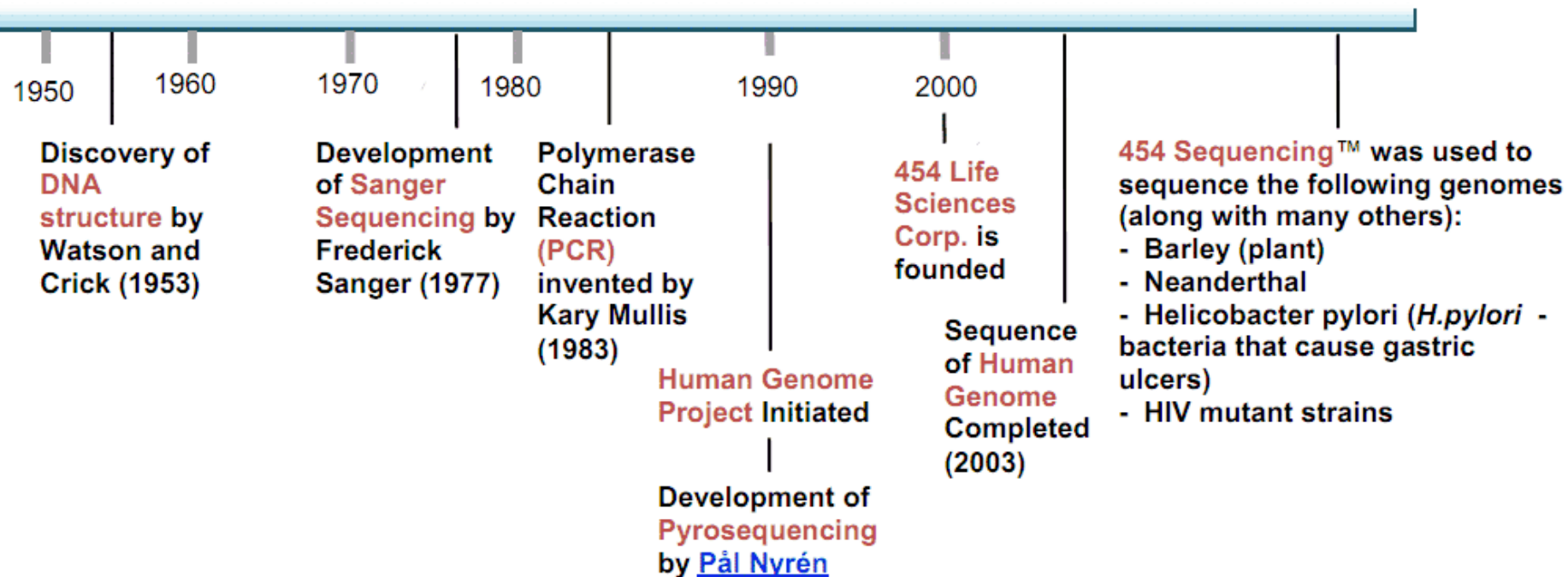
1. Techniki sekwencjonowania
2. Problemy bioinformatyczne
3. Szkic algorytmów składania sekwencji
4. Przechowywanie danych

Techniki sekwencjonowania



Źródło: http://genomics.energy.gov/gallery/basic_genomics/detail.np/detail-16.html

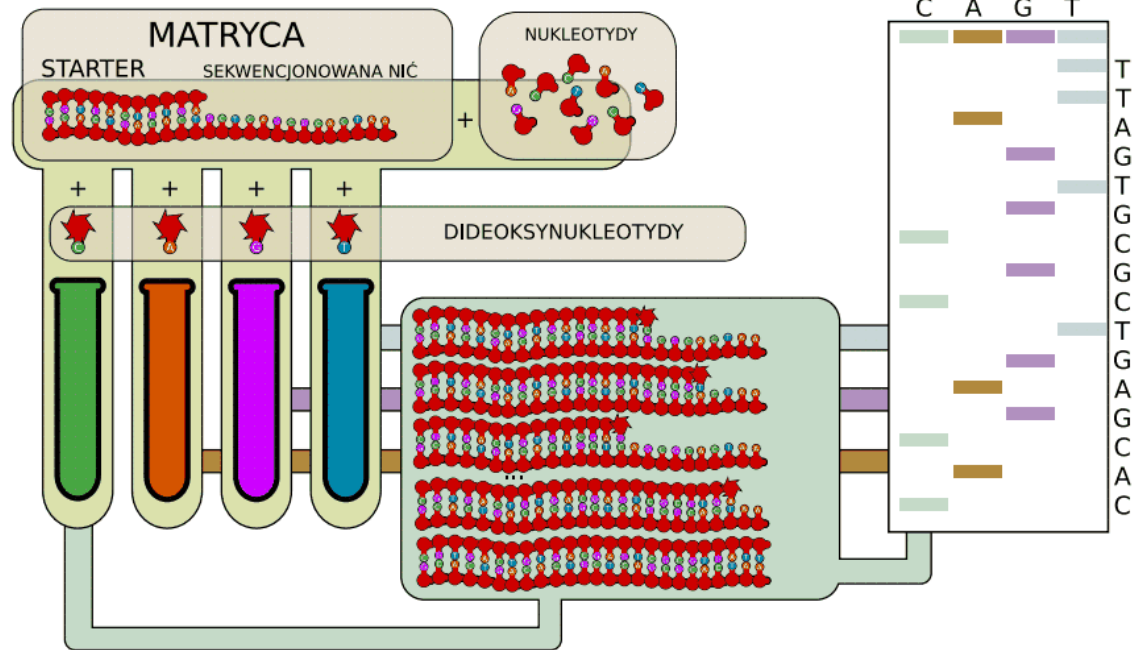
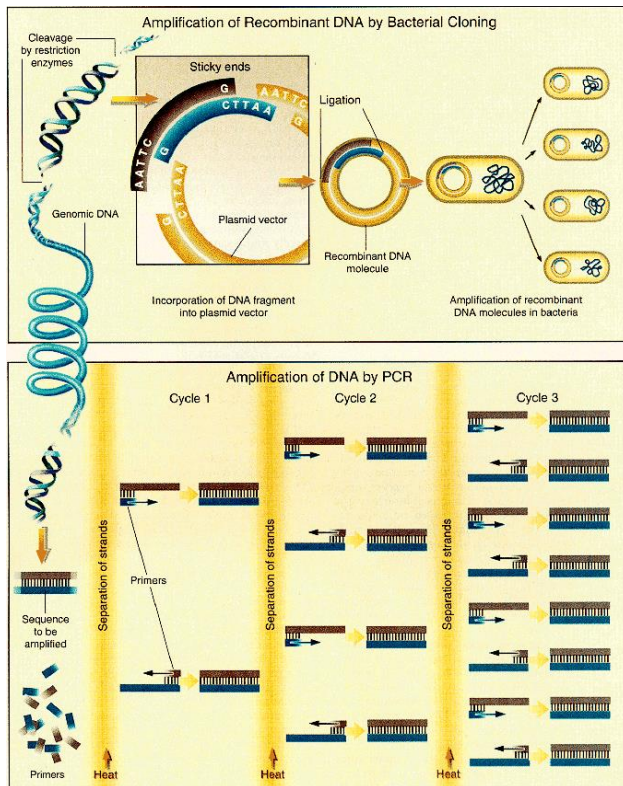
Rozwój metod sekwencjonowania



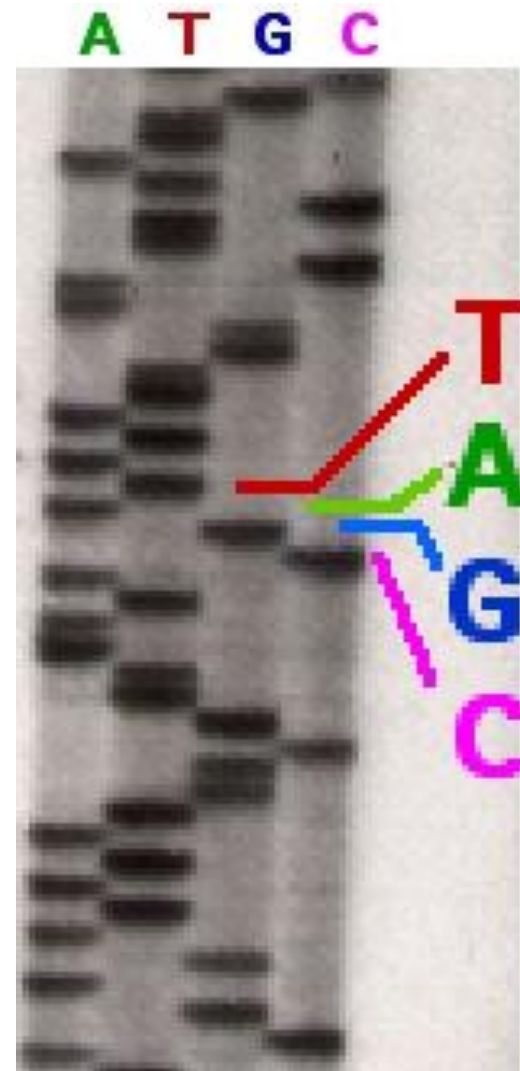
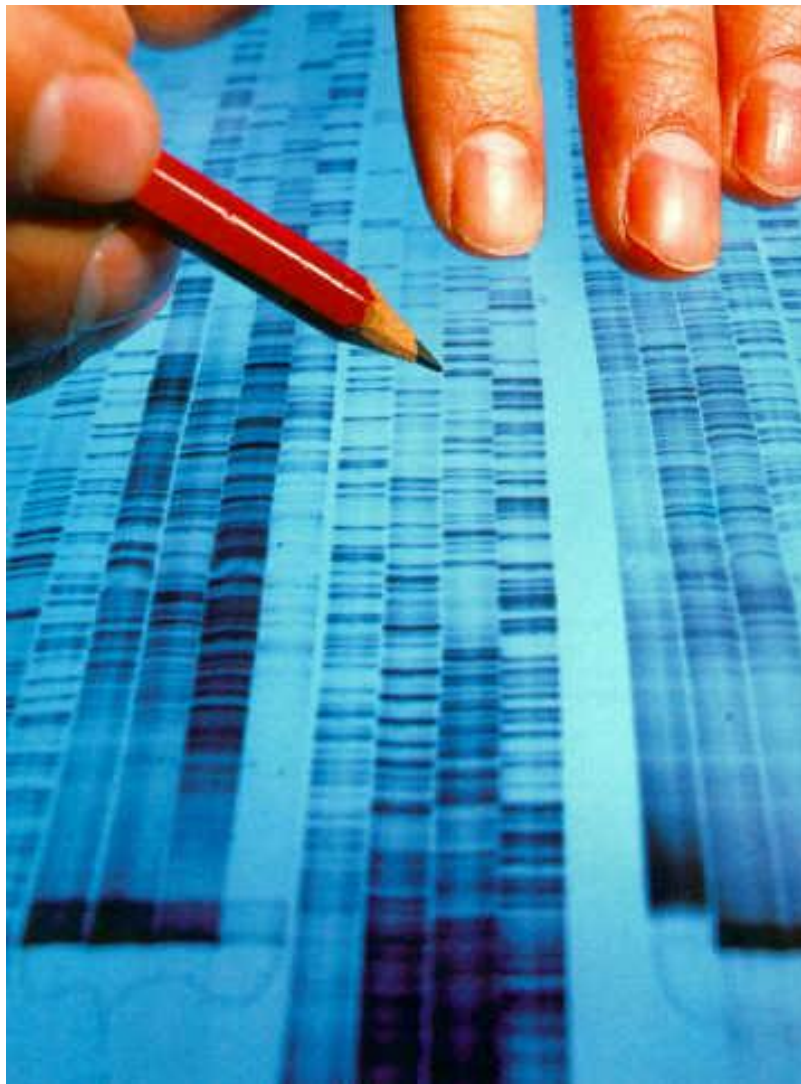
1977 – Sanger, metoda syntezy i DD-terminacji DNA;
1977 – Maxam i Gilbert, metoda chemicznej degradacji;
1981 – Sanger, Shotgun (hierarchiczny);
1987 – automatyzacja procesu;
1995 – Venter, Shotgun całego genomu;
2005 – sekwencjonowanie nowej generacji;
2014 – trzecia generacja

Proces sekwencjonowania

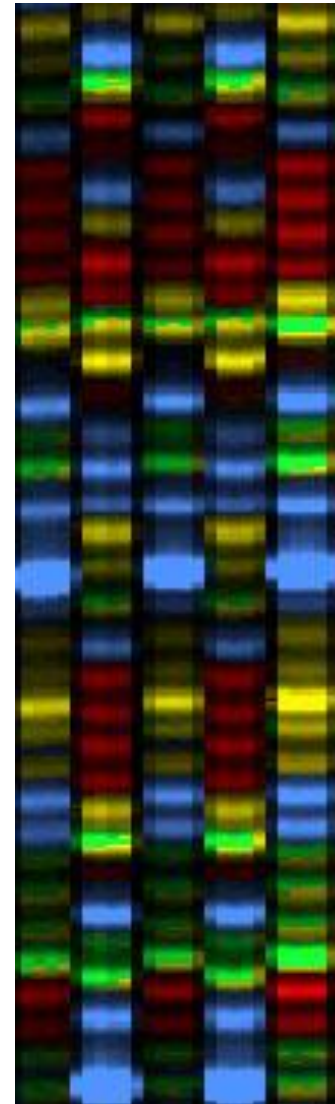
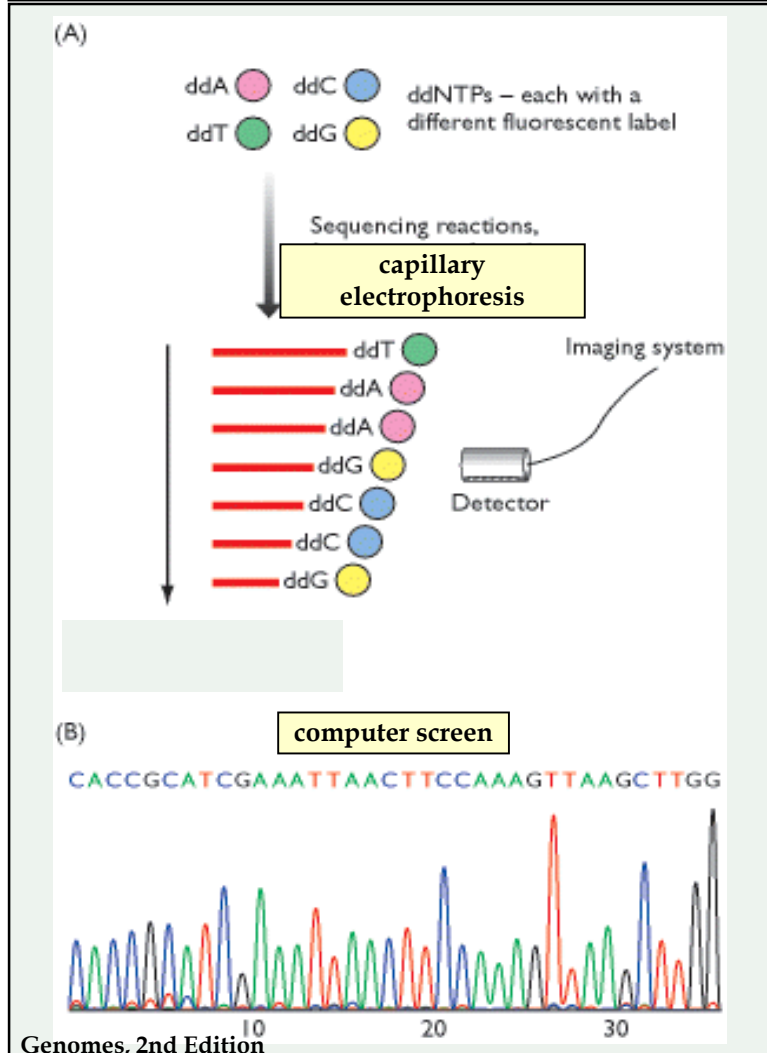
1. Zwiokrotnienie (amplifikacja) badanej próbki
2. Reakcje chemiczne pozwalające na wyznakowanie poszczególnych nukleotydów
3. Rozdział na żelu (elektroforeza) lub odczyt czytnikiem fluorescencji



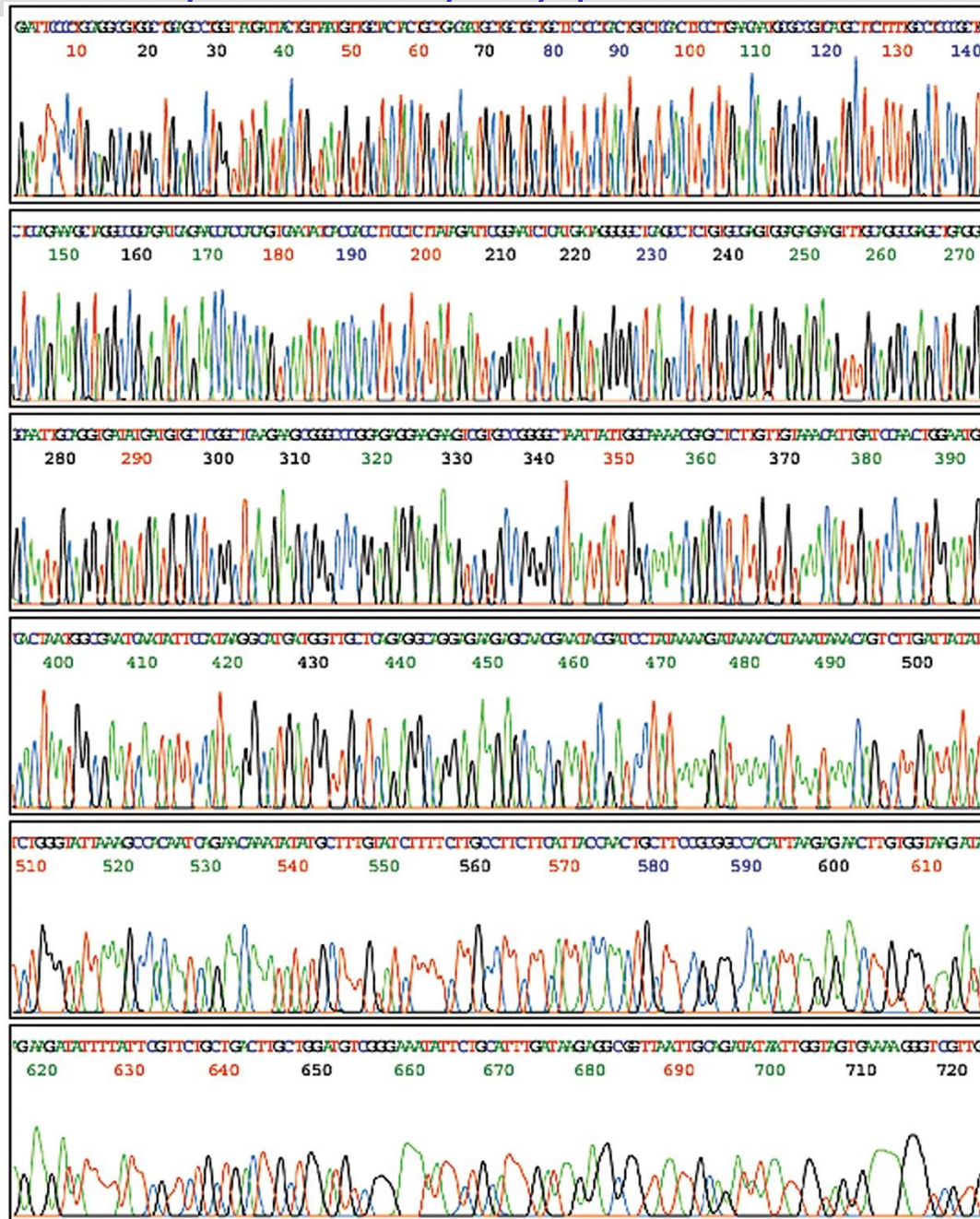
Odczytywanie sekwencji – ręczne



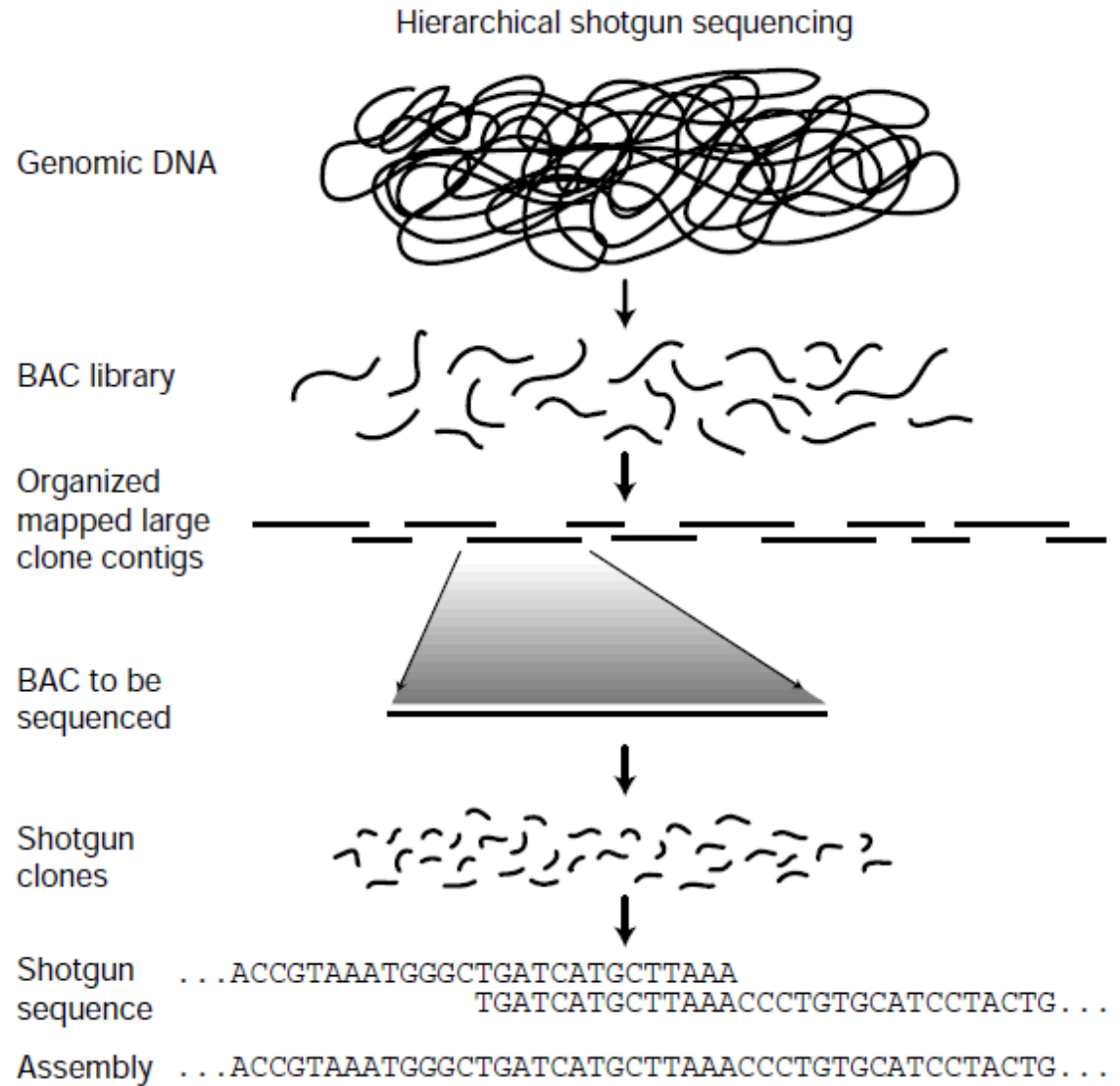
Fluorescent dyes (1986 - ...)



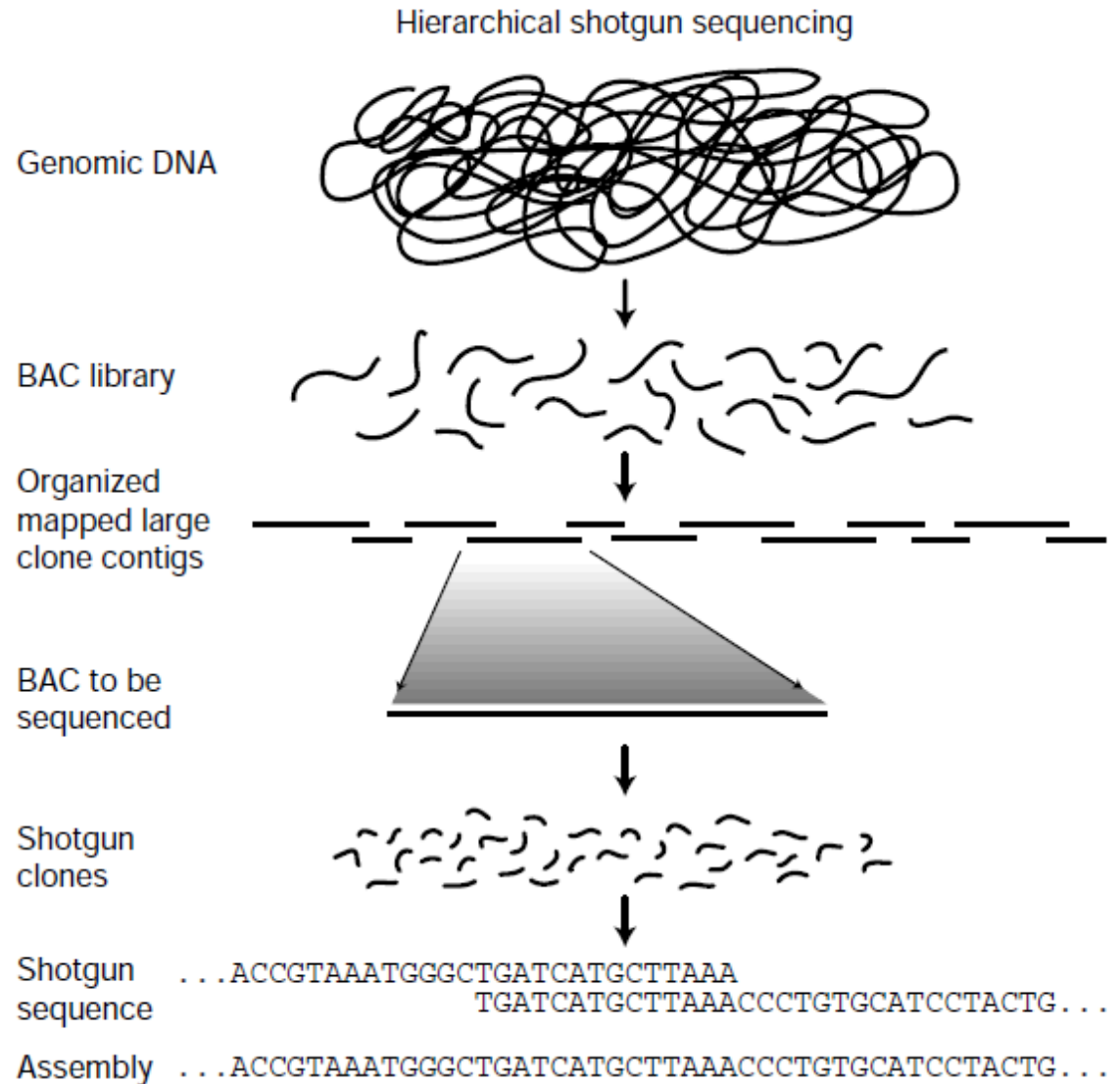
Odczytywanie sekwencji – detekcja sygnału



namnożone
fragmenty DNA są
losowo dzielone na
mniejsze odcinki,
każdy z nich
sekwencjonowany
jest oddzielnie



poszczególne
fragmenty
częściowo
zachodzą na siebie,
co umożliwia
odtworzenie pełnej
sekwencji



Sekwencjonowanie całego genomu – czy to jest możliwe?

rok	liczba zmapowanych genów	przewidywany czas potrzebny do zsekwencjonowania całego genomu
1970	-	niemożliwe
1980	3	~4 mln lat
1990	12	~1000 lat
2000	~25000	wersja robocza
2005	~30000	nowa wersja robocza
2007	31,784 / 30,384	wyzwanie „\$1000 genome”

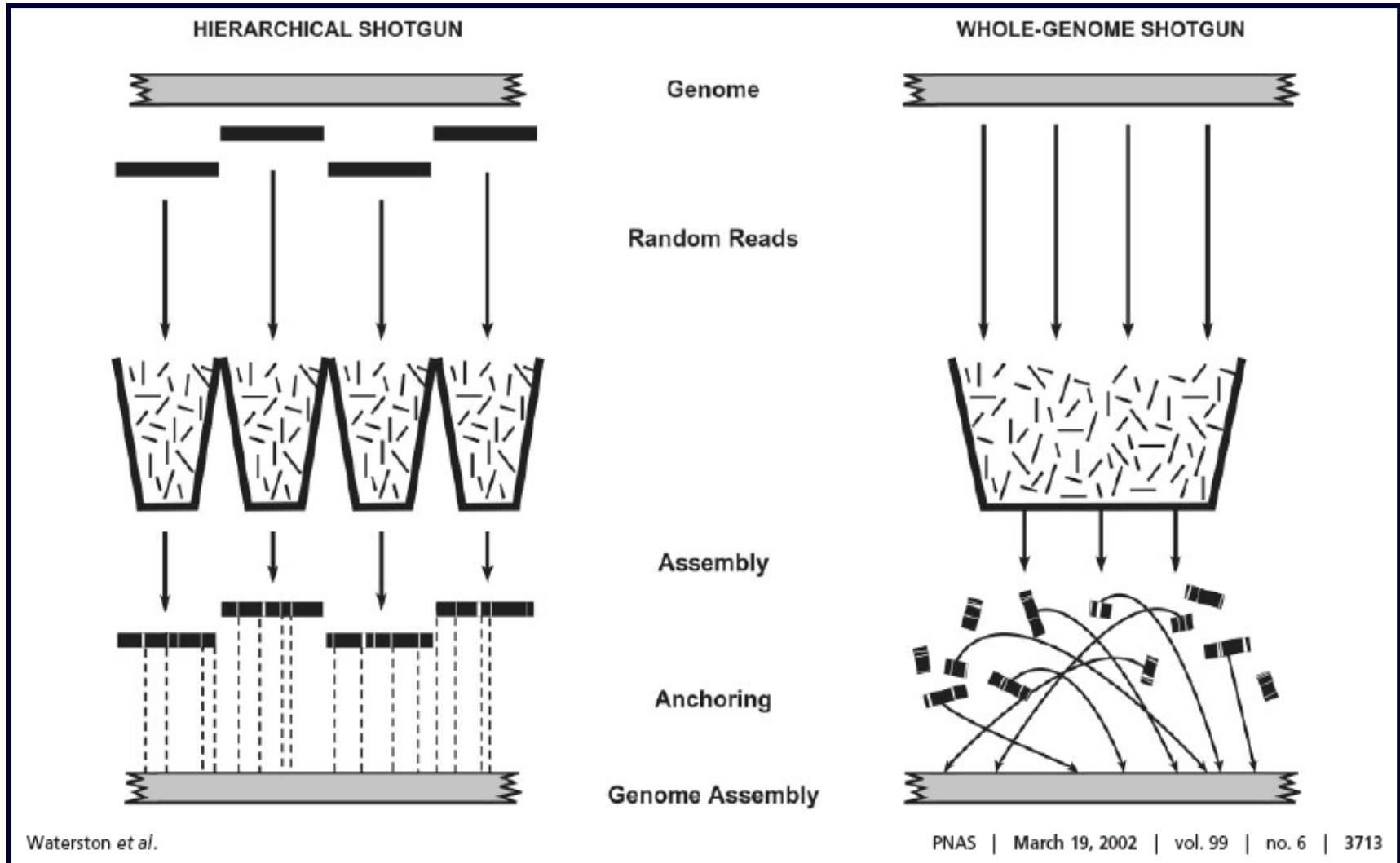
Źródło: www.cbs.dtu.dk/phdcourse/cookbooks/27Apr_1_Genomics.ppt



Human Genome Project

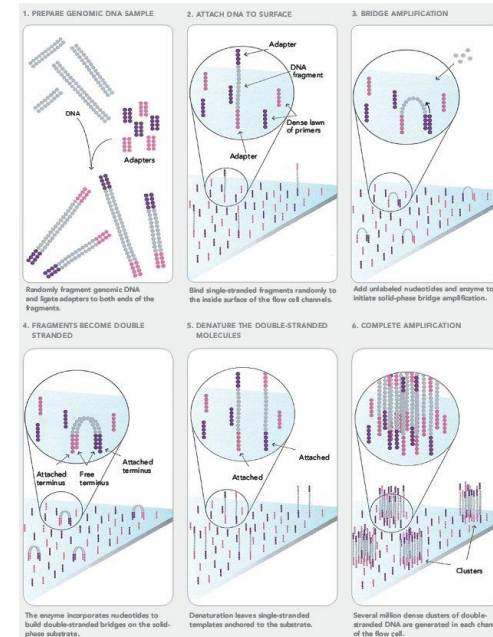
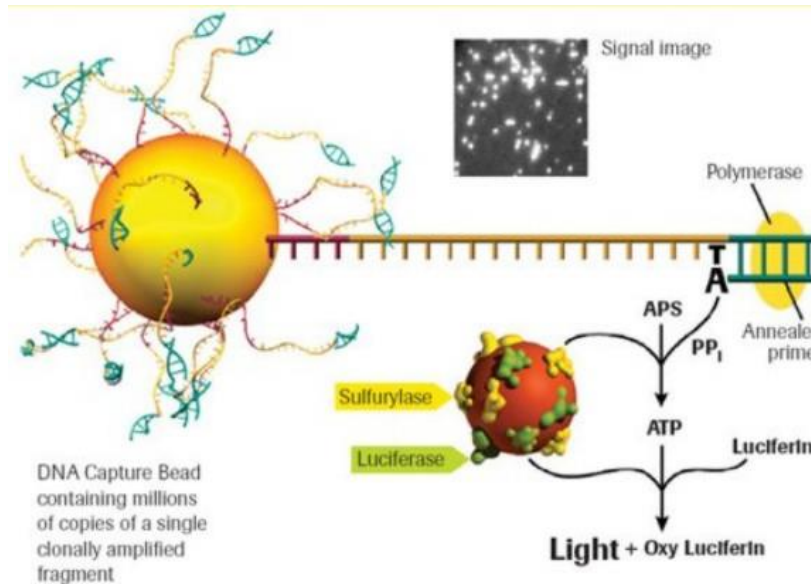
"Genome race"

Celera Genomics (C.Venter)



Nowe techniki sekwencjonowania (*next generation sequencing*)

1. Technika 454 – pirosekwencjonowanie (na matrycy badanej nici syntetyzowana jest druga nić, co powoduje emisję kwantów światła)



2. ION Torrent – mierzona jest zmiana pH wywołana wbudowaniem nukleotydu

3. Illumina (Solexa) – sekwencjonowanie w oparciu o znakowane nukleotydy

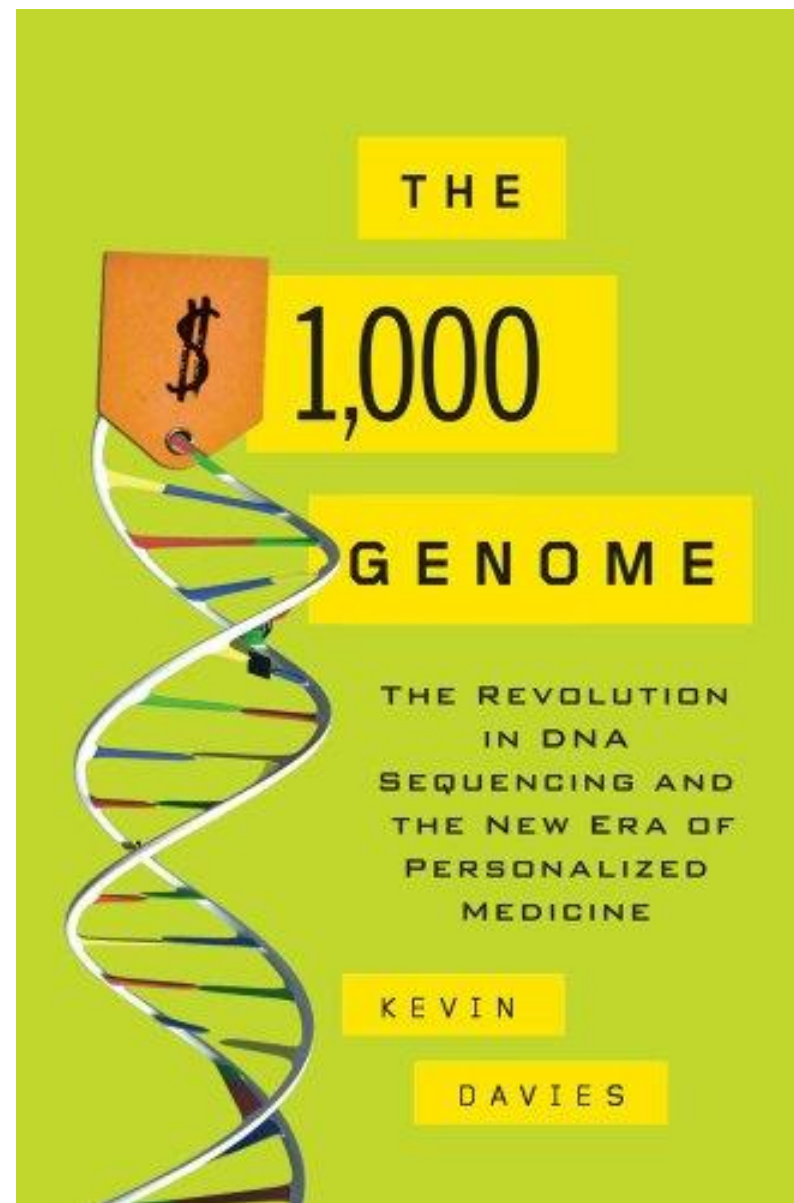
Projekt Sekwencjonowania
Genomu Ludzkiego (*HGP*)
kosztował 3 mld \$

Czy można go zmniejszyć do
1 tys \$?

Wyzwanie zostało zrealizowane
w 2014 roku

Illumina HiSeq X Ten System

(koszt odczynników: \$797,
amortyzacja: \$137,
przygotowanie próbki: \$55–\$65)



Sekwencjonowanie pojedynczych cząsteczek DNA

- polimeraza DNA syntetyzuje pojedynczą nić DNA ze znakowanych fluorescencyjnie dNTP, sygnał z pojedynczych cząsteczek odczytywany jest w czasie rzeczywistym (Pacific Biosciences)
- egzonukleaza odcina zasady z cząsteczki DNA, w trakcie ich przejścia przez por w błonie lipidowej utworzony przez białko hemolizynę, odczytywana jest zmiana przewodnictwa elektrycznego, specyficzna dla zasady (Oxford Nanopores)
- w obu technikach miliony sekwencji są odczytywane jednocześnie na mikrochipie

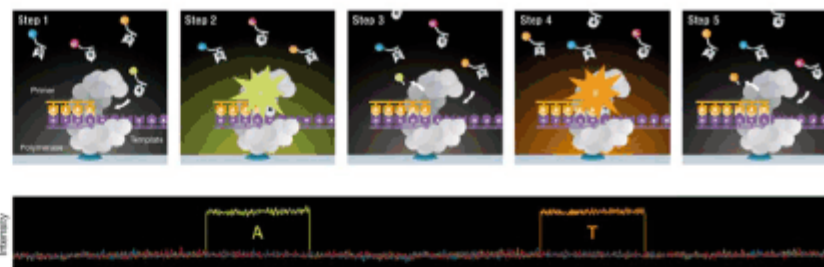
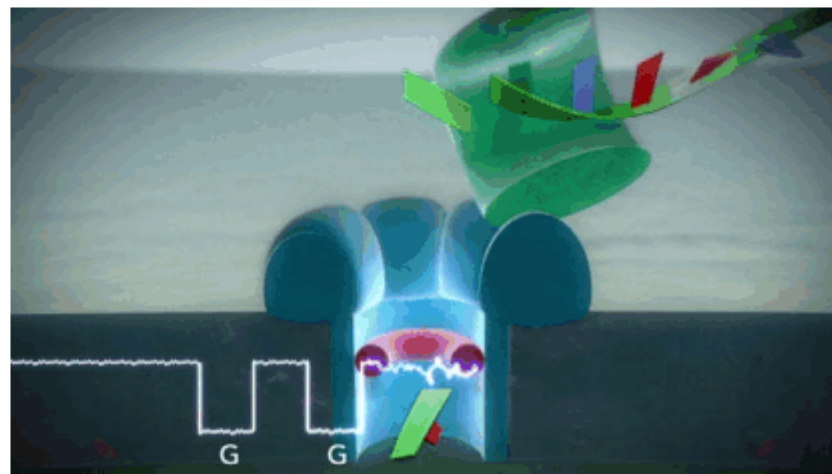


Figure 14. Generation of sequence data.



Wybrane charakterystyki metod sekwencjonowania genomów

Charakterystyka	Sekwencjonowanie WGS						
	HGP* - Sanger	Venter - Sanger	"next generation sequencing methods"				
			454 Titanium	solexa	solid	tSMS	HANS **
Długość pojedynczego odczytu [pz]	750	750	~ 400	~ 35	~ 50	~ 200	~ 100 000
Koszt USD/1pz	1	0,1	< 0,001	<0,001	<0,001	< 0,0001	
Koszt USD/hg*** [mln]	4 000			100 tys.		< 10 000	~ 100
Czas sekwencjonowania genomu ludzkiego	10 lat	1,5 lat	3 mies.				1 godz.
Czas do całkowitego zamknięcia genomu	~15 lat	?					1 godz.

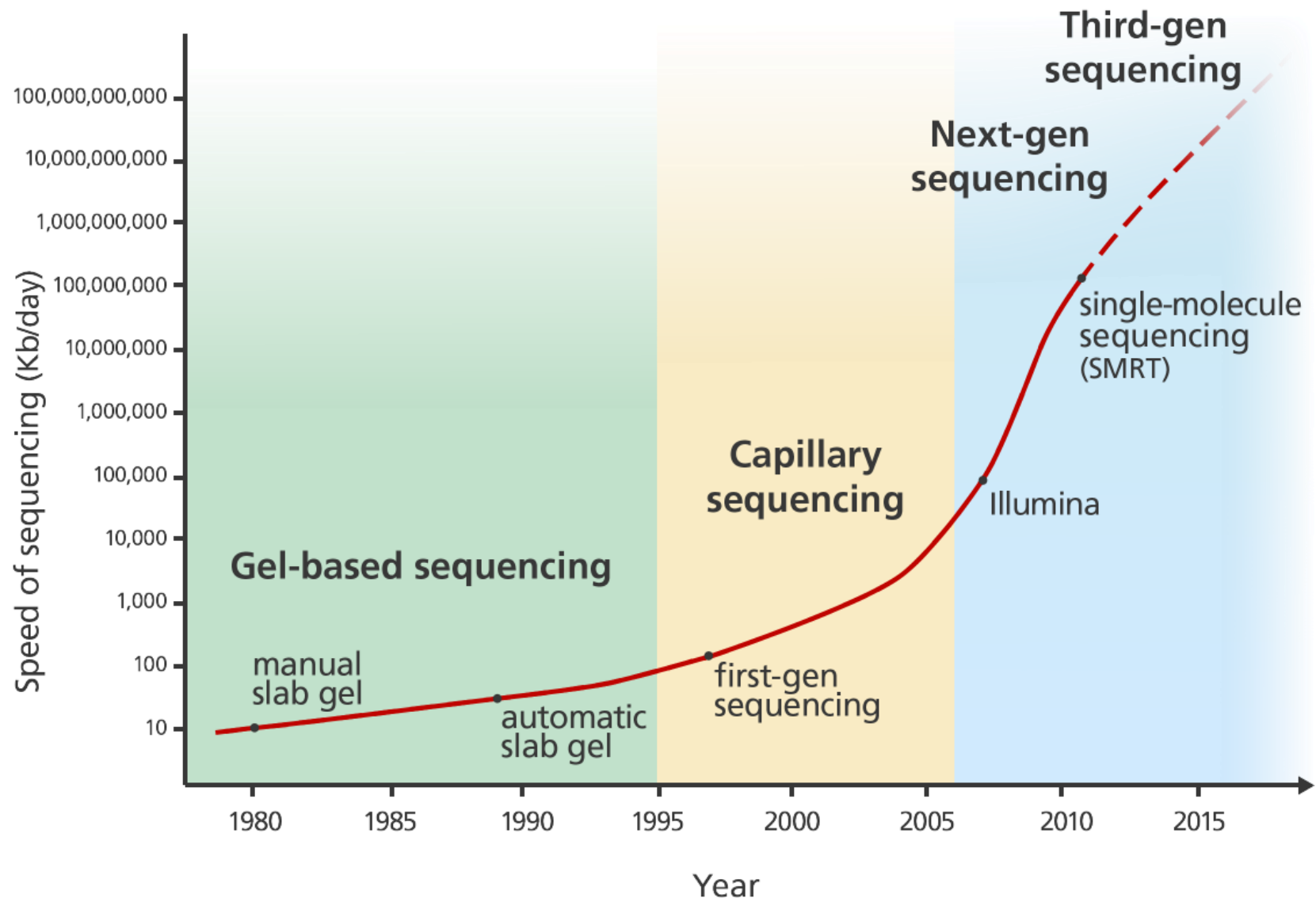
* Human Genome Project; ** Hybridization-Assisted Nanopore Sequencing or Electronic, solid-state DNA sequencing; *** human genome

Searching for Cheaper Genome Sequencers


Company	Format	Read Length (bases)	Expected Throughput MB (million bases)/day
454 Life Sciences	Parallel bead array	100	96
Agencourt Bioscience	Sequencing by ligation	50	200
Applied Biosystems	Capillary electrophoresis	1000	3–4
Microchip Biotechnologies	Parallel bead array	850-1000	7
NimbleGen Systems	Map and survey microarray	30	100
Solexa	Parallel microchip	35	500
LI-COR	Electronic microchip	20,000	14,000
Network Biosystems	Biochip	800+	5
VisiGen Biotechnologies	Single molecule array	NA	1000

Generation next. Companies racing for the \$1000 genome sequence strive simultaneously for low cost, high accuracy, the ability to read long stretches of DNA, and high throughput.

Prędkość sekwencjonowania



Projekt HapMap



International HapMap Project

[Home](#) | [About the Project](#) | [Data](#) | [Publications](#) | [Tutorial](#)

中文 | [English](#) | Français | 日本語 | Yoruba

About the HapMap

- [What is the HapMap?](#)
- [Origins of Haplotypes](#)
- [Health Benefits](#)
- [Populations Sampled](#)
- [Ethical Issues](#)
- [Consent Forms](#)
- [Community Advisory Groups\(CAG\)](#)
- [Data Release Policy](#)
- [Guidelines For Data Use](#)
- [Guidelines For Referring to HapMap Populations](#)

Project Information

- [Home](#)
- [Project Data](#)
- [HapMap Mailing List](#)
- [HapMap Project Participants](#)

Useful Links

- [HapMap Project Press Release](#)
- [NHGRI HapMap Page](#)
- [NCBI Variation Database \(dbSNP\)](#)
- [Japanese SNP Database \(JSNP\)](#)

About the HapMap

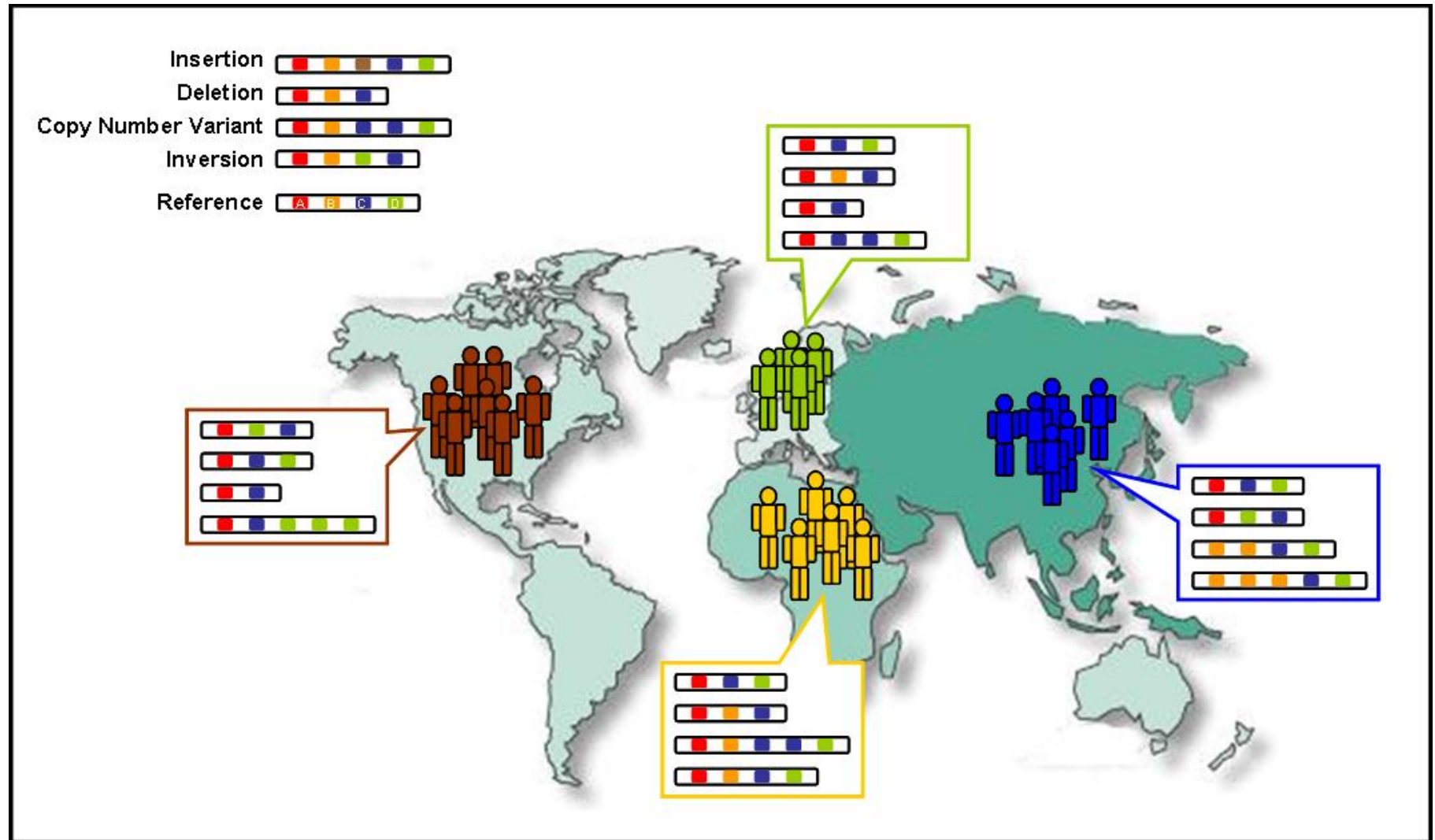
The International HapMap Project is a multi-country effort to identify and catalog genetic similarities and differences in human beings. Using the information in the HapMap, researchers will be able to find genes that affect health, disease, and individual responses to medications and environmental factors. The Project is a collaboration among scientists and funding agencies from Japan, the United Kingdom, Canada, China, Nigeria, and the United States. [See [Participating Groups](#) and [Initial Planning Groups](#).] All of the information generated by the Project will be released into the public domain.

The goal of the International HapMap Project is to compare the genetic sequences of different individuals to identify chromosomal regions where genetic variants are shared. [See [What is the HapMap?](#)] By making this information freely available, the Project will help biomedical researchers find genes involved in disease and responses to therapeutic drugs. [See [How Will the HapMap Benefit Human Health?](#)] In the initial phase of the Project, genetic data are being gathered from **four populations** with African, Asian, and European ancestry. Ongoing interactions with members of these populations are addressing potential **ethical issues** and providing valuable experience in conducting research with identified populations.

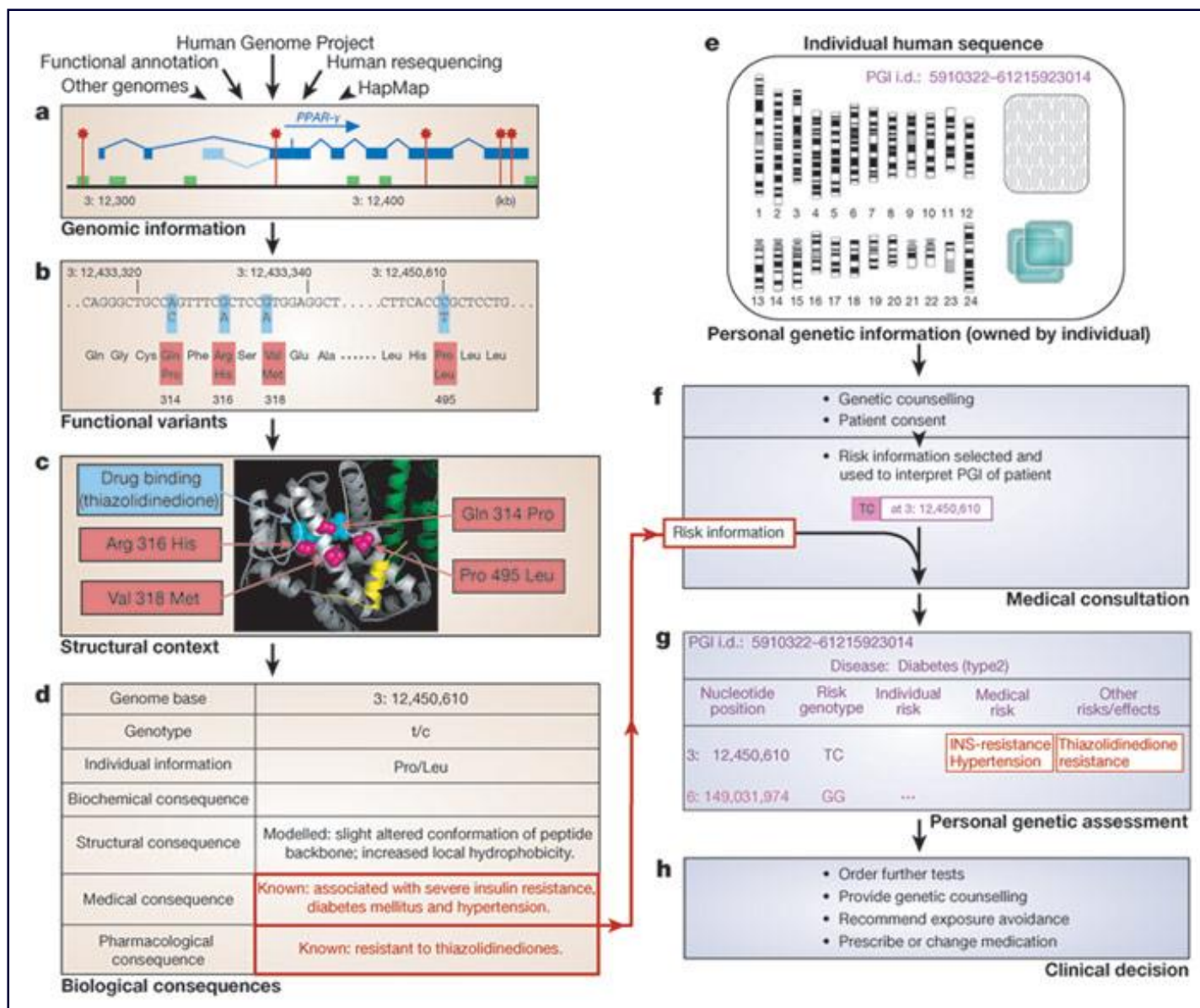
Public and private organizations in six countries are participating in the International HapMap Project. Data generated by the Project can be **downloaded** with minimal constraints. [See [Data Release Policies](#).] The Project officially started with a meeting in October 2002 (<http://genome.gov/10005336>) and is expected to take about three years.

[Home](#) | [About the Project](#) | [Data](#) | [Publications](#) | [Tutorial](#)

Please send questions and comments on website to hapmap-help@ncbi.nlm.nih.gov



Możliwości – medycyna personalizowana



- Określenie ryzyka zachorowania
- Określenie stopnia odporności
- Indywidualna strategia leczenia (oporność / wrażliwość na niektóre leki)

Table 3 | SNPs matching HGMD mutations causing disease or other phenotypes

HGMD accession	Chromosome	Coordinate	HUGO symbol	Gene name	Cytogenetic	Phenotype	Zygosity
CM003589	1	97937679	DPYD	Dihydropyrimidine dehydrogenase	1q22	Dihydropyrimidine dehydrogenase deficiency	Heterozygous
CM950484	1	157441978	FY	Duffy blood-group antigen	1q	Duffy blood group antigen, absence	Homozygous*
CM942034	4	619702	PDE6B	Phosphodiesterase 6B, cGMP-specific, rod, beta	4p16.3	Retinitis pigmentosa 40	Heterozygous
CM021718	9	36208221	GNE	UDP-N-acetylglucosamine 2-epimerase	9p	Myopathy, distal, with rimmed vacuoles	Heterozygous
CM980633	10	50348375	ERCC6	Excision repair cross-complementing rodent repair deficiency, complementation group 6 protein (CSB)	10q	Cockayne syndrome	Homozygous†
CM050716	11	76531431	MYO7A	Myosin VIIA	11q13.5	Usher syndrome 1b	Homozygous†
CM950928	12	46812979	PFKM	Phosphofructokinase, muscle	12q13.3	Glycogen storage disease 7	Homozygous*
CM032029	14	20859880	RPGRIP1	Retinitis pigmentosa GTPase regulator interacting protein 1	14q11	Cone-rod dystrophy	Heterozygous
CM984025	19	18047618	IL12RB1	Interleukin-12 receptor, beta 1	19p13.1	Mycobacterial infection	Heterozygous
CM024138	19	41014441	NPHS1	Nephrosis-1, congenital, Finnish type	19q	Congenital nephrotic syndrome, Finnish type	Heterozygous
CM910052	22	49410905	ARSA	Arylsulphatase A	22q	Metachromatic leukodystrophy	Heterozygous

* Coverage at these SNP positions is less than 5. However, both produce benign phenotypes.

† Coverage at these SNP positions is greater than 5. Both would produce severe phenotypes if

HGMD – Human Gene Mutation Database
<http://www.hgmd.org/>

Problemy bioinformatyczne

1. Zbieranie odczytów
2. Identyfikacja kontigów
3. Łączenie kontigów (*scaffold*)
4. Tworzenie konsensusu
5. Kompletny chromosom
6. Kompletny genom

Etapy posekwencyjne

- lokalizacja genów
- przewidywanie funkcji
- przechowywanie danych

Whole Genome Shotgun Sequencing Method



Genomic DNA



Sequence Each Fragment
with Shotgun Approach

GCATTTTCGAGTTACCTGGACAACCAAGTG

CCAGTGGTACTGAGGACGCAAGAGGCTTGA

GCTTGATTGGCCATAATAGTATAT

Align Contiguous Sequences

GCATTTTCGAGTTACCTGGACAACCAAGTGGTACTGAGGACGCAAGAGGCTTGATTGGCCATAATAGTATAT

Generate Finished Sequence

powtórzenia – genom ludzki zawiera mnóstwo powtarzających się sekwencji, niektóre pojawiają się w genomie więcej niż 100000 razy.

Przykładowo powtórzenie „Alu” ma długość 300 nukleotydów, pojawia się 1000000 razy w ludzkim genomie.

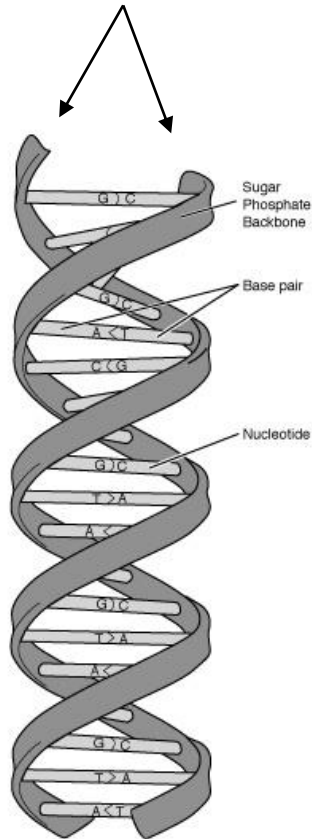


przerwy – niektórych fragmentów DNA nie da się zsekwencjonować.

błędy w sekwencjonowaniu – związane zarówno z ograniczeniami technologicznymi jak i z ludzkimi pomyłkami

nieznana orientacja – sekwencjonujemy DNA dwuniciowe; nie wiadomo, z której nici pochodzi dany odczyt

Fragment może pochodzić
z dowolnego łańcucha



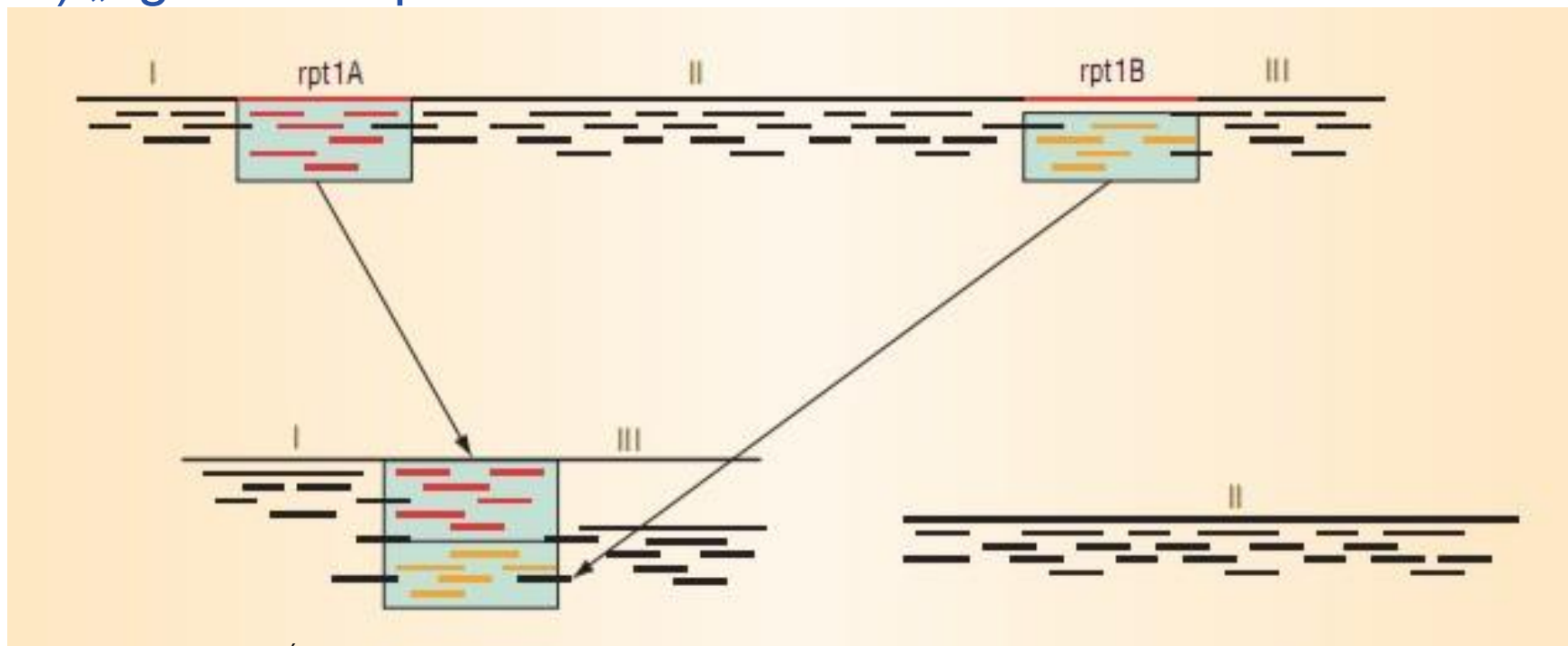
CACGT	→	CACGT
ACGT	→	-ACGT
ACTACG	←	--CGTAGT
GTACT	←	-----AGTAC
ACTGA	→	-----ACTGA
CTGA	→	-----CTGA

jeśli odczyt pochodzi z drugiej nici, musimy składać
sekwencję komplementarną czytaną od końca;

ale tego, z której nici jest dany fragment, nie wiemy

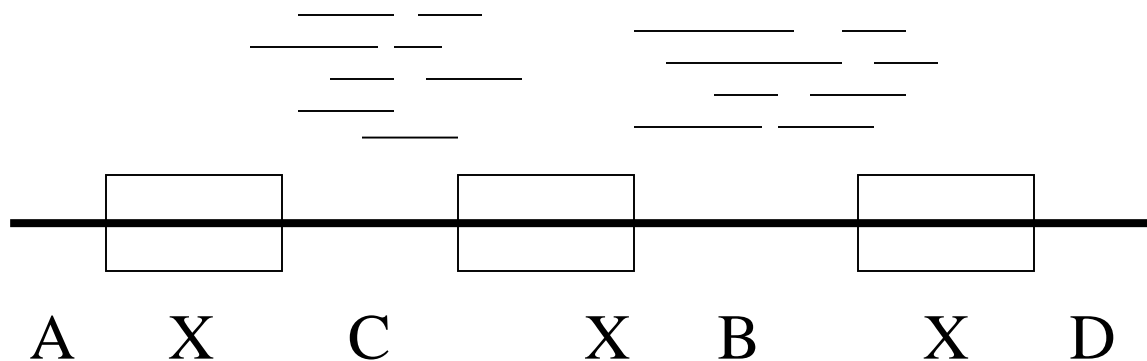
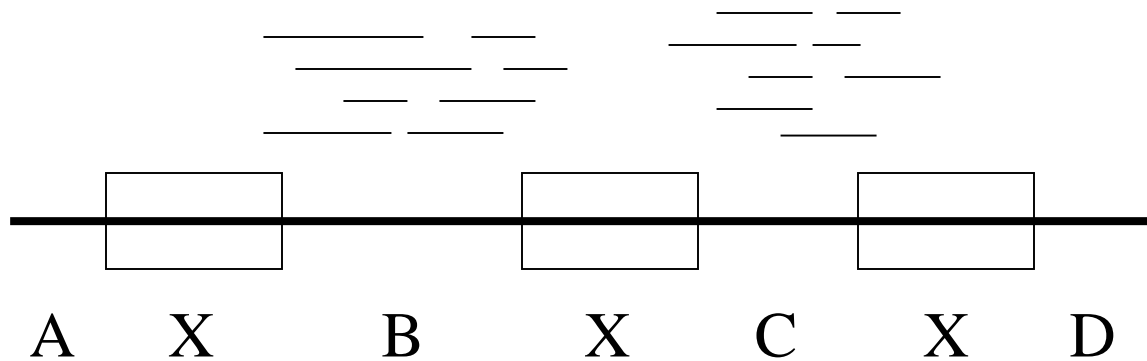
Krótkie powtórzenia nie stanowią dużego problemu – jeśli powtarzająca się sekwencja jest krótsza od długości analizowanego klonu, zapewne uda się właściwie dopasować końce sekwencji. Jednak powtórzenie dłuższe od długości odczytu może być błędnie złożone.

a) „zgubienie” powtórzenia

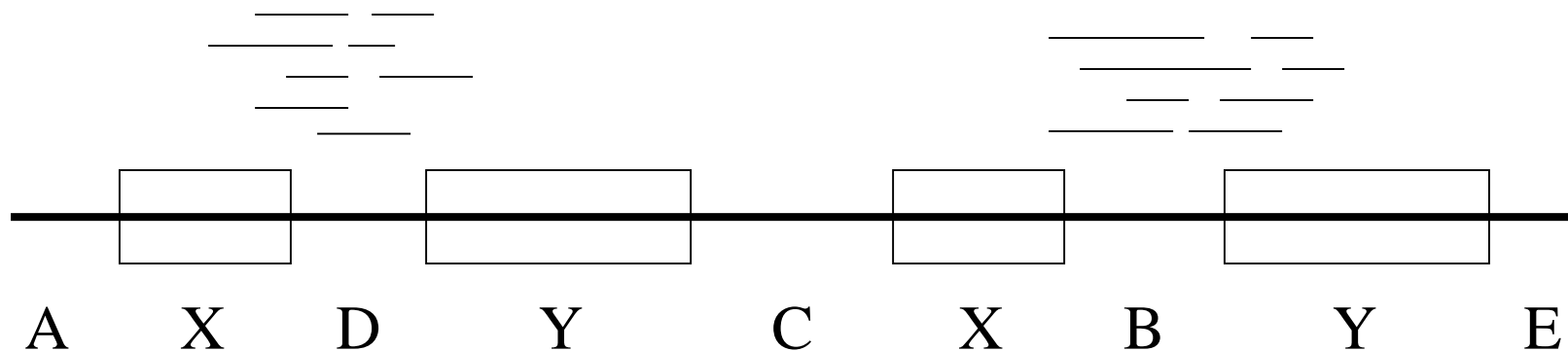
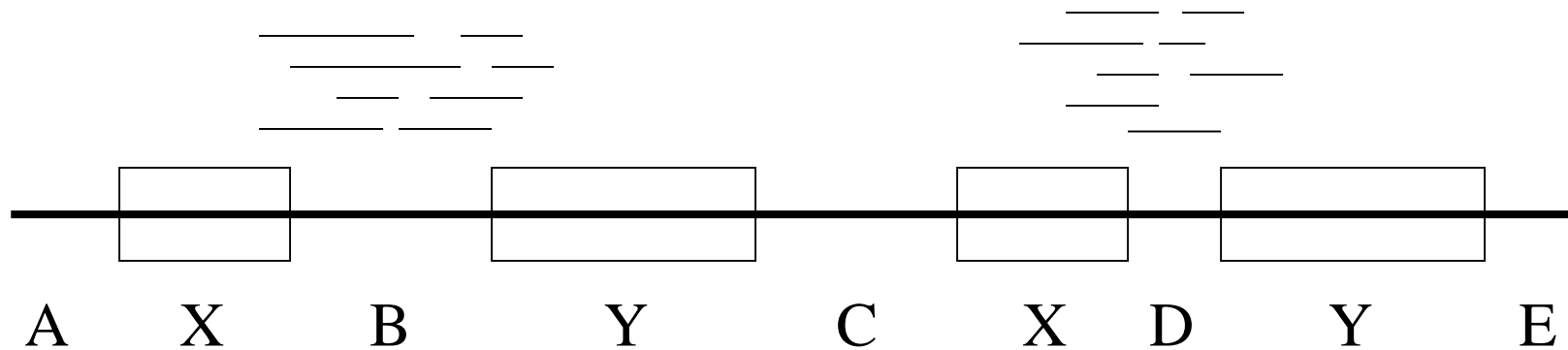


Źródło: Pop, M.; Salzberg, S.L.; Shumway, M.; Genome Sequence Assembly: Algorithms and Issues, 2002, IEEE Computer 35(7):47-54

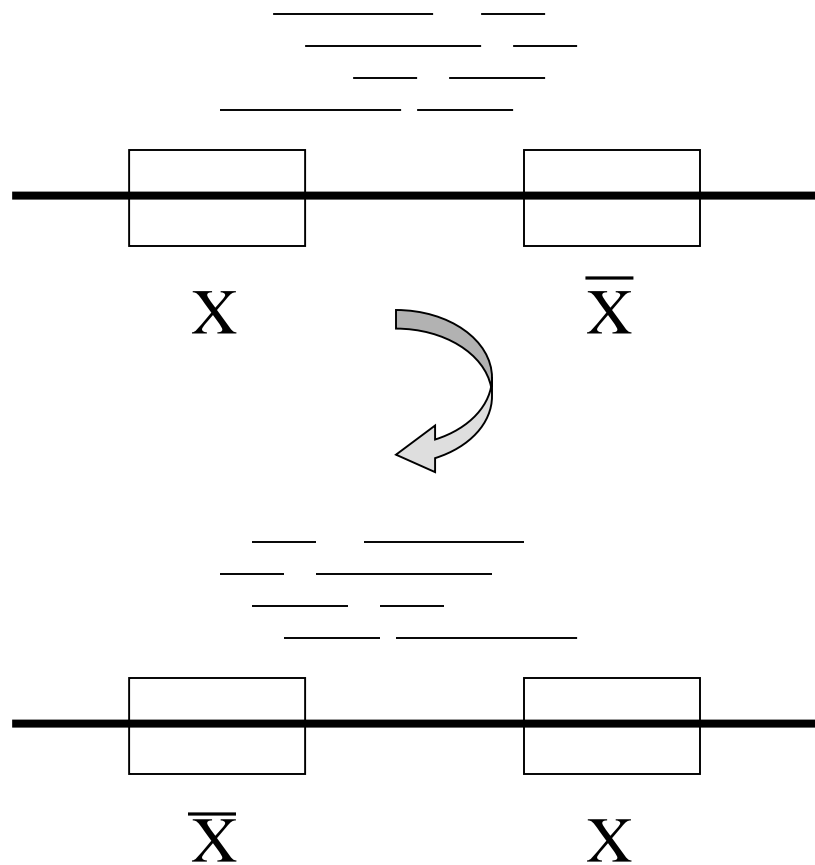
b) zmiana kolejności



b) zmiana kolejności



c) odwrócone powtórzenia



--ACCGT--
----CGTGC
TTAC-----
-T**G**CCGT-
TTACCGTGC

błędny odczyt nukleotydu

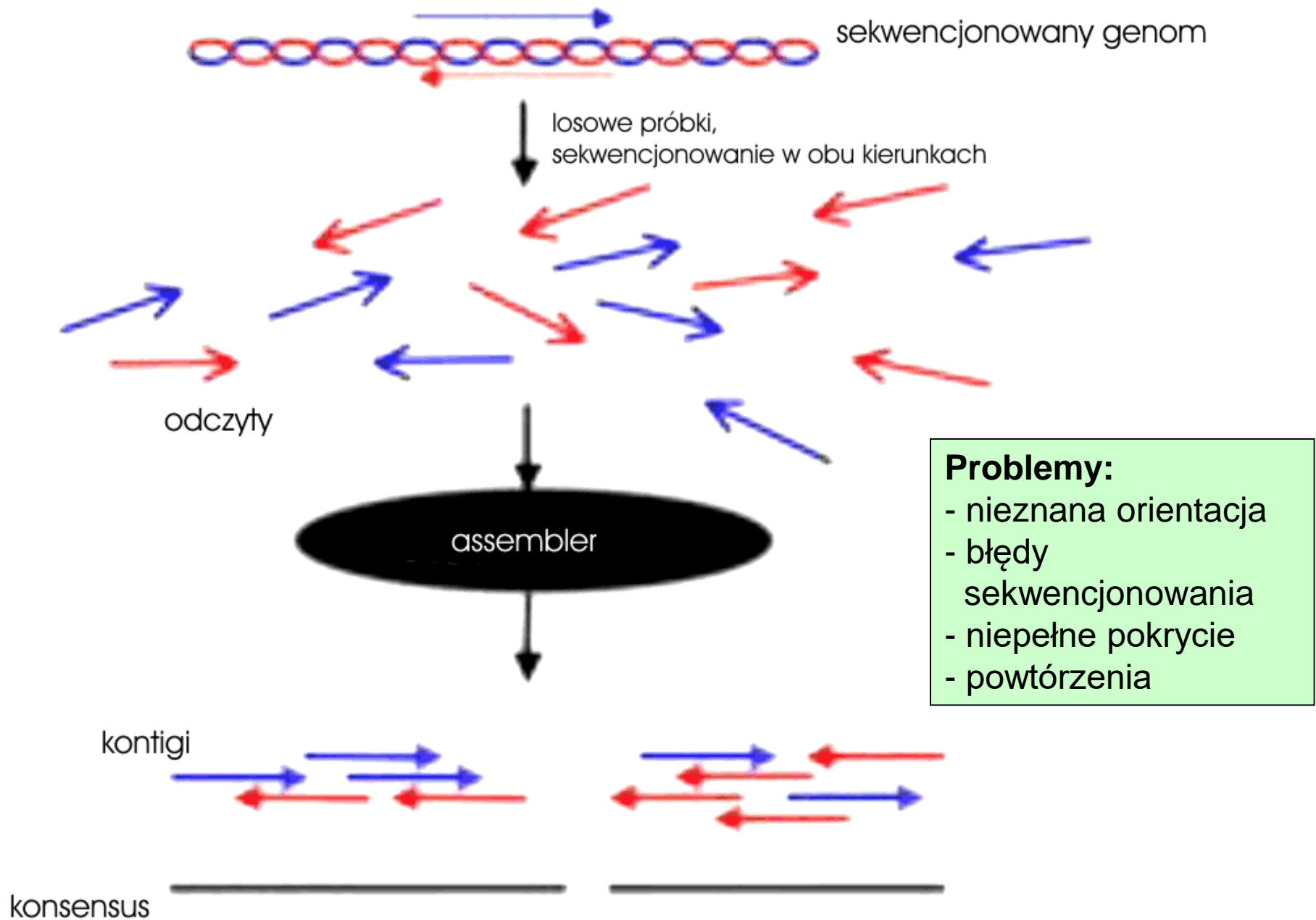
--ACC-GT--
----C**A**GTGC
TTAC-----
-TACC-GT-

TTACC-GTGC

błędna insercja

--ACCGT--
----CGTGC
TTAC-----
-TAC-GT-
TTACCGTGC

błędna delecja



Szkic algorytmów składania sekwencji

Definicja:

Dany jest zbiór słów $P = \{s_1, s_2, \dots, s_n\}$

Najkrótszym wspólnym nadśłowem zbioru P nazywany słowo S , najkrótsze spośród spełniających warunek, że każde słowo $s_i \in P$ jest podśłowem S

Założmy dodatkowo, bez straty ogólności, że słowa z P , nie zawierają się w sobie.

Przykład:

$P = \{fabcc, efab, bccla\}$.

Słowa $bcclabccefabcc$ i $efabccla$ są nadśłowami P .
 $efabccla$ jest najkrótszym nadśłowem P .

Problem jest NP-zupełny.

Klasyczny algorytm dokładny: $O(2^k)$ czas, $O(2^k)$ pamięć
(Held,Karp,1962–TSP).

Zachłanny algorytm aproksymacyjny:

4-aproksymacja (Blumetal.,1991)

3,5-aproksymacja (Kaplan,Shafrir,2005)

2-aproksymacja? (hipoteza,Blumetal.).

Algorytm 2,5-aproksymacyjny

(Breslaueretal.,1997+Kaplanetal.,2005).

ρ -aproksymacja

Algorytm A nazywamy ρ -aproksymacyjnym, jeśli dla dowolnych poprawnych danych wejściowych x , $A(x) \in F(x)$ oraz

$$\max \left\{ \frac{c(A(x))}{c_{OPT}(x)}, \frac{c_{OPT}(x)}{c(A(x))} \right\} \leq \rho, \quad \rho \geq 1$$

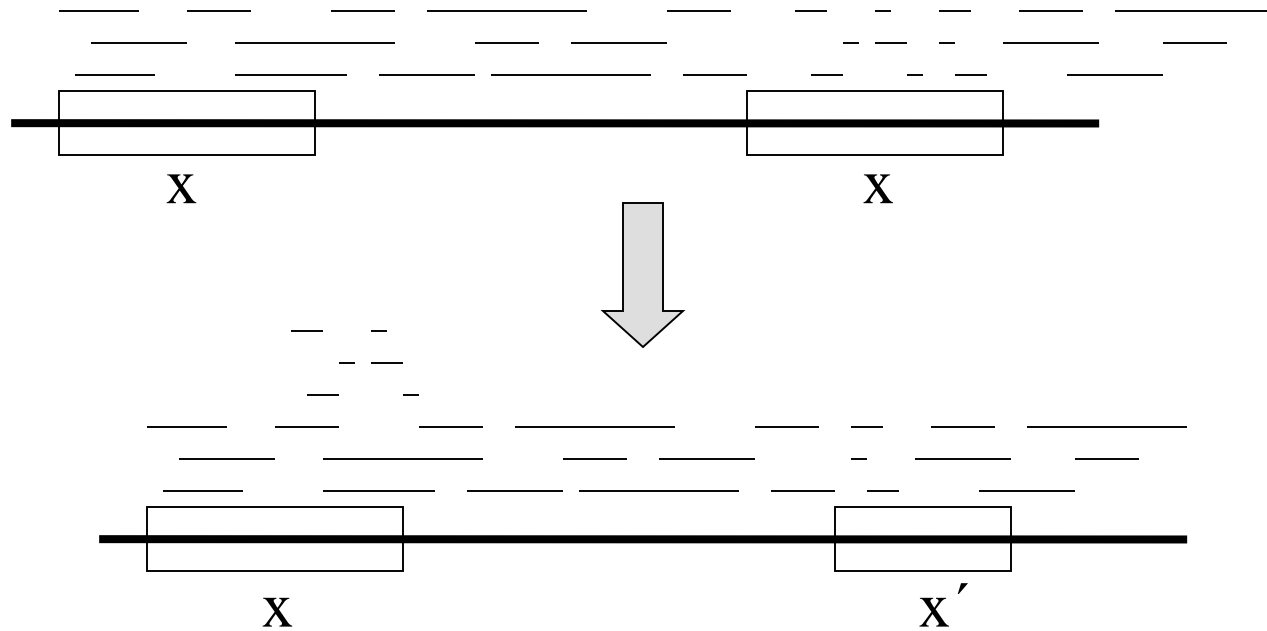
Wartość ρ określa ile razy otrzymane rozwiązanie jest gorsze od optimum.

W przypadku gdy algorytm zwraca rozwiązanie optymalne, $\rho = 1$. Jeżeli rozwiązanie może być dowolnie odległe od optimum, to wartość ρ jest nieskończonością.


```
S ← {s1, s2, ..., sk}  
while |S| > 1 do  
    (si, sj) ← para si, sj ∈ S maks. ov(si, sj)  
    u ← sklej (si, sj)  
    S ← S \ {si, sj} ∪ {u}
```

Da się go zaimplementować w czasie $O(n \log |\Sigma|)$ (Tarhio, Ukkonen, 1986), a nawet $O(n + \text{suf})$, gdzie suf to czas konstrukcji tablicy sufiksowej (Kociumaka, 2010).

- orientacja odczytów musi być znana
- nie uwzględnia pokrycia
- nie uwzględnia błędnych odczytów
- zakłada kompletność sekwencji



Uwzględnia błędy odczytu i nieznaną orientację.

Definicja:

f jest przybliżonym podciągiem S na poziomie błędu ε ,

gdy $d_s(f, S) \leq \varepsilon \times |f|$

gdzie: d_s – odległość mierzona jako stopień niedopasowania

(np. dopasowanie: 0, niedopasowanie: 1, przerwa: 1)

Zadanie:

Znaleźć najkrótsze możliwe słowo S , takie że dla każdego $f \in P$:

$$\min(d_s(f, S), d_s(\bar{f}, S)) \leq \varepsilon |f|$$

- Wejście: $\mathcal{P} = \{\text{ATCAT}, \text{GTCG}, \text{CGAG}, \text{TACCA}\}$
 $\varepsilon = 0.25$

- Wyjście:

A**T**GAT

ATCAT

-----CGAC

GTCG

-CGA**G**

----TAC**C**A

ACGATACGAC

$$d_s(\text{CGAG}, \text{ACGATACGAC}) = 1$$
$$= 0.25 \times 4$$

Odległość akceptowalna dla $\varepsilon = 0.25$

Uwzględniane są również przerwy w odczytach.

AT-GA-----
ATCGATAGAC

$$d_s = 1$$

Ten model jednak nadal ma wady:
uwzględnia błędy odczytu i nieznaną orientację,
ale:

- nie uwzględnia powtórzeń
- nie modeluje pokrycia
- zawsze generuje pojedynczy kontig (nie uwzględnia luk w sekwencji)

Wejście:

Zbiór słów nad alfabetem {A,C,G,T}

Wyjście:

Najkrótsze wspólne nadśłowo (poszukiwana sekwencja?)

Przykład:

{**ACGTAC**, **CATAC**, TACAT} -> TACAT**ACGTAC**

Problem najkrótszego wspólnego nadśłowa (SCS) można sprowadzić do problemu komiwojażera (*TSP*).

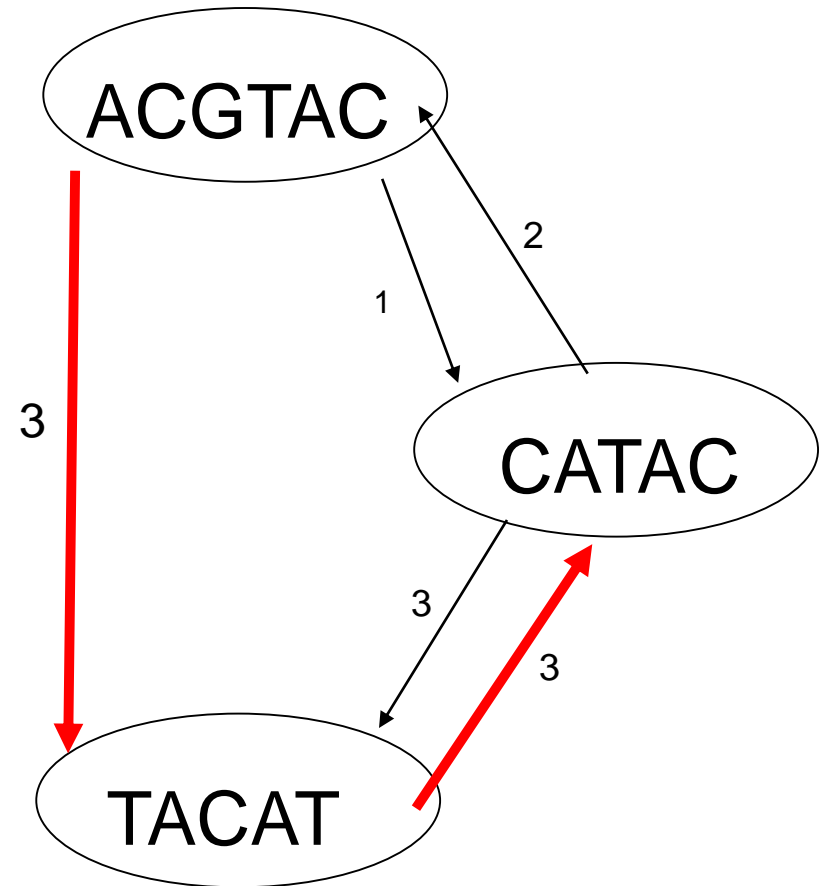
Przykład:

{ACGTAC, CATAC, TACAT}

- wierzchołki grafu reprezentują odczyty (*reads*)
- krawędzie grafu opisują nałożenia (*overlaps*)
- wagi to długości nakładających się prefiksów/sufiksów

Szukamy ścieżki Hamiltona o największym koszcie.

Rozwiązaniem TSP jest SCS.



1. Overlap-layout-consensus

J.D. Kececioglu and E.W. Myers, "Combinatorial Algorithms for DNA Sequence Assembly," *Algorithmica*, vol. 13, 1995, pp. 7-51.

wierzchołki reprezentują odczyty

- krawędzie reprezentują nałożenia (overlaps)

Szukamy takiej ścieżki, na której każdy wierzchołek występuje przynajmniej raz

2. Eulerian path

P.A. Pevzner, H. Tang, and M.S. Waterman, "An Eulerian Path Approach to DNA Fragment Assembly," *Proc. Nat'l Academy of Science USA*, vol. 98, no. 17, 2001, pp. 9748-9753.

Krawędzie reprezentują odczyty.

Szukamy ścieżki Eulera (ścieżki przechodzącej przez wszystkie krawędzie grafu).

k – ustalony parametr

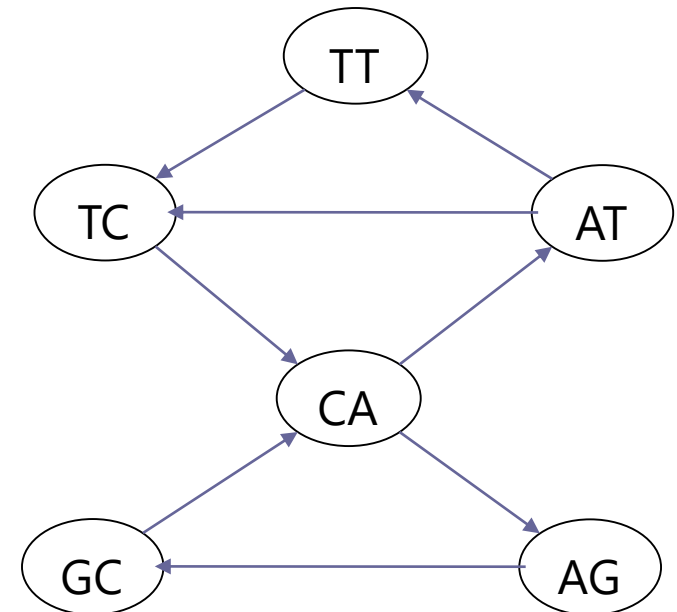
Wierzchołki to $(k-1)$ -krotki.

Krawędzie to k -krotki.

Zbiór k -krotek nazywamy k -spektrum.

Znalezienie najkrótszego słowa dla danego k -spektrum jest równoważne rozwiązaniu problemu chińskiego listonosza (*Chinese Postman Problem*)

{AGC, ATC, ATT, CAG, CAT, GCA, TCA, TTC}



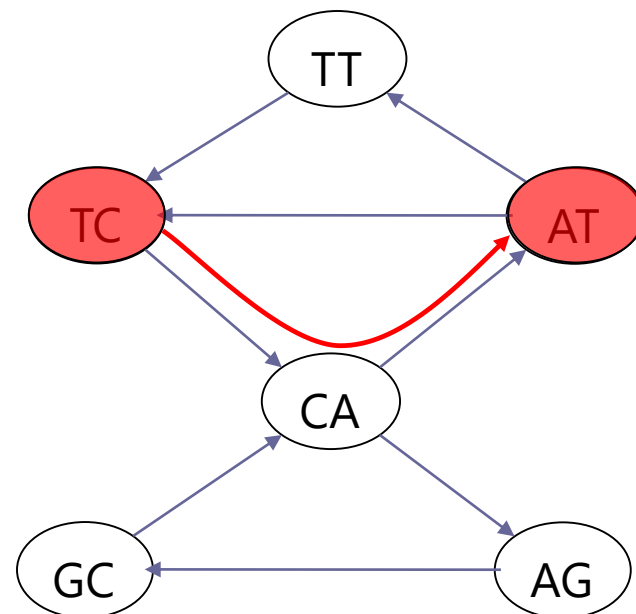
Problem chińskiego listonosza

„Dana jest sieć ulic oraz poczta. Aby listonosz dostarczył korespondencję musi przejść wzdłuż każdej ulicy co najmniej raz i powrócić do punktu wyjścia.”

Rozwiązaniem problemu jest cykl Eulera.

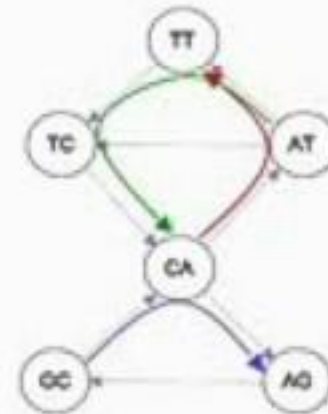
Jeśli graf nie jest Eulerowski, odpowiednio go przekształcamy:

{AGC, ATC, ATT, CAG, CAT, GCA, TCA, TTC}



Problem: znaleźć ścieżkę zawierającą wszystkie szlaki odczytów.

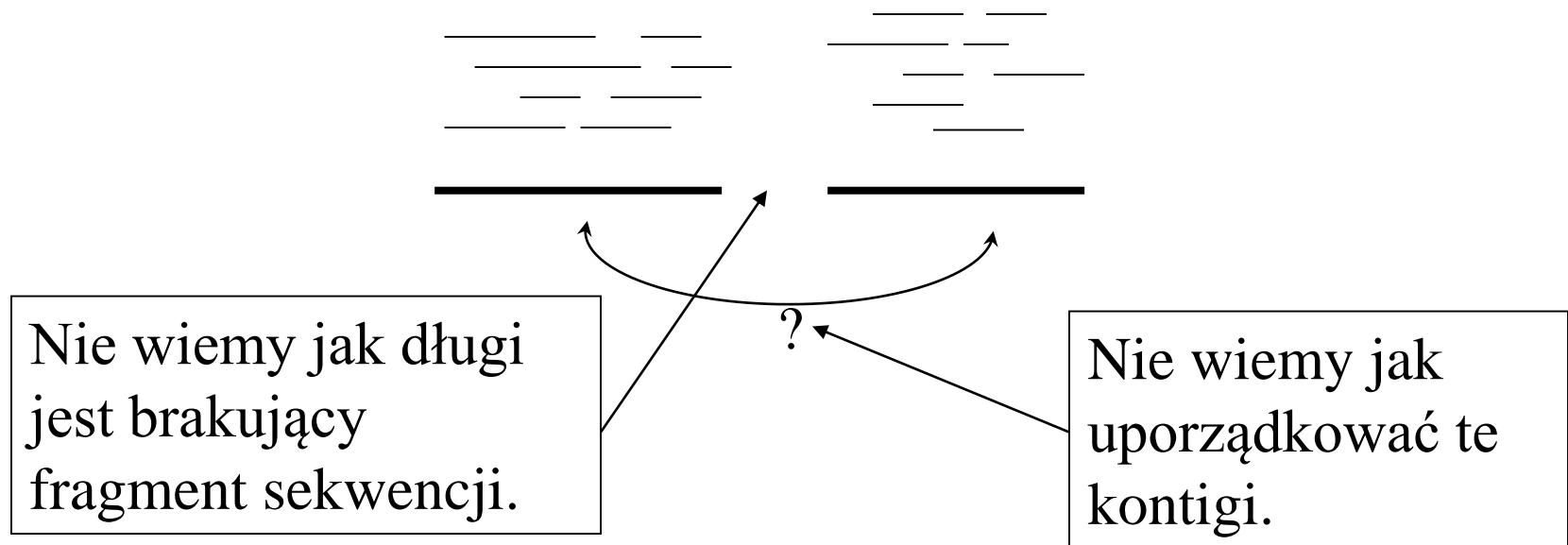
[AGC, **ATTCA**, **CATT**,
GCAG, ATC]

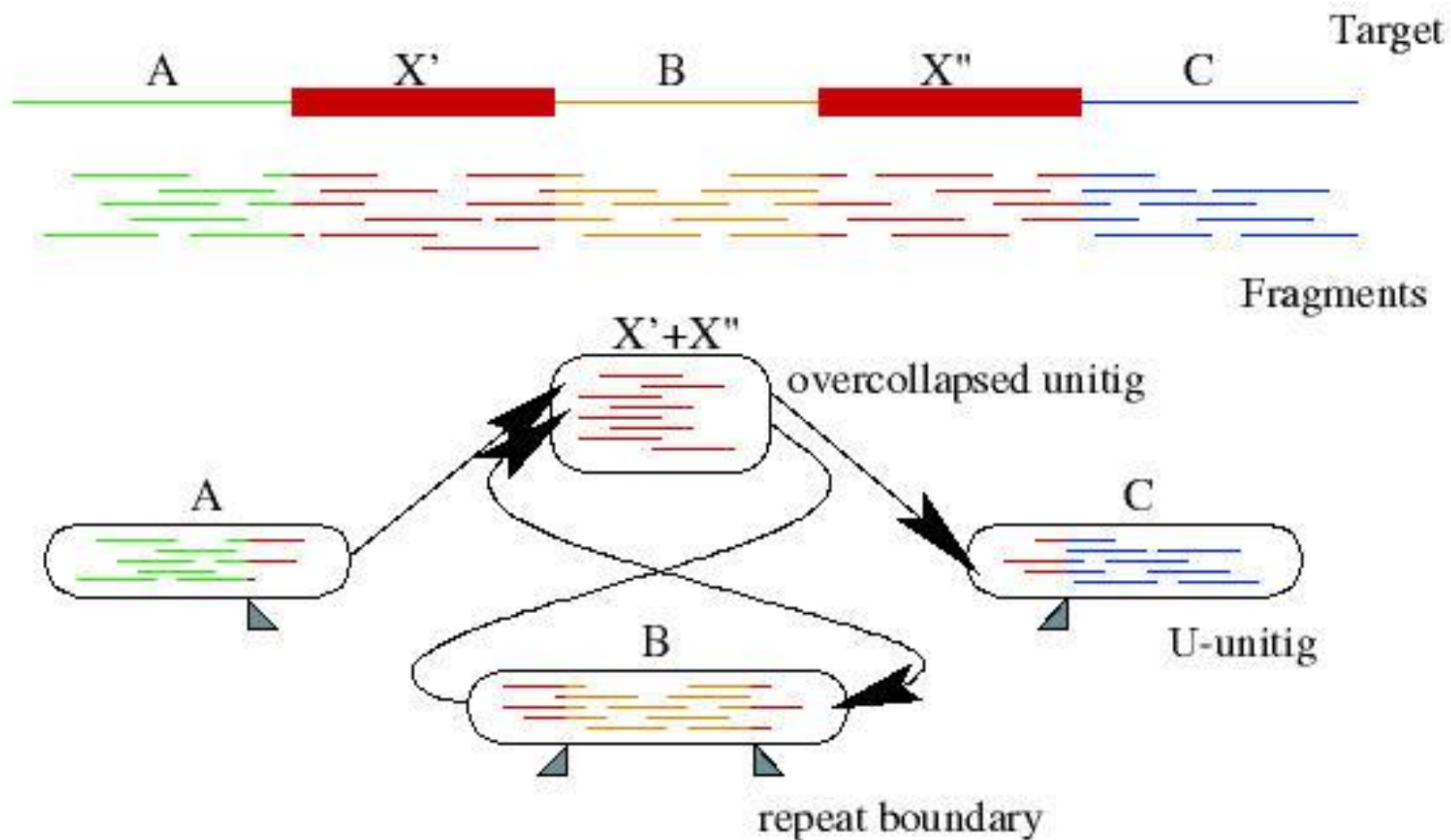


Wady podejścia opartego na grafach De Bruijna:

- arbitralny podział na k -krotki;
- algorytm wrażliwy na błędy w sekwencjonowaniu
- nieefektywny pamięciowo (jeden wierzchołek na każdą k -krotkę)

Czasem nie jesteśmy w stanie połączyć wszystkich fragmentów w jedną ciągłą sekwencję.



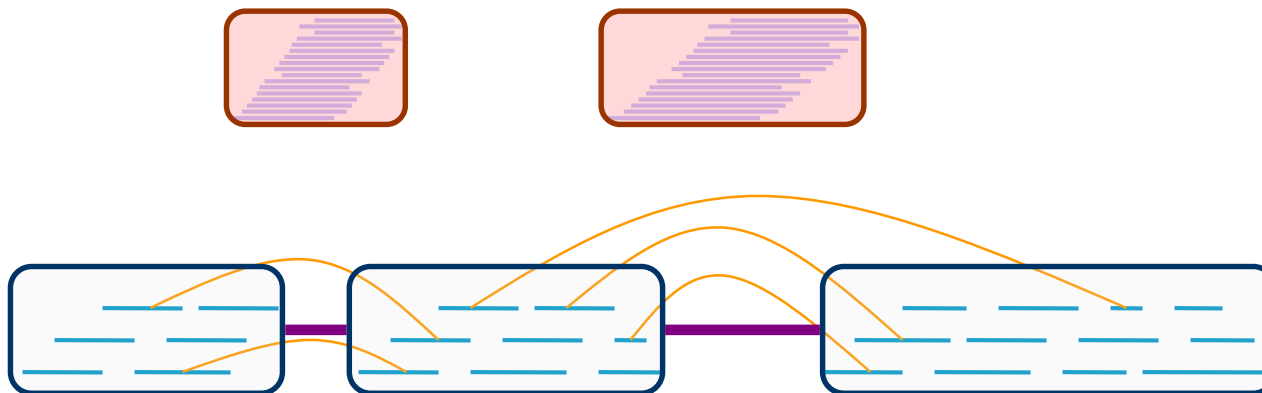


Znajdujemy linki między unikalnymi kontigami.

Łączymy je ze sobą.



Pozostałe luki wypełniamy powtarzającymi się sekwencjami.



Some Terminology

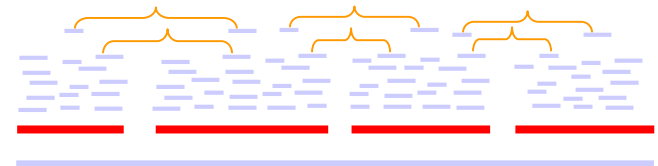
read a 25 - 200 long word that comes out of sequencer

mate pair a pair of reads from two ends of the same insert fragment

contig a contiguous sequence formed by several overlapping reads with no gaps

supercontig (scaffold) an ordered and oriented set of contigs, usually by mate pairs

consensus sequence sequence derived from the multiple alignment of reads in a contig



..ACGATTACAATAGGTT..

Średnia liczba odczytów zawierających dany fragment DNA (oczywiście, dla poszczególnych odcinków może być ona mniejsza lub większa).

inaczej

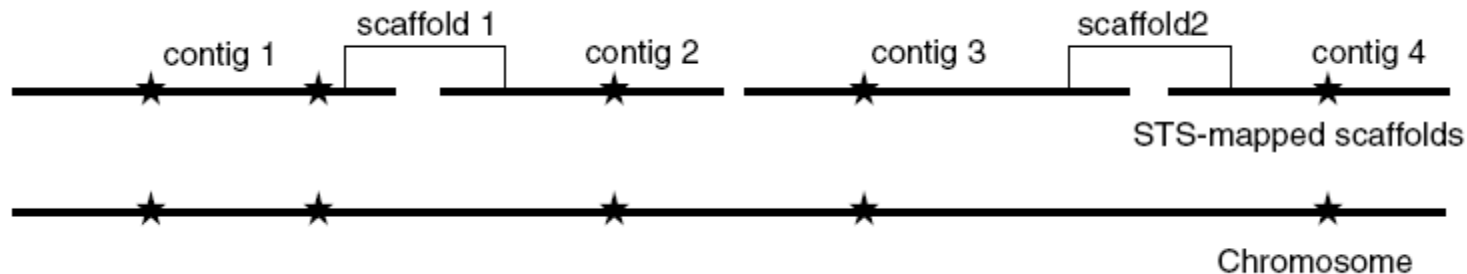
łączny rozmiar sekwencji poddanych analizie w porównaniu do długości całego genomu.

Lander & Waterman (1988) – dla idealnego projektu (bez „trudnych” obszarów) pokrycie 8x-10x jest wystarczające do skompletowania genomu. // sekwencjonowanie shotgun

W metodach nowej generacji, gdzie odczyty są znacznie krótsze, stosuje się większe pokrycie (np. 50x-100x)

- zamknięcie przerw pomiędzy kontigami;
- korekcja niepoprawnych dopasowań;
- powtórne sekwencjonowanie obszarów o niskim pokryciu lub niskiej jakości

Zwykle jest to najbardziej czasochłonny etap prac.



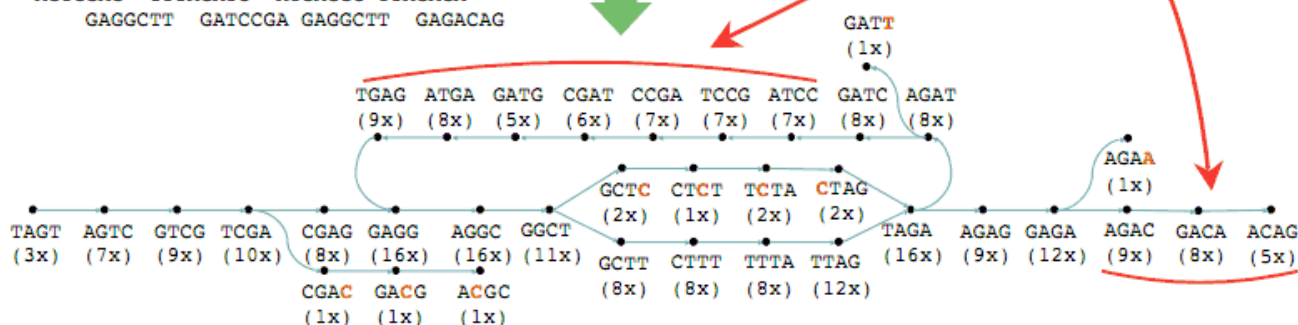
Metoda bazująca na grafach De Brujina – podsumowanie

```
TAGTCGAGGGCTTTAGATCCGATGAGGCTTTAGAGACAG
AGTCGAG CTTTAGA CGATGAG CTTTAGA
GTCGAGG TTAGATC ATGAGGC GAGACAG
GAGGCTC ATCCGAT AGGCTTT GAGACAG
AGTCGAG TAGATCC ATGAGGC TAGAGAA
TAGTCGA CTTTAGA CCGATGA TTAGAGA
CGAGGCT AGATCCG TGAGGCT AGAGACA
TAGTCGA GCTTTAG TCCGATG GCTCTAG
TCGACGC GATCCGA GAGGCTT AGAGACA
TAGTCGA TTAGATC GATGAGG TTTAGAG
GTCGAGG TCTAGAT ATGAGGC TAGAGAC
AGGCTTT ATCCGAT AGGCTTT GAGACAG
AGTCGAG TTAGATT ATGAGGC AGAGACA
GGCTTTA TCCGATG TTTAGAG
CGAGGCT TAGATCC TGAGGCT GAGACAG
AGTCGAG TTTAGATC ATGAGGC TTAGAGA
GAGGCTT GATCCGA GAGGCTT GAGACAG
```

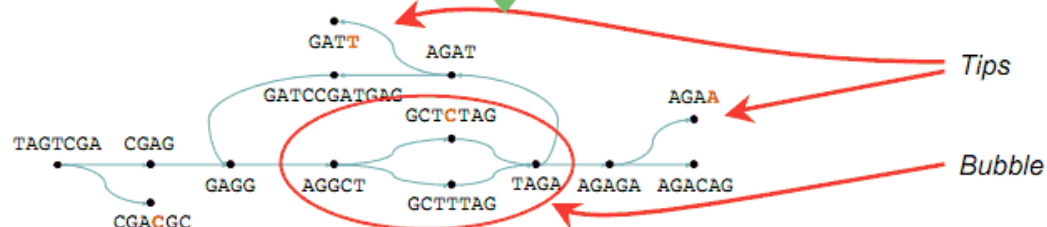
1. Sequencing
(e.g. Solexa, 454...)

2. Hashing

Linear stretches



3. Simplification of linear stretches



4. Error removal



Wejście:

Duża liczba (miliony) krótkich (po kilkadziesiąt znaków) odczytów.

Problemy:

- nieznana orientacja (bezpośredni odczyt lub komplementarna odwrócona sekwencja)
- błędy w danych wejściowych (nieprawidłowa, brak lub nadmiarowa litera)
- powtórzenia (pokrywające się odczyty pochodzące z różnych miejsc genomu)
- luki (nie wszystkie fragmenty są sekwencjonowane)

Metody:

- algorytmy słowne
- podejścia grafowe:
 - overlap-layout-consensus (szukamy ścieżki Hamiltona)
 - grafy de Brujina (szukamy ścieżki Eulera)

Etapy:

- identyfikacja kontigów
- łączenie kontigów
- tworzenie konsensusu

TIGR Assembler,

G.G. Sutton et al., "TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects," Genome Science and Technology, 1995, vol. 1, pp. 9-19.

Phrap,

P. Green, "Phrap Documentation: Algorithms," Phred/Phrap/Consed System Home Page; <http://www.phrap.org> (current June 2002).

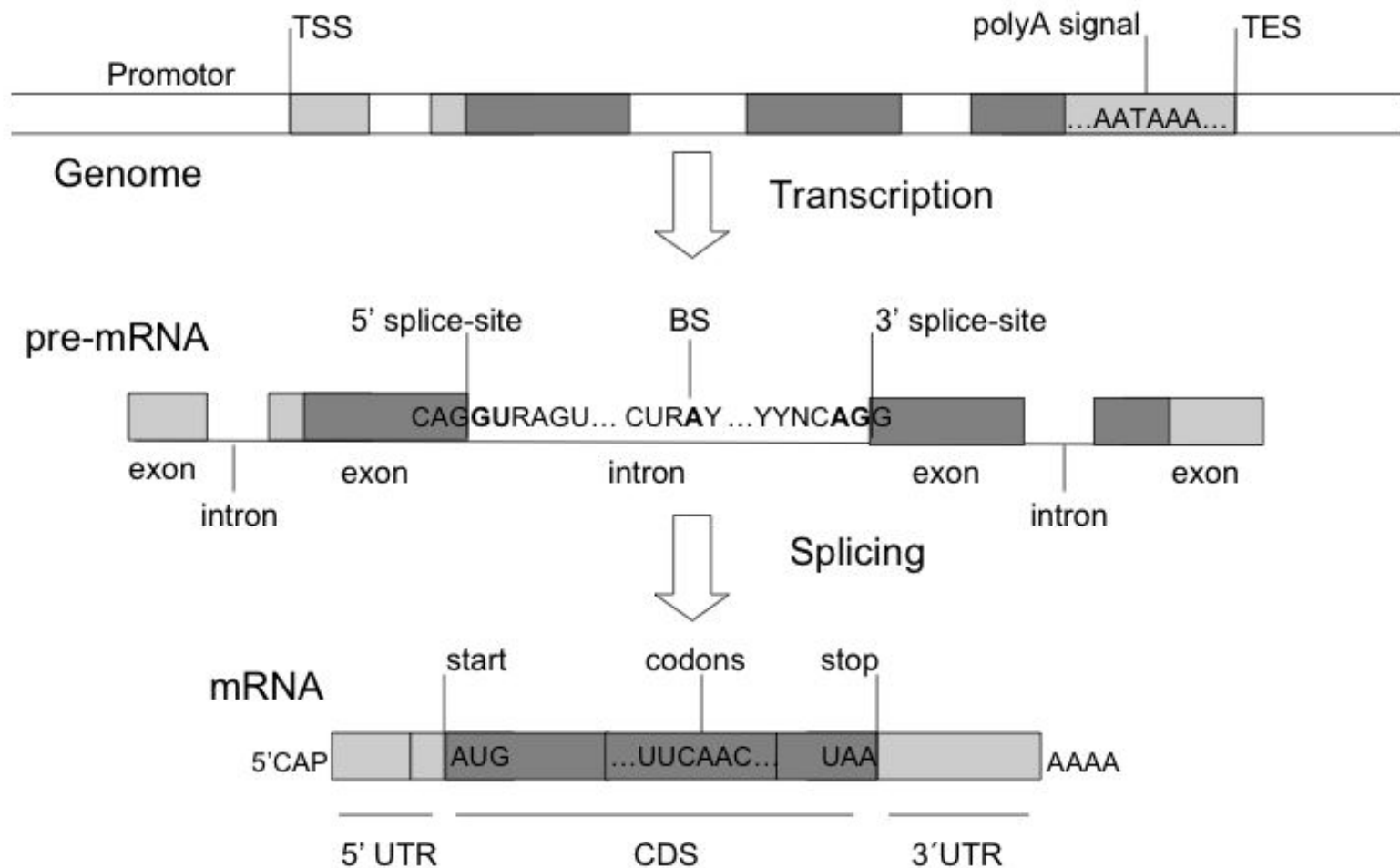
CAP3

<http://genome.cshlp.org/content/9/9/868.long>

Euler

Velvet

Przewidywanie genów



Przechowywanie danych

- Przechowywanie danych
- Transmisja danych
- Kompresja danych

Laboratoria nie przechowują danych z sekwencjonowania, ale zamrożone próbki genu.

Bardziej opłaca się ponownie zsekwencjonować daną próbkę, niż przechowywać surowe dane.

Laboratorium poprosiło pracowników z działu IT o zestawienie połączenia sFTP, po których chcieli przesłać dane sekwencyjne do współpracującego laboratorium.

IT zamiast FTP wystawił aplikację webową zdolną przesyłać maksymalnie 1 plik jednocześnie (pojedynczy eksperyment zawierał ponad 16 000 plików).

Ostatecznie udało się uzyskać sFTP. Łącze szybko się zapchało, ponieważ alokowano tylko 120 GB transferu (roczne potrzeby: 1-2 TB).

