

PRZEDMIOT OBIERALNY: BIOINFORMATYKA, LABORATORIUM

Wersja pliku: v 1.5.0, 16 marca 2025

Autor: dr inż. Marcin Radom

PODSTAWOWE TERMINY I DEFINICJE

DNA – kwas deoksyrybonukleinowy.

W ogólności, zbudowany z 4 rodzajów nukleotydów: **adeniny**, **cytozyny**, **guaniny**, **tyminy**.

Informatyczno-matematycznie: DNA możemy traktować jako ciąg znaków o długości n zbudowany nad alfabetem $\Sigma_{\text{DNA}} = \{A, C, G, T\}$.

*Dodatkowy komentarz: RNA (kwasem rybonukleinowym) nie będziemy się w ogóle przejmować. Z punktu widzenia budowy obowiązują te same reguły z jednym wyjątkiem: zamiast **tyminy** (T) która buduje DNA, w cząsteczkach RNA zamiast niej jest **uracyl** (U).*

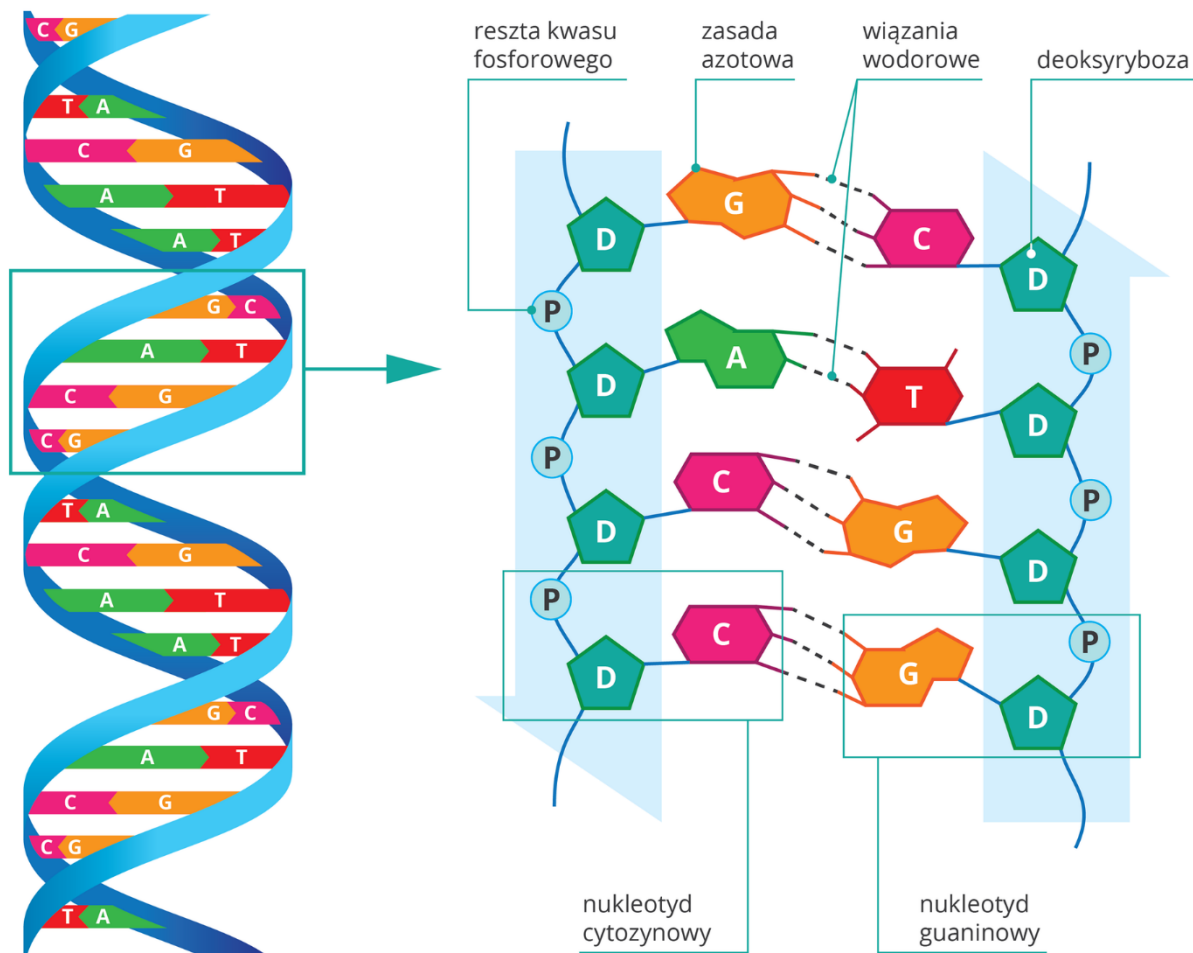
KOMPLEMENTARNOŚĆ.

DNA ma kształt podwójnej helisy, a innymi słowy zbudowany jest z dwóch nitek wzajemnie wokół siebie zawiniętych. Jeżeli na ten samej pozycji po jednej stronie jest litera A (**adenina**), wtedy po drugiej stronie musi znajdować się T (**tymina**). Jeżeli po jednej stronie jest litera C (**cytozyna**) – wtedy po drugiej G (**guanina**). Jest to spowodowane wiązaniami chemicznymi – **adenina** łączy się z **tyminą**, **cytozyna** z **guaniną**. Jeżeli po jednej stronie jest np. adenina a po drugiej z jakiegoś powodu znalazłaby się cytozyna (zamiast **tyminy**, czyli mamy wiązanie A-C) to ono oczywiście jest możliwe, ale jest ono bardzo bardzo słabe, a co ważniejsze – jest błędem z naszego punktu widzenia (z punktu widzenia białek odpowiedzialnych za naprawy DNA w komórkach – też)

Przykład: dla łańcucha (pojedynczej nitki DNA) o strukturze A-T-C-T-T-G-A-C-C, po drugiej stronie powinna być wokół niego zawinięta nitka o budowie: T-A-G-A-A-C-T-G-G, co daje taki mniej więcej obrazek:

A	T	C	T	T	G	A	C	C
T	A	G	A	A	C	T	G	G

Gdzie kreski oznaczają wiązania pomiędzy nukleotydami. Całość wygląda mniej więcej tak malowniczo:



Źródło: Dariusz Adryan, licencja: CC BY 3.0.

Cecha ta (komplementarność) jest podstawą działania dla wielu procesów związanych z DNA, np. jego replikacją (reakcja PCR – *polymerase chain reaction*, reakcja łańcuchowa polimerazy umożliwiającą powielenie fragmentów DNA) czy naprawą odbywającą się w zasadzie non stop w każdym jądrze komórkowym, bo tam też w ramach komórki DNA przebywa.

Tyle o budowie DNA w wersji skróconej i podstawowej. Kolejne definicje, mimo że będą co chwila ocierać się o budowę DNA, wchodzi już w skład słownika metody *Sekwencjonowania przez Hybrydyzację* (*Sequencing by hybridization*). Metoda ta nie jest współcześnie już tak popularna jak była kiedyś (np. nie nadaje się do sekwencjonowania całych genomów), tym niemniej jest wciąż bardzo interesująca z informatycznego punktu widzenia.

SEKWENCJONOWANIE PRZEZ HYBRYDYZACJĘ (SBH)

OLIGONUKLEOTYD

Oligonukleotyd to krótki, *jednoniciowy* fragment DNA o długości k , zazwyczaj od kilku do kilkunastu nukleotydów.

Przykład: A T C G G T jest oligonukleotydem o długości 6, jest to właśnie krótki, jednoniciowy fragment DNA (zazwyczaj sztucznie utworzony do jakiegoś celu).

MACIERZ HYBRYDYZACJI

Inne nazwy: *mikromacierz*, *mikrochip* – upraszczając zagadnienie: możemy sobie ją wyobrazić jako pewną płytkę, na której w oddzielnych komórkach znajdują się wielokrotnie sklonowane oligonukleotydy. WSZYSTKIE możliwe o danej długości zbudowane nad alfabetem $\Sigma_{\text{DNA}} = \{A, C, G, T\}$, czyli dla $k=1$ są tylko 4 takie komórki, chociaż ciężko mówić, że pojedynczy nukleotyd jest oligonukleotydem (*oligo: kilka, trochę*), dla $k=2$ – 16 komórek, czyli 4^2 , itd.

Na przykład dla $k = 3$, macierz ta zawiera $|S| = 4^3 = 64$ komórki, wszystkie kombinacje oligonukleotydów od AAA do TTT rozdzielone po osobnych komórkach macierzy. *Tutaj pewien komentarz rozdziwiający rzeczywistość biologiczną od informatycznej: w tej pierwszej - biologicznej, w każdej komórce, np. tej o której wiemy, że ma być tam 'TCA', byłyby miliony kopii TCA, a nie tylko jedna, samotna cząsteczka TCA. Wynika to z tego, że nie byłibyśmy w stanie wykryć faktu przyłączenia się jednej jedynej cząstki do innej nici DNA. Gdy kopie tego samego oligonukleotydu w komórce macierzy liczone są w milionach, wtedy sygnał pochodzący z faktu ich przyłączania się do jakiejś nici DNA jest możliwy do wykrycia (po co całe to przyłączanie i sygnały – o tym będzie dalej). Natomiast w naszej rzeczywistości informatycznej, macierz hybrydyzacji dla $k=3$ można sobie wyobrazić jako macierz (wektor właściwie) 64 pól, z których każde odpowiada za 1 nukleotyd: od AAA do TTT.*

SBH – SCHEMAT EKSPERYMENTU HYBRYDYZACYJNEGO

Sekwencjonowanie przez hybrydyzację – jest to metoda odczytu DNA. W uproszczeniu działa to tak: sklonowane (wspomniana reakcja PCR) jednoniciowe DNA jest „wrzucane” (wlewane/whatever) na macierz hybrydyzacji. Pewne komórki macierzy przyczepiają do siebie fragmenty DNA, w idealnej sytuacji: tylko te komórki w których znajdują się odpowiednie, czyli komplementarne oligonukleotydy. Do czego komplementarne? Ano do fragmentów tego sklonowanego i analizowanego DNA.

Przykład: jednoniciowe DNA o strukturze: **A T T A G T A T** „przyczepi się” (zazwyczaj, o tym – za chwilę) do komórki chipu zawierającej oligonukleotyd **A A T C A**, ponieważ **AATCA** (na płytce) jest komplementarnym łańcuchem do podłańcucha **TTAGT** (w wejściowym DNA).

SPEKTRUM DNA

Spektrum DNA jest to lista tych komórek mikrochipu SBH (patrz: macierz hybrydyzacji) które w wyniku eksperymentu hybrydyzacyjnego przyczepiły się do fragmentów DNA, które „wylaliśmy” na mikrochip.

Bardziej naukowo: spektrum określamy jako zbiór S – oligonukleotydów (z komórek mikrochipu które przyczepiły się do DNA) z których każdy z nich jest o długości k . Liczność spektrum to moc zbioru S , czyli: $|S|$. Elementy spektrum stanowią zbiór danych do budowy grafu, na podstawie którego będziemy próbowali zrekonstruować DNA.

Komentarz: to jest właśnie projekt do zrobienia. Algorytm, który z fragmentów DNA odtworzy jak najdokładniej oryginalne DNA. Oczywiście jedną nitkę, bo druga – komplementarna jest trywialna do odtworzenia (patrz: zasada komplementarności).

Na tym jednak historia się nie kończy. Są dwa rodzaje spektrum: idealne oraz z błędami. Błędów są dwa rodzaje (pozytywne i negatywne), z czego jeden rodzaj (negatywne) dzieli się dalej na dwa potencjalne „źródła” tychże błędów (z hybrydyzacji i z powtórzeń). Kontynuujemy tą jakże fascynującą lekturę dalej.

IDEALNE SPEKTRUM DNA

Jest to spektrum bez błędów hybrydyzacji (czykolwiek są, o tym dalej). Dla takiego przypadku, tj., jeżeli wiemy, że faktycznie dysponujemy idealnym spektrum, wiemy, że do rekonstrukcji musimy użyć wszystkich jego elementów. Wiemy też, ile ich jest. Jeżeli długość DNA to n , a długość oligonukleotydów z których spektrum musi się składać do k , wtedy **liczność spektrum idealnego to $|S| = n - k + 1$**

Przykład: dla DNA o długości $n=7$: **A T C T G T A** łańcuch komplementarny to **T A G A C A T**. Dla długości oligonukleotydów $k = 3$, idealne spektrum dla tego fragmentu DNA zawiera komórki z macierzy hybrydyzacji oznaczone jako **TAG**, **AGA**, **GAC**, **ACA**, **CAT**, czyli są to komplementarne oligonukleotydy do fragmentów oryginalnego DNA, odpowiednio do: **ATC**, **TCT**, **CTG**, **TGT**, **GTA**.

Liczność idealnego spektrum to $|S| = n - k + 1$, czyli dla $n = 7$, $k = 3$ mamy więc: $|S| = 7 - 3 + 1 = 5$.

Uwaga praktyczna: z punktu widzenia projektu do zrobienia, to czy algorytm odtwarza oryginalne DNA czy komplementarną do niego nie bezpośrednio z elementów spektrum – **nie ma to znaczenia**. Jak mamy jedną nitkę DNA, drugą odtworzymy w trywialny sposób z czasie liniowo zależnym od wielkości n (długości DNA) – zgodnie z regułą komplementarności.

BŁĘDY HYBRYDYZACJI

Błędy negatywne (brak czegoś w spektrum, precyzyjnie: niektórych fragmentów naszego badanego DNA)

wynikające z powtórzeń w samym DNA

wynikające z niedostatków technologicznych eksperymentu hybrydizacyjnego

Błędy pozytywne (dodatkowe, fałszywe elementy w spektrum)

wynikające z niedostatków technologicznych eksperymentu hybrydizacyjnego

Najpierw omówimy **błędy negatywne wynikające z powtórzeń**. Chodzi tutaj o powtórzenia tych samych fragmentów pewnego łańcucha DNA w różnych miejscach tego łańcucha. Na przykład we fragmencie: **A T C T T A G T C T** i $k=3$ powtarza się fragment **TCT** na dwóch różnych pozycjach. Im dłuższe DNA (n) i im mniejsze k (długość oligonukleotydów użytych w macierzy hybrydizacji w SBH) tym więcej tego typu powtórzeń będzie się zdarzać. Nawet w najbardziej podstawowej wersji algorytm sekwencjonujący DNA na podstawie informacji z macierzy/spektrum DNA (patrz punkt niżej) musi sobie z tym poradzić. A dlaczego w ogóle jest to błąd negatywny (wynikający z powtórzeń)? Dlatego, że spektrum **NIE JEST** multizbiorem, tylko zwykłym zbiorem. To, że jedna komórka mikrochipu SBH przyłączyła DNA na dwóch lub więcej pozycjach nie jest wykrywane. W związku z tym w spektrum dostajemy tylko **jeden** oligonukleotyd komplementarny to **TCT** (czyli będzie to **AGA**). Nie wiemy niestety tego, że gdy już ze spektrum zrobimy graf (o tym dalej), to ten wierzchołek musi być odwiedzony dwa razy (albo jeśli nawet odwiedzimy go raz, to w inny sposób musimy skompensować to, że odpowiada on za jeszcze inny, dalszy fragment rekonstruowanego DNA).

Ogólne błędy hybrydizacji:

1. **Błędy negatywne** - w spektrum brakuje pewnych elementów, które w teorii składają się na prawdziwe, poszukiwane DNA. Nie wynika to z powtórzeń, ale błędnego „odczytania” komórki chipu, która na przykład w teorii powinna przyłączyć się do DNA (bo jest komplementarna, tj. oligonukleotydy w tej komórce są), ale w praktyce nie zostało to wykryte w eksperymencie hybrydizacyjnym. Być może faktycznie do przyłączenia nie doszło albo z jakichś powodów było ich za mało, sygnał z tej komórki był za słaby i nie został wykryty. Powody nas nie interesują, tylko efekt: w spektrum brakuje pewnych oligonukleotydów, które byłyby tam w przypadku spektrum idealnego.
2. **Błędy pozytywne** - spektrum zawiera pewne dodatkowe oligonukleotydy, które jednak nie występują w badanym (sekwencjonowanym) DNA. Innymi słowy badane DNA w teorii nie powinno przyłączyć się do pewnej komórki (reguła komplementarności), ale z jakichś powodów komórka mikrochipu została uwzględniona jako część spektrum DNA.

W przypadku **tylko** błędów negatywnych (dowolnego typu), licznosc spektrum dana jest **nierównością**: $|S| < n - k + 1$ (czyli wiemy, że brakuje w nim elementów w stosunku do spektrum idealnego). W przypadku **tylko** błędów pozytywnych licznosc spektrum dana jest **nierównością**: $|S| > n - k + 1$.

W przypadku obu rodzaju błędów na raz, jak można się domyślić, rozpoczyna się prawdziwa „zabawa” w składanie DNA. I jak również można się domyślić, to ten przypadek jest tym realnym, z którym nasze algorytmy będą musiały się mierzyć.

GRAF ZE SPEKTRUM DNA.

Elementy spektrum to wierzchołki grafu – innymi słowy oligonukleotydy ze spektrum (albo ich komplementarne odpowiedniki, wszystko jedno, byle się na którąś wersję zdecydować) i ich liczność odpowiadają za zbiór $|V|$ wierzchołków grafu, który posłuży algorytmom do rekonstrukcji DNA.

UWAGA PRAKTYCZNA / INFORMATYCZNA. Eksperyment hybrydyzacyjny trzeba jakoś odtworzyć na komputerze. Jest to w sumie proste. Jeśli chcemy spektrum idealne, bierzemy DNA o długości n i tnimy je na wszystkie możliwe kawałki o długości k , będzie ich w sumie $n - k + 1$ jak wynika ze wzoru na spektrum idealne. Otrzymujemy z takiego prostego algorytmu sekwencję wszystkich kawałków DNA – stety-niestety: **UPORZĄDKOWANĄ**. Problem rekonstrukcji z grafu w zasadzie nie istnieje, bo przecież znamy kolejność wierzchołków. **NIE TAK TO WYGLĄDA** w eksperymencie SBH – tam dostajemy (nawet w przypadku spektrum idealnego) zbiór oligonukleotydów, ale który z nich jest pierwszy, który ostatni w DNA i jaka kolejność jest pomiędzy – nie mamy pojęcia i na tym właśnie polega problem do rozwiązania. Dlatego jest chyba oczywiste, że w Państwa projekcie po „pocięciu” DNA na kawałki symulujące efekty prawdziwego SBH w laboratorium biologicznym, musimy te fragmenty losowo wymieszać w naszym spektrum. Inaczej nie ma w ogóle czego obliczać...

Łuki grafu są z wagą oznaczającą ilość nakładających się liter podczas przechodzenia z wierzchołka do wierzchołka. Wyjaśniają to poniższe przykłady, ale najpierw przyjęta w dalszym opisie notacja (do tematu wrócę pod koniec)

waga łuku = długość_oligonukleotydów – liczba_liter_ich_nałożenia

Przykład: z wierzchołka **ATGT** prowadzi łuk o wadze 1 ($w = 4 - 3 = 1$) do wierzchołka o nazwie **TGTC**, ponieważ fragment **TGT** nakłada się na siebie przy złożeniu ATGT i TGTC: oba wierzchołki (każdy o $k=4$) tworzą wtedy łańcuch 5-literowy **ATGTC**. Dodatkowo z tego samego **ATGT** do **TGTC** prowadzi też drugi, równoległy łuk o wadze 3 ($w = 4 - 1 = 3$), ponieważ nałożenie można wykonać tylko na ostatnim nukleotydzie w **ATGT** i pierwszym nukleotydzie w **TGTC**. Otrzymamy wtedy **ATGTGTC**, czyli łańcuch 7-elementowy.

W ogólności, jeżeli długość oligonukleotydu definiującego nazwę wierzchołka to k , wtedy maksymalne nałożenie możliwe jest na $k-1$ ostatnich/pierwszych liter dwóch wierzchołków i **wagę takiego łuku określamy jako 1** (rozumiemy przez to, że w wyniku takiego przejścia tym łukiem nowy łańcuch rekonstruowanego DNA powiększy się tylko o 1 nukleotyd). Minimalne możliwe nałożenie możliwe jest przez łuk o **wadze $k-1$** , wtedy po przejściu takim łukiem rekonstruowane DNA powiększy się dokładnie o $k-1$ nowych liter (odpowiada ono nałożeniu dwóch oligonukleotydów na siebie na jednej literze: odpowiednio ostatniej pierwszego i pierwszej drugiego) – tak jak w prostym przykładzie powyżej.

Inny przykład: z wierzchołka **TAAAA** do **AAAAC** prowadzą 4 łuki, o wagach: 1 (**AAAA**), 2 (**AAA**), 3(**AA**) oraz 4(**A**) – gdzie w nawiasie jest podany potencjalny nakładający się na siebie podłańcuch obu wierzchołków (odpowiednio: podłańcuch końcowy dla **TAAAA** oraz początkowy dla **AAAAC**). Minimalna waga łuków, czyli maksymalne nałożenie na $k-1$ literach alfabetu Σ_{DNA} (innymi słowy: chodzenie w grafie tylko po łukach o wadze 1) jest możliwa do stosowania tylko w przypadku, gdy brak *jakichkolwiek* błędów, nawet tych wynikających z powtórzeń. Czyli praktycznie niemal nigdy – takie przypadki SBH gdzie spektrum jest idealne są niezwykle rzadkie i zakładanie, że z takim mamy do czynienia należy traktować jako skrajnie wyjątkowy scenariusz.

Wagi raz jeszcze: możliwa jest oczywiście inne wyliczanie wagi, według wzoru:

$$\text{waga_łuku} = \text{liczba_liter_nałożenia_oligonukleotydów}$$

Wtedy np. łuk pomiędzy **ATGT** a **TGTC** ma wagę 3. *Jest to w zasadzie kwestia techniczna, to, jak to dokładnie będzie wyglądać w projekcie zależy od tego co sami Państwo zdecydują – celem algorytmu jest rekonstrukcja DNA, jakkolwiek to robi (choć specyfikacja jaki to powinien być algorytm będzie tematem dalszych materiałów).*

Przykład raz jeszcze: dla DNA o długości $n=7$: **A T C T G T A** łańcuch komplementarny to **T A G A C A T**. Dla długości oligonukleotydów $k = 3$, idealne spektrum dla tego fragmentu DNA zawiera komórki z macierzy hybrydyzacji oznaczone jako **TAG**, **AGA**, **GAC**, **ACA**, **CAT**, czyli są to komplementarne oligonukleotydy do fragmentów oryginalnego DNA, odpowiednio do: **ATC**, **TCT**, **CTG**, **TGT**, **GTA**.

Idealna wielkość spektrum to $|S| = n - k + 1$, czyli dla $n = 7$, $k = 3$, $|S| = 7-3+1 = 5$. W takim wypadku widać, że nie ma tutaj żadnych błędów, nawet negatywnych wynikających z powtórzeń. Gdyby takie istniały, to $|S| < n - k + 1$. W przypadku, gdyby były TYLKO błędy pozytywne hybrydyzacji, wtedy $|S| > n - k + 1$. W obu wypadkach należy odwiedzić **odpowiednią liczbę wierzchołków**, aby zrekonstruować DNA o długości n .

- Jaka to *odpowiednia* liczba? **Przy braku błędów:** $|S|$, czyli wszystkie ze spektrum.
- Przy błędach negatywnych wynikających z powtórzeń:** $|S|$ czyli wszystkie ze spektrum, przy czym w tym wypadku **niektóre wierzchołki odwiedzamy więcej niż raz**. Ale wciąż przyjmujemy, że odwiedzamy wszystkie, i wtedy wciąż możemy budować graf o wagach 1 (maksymalne nałożenie wierzchołków na $k-1$ literach).
- Przy ogólnych błędach pozytywnych** (i tylko takich): $n - k + 1$ wierzchołków, bo samo spektrum ma wtedy więcej elementów: $|S| > n - k + 1$. Jest chyba absolutnie oczywiste, że graf dla takiego przypadku musi już uwzględniać wszystkie możliwe do stworzenia łuki pomiędzy każdą parą wierzchołków, czyli ich wagi będą od 1 do $k-1$.
- Przy błędach pozytywnych i negatywnych równocześnie** (inne sformułowanie: w spektrum brakuje prawdziwych elementów DNA i jednocześnie mamy w nim fałszywki: jak żyć? jak żyć?? 😞)

przechodzimy przez tyle wierzchołków, aby zrekonstruować DNA o długości n . Niestety przy tak ogólnym sformułowaniu tej liczby, wyniki, czyli uzyskane DNA najbardziej będą odstawać od oryginału.

Istnieje też problem **rozwiązań niejednoznacznych**, tj. graf może mieć wiele ścieżek będących pewnymi rekonstrukcjami DNA, i nie wiemy w sumie które z nich to oryginał (gdybyśmy go nie znali). Ten problem występować może w każdym przypadku, **nawet** wtedy, gdy spektrum jest idealne!

Podsumowując, potrzebujemy algorytm, który dla dowolnego spektrum, także tego z błędami, będzie nam rekonstruować DNA z kawałków (oligonukleotydów), czyli szukać ścieżki w grafie ze spektrum o odpowiedniej długości (odpowiedniej, czyli takiej, która odtworzy DNA o długości n).

ALGORYTM MA NASTĘPUJĄCE DANE WEJŚCIOWE:

- Długość DNA – n
- Długość oligonukleotydów – k
- Spektrum DNA, czyli lista oligonukleotydów, z błędami jeżeli są:
- Informacje jakie błędy są w spektrum, tj. dokładnie, ile pozytywnych i ile negatywnych. Jak są negatywne wynikające z powtórzeń w całym zbiorze błędów negatywnych, można ten zero-jedynkowy fakt też odnotować w danych wejściowych: czyli algorytm wie, że są powtórzenia, ale nie wie które i ile.
- Początkowy oligonukleotyd – wierzchołek początkowy dla ścieżki w grafie (to ostatnie nie jest żadnym szczególnym oszustwem czy jakimś wielkim ułatwieniem, w normalnych eksperymentach SBH też wiedzieliśmy od czego startujemy, wynika to z innych cech eksperymentu hybrydyzacyjnego).