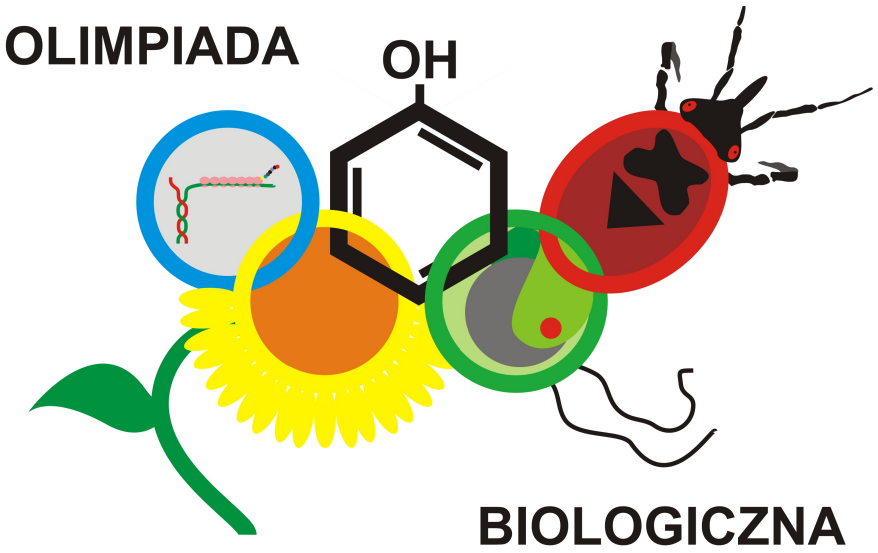
Olimpiada Biologiczna



Informator – narzędzia bioinformatyczne

Katarzyna Grudziąż, Takao Ishikawa

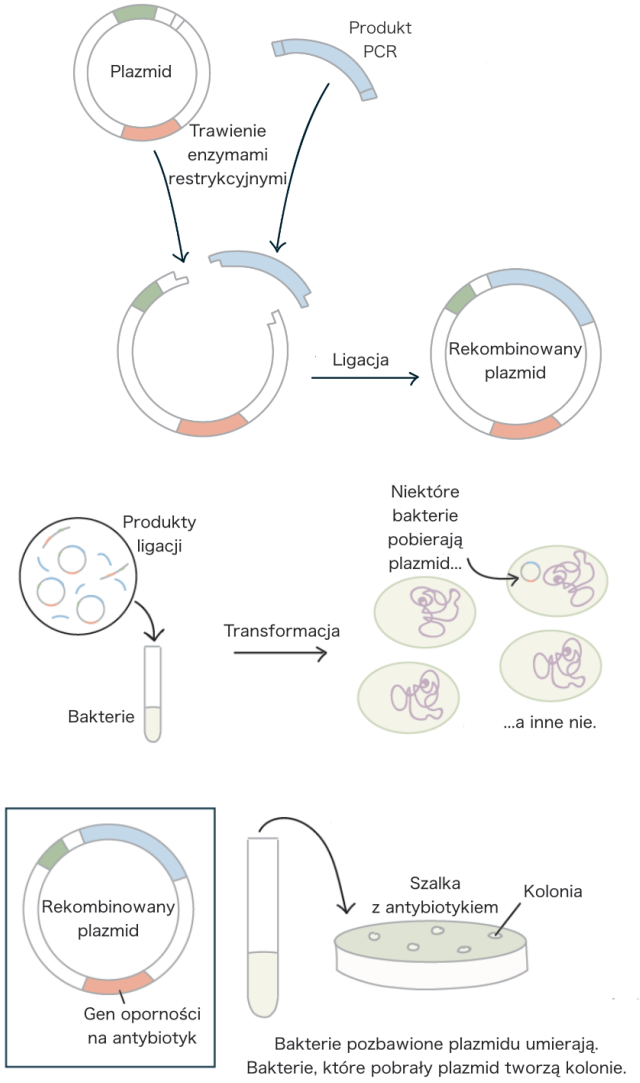
Warszawa, luty 2019 r.

Bioinformatyka to dziedzina nauki łącząca w sobie wiedzę z dziedziny biologii z wykorzystaniem narzędzi informatycznych. Do głównych zadań bioinformatyki należy przechowywanie w bazach danych informacji o sekwencjach DNA, RNA i białek oraz strukturach makrocząsteczek, a także tworzenie narzędzi służących do analizy tych informacji. Warto zwrócić uwagę na to, że skala uzyskiwanych obecnie informacji daleko przekracza możliwości ręcznego przetwarzania. Na przykład, genom bakterii *Escherichia coli* liczy ponad 4,6 milionów par zasad. Porównywanie tego genomu z genomami innych bakterii jest więc zadaniem dla programów komputerowych.

Klonowanie *in silico*

Klonowanie DNA polega na przeniesieniu wybranego fragmentu DNA z organizmu dawcy do organizmu gospodarza i powieleniu. Celem klonowania może być po prostu namnożenie fragmentu DNA w komórkach gospodarza albo przygotowanie konstruktu do ekspresji RNA lub białka.

Do powielenia wybranego fragmentu wykorzystuje się technikę PCR, następnie produkt reakcji (tzw. wstawkę, ang. *insert*) łączy się z wektorem, czyli specjalną cząsteczką DNA, umożliwiającą wprowadzenie wstawki do organizmu gospodarza. Robi się to poprzez trawienie wektora i wstawki odpowiednimi enzymami restrykcyjnymi, a następnie ligację (Rys. 1). Wektorami, w zależność od gospodarza, mogą być konstrukty stworzone przez modyfikowanie plazmidów, fagów lub mogą to być sztuczne chromosomy. W dalszej części informatora zajmiemy się najczęściej stosowanym w praktyce laboratoryjnej układem: klonowaniem genu w bakteriach przy pomocy wektora plazmidowego.



Rys. 1. Przebieg klonowania DNA.

Zanim przeprowadzi się procedurę klonowania w laboratorium trzeba ją dobrze zaplanować. W tym celu wykonuje się klonowanie *in silico* (czyli w komputerze). Korzystając z narzędzi do edytowania sekwencji DNA, wybiera się najlepszy do danego celu wektor, projektuje się startery do PCR i przygotowuje się sekwencję ostatecznego konstruktu.

1. Wektory i ich mapy

Wektory plazmidowe powstały przez wprowadzenie modyfikacji do plazmidów – naturalnie występujących u bakterii kolistych cząsteczek DNA. Wektory plazmidowe zawierają zwykle następujące elementy:

- miejsce ułatwiające klonowanie, tzw. polilinker – odcinek w wektorze zaprojektowany z myślą o umieszczeniu w nim wstawki. Zawiera on miejsca rozpoznawane przez wiele różnych enzymów restrykcyjnych. W wektorach ekspresyjnych (służących do wyrażania białka w komórce gospodarza) polilinker jest poprzedzony przez promotor (do którego przyłącza się polimeraza RNA) i operator (umożliwiający regulację transkrypcji), a za polilinkerem występuje terminator (odcinek, na którym kończy się transkrypcja). W polilinkerze mogą znajdować się także sekwencje umożliwiające dodanie znaczników (ang. *tags*) – łańcuchów peptydowych ułatwiających oczyszczanie i identyfikację rekombinowanego białka. Funkcję znacznika mogą pełnić duże polipeptydy, np. GFP (białko zielonej fluorescencji) czy GST (*S*-transferaza glutationowa), albo mniejsze, np. His-tag, czyli ciąg sześciu reszt histydynowych.
- marker selekcyjny – fragment DNA umożliwiający odróżnienie komórek gospodarza, które pobrały plazmid, od tych, które tego nie zrobiły. Najczęściej jest to gen warunkujący oporność bakterii na antybiotyk.
- miejsce początku replikacji – umożliwia powielanie wektora w komórce gospodarza.
- mogą też zawierać dodatkowe elementy, jak np. represor operonu laktozowego, który zapobiega produkcji białka przed indukcją.

pET-28a-c(+) Vectors

TB074 12/98

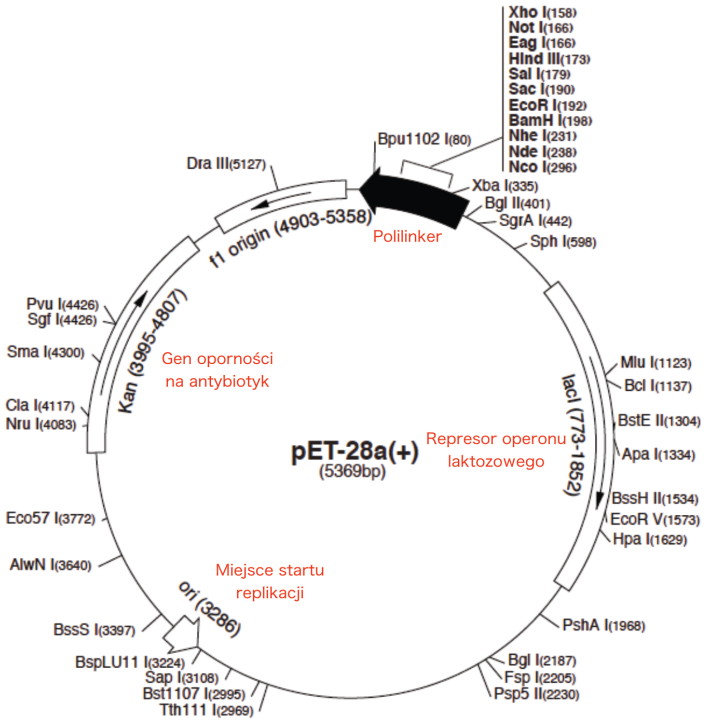
	Cat. No.
pET-28a DNA	69864-3
pET-28b DNA	69865-3
pET-28c DNA	69866-3

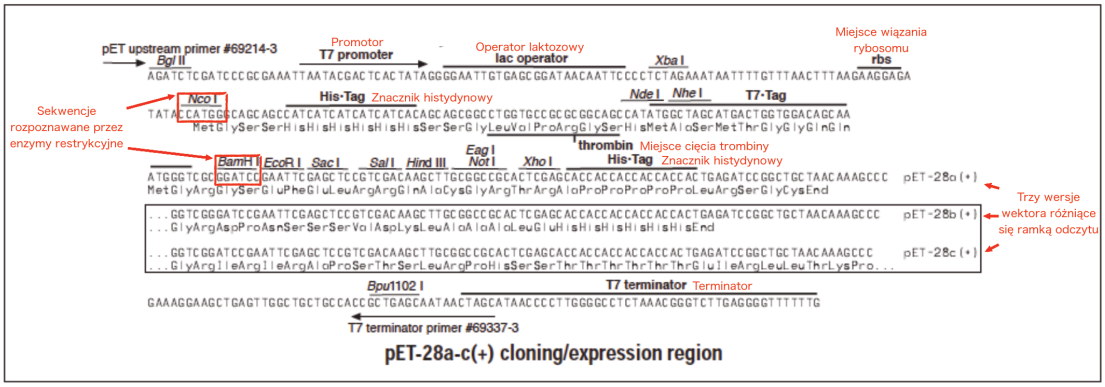
The pET-28a-c(+) vectors carry an N-terminal His•Tag®/thrombin/T7•Tag® configuration plus an optional C-terminal His•Tag sequence. Unique sites are shown on the circle map. Note that the sequence is numbered by the pBR322 convention, so the T7 expression region is reversed on the circular map. The cloning/expression region of the coding strand transcribed by T7 RNA polymerase is shown below. The f1 origin is oriented so that infection with helper phage will produce virions containing single-stranded DNA that corresponds to the coding strand. Therefore, single-stranded sequencing should be performed using the T7 terminator primer (Cat. No. 69337-3).

Najważniejsze elementy wektora wraz lokalizacją

pET-28a(+) sequence landmarks	
T7 promoter	370-386
T7 transcription start	369
His•Tag coding sequence	270-287
T7•Tag coding sequence	207-239
Multiple cloning sites (*Bam*H I - *Xho* I)	158-203
His•Tag coding sequence	140-157
T7 terminator	26-72
lacI coding sequence	773-1852
pBR322 origin	3286
Kan coding sequence	3995-4807
f1 origin	4903-5358

The maps for pET-28b(+) and pET-28c(+) are the same as pET-28a(+) (shown) with the following exceptions: pET-28b(+) is a 5368bp plasmid; subtract 1bp from each site beyond *Bam*H I at 198. pET-28c(+) is a 5367bp plasmid; subtract 2bp from each site beyond *Bam*H I at 198.





Rys. 2. Mapa komercyjnie dostępnego wektora z serii pET28 (oferowanego przez firmę Novagen).

Schematycznie przedstawienie powyższych miejsc w wektorze nazywa się mapą wektora (Rys. 2). W dokumentacji często przedstawia się też w sposób szczegółowy sekwencję polilinkera, żeby ułatwić zaplanowanie klonowania. W wektorze z serii pET28 występuje np. fragment kodujący miejsce cięcia przez trombinę – wysoce specyficzną proteazę. Jest ono wprowadzone po to, aby w razie potrzeby z otrzymanego rekombinowanego białka można było w sposób specyficzny odciąć His-tag znajdujący się na końcu N rekombinowanego białka.

Warto zwrócić uwagę na to, że wektor pET28 występuje w trzech wersjach: pET28a, pET28b i pET28c. Różnią się one liczbą dodatkowych nukleotydów znajdujących się między fragmentem kodującym T7-tag a sekwencją rozpoznawaną przez enzym restrykcyjny BamHI. pET28c jest podstawową wersją, zaś w pET28b i pET28a są dodane, odpowiednio, jeden i dwa nukleotydy. Powoduje to przesunięcie ramki odczytu, dlatego dysponując kolekcją trzech wektorów z serii pET28, zawsze można wybrać właściwy, który pozwoli na utworzenie konstruktu wyrażającego rekombinowane białko kodowane przez wstawkę w fuzji z His-tag czy T7-tag.

2. Sekwencje kodujące białka

Geny kodujące białka przechowuje się w bazach danych w postaci sekwencji mRNA. W bioinformatyce jednak zamiast znaku „U” oznaczającego uracyl używa się „T” oznaczającego tyminę, niezależnie od tego czy zapis dotyczy sekwencji DNA czy RNA. Aby w sekwencji mRNA odnaleźć odcinek kodujący białko, trzeba odszukać otwartą ramkę odczytu, a więc taki fragment sekwencji, który zaczyna się od kodonu START (ATG) i kończy jednym z kodonów STOP (TAA, TAG, TGA).

Jednym z formatów służącym do przechowywania sekwencji nukleotydowych i aminokwasowych jest format FASTA. W tym formacie nagłówki zawierające opis sekwencji zaczynają się od znaku większości „>”, a następnie zawierają dowolne znaki do końca linii. W kolejnych liniach znajduje się sekwencja zapisana w kierunku od końca 5’ do końca 3’ w przypadku sekwencji nukleotydowych oraz od końca N do końca C dla sekwencji aminokwasowych (w obydwu przypadkach odpowiada to naturalnemu kierunkowi syntezy w komórce). Nukleotydy oznacza się literami A, T, C oraz G, a reszty aminokwasowe za pomocą jednoliterowych skrótów nazw aminokwasów. W jednym pliku może znajdować się więcej niż jedna sekwencja, ale każda musi mieć osobny nagłówek znajdujący się w nowej linii. Warto zwrócić uwagę, że w przypadku DNA, podawana jest sekwencja tylko jednej z nici. Pliki można redagować za pomocą dowolnego edytora tekstu, np. Notatnika Windows. Najbezpieczniej jest unikać polskich znaków w nagłówkach sekwencji i korzystać tylko ze znaków ASCII – nie wszystkie programy są w stanie poprawnie

zinterpretować pliki zapisane w coraz popularniejszym kodowaniu UTF-8, umożliwiającym zapis znaków diakrytycznych.

Przykładowa sekwencja nukleotydów w formacie FASTA wygląda następująco:

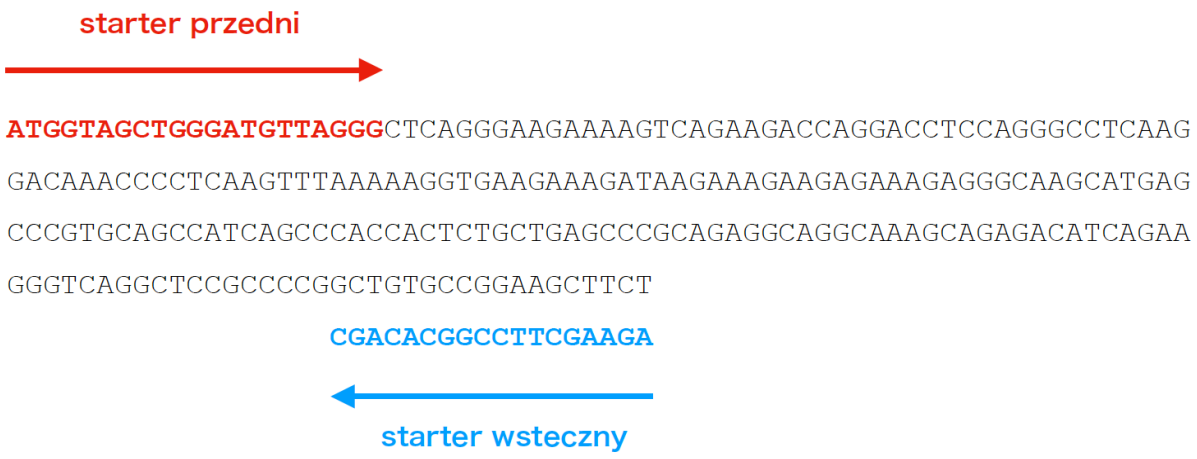
```
>Fragment DNA kodujacy His-tag|Wektor pET28a
CATCATCATCATCATCAC
```

Przykładowy sekwencja aminokwasowa w formacie FASTA wygląda następująco:

```
>His-tag|Wektor pET28a
HHHHHH
```

3. Dopasowanie starterów

Do przeprowadzenia reakcji PCR potrzebna jest para starterów, które wyznaczają początek i koniec powielanego fragmentu. Starter wyznaczający lewy kraniec sekwencji (od końca 5') nazywa się przednim (ang. *forward*) i odpowiada początkowi sekwencji, starter po prawej stronie (od końca 3') nazywa się wstecznym (ang. *reverse*) i jest komplementarny do końca sekwencji, którą chcemy powielić (Rys. 3). Do pewnych zastosowań projektuje się startery z tzw. over-hangami, które są dodatkową częścią startera, niehybrydyzującą z matrycowym DNA. Niżej umieszczona instrukcja odnosi się do podstawowej części startera, tej hybrydyzującej z matrycą.



Rys. 3. Startery do reakcji PCR. Strzałki wskazują kierunek 5'→3'.

Wybierając startery do klonowania, należy kierować się następującymi zasadami:

1. Długość części podstawowej (hybrydyzującej z matrycą) powinna wynosić od 18 do 25 nukleotydów.
2. Zawartość par GC w części podstawowej powinna wynosić 50–60%.
3. Temperatura przyłączania się starterów powinna zawierać się w przedziale od 50 do 65°C.
4. Różnica między temperaturami przyłączenia starterów do matrycy nie powinna być większa niż 5°C.
5. Na końcu 3’ startera powinna znaleźć się guanina lub cytozyna. Zapewnia to stabilniejsze wiązanie z matrycowym DNA.
6. Sekwencje starterów podaje się od końca 5’ do końca 3’ (zarówno startera przedniego, jak i wstecznego!)

4. Wybór enzymów restrykcyjnych

Do klonowania wykorzystuje się enzymy restrykcyjne, których miejsca cięcia występują w cząsteczce wektora tylko raz i są obecne w polilinkerze. Chociaż klonowanie można przeprowadzić wykorzystując tylko jeden enzym, lepiej jest użyć pary enzymów. Zapobiega to autoligacji wektora lub wstawki (zamknięcie się w kolistą cząsteczkę bez udziału innej), a także wstawieniu wstawki do wektora w niewłaściwej orientacji. Wybierając enzymy, trzeba uwzględnić ich położenie w polilinkerze. Niektóre miejsca dla enzymów są położone tak, żeby posłużyć do przygotowania konstruktu ze znacznikiem na końcu N białka, inne są lepsze, jeśli chcemy dołączyć znacznik na końcu C. Ponadto należy sprawdzić czy miejsca cięcia enzymów, które planuje się wykorzystać nie występują w sekwencji wstawki. Powinno się także wziąć pod uwagę warunki przeprowadzania trawienia. Najwygodniej jest używać enzymów, które wykazują maksimum aktywności w podobnych warunkach (najlepiej w tym samym buforze reakcyjnym i w tej samej temperaturze).

Miejsca rozpoznawane przez enzymy restrykcyjne można dodać na końcu 5’ startera. Ta część nie będzie hybrydyzować z matrycą i nazywa się over-hangiem. Trzeba też pamiętać o tym, że wiele enzymów restrykcyjnych nie potrafi przecinać DNA blisko krańca cząsteczki. Dlatego oprócz miejsca restrykcyjnego na 5’ końcu należy dodać kilka dodatkowych reszt nukleotydowych (dokładna liczba zależy od rodzaju enzymu, sekwencja może być dowolna, ponieważ i tak zostanie odcięta w wyniku działania enzymu restrykcyjnego).

5. Przygotowanie sekwencji gotowego konstruktu

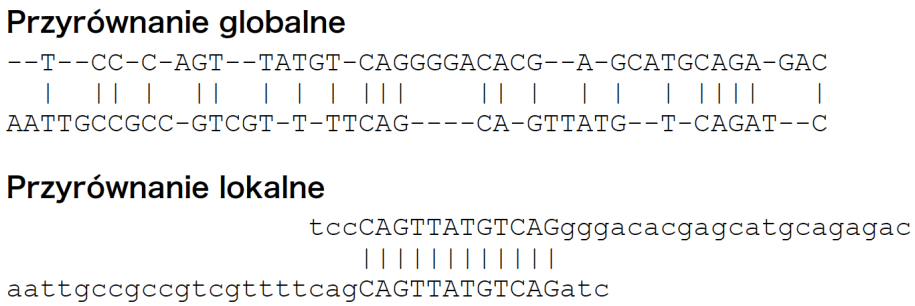
Aby przygotować sekwencję gotowego konstruktu, należy odnaleźć w sekwencji wektora miejsca rozpoznawane przez wybrane wcześniej enzymy restrykcyjne, wyciąć fragment sekwencji znajdujący się pomiędzy nimi, a następnie wkleić fragment DNA, który planujemy sklonować w odpowiedniej orientacji.

Przebieg klonowania *in silico* możesz obejrzeć w poradniku filmowym dostępnym na kanale YouTube Komitetu Głównego Olimpiady Biologicznej.

6. Przyrównywanie sekwencji

Występowanie w DNA, RNA lub białkach bardzo podobnych do siebie rejonów może być konsekwencją wspólnego pochodzenia lub pełnienia podobnej funkcji (konwergencja). Odnajdywanie takich miejsc opiera się na przyrównywaniu (ang. *alignment*) ze sobą sekwencji. Przyrównanie polega na wprowadzeniu do sekwencji przerw (ang. *gaps*), aby po ustawieniu sekwencji jedna pod drugą takie same lub podobne reszty nukleotydowe lub aminokwasowe znajdowały się w tych samych pozycjach. Pozwala to na łatwą identyfikacje podobnych do siebie miejsc w sekwencjach.

Oczywiście większość sekwencji daje się do siebie przyrównać, jeśli wprowadzić do nich odpowiednio dużą liczbę przerw. Dlatego, aby ocenić stopień podobieństwa dwu sekwencji algorytmy do przyrównywania przyznają punkty za dopasowane reszty i odejmują je za wprowadzone przerwy. Istnieje wiele algorytmów służących do przyrównywania sekwencji, ale można wśród nich wyróżnić dwie główne metody: globalne i lokalne. Przyrównanie lokalne służy do odnajdowania identycznych lub bardzo podobnych fragmentów w obrębie większych sekwencji. Z kolei przyrównanie globalne polega na jak najlepszym dopasowaniu sekwencji na całej długości (Rys. 4).



Rys. 4. Porównanie przyrównania globalnego i lokalnego na przykładzie pary tych samych sekwencji.

Algorytmy do przyrównywania sekwencji można wykorzystać, przeszukując bazę danych interesująca nas sekwencją, aby odnaleźć najbardziej podobne sekwencje przechowywane w bazie danych. Może to pomóc w identyfikacji funkcji i pochodzenia ewolucyjnego badanego fragmentu DNA lub białka.

Jednym z programów, który służy do porównywania sekwencji nukleotydowych i aminokwasowych z sekwencjami przechowywanymi w bazie danych jest BLAST. Są różne wersje tego programu:

- *Nucleotide BLAST* – porównuje sekwencję nukleotydową z sekwencjami nukleotydowymi z bazy danych;
- *Protein BLAST* – porównuje sekwencję aminokwasową z sekwencjami aminokwasowymi z bazy danych;
- *blastx* – tłumaczy sekwencję nukleotydową na aminokwasową w trzech różnych ramkach odczytu i porównuje z sekwencjami aminokwasowymi z bazy danych;
- *tblastn* – tłumaczy sekwencję aminokwasową na nukleotydową w różnych wariantach, które mogą kodować podaną przez użytkownika sekwencję aminokwasową i porównuje z sekwencjami nukleotydowymi z bazy danych

Więcej o programie BLAST dowiesz się z poradnika filmowego dostępnego na kanale YouTube Komitetu Głównego Olimpiady Biologicznej.

7. Przyrównania wielu sekwencji

Przyrównanie wielu sekwencji pozwala wykryć pokrewieństwo ewolucyjne sekwencji oraz na identyfikację motywów lub pojedynczych reszt zachowanych w ewolucji. Takie fragmenty często pełnią istotne funkcje, w przeciwieństwie do reszt, które zmieniają się w szybszym tempie.

Do przyrównywania wielu sekwencji służy np. program Clustal (Rys. 5).

Więcej o programie Clustal dowiesz się z poradnika filmowego dostępnego na kanale YouTube Komitetu Głównego Olimpiady Biologicznej.

```
sp|Q6NVM0|H10_XENTR      MTENSAAAPAGKPKRSKASKKATDHPKYSDMILAAVQAEKSRSGSSRQSIQKYIKNHYKV   60
sp|P10922|H10_MOUSE      MTENSTSAPAAKPKRAKASKKSTDHPKYSDMIVAAIQAEKNRAGSSRQSIQKYIKSHYKV   60
sp|P43278|H10_RAT        MTENSTSTPAAKPKRAKAAKKSTDHPKYSDMIVAAIQAEKNRAGSSRQSIQKYIKSHYKV   60
sp|P07305|H10_HUMAN      MTENSTSAPAAKPKRAKASKKSTDHPKYSDMIVAAIQAEKNRAGSSRQSIQKYIKSHYKV   60
sp|Q5NVN9|H10_PONAB      MTENSTSAPAAKPKRAKASKKSTDHPKYSDMVVAAIQAEKNRAGSSRQSIQKYIKSHYKV   60
                         *****:::**.****:**:**:*********::**:****.*:***********.****

sp|Q6NVM0|H10_XENTR      GENADSQIKLSIKRLVTSGTLKQTKGVGASGSFRLAKADEGKKPA--KKPKKEIKKAASP   118
sp|P10922|H10_MOUSE      GENADSQIKLSIKRLVTTGVLKQTKGVGASGSFRLAKGDEPKRSVAFKKTKKEVKKVATP   120
sp|P43278|H10_RAT        GENADSQIKLSIKRLVTTGVLKQTKGVGASGSFRLAKGDEPKRSVAFKKTKKEVKKVATP   120
sp|P07305|H10_HUMAN      GENADSQIKLSIKRLVTTGVLKQTKGVGASGSFRLAKSDEPKKSVAFKKTKKEIKKVATP   120
sp|Q5NVN9|H10_PONAB      GENADSQIKLSIKRLVTTGVLKQTKGVGASGSFRLAKSDEPKKSVAFKKTKKEIKKVATP   120
                         *****************:*.*****************.** *: .  ** ***:**.*:*

sp|Q6NVM0|H10_XENTR      KKAAKPKKAAKSPAKAKKPKVAEKKVKKPAKKKPAPSPKKAKKTKTVKAKPVRASRVKKA   178
sp|P10922|H10_MOUSE      KKAAKPKKAASKAPS-KKPKA---TPVKKAKKKPAATPKKAKKPKVVKVKPVKASKPKKA   176
sp|P43278|H10_RAT        KKAAKPKKAASKAPS-KKPKA---TPVKKAKKKPAATPKKAKKPKIVKVKPVKASKPKKA   176
sp|P07305|H10_HUMAN      KKASKPKKAASKAPT-KKPKA---TPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKA   176
sp|Q5NVN9|H10_PONAB      KKASKPKKAASKAPT-KKPKA---TPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKA   176
                         ***:******..  . ****.   .  * **** * :****** * **.***:**: ***

sp|Q6NVM0|H10_XENTR      KPSKPKAKASPKKSGRKK     196
sp|P10922|H10_MOUSE      KTVKPKAKSSAKRASKKK     194
sp|P43278|H10_RAT        KPVKPKAKSSAKRASKKK     194
sp|P07305|H10_HUMAN      KPVKPKAKSSAKRAGKKK     194
sp|Q5NVN9|H10_PONAB      KPVKPKAKSSAKRAGKKK     194
                         *  *****:* *::.:**
```

Rys. 5. Przyrównanie wielu sekwencji aminokwasowych. Gwiazdką „*” zaznaczono miejsca, w których wszystkie reszty są identyczne. Dwukropkiem „:”, w których jedna reszta jest bardzo podobna do pozostałych, a kropką „.”, w których jedna reszta jest podobna w niewielkim stopniu do pozostałych. Symbol „–” oznacza przerwę wprowadzoną przez program w celu uzyskania najlepszego przyrównania.

8. Drzewa filogenetyczne

Zakładając, że obserwowanie różnice między dwoma sekwencjami są proporcjonalne do czasu jaki upłynął od rozdzielenia się linii ewolucyjnych, a więc do czasu występowania ostatniego wspólnego przodka dwóch organizmów, analiza porównywanych sekwencji może

posłużyć do konstruowania drzew filogenetycznych metodą UPGMA. Więcej o szacowaniu filogenezy możesz dowiedzieć się z części statystyczno-filogenetycznej poradnika.

9. Bazy danych zawierające informacje o białkach

UNIPROTKB

UniProtKB to baza danych sekwencji białkowych. Składa się z dwóch części: UniProtKB/Swiss-Prot i UniProtKB/TrEMBL. UniProtKB/Swiss-Prot to baza danych nadzorowana, co oznacza, że informacje, które się w niej znajdują zostały sprawdzone przez ekspertów z dziedziny biologii molekularnej i medycyny. Można w niej znaleźć sekwencje aminokwasowe znanych białek, krótki opis funkcji i budowy domenowej białka, miejsc oddziaływania i modyfikacji posttranslacyjnych oraz lokalizacji w komórce i tkankach wraz z odniesieniami do literatury. UniProtKB/TrEMBL to baza danych, w której znajdują się sekwencje jeszcze niesprawdzone przez specjalistów i nie tak dokładnie opisane.

Adres bazy danych: www.uniprot.org

PDB Protein Data Bank

PDB to baza danych zawierająca informacje o trójwymiarowych strukturach dużych cząsteczek o znaczeniu biologicznym, głównie białek, kwasów nukleinowych oraz ich kompleksów. Informacje o strukturze makrocząsteczek zdobywane są za pomocą krystalografii rentgenowskiej, spektroskopii NMR lub kriomikroskopii elektronowej. Obecnie w PDB zdeponowanych jest ponad 150 000 struktur.

Ważną częścią PDB jest portal edukacyjny przybliżający zagadnienia związane z biologią strukturalną uczniom liceum i wszystkim zainteresowanym.

Adres bazy danych: www.rcsb.org

Adres portalu edukacyjnego: pdb101.rcsb.org

Pfam

Pfam to baza danych rodzin białkowych, czyli grup tworzonych przez spokrewnione ze sobą białka. Rodziny białkowe są identyfikowane przez przyrównanie i analizę wielu sekwencji.

Adres bazy danych: pfam.xfam.org