Przegląd Medyczny Uniwersytetu Rzeszowskiego
Rzeszów 2005, 3, 284–293

Damian Mazur

Zastosowanie systemów informatycznych w medycynie – bioinformatyka

Z Politechniki Rzeszowskiej, Zakład Podstaw Elektrotechniki i Informatyki

W pracy opisano Toolbox Bioinformatica wchodzący w skład oprogramowania MATLAB oraz jego zastosowanie do przetwarzania ogromnych zasobów baz danych zawierających informacje na temat różnego rodzaju DNA, enzymów, genów itp. Toolbox ten umożliwia przeszukiwanie baz danych DNA w poszukiwaniu podobnych sekwencji kodu, poszukiwanie różnic poszczególnych organizmów i określenia ich pochodzenia. Zawiera narzędzia do analizowania wirusów, chorób. Pozwala badać, do tej pory, nieuleczalne choroby, takie jak choroba Parkinsona, lub choroba raka (różnych organów). W tym celu wykorzystuje się specjalistyczne języki programowania, np: BioPerl, BioJava, BioPython, BioRuby, BioPipe, BioSQL/OBDA

Słowa kluczowe: systemy informatyczne, MATLAB, bioinformatyka

Application of informatics systems in medicine- bioinformatics

Special This paper discusses Toolbox Bioinformatica included in MATLAB software as well as its application for the processing of very big data base resources containing information on various DNA types, enzymes, genes etc. This Toolbox makes it possible to scan DNA data base for similar code sequences, scan for differences between particular organisms and defining their origins. It gives tools for analysis of virii, and diseases, and makes it possible to study incurable since now, diseases, as, Parkinson disease, or cancer disease (of various organs). For this purpose, specialised programming languages are used e.g., BioPerl, BioJava, BioPython, BioRuby, BioPipe, BioSQL/OBDA.

Key words: application of informatics, MATLAB, Bioinformatyka

WPROWADZENIE

Od momentu pojawienia się pierwszych komputerów, zaczęły one sukcesywnie wchodzić w każdą dziedzinę życia. Wraz z bardzo szybkim wzrostem ich mocy obliczeniowej, zakres ich zastosowania oraz efektywność rosła coraz bardziej. Szczególnie przydatne okazało się połączenie nauk biologicznych z komputerem. Najbardziej znanym zastosowaniem komputerów, od którego zaczęła się poważna symbioza biologii (a właściwie genetyki) i maszyn cyfrowych, był Human Genom Project (HGP). Było to przedsięwzięcie mające na celu zsekwencjonowanie ludzkiego genomu. Wymagało to olbrzymich ilości danych, które trzeba było gdzieś przechowywać i przetwarzać. Z pomocą przyszły nowoczesne superkomputery i sieć Internetu. Aby uzmysłowić sobie ogrom informacji, z jakim należało pracować przy HGP, przytoczę dość obrazowy opis 'rozmiaru' ludzkiego genomu. Autorem jest Irena Roterman-Konieczna, z Zakładu Biostatystyki i Informatyki Medycznej – Collegium Medicum – Uniwersytetu Jagiellońskiego:

Podwójna nić DNA człowieka zawiera 3 miliardy par nukleotydów. Gdyby cały DNA człowieka wymodelować za pomocą zamka błyskawicznego, przyjmując, że 1 ząbek odpowiadałby jednemu nukleotydowi, a 300 ząbków przypadałoby na 1 m długości zamka, to zamek taki powinien mieć długość 10 000 km (odległość z Warszawy do Nowego Yorku i z powrotem).

Naukowcy mają nadzieję, że dzięki HGP w przyszłości będziemy mogli leczyć choroby i patologie wpływające niekorzystnie na nasze zdrowie, jeszcze przed ich powstaniem. Wszystkie zaś pojawiające się choroby będziemy usuwać na poziomie molekularnym, mogąc dokładnie określić, co leży u podstaw ich powstania i działania. W niezależną dyscyplinę badawczą bioinformatyka przekształciła się w chwili, kiedy wykorzystując zmagazynowane informacje, potrafiła przetworzyć je tak, aby wykreować własne modele procesu przesyłania informacji w organizmie żywym. Stało się to możliwe w odniesieniu do DNA w momencie, kiedy opracowane zostały specjalistyczne programy komputerowe przeznaczone do analizy uzyskanych eksperymentalnie sekwencji nukleotydów.

O tym, że metody bioinformatyki już zaczynają przynosić wymierne efekty, niech świadczy choćby istnienie enzymu cathepsin K., który być może będzie mógł walczyć z osteoporozą, chorobą kości dotykającą miliony ludzi. Na początku 1990 r. naukowcy z GlaxoSmithKline wyizolowali materiał genetyczny z komórek ludzi chorych na raka kości. Przekazali ten materiał do testów firmie, która brała udział w HGP, a ta przeprowadziła szereg testów poszukujących protein, które geny z dostarczonych próbek kodowały. Jeden z takich testów wykazał bliskie podobieństwo pomiędzy komórkami raka kości, a wcześniej wyizolowanymi cząsteczkami nazwanymi cathepsins. W tej chwili trwają badania nad środkiem blokującym, skierowanym na ten cel.

1. Techniki i modele komputerowe do przetwarzania informacji biomedycznej

Bioinformatyka jest nową dziedziną nauk biologicznych, która stosuje techniki i modele komputerowe do przetwarzania informacji gromadzonej w licznych biologicznych bazach danych. W ostatnim okresie odnotowuje się lawinowy przyrost wyników i danych eksperymentalnych, które deponowane są w takich bazach. Przykładowo, wielkość bazy GenBank (gromadzącej sekwencje nukleotydów) czy SwissProt (sekwencje aminokwasowe) ulega podwojeniu średnio, co 15 miesięcy. O bogactwie i różnorodności gromadzonej informacji decyduje przede wszystkim wielkość i złożoność projektów badawczych, w których bada się ekspresję genów, strukturę kodowanych przez nie białek oraz sieć wzajemnych oddziaływań między biocząsteczkami. Bogactwo informacji dostępnej w biologicznych bazach danych sprowadza wiele wyzwań współczesnej biologii molekularnej i biotechnologii do wyzwań natury obliczeniowej. Stąd, podstawowym celem, przed jakim staje bioinformatyka jest wykorzystanie i rozwijanie technik i modeli komputerowych, zmierzające do zrozumienia i sklasyfikowania informacji związanej z biocząsteczkami.

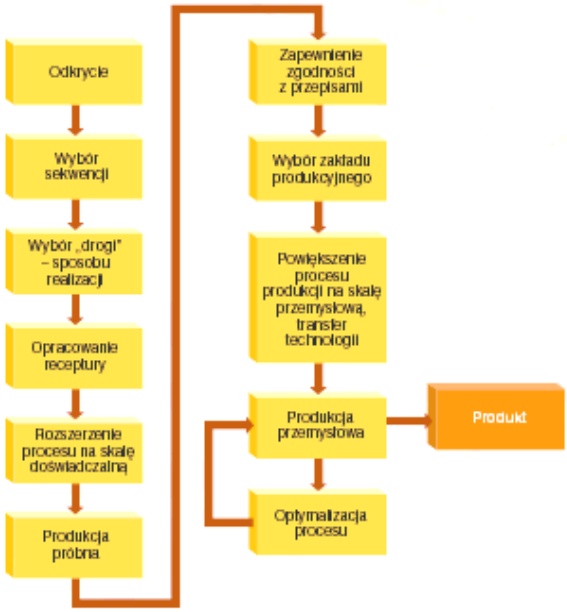
1. PODSTAWOWE ZAGADNIENIA BIOINFORMATYKI:

- Katalogowanie informacji biologicznych (bazy danych, wyszukiwanie sekwencji, adnotacji, danych numerycznych w bazach danych)
- Analiza sekwencji DNA (składanie sekwencji, adnotacja, wyszukiwanie sekwencji kodujących, regulatorowych i repetytywnych, motywów, markerów)
- Analiza sekwencji genomów, porównywanie genomów
- Ustalanie ewolucyjnych relacji pomiędzy zbiorami sekwencji / organizmów
- Genotypowanie (używane między innymi do wyszukiwania genów odpowiedzialnych za choroby genetyczne, w ustalaniu ojcostwa, kryminalistyce)
- Analiza ekspresji genów (głównie ang. microarrays analysis)
- Analiza sekwencji białek (porównywanie sekwencji, wyszukiwanie domen i motywów, przewidywanie własności fizyko-chemicznych, drugo- i trzeciorzędowej struktury białka, lokalizacji w obrębie komórki)
- Katalogowanie funkcji genów/białek, analiza dróg metabolicznych (np metabolizm lipidów) oraz dróg sygnałowych (np od receptora na powierzchni komórki poprzez kaskadę kinaz do czynników transkrypcyjnych)
- Modelowanie układów biologicznych (np. kinetyka szeregu reakcji enzymatycznych w komórce)
- virtual docking - np. używając trójwymiarowej struktury aktywnego centrum enzymu (pocket) przeszukuje się w komputerze tysiące małych cząsteczek, z których kilka-kilkanaście ('kluczy') będzie miało kształt mieszczący się w centrum aktywnym.

Pierwszy krok w kierunku odkrywania nowych leków.

2. MORFOMETRIA - ANALIZA OBRAZU

Łańcuch wartości innowacji i technologii rozpoczyna się z odkryciem nowej molekuły. Firmy biotechnologiczne inwestują znaczne środki, by poszerzać swoją wiedzę w dziedzinach takich, jak



Źródło: Cancer Engineering w opraciu o nane z firmy Aspan Technology

RYC. 1. W przemyśle biotechnologicznym przepływ informacji o procesie opracowywania nowego produktu zazwyczaj postępuje według powyższego schematu.

rekombinacja DNA i wiedza o wzorcach genetycznych, tzw. genomika.

Aby zwiększyć tempo odkrywania nowych jednostek chemicznych, kilka firm farmaceutycznych poszerzyło swoje zaplecza badawczo-rozwojowe przez podjęcie współpracy z firmami specjalizującymi się w wykonywaniu badań na zlecenie, takimi jak Avantium Technologies.

Aby coraz skuteczniej pomagały klientom, oddziały farmaceutyczne firmy Avantium zostały ostatnio wyposażone w urządzenia do klasyfikowania związków chemicznych klasy IV, do badania powstawania soli i właściwości polimorficznych (związków). Są to udogodnienia, na których samodzielne opracowanie i utrzymywanie może sobie pozwolić niewiele firm. Wszystko to ma na celu pomoc w zredukowaniu stanu przejściowego pomiędzy etapem badawczo--rozwojowym a produkcją.

Po odkryciu nowej cząsteczki decydującym elementem wprowadzania produktu na rynek jest zazwyczaj opracowanie samego produktu i procesu produkcji. Wiele elementów procesu jest realizowanych równocześnie: odkrycie, testowanie biologiczne, przygotowanie do doświadczalnego zastosowania nowego leku; próby kliniczne oraz jego zastosowanie właściwe. Wszystkie te elementy składają się na proces przygotowania leku, jak to widać na schemacie.

Firmy biotechnologiczne, które z powodzeniem ujednoliciły i skompresowały przebieg procesu, donoszą o oszczędnościach produkcyjnych rzędu 40% w przypadku procesów zoptymalizowanych w sposób utrzymujący taką samą ścieżkę syntezy. Oszczędności produkcyjne wzrastają aż do 65% w przypadku procesów zoptymalizowanych tak, by możliwe były zmiany ścieżki syntezy. Zachęciło to kolejne firmy biotechnologiczne do zwiększenia środków na optymalizację procesu wprowadzania produktu na rynek oraz „mechanikę" transferu wiedzy i technologii. Chociaż między działami badawczo-rozwojowymi a produkcją mogą pojawiać się rozbieżności terminologiczne, największe wyzwania związane z transferem wiedzy i technologii to:

- Zapewnienie stosowania zgodnych danych, metod, modeli i narzędzi (takich jak obliczanie gęstości czy pojemności cieplnej;
- Zagwarantowanie bezpiecznego zwiększania skali danego procesu, szczególnie jeśli są w nim używane rozpuszczalniki;
- Zidentyfikowanie wpływu ciepła, światła i/lub pobudzenia fizycznego na przechowalność produktu;
- Ustalenie właściwych działań procesowych (wstrząsanie zamiast mieszania czy tłuczenia lub określanie właściwych prędkości mielenia niezbędnych do uzyskania cząstek o właściwej wielkości);
- Określenie wpływu wielkości partii oraz czasu przygotowania na jakość produktu (czas podgrzewania lub chłodzenia laboratoryjnej, pilotowej partii w laboratorium zakładu doświadczalnego w porównaniu z czasem podgrzewania bądź chłodzenia dużej, przemysłowej partii

produktu w produkcyjnym zakładzie przemysłowym);
- Zapewnienie, że możliwości sprzętowe są odpowiednie (wielkość chłodziarek jest odpowiednia do wymogów lub możliwości odzysku rozpuszczalników).

3. OMÓWIENIE OPROGRAMOWANIA GAUSSIAN 03

Omówienie oprogramowania gaussian 03 – jest systemem przeznaczonym do obliczeń orbitali molekularnych przy użyciu metod półempirycznych i ab initio. Gaussian 03 służy do określania wielu własnosci molekuł i reakcji, włączając w to: energię i strukturę molekularną, energię i strukturę stanów przejściowych, częstotliwość drgań, widma IR i Ramana, własności termochemiczne, energię wiązań i reakcji, orbitale molekularne, ładunki atomowe, momenty multipolowe, osłanianie NMR i wrażliwość magnetyczną, powinowactwo elektronowe i potencjały jonizacyjne, polaryzowalność i hyperpolaryzowalność, potencjały elektrostatyczne i gęstość elektronową.

Obliczenia mogą być prowadzone na układach w stanie gazowym lub roztworach, w stanie podstawowym lub stanie wzbudzonym. Gaussian 03 jest potężnym narzędziem do analizy wpływu podstawników, badania mechanizmów reakcji, obliczeń powierzchni energii potecjalnej i energii wzbudzenia.

4. ANALIZA MATERIAŁU GENETYCZNEGO, GENÓW I KODOWANIE BIAŁEK W MATLABIE

Bioinformatics Toolbox – oprogramowanie umożliwiające wykorzystanie MATLABa do analizy materiału genetycznego, badania obecności genów i kodowania białek. Oferuje on obliczenia modelowania nowych algorytmów i prowadzenia badań nad lekami, badań z zakresu inżynierii genetycznej oraz innych projektów genomicznych i proteomicznych. Obecnie najnowsza wersja Toolboxa Bioinformatics jest w wersji 2.0 i wymaga do swego działania dodatkowo toolboxa Statistics Toolbox.

Poniżej znajdują się podstawowe zastosowania Toolboxa Bioinformatics:
- dostępność do internetowych baz danych z informacjami na temat DNA ludzi, oraz innych baz danych oferujących dane biologiczne służące do dalszych analiz
- potrafi się odwoływać i zapisywać do wielu źródeł i układów danych
- określa statystyczne cechy danych
- manipuluje i szereguje kolejności występowania danych
- modeluje wzory w kolejności biologicznej używając do tego sieci neuronowej opartej o model Markova (profil HMM)
- odczytuje, normalizuje i wizualizuje dane
- tworzy i manipuluje filogenetyczne dane drzewa
- zawiera wbudowany interfejs do komunikowania się z innym specjalistycznym oprogramowaniem bioinformatic (BioPearl i BioJava)

Przykładowe zastosowania bioinformaics.
Kolejność DNA

Analiza kolejności jest procesem, który jest używany, by znaleźć informację o nukleotydach albo kolejności aminokwasu używający różnych metod obliczeniowych. Zadanie polega na wyznaczeniu kolejności, analizie i identyfikacji genu, określa podobieństwo dwóch genów. Znajduje białka skłonne do łączenia się z różnymi genami. Funkcja ta pozwala określić podobieństwa danego genu z genem w innym danym i badanym organizmie.

Przykład zastosowany w Toolboxie:

a) Kolejność Dana:
- rozpoczynając od kolejności DNA, oblicza dana dla zawartości nukleotydów.

b) *Kolejność Wyrównanie*
- rozpoczynając od kolejności DNA dla ludzkiego genu, osiedla i weryfikuje odpowiednie geny w modelowym organizmie.

c) *Szacowanie znaczenia wyrównania*
- Ten przykład przedstawia metodę, która może zostać użyta do badania znaczenia kolejności wyrównań. Program Matlab uzyskuje dostęp do danych NCBI. W tym przykładzie program pracuje bezpośrednio z danymi białka, używa getgenpept zamiast getgenbank, by pobrać dane z NCBI. Matlab czyta informacje o ludzkim białku.

d) Wizualizacja danych Microarray

Służy do pokazania różnych dróg analizowania danych zawartych w Microarray. Przykład używa danych microarray do badania wyrażenia genu w mózgach myszy. Wykorzystuje metodę Multiplex trójwymiarowego wyrażenie genu mózgu, sporządzając mapę w modelu choroby Parkinsona u myszy. Dane są pobierane z bazy zawartej w Internecie.

e) Specjalistyczne języki programowania:
- Wraz z rosnącym wpływem komputerów na badania i przetwarzanie informacji, środowiska naukowe zaczęły opracowywać specjalistyczne narzędzia. Kwestią czasu było powstanie specjalizowanych języków programowania, bądź

też dodatków do języków już istniejących. Jedną z organizacji zajmujących się bioinformatyką jest Open Bioinformatics Foundation, która opiekuje się m.in. takimi projektami, jak: BioPerl, BioJava, BioPython, BioRuby, BioPipe, BioSQL/OBDA, MOBY.

- Bioperl – podając odpowiednie argumenty można dane przetworzyć i wyszukać za pomocą funkcji BLAST. Perl i moduły Bioperl można znaleźć na stronach: http://www.perl.com i http://bioperl.org/. Moduły pokazują proces szukania leku na chroniczną białaczkę, oraz leku pomocnego w leczeniu guzów (GIST). (Gleevec(tm) (STI571 albo mesylate imatinib) był najpierw chwalonym lekiem, który wyłączał gen odpowiedni za powstawanie raka białka, a następnie powodował chroniczną białaczkę myelogenous (CML)).
- Wykorzystanie oprogramowania SOAP. Używa algorytmu wyrównania danych pobranych z GenBank. Oprogramowanie pokazuje możliwość dostępu do baz NCBI z poziomu MATLABA za pomocą strony http://www.ncbi.nlm.nih.gov/
- Analiza klastera Microarray. Umożliwia użycie danych genów zawartych w tablicach Microarray do badania funkcji komórek, porównuje różnice między zdrową i chorą tkanką i pokazuje zmiany spowodowane stosowanie leków (podawanie leków dla organizmu, z którego pochodzą badane tkanki i materiał biologiczny). Przykładem ten służy do zaznajomienia się z możliwościami do analizowania i wizualizowania wzorów wyrażenia genów.

5. BAZY DANYCH, ANALIZA INFORMACJI GENETYCZNYCH

Badanie ewolucji gatunku:

Dzieje ewolucji organizmów są zapisane w ich genomach. W przeszłości ustalenie pokrewieństw ewolucyjnych było możliwe tylko dzięki analizie wpływu informacji genetycznej obserwowanego jako ujawniające się cechy organizmów. Historię biologiczną odtwarzano więc początkowo na podstawie badań morfologicznych skamieniałości oraz porównawczej anatomii, fizjologii i embriologii współcześnie żyjących gatunków. W połowie naszego stulecia, kiedy opracowano metody określania sekwencji aminokwasowej białek, innych wskaźników pokrewieństwa ewolucyjnego poszukiwano w chemii porównawczej białek. Niemal natychmiast stało się jasne, że pewne białka, które w różnych organizmach pełnią taką samą funkcję, mają również bardzo podobne sekwencje aminokwasowe. Odkrycie to dowiodło, że wiele różnych rodzajów organizmów miało wspólną przeszłość ewolucyjną, czego nie sposób ocenić na podstawie przypadkowych i rozproszonych danych pochodzących z badań skamieniałości.

Nasza obecna wiedza o strukturze DNA oraz procesach genetycznych jest całkowicie zgodna z koncepcją wspólnej historii ewolucyjnej żywych organizmów naszej planety. A co ważniejsze, wiedza ta dostarcza wskazówek, w jaki sposób zachodzi ewolucja biologiczna. Porównując strukturę i funkcję genów oraz organizację genomów różnych żyjących obecnie organizmów, można wydedukować, co działo się w przeszłości; podobnie jak detektyw, który na podstawie pozostawionych na miejscu zbrodni przedmiotów oraz znajomości, choćby pobieżnej, ludzkiego charakteru może zrekonstruować fakty.

Podczas gdy badania nad dawnymi formami życia będą nadal polegać na analizie morfologicznej skamieniałości, coraz ważniejszą metodą ustalania pokrewieństw między organizmami współczesnymi a ich przodkami staje się obecnie porównywanie sekwencji DNA.

Projektowanie lekarstw:

Dzięki modelowaniu struktury 3D białek na podstawie sekwencji aminokwasowych i analizie oddziaływania cząsteczek chemicznych z białkami naukowcy są w stanie projektować nowe leki. Spis najpopularniejszych narzędzi można znaleźć: http://www.expasy.org/ tools/#secondary

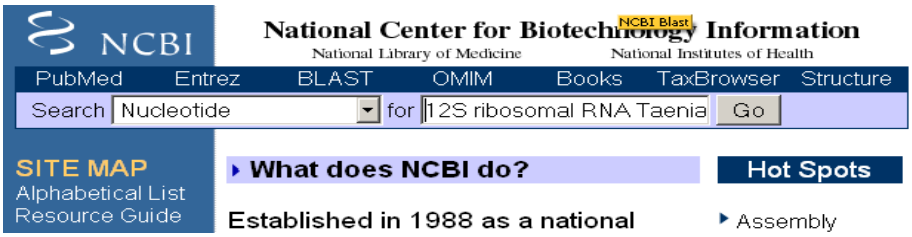
Badanie nowych genów:

Jeśli naukowiec dokona zsekwencjonowania jakiegoś genu, pierwsze, co robi, to analiza sekwencji – czyli porównanie genu z bazą. Robi się to najczęściej za pomocą programu BLAST (Basic Local Alingment Search Tool). Jeśli znajdziemy gen choć trochę podobny do naszego, mamy punkt wyjściowy do badań nad jego właściwościami.

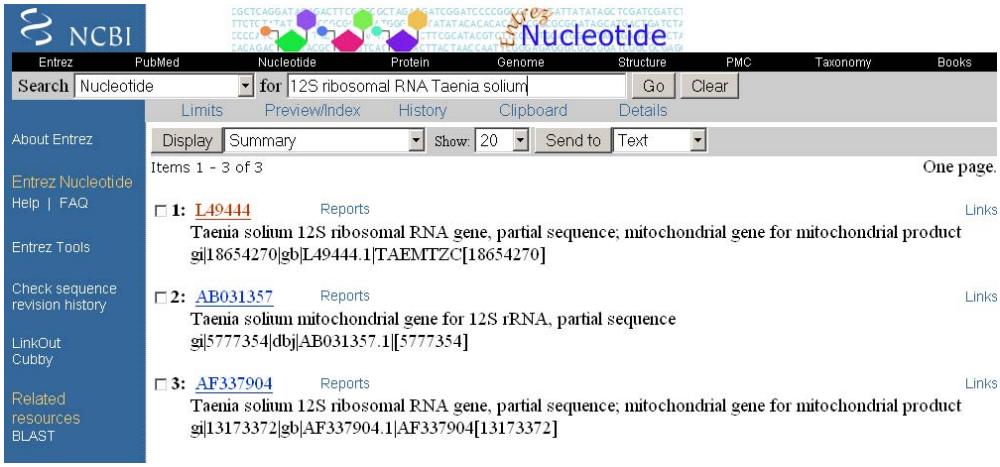
6. PRAKTYCZNE ZASTOSOWANIE NARZĘDZI BIOINFORMATYKI

Tworzenie metody diagnostycznej opartej na biologii molekularnej, do wykrycia schorzenia bakteryjnego albo zakażenia pasożytami wewnętrznymi. Poniżej przedstawiono algorytm postępowania:

- Dlaczego ta metoda? Metoda jest prosta do zrealizowania, szybka, tania i bardzo dokładna. Mało przykra dla pacjentów (nieinwazyjna, pobieramy tylko krew).



RYC. 2

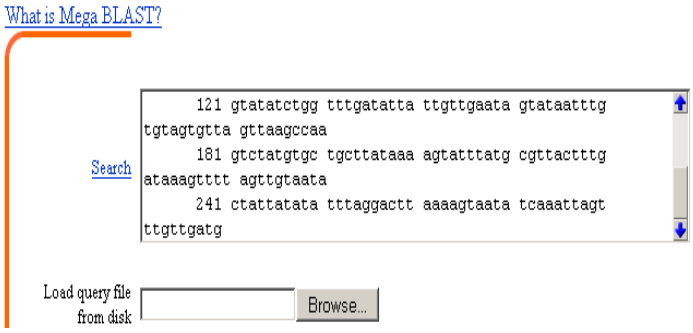


RYC. 3

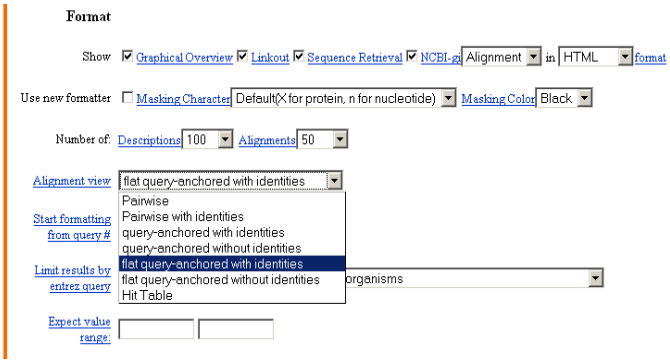
ORIGIN

1 gtattttgta atattatctt agtttagtaa ctaaaatggt ttggcagtga gtgattcttt
61 ttaggggaag gtgtagtgta aaggatgttc cgcctattaa tttacttta ttatgttggt
121 gtatatctgg tttgatatta ttgttgaata gtataatttg tgtagtgtta gttaagccaa
181 gtctatgtgc tgcttataaa agtatttatg cgttactttg ataaagtttt agttgtaata
241 ctattatata tttaggactt aaaagtaata tcaaattagt ttgttgatg

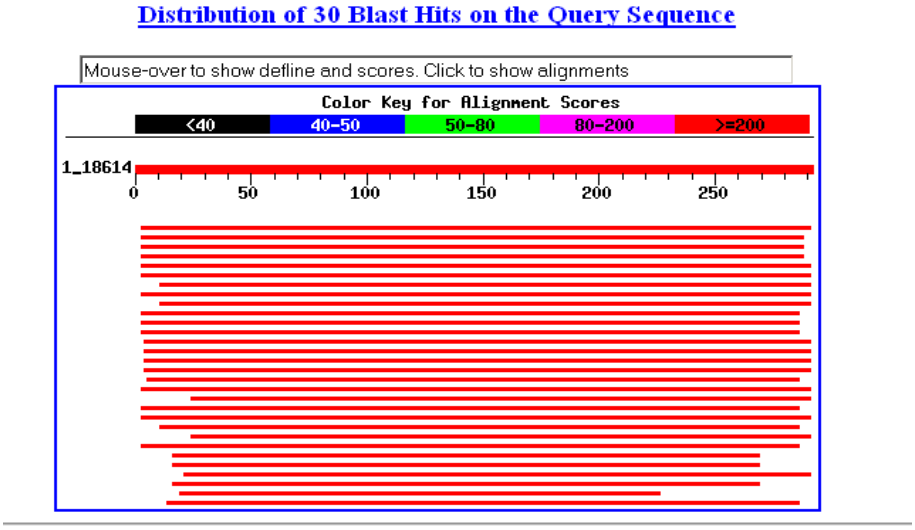
RYC. 4



RYC. 5



RYC. 6



RYC. 7

Score E

Sequences producing significant alignments: (bits) Value

gi|18654270|gb|L49444.1|TAEMTZC Taenia solium 12S ribosomal... 557 e-156

gi|21388751|dbj|AB086256.1| Taenia solium mitochondrial DNA... 545 e-152

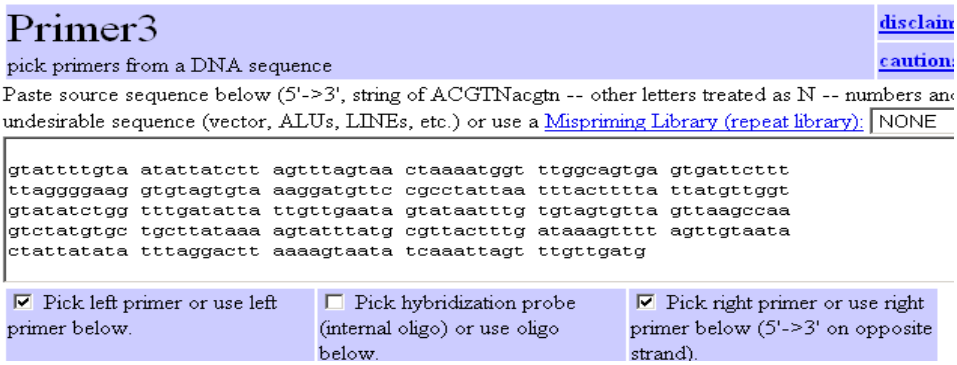
gi|13173372|gb|AF337904.1|AF337904 Taenia solium 12S riboso... 545 e-152

gi|5777354|dbj|AB031357.1| Taenia solium mitochondrial gene... 545 e-152

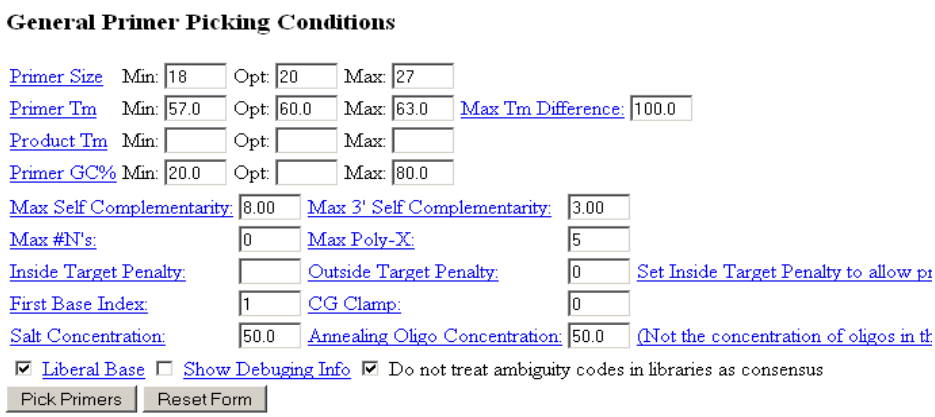
gi|5777352|dbj|AB031355.1| Taenia saginata mitochondrial ge... 437 e-120

gi|5777353|dbj|AB031356.1| Taenia asiatica mitochondrial ge... 430 e-117 [...]

RYC. 8



RYC. 9



RYC. 10

```
Primer3 Output
Using 1-based sequence positions
OLIGO          start  len     tm     gc%   any   3'  seq
LEFT PRIMER       37   20   59.14   45.00  3.00  1.00 tggtttggcagtgagtgatt
RIGHT PRIMER     193   21   58.21   47.62  5.00  2.00 gcagcacatagacttggctta
SEQUENCE SIZE: 289
INCLUDED REGION SIZE: 289
PRODUCT SIZE: 157, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 1.00
   1 gtattttgtaatattatcttagtttagtaactaaaatggtttggcagtgagtgattcttt
                          >>>>>>>>>>>>>>>>>>>>
  61 ttaggggaaggtgtagtgtaaaggatgttccgcctattaatttacttttattatgttggt
 121 gtatatctggtttgatattattgttgaatagtataatttgtgtagtgttagttaagccaa
                                    <<<<<<<<
 181 gtctatgtgctgcttataaaagtatttatgcgttactttgataaagttttagttgtaata
       <<<<<<<<<<<<<
 241 ctattatatatttaggacttaaaagtaatatcaaattagtttgttgatg
```

RYC. 11

Baza NCB – wyszukujemy sekwencję nukleotydową dla genu kodującego dane białko (12S ribosomal RNA Taenia solium) u danego organizmu (Teania solium – tasiemiec uzbrojony). (Operacje przeprowadzamy dla wszystkich podejrzanych gatunków) (ryc. 2).

Uzyskujemy trzy wyniki, wybieramy pierwszy z nich (ryc 3).

Otrzymujemy informację o sekwencji białka, którą przedstawiono na ryc. 4.

- Program BLAST służy do tego, by porównać dwa genomy – człowieka i zarazka oraz znaleźć taki fragment DNA, który będzie charakterystyczny dla danego pasożyta. Wykonuje się to, aby uzyskać 100% pewności, że wybrane do testu białko na pewno nie znajduje się w człowieku. Następnie wkleja się tę sekwencję do BLASTa (opcja Nucleotide)
- Quickly search for highly similar sequences (megablast) (ryc. 5).

Możemy wybrać jeden ze sposobów przedstawienia wyników naszego przeszukiwania (ryc. 6).

Wynikiem poszukiwania jest sekwencja, której na pewno nie znajdziemy w organizmie człowieka. Jest ona charakterystyczna dla organizmu obcego (Teania solium).(ryc. 7).

- Pod graficznym przedstawieniem otrzymujemy nazwy organizmów, u których występuje dana sekwencja. Możemy upewnić się, że na pewno nie jest to sekwencja występująca u człowieka, czyli jest dla nas sekwencją charakterystyczną. W naszym przypadku otrzymujemy głównie różne rodzaje tasiemców (ryc. 8).

Powyższe fragment sekwencji wykonujemy dla grupy pasożytów, którą podejrzewamy, że mogą być przyczyną tej infekcji.

- Następnie projektuje się tzw. startery, które służą do identyfikacji danego fragmentu w pobranym materiale genetycznym (służy temu np program Primer 3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). Startery muszą pasować do sekwencji oraz muszą mieć charakterystyczne parametry, które spowodują że nie będą się łączyły same z sobą.
- Startery są to krótkie fragmenty kwasu nukleotydowego, do którego przyczepia się enzym (polimeraza), aby móc dołączyć nukleotydy na zasadzie komplementarności zasad azotowych. Rozdzielony łańcuch DNA składa się z dwóch komplementarnych sekwencji, dzięki czemu wiemy, czego „brakuje" (ryc. 9).
- Ważnym parametrem jest sposób traktowania startera: jeśli wybierzemy dla wszystkich starterów podobne temperatury, za jednym razem będziemy mogli sprawdzić więcej możliwości oszczędzając czas. Warto też wybrać optymalną długość sekwencji (ryc. 10).

Klikamy „Pick Primers" i otrzymujemy opis naszego startera (ryc. 11).

Strzałki pod sekwencją oznaczają usytuowanie naszego startera w wybranej sekwencji.

- Łączymy ze sobą pobrane DNA, startery i enzym, który posłuży do dobudowywania cegiełek DNA. To wszystko dzieje się w Termocyklerze, który ustala temperatury i czas trwania kolejnych faz reakcji, by wszystko zaszło optymalnie.
- Następnie wybieramy mieszaninę i wrzucamy ją na specjalny żel. Żel jest tak zbudowany, że cząsteczki różnej wielkości przemieszczają się w polu elektrycznym z różną prędkością. Przy okazji, żel posiada związek, który po związaniu się z DNA zaczyna świecić w świetle UV.
- Jeśli trafiliśmy dobrze i dany starter jest starterem naszego pasożyta, to wynikiem naszego testu jest obecność świecących w świetle UV prążków w żelu, które to prążki odpowiadają namnożonemu przez nas w reakcji PCR danemu fragmentowi DNA.

- Wiedząc, co było przyczyną infekcji możemy dobrać odpowiednie środki leczące.

7. KRYTERIA OSZACOWANIA PODOBIEŃSTWA SEKWENCJI BIAŁKOWYCH

- Zawartość % pozycji identycznych
- Długość porównywanych sekwencji
- Rozmieszczenie identycznych pozycji wzdłuż porównywanych sekwencji
- Typ reszt w postaci konserwatywnych
- Podobieństwo strukturalne/genetyczne aminokwasów w nieidentycznych pozycjach
- Kryterium identyczności (podobieństwa strukturalnego lub genetycznego)
- Zasadnicze kryteria oszacowania podobieństwa porównywanych sekwencji
- Procent identyczności (względny udział odpowiadających sobie pozycji obsadzonych tymi samymi resztami)
- Długość porównywanych sekwencji (liczba porównywanych pozycji)
- Rozmieszczenie identycznych pozycji wzdłuż porównywanych sekwencji
- Typ reszt okupujących pozycje konserwatywne (sekwencje białkowe)
- Relacje genetyczne/strukturalne między resztami znajdującymi się w odpowiadających sobie nieidentycznych pozycjach (sekwencje białkowe)

Procedura oszacowania stopnia podobieństwa porównywanych sekwencji:

Bardzo często oszacowanie stopnia podobieństwa porównywanych sekwencji sprowadzane jest jedynie do określenia względnego udziału pozycji identycznych. Pozostałe kryteria analizy zazwyczaj nie są w ogóle brane pod uwagę (np. bezwzględna długość sekwencji, dystrybucja identycznych pozycji wzdłuż łańcucha). Podejście takie jest niekompletne i stwarza ryzyko błędnej interpretacji otrzymanych wyników. Przedstawiona niżej metoda oparta jest na prawdopodobieństwie przypadkowego pojawienia się zadeklarowanego stopnia identyczności. Uwzględnia ona podstawowe parametry mające znaczenie dla opisu faktycznego związku między porównywanymi sekwencjami.

Liczbę wszystkich możliwych stopni identyczności dla danych dwóch sekwencji opisuje poniższe równanie:

$$T = x^{2n} = \sum_{a=0}^{n} \binom{n}{a} x^a \left(x(x-1)\right)^{n-a}$$

gdzie:

x – ilość rodzajów jednostek występujących w sekwencjach (20 dla białek; 4 dla kwasów nukleinowych)

n – długość sekwencji (liczba porównywanych par pozycji)

a – ilość pozycji identycznych

Prawdopodobieństwo przypadkowego wystąpienia zadeklarowanego minimalnego stopnia identyczności (dla zadeklarowanej wartości a lub wyższej) określa wzór:

$$P_{an} = \frac{\sum_{k=a}^{n} \binom{n}{k} x^k \left(x(x-1)\right)^{n-k}}{x^{2n}}$$

Prawdopodobieństwo przypadkowego wystąpienia zadeklarowanego stopnia identyczności (odpowiadającemu dokładnie wartości a) określa wzór:

Błąd!

$$P_a = \frac{\binom{n}{a} x^a \left(x(x-1)\right)^{n-a}}{x^{2n}}$$

Zalety i wady metody oszacowania stopnia identyczności opartej na prawdopodobieństwie jego przypadkowego wystąpienia.

Zalety:

- Metoda uwzględnia istotne parametry identyczności, co ma wpływ na wiarygodność otrzymanych wyników.
- Metoda może być standardowym narzędziem analizy teoretycznej dla wszystkich rodzajów białek i kwasów nukleinowych.
- Podstawy teoretyczne algorytmu są proste i logicznie spójne.
- Część obliczeniową metody można w pełni zautomatyzować, jak również można ją modyfikować w dowolnej części, w zależności od potrzeb konkretnej analizy.
- Prezentowane podejście pozwala zarówno na analizę pokrewieństwa w obrębie danej rodziny, jak również umożliwia analizę porównawczą między niespokrewnionymi rodzinami.

Wady:

- Wartości pojawiające się w trakcie obliczeń są niezwykle wysokie, co stwarza konieczność zastosowania odrębnych aplikacji umożliwiających operacje na takich wartościach.
- Na obecnym etapie rozwoju metody analiza porównawcza długich sekwencji zajmuje bardzo dużo czasu.
- Kierunek dalszego rozwoju metody ma na celu bardziej szczegółowe uwzględnienie wszystkich istotnych parametrów identyczności, podobieństwa i pokrewieństwa, osiągnięcie uni-

wersalności aplikacyjnej i maksimum informacji możliwej do uzyskania w wyniku kompleksowej analizy teoretycznej.

- Zasadnicze kryteria oszacowania podobieństwa porównywanych sekwencji
- Procent identyczności (względny udział odpowiadających sobie pozycji obsadzonych tymi samymi resztami)
- Długość porównywanych sekwencji (liczba porównywanych pozycji)
- Rozmieszczenie identycznych pozycji wzdłuż porównywanych sekwencji

PODSUMOWANIE

Nowoczesna medycyna, w szczególności zaś genetyka, nie może obejść się bez narzędzi bioinformatycznych. Ich zastosowanie sprawia, że postęp w genetyce i naukach pokrewnych stał się w ostatnich latach tak szybki. W Matlabie zaimplementowane moduły Bioperl służą do sekwencyjnego przeszukiwania funkcją BLAST. Za jego pomocą można analizować białka oraz można przeprowadzać różne badania nad tymi białkami. MW_BLAST.pl jest programem Perl opartym na module Bioperl RemoteBlast. Program rozpoczyna działanie od utworzenia pliku FASTA dla każdej sekwencji. Przeszukiwanie funkcją BLAST może zająć dużo czasu, zanim pojawi się wynik. Przeszukuje on funkcją BLAST trzy sekwencje. MW_BLAST.pl zapisuje wyniki działania programu na dysku w trzech plikach ABL1., Kit.i PDGFRA. Proces przeszukiwania może trwać ok. 15 minut a nawet dużo więcej. Następnie wyświetlony zostaje raport oraz szukanie wyników >=100. Można wtedy zidentyfikować znalezione białko do dalszego badania, identyfikując nowe cele dla terapii lekowej. Za jego pomocą można uzyskać sekwencyjny dostęp i przeprowadzić wyrównanie metodą Smitha-Waterman na domenie kinazy ludzkich receptorów insuliny (pdb 1IRK) i ludzkiej kinazay limfocytu (pdb 3LCK).

Matlab z dodatkowym toolboxem Bioinformatics oferuje dodatkowe narzędzia do badania nukleotydów i kolejności aminokwasów. Na przykład funkcja pdbdistplot pokazuje odległości między atomami i aminokwasami w strukturze PDB, a funkcja ramachandran generuje wątek kąta torsji PHI i kąta torsji PSI kolejności białek. Funkcja Proteinplot dostarcza graficznego interfejsu użytkownika (GUI) do łatwej sekwencji i różne własności, takie jak hydrophobicity. Istnieje w nim możliwość porównywania sekwencji genetycznych różnych organizmów (w tym ludzi i zwierząt, a co za tym idzie szukaniu zastępczych tkanek – możliwość wszczepienia ludziom, testowanie leków na organizmach zwierząt i sprawdzaniu działania na ich organizm, a co za tym idzie dzięki bazie danych możliwość symulowania tego leku na ludzkim organizmie – bez faktycznego podania go człowiekowi).

PIŚMIENNICTWO

1. Watała C.: *Biostatystyka – wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych,* Alfa- medica 2002 Łódź
2. http://www.mariner.hg.pl/strony/bioinformatyka/
3. http://www.lmb.uni-muenchen.de/groups/bioinformatics/bioinfo.html
4. http://www.geocities.com/bioinformaticsweb/tutorial.html
5. http://www.bioperl.org/
6. http://www.ensembl.org/Docs/enstour/
7. http://www.wiw.pl/Biologia/Genetyka/JezykGenow/Esej.asp?base=r&cp=1&ce=35
8. http://www.ncbi.nlm.nih.gov/entrez/
9. ExPASy http://www.expasy.org/
10. Swiss-Prot database http://www.expasy.org/sprot/
11. NCBI http://www.ncbi.nlm.nih.gov/
12. EBI http://www.ebi.ac.uk
13. GenomeNet http://www.genome.ad.jp/
14. EMBNet http://www.ch.embnet.org/
15. Protein Data Bank (PDB) http://www.rcsb.org/pdb/ index.html
16. DBTSS: Database of Transcriptional Start Sites http://dbtss.hgc.jp/index.html
17. Human Genome Centers and Contacts: Baylor College of Medicine Human Genome Center: http://www.hgsc. bcm. tmc.edu
18. http://genome1.ccc.columbia.edu/%7Egenome
19. http://lpg.nci.nih.gov/CHLC
20. http://eri.uchsc.edu/chromosome21
21. http://www.cephb.fr/
22. http://www.genethon.fr/genethon_en.html
23. http://www.ncbi.nlm.nih.gov/BLAST/
24. SWISS-MODEL http://www.expasy.org/swissmod/ SWISS-MODEL.html
25. ACCELRYS http://www.accelrys.com/
26. Tripos http://www.tripos.com
27. FQS Semihomology tool http://www.fqspl.com.pl/sh/
28. Pedro's BioMolecular Research Tools http:// www. biophys. uniduesseldorf. de/ BioNet/ Pedro /research_ tools. html

Damian Mazur
ul. Niezapominajek 33
35-604 Rzeszów
Tel. 604 459 775
E – mail: dmazu@wp.pl