

Narzędzie do analizy sekwencji – BLAST

(Rozdział 16)

Streszczenie

Porównanie sekwencji nukleotydów lub białek tych samych lub różnych organizmów jest bardzo potężnym narzędziem w biologii molekularnej. Poprzez znalezienie podobieństw między sekwencjami, naukowcy mogą poznawać funkcje nowo zsekwencjonowanych genów, nowe rodziny genów, badać powiązania ewolucyjne. Aktualnie, kiedy cały genom jest zsekwencjonowany, wyszukiwanie podobieństw może być użyte do przewidywania lokalizacji i funkcji kodu białkowego oraz regulacji regionów transkrypcyjnych w genomowym DNA

Basic Local Alignment Search Tool (Blast) (1, 2) jest narzędziem najczęściej wykorzystywanym do obliczania podobieństwa sekwencji. BLAST wchodzi w wariacje w różnych kolejnościach zapytania w stosunku do różnych baz danych. Wszystkie funkcje BLAST, jak również informacje, jak używać Blast oraz innych dokumenty pomocnicze, są wymienione na stronie głównej Blast. W tym rozdziale skupiono się bardziej na tym, jak działa BLAST, jego wydajności, oraz jak to może być wykorzystywane, a nie jak korzystać czy interpretować wyniki w Blast.

Wstęp

Sposób, w jaki większość ludzi używa BLAST jest wpisanie sekwencji nukleotydowych lub białkowych zapytanie wobec wszystkich (lub jej części) publicznych sekwencyjnych baz danych, wklejanie sekwencji w polu tekstowym na jednej ze stron internetowych Blast. Ten wysyła zapytanie przez Internet, wyszukiwanie odbywa się w bazach danych i serwerach NCBI, a wyniki są wysyłane z powrotem do osoby szukającej w wybranym formacie ekranu. Jednak wiele przedsiębiorstw biotechnologicznych, naukowców genetycznych oraz osoby zajmujące się bioinformatyką może użyć "stand-alone" Blast czyli zapytanie własne, lokalne bazy danych lub jeśli chcą dostosować BLAST, aby lepiej odpowiadał ich potrzebom. Stand-alone BLAST występuje w dwóch formach: pliki wykonywalne, które mogą być uruchamiane z linii poleceń lub Standalone WWW BLAST Server, która pozwala użytkownikom tworzyć własne w wersji domowej stron internetowych BLAST.

Istnieje wiele różnych dostępnych odmian BLAST wykorzystywanych do porównywania różnych sekwencji, np. zapytanie o DNA bazy danych DNA, zapytanie o białka białkowej bazy danych i zapytanie o DNA, przetłumaczona na wszystkie sześć ramek odczytu, do bazy danych sekwencji białkowych. Inne modyfikacje BLAST, takich jak PSI-BLAST (do wyszukiwania powtarzających się podobieństw sekwencji białka przy użyciu poszczególnych pozycji macierzy) i RPS-BLAST (do wyszukiwania domen białka w Conserved Domains Database, rozdział 3) wykonywanie porównań w zależności od profilu sekwencji.

Ten rozdział po raz pierwszy opisze architekturę BLAST - jak to działa na stronie NCBI – a następnie przejdzie do opisu różnych wyjść Blast. Najlepiej poznany jest wyświetlanie domyślne ze stron internetowych BLAST, tak zwanych "tradycyjnych raportów". Tak jak otrzymanie tradycyjnych raportów, wyniki mogą być również dostarczone w formie wyjściowej, takich jak hit table (patrz poniżej), XML lub ASN.1. Optymalny wybór formatu wyjściowego zależy od aplikacji. W końcowej części omówiono rozdział stand-alone BLAST i możliwości jego dostosowania. Istnieje wiele interfejsów do BLAST, które często nie są wykorzystywane przez użytkowników, ale mogą prowadzić do wniosków bardziej wydajnych i niezawodnych.

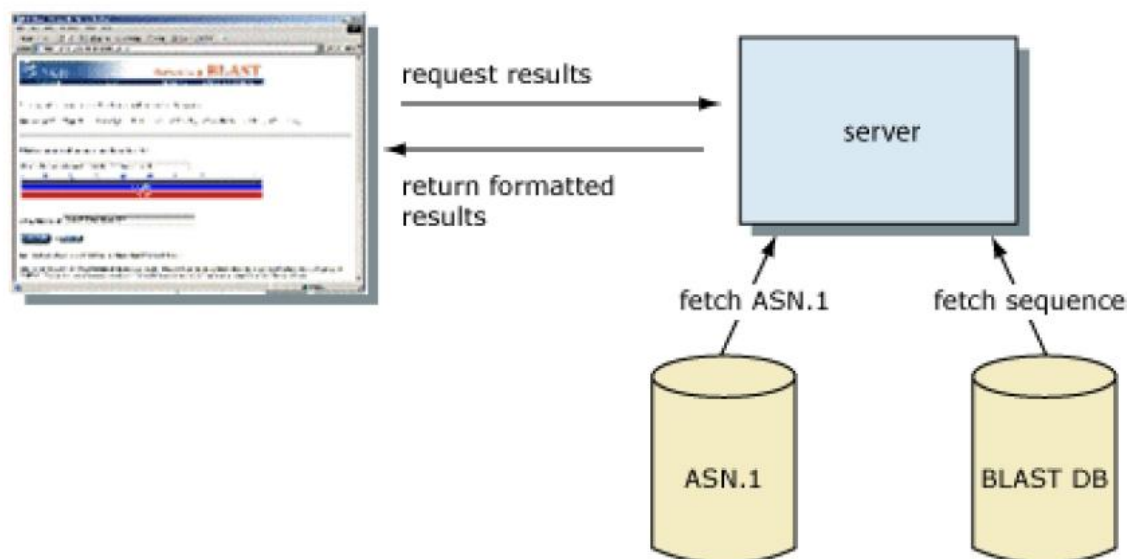
Jak działa BLAST: Podstawy

Algorytm BLAST jest programem heurystycznym, co oznacza, iż opiera się on na pewnych inteligentnych skrótach, których zadaniem jest przeprowadzenie szybszego wyszukiwania. BLAST dokonuje „lokalnego” uporządkowania. Większość złożonych białek występujących w naturze ma złożony charakter, funkcjonalne domeny są często powtarzane w tym samym białku, jak również wśród białek pochodzących od różnych gatunków. Algorytm BLAST pozwala znaleźć te domeny lub krótsze odcinki podobieństwa sekwencji. Lokalne uporządkowanie sugeruje również, iż mRNA może zostać uporządkowane łącznie z częścią genomowego DNA, co często jest wymagane przy składaniu i analizie genomu. Jeśli zamiast użycia BLAST rozpoczniemy od próby uporządkowania dwóch sekwencji wzdłuż ich całych długości (znanych jako globalne uporządkowanie), kilka podobieństw może zostać wykrytych, szczególnie w odniesieniu do domen i wzorów.

Kiedy kwerenda jest poddawana obróbce przy użyciu jednej ze stron internetowych BLAST, sekwencja, łącznie z każdą dodatkową wprowadzoną informacją, taką jak np. nazwa bazy danych, która ma zostać przeszukana, wielkością słowa, przewidywaną wartością, itp. Zostaje dostarczona do algorytmu znajdującego się na serwerze BLAST. BLAST przeszukuje najpierw tabelę zawierającą wszystkie „słowa” (krótkie sekwencje, które dla białek domyślnie posiadają trzy litery) oraz „słowa sąsiadujące”, czyli słowa podobne do występujących w sekwencji kwerendy. Baza danych sekwencji jest następnie skanowana w poszukiwaniu wymaganych sekwencji. Kiedy porównanie jest zidentyfikowane, zostaje ono użyte do zainicjowania wydłużeń pozbawionych lub posiadających luki „słów”.

BLAST nie przeszukuje bazy GenBank złożonej z oddzielnych plików (ani innych podzbiorów bazy GenBank składających się z oddzielnych plików) bezpośrednio. Raczej to sekwencje są stworzone dla bazy danych BLAST. Każda pozycja jest podzielona, tworzone są dwa pliki, w których pierwszy zawiera jedynie główne informacje, a drugi informacje dotyczące sekwencji. Są to dane wykorzystywane przez algorytm. Jeżeli BLAST używany jest w trybie „samotnym”, plik z danymi może składać się z lokalnych, prywatnych danych pobranych z baz NCBI BLAST lub połączenia obu.

Po tym, jak algorytm przeszukał wszystkie możliwe „słowa” zawarte w sekwencji kwerendy i maksymalnie je rozszerzył, konstruuje on najlepsze uporządkowanie dla każdej kwerendy - pary sekwencji i zapisuje te informacje do struktury danych SeqAlign (w ASN.1; używanej także przez Sequin, zobacz: Rozdział 12.). Struktura SeqAlign sama w sobie nie zawiera informacji o sekwencji; raczej odwołuje się zawartych w bazie danych BLAST.



Jak wyniki BLAST stron internetowych są montowane

System QBLAST znajduje się na serwerze BLAST dokonuje wyszukiwania, pisze informacji o dostosowaniu sekwencji ASN.1. Wyniki mogą być sformatowane przy wymianie ASN.1 (pobranie ASN.1) i pobieraniu sekwencji (kolejność pobrań) z bazy danych BLAST. Ponieważ realizacja algorytmu wyszukiwania jest oddzielona od formatowania wyników mogą być one dostarczane w różnych formatach bez konieczności ponownego uruchamiania wyszukiwania. BLAST formatujący, który znajduje się na serwerze BLAST, może korzystać z informacji w SeqAlign znaleźć i pobrać podobne sekwencje oraz wyświetlić je w różny sposób. Gdy jedno zapytanie zostanie zakończone, wyniki mogą być przekształcone bez konieczności ponownego wykonania wyszukiwania. Jest to możliwe ze względu na system QBLAST.

Wyniki BLAST i statystyki

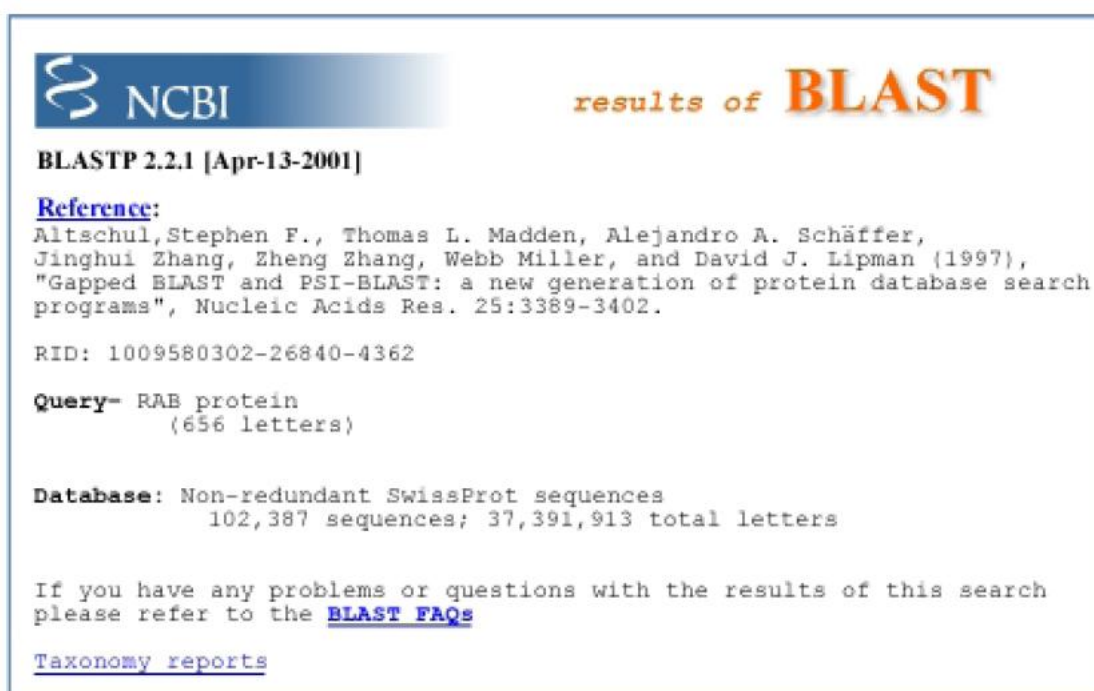
Gdy BLAST znalazł podobne sekwencje na zapytanie w bazie danych, warto mieć jakiś pomysł na to, czy wyrównanie jest "dobre" i czy przedstawia możliwe biologiczne związki, czy też zauważyć podobieństwo wynikające z szansy. BLAST wykorzystuje teorie statystyczne do wynalezienia wyniku i oczekiwanej wartości (E-value) dla każdej pary dostosowanie (zapytanie do trafienia).

Wynik wskazuje, jak dobra jest dostosowanie; im wyższy wynik, tym lepiej wyrównania. Ogólnie rzecz biorąc, wynik ten jest obliczany z wzoru uwzględniającego dostosowanie pozostałości podobne lub identyczne, jak również wszelkie luki wprowadzone w celu dostosowania sekwencji. Kluczowym elementem w tej kalkulacji jest "macierz substytucji", która przypisuje ocenę na dostosowanie ewentualnych parę pozostałości. Macierz BLOSUM62 jest domyślną wartością dla większości programów BLAST, wyjątkami są blastn i MegaBLAST (programy, które wykonują porównanie nukleotyd – nukleotyd, więc nie trzeba używać specyficznego białka macierzy). Ocenę Bit są znormalizowane, co oznacza, że wyniki z różnych linii trasowania można porównywać, nawet jeśli różne matryce punktacji zostały wykorzystane.

E-wartość daje wskaźnik istotności statystycznej danego dostosowania parami i odzwierciedla rozmiar bazy danych i stosowany system punktacji. Niższe E-wartość, większy hit. Dopasowanie sekwencji, które e-wartość 0,05 oznacza, że podobieństwo 5, 100 (1 na 20) możliwość wystąpienia przez przypadek sam.

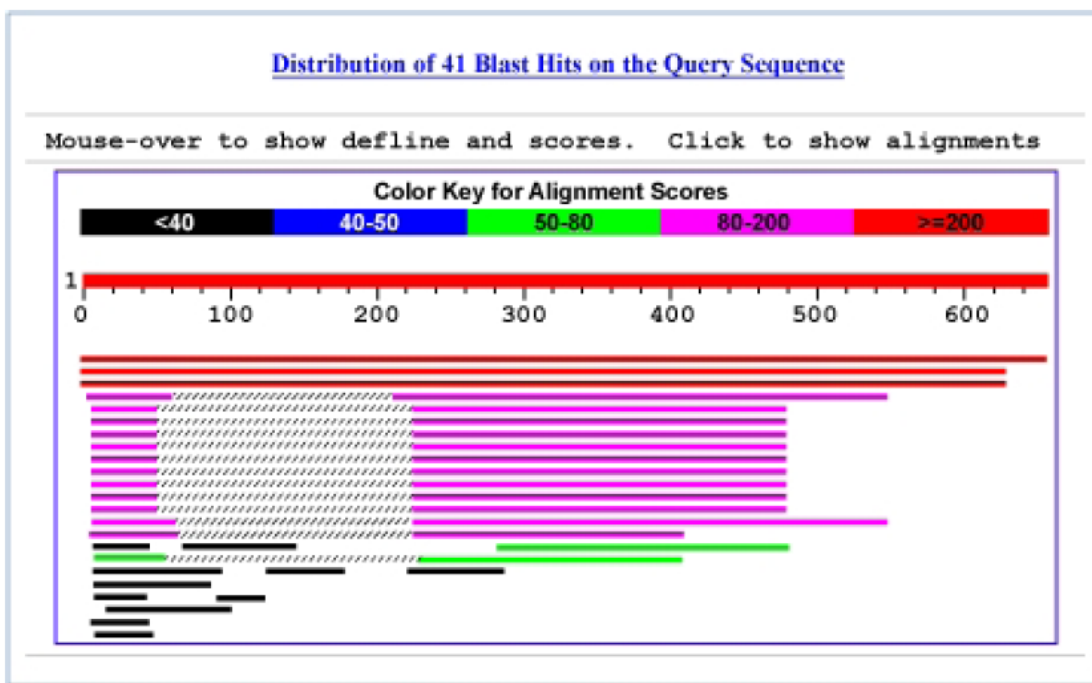
Dane wyjściowe BLAST: 1 Tradycyjny raport

Większość użytkowników Blast jest zaznajomiona z tzw. „tradycyjnym” raportem. Raport składa się z trzech głównych sekcji; (1) nagłówek, który zawiera informacje o sekwencji kwerendy znalezionej w bazie danych (rysunek 1). W sieci jest również przegląd graficzny (rysunek 2); (2) liniowy opis każdej znalezionej w bazie danych sekwencji pasujących do kwerendy, to pozwala szybko przejrzeć dysk (rysunek 4); (3) ustawienie dla każdej pasującej sekwencji (rysunek 5) (może być więcej niż jedno ustawienie dla sekwencji, która pasuje).



Nagłówek raportu BLAST

Górna linia informuje o typie programu (w tym przypadku BLASTP), wersja (2.2.1.). Analizowany artykuł, który Blast opisuje jest wymieniony, na żądanie ID, definicja sekwencji kwerendy oraz podsumowanie sprawdzonych baz danych. Raport taksonomiczny z głównymi informacjami znajduje się w taksonomicznej bazie danych (rozdział 4).



Graficzny przegląd wyników

Sekwencja kwerendy jest reprezentowana przez *numerowaną czerwoną linię* na *górze* rysunku. Trafienie pokazuje ustawienie do kwerendy *poniżej* czerwonej linii. Ustawienie pokazuje sekwencje najbardziej podobne do kwerendy. W tym przypadku są trzy najwyżej punktowane, które najbardziej pasują do kwerendy. Następne dwanaście linii reprezentują niżej punktowane kwerendy pasujące w dwóch regionach, od około 3-60 i 220-500. *Zakreskowane skośnie części linii* wskazują, że dwa podobne regiony to są te same białka ale to pokazuje, że te regiony nie pasują. Reszta linii ukazuje najniżej punktowane ustawienia. Najechnanie na linii myszką pokaże opis linii, który będzie widoczny w oknie nad grafiką.

Sequences producing significant alignments:			Score (bits)	E Value
(a)	(b)	(c)	(d)	
gi 116365 sp P26374 RAE2_HUMAN	Rab proteins geranylgeranyl...	1216	0.0	
gi 21431807 sp P24386 RAE1_HUMAN	Rab proteins geranylgeranyl...	879	0.0	
gi 585775 sp P37727 RAE1_RAT	Rab proteins geranylgeranyltra...	846	0.0	
gi 13626886 sp Q61598 GDIC_MOUSE	RAB GDP dissociation inhib...	127	5e-29	
gi 729566 sp P39958 GDI1_YEAST	SECRETORY PATHWAY GDP DISSOC...	127	5e-29	
gi 13626813 sp O97556 GDIB_CANFA	Rab GDP dissociation inhib...	126	1e-28	
gi 13638229 sp P50397 GDIB_MOUSE	RAB GDP dissociation inhib...	125	3e-28	
gi 1707888 sp P50398 GDIA_RAT	RAB GDP dissociation inhibito...	124	7e-28	
gi 121108 sp P21856 GDIA_BOVIN	Rab GDP dissociation inhibit...	124	7e-28	
gi 21903424 sp P50396 GDIA_MOUSE	Rab GDP dissociation inhib...	124	7e-28	
gi 13626812 sp O97555 GDIA_CANFA	RAB GDP dissociation inhib...	124	8e-28	
gi 1707886 sp P31150 GDIA_HUMAN	Rab GDP dissociation inhibi...	123	9e-28	
gi 13638228 sp P50395 GDIB_HUMAN	Rab GDP dissociation inhib...	122	2e-27	
gi 1707891 sp P50399 GDIB_RAT	RAB GDP DISSOCIATION INHIBITO...	121	5e-27	
gi 1723467 sp Q10305 YD4C_SCHPO	Putative secretory pathway ...	120	8e-27	
gi 585776 sp P32864 RAEP_YEAST	RAB proteins geranylgeranyl...	97	7e-20	
gi 10720243 sp O93831 RAEP_CANAL	RAB proteins geranylgeranyl...	74	9e-13	
gi 2498411 sp Q49398 GLF_MYCGE	UDP-galactopyranose mutase	35	0.63	
gi 11135401 sp Q9XBQ9 STHA_AZOFI	Soluble pyridine nucleotid...	34	1.0	
gi 11135075 sp O05139 STHA_PSEFL	Soluble pyridine nucleotid...	33	1.3	
gi 11135195 sp P57112 STHA_PSEAE	Soluble pyridine nucleotid...	33	1.8	
gi 22257022 sp Q8TJZ8 RLA0_PYRFU	Acidic ribosomal protein P...	33	2.1	
gi 3915516 sp P94488 YNAJ_BACSU	Hypothetical symporter ynaJ	32	3.4	
gi 231788 sp P30599 CHS2_USTMA	CHITIN SYNTHASE 2 (CHITIN-UD...	32	3.7	
gi 2498412 sp P75499 GLF_MYCPN	UDP-galactopyranose mutase	32	4.2	
gi 547891 sp P36225 MAP4_BOVIN	Microtubule-associated prote...	32	4.2	
gi 586602 sp P37747 GLF_ECOLI	UDP-galactopyranose mutase	32	4.6	

Linijowe opisy w raporcie

Każda linia składa się z czterech pól: (a) liczba gi opisuje bazę danych, numer akcesyjny i nazwę miejsca pasującej sekwencji oddzieloną pionowymi liniami (Appendix I); (b) krótki opis sekwencji, definicję. Zawiera również informacje o organizmie, z którego sekwencja pochodzi, typ sekwencji (np. mRNA albo DNA) oraz informacje o funkcjonowaniu bądź fenotypie. Definicja jest często okrojona w jednej linii by widok był czytelny; (c) wynik ustawienia jest ukazany w bitach. Wysoko punktowane wyniki są umieszczone na szczycie listy; (d) wartość E oszacowuje statystyczne znaczenie. Dla pierwszego wyniku z listu liczba gi wynosi 116365, opis bazy danych sp (dla SWISS-PROT), numer akcesji P26374, nazwa miejsca (locus) RAE2_HUMAN, definicja Rab białka, wynik to 1216 oraz wartość E wynosi 0.0. Siedemnaście pierwszych wyników ma małą wartość E i są albo RAB białkami albo GDP inhibitorami dysocjacji. Pozostałe pasujące dane mają znacznie wyższą wartość E 0.5 i więcej co oznacza, że te sekwencje mogły być dopasowane przypadkiem.


```

>gi|116365|sp|P26374|RAE2_HUMAN Rab proteins geranyltransferase component A 2 (Rab escort
protein 2) (REP-2) (Choroideraemia-like protein)
Length = 656

Score = 846 bits (2186), Expect = 0.0
Identities = 432/632 (68%), Positives = 489/632 (77%), Gaps = 13/632 (2%)

Query: 1 MADNLPTEFDVVIIGTGLPESILAAACSRSGQVRVHIDSRSYGGNWASFSFSGLLSWLK 60
      MADNLP++FDV++IGTGLPESI+AAACSRSGQVRVH+DSRSYGGNWASFSFSGLLSWLK
Sbjct: 1 MADNLPSPDFDVIIGTGLPESIIAAACSRSGQVRVHVD+SRSYGGNWASFSFSGLLSWLK 60

Query: 61 EYQQNDIGEESTVWQDLIHETEEAITLRKKDETIQHTFAFFYASQDMEDNVEEIGALQ 120
      EYQ+NND+ E++ +WQ+ I E EEAI L KD+TIQH E F YASQD+ +VEE GALQ
Sbjct: 61 EYQENNDVVITENS-MWQEIQILENEEAIPSSKDKTIQHVVEVFCYASQDLHKDVEEAGALQ 119

Query: 121 KNPSLGVS----NTFTEVLDSALPEESQLSYFNSDEMPAKHTQKSDTEISLEVTDVEESV 176
      KN + S S LP + S E+PA+ +Q E S EV D E +
Sbjct: 120 KNHASVTSQAQSAEAAEAARTSCLP+AVEFLSMGSCIPAEQSQCPGPSSPEVNDARATG 179

Query: 177 EKEKYCGDKTCMHTVXXXXXXXXXXXXTVEDKADEPIRNRITYSQIVKEGRRFNIDLVSQ 236
      +KE + V+D + P +NRITYSQI+KEGRRFNIDLVS+
Sbjct: 180 KKENSDAKSS-----TEEPSENVFKVQDNTETPKKNRITYSQI+KEGRRFNIDLVSQ 231

Query: 237 LLYSQGLLIDLLIKSDVSRYVEFKNVTRILAFREGKEVEQVPCSRADVFNKELTMVEKRM 296
      LLYS+GLLIDLLIKS+VSRY EFKN+TRILAFREG VEQVPCSRADVFNK+LTMVEKRM
Sbjct: 232 LLYSRGLLIDLLIKSNVSRYAEFKNITRILAFREGTVEQVPCSRADVFNKQLTMVEKRM 291

Query: 297 IMKFLTFCLYEYQHPDEYQAFRCQSFSEYLTKEKLTLPNLQHFVLHSIAMTSESSCTTIDG 356
      IMKFLTFC+EYE+HPDEY+A+ +FSEYLT+KLT+PNLQ+FVLHSIAMTSE++ T+DG
Sbjct: 292 IMKFLTFCVEYEEHPDEYRAYEGTTFSEYLTQKLT+PNLQYFVLHSIAMTSETTCTVDG 351

Query: 357 LNATKNFLQCLGRFGNTPFLLFLYQGGEIPQGFRCMCAVFGGIYCLRHVQCFCVVDKESG 416
      L ATK FLQCLGR+GNTPFLLFLYQGGE+PQ FCRMCVAVFGGIYCLRH VQC VVDKES
Sbjct: 352 LKATKKFLQCLGRYGNTPFLLFLYQGGEIPQGFRCMCAVFGGIYCLRHVQCLVVDKESR 411

Query: 417 RCKAIIDHFGQRINAKYFIVEDSYLSEETCSNVQYQISRAVLITDQSILKTDLDQQTISI 476
      +CKA+ID PGQRI +K+PI+EDSYLSE TCS VQY+QISRAVLITD S+LKT DQQ SI
Sbjct: 412 KCKAVIDQPGQRIISKHFIIEDSYLSENTCSRVQYQISRAVLITDGSVLKTDADQQVSI 471

Query: 477 LIVPPAEFGACAVRVTELCSSMTCMKDTYLVHLTCSSSKTAREDLERSVVKLFTPYTET 536
      L VP EPG+ VRV ELCSSMTCMK TYLVHLTC SSKTAREDLR VV+KLFTPYTE
Sbjct: 472 LAVPAEFGSGFGRVRIELCSSMTCMKGTYLVHLTCSSSKTAREDLERVVQKLTFTPYTEI 531

Query: 537 EINEEELTKPRLLWALYFNMRRDSSGISRSSYNGLPNNVYVCSGPDGGLNEHAVKQAEATL 596
      E E++ KPRLLWALYFNMRRDSS ISR YN LPSNVYVCSGPD GLGN++AVKQAEATL
Sbjct: 532 EAEENEQVEKPRLLWALYFNMRRDSSDISRDCYNDLPSNVYVCSGPDGGLNDNAVKQAEATL 591

Query: 597 FQXXXXXXXXXXXXXXXXXGDDKQPEAP 628
      FQ DGD Q E P
Sbjct: 592 FQICPNEDFCFAPPNPEDIVLDGSSQGEVP 623

```

Pairwise alignment sekwencji aminokwasowych w raporcie BLAST.

Alignment jest poprzedzony identyfikatorem sekwencji, nazwą określonej sekwencji (białka) oraz wartością długości sekwencji aminokwasowej. Następnie występuje wynik końcowy (surowy wynik występuje w nawiasach), a po niej wartość E. Następna linijka zawiera informacje o liczbie identycznych reszt aminokwasowych w szeregu porównawczym (Identities), liczbę konserwatywnych substytucji (Positives) oraz jeśli występują - liczbę luk w szeregu. Pod powyższymi danymi występuje szereg, w którym porównane są 2 sekwencje - u góry sekwencja podana w zapytaniu (query), na dole sekwencja z bazy danych BLASTA (subject), którą z nią porównano. Liczby z prawej i lewej strony sekwencji to odnośniki do określonej lokacji aminokwasów w sekwencji peptydu. Myślniki w sekwencji wskazują insercje bądź delecje. Reszty aminokwasowe, które zostały zamaskowane z powodu małej kompleksowości zostały oznaczone literą X (patrz linijka 4 i ostatnia). Linijka pomiędzy dwoma sekwencjami ukazuje podobieństwa pomiędzy tymi dwoma sekwencjami. Jeśli obie sekwencje w określonym miejscu mają identyczne reszty aminokwasowe to są one ukazane odpowiednio w linijce między nimi. Konserwatywne substytucje definiowane przez macierz substytucji oznaczone są symbolem "+". Raport z narzędzia BLAST został tak opracowany, aby być jak najbardziej przyjaznym czytelnikowi w przeciwieństwie do analiz komputerowych, które bywają zawiłe. Przykładem może być pierwsza linijka opisowa dająca szybką przeglądową informację na temat wyników wyszukiwania - jest ona jednak znacznie

skrótowa ze względu na małą ilość miejsca w raporcie. Również dla wygody informacje, które można odczytać z szeregu porównań sekwencji są podane w określonej linii na początku raportu (wartości E, wyniki, opisy), co upraszcza sprawę czytelnikowi, który nie musi szukać tych informacji, przeglądając cały raport. Do raportu można też dodać określone funkcje, np. linki do rekordów z bazy Entrez Gene (rozdział 19), co skutkuje zmianą formatu pliku. Ułatwia to użytkownikowi znaleźć i wykorzystać określoną informację, używając narzędzi, które odczytują format BLAST'a.

Domyślnie, maksimum wyświetlanych sekwencji wynosi 500, jednak opcję tą można dostosować do własnych potrzeb w opcjach zaawansowanych narzędzia BLAST w sekcji "Alignment options". Wiele komponentów z raportu wyników BLAST'a wyświetlanego w Internecie jest połączonych z tym samym typem informacji z różnych miejsc ze strony, m.in. z dodatkowymi informacjami, włączając pomoc narzędzia BLAST, rekordy sekwencji ENTREZ, pasujące do porównywanych sekwencji w BLAST. Rekordy te zapewniają więcej informacji o danej sekwencji, włączając w to wyszukiwanie dotyczących ich abstraktów z bazy PubMed.

Wynik BLAST'a: 2. Tabela wyników.

Chociaż tradycyjny raport BLAST'a jest idealny dla prześledzenia opisu jednego genu czy białka, naukowcy często z określonej przyczyny czy potrzeby dokonują wiele porównań za pomocą narzędzia BLAST i potrzebują jedynie małego podzbioru informacji zawartego w tradycyjnym raporcie BLAST. Ponadto w przypadkach gdy dane wyjściowe BLAST'a zostaną dalej przekształcone, wydaje się nieprawdopodobnym zanalizować tradycyjny raport. Raport tradycyjny jest jedynie formatem wyświetlanym bez specjalnie formalnej struktury czy zasad, poprawki do raportu można wprowadzić, zmieniając komendy HTML w źródle pliku. Format tabeli wyników zapewnia prostą alternatywę (rysunek 6.).


```
# BLASTN 2.2.1 [Aug-1-2001]
# Database: ecoll
# Query:gi|4730899|dbj|AP000130.1|Homo sapiens genomic DNA of 21q22.1, GART and AML, f43D11-11988 region, segment 5/10.
# Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit
gi|4730899|dbj|AP000130.1|AP000130      gi|2367099|gb|AE000133.1|AE000133      100.00  1198  0  0  52913  54110  10943
gi|4730899|dbj|AP000130.1|AP000130      gi|1789919|gb|AE000427.1|AE000427      100.00  1198  0  0  52913  54107  2347
gi|4730899|dbj|AP000130.1|AP000130      gi|1789607|gb|AE000401.1|AE000401      100.00  1198  0  0  52913  54107  6037
gi|4730899|dbj|AP000130.1|AP000130      gi|1788338|gb|AE000294.1|AE000294      100.00  1198  0  0  52913  54107  5700
gi|4730899|dbj|AP000130.1|AP000130      gi|1787588|gb|AE000231.1|AE000231      100.00  1198  0  0  52913  54107  4146
gi|4730899|dbj|AP000130.1|AP000130      gi|1786875|gb|AE000170.1|AE000170      100.00  1198  0  0  52913  54107  2321
gi|4730899|dbj|AP000130.1|AP000130      gi|1786751|gb|AE000160.1|AE000160      100.00  1198  0  0  52913  54107  9133
gi|4730899|dbj|AP000130.1|AP000130      gi|1788508|gb|AE000308.1|AE000308      99.92   1198  1  0  52913  54107  11740
gi|4730899|dbj|AP000130.1|AP000130      gi|2367181|gb|AE000381.1|AE000381      99.83   1198  2  0  52913  54107  2030
gi|4730899|dbj|AP000130.1|AP000130      gi|1788298|gb|AE000291.1|AE000291      99.58   1198  5  0  52910  54107  5290
gi|4730899|dbj|AP000130.1|AP000130      gi|1787633|gb|AE000234.1|AE000234      93.21   1105  72  3  52912  54014  1146
```

BLAST format outputowy tabeli trafień.

Przedstawione są wyniki przeszukania bazach danych E.coli z użyciem ludzkiej sekwencji jako kwerendy. Linijka zaczynająca się znakiem # powinna być rozważana i zignorowana. *The last comment line* – pole w tabeli.

Klasyfikowanie nowych sekwencji człowieka Expressed Sequence Tags (ESTs) kontaminowanych z wektorem E.coli jest dobrym przykładem, kiedy bardziej pożądane jest użycie outputowych tabeli trafień niż tradycyjnych raportów. W tym okienku dokładna wysokość E- wartości progu powinna być stosowana do różnic pomiędzy kontaminacją sekwencji E.coli i ludzkiej sekwencji. Ludzka ESTs znaleziona blisko sekwencji E.coli i ich połączenie może być zbyteczne bez wcześniejszego zbadania. (Graniczne okienka wymagają wcześniejszego zbadania przez naukowców).

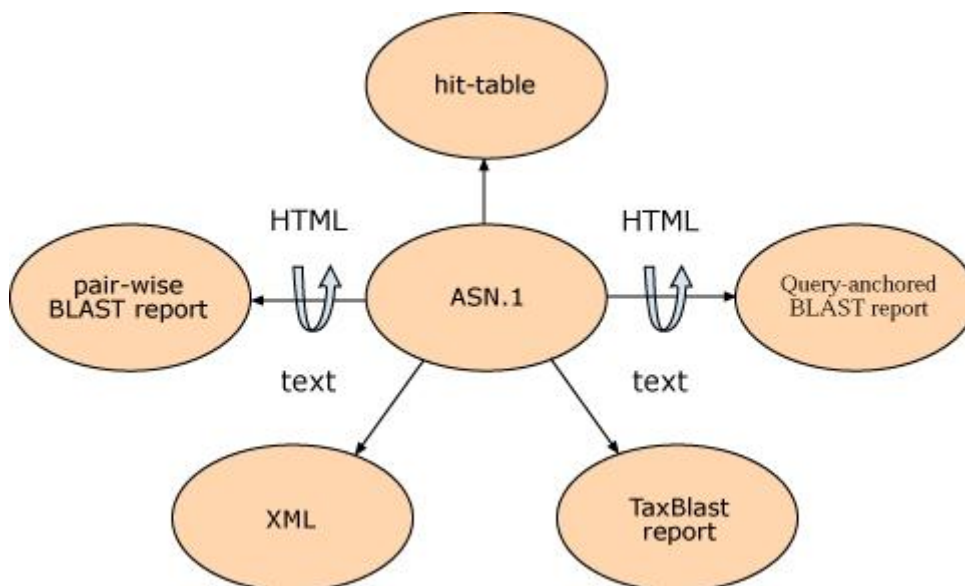
W tym celu outputowa tablica trafień jest bardziej użyteczna niż tradycyjny raport; zawiera tylko odpowiednie informacje w bardziej formalnej strukturze. Outputowa tablica trafień nie zawiera linii sekwencji ani definicji, ale dla innych sekwencji dopasowania, listę zidentyfikowanych sekwencji, start i stop punkty przyłączenia sekwencji podobnej, procent zidentyfikowanych połączeń i E- wartość.

BLAST Output: 3. Struktura outputowa.

Jest kilka przeszkód w procedurze tworzenia raportu BLASTa i prostszej tablicy trafień. Nie ma innej drogi do automatycznego sprawdzenia skróconych lub inaczej uszkodzonych outputów w okienku jak duży numer sekwencji, która będzie klasyfikowana. (Tak się dzieje, gdy dysk jest pełny). Nie ma rygorystycznego sprawdzania zmian składni w outputcie, tak samo jak nowych rys, które mogą być tworzone fałszywie. Struktura outputowa poprzez automatyczne i rygorystyczne nakreśla błędy składni i zmiany. Obydwie XML i ASN.1 są przykładem struktury outputowej, w której środku są budowane kreski do naprawy i skompletowania składni i struktury. (W okienku od XMLa, przykładowo, to jest zapewnienie niezbędne do łączenia identyfikatorów i DTD). W raportach tekstowych często nie ma zapewnienia, ale być może (niekompletny) opis pliku będzie dopisany później.

ASN.1 jest używany jako BLAST Serwer

Tak samo dobrze jak tablica trafień i tradycyjny raport pokazany w HTMLu, tak wyniki BLASTa mogą być zapisane w XML i ASN.1 (Rysunek 7) i co więcej, format, który daje BLAST może być zmieniany bez ponownego wyszukiwania.



Różne formaty, które mogą być tworzone z ASN.1

Zauważmy, że niektóre węzły mogą być wyświetlane zarówno jako HTML i tekst. XML jest również tworzony jest również tak zorganizowanym formatem, który może być wytworzony z ASN.1 ponieważ zawiera równoznaczne informacje.

Zmiana taka jest możliwa, ponieważ kiedy badacz przegląda stronę w sieci Web z wynikiem BLAST w NCBI to właśnie HTML sprawia, że strona została stworzona z ASN.1 (ryc. 7). Chociaż sformatowane wyniki są wymagane na serwerze, to informacje na temat dopasowania są pobrane z dysków w ASN.1, gdzie są odpowiednie sekwencje z bazy danych BLAST (rys. 1). Formatowanie przy użyciu serwera BLAST nie zestawia tych wyników razem jako sprawozdanie BLAST. Szukanie w BLAST zostało odłączone od sposobu w jaki formatowane są wyniki, dzięki czemu możemy uzyskać różne formaty wyjściowe z tego samego wyszukiwania. Ścisłe wewnętrzne zatwierdzenie w ASN.1 zapewnia, że formaty wyjściowe można sporządzić w sposób zawsze wiarygodny.

Informacje o ustawieniach zawarte są w SeqAlign

SeqAlign jest obiektem ASN.1 zawierającym informacje na temat dostosowania wyszukiwania BLAST. SeqAlign nie zawiera rzeczywistych sekwencji, które znajdowały się w dopasowaniu ale zawiera start, stop oraz luki informacyjne, jak również punkty, E-wartości, identyfikatory sekwencji oraz część informacji (DNA).

Jak wyżej wspomniano, rzeczywiste sekwencje bazy danych BLAST są pobierane, gdy zajdzie taka potrzeba. Oznacza to, że identyfikator musi jednoznacznie zidentyfikować kolejność w bazie danych. Ponadto zapytana sekwencja nie może mieć takiego samego identyfikatora, jak którykolwiek inny w bazie danych, chyba, że zapytania te znajdują się w wyżej wymienionej bazie. Jeżeli ktoś samodzielnie korzysta z niestandardowych baz danych BLAST to możliwe jest określenie aby każda sekwencja była identyfikowana przy pomocy opcji **-O** z formatdb (program, który konwertuje pliki FASTA w formacie bazy BLAST). To również spis wpisów identyfikatora. Podobnie opcja **-J** (niestandardowa) w programach blastall, blastpgp, megablast, albo rpsblast zapewniają, że zapytanie nie korzysta już z identyfikatora bazy danych poszczególnych sekwencji. Jeżeli opcja **-O** i **-J** nie są wykorzystywane to BLAST przypisuje unikalne identyfikatory (dla danej serii) do wszystkich sekwencji użytkownika.

Baza danych BLAST lub plik FASTA z sieci NCBI, która zawiera już numery gi spełnia kryterium niepowtarzalności. Unikalne identyfikatory są normalnie problemem tylko wtedy, gdy wytworzenie i ostrożność baz danych klienta nie wymaga przydzielania identyfikatorów. Identyfikator dla zapisu FASTA to pierwszy leksem (chodzi o litery aż do pierwszej spacji) po znaku > w linii definicji. Najprostszym przypadkiem jest zwyczajny unikalny leksem (np. 1, 2 itd.), ale możliwe jest zbudowanie bardziej skomplikowanych identyfikatorów, które mogą, np., opisywać źródło danych. Dla identyfikatora FASTA, aby składnia została poddana wiarygodnej analizie, konieczne jest naśladowanie określonej syntax (patrz Aneks 1).

Więcej informacji na temat SeqAlign stworzoną przez BLAST może być znalezione tutaj lub ściągnięte w formie prezentacji PowerPoint, jak również podręcznika NCBI Toolkit Software Developer's (oprogramowanie zestawu narzędzi producenta).

XML

XML i ASN.1 Są to zbudowane języki i mogą wyrażać te same informacje, dlatego możliwe jest przedstawienie SeqAlign w XML. Niektórzy użytkownicy nie znajdują dogodnie informacji w formacie SeqAlign, ponieważ nie zawiera aktualnych informacji o sekwencjach, i kiedy sekwencja jest ładowana do pamięci z bazy danych BLAST, pakowane są 2 lub 4 bazy na bajt. Typowo, Ci użytkownicy są zaznajomieni ze sprawozdaniem BLAST i chcą czegoś podobnego, ale w formacie, który może składnię poddać wiarygodnej analizie. XML stworzony przez BLAST spełnia te wymagania, zawierając zapytanie i bazę danych sekwencji, linijki definicji sekwencji, pozycje start i stop w ustawieniach (jedno przesunięcie), jak również rezultaty, E-wartości, i procent dopasowania. Jest tam publiczny DTD dla przesyłanych danych XML.

Kod BLAST

Kod BLAST jest częścią zestawu narzędzi (Toolkit) NCBI, która ma dużo nieskomplikowanych funkcji, co czyni platformę samodzielną; Zestaw narzędzi jest wspierany działaniem Linux i dużymi wyborami w UNIX, NT i MacOS. Aby użyć Zestawu narzędzi (Toolkit), należy wpisać funkcję „Main” (główne), która w Zestawie Narzędzi również jest nazwana „main”. Kod BLAST jest zawarty głównie w katalogu narzędzi (zobacz Aneks 2 dla przykładu).

Kod BLAST ma zamysł modułowy. Na przykład, Aplikacja programowania interfejsu (Application Programming Interface – API) dla odzyskania baz danych BLAST jest niezależnym silnikiem obliczeniowym. Silnik obliczeniowy jest niezależny od formatowania; dlatego możliwe jest (jak wymieniono powyżej) liczenie rezultatów raz, ale przeglądanie ich w wielu różnych formach.

Readdb API

Readdb API może być używany do łatwiejszego wyciągania informacji z baz danych BLAST. Wśród dostępnych danych są data wytworzenia bazy danych, tytuł, liczba liter, liczba sekwencji, i najdłuższa sekwencja. Także dostępna jest sekwencja i opis wpisu. Najnowsza wersja baz danych BLAST zawiera również taxid (liczbę całkowitą niektórych wyszczególnionych terminali drzewa taksonomicznego NCBI; patrz rozdział 4). Użytkownicy zdecydowanie zachęcają do stosowania raczej readdb API niż czytania plików związanych z bazą danych, ponieważ są to pliki z zastrzeżeniem zmian. API, z drugiej strony, wspiera najnowszą wersję, i będzie to zamach na produkowanie starszych wersji. Zobacz Aneks 2 z uwagi na przykład prostego programu (db2fasta.c) który demonstruje zastosowanie readdb API.

Wykonanie wyszukiwania narzędziem BLAST z zawołaniami funkcji C

Tylko kilka zapytań funkcyjnych jest potrzebnych do wykonania wyszukiwania BLAST. Aneks 3 przedstawia wyjątek w programie demonstracyjnym doblast.c.

Formatowanie SeqAlign

MySeqAlignPrint (przywołany w przykładzie w Aneksie 3) jest prostą funkcją do drukowania podglądu SeqAlign (patrz Aneks 4).

Aneks 1. Identyfikatory FASTA

Składnia (syntax) linii definicji FASTA używanej w bazach danych NCBI BLAST jest zależna od bazy danych z której otrzymano każdą sekwencję (patrz Rozdział 1 w „GenBank”). Tabela 1 przedstawia jak są identyfikowane bazy danych będące źródłami sekwencji.

Tabela 1

Identyfikatory baz danych w liniach definicji FASTA

Nazwa bazy danych	Identyfikator syntax
GenBank	gb accession locus
EMBL Data Library	emb accession locus
DDBJ, DNA Database of Japan	dbj accession locus
NBRF PIR	pir entry
Protein Research Foundation	prf name
SWISS-PROT	sp accession entry name
Brookhaven Protein Data Bank	pdb entry chain
Patents	pat country number
GenInfo Backbone Id	bbs number
General database identifier ^a	gnl database identifier
NCBI Reference Sequence	ref accession locus
Local Sequence identifier	lcl identifier

^a gnl pozwala bazom danych, nie wymienionym w tym zestawieniu, na używanie tej samej syntax identyfikującej. Jest używane dla sekwencji w śledzących bazach danych, np. gnl|ti|53185177. Kombinacja pól: drugiego i trzeciego powinna być unikatowa.

Na przykład, jeśli identyfikatorem sekwencji w wyniku BLAST jest gb|M73307|AGMA13GT, to znacznik gb wskazuje, że sekwencja pochodzi z bazy GenBank, M73307 jest numerem dostępu (GenBank Accession), a AGMA13GT jest umiejscowieniem w GenBank (tzw. locus).

Znak „|” oddziela różne pola. W pewnych przypadkach pole jest pozostawione puste, mimo że oryginalna specyfikacja wołała o włączenie tego pola. Aby te identyfikatory były kompatybilne wstecz, ze starszymi parserami (analizatorami składniowymi), puste pole jest oznaczane dodatkowym znakiem „|” (| |).

Identyfikator gi jest przypisany do każdej sekwencji w bazach danych NCBI. Jeśli sekwencja pochodzi z bazy danych NCBI, to numer gi pojawia się na początku identyfikatora. Na przykład: gi|16760827|ref|NP_456444.1 wskazuje sekwencję referencyjną z numerem gi 16760827 i numerem dostępu (Accession) NP_456444.1. (W autonomicznym BLAST, lub wtedy, gdy BLAST działa z wiersza poleceń, do wyświetlenia numeru gi należy użyć opcji -I.)

Powodem dodawania identyfikatora gi jest ustanowienie jednolitej, stałej konwencji oznaczania. Jeśli nukleotydowa lub proteinowa sekwencja zmienia się (np. jeśli jest edytowana przez osobę, która ją zamieściła), to jest przydzielany nowy numer gi, ale numer dostępu (Accession) tego rekordu pozostaje niezmienny. W ten sposób identyfikator gi zapewnia mechanizm do identyfikacji właściwej sekwencji, która była użyta lub uzyskana w danych badaniach. Jest również pomocny przy tworzeniu powiązań krzyżowych między różnymi bazami danych Entrez (Rozdział 15).

Aneks 2. Readdb API

Prosty program (db2fasta.c) demonstrujący użycie readdb API.

```
Int2 Main (void)
{
    BioseqPtr bsp;
    Boolean is_prot;

    ReadDBFILEPtr rdfp;
    FILE *fp;
    Int4 index;
    if (! GetArgs ("db2fasta", NUMARG, myargs))
    {
        return (1);
    }
    if (myargs[1].intvalue)
        is_prot = TRUE;
    else
        is_prot = FALSE;
    fp = FileOpen("stdout", "w");
    rdfp = readdb_new(myargs[0].strvalue, is_prot);
    index = readdb_acc2fasta(rdfp, myargs[2].strvalue);
    bsp = readdb_get_bioseq(rdfp, index);
    BioseqRawToFasta(bsp, fp, !is_prot);
    bsp = BioseqFree(bsp);
    rdfp = readdb_destruct(rdfp);
```



```

        return 0;
    }

```

Zauważ:

1. Przydzielniki obiektów `Readdb_new` do czytania dla bazy danych
2. `Readdb_acc2fasta` podaje zwykły numer (zero offsetowy) z rekordu danego z identyfikatora FASTA (np.: GB | AAH06776.1 | AAH0676)
3. `Readdb_get_bioseq` podaje `BioseqPtr` (który zawiera sekwencje, rodzaj i identyfikator) dla tych rekordów
4. `BioseqRawToFasta` składa sekwencje jak FASTA

Aneks 3. Demonstracyjny fragment formularza programu `doblast.c`

```

/* Get default options. */
options = BLASTOptionNew(blast_program, TRUE);
if (options == NULL)
    return 5;

options->expect_value = (Nlm_FloatHi) myargs [3].floatvalue;

/* Perform the actual search. */
seqalign = BioseqBlastEngine(query_bsp, blast_program, blast_database,
options,

    NULL, NULL, NULL);

/* Do something with the SeqAlign... */
MySeqAlignPrint(seqalign, outfp);

/* clean up. */
seqalign = SeqAlignSetFree(seqalign);
options = BLASTOptionDelete(options);

```

```
sep = SeqEntryFree(sep);
```

```
FileClose(infp);
```

```
FileClose(outfp);
```

Główne kroki to:

1. BLASTOptionNew przydzielniki BLASTOptionBlk z brakiem wartości dla specyficznego programu (np.: blastp); argument boolowski określa szukanie.
2. Expect_value (wartość oczekiwana) z BLASTOptionBlk jest zmieniona na określoną niedomyślną wartość (command-line) polecenia
3. BioseqBlastEngine spełnia rolę wyszukiwania w BioseqPtr (query_bsp). BioseqPtr uzyskuje z bazy danych BLAST, Entrez lub z FASTA używając funkcji FastaToSeqEntry.

Struktura BLASTOptionBlk zawiera duże numery elementów zbiorowych. Najbardziej użyteczne opcje dla każdego znajdują się w tabeli 2.

Tabela 2

Najczęściej używane opcje BLAST w BLASToptionBlk

Typ	element	Opis
Nlm_FloatHi	expect_value	Spodziewana odcięta wartość
Int2	wordsize	Liczba oznaczeń użytych do przygotowania słów dla znalezienia tabeli
Int2	penalty	Strata rozbieżności (tylko dla blastn i MegaBLAST)
Int2	reward	Dopasowanie satysfakcjonujących (tylko dla blastn i MegaBLAST)
CharPtr	matrix	Matryca użyta dla porównywania (nie dla blastn lub MegaBLAST)
Int4	gap_open	Cena istnienia luki
Int4	gap_extend	Cena rozbudowania luki na większe ilości oznaczeń
CharPtr	filter_string	Opcje filtrowania (np.: L, mL)
Int4	histlist_size	Liczba sekwencji bazy do zapisania
Int2	number_of_cpus	Liczba CPU do użycia

Aneks 4. Funkcje do wydruku widoku SeqAlign: MySeqAlign Print

```
#define BUFFER_LEN 50

/*
Print a report on hits with start/stop. Zero-offset is used.
*/
static void MySeqAlignPrint(SeqAlignPtr seqalign, FILE *outfp)
{
    Char query_id_buf[BUFFER_LEN+1], target_id_buf[BUFFER_LEN+1];
    SeqIdPtr query_id, target_id;
    while (seqalign)
    {
        query_id = SeqAlignId(seqalign, 0);
        SeqIdWrite(query_id, query_id_buf, PRINTID_FASTA_LONG, BUFFER_LEN);

        target_id = SeqAlignId(seqalign, 1);
        SeqIdWrite(target_id, target_id_buf, PRINTID_FASTA_LONG, BUFFER_LEN);

        fprintf(outfp, "%s:%ld-%ld\t%s:%ld-%ld\n",
            query_id_buf, (long) SeqAlignStart(seqalign, 0), (long) SeqAlignStop(seqalign, 0),
            target_id_buf, (long) SeqAlignStart(seqalign, 1), (long) SeqAlignStop(seqalign, 1));
        seqalign = seqalign->next;
    }
    return;
}
```

Zauważ:

1. SeqAlignId dostaje identyfikator sekwencji dla zerowych identyfikatorów (zero offsetowych). Aktualnie jest to struktura C.
2. Format informacji SeqIdWrite pyta w query_id w identyfikatorze FASTA, oraz ich miejsca w query_buf.

3. Powrót SeqAlignStart oraz SeqAlignStop do startowych wartości zerowych i pierwszych (lub pierwszych i drugich) sekwencji.

Wszystkie te funkcje znajdują się na poziomie funkcji zaawansowanych i nie jest istotne, aby je zapisywać w funkcjach mniej zaawansowanych do przeprowadzenia ASN.1.

Bibliografia

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. J Mol Biol 1990;215:403–410. [PubMed: 2231712]
2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402. [PubMed: 9254694]146917

Wykonali:

Aneta Pułtorak

Aneta Spóz

Anna Ziółkowska

Bartłomiej Kocbach

Dominik Gmurczyk

Grzegorz Skrobisz

Izabela Mazurek

Joanna Kraszevska

Magdalena Żabińska

Natalia Laskowska