А.В. Бутвиловский,Е.В. Барковский,В.Э. Бутвиловский

Сравнительная характеристика вариантов ncbi blast-анализа ряда митохондриальных ферментов различных животных

Белорусский государственный медицинский университет

Статья посвящена теоретическим аспектам методов поиска аминокислотных и нуклеотидных последовательностей в базах данных сети Интернет. Проведена сравнительная характеристика двух вариантов NCBI BLAST-анализа.

Ключевые слова: BLAST-анализ, поиск аминокислотных и нуклеотидных последовательностей.

Сравнение аминокислотных и нуклеотидных последовательностей является важным звеном исследований молекулярной биологии, которое позволяет идентифицировать семейства генов, относить к ним секвенированные последовательности, устанавливать их структурные и функциональные взаимоотношения. В настоящее время, когда секвенируются целые геномы, значение подобных исследований постоянно возрастает.

К программам, используемым для сравнения последовательностей с последующим определением их сходства, относятся: ALIGN, AMAS, BLAST, BLAT, CLUSTAL, DiAlign, FASTA, HI, HMMER, MAP, MGA, OWEN, PipMaker, MultiPipMaker, T-Coffee и др.

Наиболее часто используются программы CLUSTAL и BLAST (Basic Local Alignment Search Tool, основное средство поиска, основанное на локальных выравниваниях). Первое поколение BLAST программ появилось в начале 90-х годов прошлого столетия [1]. Второе поколение программ данной серии представлено двумя вариантами: WU-BLAST 2 (Washington University BLAST 2 [5]) и NCBI BLAST 2 (National Center for Biotechnology information BLAST 2 [3]).

Эти программы являются самыми быстрыми (скорость поиска на порядок выше программы FASTP и других алгоритмов) и чувствительными (определяют даже незначительное сходство последовательностей). Программы серии BLAST продолжают модифицироваться, при этом скорость поиска у последних версий BLAST приблизительно в 3 раза выше скорости оригинала.

Классификация программ серии BLAST.

Семейство программ серии BLAST делится на 5 основных групп:

1. Нуклеотидные – предназначены для сравнения изучаемой нуклеотидной последовательности с базой данных секвенированных нуклеиновых кислот и их участков:

• megablast – быстрое сравнение с целью поиска высоко сходных последовательностей,

• dmegablast – быстрое сравнение с целью поиска дивергировавших последовательностей, обладающих незначительным сходством,

• blastn – медленное сравнение с целью поиска всех сходных последовательностей и др..

2. Белковые – предназначены для сравнения изучаемой аминокислотной последовательности белка с имеющейся базой данных белков и их участков.

• blastp – медленное сравнение с целью поиска всех сходных последовательностей,

• cdart – сравнение с целью поиска гомологичных белков по доменной архитектуре,

• rpsblast – сравнение с базой данных консервативных доменов,

• psi-blast – сравнение с целью поиска последовательностей, обладающих незначительным сходством,

• phi-blast – поиск белков, содержащих определенный пользователем паттерн и др.

3. Транслирующие – способны транслировать нуклеотидные последовательности в аминокислотные:

• blastx – переводит изучаемую нуклеотидную последовательность в кодируемые аминокислоты, а затем сравнивает ее с имеющейся базой данных аминокислотных последовательностей белков,

• tblastn – изучаемая аминокислотная последовательность сравнивается с транслированными последовательностями базы данных секвенированных нуклеиновых кислот,

• tblastx – переводит изучаемую нуклеотидную последовательность в аминокислотную, а затем сравнивает ее с транслированными последовательностями базы данных секвенированных нуклеиновых кислот.

4. Геномные – предназначены для сравнения изучаемой нуклеотидной последовательности с базой данных секвенированного генома какого-либо организма (человека, мыши и др.)

5. Специальные – прикладные программы, использующие BLAST:

• bl2seq – сопоставление двух последовательностей по принципу локальных выравниваний,

• VecScreen – определение сегментов нуклеотидной последовательности нуклеиновой кислоты, которые могут иметь векторное происхождение и др.

Принципы работы BLAST.

Все выравнивания принято делить на глобальные (последовательности сравниваются полностью) и локальные (сравниваются только определенные участки последовательностей). Программы серии BLAST производят локальные выравнивания, что связано с наличием в различных белках сходных доменов и паттернов. Кроме этого локальное выравнивание позволяет сравнить иРНК с геномной ДНК. В случае глобального выравнивания обнаруживается меньшее сходство последовательностей, особенно их доменов и паттернов.

После введения изучаемой нуклеотидной или аминокислотной последовательности (запрос) на одну из web-страниц BLAST, она вместе с другой входной информацией (база данных, размера «слова» (участка), значение величины E и др.) поступает на сервер. BLAST создает таблицу всех «слов» (в белке – это участок последовательностей, который по умолчанию состоит из трех аминокислот, а для нуклеиновых кислот из 11 нуклеотидов) и сходных «слов».

Затем в базе данных проводится их поиск. Когда обнаруживается соответствие, то делается попытка продлить размеры «слова» (до 4 и более аминокислот и 12 и более нуклеотидов) сначала без гэпов (пробелов), а затем с их использованием. После максимального продления размеров всех возможных «слов» изучаемой последовательности, определяются выравнивания с максимальным количеством

совпадений для каждой пары запрос – последовательность базы данных, и полученная информация фиксируется в структуре SeqAlign. Форматер, расположенный на сервере BLAST, использует информацию из SeqAlign и представляет ее различными способами (традиционным, графическим, в виде таблицы).

Для каждой обнаруженной в базе данных программами BLAST последовательности необходимо определить, насколько она сходна с изучаемой последовательностью (запрос) и значимо ли это сходство. Для этого BLAST вычисляет число битов и величину E (expected value, E-value) для каждой пары последовательностей [7].

При определении сходства ключевым элементом является матрица замен, так как она определяет показатели сходства для любой возможной пары нуклеотидов или аминокислот. В большинстве программ серии BLAST используется матрица BLOSUM62 (Blocks Substitution matrix 62% identity, блоковая матрица замен с 62% идентичности) [8]. Исключением являются blastn и megablast (программы, которые выполняют нуклеотид – нуклеотидные сравнения и не используют матрицы аминокислотных замен).

С помощью модифицированных алгоритмов Смита-Уотермана [9] или Селлерса [10] определяются все пары сегментов (продленные «слова»), которые нельзя увеличить, так как это приведет к уменьшению показателей сходства. Такие пары продленных «слов» называются парами сегментов с максимальным сходством (high-scoring segment pairs, HSP). В случае достаточно большой длины изучаемой последовательностей (m) и последовательности базы данных (n) показатели сходства HSP характеризуются двумя параметрами K (размера области поиска) и ? (системы подсчета). Эти показатели необходимо указывать при приведении показателей сходства изучаемой последовательности и последовательности базы данных (S).

Для сравнения показателей сходства различных выравниваний независимо от используемой матрицы, их необходимо преобразовать. Для получения преобразованного показателя сходства (числа битов, Sr) используют формулу:

$$Sr = (?S - \ln K)/\ln 2 \quad (1).$$

Величина Sr показывает, насколько сходны последовательности (чем больше число битов, тем больше сходство). Так как в формулу расчета Sr заложены показатели K и ?, то нет необходимости указывать их при приведении значений Sr. Величина E (E-value), соответствующая показателю Sr, показывает достоверность данного выравнивания (чем ниже значение E, тем достовернее выравнивание). Она определяется по формуле:

$$E = mn\,2 - Sr \quad (2).$$

Программы BLAST преимущественно определяют значение E, а не P (вероятности наличия хотя бы одного HPS с показателем, превышающим или равным S). Но при $E < 0{,}01$ значения P и E почти идентичны [7].

Величина E определяется по формуле (2) при сравнении лишь двух аминокислотных или нуклеотидных последовательностей. Сравнение изучаемой последовательности длиной m с множеством последовательностей базы данных может основываться на двух положениях. Первое положение состоит в том, что все последовательности базы данных одинаково сходны с изучаемой. Это подразумевает, что значение E для выравнивания с короткой последовательностью, содержащейся в базе данных, следует приравнять со значением E для выравнивания с длинной последовательностью. Для вычисления значения E по базе данных необходимо

умножить значение E, полученное при попарном сравнении, на число последовательностей в ней. Второе положение заключается в том, что изучаемая последовательность более сходна с короткими, а не с длинными последовательностями, потому что последние часто состоят из различных участков (многие белки состоят из доменов). Если предположить, что вероятность сходства пропорциональна длине последовательности, то попарное значение E для последовательности базы данных длиной n надо умножить на N/n, где N – общая длина аминокислот или нуклеотидов в базе данных. Программы BLAST преимущественно используют этот подход для вычисления значений E по базе данных.

Теоретически локальное выравнивание может начинаться с любой пары нуклеотидов или аминокислот выровненных последовательностей. Однако HPS, как правило, не начинаются близко к краю (началу или концу) последовательностей. Для коррекции такого краевого эффекта необходимо вычислять эффективную длину последовательностей [2]. В случае последовательностей длиной более 200 остатков происходит нейтрализация краевого эффекта.

Рассмотренные выше показатели разрабатывались для не содержащих гэпов местных выравниваний. Однако в ходе последующих исследований [2, 3, 4] было установлено, что эти показатели могут использоваться и для выравниваний, содержащих гэпы.

Цель исследования: сравнить показатели вариантов NCBI BLAST-анализа ряда митохондриальных ферментов дыхательной цепи различных животных.

Материал и методы

В качестве запросов на серверах, предоставляющих возможность проведения вариантов NCBI BLAST-анализа (NCBI, http://www.ncbi.nlm.nih.gov и NPS, http://npsa-pbil.ibcp.fr) использованы аминокислотные последовательности ряда ферментов дыхательной цепи человека [6] – НАДН-дегидрогеназ (субъединицы 1, 2, 3, 4, 4L, 5, 6), цитохрома b, АТФазы 6.

BLAST-анализ проведен при стандартных условиях на матрице BLOSUM 62. Полученные показатели сходства Sr и достоверности выравниваний E выборочно проанализированы для соответствующих ферментов приматов (человека (Homo sapiens, H.s.), шимпанзе (Pan troglodytes, P.t.), бабуина (Papio hamadryas, P.h.)), парнокопытных (быка (Bos indicus, B.i.), козла (Capra hircus, C.h.), свиньи (Sus scrofa, S.s.)), непарнокопытных (лошади (Equus caballus, E.c.)), хищных (кошки (Felis catus, F.c.), медведя (Ursus arctos, U.a.), собаки (Canis familiaris, C.f.)), грызунов (крысы (Rattus norvegicus, R.n.), мыши (Mus musculus, M.m.), белки (Sciurus vulgaris, S.v.)), зайцеобразных (кролика (Oryctolagus cuniculus, O.c.)), птиц (петуха (Gallus gallus, G.g.)), рептилий (аллигатора (Alligator mississippiensis, A.m.)), земноводных (лягушки (Xenopus laevis, X.l.)), рыб (данио (Danio rerio, D.r.)), ланцетника (Branchiostoma floridae, B.f.), круглых червей (аскариды (Ascaris suum, A.s.), трихинеллы (Trichinella spiralis, T.s.), нематоды (Caenorhabditis elegans, C.e.)). В случае наличия нескольких выравниваний белков одного организма учтены выравнивания с максимальным числом битов.

Результаты и обсуждение

Полученые показатели вариантов NCBI BLAST-анализа представлены в табл. 1 и 2.

Таблица 1

Значения показателей сходства Sr и их достоверность, полученные на сервере NPS, для ряда ферментов дыхательной цепи различных животных

Фермент/ организм	НАДН-ДГ 1	НАДН-ДГ 2	НАДН-ДГ 3	НАДН-ДГ 4	НАДН-ДГ 4L	НАДН-ДГ 5	НАДН-ДГ 6	Цитохром *b*	АТФаза 6
H.s.	499 (e-141)	419 (e-117)	126 (2e-29)	726 (0)	191 (6e-49)	996 (0)	250 (1e-66)	619 (e-177)	345 (5e-95)
P. t.	479 (e-135)	402 (e-112)	120 (1e-27)	695 (0)	189 (2e-48)	935 (0)	240 (1e-63)	592 (e-169)	332 (5e-91)
P. h.	431 (c-120)	313 (5c-85)	107 (8c-24)		151 (5c-37)	802 (0)	209 (2c-54)		289 (5c-79)
B. i.									
C. h.		278 (1e-74)						517 (e-146)	290 (3e-78)
S. s.	414 (e-115)	282 (1e-75)	100 (1e-21)	574 (e-163)	156 (1e-38)	751 (0)	156 (2e-38)	518 (e-146)	288 (7e-78)
E. c.	414 (e-115)	283 (e-76)	99,4 (2e-21)	572 (e-163)	150 (6e-37)	747 (0)	161 (8e-40)	515 (e-146)	288 (7e-78)
F. c.	410 (e-114)	264 (3e-70)	97,8 (7e-21)	558 (e-158)	143 (1e-34)	731 (0)	154 (2e-37)	516 (e-146)	288 (7e-78)
U. a.									
C. f.	409 (e-114)	280 (5e-75)	93,6 (1e-19)	577 (e-164)	147 (9e-36)	723 (0)	155 (5e-38)	533 (e-151)	293 (4e-79)
R. n.	398 (e-110)	238 (3e-62)	94,7 (6e-20)	533 (e-151)	137 (1e-32)	665 (0)	126 (4e-29)	508 (e-143)	280 (2e-75)
M. m.	405 (e-113)	239 (7e-63)	90,5 (1e-18)	528 (e-149)	131 (4e-31)	671 (0)	130 (2e-30)	508 (e-143)	285 (1e-76)
S. v.									
O. c.	412 (e-115)	287 (e-77)	97,1 (1e-20)	572 (e-163)	152 (3e-37)	741 (0)	174 (9e-44)	526 (e-149)	286 (3e-77)
G. g.	370 (e-102)	196 (7e-50)	69,3 (3e-12)	466 (e-131)	99,8 (2e-21)	568 (e-161)	77,4 (2e-14)		194 (2e-49)
A. m.	335 (1e-91)		59,7 (2e-09)						184 (2e-46)
X. l.	348 (2e-95)	211 (3e-54)	79,3 (2e-15)	454 (e-127)	67,8 (7e-12)	580 (e-165)	64,3 (2e-10)		176 (4e-44)
D. r.									
B. f.	314 (3e-85)	101 (4e-21)	51,6 (5e-07)	344 (4e-94)	68,9 (3e-12)	479 (e-134)	34,3 (0,18)		162 (7e-40)
A. s.	146 (9e-35)			171 (6e-42)		188 (5e-47)			39,3 (0,009)
T. s.									
C. e.	148 (2e-35)			175 (2e-43)		187 (1e-46)			41,6 (0,002)

Примечание. Значения E указаны в скобках.

Таблица 2

Значения показателей сходства Sr и их достоверность, полученные на сервере NCBI, для ряда ферментов дыхательной цепи различных животных

Фермент/ организм	НАДН-ДГ 1	НАДН-ДГ 2	НАДН-ДГ 3	НАДН-ДГ 4	НАДН-ДГ 4L	НАДН-ДГ 5	НАДН-ДГ 6	Цитохром b	АТФаза 6
H.s.	499 (5e-140)	419 (6e-116)	128 (7e-29)	726 (0)	191 (9e-48)	996 (0)	251 (1e-65)	619 (5e-176)	349 (5e-95)
P. t.	479 (8e-134)	406 (7e-112)	120 (2e-26)	695 (0)	189 (3e-47)	935 (0)	240 (1e-62)	592 (1e-167)	332 (6e-90)
P. h.	431 (2e-119)	313 (6e-84)	107 (1e-22)	578 (2e-163)	151 (8e-36)	802 (0)	209 (3e-53)	519 (7e-148)	289 (6e-77)
B. i.		281 (2e-74)	98 (6e-20)	570 (3e-161)	152 (3e-36)	756 (0)	163 (2e-39)		294 (1e-78)
C. h.		278 (2e-73)		570 (4e-161)	154 (1e-36)	740 (0)	164 (2e-39)	518 (2e-145)	290 (4e-77)
S. s.	409 (1e-112)	282 (1e-74)	100 (2e-20)	577 (3e-163)	159 (4e-38)	756 (0)	156 (3e-37)	521 (2e-148)	291 (2e-77)
E. c.	414 (2e-114)	284 (4e-75)	99 (3e-20)	572 (8e-162)	150 (1e-35)	747 (0)	161 (8e-39)	517 (2e-145)	288 (8e-77)
F. c.		264 (3e-69)	97 (1e-19)	558 (2e-157)	143 (2e-33)	731 (0)	154 (2e-36)	521 (2e-146)	288 (8e-77)
U. a.		262 (1e-68)	96 (2e-19)	561 (3e-158)	152 (4e-36)	728 (0)	160 (2e-38)	522 (8e-147)	280 (4e-74)
C. f.	411 (2e-113)	282 (1e-74)	93 (2e-18)	577 (3e-163)	147 (1e-34)	733 (0)	155 (6e-37)	533 (3e-150)	293 (4e-78)
R. n.	411 (2e-113)	238 (2e-61)	94 (9e-19)	533 (7e-150)	139 (4e-32)	679 (0)	126 (4e-28)		284 (1e-75)
M. m.	405 (8e-112)	239 (9e-62)	90 (2e-17)	528 (1e-148)	136 (2e-31)	674 (0)	130 (3e-29)		285 (9e-76)
S. v.	416 (6e-115)	288 (2e-76)	94 (1e-18)	547 (4e-154)	147 (1e-34)	700 (0)	163 (3e-39)	521 (1e-146)	293 (2e-78)
O. c.		287 (4e-76)	97 (2e-19)	572 (8e-162)	152 (4e-36)	741 (0)	174 (9e-43)	526 (7e-148)	286 (4e-76)
G. g.			69 (3e-12)	467 (e-131)	99 (3e-20)	572 (e-161)	77 (2e-13)		194 (2e-48)
A. m.	335 (2e-90)			409 (1e-112)	92 (3e-18)	528 (2e-148)	46 (3e-04)		
X. l.	348 (2e-94)		79 (4e-14)	454 (4e-126)	67 (1e-10)	580 (4e-164)	64 (2e-09)		
D. r.			75 (5e-13)	491 (3e-137)	102 (3e-21)	584 (1e-166)	82 (8e-15)		
B. f.					68 (5e-11)	486 (1e-135)	34 (1.9)		
A. s.									
T. s.									
C. e.					54 (1e-06)				

Примечание. Значения E указаны в скобках.

Отсутствие данных по результатом BLAST-анализа, проведенного на сервере NPS, для некоторых ферментов и для некоторых животных (пустые ячейки), вероятно, обусловлено отсутствием соответствующих последовательностей в используемой базе данных белков. Это предположение подтверждает тот факт, что ее размер сравнительно мал (206586 белков или их участков общей длиной 75181616 аминокислотных остатков).

Отсутствие данных (пустые ячейки) для некоторых ферментов и для некоторых животных наблюдается и при использовании второго варианта BLAST-анализа (NCBI), использующего значительно бульшую базу данных (3292813 белков или их участков общей длиной 1128164434 аминокислотных остатков). Практически полное отсутствие найденных последовательностей ортологичных белков филогенетически удаленных от человека организмов объясняется тем, что NCBI BLAST выдает только показатели сходства для 1000 выравниваний и в большинстве случаев не доходит до ланцетника, аскариды и др. организмов.

Полученные данные свидетельствуют о том, что показатели сходства последовательностей, в основном, ниже у филогенетически отдаленных организмов. Оба варианта BLAST-анализа дают близкие значения показателей Sr, отличаясь лишь значениями E. Так значения E, полученных на сервере NPS, в среднем на порядок ниже таковых, полученных на сервере NCBI, что обусловлено различиями в размере используемых баз данных. Несмотря на это, результаты, полученные на большей базе данных, обладают большей биологической значимостью. Установлено, что для достижения поставленной нами цели более удобно использовать NCBI-BLAST, поскольку он предоставляет возможность предоставления выходных данных в виде

TAXBLAST (taxonomy BLAST), отображающего результаты в соответствии с таксономическими категориями.

Таким образом, использование вариантов BLAST-анализа основывается на знании его теоретических основ и цели применения.

Выводы

1. Преобразованные показатели сходства аминокислотных последовательностей ферментов дыхательной цепи человека и различных животных, полученные двумя вариантами NCBI BLAST-анализа, в большинстве случаев совпадают.

2. Выравнивания, проведенные на сервере NPS, более достоверны, чем таковые проведенные на сервере NCBI, что связано с размерами используемых баз данных.

Литература

1. Altschul S.F., Gish W., Milleer W., Myers E.W., Lipman D.J. //J. Mol. Biol. – 1990. – Vol. 215. – P. 403-410.

2. Altschul S.F., Gish W. //Meth. Enzymol. – 1996. – Vol. 266. – P. 460-480.

3. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. //Nucleic Acids Res.. – 1997. – Vol. 25(17). – P. 3389-3402.

4. Arratia R., Wateman M.S. //Ann. Appl. Prob. – 1994. – Vol. 4. – P. 200-225.

5. Gish W. (1996-2002) http: //blast.wustl.edu.

6. Ingman M., Kaessmann H., Paabo S., Gyllensten U. //Nature. – 2000. – Vol. 408. – P. 708-713.

7. Karlin S., Altschul S. F. //PNAS. – 1993. – Vol. 90. – P. 5873-5877.

8. Pearson W.R. //Prot. Sci. – 1995. – Vol. 4. – P. 1145-1160.

9. Smith T.F., Waterman M.S. //J. Mol. Biol. – 1981. – Vol. 147. – P. 195-197.

10. Sellers P.H. //Bull. Math. Biol. – 1984. – Vol. 46. – P. 501-514