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

Выравнивание аминокислотных и нуклеотидных последовательностей

Белорусский государственный медицинский университет

Статья посвящена теоретическим аспектам выравнивания аминокислотных и нуклеотидных последовательностей. Рассмотрены основные принципы работы программ Clustal.

Ключевые слова: выравнивание, серия программ Clustal, матрицы выравнивания.

Выравнивание аминокислотных или нуклеотидных последовательностей – это процесс сопоставления сравниваемых последовательностей для такого их взаиморасположения, при котором наблюдается максимальное количество совпадений аминокислотных остатков или нуклеотидов. Различают 2 вида выравнивания: парное (выравнивание двух последовательностей ДНК, РНК или белков) и множественное (выравнивание трех и более последовательностей). Наиболее популярной серией программ для множественного выравнивания последовательностей является Clustal. Первая программа серии Clustal была создана Д.Хиггинсом в 1988 году [8]. Затем она была усовершенствована Д. Фенгом, Р. Дулиттл и В. Тейлором путем добавления прогрессивного выравнивания, то есть созданием множественного выравнивания в результате серий попарных выравниваний, следуя ветвлению направляющего дерева, построенного методом UPGMA [3, 10].

В 1992 году появилась второе поколение программ Clustal. Программа, названная Clustal V, отличалась способностью проводить сопоставления существующих выравниваний и построением направляющего дерева методом NJ [6, 7, 9].

Третье поколение программ, появившееся в 1994 году и названное Clustal W, стало значительно проще в работе благодаря усовершенствованному алгоритму [12]. Кроме этого появилась возможность выбирать матрицы сравнения аминокислот и нуклеотидов, а также устанавливать штрафы за внесение пробелов. Следует отметить, что высокая совместимость программ этого поколения с другими пакетами программ обусловлена за счет предоставления результатов выравнивания в виде формата FASTA. Последним представителем серии является программа Clustal X, для которой характерен более удобный интерфейс и более легкая оценка результатов выравниваний [11]. В настоящее время именно последние программы серии Clustal этого поколения (версия 1.83) позволяют создавать наиболее биологически корректные множественные выравнивания дивергировавших последовательностей [1].

Программы третьего поколения серии Clustal доступны на многих серверах (http://npsa-pbil.ibcp.fr, http://www.ebi.ac.uk) в двух вариантах – интерактивном и почтовом. Интерактивный вариант предполагает ожидание пользователем получения результатов выравнивания (целесообразно применять при небольшом (<100) количестве последовательностей), а почтовый – по электронной почте (применяется при большом числе последовательностей).

Принципы работы CLUSTAL

Первоначально необходимо ввести на одном из серверов изучаемые аминокислотные или нуклеотидные последовательности в одном из 7 возможных форматов (NBRF/PIR, EMBL/SWISSPROT, Pearson (Fasta), Clustal (*.aln), GCG/MSF (Pileup), GCG9/RSF, GDE).
Наиболее часто используется формат FASTA, сущность которого заключатся во введении знака > перед названием каждой последовательности, а затем (с новой строки) однобуквенном обозначении аминокислот и нуклеотидов. Суммарная длина вводимых последовательностей не должна превышать 40000 для WWW и 60000 для e-mail серверов.
При использовании данной программы выравнивание состоит из трех этапов: парных выравниваний, построения направляющего дерева и множественного выравнивания.
1. В ходе парных выравниваний предварительно сравниваются все возможные пары изучаемых последовательностей. На основании проведенных сравнений вычисляются показатели сходства в соответствии с выбранными матрицами. Существуют 2 разновидности парного выравнивания: медленное (slow) и быстрое (fast). Медленное выравнивание является более точным, но его не рекомендуется применять в случае большого количества (более 20) последовательностей значительной длины (более 1000 остатков). Медленное выравнивание характеризуется 4 параметрами:
- штрафом на внесение делеции (gap open penalty). Уменьшение этого параметра способствует внесению разрывов в выравнивание, что ухудшает качество. Увеличение – приводит к тому, что выравнивание будет представлять собой длинные участки последовательностей почти без вставок или делеций.
· штраф на продолжение делеции (gap extension penalty). Этот параметр контролирует возможность внесения длинных вставок или делеций.
- матрица сравнений нуклеотидов (DNA weight matrix, Clustal W 1.6). В наиболее широко используемой матрице DNA identity (рис. 1) совпадение нуклеотидов оценивается в 1 балл, а несовпадение – -10000 баллов. Такой высокий штраф за несоответствие облегчает внесение пробелов.

	A	T	G	C
A	1			
T	-10000	1		
G	-10000	-10000	1	
C	-10000	-10000	-10000	1

Рис. 1. Матрица DNA identity.
- матрица сравнения аминокислот (protein weight matrix) – PAM, Blosum и Gonnet.
Выбор матрицы оказывает большое влияние на получаемые результаты, так как каждая матрица представляет отражение отдельных эволюционных гипотез. Известно, что все замены аминокислот не являются равновероятными и в ходе эволюции чаще происходят замены на сходные по физико-химическим свойствам аминокислоты. Так в ходе эволюции гидрофобный изолейцин достаточно часто заменяется на гидрофобный валин и редко на гидрофильный цистеин. Исследования эволюционных изменений различных белковых семейств позволили установить частоты фиксированных мутаций аминокислот и

нуклеотидов и обобщить полученную информацию в виде матриц. В настоящее время используются серии белковых матриц Blosum, PAM и Gonnet [2, 4, 5]. Матрицы серии Blosum (рис. 2) преимущественно используются при проведении локальных выравниваний (поиск сходных последовательностей по базам данных).

	G	P	D	E	N	H	Q	K	R	S	T	A	M	V	I	L	F	Y	W	C
G	7																			
P	-2	9																		
D	-1	-1	7																	
E	-2	0	2	6																
N	0	-2	2	0	6															
H	-2	-2	0	0	1	10														
Q	-2	-1	0	2	0	1	6													
K	-2	-1	0	1	0	-1	1	5												
R	-2	-2	-1	0	0	0	1	3	7											
S	0	-1	0	0	1	-1	0	-1	-1	4										
T	-2	-1	-1	-1	0	-2	-1	-1	-1	2	5									
A	0	-1	-2	-1	-1	-2	-1	-1	-2	1	0	5								
M	-2	-2	-3	-2	-2	0	0	-1	-1	-2	-1	-1	6							
V	-3	-3	-3	-3	-3	-3	-3	-2	-2	-1	0	0	1	5						
I	-4	-2	-4	-3	-2	-3	-2	-3	-3	-2	-1	-1	2	3	5					
L	-3	-3	-3	-2	-3	-2	-2	-3	-2	-3	-1	-1	2	1	2	5				
F	-3	-3	-4	-3	-2	-2	-4	-3	-2	-2	-1	-2	0	0	0	1	8			
Y	-3	-3	-2	-2	-2	2	-1	-1	-1	-2	-1	-2	0	-1	0	0	3	8		
W	-2	-3	-4	-3	-4	-3	-2	-2	-2	-4	-3	-2	-2	-3	-2	-2	1	3	15	
C	-3	-4	-3	-3	-2	-3	-3	-3	-3	-1	-1	-1	-2	-1	-3	-2	-2	-3	-5	12

Рис. 2. Матрица Blosum 45

Матрицы серии PAM (рис. 3), предложенные М. Дэйхофф, широко используются с 70-х годов. Основными отличиями матриц PAM и Blosum являются: 1) использование матрицами PAM простой эволюционной модели (подсчет замен на ветвях филогенетического древа); 2) матрицы PAM основаны на учете мутаций по принципу глобального выравнивания (в высококонсервативных и высокомутабельных участках), а матрицы Blosum – локального (только высококонсервативных участков); 3) для матриц PAM замены в группах последовательностей подсчитываются сходным образом.

	C	G	P	S	A	T	D	E	N	Q	H	K	R	V	M	I	L	F	Y	W
C	12																			
G	-3	5																		
P	-3	-1	6																	
S	0	1	1	1																
A	-2	1	1	1	2															
T	-2	0	0	1	1	3														
D	-5	1	-1	0	0	0	4													
E	-5	0	-1	0	0	0	3	4												
N	-4	0	-1	1	0	0	2	1	2											
Q	-5	-1	0	-1	0	-1	2	2	1	4										
H	-3	-2	0	-1	-1	-1	1	1	2	3	6									
K	-5	-2	-1	0	-1	0	0	0	1	1	0	5								
R	-4	-3	0	0	-2	-1	-1	-1	0	1	2	3	6							
V	-2	-1	-1	-1	0	0	-2	-2	-2	-2	-2	-2	-2	4						
M	-5	-3	-2	-2	-1	-1	-3	-2	0	-1	-2	0	0	2	6					
I	-2	-3	-2	-1	-1	0	-2	-2	-2	-2	-2	-2	-2	4	2	5				
L	-6	-4	-3	-3	-2	-2	-4	-3	-3	-2	-2	-3	-3	2	4	2	6			
F	-4	-5	-5	-3	-4	-3	-6	-5	-4	-5	-2	-5	-4	-1	0	1	2	9		
Y	0	-5	-5	-3	-3	-3	-4	-4	-2	-4	0	-4	-5	-2	-2	-1	-1	7	10	
W	-8	-7	-6	-2	-6	-5	-7	-7	-4	-5	-3	-3	2	-6	-4	-5	-2	0	0	17

Рис. 3. Матрица PAM 250.

Матрицы этих двух серий сопоставимы следующим образом PAM 100 – Blosum 90, PAM 120 – Blosum 80, PAM 160 – Blosum 60, PAM 200 – Blosum 52, PAM 250 – Blosum 45. Наиболее часто используются матрицы Blosum 62 и PAM 160 (при среднем сходстве последовательностей). При выравнивании близко родственных последовательностей следует использовать матрицы Blosum с большим порядковым номером и матрицы PAM с меньшим номером.

Матрицы Gonnet (рис. 4) представляют собой усовершенствованный вариант матриц Дэйхофф, основанный на большей базе данных. Использование этой матрицы наиболее целесообразно для инициальных сравнений.

Рис. 4. Матрица Gonnet

Быстрое выравнивание (последовательности выравниваются с помощью поиска длинных сходных участков «к-плетов», затем эти наиболее сходные участки образуют «блоки» выравнивания). Быстрое выравнивание также характеризуется 4 параметрами:

· размер идентичного участка (K-tuple size). По умолчанию равен 1 для белков и 2 для НК. Для увеличения скорости (но уменьшения точности) можно увеличить до 2 для белков и 4 для НК. Для выравнивания последовательностей длиной более 1000 аминокислот или нуклеотидов необходимо уменьшать этот параметр.

- штраф на введение делеции (gap penalty). При быстром выравнивании не оказывает существенного влияния на его скорость и точность.

- число непрерывно совпадающих к-плетов (top diagonals) на участке парного выравнивания (если к=1, то это просто длина совпадающего сегмента). Для

построения выравнивания выбираются только сегменты, превышающие это порог. Для увеличения скорости можно уменьшить этот параметр, а для увеличения точности – увеличить.
- длина сегмента, включающего “наилучший выровненный сегмент” (window size). Для увеличения скорости надо уменьшать этот параметр, для увеличения точности – увеличивать.
2. Построение на основании попарных сравнений направляющего дерева (guide-tree). Первоначально методом NJ (neighbor-joining, связывания ближайших соседей) строится бескорневое дерево. Затем устанавливается корень по методу Томсона-Хиггинса-Гибсона таким образом, чтобы значения длин ветвей по отношению к корню остались неизменными.
3. Множественное выравнивание является основой программ Clustal, однако детали его очень сложны. Каждый этап множественного выравнивания состоит из сопоставления двух последовательностей или выравниваний, выполняемого в соответствии с ветвлением дендрограммы. Основными параметрами множественного выравнивания являются:
- штрафы за внесение делеции (gap penalties) устанавливаются как в попарном выравнивании.
- отсрочка различающихся последовательностей (delay divergent sequences) обеспечивает первоочередное выравнивание более сходных последовательностей.
- вес транзиций (transition weight) (А-Г или Ц-Т) имеет значения между 0 и 1. Если вес равен 0, то транзиция рассматривается как несовпадение. Если вес равен 1, то транзиция рассматривается как совпадение (алфавит из 4-буквенного вырождается в двухбуквенный пурин-пиримидин). Для слабо сходных последовательностей вес транзиций должен быть близок к 0, для близкородственных – к 1.
- матрицы сравнения нуклеотидов или аминокислот.
Полученное выравнивание может быть отображено в черно-белой или цветной гамме. Идентичные аминокислотные остатки или нуклеотиды отмечаются звездочкой (*), консервативные замены – двоеточием (:), а полуконсервативные – точкой (.). Консервативность и полуконсервативность аминокислотных замен определяются в соответствии с таблицей 1. Если заменяемые аминокислоты расположены в одной группе, то замена считается консервативной. Результаты выравнивания можно загрузить с помощью приложения Jal View для последующего анализа последовательностей.
Таблица 1
Группы аминокислот, используемые для определения консервативности и полуконсервативности замены при выравнивании последовательностей

Аминокислоты	Цветовое обозначение
AVFPMILW	красный
DE	синий
RHK	сиреневый
STYHCNGQ	зеленый
Другие	серый

Используемые по умолчанию параметры выравниваний

В таблице 2 приведены параметры, используемые по умолчанию на двух серверах (http://npsa-pbil.ibcp.fr и http://www.ebi.ac.uk).

Таблица 2

Используемые по умолчанию параметры Clustal W

Параметр / сервер		http://npsa-pbil.ibcp.fr	http://www.ebi.ac.uk
Медленное попарное выравнивание	штраф за внесение делеции	15 / 10	15 / 10
	штраф за продолжение делеции	6.66 / 0.1	6.66 / 0.2
	матрица сравнений нуклеотидов / аминокислот	IUB / Gonnet	DNA identity / Gonnet
Быстрое попарное выравнивание	размер участка максимального совпадения	2 / 1	.
	штраф на внесение делеции	5 / 3	15 / 10
	число непрерывно совпадающих k-плетов	4 / 5	.
	длина сегмента, включающего "наилучший выровненный сегмент"	4 / 5	.
Множественное выравнивание	штрафы за внесение делеции	15 / 10	15 / 10
	матрицы сравнения нуклеотидов и аминокислот	IUB / Gonnet	DNA identity / Gonnet

Примечание. Первое значение в каждой ячейке относится к последовательностям нуклеиновых кислот, второе – к последовательностям белков.

При выравнивании последовательностей нет необходимости указывать все используемые параметры. Как правило, достаточно указать сервер, на котором проводилось выравнивание, и отметить стандартность его условий.

Применение программ Clustal

Основным предназначением выравниваний, проведенных с помощью программ Clustal, является вычисление на их основании эволюционных дистанций между аминокислотными и нуклеотидными последовательностями, синонимичной и несинонимичной дистанций, определение характера и типа аминокислотных замен и т. д.

В ходе выравнивания также выявляются консервативные участки последовательностей, которые могут являться элементами вторичной структуры, сайтами связывания лигандов и другими функциональными мотивами. Это используется для предсказания вторичной и третичной структуры и функции белков, а также для идентификации новых представителей белковых семейств.

Кроме этого, программы Clustal используются для построения дендрограмм, показывающих филогенетические отношения сравниваемых последовательностей без учета (кладограммы) или с учетом длин ветвей (филограммы).

Дот-матрицы

При высоком сходстве последовательностей и небольшом числе делеций выравнивание последовательностей может быть получено с помощью точечных матриц гомологии (дот-матриц). При этом одна из последовательностей располагается горизонтально, а другая – вертикально. Если символы в строке и колонке совпадают, то на их пересечении ставится точка. При сопоставлении двух идентичных последовательностей поставленные точки образуют сплошную линию, при наличии делеций – линия разрывается и ее участки смещены влево или вправо. В случае большой длины последовательностей они разбиваются на подслова длины n, точка в позиции i, j ставится только, когда в двух подсловах, начинающихся в позициях i и j, совпадает не менее k символов.

Литература

1. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., Thompson, J.D. //Nucl. Acids Res. – 2003. – Vol. 31 913). – P. 3497-3500.
2. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C. // In atlas of protein sequence and structure. – 1978, NBRF, Washington.-Vol. 5, suppl. 3 (Dayhoff, M.O., ed.). – P. 345-352.
3. Feng, D.F., Doolitle, R.F. //J. Mol. Evol. – 1987. – Vol. 25. – P. 351-360.
4. Gonnet, G.H., Cohen, M.A., Benner, S.A. // Science. – 1992. – Vol. 256. – P. 1443-1445.
5. Henikoff, S., Henikoff, J.G. //Proc. Natl. Acad. Sci. – 1992. – P. 10915-10919.
6. Higgins, D.G. //Methods Mol. Biol. – 1994. – Vol. 25. – P. 307-318.
7. Higgins, D.G., Bleasby, A.J., Fucks, R. //Comput. Appl. Boisci. – 1992. – Vol. 8. – P. 189-191.
8. Higgins, D.G., Sharp, P.M. //Gene. – 1988. – Vol. 73. – P. 237-244.
9. Saitou, N., Nei, M. //Mol. Biol. Evol. – 1987. – Vol. 4. – P. 406-425.
10. Taylor, W.R. //J. Mol. Evol. – 1988. – Vol. 28. – P. 161-169.
11. Tompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G. //Nucl. Acids Res. – 1997. – Vol. 25. – P. 4876-48882.
12. Tompson, J.D., Higgins, D.G., Gibson, T.J. //Nucl. Acids Res. – 1994. – Vol.22. – P.4673 – 4680.