Д.Н. Штокало, Е.С Черемушкин

ПОСТРОЕНИЕ ПРОГРАММНОГО КОМПЛЕКСА “REGULATORY SEQUENCES ANALYZER” ДЛЯ РАСПОЗНАВАНИЯ ЦИС-ЭЛЕМЕНТОВ В ПОСЛЕДОВАТЕЛЬНОСТЯХ ДНК[1]

ВВЕДЕНИЕ

Перед авторами стояла следующая задача — наглядная, удобная для пользователя реализация алгоритмов **Match** и **CoMatch**, т.е. алгоритмов распознавания потенциальных цис-элементов последовательности ДНК с использованием весовых матриц. Поиск потенциальных цис-элементов должен производиться на прямой либо обратной цепи последовательности.

Наглядная реализация подразумевает возможность динамически в процессе выполнения программы осуществлять выбор искомых сайтов, выбор порогового значения, мгновенный и понятный вывод результатов в виде прокручиваемой ДНК последовательности и выделение цветом найденных сайтов.

1. АЛГОРИТМЫ И ИСПОЛЬЗУЕМЫЕ ДАННЫЕ

Match и **CoMatch** относятся к алгоритмам весовых матриц. Основная идея алгоритма весовых матриц [1, 2] заключается в приписывании четырех весов каждой позиции сайта в соответствии с четырьмя нуклеотидам A, T, G и C. Эти веса связаны с частотой встречаемости конкретного нуклеотида в конкретной позиции.

1.1. Алгоритм Match

Пусть $F = \left| f_{ij} \right|$ — l×4-матрица нуклеотидных частот, f_{ij} — абсолютная частота встречаемости i-го нуклеотида на j-ой позиции в обучающей

[1] cher@bionet.nsc.ru

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (грант № 05-08-33448).

выборке выровненных нуклеотидных фрагментов, кодирующих известные сайты связывания ($I = 1, .., l; j = 1, .., 4$). Далее определяется весовая матрица W. Существует много способов определения элементов весовой матрицы w_{ij}. Рассмотрим способ, реализованный в “Regulatory Sequence Analyzer”:

$$w_{ij} = I(i) * f_{ij}, \quad \text{где} \quad I(i) = \left| \sum_{B \in \{A,C,G,T\}} f_{i,B} * \ln(4 * f_{i,B}) \right|.$$

Таблица

Пример частотной(F) и весовой (W) матрицы

		позиции сайта							
		123	124	125	126	127	128	129	130
	‘A’	49	0	288	26	77	67	45	50
F:	‘C’	48	303	0	81	95	118	85	96
	‘G’	69	0	0	116	0	46	73	56
	‘T’	137	0	15	80	131	72	100	101
	‘A’	9.7	19.4	30.9	0	75	0	0	0
W:	‘C’	19.4	9.7	0	74.9	0		12.5	0
	‘G’	19.4	19.4	10.2	0	0	49.9	50	0
	‘T’	0	0	10.2	0	0	12.5	0	74.9

Процедура распознавания функционального сайта (характеризуемого весовой матрицей W) в произвольном нуклеотидном фрагменте длины L заключается в сопоставлении величины *match* и заранее заданного порогового значения $match^{(crit)}$:

$$match = \frac{x - x_{\min}}{x_{\max} - x_{\min}},$$

где $x_{\max} = \sum_{i=1}^{L} \max_j(w_{ij})$, $x_{\min} = \sum_{i=1}^{L} \min_j(w_{ij})$, а значение x оценивает степень близости тестируемого фрагмента и обучающей выборки:

$$x = \sum_{i=1}^{L} w_{i,B}, \quad B \in \{A, C, G, T\}.$$

Все потенциальные сайты в заданной нуклеотидной последовательности распознаются с помощью применения вышеизложенного алгоритма к каждому скользящему окну из этой последовательности.

1.2. Алгоритм распознавания двойных сайтов CoMatch

Сайты связывания некоторых транскрипционных факторов состоят из двух полусайтов с варьирующимся расстоянием между ними. Расстояние между полусайтами зависит от типа фактора, узнающего этот сайт. Полусайты могут иметь схожую структуру. Так как сайт состоит из 2-х консервативных доменов с варьирующим расстоянием между ними (рис. 1.1), то зададим double-core модели распознавания M_k следующим образом[4]:

$$M_k = \langle m_1, m_2, d_1, d_2 \rangle,$$

где m_1 и m_2 — весовые матрицы [2], d_1, d_2 — минимальное и максимальное расстояния между половинками сайтов. Пусть $w_1(i)$ и $w_2(j)$ веса m_1 и m_2 в позиции i и j соответственно на последовательности. Сайт считается распознанным, если вес $w = \frac{w_1(i) + w_2(j)}{2}$ больше заданного порога C и расстояние между половинками сайтов — $d \in [d_1, d_2]$.

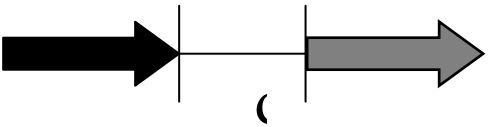

Рис. 1.1. Сайт состоит из 2-х консервативных доменов с варьирующим расстоянием между ними

Распознавание двойных сайтов будем производить следующим образом. Если на последовательности был распознан характерный полусайт, рассмотрим, какой максимальный вес w_k распознавания дает каждая из моделей M_k. Если модель M_k не распознана в данном районе, то считаем $w_k = 0$. Рассмотрим алгоритм получения моделей M_k. Пусть $S = (S_1, ..., S_m)$ — обучающая выборка последовательностей сайтов. Для каждого подмножества $S' = (S_1', ..., S_m')$ множества S зададим два набора подпоследовательностей $S^1 = (s_1^1, ..., s_n^1)$ и $S^2 = (s_1^2, ..., s_n^2)$, $s_i^1, s_i^2 \in S_i'$, длина s_i^j равна 6.

Найдем с помощью классической процедуры гиббс самплинга [5] S^1 и S^2 такие, что s_i^1 похожи друг на друга и s_i^2 похожи между собой. По S^1 и S^2 создадим соответствующие матрицы m_1 и m_2. Затем выберем расстояния $d_1 = \min_i \left(d(s_i^1, s_i^2) \right)$ и $d_2 = \max_i \left(d(s_i^1, s_i^2) \right)$. Выберем начальное подмножество $S_{[0]}$, называемое базовой выборкой. Теперь построим модель $M_{[0]}$ и добавим в $S_{[0]}$ последовательность из $S \setminus S_{[0]}$, для которой вес $w_{[0]}$ модели $M_{[0]}$ максимален. Таким образом получим модель $M_{[1]}$. Будем продолжать процедуру добавления до тех пор, пока вес $w_{[k]}$ превышает изначально заданный порог C (рис. 1.2).

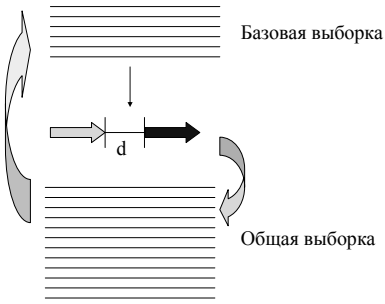

Рис 1.2

После окончания процедуры получим модель M, описывающую выборку S. Таким образом получим различные модели $M_1,...,M_T$ для различных классов сайтов.

1.3. Библиотека матриц, используемая программой при реализации алгоритма Match

Файл с матричной библиотекой [3], открываемый в главном окне программы, имеет расширение .dat и содержит в себе записи, перечисляющие частотные матрицы F, снабженные комментарием.

Пример записи, взятой из библиотеки матриц:

```
//
AC  M00028
XX
ID  I$HSF_01
XX
DT  18.10.1994 (created); ewi.
DT  16.10.1995 (updated); ewi.
CO  Copyright (C), Biobase GmbH.
XX
NA  HSF
XX
DE  heat shock factor (Drosophila)
XX
BF  T00386; HSTF; Species: fruit fly, Drosophila melanogaster.
XX

P0      A     C     G     T
01     31    14     4     1     A
02      0     0    50     0     G
03     48     2     0     0     A
04     43     1     5     1     A
05     23     1    14    12     N
XX
BA  50 functional genomic HSEs
XX
CC  not included are sequences with more than 2 mismatches at positions 2 to 4;
CC  the matrix only describes the properties of the basic 5-bp unit of which
CC  three have to be present to constitute a minimal HSE
XX
RN  [1]
```

```
RX  MEDLINE; 94167242.
RA  Fernandes M., Xiao H., Lis J. T.
RT  Fine structure analyses of the Drosophila and Saccharomyces heat shock
RT  factor-heat shock element interactions
RL  Nucleic Acids Res. 22:167-173 (1994).
XX
//
```

Необходимо, чтобы запись содержала строки, начинающиеся с управляющих символов:

```
AC          — начало записи
ID          — имя матрицы
P0          — начало матрицы
XX          — конец матрицы
//          — конец записи
```

1.4. Библиотека двойных сайтов, используемая программой при реализации алгоритма CoMatch

Пользователю необходимо открыть два библиотечных файла в главном окне программы. Первый — с расширением .clib, второй — формата .matrixlib.

Файл формата .clib содержит информацию о двойном сайте, т.е. его имя, имена составляющих матриц, номера матриц в файле .matrixlib и др.

Файл формата .matrixlib должен перечислять матрицы, упомянутые в файле .clib. В отличие от .dat, файл .matrixlib содержит не частотные матрицы F, а уже весовые, т.е. W.

Открыв файл .clib, программа автоматически попросит открыть файл .matrixlib.

Фрагмент файла в формате .clib:

```
AC  E00001
ID  V$C_AR_G01
I1  M00001
M1  V$AR_HG01
S1  0.0
I2  M00002
M2  V$AR_HG02
S2  0.0
OR  > >
DI  0 6
//
```

Фрагмент файла в формате .matrixlib:

```
AC  M00001
ID  V$AR_HG01
NA
MATR_LENGTH  6
CORE_START  1
CORE_LENGTH 5
MAXIMAL 5421.14879048849
MINIMAL 0
THRESHOLD 0.5
WEIGHTS
1 A:654.318988544113 C:0 G:373.89656488235 T:0
2 A:821.353332948717 C:0 G:308.007499855769 T:0
3 A:144.009999807692 C:108.007499855769 G:144.009999807692 T:0
4 A:120.281615774509 C:420.98565521078 G:0 T:120.281615774509
5 A:1398.72445341174 C:0 G:0 T:139.872445341174
6 A:0 C:0 G:1981.75636056544 T:0
//
AC  M00002
ID  V$AR_HG02
NA
MATR_LENGTH  6
CORE_START  1
CORE_LENGTH 5
MAXIMAL 8008.98919012654
MINIMAL 5.33833323717916
THRESHOLD 0.5
WEIGHTS
1 A:0 C:0 G:0 T:1981.75636056544
2 A:0 C:0 G:1398.72445341174 T:139.872445341174
3 A:0 C:0 G:0 T:1981.75636056544
4 A:8.00749985576874 C:5.33833323717916 G:5.33833323717916
T:10.6766664743583
5 A:0 C:1981.75636056544 G:0 T:0
6 A:373.89656488235 C:0 G:0 T:654.318988544113
//
```

1.5. Поиск цис-элементов на комплиментарной цепи данной последовательности

Если пользователь пометил «галочкой» пункт Reverse DNA на главном окне, то после нажатия кнопки «Calculate» программа произведёт поиск цис-элементов на комплиментарной цепи ДНК, т.е. программа вычислит комплиментарную цепь и примет её за основную. Более эта пометка на работу программы не повлияет.

2. РАБОТА С ПРОГРАММОЙ

При запуске программы проследите за наличием в той же директории папки Pictures. Запущенная программа предоставляет *Главное окно* для выбора желаемого алгоритма и файлов с биологическими данными.

2.1. Главное окно

Настройки Главного окна повлияют на результат работы программы только после нажатия кнопки “Calculate”.

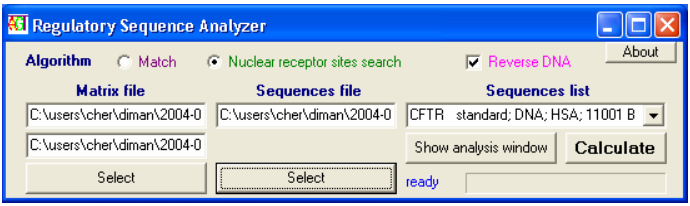

Рис. 2.1. Главное окно программы

2.1.1. В колонке “Algorithm” производится выбор алгоритма Match (см. 1.1) либо CoMatch (см. 1.2), а также, при желании произвести поиск на комплиментарной цепи последовательности, помечается «галочкой» пункт “Reverse DNA” (см. 1.5).

2.1.2. В колонке “Matrix file” производится открытие библиотек с данными о сайтах. Если отмечен пункт Match, программа позволяет открыть файлы с расширением .dat (см. 1.3). Второе редакционное поле колонки “Matrix file” при этом блокируется. Если отмечен пункт CoMatch, возможно открытие файлов композиционной библиотеки формата .clib, а также при

необходимости открытие файлов матричной библиотеки .matrixlib (см. 1.4).

2.1.3. В колонке “Sequences file” производится выбор файла библиотеки последовательностей формата .embl. После выбора файла происходит автоматическое построение списка имён последовательностей ДНК, содержащихся в embl файле. Этот список приготовлен для выбора в “Sequences list”.

2.1.4. При открытии необходимых файлов становится возможен обсчёт данных и немедленный вывод результатов в автоматически открываемом *Окне результатов*. Для этого нажмите кнопку “Calculate”. Стадия готовности результата будет отражена заполняющейся полосой “ready”.

2.1.5. Кнопка “Show analysis window” предназначена для показа *Окна результатов*. Если до момента нажатия этой кнопки они ещё не были получены программой, нажатие не произведёт никакого эффекта.

2.2. Окно результатов

На любое действие пользователя с элементами этого окна программа реагирует немедленно и полностью отражает результаты, соответствующие статусу интерфейса.

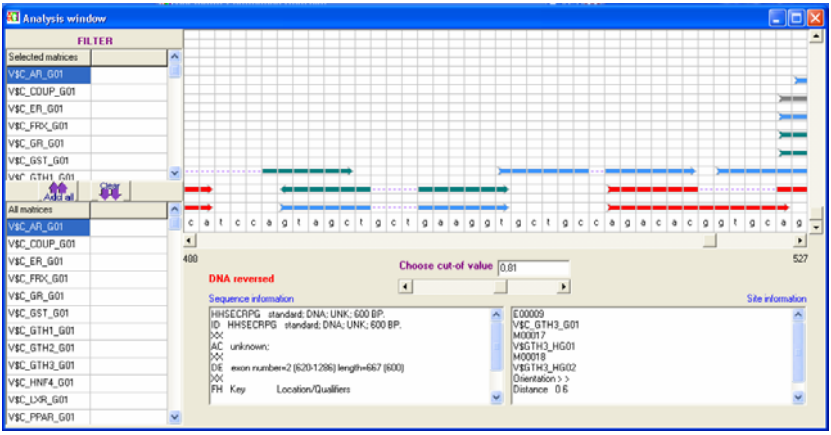

Рис. 2.2. Окно результатов

В окне расположены:

- две таблицы слева для фильтрации сайтов;
- два поля внизу для отображения информации о последовательности и каком-либо выбранном сайте;
- центральный скроллер для динамического выбора порога “cut-off value”;
- поле для наблюдения последовательности и стрелок, соответствующих найденным программой сайтам.

2.2.1. Таблица “All matrices” содержит список всех матриц (моделей), которые искала программа. “Клик” на ячейку левой колонки добавляет соответствующую матрицу (модель) в таблицу “Chosen matrices”. Если “Chosen matrices” уже имеет данную матрицу (модель), добавления не происходит. Нажатие кнопки “Add all” влечёт занесение в “Chosen matrices” списка всех имеющихся матриц (моделей). “Клик” на ячейку правой колонки выводит на поле “Site information” информацию о данной матрице (модели).

2.2.2. Таблица “Chosen matrixes” содержит список матриц (моделей), результат поиска которых желает видеть пользователь. Пользователю видны только стрелки, соответствующие матрицам (моделям) из данной таблицы. Клик на ячейку левой колонки удаляет из таблицы соответствующую матрицу (модель), кнопка “Clear” очищает таблицу. “Клик” на ячейку правой колонки выводит на поле “Matrix information” информацию о данной матрице (модели) и, кроме того, выделяет стрелки соответствующего сайта красным цветом.

2.2.3. Центральный скроллер выбора порога “cut-off value” позволяет выбрать порог в интервале [0.5, 1]. Результат изменения порога немедленно виден на поле результатов. Повышающийся порог уменьшает количество выводимых стрелок, понижающийся — увеличивает.

2.2.4. На поле результатов виден участок последовательности, как правило, длиной в 40 символов. Выше расположены изображения сайтов в виде стрелок. Покрываемый стрелкой интервал последовательности соответствует найденному цис-элементу. “Клик” по стрелке выделит красным цветом все стрелки, соответствующие этой матрице (модели), и заполнит “Site information” информацией. Стрелка, обращённая вправо, означает положение найденного цис-элемента на прямой цепи ДНК, обращённая влево — на обратной цепи. Поле вывода снабжено скроллерами прокрутки, так что пользователю сразу легко доступна итоговая информация обо всей последовательности.

ЗАКЛЮЧЕНИЕ

Реализована программная система для визуализации результатов работы алгоритмов **Match** и **CoMatch**, т.е. алгоритмов распознавания потенциальных цис-элементов последовательности ДНК с использованием весовых матриц. Поиск потенциальных цис-элементов производится на прямой либо обратной цепи последовательности. Наглядная реализация позволяет в процессе выполнения программы динамически осуществлять выбор искомых сайтов, выбор порогового значения, мгновенный и очевидный вывод результатов в виде прокручиваемой ДНК последовательности и отмеченных цветной стрелкой найденных сайтов.

СПИСОК ЛИТЕРАТУРЫ

1. **Schneider T., Stephens R.** Sequence logos: a new way to display consensus sequences // Nucleic Acids Res. — 1990. — Vol. 18. — P. 6097–6100.
2. **Kel A.E., Gossling E., Reuter I., et al.** MATCH: A tool for searching transcription factor binding sites in DNA sequences // Nucleic Acids Res. — 2003. — Vol. 31, N 13. — P. 3576–3579.
3. **Wingender E., Chen X., Fricke E., et al.** The TRANSFAC system on gene expression regulation // Nucleic Acids Res. — 2001. — Vol. 29, N 1. — P. 281–283.
4. **Черемушкина Е., Черемушкин Е., Чекменев Д., Кель О.** Метод идентификации сайтов ядерных рецепторов // Тез. конференции-конкурса «Технологии Microsoft в информатике и программировании», 21–23 февраля 2004, Новосибирск. — Новосибирск, 2004. — С. 137–139.
5. **Lawrence, C.E., Altschul, S.F., Bogouski, M.S., et al.** Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment // Science. — 1993. — Vol. 262, Iss. 5131. — P. 208–214.