Е. С. Черемушкин, Т. Г. Коновалова,
Ф. А. Мурзин, А. Э. Кель

СИСТЕМА РАСПОЗНАВАНИЯ ЦИС-ЭЛЕМЕНТОВ НА ПОСЛЕДОВАТЕЛЬНОСТЯХ ДНК*

ВВЕДЕНИЕ

В настоящее время проведены основные экспериментальные работы по секвенированию нуклеотидных последовательностей для достаточно большого числа организмов [1,2]. Для хранения получаемой первичной информации создан и постоянно пополняется ряд баз данных как специализированных, так и широкого профиля. В то же время, несмотря на наличие большого количества отсеквенированных последовательностей, наши представления о принципах их организации весьма ограничены. Поэтому одним из ведущих направлений молекулярной биологии в последнее время становится компьютерный анализ генетических текстов.

Проблематика понимания структурно-функциональной организации генома эукариот включает в себя широкий круг проблем. Наряду с такими вопросами, как распознавание интронов, экзонов или сайтов сплайсинга, существует все увеличивающийся круг задач, связанных с регуляцией транскрипции генов позвоночных. В последнее время появилось большое количество разнообразных данных (таких как SNP или паттерны экспрессии), позволяющих углубить понимание механизмов регуляции экспрессии генов. Одним из базовых понятий, занимающих ключевую роль в процессах транскрипции, является понятие транскрипционных факторов, которые представляют собой регуляторные белки, обладающие способностью распознавания специфических коротких участков ДНК. Поэтому детальному изучению и распознаванию соответствующих нуклеотидных фрагментов, называемых цис-элементами или сайтами связывания транскрипционных факторов (ССТФ или сайтами), отводится большое внимание. Несмотря на разнообразие подходов, проблема построения точных методов распознавания ССТФ в настоящее время не может считаться окончательно решенной. Причина этого состоит в большом разнообразии контекстных, физико-

* Работа выполнена при финансовой поддержке Министерства образования РФ (грант № E02-1.0-42).

химических и конформационных особенностей ССТФ; механизмов ДНК-белковых взаимодействий между ССТФ и транскрипционными факторами; специфичности контекста, окружающего ССТФ, степени консервативности нуклеотидного контекста в эволюции. Поэтому перспективным представляется применение методов, ориентированных в каждом конкретном случае на специфическую информацию, которой обладает биолог.

С этой целью был разработан комплекс алгоритмов идентификации цис-элементов и объектно-ориентированная среда, реализующая эти методы.

1. СУЩЕСТВУЮЩИЕ ДАННЫЕ

В распоряжении экспериментатора находится различная генетическая информация. От количества информации зависит качество распознавания.

Регуляторная последовательность. Это классический объект, которым обладает каждый экспериментатор. Задача, которую он решает с помощью определенного метода, — предсказание потенциальных сайтов определенного типа на этой последовательности. Отметим, что на основе только этой информации предсказание будет крайне неточным.

Гомологичные регуляторные последовательности разных организмов. В настоящее время экспериментатор зачастую обладает информацией об эволюционном сходстве изучаемого участка последовательности ДНК одного организма с некоторыми участками ДНК других организмов. Имея соответствующий метод, он может получить более точное предсказание цис-элементов на этом наборе последовательностей. В зависимости от уровня гомологии целесообразно использовать различные методы.

Функционально-связанные последовательности. Ввиду больших объемов аннотированной ДНК экспериментатор часто имеет набор функционально-зависимых промоторов генов, например, генов, вовлеченных в один биологический процесс. Таким образом, используя специфический метод, он имеет возможность предсказать транскрипционные факторы, регулирующие гены его выборки. Также может присутствовать несколько выборок.

Паттерн экспрессии генов. В настоящее время получили распространение паттерны экспрессии генов (expression pattern, microarray experiments). Используя специфическую технологию, экспериментатор получает набор чисел, соответствующих уровню экспрессии для большого

количества генов (порядка 20 000). Точность метода в данный момент не очень велика, но достаточна для широкого и эффективного использования этой информации. Зачастую одновременно используют результаты двух экспериментов: здоровой и больной клетки. Имея соответствующий метод, экспериментатор может предсказать цис-элементы, нарушающие нормальную деятельность организма, чтобы впоследствии воздействовать на соответствующие транскрипционные факторы.

Однонуклеотидные полиморфизмы. Однонуклеотидные полиморфизмы (Single-Nucleotide Polymorphism) — это различия в ДНК между индивидами одного вида. Они характеризуют индивидуальные особенности или особенности популяции (этноса). Полиморфизмы в регуляторных областях могут влиять на регуляцию. Таким образом, используя соответствующий метод, можно предсказывать изменение регуляции в связи с SNP.

2. МЕТОДЫ РАСПОЗНАВАНИЯ ЦИС-ЭЛЕМЕНТОВ

Авторами настоящей работы была создана объектно-ориентированная система, реализующая как созданные ранее, так и разработанные авторами методы. Данная среда является инструментарием для решения широкого круга задач распознавания цис-элементов.

2.1. Метод весовых матриц

Основная идея метода весовых матриц [3] заключается в приписывании четырех весов каждой позиции сайта в соответствии с четырьмя нуклеотидами A, T, G и C. Эти веса связаны с вероятностью появления конкретного нуклеотида в конкретной позиции.

Пусть $F = \left| f_{ij} \right|$ — 4×1 матрица нуклеотидных частот, f_{ij} — абсолютная частота встречаемости *i*-го нуклеотида на *j*-ой позиции в обучающей выборке выровненных нуклеотидных фрагментов, кодирующих известные сайты связывания (i=1,...,4; j=1,...,l). Элементы w_{ij} весовой матрицы W определяются соотношением:

$$w_{ij} = \ln\left(\frac{f_{ij}}{e_{ij}} + \frac{s}{100} \right) + c_i ,$$

где e_{ij} — ожидаемые частоты, соответствующие величинам f_{ij}, c_i — нуклеотидно-специфические константы, s — параметр сглаживания, измеряемый в процентах.

Таблица 1

Матрица нуклеотидных частот (F) и весовая матрица (W), вычисленные для кэп-сайта[1]

		позиции сайта							
		-2	-1	0	1	2	3	4	5
F:	‘A’	49	0	288	26	77	67	45	50
	‘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
W:	‘A’	-1.1	-5.3	0.0	-1.5	-0.7	-0.6	-0.9	-0.8
	‘C’	-1.2	0.0	-5.2	-0.4	-0.5	0.0	-0.3	-0.2
	‘G’	-0.8	-5.3	-5.2	0.0	-4.6	-0.9	-0.4	-0.7
	‘T’	0.0	-5.3	-2.7	-0.3	0.0	-0.4	0.0	0.0

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

$$match = 100 \times \frac{x - x_{\min}}{x_{\max} - x_{\min}},$$

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

$$x = \sum_{j=1}^{L} w_{ij}.$$

[1] Идея метода весовых матриц заключается в приписывании четырех весов каждой позиции сайта в соответствии с четырьмя нуклеотидами A, T, G и C. Эти веса связаны с вероятностью появления конкретного нуклеотида в конкретной позиции.

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

2.2. Метод распознавания двойных сайтов

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), то зададим double-core модели распознавания M_k следующим образом:

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

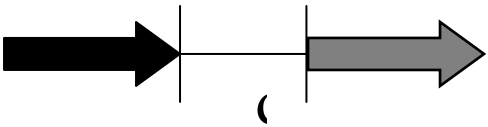

Рис.1. Сайт состоит из 2 консервативных доменов с варьирующимся расстоянием между ними

При этом m_1 и m_2 — весовые матрицы [3], 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 $w = \dfrac{w_1(i) + w_2(j)}{2}$ больше заданного порога c и расстояние между половинками сайтов $d \in [d_1, d_2]$.

Распознавание сайтов NR 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 метод 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 широко используемой в биоинформатике процедуры гиббссамплинга [4] S^1 и S^2 такие, что s_i^1 похожи между собой в терминах расстояния между последовательностями, и s_i^2 похожи между собой. На основе 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 коровой 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.

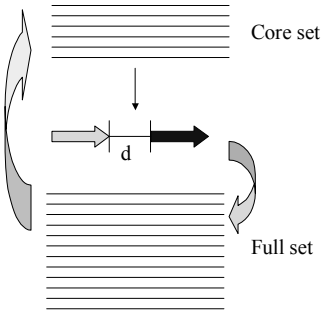

Рис.2. Процесс построения модели $M_{[i]}$

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

2.3. Филогенетический футпринт

Мы разработали алгоритм для выравнивания двух или более нуклеотидных последовательностей. Метод основывается на предположении, что в процессе эволюции цис-элементы более консервативны, чем другие участки промоторных последовательностей. Алгоритм схож с общепринятым алгоритмом Недельмана—Вунша[5]. Основные изменения сделаны в способе подсчета весов на нуклеотидные замены и в штрафах на делеции.

Штраф на делеции, при вставке гэпа в $\overline{S}^1$ между $k-1$ и k над позицией l в $\overline{S}^2$:

$$GAP\left(\overline{S}^1,\overline{S}^2,k,l\right)=\frac{G\left(\overline{S}^1,k\right)+R\left(\overline{S}^2,l\right)}{2}.$$

Штраф на замену:

$$SUB\left(\overline{S}^1,\overline{S}^2,k,l\right)=\mathrm{Z}\left(s_k^1,s_l^2\right),$$

где

$$G\left(\overline{S}^1,k\right)=\mathrm{Y}\left(s_{k-1}^1,s_k^1\right),$$

$$R\left(\overline{S}^2,l\right)=\frac{\mathrm{Y}\left(s_{l-1}^2,s_l^2\right)+\mathrm{Y}\left(s_l^2,s_{l+1}^2\right)}{2}$$

$$\mathrm{Y}\left(a,b\right)=\frac{C_{gap}}{N}+W_{gap}\cdot s_{gap}\left(a,b\right),$$

$$\mathrm{Z}\left(a,b\right)=\frac{\Delta}{N}\cdot C_{sub}-W_{sub}\cdot\frac{\sum_{i=1}^{3}\lambda_i\cdot s_i\left(a,b\right)}{\sum_{i=1}^{3}\lambda_i},\ \text{для}\ \ a,b\in\Sigma\times\Phi,$$

где

$$\Delta=\begin{cases}1,\gamma\left(a\right)\neq\gamma\left(b\right)\\0,\gamma\left(a\right)=\gamma\left(b\right)\end{cases},$$

$$s_{gap}\left(a,b\right)=\begin{cases}\left(\vec{\varphi}\left(a\right)+\vec{\varphi}\left(b\right)\right)^2,\gamma\left(a\right)=\gamma\left(b\right)\\\vec{\varphi}\left(a\right)^2+\vec{\varphi}\left(b\right)^2,\gamma\left(a\right)\neq\gamma\left(b\right)\end{cases},$$

$$s_1\left(a,b\right)=s_{gap}\left(a,b\right),$$

$$s_2(a,b) = \begin{cases} 0, m > C_{\min} \\ (C_{\min} - m)/C_{\min}, m \le C_{\min} \end{cases},$$

где

$$m = \min_i \left| \varphi_i(a) - \varphi_i(b) \right|,$$

$$s_3(a,b) = \max_i \left(\varphi_i(a) \cdot \varphi_i(b) \right),$$

$\gamma(a) \in \Sigma$ — нуклеотид,

$\vec{\varphi}(a) \in \Phi$ — вектор весов для матрицы,

C_{corr}, C_{gap}, W_{corr}, W_{gap}, λ_i — константы,

N — количество последовательностей.

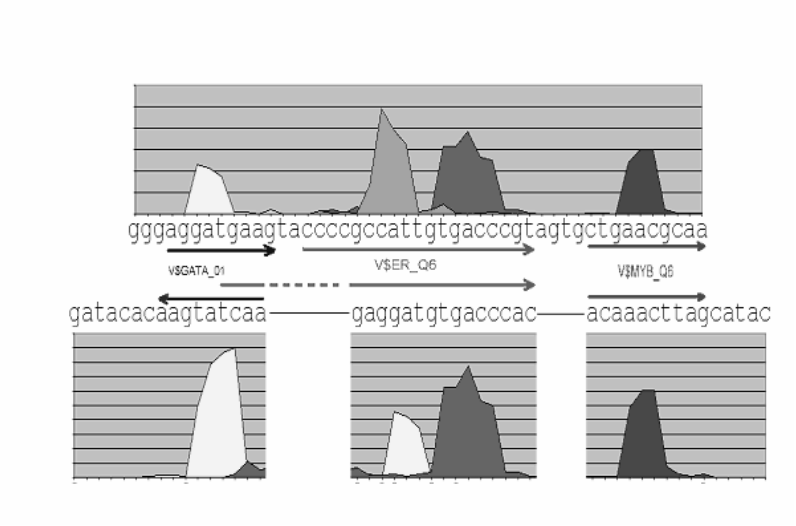

Рис. 3. Демонстрация работы алгоритма на примере двух сгенерированных последовательностей

Профили выравниваются друг с другом наряду с последовательностями. Стрелками обозначены потенциальные ССТФ и соответствующие им транскрипционные факторы.

2.4. Метод антифутпринта

При большой гомологии последовательностей имеет смысл рассматривать не сходные сайты, а различия. Считается вероятным, что эти различия легли в основу разницы между данными видами.

2.5. Метод анализа группы последовательностей

Определим композиционный модуль (КМ) как набор факторов с некоторыми параметрами (такими как вес матрицы). Зададим целевую функцию модуля F(S), характеризующую присутствие этого комплекса в последовательности S. Алгоритм получает на вход два набора последовательностей: анализируемый и фоновый. Далее выбирается комплекс, максимизирующий $R=(F_+ - F_-)/(\delta_+ + \delta_-)$, где F_+ и F_- средние, а δ_+ и δ_- — дисперсии распределения F на анализируемой и фоновой выборках соответственно [6].

2.6. Метод поиска цис-элементов на основе данных с паттернов экспрессии

Другой подход заключается в поиске цис-элементов на основе набора промоторов и соответствующих им значений, характеризующих уровни экспрессии. При этом ищутся, как и в предыдущем случае, композиционные модули. За целевую функцию R берется корреляция F и уровня экспрессии.

2.7. Метод поиска цис-элементов с учетом контекста

Предположим, что имеются две выборки последовательностей: позитивная $Q = \{q_1,...,q_m\}$ и негативная $T = \{t_1,...,t_k\}$. Позитивная выборка содержит последовательности, в которых присутствуют цис-элементы заданного типа, а негативная содержит последовательности, где таких цис-элементов нет. С помощью некоторого метода осуществим поиск сайтов и используем информацию о позитивной и негативной выборках для фильтрации сайтов. Введем правило $f(s) \in R$ такое, что если $f(s) > 0$, то s считается распознанным цис-элементом, иначе — не является. Зададим *f(s)* следующим образом: $f(s) = \sum_{i=0}^{N} f_i(s)$, где $f_i(s) = c_i^1$, если в районе $\left[p_i^1, p_i^2\right]$ присутствует последовательность (блок) s_i, и $f_i(s) = c_i^2$, если не присутст-

вует. Блок-моделью назовем тройку $< f_i, p_i^1, p_i^2 >$. Итак, по выборкам Q и T получаем блок-модели, а затем используем их при распознавании.

Для получения блок-моделей применим критерий максимального правдоподобия: $c_i^1 = \log(fr_i^1) - \log(fr_i^2)$, $c_i^2 = \log(1 - fr_i^1) - \log(1 - fr_i^2)$, где fr_i^1 — частота встречаемости блока s_i в районе $\left[p_i^1, p_i^2 \right]$ в выборке $Q = \{q_1, ..., q_m\}$, а fr_i^2 — частота встречаемости блока s_i в районе $\left[p_i^1, p_i^2 \right]$ в выборке $T = \{t_1, ..., t_k\}$. Далее выберем N моделей с наибольшей разностью $\left| c_i^1 - c_i^2 \right|$. По этим моделям будем проводить фильтрацию сайтов, найденных на произвольной последовательности. Если $f(s) > 0$, то s удовлетворяет фильтру, иначе — не удовлетворяет.

3. МОДУЛЬ СРАВНЕНИЯ МЕТОДОВ ПОИСКА

Качество распознавания может быть оценено распределением двух величин: ошибкой предсказания первого (*FP*) и второго (*FN*) родов. В зависимости от параметров поиска получим распределение этих ошибок. Введем следующие величины: значение предсказания $\alpha = 1 - FP$ и чувствительность $\beta = 1 - FN$. Пусть $S = \{s_1, ..., s_n\}$ — известные экспериментальные сайты, $Q = \{q_1, ..., q_m\}$ — сайты, найденные определенным методом. Обозначим $s_i \approx q_j$, это значит, что сайт s_i совпадает с сайтом q_j (распознан сайтом q_j). Пусть $Q' = \{q_j \in Q \mid \exists s_i \in S, s_i \approx q_j\}$ — правильно распознавшие сайты, $S' = \{s_i \in S \mid \exists q_j \in Q, s_i \approx q_j\}$. Тогда

$$\alpha = \frac{|Q'|}{|Q|}, \beta = \frac{|S'|}{|S|}.$$

Основная проблема состоит в том, что далеко не все сайты в геноме открыты и не все из открытых содержатся в соответствующих базах данных. Пусть $T = \{t_1, ..., t_k\}$ — неизвестные сайты. В объединении с известными неизвестные сайты дают все множество сайтов $S^* = S \cup T$. Перепишем значение предсказания и чувствительность с учетом неизвестных сайтов.

$$\alpha^* = \frac{|Q'^*|}{|Q|},\ \beta^* = \frac{|S'^*|}{|S^*|},$$

где $Q'^* = \{q_j \in Q \mid \exists s_i \in S^*, s_i \approx q_j\}$, $S'^* = \{s_i \in S^* \mid \exists q_j \in Q, s_i \approx q_j\}$. Можем записать, что $S'^* = S' \cup T'$ и $Q'^* = Q' \cup Q'_T$. Пусть неизвестных сайтов в k_T раз больше, чем известных $|T| = k_T |S|$. Пусть методы распознают меньший процент неизвестных сайтов, чем известных

$$\frac{|Q'_T|}{|Q'|} = k_\alpha \frac{|T|}{|S|} = k_\alpha \cdot k_T,\ k_\alpha \in (0,1].$$

Пусть также количество распознанных неизвестных сайтов зависит от количества распознанных известных сайтов, аналогично

$$\frac{|T'|}{|S'|} = k_\beta \frac{|T|}{|S|} = k_\beta \cdot k_T,\ k_\beta \in (0,1].$$

Тогда получим, что

$$\alpha^* = \alpha \cdot (1 + k_\alpha \cdot k_T),\ \beta^* = \beta \cdot \frac{1 + k_\beta \cdot k_T}{1 + k_T}.$$

Заметим, что в общем случае для различных транскрипционных факторов получаются различные константы. Константа k_T не зависит от исследуемого метода поиска. Так как сайты из множества T неизвестны, то предположим, что остальные константы k_α и k_β также не зависят от исследуемого метода. Тогда для сравнительного анализа методов достаточно использовать распределение $\langle \alpha, \beta \rangle$, имея в виду, что это не абсолютная, а относительная оценка методов. Качество метода распознавания варьируется для разных факторов, для разных групп последовательностей так же, как и для параметров метода. Параметры $\langle \alpha, \beta \rangle$ несравнимы для различных групп факторов и групп последовательностей, но сравнимы внутри одной группы факторов и последовательностей.

Сравнение реализовано в системе в виде модуля. Для добавления нового тестируемого метода достаточно реализовать функцию с использованием реализованных механизмов подсчета статистики. Если метод требует ис-

пользования дополнительных данных, то эти данные тоже должны быть добавлены таким образом, чтобы для подсчета статистики в методах использовался один и тот же набор генов и транскрипционных факторов. В процессе работы метода статистика сохраняется в виде, удобном для визуализации.

4. ОБЬЕКТНО-ОРИЕНТИРОВАННАЯ СИСТЕМА ПОИСКА ЦИС-ЭЛЕМЕНТОВ

Среда GRESA DT имеет иерархическую структуру. Вся функциональность разбита на классы, а классы сгруппированы в 3 основных пакета: ядро, набор общепринятых инструментов, набор экспериментальных инструментов.

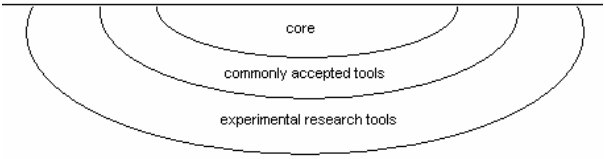

Рис. 4. Структура среды GRESA DT

Пакет «ядро» состоит из классов, представляющих основные общепринятые объекты биоинформатики регуляторных последовательностей ДНК.

Последовательность — последовательность ДНК. Представляет собой линейную последовательность нуклеотидов, обозначаемых буквами A,C,G,T. Также имеет название, описание и привязку к геному, т.е. номер хромосомы, стартовую позицию на хромосоме и направление “+” или “–”.

Сайт — подпоследовательность цепочки ДНК длиной, как правило, 10-20 нуклеотидов, имеющая позицию, длину, направление.

Фактор — объект, реализующий свойства транскрипционного фактора. Транскрипционный фактор — это белок, который связывается с сайтом на ДНК.

Выравнивание — несколько выровненных последовательностей. В каждой из них между нуклеотидами могут быть вставлены гэпы (промежутки). Выравнивание отражает эволюционное сходство последовательностей.

Набор последовательностей, сайтов, факторов — классы, в которых реализованы в основном сохранение и загрузка из общепринятых форматов, а также набор вспомогательных классов. Над объектами реализованы классические операции, такие как получение комплиментарной последовательности, поиск и др.

Набор общепринятых инструментов состоит из таких приложений, как MATCH, COMATCH (поиск композиционных модулей), footprint, CM SEARCH и др.

- MATCH — метод поиска сайтов на основе весовых матриц. Самый широко используемый в настоящее время метод.
- COMATCH — метод поиска композиционных элементов и сайтов с двумя доменами.
- FOOTPRINT — метод, учитывающий эволюционное сходство последовательностей. Вначале производится выравнивание последовательностей, а затем поиск сайтов, которые встретились на обеих последовательностях в одном и том же блоке выравнивания.
- CM SEARCH — метод поиска композиционных модулей, регулирующих группу генов. Для данной группы генов ищется общий модуль, предположительно регулирующий эти гены.

Набор экспериментальных инструментов состоит из еще не опубликованных приложений, находящихся в стадии разработки. Среди них наработки по поиску сайтов с использованием контекста, средства оценки качества распознавания методов.

Разработка и применение. Среда GRESA DT постоянно дополняется и развивается. Разработка среды по технологии Extreme Programming дает возможность постоянно поддерживать рабочую версию. Стабильность, при довольно большой и распределенной группе разработчиков, поддерживается за счет большого количества автоматизированных тестов. Жизненный цикл отдельного приложения состоит из этапов, когда приложение находится в стадии экспериментальной разработки, затем переходит в стабильную стадию.

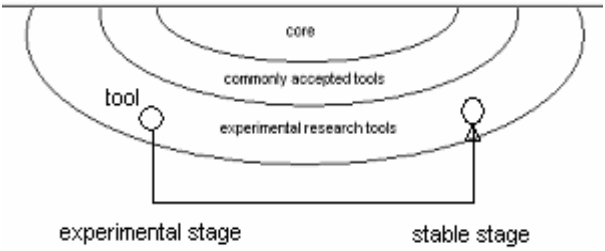

Рис. 5. Жизненный цикл отдельного приложения

Далее оно может перейти в набор общепринятых инструментов. Любой член команды может вносить изменения в любой класс, главное — сохранить успешное выполнение тестов.

На данный момент GRESA DT используется для обработки регуляторных ДНК последовательностей. Применение охватывает широкий круг задач распознавания сайтов. Также реализованы некоторые методы предсказания регуляции на основе предсказанных сайтов. Есть возможность построения комбинаций методов, например, как основу footprint-поиска сайтов можно взять либо результат match-поиска, либо результаты какого-либо другого метода, поддерживающего нужный формат записи. В GRESA DT также поддерживается сравнительное тестирование методов. Система тестирования оценивает качество распознавания сайтов. В данный момент в качестве выборки для тестирования используется база данных TRANSFAC [7].

Реализация и системные требования. Система реализована на языке C++ с использованием среды Microsoft Visual Studio. Операционная система Windows. Использованы технологии разработки ПО Extreme Programming и Microsoft Solution Framework. Системные требования зависят от задачи.

5. ЗАКЛЮЧЕНИЕ

Разработан и реализован инструментарий, позволяющий производить полноценный поиск цис-элементов, наиболее полно использующий данные, имеющиеся у экспериментатора. Каждый из методов может быть использован как в отдельности, так и в качестве дополнительного фильтра результа-

тов другого метода. Существует возможность простого и эффективного создания новых алгоритмов на базе уже существующих. Система эффективно используется в нескольких организациях.

СПИСОК ЛИТЕРАТУРЫ

1. Doolittle R. F. Microbial genomes opened up // Nature. — 1997. — Vol. 392. — P.339–342.
2. Maley L. E., Marshall C. R. The coming of age of molecular systematics // Science. — 1998. — Vol. 279. — P. 505–506.
3. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. MATCH: A tool for searching transcription factor binding sites in DNA sequences // Nucleic Acids Res. — 2003. — Vol. 31, N 13. — P. 3576–3579.
4. Lawrence, C.E., Altschul, S.F., Bogouski, M.S., Liu, J.S., Neuwald, A.F., and Wooten, J.C. Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment // Science. — 1993. — Vol. 262. — P. 208–214.
5. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins // J. Mol. Biol. — 1970. — Vol. 48, N 3. — P.443–53
6. Kel-Margoulis OV, Ivanova TG, Wingender E, Kel AE. Automatic annotation of genomic regulatory sequences by searching for composite clusters // Pac. Symp. Biocomput. — 2002. — Vol. 7. — P. 187–198.
7. Wingender E., Chen X., Fricke E., Geffers R., Hehl R., Liebich I., Krull M., Matys V., Michael H., Ohnhäuser R., Prüß M., Schacherer F., Thiele S. and Urbach S. The TRANSFAC system on gene expression regulation // Nucleic Acids Res. — 2001. — Vol. 29. — P. 281–283.