БИОФИЗИКА, 2019, том 64, № 3, с. 446–456

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 577.212.2

БАЗА ДАННЫХ ПОТЕНЦИАЛЬНЫХ СДВИГОВ РАМКИ СЧИТЫВАНИЯ В КОДИРУЮЩИХ ПОСЛЕДОВАТЕЛЬНОСТЯХ ИЗ РАЗЛИЧНЫХ ГЕНОМОВ ЭУКАРИОТ

© 2019 г. Ю.М. Суворова*, В.М. Пугачева*, Е.В. Коротков* **

**Федеральный исследовательский центр «Фундаментальные основы биотехнологии» РАН, 119071, Москва, Ленинский просп., 33/2*

***Национальный исследовательский ядерный университет «МИФИ», 115409, Москва, Каширское шоссе, 31*

E-mail: genekorotkov@gmail.com

Поступила в редакцию 06.02.2019 г.
После доработки 06.02.2019 г.
Принята к публикации 25.02.2019 г.

Разработан новый банк данных, содержащий потенциальные сдвиги рамки считывания (potential reading frame shifts). Для поиска потенциальных сдвигов рамки считывания использован новый математический метод, основанный на использовании генетического алгоритма и динамического программирования. Банк данных содержит координаты потенциальных сдвигов рамки считывания для кодирующих последовательностей 76 эукариотических геномов из геномного браузера Ensembl, версия 86. База данных расположена по адресу: http://victoria.biengi.ac.ru/cgi-bin/frameshift/index.cgi. Из всех проанализированных геномов примерно 23% кодирующих последовательностей имеют сдвиг рамки считывания. Ошибки первого и второго рода составляют приблизительно 11% и 30%. Одновременно с банком данных создан Web-сервер для поиска потенциальных сдвигов рамки считывания, который находится по адресу: http://victoria.biengi.ac.ru/fsfinder. Сервер может быть использован для поиска потенциальных сдвигов рамки считывания во вновь определенных кодирующих последовательностях.

Ключевые слова: сдвиг рамки считывания, динамическое программирование, позиционно-весовая матрица, кодирующая последовательность.

DOI: 10.1134/S0006302919030049

Одной из известных мутаций является сдвиг рамки считывания в гене, который может приводить к созданию мутантных белков [1]. С этой мутацией могут быть связаны различные наследственные и онкологические заболевания [2–5]. Сдвиг рамки считывания в гене создают вставки и делеции оснований, не кратные трем основаниям [6,7], и сдвиги границ экзонов и интронов [8]. Аминокислотная последовательность полностью меняется после позиции сдвига рамки [9], что может привести к изменению или потере функции, либо не вызвать никаких изменений. Поиск сдвигов рамки считывания важен, потому что сдвиги рамки считывания могут вносить большой вклад в эволюцию белковых последовательностей. Современные базы данных содержат последовательности, в которых могут быть сдвиги рамки считывания из-за неточностей при секвенировании и сборке геномов [10]. Ошибки сдвига рамки считывания в ридах очень часто наблюдаются для пиросеквенирования и PacBio. Это приводит к тому, что после точки сдвига рамки считывания возникает ошибочная аминокислотная последовательность, что затрудняет аннотацию генов. Существуют специальные программы для поиска и коррекции сдвигов рамки считывания для этих методов [11,12]. Поиск таких ошибок необходим для более точной аннотации генов. Образование альтернативного белка возможно при программируемом сдвиге рамки считывания. В этом случае сама рибосома программируемо сдвигается на одно основание [13]. Такие сдвиги рамки считывания лежат вне задач данного исследования. В данной работе мы создали базу данных, где содержатся потенциальные сдвиги рамки считывания (potential reading frame shifts), связанные с возможными вставками или делециями нуклеотидов

Сокращения: cds – кодирующая последовательность, HMM – hidden Markov models.

не кратные трем основаниям в кодирующих последовательностях (cds).

Для поиска сдвига рамки считывания существует два класса математических методов. Первый класс содержит методы, основанные на выравнивании последовательностей. Для поиска сдвига рамки считывания проводится поиск аминокислотных последовательностей, гомологичных аминокислотным последовательностям, полученным из изучаемой cds, но в других рамках считывания. Если среди найденных последовательностей есть хотя бы одна подобная, то можно говорить о существовании сдвига рамки считывания в исследуемой последовательности. Для поиска подобий часто используется программа Blast [7, 9, 14]. Однако если подобные последовательности не обнаруживаются, то определить существование сдвига рамки считывания в cds невозможно. Также помешать сделать вывод о существовании сдвига рамки считывания может низкий уровень подобия. Для работы с слабо подобными последовательностями были разработаны специальные программы [15]. Обнаружение сдвигов рамки считывания методами второго класса происходит посредством анализа cds *ab initio* [16–18]. На основе этих методов были разработаны программы FrameD [16, 19], FragGeneScan [20], программа HMM-frame [21]. Программа GeneTack часто применяется для поиска сдвигов рамки считывания [17, 22]. Этой программой могут быть найдены сдвиги рамки считывания, возникшие в результате мутаций или ошибок секвенирования. Она ищет сдвиги там, где предполагается эволюционное происхождение двух подряд идущих генов от одного гена, т.е. сдвиг рамки считывания разбивает один длинный ген на два новых. В программе GeneTack используются скрытые модели Маркова (hidden Markov models – HMM) и алгоритм Витерби. Для работы программы GeneTack требуется обучающая выборка для настройки программы, что существенно ограничивает ее возможности. В обучающей выборке статистические свойства каждой cds усредняются, что увеличивает число ошибок первого рода и значительно уменьшает мощность метода. Программа GeneMarkS [23] используется для обучения. GC-содержание генов также оказывает сильное влияние на поиск сдвигов рамки считывания.

Триплетная периодичность cds часто используется в методах по поиску сдвигов рамки считывания. Почти все cds содержат триплетную периодичность, которая возникает из-за неравномерного использования кодонов различными организмами [24, 25]. Многие программы по поиску кодирующих районов используют это свойство [26, 27]. Также известно [18], что сдвиг фазы триплетной периодичности возникает одновременно со сдвигом рамки считывания. Для поиска сдвига фазы триплетной периодичности используются преобразование Фурье [28], динамическое программирование [18], вэйвлет-преобразование [29] и другие методы [30]. Метод преобразования Фурье применим только для последовательностей длиннее 750 оснований и с сильно выраженной триплетной периодичностью [28]. Применение динамического программирования [18] также не может выявить сдвиги рамки считывания при слабо выраженной триплетной периодичности.

В данной работе мы применили новый метод поиска потенциальных сдвигов рамки считывания в cds из геномов эукариот. Применяемый подход лишен недостатков HMM, связанных с тем, что для HMM требуется использование выборки генов. Такая выборка может сильно усреднить частоты *k*-слов, которые встречаются в отдельных генах, что в свою очередь может сильно усреднить корреляции между нуклеотидами по сравнению с корреляциями, которые наблюдаются в отдельных генах. HMM также настраивается на корреляции нуклеотидов, которые характерны для обучающей выборки. В то же время корреляции нуклеотидов в анализируемом гене могут быть совершенно другими, чем в обучающей выборке. Это будет приводить к тому, что при помощи HMM не получится достоверно зарегистрировать сдвиги фазы. Ранее нам удалось при помощи генетического алгоритма определить для каждого гена наилучшую матрицу триплетной периодичности [31]. В данной работе этот подход был расширен посредством использования матрицы, учитывающей корреляцию соседних оснований. Это означает, что наилучшие корреляции нуклеотидов могут быть определены для каждого изучаемого гена индивидуально.

В данной работе мы провели финальное выравнивание изучаемой последовательности относительно найденной матрицы. В ходе этого выравнивания мы определяли места потенциальных сдвигов рамки считывания. В cds из генома *A. thaliana* было обнаружено 9930 cds, имеющих потенциальные сдвиги рамки считывания (при уровне ошибок первого и второго рода в 11% и 30%). Это составляет примерно 21% от всех cds из этого организма. Аналогичные результаты были получены для 76 геномов, накопленных в базе данных по адресу: ftp://ftp.ensembl.org/pub/release-86/embl/. Все полученные результаты собраны в базу данных потенциальных сдвигов рамки считывания, которая находится по адресу: http://victoria.biengi.ac.ru/cgi-bin/frameshift/index.cgi. Одновременно с банком данных был создан сервер для поиска потенциальных сдвигов рамки считывания в любых cds, который находится по

Таблица 1. Матрица триплетной периодичности $T(3,4)$, полученная для последовательности S=atcgtagctgacagtcga длиной 18 нуклеотидов

	1	2	3
a	2	1	2
t	0	2	2
c	1	1	2
g	3	2	0

адресу: http://victoria.biengi.ac.ru/fsfinder. На этом сервере можно ввести любую cds и получить координаты возможных сдвигов рамки считывания.

МАТЕМАТИЧЕСКИЕ МЕТОДЫ И АЛГОРИТМЫ

Общее описание метода. В данной работе мы рассматривали триплетную периодичность как статистическое свойство, позволяющее обнаружить сдвиги рамки в анализируемых генах. Триплетная периодичность задается матрицей $T(3,4)$, где признаками столбцов являются позиции оснований в кодонах (1, 2 или 3), а признаками строк являются нуклеотиды [32]. Матрица $T(3,4)$ хорошо учитывает отличие частот оснований в каждой позиции кодона от частот оснований во всей нуклеотидной последовательности. Рассмотрим последовательность S = atcgtagctgacagtcga длиной $N = 18$. Для заполнения матрицы создаем последовательность A, элементы которой $a(i) = i\text{mod}3 + + 1$ для всех i от 1 до N. Здесь mod обозначает остаток от деления числа i на 3. Тогда последовательность A = 123123123123123123. Матрица триплетной периодичности $T(3,4)$ заполняется как $t(s(i),a(i)) = t(s(i),a(i)) + 1$ для всех i от 1 до N. Результат показан в табл. 1. Сумма всех элементов матрицы равна N. Однако матрица T не учитывает корреляцию между соседними основаниями. Пусть последовательность S имеет четыре триплета: ATA, TAT, CGC, GCG. Пусть также они будут расположены с равной вероятностью и в произвольном порядке. В этом случае частоты оснований в каждом столбце матрицы T будут равны частотам оснований во всей последовательности S. Это означает, что найти корреляцию между последовательностями оснований ДНК при помощи матрицы T невозможно. Однако это можно сделать, если использовать матрицу $M(k,n)$. У этой матрицы признаками столбцов являются пары оснований, которые встречаются в позициях $i - 1$ и i последовательности S, i пробегает от 1 до N. В последовательности A в этом случае встречаются сочетания 12, 23 и 31. Пронумеруем столбцы матрицы M следующим образом: $k = 1$ соответствует сочетанию 12, $k = 2$ соответствует сочетанию 23, $k = 3$ соответствует сочетанию 31 в последовательности A. Таким образом, номер столбца k матрицы M совпадает со значением $a(i - 1)$. Номер строки матрицы M для каждой пары оснований рассчитаем как $n = s(i - 1) + (s(i) - 1)\times 4$. Матрица M заполняется как $m(a(i - 1), n)) = m(a(i - 1), n) + 1$ для i от 1 до N. В результате мы получаем матрицу размерности 3×16. В данной работе мы использовали перекодировку нуклеотидов в числа как: $1 = a$, $2 = t$, $3 = c$, $4 = g$. Матрица $M(k,n)$ учитывает как неоднородность использования нуклеотидов в матрице T, так и корреляцию соседних оснований в последовательности S. Введем следующие обозначения. Пусть S_0 – последовательность, которая была в cds до сдвига рамки считывания. Последовательность S образовалась из последовательности S_0 при сдвиге рамки считывания посредством делеции одного основания в центре последовательности. Для последовательности S_0 можно рассчитать матрицу $M_0(k,n)$, а для последовательности S – матрицу $M(k,n)$. Если бы матрица $M_0(k,n)$ была известна, то относительно несложно найти местоположение вставки или делеции в последовательности S. Это можно сделать, если построить глобальное выравнивание при помощи динамического программирования для последовательности S относительно матрицы $W_0(k,n)$. Матрица $W_0(k,n)$ есть позиционно-весовая матрица, которая создается с использованием частотной матрицы $M_0(k,n)$ по формуле (1) (см. ниже). По этой же формуле можно также рассчитать $W(k,n)$ для частотной матрицы $M(k,n)$.

Такое выравнивание можно сделать так, как это было выполнено в работе [33]. В результате мы рассчитываем максимальное значение функции сходства mF_0 последовательности S относительно матрицы $W_0(k,n)$. Однако нет сохраненной матрицы $M_0(k,n)$, мы не можем построить $W_0(k,n)$, и местоположение вставки или делеции оснований в последовательности S является неизвестным. Это также значит, что мы не можем рассчитать функцию сходства mF_0. Задача в таком случае и состоит в том, чтобы найти такую матрицу $W_2(k,n)$, которая была бы наиболее близка к матрице $W_0(k,n)$. Для нахождения матрицы $W_2(k,n)$ будем использовать генетический алгоритм, который был разработан нами в работе [31]. Начало генетического алгоритма состоит в том, что мы генерируем определенное множество WR случайных позиционно-весовых матриц размерностью 3 на 16. Эти матрицы являются «организмами» для генетического алгоритма, а функция сходства выполняет роль целевой функции. Затем мы применяем генетический алгоритм с целью

создания такой матрицы $W_2(k,n)$, которая будет максимизировать функцию сходства mF.

Пусть матрица $W_2(k,n)$ имеет максимальное значение функции сходства mF_1 при проведении глобального выравнивания последовательности S. При помощи генетического алгоритма мы ищем такую матрицу $W_2(k,n)$, которая имеет функцию сходства mF_1, как угодно близкую к значению mF_0. Это означает, что разница $(mF_0 - mF_1)$ стремится к нулю при стремлении числа итераций генетического алгоритма к бесконечности. Глобальное выравнивание также позволяет определить координаты потенциальных сдвигов рамки считывания в последовательности S. Блок-схема алгоритма показана на рис. 1. Рассмотрим более детально по пунктам этот алгоритм.

Расчет позиционно-весовых матриц и создание множества случайных матриц. Пусть последовательность S содержит cds. Для создания случайной матрицы $W(3,16)$ из множества случайных матриц WR мы случайно перемешивали нуклеотиды в последовательности S и создавали последовательность SR. Алгоритм случайного перемешивания был подробно описан ранее [31]. Затем при помощи последовательности SR мы заполняли матрицу $M(3,16)$, как это было описано в пункте «Общее описание метода». Затем мы рассчитывали матрицу $WR(3,16)$ по формуле:

$$WR(k,n) = \frac{M(k,n) - (N-1)p_2(n)}{\sqrt{(N-1)p_2(n)(1-p_2(n))}}. \qquad (1)$$

Здесь $p_2(n) = p(l)p(m)$, где $p(l)$ и $p(m)$ – вероятность встретить нуклеотид типа l или m в последовательности SR ($l \in$ {a,t,c,g}, $m \in$ {a,t,c,g}); $p(n) = q(n)/N$, $q(n)$ – количество нуклеотидов типа l в последовательности SR; N – длина последовательности SR.

При таком расчете мы учитываем как неоднородности в частотах нуклеотидов в каждой позиции кодона в последовательности SR, так и корреляцию соседних оснований. Это связано с тем, что рассчитывается ожидаемое число каждой из 16 пар соседних оснований для каждой позиции k матрицы $W(3,16)$. Последовательность S случайно перемешивалась 500 раз и всего было создано 500 матриц для множество WR.

Оптимизация матрицы $W_2(3,16)$ генетическим алгоритмом и динамическим программированием. Функция сходства F в динамическом программировании выступает как целевая функция для генетического алгоритма. Пусть мы имеем ген или cds в виде последовательности S. Длина последовательности равна N. Последовательность S выравнивалась относительно трансформированной

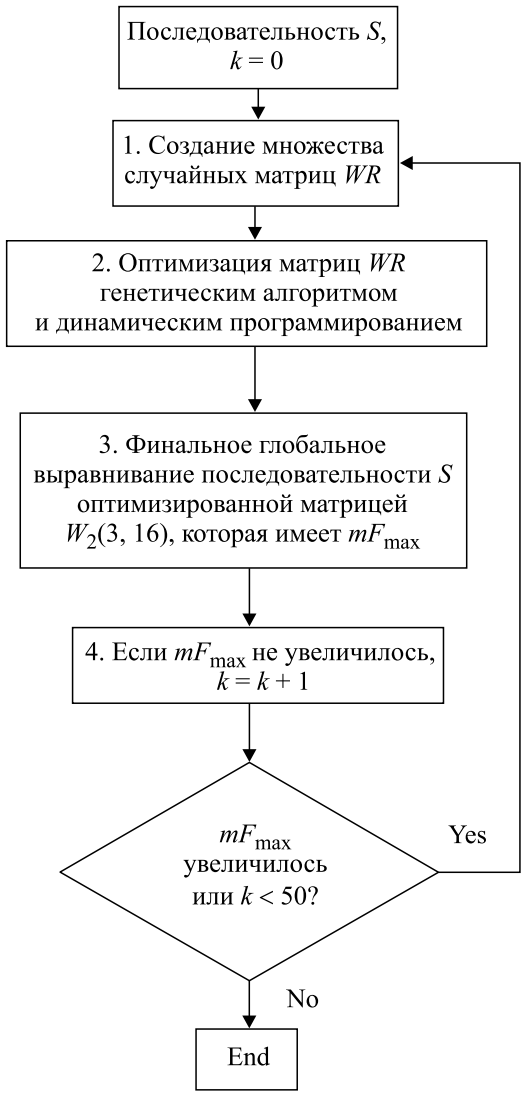

Рис. 1. Блок-схема алгоритма, применяемого для процедуры оптимизации случайных матриц $M(3,16)$ из множества MR с целью поиска матрицы с наибольшим mF_{max}.

матрицы $W^t(3,16)$. Суть трансформации состоит в том, чтобы сумма квадратов элементов матрицы была равна R^2. При трансформации поддерживалось постоянство константы K_d , что позволяет сохранять среднее значение элементов матрицы W^t в расчете на один нуклеотид последовательности S. Метод трансформации матрицы W_2 в матрицу W^t, выбор констант R^2 и K_d и генетический алгоритм подробно был рассмотрен нами в работе [31]. Для этого использовали следующую итеративную процедуру:

$F(i,j-4)$	$F(i,j-3)$	$F(i,j-2)$	$F(i,j-1)$	$F(i,j)$
$F(i-1,j-4)$	$F(i-1,j-3)$	$F(i-1,j-2)$	$F(i-1,j-1)$	$F(i-1,j)$
$F(i-2,j-4)$	$F(i-2,j-3)$	$F(i-2,j-2)$	$F(i-2,j-1)$	$F(i-2,j)$
$F(i-3,j-4)$	$F(i-3,j-3)$	$F(i-3,j-2)$	$F(i-3,j-1)$	$F(i-3,j)$
$F(i-4,j-4)$	$F(i-4,j-3)$	$F(i-4,j-2)$	$F(i-4,j-1)$	$F(i-4,j)$

Рис. 2. Иллюстрация для выбора индекса k при делеции двух оснований в последовательности S. В этом случае $k = s(j\text{-}3)$. (Это третий вариант выбора k в пункте «Оптимизация матрицы W2(3,16) генетическим алгоритмом и динамическим программированием»).

$$F(i,j) = \max\begin{Bmatrix} F(i-1,j-1) + W^t(a(i),n) \\ F(i,j-1) - d \\ F(i-1,j) - d \end{Bmatrix} \quad (2)$$

где i меняется от 1 до $(N + 2)$, а j меняется от 1 до N. Здесь $n = s(k) + (s(j) - 1)\times 4$, а значение $a(i)$ было определено в пункте «Общее описание метода». Введение параметра n связано с тем, что в матрице W и W^t учитываются пары оснований. Это означает, что мы должны определить предыдущее основание, которое мы уже включили в выравнивание. Поэтому индекс k рассчитывался по уже созданным переходам в матрице F в зависимости от того, какое предыдущее основание из последовательности S было включено в выравнивание. В зависимости от индекса k мы брали значение n. Есть три варианта для индекса k.

1. Если предыдущее основание из последовательности S (которое мы уже взяли в выравнивание) есть $s(j-1)$, то мы берем $k = j\text{-}1$ и $n = s(j\text{-}1) + (s(j)\text{-}1)\times 4$. Это соответствует переходам по ячейкам матрицы $F(i\text{-}2,j\text{-}2) \rightarrow F(i\text{-}1,j\text{-}1) \rightarrow F(i,j)$.

2. Если предыдущее основание в выравнивании есть $s(j\text{-}2)$, то мы берем $k = j\text{-}2$ и $n = s(j\text{-}2) + (s(j) - 1)\times 4$. Это соответствует переходам из ячейки $F(i\text{-}2,j\text{-}3) \rightarrow \rightarrow F(i\text{-}1,j\text{-}2) \rightarrow F(i\text{-}1,j\text{-}1) \rightarrow F(i,j)$. Эти переходы соответствует пропуску основания $s(j\text{-}1)$.

3. Если предыдущее основание есть $s(j\text{-}3)$, то мы берем $k = j\text{-}3$ и $n = s(j\text{-}3) + (s(j) - 1)\times 4$. Это соответствует переходам из ячейки $F(i\text{-}2,j\text{-}4) \rightarrow F(i\text{-}1, j\text{-}3) \rightarrow F(i\text{-}1,j\text{-}2) \rightarrow F(i\text{-}1,j\text{-}1) \rightarrow F(i,j)$.

На рис. 2 схематично показаны эти переходы. Они соответствуют пропуску основания $s(j\text{-}1)$ и $s(j\text{-}2)$ в выравнивании. Так как мы изучаем триплетную матрицу, то делеции больше чем два основания в последовательности S блокировались. Для них мы брали $d = 10000$, что исключает создание таких вставок или же делеций.

Одновременно с делециями в последовательности S могут происходить делеции одного или двух столбцов матрицы $W^t(a(i),n)$. Введение таких столбцов соответствует переходам $F(i\text{-}3,j\text{-}2) \rightarrow F(i\text{-}2, j\text{-}1) \rightarrow \mathrm{F}(i\text{-}1,j\text{-}1) \rightarrow F(i,j)$ и $F(i\text{-}4,j\text{-}2) \rightarrow F(i\text{-}3,j\text{-}1) \rightarrow F(i\text{-}2, j\text{-}1) \rightarrow F(i\text{-}1,j\text{-}1) \rightarrow F(i,j)$. Для этих переходов мы не учитывали корреляции между соседними основаниями. Вместо матрицы $W^t(a(i),n)$ для этих двух переходов мы использовали матрицу $E^t(x,n)$:

$$E^t(x,n) = 0,25\sum_{i=1,4} W_2^t(x, i + (n-1)\times 4) \quad (3)$$

Здесь уже n меняется от 1 до 4 и при использовании этой матрицы $n = s(j)$. Это означает, что при делециях столбцов корреляция соседних оснований ДНК не учитывается. Это вполне допустимо в том случае, когда число делеций или же вставок небольшое.

Нулевые строка и столбец матрицы F заполнялись отрицательными числами: $F(0,j)$ и $F(i,0)$ равны 0 для i от 1 до N+2, а j от 1 до N, а значения $F(0,0)$, $F(1,0)$, ..., $F(2,0)$ также равны 0. Матрица использовалась также для перехода из нулевого столбца в первый столбец матрицы F и из нулевой строки в первую строку матрицы F. Выбор цены за вставку или делецию d (d = 25,0) был сделан так, как мы это делали ранее [31]. Матрица обратных переходов заполнялась одновременно с матрицей F, как это обычно делается при поиске глобального выравнивания. Потом мы строили выравнивание последовательности S относительно матрицы W^t по матрице обратных переходов и определяли максимальное значение функции сходства mF в одной из трех ячейках матрицы $F(N,N)$, $F(N+1,N)$, $F(N+2,N)$.

Целевая функция для генетического алгоритма была выбрана как mF. «Организм» в генетическом алгоритме представляет собой матрицу $W(3,16)$. Генетический алгоритм был рассмотрен ранее в работе [31], там можно найти и подробности его применения. Множество случайных матриц WR (пункт «Расчет позиционно-весовых матриц и создание множества случайных матриц WR») было сформировано, оно содержало 500 матриц, каждая размерностью 3 на 16. Это множество участвовало в работе генетического алгоритма. В результате работы генетического алгоритма определялась одна матрица из множества WR, которая имеет максимальное значение mF. Назовем его $mF_{\max}$. Окончание генетического алгоритма происходило после того, как значение $mF_{\max}$ переставало увеличиваться на протяжении

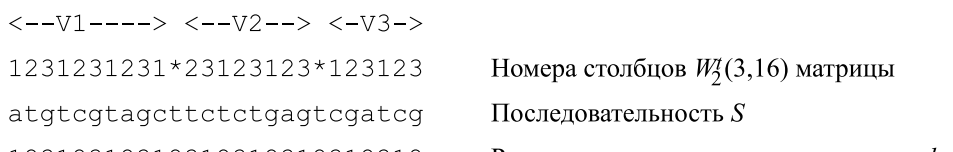

Рис. 3. Схема разделения $mF_{\max}$ на $V1$, $V2$ и $V3$ (см. пункт «Расчет количественной меры для поиска сдвига фазы триплетной периодичности»).

50 итераций. В среднем до этого момента требуется около $9 \cdot 10^3$ итераций.

В итоге в конце работы генетического алгоритма мы имеем значение $mF_{\max}$ и соответствующую матрицу $W(3,16)$. Назовем ее $W_{\max}(3,16)$. Кроме матрицы мы имеем построенное выравнивание последовательности S относительно столбцов матрицы $W_{\max}(3,16)$. Именно это выравнивание позволяет нам определить координаты потенциальных сдвигов рамки считывания в последовательности S.

Расчет количественной меры для поиска сдвига фазы триплетной периодичности. Значение $mF_{\max}$, которое характеризует триплетную периодичность в последовательности S, само по себе мало информативно. Интерес представляет статистическая значимость найденных сдвигов фазы триплетной периодичности. Вставки или делеции символов в выравнивании последовательности S относительно матрицы $W_{\max}(3,16)$ указывают на координаты сдвиги фазы. Значение $mF_{\max}$ было разделено на три части для расчета количественной меры сдвига фазы триплетной периодичности. Участки последовательности S, где позиции столбцов матрицы и рамка считывания в последовательности S совпадают между собой, принадлежат первой части. На рис. 3 этот район обозначен как $V1$. Районы последовательности S, где имеются совпадения $1 \to 2$, $2 \to 3$, $3 \to 1$ принадлежат второй части (рис. 3, $V2$). Совпадения вида $1 \to 3$, $2 \to 1$, $3 \to 2$ принадлежат к третьей части (рис. 3, $V3$). Сумма $V1 + V2 + V3\text{-}kd$ равна $mF_{\max}$, где k – число вставок или делеций, d – цена за вставку или делецию из формулы 2. Матрица $W_{\max}(3,16)$ не имеет связи с рамкой считывания. Поэтому циклические перестановки матрицы $W_{\max}(3,16)$ производились для того, чтобы $V1 \geqslant V2$ и $V1 \geqslant V3$. Мы выбрали сумму $V2 + V3$ как количественную меру характеризующую статистическую значимость сдвигов фазы триплетной периодичности. Для каждой анализируемой cds было рассчитано значение $V = V2 + V3$. Пороговое значение для $V = V0$ было рассчитано. Если $V < V0$, то сдвиг рамки считывания в cds отсутствует. Свое пороговое $V0$ выбиралось для каждой cds. Для расчета V0 мы создали множество последовательностей SR. Последовательности генов имеют различную длину и различную триплетную периодичность [18], поэтому для каждого гена необходим свой расчет $V0$. Во множестве SR содержится 10^3 последовательностей S, кодоны в которых были случайно перемешаны. Таким перемешиванием уничтожаются все возможные сдвиги фазы триплетной периодичности. Мы выбирали значение $V0$ для каждого гена, при котором было не более 20 последовательностей во множестве SR, для которых $V > V0$. Только такие гены имеют статистически значимые вставки или делеции. Сдвиги фазы триплетной периодичности отсутствовали в генах, если $V2 + V3 = 0$.

РЕЗУЛЬТАТЫ И ДИСКУССИЯ

Геномы, которые были изучены. Всего было изучено 76 эукариотических геномов из базы данных Ensembl (последовательности cds для каждого генома были взяты из ftp://ftp.ensembl.org/pub/release-86/embl/). Из рис. 4 видно, что примерно 23% cds в среднем из каждого эукариотического организма содержат сдвиги рамки считывания. Созданный банк данных находится по адресу: http://victoria.biengi.ac.ru/cgi-bin/frameshift/index.cgi. Первая страница базы данных показана на рис. 5. Из этого рисунка видно, что пользователю предоставляется возможность выбрать геном, для которого он хочет найти cds с потенциальным сдвигом рамки считывания. Уже в выбранном организме предоставляется возможность выбрать cds либо общим списком, либо произвести поиск нужной cds по символу гена, либо же по идентификатору транскрипта. Если транскрипт выбран, то пользователю предоставляется следующая информация: 1) матрица $W_{\max}(3,16)$; 2) координата найденных сдвигов фазы; 3) выравнивание cds относительно матрицы $W_{\max}(3,16)$.

Рассмотрим пример этой информации для транскрипта ENST00000583951 из генома *Dasypus novemcinctus* (девятипоясный броненосец). Это транскрипт гена (AKAP10), который кодирует A-kinase anchoring protein 10 [34]. В этом транскрипте была совершена вставка нуклеотида *c* в 71-й позиции. В

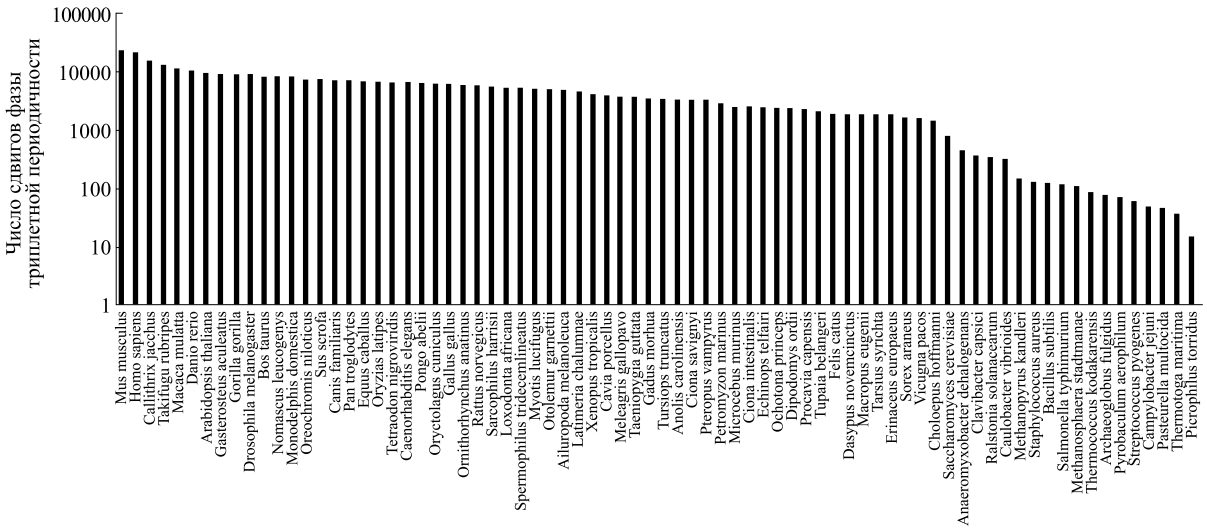

Рис. 4. По оси абсцисс показаны геномы эукариотических организмов, которые были проанализированы и которые содержатся в базе данных. По оси ординат показано число сдвигов фазы триплетной периодичности, которое было найдено в cds в каждом эукариотическом организме.

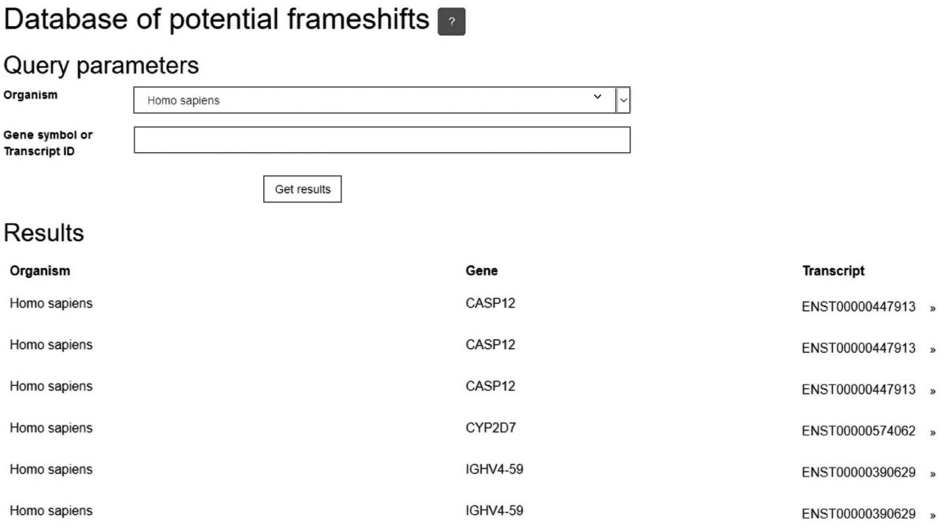

Рис. 5. Первая страница базы данных данных потенциальных сдвигов рамки считывания в кодирующих последовательностях. База данных расположена по адресу: http://victoria.biengi.ac.ru/cgi-bin/frameshift/index.cgi.

результате этого произошел сдвиг фазы триплетной периодичности после 71-й позиции. Матрица $W_{\max}(l,n)$ ($l = 1,2,3$, $n = 1,2,...,16$) показана в табл. 2. Для удобства восприятия она была переведена в трехмерную матрицу $W^3_{\max}(i,j,k)$. Индекс i в этой матрице соответствует индексу l матрицы $W_{\max}(l,n)$. Индекс j матрицы получается как $j = n - 4\times\text{int}((n\text{-}1)/4)$, где $n = 1,2,...,16$. Индекс k рассчитывается как: $k = 4\times\text{int}((n\text{-}1)/4)$. Индекс j и k показывают пару оснований с номером $n = j + (k\text{-}1)\times 4$ (см. пункты «Общее описание метода» и «Оптимизация матрицы W2(3,16) генетическим алгоритмом и динамическим программированием»). Из табл. 2 видно, что веса пар могут существенно отличаться. Например, вес пары TG для $i = 1$ (т.е. в первой фазе) составляет 11,6. Это означает, что вес пары TG в том случае, если T будет в третьей фазе, а G будет в первой фазе (см. пункт «Общее описание метода»), будет равен 11,6. Присутствуют в матрице также и большие отрицательные значения. Например для $i = 2$ пара TA имеет вес, равный –6,7. Это означает, что вес пары TA будет равен –6,7 в том случае, если T будет в первой фазе, а A будет во второй фазе. Таким образом, мы

Таблица 2. Матрица $W_{max}(3,16)$, полученная для поиска возможных сдвигов рамки считывания последовательности транскрипта ENST00000583951

		A	T	C	G
1	A	−3,4	0,7	−2,4	2,4
1	T	−3,8	−2,3	−0,6	11,6
1	C	−2,0	0,5	1,3	−4,5
1	G	−2,2	−1,9	−0,2	−0,9
2	A	−1,7	−0,2	−3,5	−1,1
2	T	−6,7	−0,2	6,3	−3,0
2	C	0,1	−2,9	2,5	−3,5
2	G	5,7	−1,0	0,9	0,0
3	A	1,8	−0,2	−1,4	1,5
3	T	−4,8	−1,3	0,6	−0,1
3	C	−3,0	13,0	−1,2	−3,5
3	G	−3,1	−2,9	−1,2	−0,9

учитываем не только отличие частот по позициям периода от средних частот оснований в последовательности S, но также учитываем корреляцию соседних оснований. Полученное с помощью матрицы $W^3_{max}(i,j,k)$ выравнивание cds для транскрипта ENST00000583951 показано в табл. 3.

На рис. 6 показано распределение числа cds по суммарному числу потенциальных сдвигов рамки считывания. Видно, что наибольшее число cds содержат один или два потенциальных сдвига рамки считывания. Количество cds, которое содержит более пяти сдвигов рамки считывания, является незначительным.

Оценка числа ошибок первого и второго рода. Для определения количества ошибок первого и второго рода мы взяли все кодирующие последовательности из генома *A. thaliana*, число которых составляет 48322 cds. Затем из этих последовательностей множество последовательностей *RN* было создано путем случайного перемешивания кодонов. В этом множестве заведомо не будет сдвигов фазы триплетной периодичности при сохранении статистических свойств исходной кодирующей последовательности. Множество *RN* позволяет оценить число ошибок первого рода (true positive). Мы применили разработанный нами подход к анализу множества *RN* и оказалось, что в этих последовательностях можно найти 1098 последовательностей со сдвигом фазы триплетной периодичности, где $V2 + V3$ было больше порогового уровня. В них содержится 1549 сдвигов рамки считывания. Если учесть, что в геноме *A. thaliana* нашим методом было найдено 9930 cds, то уровень ошибок первого рода составляет примерно 15%. Мы проделали такое изучение для всех проанализированных геномов и максимальное значение ошибок первого рода не превышало 23%.

Интересно также определить число ошибок второго рода и мощность метода. Для этого мы создавали множество последовательностей *RD* из кодирующих последовательностей из генома *A. thaliana* с длиной более 500 нуклеотидов. Кодоны в этих последовательностях были случайно перемешаны. Затем в случайное положение каждой последовательности не ближе 100 нуклеотидов от начала и конца каждой последовательности вносилась случайная делеция одного основания. Эти по-

Таблица 3. Выравнивание cds для транскрипта ENST00000583951 относительно матрицы $W^3_{max}(i,j,k)$. Вставка одного нуклеотида в позиции 71 отмечена звездочкой

```
0001 CTGGCTCTGTTGGCCCTCCTGATGAGTCTCACCCAGGGAGTTCTGACAGCTCTGCGTCTC
0001 23123123123123123123123123123123123123123123123123123123123

0061 AGGAAGATGACATTTGGAAGAGTCAGTGACTTGGGGCAATTCATCCGAGAATCTGAGCCT
0061 2312312312*312312312312312312312312312312312312312312312312

0121 GAACCTGATGTAAGGAAATCAAAAGGATCCATGTTCTCACAAGCTATGAAGAAATGGGTG
0120 123123123123123123123123123123123123123123123123123123123123

0181 CAAGGAAATACTGATGAGGCCCAGGAAGAGCTAGCTTGGAAGATTGCTAAAATG
0180 12312312312312312312312312312312312312312312312312312312
```

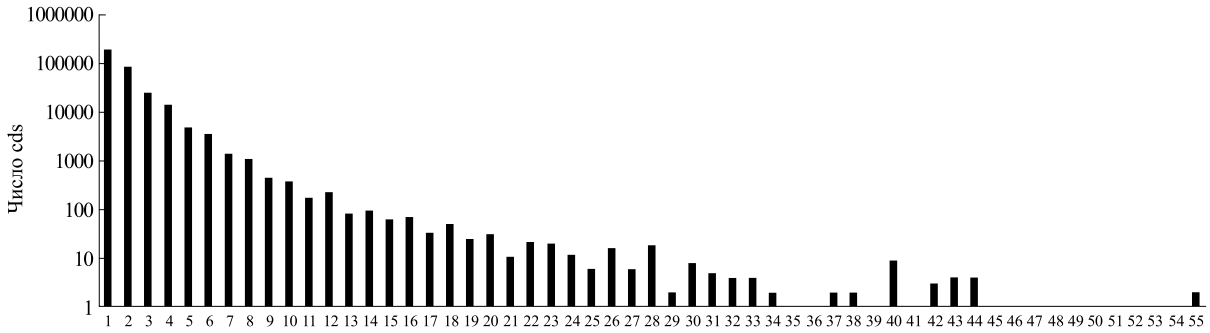

Рис. 6. Распределение числа cds, содержащихся в базе данных потенциальных сдвигов рамки считывания, по числу сдвигов триплетной периодичности.

следовательности были проанализированы с помощью разработанного алгоритма, и в итоге получены результаты, представленные в табл. 4. Из этой таблицы видно, что всего удалось выявить 28357 последовательностей во множестве *RD*, для которых ($V2 + V3) \geqslant V0$ и положение находится в интервале ±50 от искусственной делеции из 40621 последовательностей со сдвигами. Это показывает, что уровень ошибок второго рода составляет 30%, т.е. мощность метода составляет 70%. Также метод достаточно точно предсказывает местоположение сдвигов, так как только у 1128 последовательностей были найдены сдвиги вне района ±50 от точки делеции. Также следует отметить, что метод не создает значительного количества случайных сдвигов, так как на 29485 последовательностей со статистически значимыми сдвигами (28357 + 1128) приходится 29888 сдвигов, т.е. 403 сдвига обусловлены чисто случайными факторами.

Сравнительный анализ полученных результатов с более ранними публикациями. Интересно сравнить полученные нами результаты с полученными ранее результатами по поиску сдвигов рамки считывания в cds из эукариотических геномов. В первую очередь это относится к геномам, результаты изучения которых, накопленные в GeneTack database [22], были получены с использованием программы GeneTack-GM. Данная программа представляет собой объединение программ GeneMark (для выделения кодирующих последовательностей в геноме) и Genetack (для поиска потенциальных сдвигов рамки считывания) [35]. Программа Genetack ищет случаи потенциальных сдвигов рамки, которые привели к разбиению одной гипотетической кодирующей последовательности на две независимых (в современных базах данных они обычно представлены двумя отдельными генами). Результатом работы программы GeneTack-GM являются предсказанные координаты гена (обычно гипотетического гена) и координата сдвига внутри этого гена.

Всего в геноме *A. thaliana* программой GeneTack-GM [22] было найдено в mRNA 2067 потенциальных сдвигов рамки считывания, тогда как нам удалось обнаружить 14951 (табл. 5). Следует отметить, что мы в данной работе анализировали только cds, тогда как в GeneTack database содержатся данные для mRNA, которые также имеют некодирующие последовательности. Поэтому мы разделили найденные 2067 сдвигов рамки считывания на три группы. К первой группе отнесли потенциальные сдвиги рамки считывания, которые находятся только внутри cds не ближе 50 nt от начала и конца кодирующего участка. Во второй группе сдвиг содержится на расстоянии не большем 50 nt от концов cds, а в третьей группе он приходится на некодирующие районы mRNA. К первой группе относится 485, ко второй группе относится 710, а в третью группу входит 872 сдвига рамки считывания, найденные программой GeneTack-GM. Если учесть, что нам удалось вы-

Таблица 4. Поиск сдвигов фазы триплетной периодичности во множестве RD

Название организма	Общее число последовательностей в множестве *RD*	Число последовательностей с $V2 + V3 \neq 0$	Число последовательностей с $V2 + V3 \geqslant V0$		Общее число сдвигов
			внутри ±50	вне ±50	
A. thaliana	40612	31203	28357	1128	29888
Anaeromyxobacter dehalogenans	3460	3458	2975	68	3668

Таблица 5. Число cds содержащих потенциальные сдвиги рамки считывания в шести исследованных эукариотических геномах

Название организма	Число потенциальных framesfift mutations	Число cds с потенциальными framesfift mutations	Число потенциальных framesfift mutations из работы [22]
A. thaliana	14954	9930	2067
C. elegans	10411	5941	611
D. melanogaster	31873	8833	2616
H. sapiens	20795	13285	7395
R. norvegicus	9811	5768	703
X. tropicalis	6518	4228	529

явить 14951 потенциальный сдвиг рамки считывания из генома *A. thaliana*, то разработанный нами алгоритм примерно в семь раз эффективнее, чем программа GeneTack-GM. Если сравнивать результаты, относящиеся к первой группе, то такое различие будет еще больше, так как более 70% обнаруженных нами сдвигов относится к первой группе. Естественно, что такое сравнение корректно проводить только с первой группой. Похожие результаты также были получены для некоторых других эукариотических геномов (табл. 5), где мы обнаруживаем в несколько раз больше потенциальных сдвигов рамки считывания.

Такое различие в результатах может быть связано с тем, что метод HMM настраивается по множеству выбранных mRNA [22] и используются статистические свойства усредненные по множеству последовательностей и эти свойства фиксированы. Это усреднение как раз и может приводить к тому, что многочисленные сдвиги фазы будут пропускаться. В разработанном нами алгоритме такого усреднения не происходит, метод настраивается на триплетную периодичность, которая существует в каждом гене или cds. Это означает что разработанный нами математические метод персонально для каждой анализируемой cds или гена находит оптимальную корреляционную матрицу с учетом возможности вставок или делеций нуклеотидов. Финальное выравнивание изучаемой последовательности против полученной оптимизационной матрицы дает выравнивание последовательности и координаты потенциальных сдвигов рамки считывания. Важно еще отметить, что для поиска потенциальных сдвигов рамки считывания нашим методом требуются только последовательности гена или cds и никакая другая информация больше не нужна. В этом состоит основное улучшение метода по сравнению с использованием метода HMM для определения сдвигов рамки считывания.

Сдвиги рамки считывания могут принимать участие в образовании новых генов. Изучение эволюции генов привлекает внимание исследователей много лет и исследования в этой области значительно ускорились после определения последовательностей многих прокариотических и эукариотических генов. В настоящее время считается, что новые гены происходят путем дупликации существующих генов [36]. Поэтому большое количество генов объединены в семейства на основе подобия их аминокислотных или нуклеотидных последовательностей [37]. Также известно, что процессы склейки генов (gene fusion) [38], перетасовка экзонов (exon shuffling) [39], альтернативный сплайсинг (alternative splicing) [40] и горизонтальный перенос генов (lateral gene transfer) [41] являются основными механизмами по дивергенции генов и создании разнообразия генов в геноме и соответствующих им белков. Однако такими механизмами трудно быстро создать принципиально новую последовательность, а сдвиг рамки считывания может это сделать быстро. В связи с этим было предположено ранее, что мутации типа сдвиг рамки считывания могут играть существенную роль в создании новых генов [7,9]. В этих работах высказано предположение, что если белок по тем или иным причинам выведен из-под давления отбора, то сдвиг рамки считывания может закрепиться в гене. Именно это явление мы, вероятно, наблюдаем в данной работе, так как более 20% cds из различных геномов эукариот содержат потенциальные сдвиги рамки считывания. Остается только вопрос: как после сдвига рамки считывания образуется белковая последовательность, имеющая какой-либо биологический смысл? Можно предполагать, что генетический код хорошо адаптирован к таким изменениям, и он позволяет в результате сдвига рамки считывания получать биологически осмысленные последовательности.

СПИСОК ЛИТЕРАТУРЫ

1. J. D. Watson, et al., Molecular biology of the gene (Pearson Education, Inc. 2013).
2. Y. Ogura, et al., Nature **211**, 603 (2001).
3. M. C. Iannuzzi, et al., Am. J. Hum. Genet. **48** (2), 227 (1991).
4. W. K. Chung, C. Kitner, and B. J. Maron, Cardiol. Young **21**, 345 (2011).
5. X. Xu, et al., Gene 519, 343 (2013).
6. J. L. Wood and J. Chen, in *DNA Repair, Genetic Instability, and Cancer* (World Scientific Publishing Co. Pte. Ltd., 2007), pp. 1–22.
7. K. Okamura, et al., Genomics **88** (6), 690 (2006).
8. S. M. Berget, J. Biol. Chem., **270** (6), 2411 (1995).
9. J. Raes and Y. Van De Peer, Trends Genet. **21** (8), 428 (2005).
10. S. L. Sheetlin, et al., Bioinformatics **30** (24), 3575 (2014).
11. Y. Zhang and Y. Sun, BMC Bioinformatics **12**, 198 (2011).
12. N. Du and Y. Sun, Bioinformatics **32** (17), i529 (2016).
13. R. Ketteler, Frontiers in Genetics **19** (3), 242 (2012).
14. A. A. Mironov, P. S. Novichkov, and M. S. Gelfand, Bioinformatics **17** (1), 13 (2001).
15. M. Gîrdea, L. Noé, and G. Kucherov, Algorithms Mol. Biol. **5** (1), 6 (2010).
16. T. Schiex, et al., Nucl. Acids Res. 31 (13), 3638 (2003).
17. I. Antonov and M. Borodovsky, J. Bioinform. Comput. Biol. **8** (3), 535 (2010).
18. F. Frenkel and E. V. Korotkov, DNA Research 16 (2), 105 (2009).
19. J. Gouzy, S. Carrere, and T. Schiex, Bioinformatics **25** (5), 670 (2009).
20. M. Rho, H. Tang, and Y. Ye, Nucl. Acids Res. **38** (20), e191 (2010).
21. 21. Y. Zhang and Y. Sun, BMC Bioinformatics **12** (1), 198 (2011).
22. 22. I. Antonov, P. Baranov, and M. Borodovsky, Nucl. Acids Res. 41, D152 (2013).
23. R. K. Azad and M. Borodovsky, Brief. Bioinform. **5** (2),118 (2004).
24. G. Gutiérrez, J. L. Oliver, and A. Marín, J. Theor. Biol. **167** (4), 413 (1994).
25. V. R. Chechetkin and A. Yu. Turygin, J. Theor. Biol. **175** (4), 477 (1995).
26. J. Gao, et al., J. Biomed. Biotechnol. **2005** (2), 139 (2005).
27. C. Yin and S. S.-T. Yau, J. Theor. Biol. **247** (4), 687 (2007).
28. H. Masoom, et al., in *Proc. 2006 IEEE Symp. on Comput. Intelligence in Bioinformatics and Comput. Biology (CIBCB'06)* (2006). DOI: 10.1109/CIBCB.2006.330971.
29. L. Wang and L. D. Stein, BMC Bioinformatics **11** (1), 550 (2010).
30. M. A. Korotkova, N. A. Kudryashov, and E. V. Korotkov, Genomics. Proteomics. Bioinformatics **9** (4), 158 (2011).
31. V. Pugacheva, A. Korotkov, and E. Korotkov, Stat. Appl. Genet. Mol. Biol. **15** (5), 381 (2016).
32. F. E. Frenkel and E. V. Korotkov, Gene **421** (1–2), 52 (2008).
33. A. A. Laskin, Mol. Biol. (Moscow) **37** (4), 663 (2003).
34. L. J. Huang, Proc. Natl. Acad. Sci. USA **94** (21), 11184 (1997).
35. I. Antonov, et al., Nucl. Acids Res. **41** (13), 6514 (2013).
36. S. Ohno, *Evolution by Gene Duplication* (Springer Berlin Heidelberg, 1970).
37. E. V. Koonin, Rev. Genet. **39** (1), 309 (2005).
38. T. M. Thomson, et al., Genome Res. **10** (11), 1743 (2000).
39. W. Gilbert, Nature **271** (5645), 501 (1978).
40. M. Hiller, et al., Genome Biol. **6** (7), R58 (2005).
41. H. Ochman, Curr. Opin. Genet. Dev. **11** (6), 616 (2001).

A Database of Potential Reading Frame Shifts in Coding Sequences from Different Eukaryotic

Yu.M. Suvorova*, V.M. Pugacheva*, and E.V. Korotkov* **

**Institute of Bioengineering, Research Center of Biotechnology of the Russian Academy of Sciences, Leninsky prosp. 33/2, Moscow, 119071 Russia*

***National Research Nuclear University (Moscow Engineering Physics Institute), Kashirskoye Shosse 31, Moscow, 115409 Russia*

In this paper we offer a new databank containing potential reading frame shifts in coding sequences. A new mathematical method based on the use of a genetic algorithm and dynamic programming was used to search for potential shifts in the reading frame. The data bank includes coordinates of potential reading frame shifts for coding sequences of 76 eukaryotic genomes from Ensembl genomic browser, version 86. The database is located at: http://victoria.biengi.ac.ru/cgi-bin/frameshift/index.cgi. Of all the genomes analyzed, approximately 23% of coding sequences have a reading frame shift. Errors of the first and second kind are about 11% and 30%. Simultaneously with the data bank, a Web-server has been created to search for potential shifts in the reading frame, which is located at: http://victoria.biengi.ac.ru/fsfinder. The server can be used to search for potential shifts of the reading frame in newly defined coding sequences.

Keywords: reading frame shift, dynamic programming, position-weight matrix, coding sequence