УДК 577.21

С. В. Траньков[1], *В. А. Яворский*[1], *М. Бородовский*[2]

[1]Московский физико-технический институт (государственный университет)
[2]Georgia Institute of Technology, Atlanta

Анализ эффективности использования семейства алгоритмов GeneMark при аннотации геномов

В статье проводится анализ семейства алгоритмов GeneMark, которые используются для автоматизированной аннотации генов в новых прокариотических (в том числе всех бактериальных) геномах без использования сравнения (выравнивания) с известными генами и белками. Показано, что в своем классе алгоритм GeneMark является лучшим; дальнейшее улучшение этого алгоритма возможно более полным учетом регуляторных участков, располагающихся вблизи стартов трансляции, а также учетом локальных вариаций GC композиции генома, часто связанных с горизонтальным переносом генов из других микробов.

Ключевые слова: GeneMark, аннотация генома, бактериальный геном, семейство алгоритмов, скрытые марковские модели, алгоритм Витерби.

Введение

Одним из основных этапов обработки отсеквенированных геномов является их структурная аннотация, т.е. проведение процесса маркировки генов, регуляторных участков и других объектов в последовательности ДНК. В начале «геномной эры» в молекулярной биологии аннотация геномов производилась вручную путем выравнивания нуклеотидных последовательностей. Однако такой подход не позволял корректно проаннотировать геномы видов, которые не являются близкородственными, поскольку не всегда находил новые гены.

После активного внедрения в 1980-х годах технологий автоматического секвенирования начался лавинообразный рост размеров геномных банков. Например, если в 2001 году в принятой за международный стандарт базе данных GenBank находилась информация о 15 миллионах нуклеотидных последовательностях, то к концу 2011 года их количество достигло 146 миллионов. При этом количество проаннотированных бактериальных геномов достигло 1826. Понятно, что обработка такого числа геномных последовательностей невозможна без автоматизации.

Алгоритмические методы анализа последовательностей ДНК не только ускоряют исследовательский процесс, но и позволяют делать нетривиальные предсказания о функциональных свойствах геномов.

Семейство алгоритмов GeneMark берет свое начало из исследований, проводившихся еще в середине 1980-х годов в Институте молекулярной генетики Академии наук СССР [1–3]. Эти исходные работы послужили основой для дальнейшего развития и создания алгоритма и программного обеспечения под названием GeneMark в 1993 году [4, 5]. Интересно, что этот алгоритм, появившийся еще до внедрения в биоинформатику скрытых марковских моделей (HMM — Hidden Markov Model), по существу являлся аналогом «posterior decoding» алгоритма, одного из главных HMM-алгоритмов, для неявно определенной HMM, в работе [4]. Появившиеся позже программы, такие как Glimmer [6] и ORPHEUS [7], использовали различные эмпирические закономерности и корреляции, позволяющие быстро прийти к решению задачи. Однако прямое использование HMM, продемонстрировавших значительные успехи в решении задач распознавания речи и текста [8], привело к более точным алгоритмам. Эти алгоритмы были реализованы в программах ECOPARSE [9] (1994 год) и GeneMark.hmm [10] (1998 год).

В отличие от ECOPARSE, где HMM были применены только к анализу генома *E.coli*, семейство алгоритмов GeneMark успешно используется для аннотации самых разных геномов — бактериальных, вирусных, эукариотических, метагеномов и т.д. [10–14]. Оказалось, что реализованный в GeneMark.hmm алгоритм точно предсказывает 83–94% генов (т.е. предсказания 5' и 3'-концов совпадают с аннотацией GenBank).

Марковские модели

В результате секвенирования генома получается одномерная цепочка нуклеотидов, для которой задача аннотации состоит в нахождении генов с определением их границ. Набор нуклеотидов универсален практически для всех биологических видов и может быть представлен в виде последовательности символов из четырехбуквенного алфавита. Таким образом, решение задачи аннотации геномной последовательности сводится к задаче распознавания текстовых образов.

Как найти эти структуры в длинной неразмеченной последовательности? Можно подсчитывать *логарифм отношения правдоподобия* (отношения вероятности того, что нуклеотид принадлежит искомой структуре, к вероятности того, что не принадлежит) для окна, например, длины 100 нуклеотидов вокруг каждого нуклеотида в последовательности, и наносить эти значения на график. Тогда можно ожидать, что структура будет выделяться положительными значениями логарифма [1–3]. Однако этого недостаточно, если в действительности искомые структуры имеют четкие границы и различаются по длине. Более эффективным подходом было бы построить модель, включающую в себя две вспомогательные модели – для «островов» структуры и «моря» вокруг них [8].

В программе GeneMark.hmm [10] для моделирования нуклеотидной последовательности используются скрытые марковские модели [15], которые ранее были успешно применены к решению задач распознавания речи и текста. Марковские модели, описывающие кодирующую и некодирующую области ДНК, используют алфавит $\aleph = \{\mathrm{A}, \mathrm{T}, \mathrm{G}, \mathrm{C}\}$, символы которого обозначают соответствующие нуклеотиды.

Точности модели Маркова 1-го порядка, когда вероятность символа зависит только от предыдущего нуклеотида, оказалось недостаточно для удовлетворительного описания кодирующей области генома [1]. Поэтому возникла необходимость повысить порядок используемой цепи путем учета нескольких нуклеотидов перед рассматриваемым.

Известно также, что позиционные частоты нуклеотидов в трехбуквенных кодонах имеют довольно разные статистики. Поэтому для построения более точной марковской модели генов в GeneMark используются три различные цепи Маркова для того, чтобы смоделировать кодирующие области [2, 4].

В программе GeneMark.hmm [10] геном моделируется в рамках «обобщенной скрытой марковской модели». В ней нуклеотидная последовательность представляется как реализация марковского процесса со скрытыми состояниями a_i, генерирующими (испускающими) фрагменты последовательности ДНК длиной d_i нуклеотидов: $\sum_i d_i = L$. Вероятность наблюдать некую последовательность нуклеотидов как составную часть некодирующей области генома рассчитывалась, используя однородную марковскую модель нулевого порядка. Вероятность наблюдать некую последовательность нуклеотидов как составную часть кодирующей области рассчитывалась, используя трехпериодическую марковскую модель второго или пятого порядков. Вероятность того, что скрытое состояние a сгенерирует фрагмент длиной d, определялась по аналитической аппроксимации распределения длин кодирующих и некодирующих участков для генома *E.coli* [10].

Для того чтобы иметь дело одновременно с двумя нитями ДНК, было введено девять скрытых состояний. Переходы между ними разрешены по определенной схеме, учитывающей генетическую структуру бактериального генома (рис. 1).

При этом кодирующие состояния разделены на два кластера — «типичные» (к которому принадлежит большинство генов *E.coli*) и «атипичные» (как полагают, результаты горизонтального переноса генов). Начальные вероятности четырех кодирующих и одного

некодирующего состояний были положены равными 0,2; вероятности старт-кодонов были определены согласно их статистике в геноме *E.coli*: $P\,(\mathrm{ATG}) = 0,905$; $P\,(\mathrm{GTG}) = 0,090$; $P\,(\mathrm{TTG}) = 0,005$. Вероятности переходов из некодирующего состояния в типичный (атипичный) ген положены равными 0,85 (0,15) [10, 19].

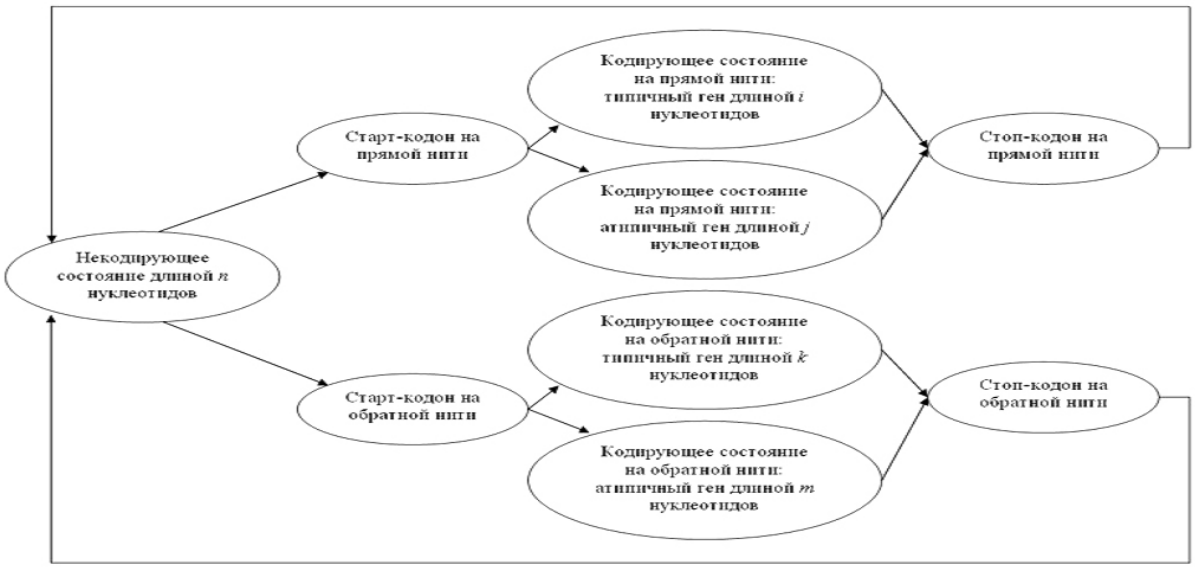

Рис. 1. Скрытая марковская модель прокариотического генома, использующаяся в программе GeneMark.hmm. Скрытые состояния обозначены овалами, а разрешенные переходы между ними — стрелками.

Параметры описанной модели определялись либо путем контролируемого обучения, либо путем неконтролируемого обучения по Витерби в программе GeneMarkS [11]. Аннотирование генома производилось путем дешифровки последовательности нуклеотидов по алгоритму Витерби [16].

Постобработка результатов дешифрования

Для улучшения качества распознавания генов производится дополнительная обработка полученных данных. Предсказанные алгоритмом Витерби гены должны быть разделены как минимум одним нуклеотидом, поэтому положение одного из двух перекрывающихся генов невозможно определить точно. Поскольку в бактериальных геномах перекрывание генов — достаточно частое событие, к проделанной аннотации дополнительно применяется процедура нахождения сайтов связывания рибосом (RBS) следующим образом. Для каждого предсказанного гена RBS искались на интервале от —19 до —4 нуклеотидов для каждого альтернативного старт-кодона между старт-кодоном, предсказанным алгоритмом Витерби, и старт-кодоном длиннейшей открытой рамки считывания для предсказанного гена. Изначально предсказанный сайт инициации трансляции переопределялся, если вес одного из кандидатов RBS оказывался выше определенного порога.

Вероятностная модель для RBS строилась следующим образом. Для каждого из 325 генов *E.coli* с известными RBS были собраны короткие последовательности, лежащие в интервале от —19 до —4 нуклеотидов. Эти 325 коротких последовательностей были подвергнуты множественному выравниванию с целью нахождения короткого отобранного в эволюции мотива, что позволило получить матрицу позиционных вероятностей нуклеотидов в модели RBS.

Эвристический поиск начального приближения

Основное время работы программы, реализующей поиск генов в нуклеотидной последовательности на основе HMM, занимает оценка значений ее параметров. Поэтому выбор «правильного» начального приближения для алгоритма обучения по Витерби помогает

значительно ускорить поиск [17].

В программе GeneMark.hmm для этого реализован эвристический метод предварительной оценки параметров HMM, используя минимальное количество информации о последовательности. В его основу положены найденные корреляции между частотами нуклеотидов в геноме и в триплете, а также зависимости между частотой аминокислот и содержанием GC в геноме [12].

При анализе 17 полных бактериальных геномов были найдены корреляции между позиционной и глобальной частотами встречаемости нуклеотида и между частотами встречаемости некоторых аминокислот и содержанием GC в геноме. Поскольку глобальные частоты нуклеотидов однозначно определяются содержанием GC в геноме, была проведена дополнительная проверка найденных корреляций на большем (319) количестве геномов. Зависимости между позиционной частотой нуклеотидов и содержанием GC в геноме хорошо аппроксимируются прямыми линиями по стандартному регрессионному методу [14].

Как было показано в работе [12], частоты только десяти из двадцати аминокислот обнаружили значимые изменения в зависимости от содержания GC в геноме (от 28.6% у *B.burgdorferi* до 65.6% у *M.tuberculosis*). Четыре аминокислоты из этих десяти кодируются триплетами вида SSN (S означает C или G и описывает *сильное спаривание*): аланин (A), глицин (G), пролин (P) и аргинин (R); их частоты увеличиваются с увеличением содержания GC. Хотя аргинин кодируется не только триплетами вида GCN, но также AGA и AGG, авторы все равно относят его к SSN-типу, поскольку четыре из шести кодирующих его триплетов SSN-типа. Пять других аминокислот кодируются триплетами вида WWN (W означает A или T и описывает *слабое спаривание*): фенилаланин (F), изолейцин (I), лизин (K), аспарагин (N) и тирозин (Y); их частоты уменьшаются с увеличением содержания GC. Метионин технически относится к WWN-типу, однако редко встречается в бактериальных геномах ($\sim$1,8%) и его частота от содержания GC зависит слабее остальных аминокислот. Аминокислоты, кодируемые комбинацией сильного и слабого спаривания на первых двух позициях триплета рассматривались как нейтральные. Из нейтральных аминокислот только частота валина обнаружила положительную корреляцию с содержанием GC, т.е. как аминокислота SSN-типа.

Параметры набора марковских моделей (трехпериодических нулевого, первого и второго порядков для кодирующей последовательности и однородных нулевого — для некодирующей) были определены следующим образом. Нахождение содержания GC в данном геноме позволяет вычислить позиционные частоты каждого из четырех нуклеотидов, используя линейные зависимости, определенные ранее. Далее, частоты встречаемости каждого из 61 кодона вычисляются как произведение трех позиционных частот составляющих его нуклеотидов:

$$f_I\left(XYZ\right) = f\left(X\right) f\left(Y\right) f\left(Z\right).$$

Затем они модифицировались с учетом вычисленной по содержанию GC частоты кодируемой триплетом аминокислоты. Например, итоговая оценка частоты кодона аланина GCT определяется следующей формулой:

$$f_{\mathrm{R}}\left(\mathrm{GCT}\right) = f_{\mathrm{alanine}}\left(\mathrm{GC}\%_{global}\right) \frac{f_I\left(\mathrm{GCT}\right)}{f_I\left(\mathrm{GCT}\right) + f_I\left(\mathrm{GCC}\right) + f_I\left(\mathrm{GCA}\right) + f_I\left(\mathrm{GCG}\right)}.$$

При подобных вычислениях для каждого из 61 кодона была получена таблица использования кодонов для входной последовательности [12].

Чтобы построить трехпериодическую марковскую модель нулевого порядка для белоккодирующей области, необходимо вычислить три переходных матрицы для каждой позиции внутри триплета. Это было сделано при применении найденной таблицы использования кодонов: вероятность обнаружить нуклеотид X в i-й позиции триплета равна сумме частот всех кодонов, имеющих нуклеотид X на этой позиции. В марковской модели нулевого порядка для некодирующей последовательности были использованы соответствующие глобальные частоты встречаемости кодонов.

Таблица 1

Точность предсказания генов программой GeneMarkS [11]

Геном	**Чувствительность, %**	**Специфичность, %**
A.fulgidus	98,5	91,8
B.subtilis	98,8	91,1
E.coli	96,9	94,5
H.influenzae	98,2	92,8
H.pylori	97,7	86,5
M.jannaschii	99,4	90,1
M.thermoautotrophicum	97,9	94,5
Synechocystis	98,7	88,8
Среднее	98,3	91,3

Для трехпериодической марковской модели первого порядка на основе таблицы использования кодонов можно определить только две матрицы вероятностей перехода из трех. Для определения величин переходных вероятностей для нуклеотидов третьей позиции одного триплета и первой позиции следующего вводилось предположение о независимом следовании триплетов друг за другом. Таким образом, вероятность $P(X \to Y)$ следования нуклеотида Y в первой позиции кодона за нуклеотидом X в третьей позиции предыдущего кодона (конфигурация $...X||Y...$) равна вероятности нуклеотида Y в первой позиции кодона в определенной выше марковской модели нулевого порядка.

Для трехпериодической марковской модели второго порядка можно определить только одну матрицу вероятностей перехода нуклеотидов в третьей позиции кодона. Чтобы найти переходные вероятности для первых двух позиций кодона, было использовано то же предположение о независимом следовании кодонов. Так, вероятность $P(XY \to Z)$ в конфигурации $...XY||Z...$ получалась равной вероятности встречи в первой позиции кодона нуклеотида Z в марковской модели нулевого порядка. Вероятность $P(XY \to Z)$ для конфигурации $...X||YZ...$ получалась равной вероятности нуклеотиду Z оказаться во второй позиции кодона при нуклеотиде Y в первой позиции, эта вероятность уже была определена ранее для марковской модели первого порядка.

Таким образом, описаны основные отличия алгоритма GeneMarkS от традиционного алгоритма дешифрования с использованием скрытых марковских моделей:

1) Использование обобщенной скрытой марковской модели.

2) Применение эвристической информации о связи между частотами нуклеотидов в данной позиции триплета и кодируемых триплетами аминокислот с содержанием GC в геноме.

3) Использование неконтролируемого обучения по Витерби.

Точность аннотации и сравнение с другими программами

Точность проводимой программой аннотации обычно описывают двумя параметрами: чувствительностью *Sn* («sensitivity»), характеризующей долю правильно предсказанных генов среди всех реально существующих генов, и специфичностью *Sp* («specificity»), характеризующей долю правильно предсказанных генов среди всех предсказанных генов. Оценка этих параметров на бактериальных геномах показала хорошую эффективность алгоритма GeneMarkS (см. табл. 1): средняя чувствительность составила 98%, а средняя специфичность — 91% [11].

Для сравнения с конкурирующими алгоритмами аннотации геномов на примере генома была выбрана программа Glimmer (версии 2.02) [6], как другая часто использующаяся

Таблица 2

Сравнение программ GeneMarkS и Glimmer 2.02 [11]

Программа	**Набор генов**	**Количество генов в наборе**	**Число генов, предсказанных точно**	**Число предсказанных генов**
Glimmer	A	4099	2556 (62,4%)	**4023 (98,1%)**
GeneMarkS	A		**3412 (83,2%)**	3962 (96,7%)
Glimmer	B	123	70 (57,0%)	112 (91,1%)
GeneMarkS	B		**102 (82,9%)**	**113 (91,9%)**
Glimmer	C	72	41 (57,0%)	66 (91,7%)
GeneMarkS	C		**64 (88,9%)**	**68 (94,4%)**
Glimmer	D	51	26 (51,0%)	45 (88,2%)
GeneMarkS	D		**46 (90,2%)**	**48 (94,1%)**
Glimmer	E	195	139 (71,3%)	**195 (100%)**
GeneMarkS	E		**184 (94,4%)**	**195 (100%)**

программа. Совокупность генов была разбита на 5 частей: полный набор генов *B.subtilis* согласно аннотации GenBank (A), три набора генов *B.subtilis* не длиннее 300 нуклеотидов с как минимум одной (B), двумя (C) и десятью (D) значимыми гомологиями, определенными анализом BLAST [20], и набор из 195 экспериментально подтвержденных генов E.coli (E). В каждом наборе производился подсчет точно предсказанных генов (предсказанные 5' и 3'-концы совпали с аннотацией) и подсчет генов, предсказанный 3'-конец которых совпал с аннотацией. Результаты сравнения показаны в табл. 2.

Из таблиц 1 и 2 видно, что GeneMarkS не только не уступает, а практически всюду выигрывает в точности предсказания.

Недостатки семейства алгоритмов GeneMark

Наибольшие потери точности наблюдаются при предсказании перекрывающихся, коротких и длинных генов. Возможно, это связано с тем, что алгоритм не успевает переходить между скрытыми состояниями на коротких дистанциях и, наоборот, заведомо переходит между состояниями на очень больших дистанциях, обусловленных биологически.

Не очень высокая специфичность алгоритма, т.е. значительная доля ложнопредсказанных генов, может быть обусловлена чисто статистическими причинами. Например, если в принятой модели некоторая подпоследовательность предшествовала гену, то на достаточно длинном межгенном участке вполне могут встретиться подпоследовательности, идентичные ей.

При обработке геномов эукариот и вирусов алгоритм сталкивается с трудностями, связанными со сложной интрон-экзонной структурой и перекрыванием генов, наличием транспозонов и т.д. Существенную роль также играет глобальная неоднородность длинной нуклеотидной последовательности: если локально ее части хорошо описываются скрытыми марковскими моделями, то в сумме точность описания обоих частей моделью с одним набором параметров сильно падает.

Наконец, алгоритм принципиально короткодействующий: для данной позиции он учитывает влияние не более чем пяти (для модели 5-го порядка) предыдущих нуклеотидов.

Повышение точности предсказания путем увеличения порядка цепи Маркова приводит к геометрическому росту числа независимых параметров модели ($\sim 4^n$). Помимо экспоненциального увеличения времени работы программы такой алгоритм потребует обширную обучающую выборку, поэтому увеличение порядка модели вычислительно неэффективно и необходимы иные пути повышения точности алгоритма. Одним из таких путей может стать поиск регуляторных участков, различных мотивов или других структур в межген-

ном пространстве.

Вывод

В настоящей статье проведен анализ семейства алгоритмов GeneMark для аннотации прокариотических геномов. Хотя по сравнению с аналогичными программами GeneMark показывает хорошие результаты, точность предсказания в ряде случаев может быть улучшена: это касается предсказания генных стартов, предсказания генов в фагах (где гены часто перекрываются). Определенной проблемой является также случайный пропуск реальных очень коротких генов, или, наоборот, появление ложно предсказанных коротких генов в длинных некодирующих участках из-за случайных аномалий в композиции последовательности. Поскольку дальнейшее усложнение модели может вызвать затруднения в оценке необходимого числа параметров, следует искать новые пути повышения точности предсказания генов. Например, одним из возможных путей решения проблемы улучшения предсказания коротких генов, генных стартов и перекрывающихся генов является разработка алгоритмов обнаружения регуляторных участков, располагающихся вблизи генных стартов. Другим направлением является учет локальных вариаций GC композиции генома, часто связанных с горизонтальным переносом генов из других микробов [18].

Литература

1. *Borodovsky M., Sprizhitsky Yu., Golovanov E., Alexandrov A.* Statistical Patterns in Primary Structures of Functional Regions in the E. Coli Genome: I. Oligonucleotide Frequencies Analysis // Molecular Biology. — 1986. — V. 20. — P. 826–833.
2. *Borodovsky M., Sprizhitsky Yu., Golovanov E., Alexandrov A.* Statistical Patterns in Primary Structures of Functional Regions in the E. Coli Genome: II. Non-homogeneous Markov Models // Molecular Biology. — 1986. — V. 20. — P. 833–840.
3. *Borodovsky M., Sprizhitsky Yu., Golovanov E., Alexandrov A.* Statistical Patterns in Primary Structures of Functional Regions in the E. Coli Genome: III. Computer Recognition of Coding Regions // Molecular Biology. — V. 20. — P. 1145–1150.
4. *Borodovsky M., McIninch J.* GeneMark: parallel gene recognition for both DNA strands // Computers & Chemistry. — 1993. — V. 17, N. 19. — P. 123–133.
5. *Borodovsky M., McIninch J.* Recognition of genes in DNA sequence with ambiguities // Biosystems. — 1993. — V. 30, N 1–3. — P. 161–171.
6. *Delcher A.L., Harmon D., Kasif S., White O., Salzberg S.L.* Improved microbial gene identification with GLIMMER // Nucleic Acids Res. — 1999. — V. 27. — P. 4636–4641.
7. *Frishman D., Mironov A., Mewes H.-W., Gelfand M.* Combining diverse evidence for gene recognition in completely sequenced bacterial genomes // Nucleic Acids Res. — 1998. — V. 26. — P. 2941–2947.
8. *Rabiner L.R., Juang B.H.* An introduction to hidden Markov models // IEEE ASSP Magazine. — 1986. — V. 3, N 1. — P. 4–16.
9. *Krogh A., Mian I.S., Haussler D.* A hidden Markov model that fins genes in *E.coli* DNA // Nucleic Acids Res. — 1994. — V. 22. — P. 4768–4778.
10. *Lukashin A., Borodovsky M.* GeneMark.hmm: new solutions for gene finding // Nucleic Acids Research. — 1998. — V. 26, N 4. — P. 1107–1115.
11. *Besemer J., Lomsadze A., Borodovsky M.* GeneMarkS – a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions // Nucleic Acids Research. — 2001. — V. 9, N 12. — P. 2607–2618.
12. *Besemer J., Borodovsky M.* Heuristic approach to deriving models for gene finding // Nucleic Acids Research. — 1999. — V. 27, N 19. — P. 3911–3920.

13. *Lomsadze A., Ter-Hovhannisyan V., Chernoff Y., Borodovsky M.* Gene identification in novel eukaryotic genomes by self-training algorithm // Nucleic Acids Research. — 2005. — V. 33, N 20. — P. 6494–6506.

14. *Zhu W., Lomsadze A., Borodovsky M.* Ab initio gene identification in metagenomic sequences // Nucleic Acids Research. — 2010. — V. 38, N 12.

15. *Durbin R., Eddy S., Krogh A., Mitchison G.* Biological sequence analysis: Probabilistic models of proteins and nucleic acids. — Cambridge University Press, 1998. — P. 54.

16. *Viterbi A.* Error bounds for convolutional codes and an asymptotically optimum decoding algorithm // Information Theory. — 1967. — V. 13, N 2. — P. 260–269.

17. *Baum L.E.* An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes // Inequalities. — 1972. — V. 3, N 1. — P. 1–8.

18. *Medigue C., Rouxel T., Vigier P., Henaut A., Danchin A.* Evidence for horizontal gene transfer in Escherichia coli speciation // J. Mol. Biol. — 1991. — V. 222. — P. 851–856.

19. *Lawrence J.G.* Selfish operons and speciation by gene transfer // Trends Microbiol. — 1997. — V. 5. — P. 355–359.

20. *Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J.* Basic local alignment search tool // J. Mol. Biol. — 1990. — V. 215, N 3. — P. 403–410.

Поступила в редакцию 02.02.2012.