УДК 519.6+576.895.42+519.2

МЕТОД ПОИСКА ЛОГИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ В СТРУКТУРЕ ГЕНОМОВ*

Г.Л. Полякова[1,3], Г.С. Лбов[2], В.А. Гусев[1], В.С. Алтынцева[3], В.А. Габриэль[3]
[1] *Учреждение Российской академии наук Институт математики СО РАН,*
e-mail: Polyakova@math.nsc.ru, karusel2007@ya.ru; vgus@math.nsc.ru;
[2] http://lbovgenady.narod2.ru;
[3] *Национальный исследовательский Новосибирский государственный университет,*
e-mail: Altynya@gmail.ru, vitek-novosib@mail.ru

Аннотация

Подтверждена гипотеза о наличии логических закономерностей в структуре генома. Для анализа геномов высших форм живых систем был использован алгоритм полного перебора конъюнкций (L - грамм) для выявления логических закономерностей, обладающих высокой относительной частотой их встречаемости в бинарной последовательности. Приведено описание логико-вероятностной модели для бинарных последовательностей и алгоритма обнаружения логических закономерностей в бинарной последовательности.

Ключевые слова: логико-вероятностная модель, бинарная последовательность, структура генома

ВВЕДЕНИЕ

Обширная библиотека расшифрованных генетических последовательностей (GenBank) является в настоящее время объектом пристального внимания математиков. Как правило, работы по математическому анализу геномов посвящены применению различных математических методов для выявления регуляторных и кодирующих участков в структуре геномов [1-3]. Этот анализ основан на сопоставлении различных символьных последовательностей исследуемых геномов с паттернами ДНК, функции которых уже известны.

В работах [4-10] авторы проводили анализ структуры генетического кода, то есть кодон-аминокислотного соответствия с использованием методов теоретико-группового анализа. Это позволило обнаружить неизвестные ранее молекулярным биологам и биохимикам закономерности в структуре кода.

Логично предположить, что соответствующие алгебраические и арифметические, то есть символьные и числовые закономерности генетического кода должны иметь отображения в нуклеотидной последовательности геномов. Для поиска числовых закономерностей необходимо представить стандартную символьную последовательность из AT и GC пар в геноме в цифровом виде. Мы воспользовались данными работы [8], в которой показано, что молекулярные массы AT и GC пар в составе двойной спирали равны соответственно 259 и 260 независимо от ориентации, то есть принадлежности нуклеотида к конкретной нити ДНК. Таким образом, чередование AT и GC пар в двойной спирали соответствует чередованию только двух чисел. Для анализа закономерностей удобно их

* Работа выполнена при финансовой поддержке РФФИ (проект № 10-01-00113-а)

представить в виде последовательностей нулей и единиц. Поиск возможных закономерностей в таком бинарном ряду чисел был проведен в рамках подхода к анализу эмпирической информации, изложенного в монографиях [11, 12, 13]. Подход сводится к построению логико-вероятностной модели объекта исследования. Под логико-вероятностной моделью понимается список логических закономерностей, обладающих достаточно большой прогнозирующей способностью (см. ниже). Целью данной работы является поиск логических закономерностей в бинарных последовательностях, сопоставленных кодирующим фрагментам генома прокариот *E.coli*, а также фрагментам X и Y хромосом эукариотического генома вида *Homo Sapiens*.

ОБНАРУЖЕНИЕ ЛОГИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ

Метод поиска (обнаружения) логических закономерностей в бинарной последовательности состоит из выбора наилучшего разбиения, полученного для слова длины l и его окружения длины s, по некоторому критерию. Для каждого разбиения определяются частоты перехода из множества окружений слова x в множество слов y для всех пар $< x, y >$ на основе анализа исходной последовательности; затем определяется порядок для всех пар по их частотам появления и выделяются пары с наибольшими частотами, которые и будут логическими закономерностями.

Пусть имеется последовательность $gene$ длины L: $gene = (nucl_1, nucl_2 ..., nucl_L)$, где $nucl_m \in D_{nucl}$, $m = 1,..., L$ и множество D_{nucl} представляет собой неупорядоченный набор значений элементов последовательности $D_{nucl} = \{a, t, c, g\}$.

Переходим к бинарной последовательности заменой в исходной последовательности символов a и t на 0, c и g на 1. Получаем последовательность $b = (b_1, b_2, ..., b_L)$, где $b_m \in \{0,1\}$, $m = 1,..., L$.

Под словом длины l в алфавите $\{0,1\}$ понимается бинарная последовательность длины l (например, при $l = 4$ «слово» имеет вид 0110; заметим, что число возможных слов длины l в этом случае равно $2^4 = 16$). Рассматриваем слова длины l.

Предположим, что частота возникновения слова $w^k = (b_k, b_{k+1}, ..., b_{k+l-1})$ длины l, начинающегося с k-й позиции в последовательности b, зависит только от некоторых из s_1 ближайших слева и s_2 ближайших справа элементов («окружения»), т.е. зависит от некоторых из элементов $b_s^k = (b_{k-s_1}, b_{k-s_1+1}, ..., b_{k-1}, b_{k+l}, ..., b_{k+l+s_2-1})$, где $s = s_1 + s_2$; s_1, s_2 – некоторые параметры, $k = s_1 + 1, ..., L - l - s_2 + 1$. Например, при $s_1 = 3$, $s_2 = 2$ для указанного выше слова «окружение» из 3 ближайших слева и 2 ближайших справа элементов может иметь вид 110(0110)01.

Логической закономерностью для указанного слова называется логическое высказывание на символах окружения, которое с большой частотой характеризует данное слово. Например, для указанного слова 0110 высказывание «(слева в 1-й позиции стоит 1) и (справа в 1-й позиции стоит 0)» появляется в рассматриваемой бинарной последовательности с относительной частотой $\overline{P} \geq d = 0{,}95$, d - параметр; а для всех других слов появляется с частотой, близкой к нулю.

Сопоставим каждому участку b_s^k набор значений некоторых бинарных переменных $X = X_1, X_2, ..., X_n$, $n = s_1 + s_2$. $D_j = \{0,1\}$ - область определения переменной X_j. Пусть $x^k = X(b_s^k) = (X_1(b_s^k), X_2(b_s^k), ..., X_n(b_s^k))$; $X_j(b_s^k)$ - значение переменной X_j для участка b_s^k. Для указанного выше примера окружение имеет вид: $(x_1 = 1, x_2 = 1, x_3 = 0, x_4 = 0, x_5 = 1)$.

Частоту встречаемости данного слова w длины l в последовательности b определим как $\overline{P}(w) = \frac{N_w}{N}$, где $N = L - l - s + 1$ – число всех слов длины l, (совпадающих и несовпадающих) которые можно получить из последовательности b сдвигом на один символ с позиции $k = s_1 + 1$ в последовательности b до позиции $k = L - l - s_2 + 1$; N_w - число повторов слова w среди всех таких N слов. Для надежности метода необходимо выбрать среди всевозможных слов длины l слово w_* наиболее высокой частоты встречаемости.

Сопоставим слову $w^k = (b_k, b_{k+1}, ..., b_{k+l-1})$, начинающегося с k-й позиции в последовательности b, $k = s_1 + 1, ..., L - l - s_2 + 1$, значение целевой (прогнозируемой) переменной Y. Будем считать, что $y^k = Y(w^k) = 1$, если слово $w^k = w_*$; $y^k = Y(w^k) = 2$, если слово $w^k \neq w_*$.

Сопоставив каждому значению x^k значение y^k, получим таблицу данных $v = \{x^k, y^k\}$, размерностью $n \times \mathrm{N}$, где $n = s_1 + s_2$, $N = L - l - s + 1$. Можно определить по таблице данных число $N_{(1)}$ объектов первого образа, и число $N_{(2)}$ объектов второго образа.

Требуется по этим наблюдениям найти логические закономерности, обладающие большой прогнозирующей способностью, для предсказания значения y в зависимости от «окружения» x. Множество таких закономерностей представляет логико-вероятностную модель, отражающую причинно-следственные взаимосвязи между характеристиками. В процессе построения закономерностей автоматически отбираются наиболее информативные характеристики.

Задача обнаружения всех закономерностей является Np - трудной задачей. Для обнаружения закономерностей используются алгоритмы класса ТЕМР [11, 12, 13, 14], которые дают возможность значительно сократить время вычислений, учитывать разнотипность переменных, перебирать конъюнкции различной длины. Эти алгоритмы обнаруживают все логические закономерности на реальных таблицах за приемлемое время.

Обозначим $J(a, E_j)$ - предикат, принимающий значения «истина» или «ложь». Предикат $J(a, E_j)$ эквивалентен утверждению: $X_j \in E_j$, $a \in \Gamma$ - объект из некоторой генеральной совокупности, описываемый характеристиками $X_1, ..., X_n, Y$; E_j является подмножеством множества значений D_j, $j = 1, ..., n$.

Назовем $S(a, E) = J(a, E_{j_1}) \wedge ... \wedge J(a, E_{j_d})$ конъюнкцией длины d. Областью истинности конъюнкции $S(a, E)$ является подмножество $E = \prod_{i=1}^{d} E_{j_i}$, $E_{j_i} \subset D_{j_i}$. Обозначим через m нормированную меру подмножества E. Для любой конъюнкции $S(a, E)$ можно

определить по таблице данных v число объектов первого образа $N_{(1,S)}$ и число объектов второго образа $N_{(2,S)}$, на которых указанная конъюнкция истинна.

Конъюнкцию $S(a,E)$ будем называть *логической закономерностью,* с большой вероятностью характеризующей первый образ, если выполняются неравенства: $\frac{N_{(1,S)}}{N_{(1)}} \geq d$, $\frac{N_{(2,S)}}{N_{(2)}} \leq b$, где δ и β - некоторые параметры; $0 \leq b < d \leq 1$. Чем больше d и меньше β, тем сильнее логическая закономерность. Множество всех закономерностей обозначим через S^*.

Конъюнкцию $S(a,E)$ будем называть *потенциальной логической закономерностью* для первого образа (обозначим ее через *S'),* если выполняются неравенства: $\frac{N_{(1,S)}}{N_{(1)}} \geq d$, $\frac{N_{(2,S)}}{N_{(2)}} > b$. Множество потенциальных закономерностей обозначим через S'. Очевидно, что из $S' \in \mathsf{S}'$ можно получить закономерность S^* последовательным присоединением предикатов, т.е. $S' \wedge J(a,E_j) \wedge \mathbf{K}$; если для некоторой конъюнкции $S(a,E)$ выполняется неравенство $\frac{N_{(1,S)}}{N_{(1)}} < d$, то конъюнкция S по определению не является закономерностью и присоединение к ней какого-либо предиката не даст закономерности (множество таких конъюнкций обозначим через S). Таким образом, любая конъюнкция $S(a,E)$ может быть трех типов: S^*, S', S.

Алгоритм обнаружения логических закономерностей состоит в последовательном выполнении следующих шагов.

На *первом шаге* рассматриваются всевозможные конъюнкции длины один, т.е. конъюнкции вида $S(a,E) = J(a,E_j)$, E_j является подмножеством множества значений D_j, $j = 1,...,n$. Если $S(a,E) \in \mathsf{S}^*$, то она включается в список закономерностей и соответствующее подмножество E_j исключается из дальнейшего перебора; если $S(a,E) \in \mathsf{S}'$, то соответствующее подмножество E_j оставляется для дальнейшего перебора; если $S(a,E) \in \mathsf{S}$, то соответствующее подмножество E_j исключается из дальнейшего перебора. Обозначим через Q_j^1 множество подмножеств E_j, оставленных для дальнейшего перебора после выполнения первого шага алгоритма.

На *втором шаге* рассматриваются всевозможные конъюнкции длины два, т.е. конъюнкции вида $S(a,E) = J(a,E_i) \wedge J(a,E_j)$, $i \neq j$, $E_i \in Q_i^1$, $E_j \in Q_j^1$. Если $S(a,E) \in \mathsf{S}^*$, то соответствующие подмножества E_i и E_j исключаются из дальнейшего перебора, и соответствующая конъюнкция включается в список закономерностей; если $S(a,E) \in \mathsf{S}'$, то соответствующие подмножества E_i и E_j оставляются для дальнейшего перебора; если $S(a,E) \in \mathsf{S}$, то соответствующие подмножества E_i и E_j исключаются из дальнейшего

перебора. Аналогично обозначаем Q_j^2 множество подмножеств E_j, оставленных для дальнейшего перебора после выполнения второго шага алгоритма.

Далее, аналогично, рассматриваются конъюнкции длины три, четыре, пять и т.д. В результате работы алгоритма получаются конъюнкции небольшой длины. Например, максимальная длина полученных конъюнкций в задаче, описанной ниже, не больше 6.

ЗАКОНОМЕРНОСТИ В ГЕНЕТИЧЕСКИХ ПОСЛЕДОВАТЕЛЬНОСТЯХ

В результате работы алгоритма может быть найдено несколько закономерностей. Вероятность образования таких закономерностей при условии равномерного распределения $P(S \mid H_0)$ различна. Из всех таких закономерностей естественно считать наилучшей ту, для которой эта вероятность минимальна.

Меру m можно рассматривать как вероятность попадания в область E при равномерном распределении, $E \subset D$; $1-m$ как вероятность попадания в область $D \setminus E$. Следовательно, вероятность $P(S \mid H_0)$ образования закономерности заданной длины при известных $N_{(1,S)}$, $N_{(2,S)}$, $N_{(1)}$, $N_{(2)}$ и m может быть вычислена следующим образом:

$$P(S \mid H_0) = C_{N_{(1)}}^{N_{(1,S)}} m^{N_{(1,S)}} (1-m)^{N_{(1)}-N_{(1,S)}} \cdot C_{N_{(2)}}^{N_{(2,S)}} m^{N_{(2,S)}} (1-m)^{N_{(2)}-N_{(2,S)}}.$$

Чем больше длина конъюнкции, тем меньше мера m и меньше вероятность $P(S \mid H_0)$; следовательно, при одинаковых значениях $N_{(1,S)}$, $N_{(2,S)}$, $N_{(1)}$, $N_{(2)}$ предпочтительней будут конъюнкции большей длины.

РЕЗУЛЬТАТЫ АНАЛИЗА ГЕНОМА E-COLI

Был проанализирован весь геном *E-coli* (4266 генов, GenBank [15]). Однако объем статьи не позволяет привести весь иллюстративный материал. Приведем здесь в качестве примера только некоторые из полученных результатов.

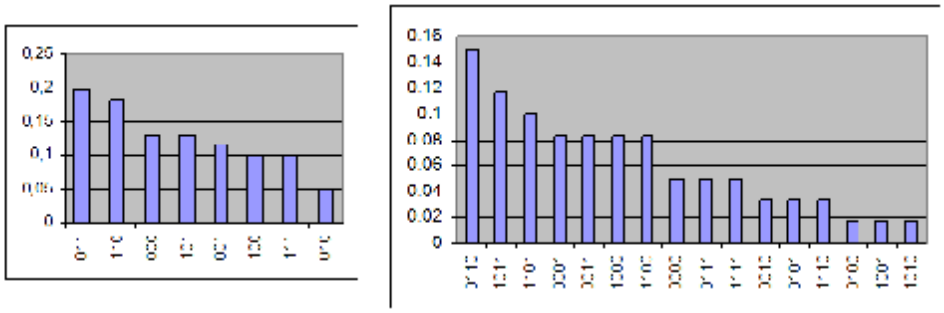

Рис. 1 Частоты встречаемости слов длины 3 и 4 в бинарной последовательности

Для кодирующей последовательности гена thrL *E-coli* (длина 63 символа) ниже приводим гистограммы частот встречаемости слов в соответствующей бинарной последовательности. Примеры гистограмм, отражающих среднюю частоту встречаемости слов соответствующей длины, приведены на Рис. 1.

Исследования проводились при длине слова от 3 до 7 символов. Ограничение длины слова обусловлено большими затратами машинного времени. Вероятность появления слова длины l при условии «чисто» случайной последовательности равна $\frac{1}{2^l}$. Вероятность появления при $l = 3, 4, 5, 6, 7$ равны $0{,}1250;\ 0{,}0625;\ 0{,}0312;\ 0{,}0156;\ 0{,}0078$. Полученные частоты закономерностей на порядок превышают указанные вероятности.

Найдены следующие закономерности:

1. Для слова 011 окружение «(слева в 10-й позиции стоит 1) и (слева в 9-й позиции стоит 0) и (справа в 7-й позиции стоит 0)» появляется в рассматриваемой бинарной последовательности с относительной частотой 0,778; для всех остальных слов длины 3, не равных данному, с относительной частотой 0. Вероятность $P(S \mid H_0)$ приближенно равна $1,83e^{-7}$
2. Для этого же слова 011 окружение «(слева в 10-й позиции стоит 1) и (слева в 9-й позиции стоит 0) и (слева в 3-й позиции стоит 0) и (справа в 7-й позиции стоит 0)» появляется в рассматриваемой бинарной последовательности с относительной частотой 0,778; для всех остальных слов длины 3, не равных данному, с относительной частотой 0. Вероятность $P(S \mid H_0)$ приближенно равна $1,49e^{-8}$.
3. Для слова 0110 окружение «(слева в 10-й позиции стоит 1) и (слева в 9-й позиции стоит 0) и (слева в 3-й позиции стоит 0) и (справа в 6-й позиции стоит 0)» появляется в рассматриваемой бинарной последовательности с относительной частотой 0,778; для всех остальных слов длины 4, не равных данному, с относительной частотой 0. Вероятность $P(S \mid H_0)$ приближенно равна $1,49e^{-8}$.

Вероятность образования первой закономерности при условии равномерного распределения $P(S \mid H_0)$ больше, чем второй и третьей закономерностей, поэтому предпочтение отдается последним.

Для кодирующей последовательности гена yahH E-coli (длина 318 символов) на Рис. 2 приведен пример гистограммы частот встречаемости слов в соответствующей бинарной последовательности и найдена закономерность.

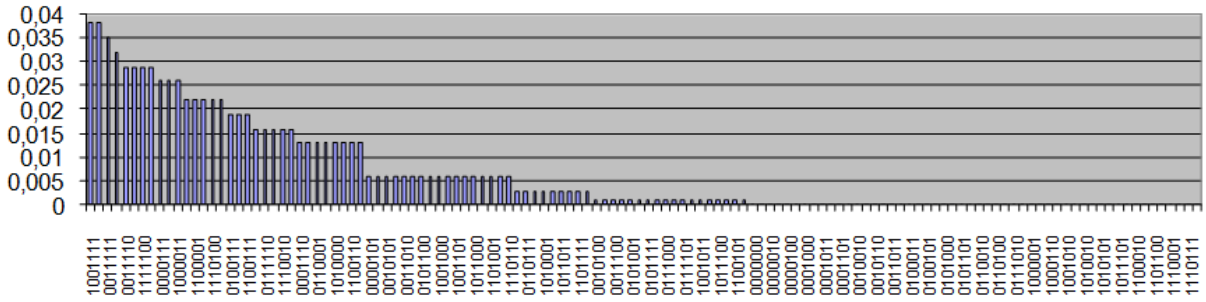

Рис. 2 Частоты встречаемости слов длины 7 в бинарной последовательности

Для слова 1001111 окружение «(слева в 10-й позиции стоит 1) и (слева во 2-й позиции стоит 1) и (справа в 1-й позиции стоит 0) и (справа в 4-й позиции стоит 0) и (справа в 7-й позиции стоит 1)» появляется в рассматриваемой бинарной последовательности с относительной частотой 0,727; для всех остальных слов длины 7, не равных данному, с относительной частотой 0,046.

Вероятность образования такой закономерности при условии равномерного распределения $P(S \mid H_0)$ приближенно равна $1,25e^{-17}$, что дает право считать ее логической закономерностью. В качестве сравнения использовались случайные последовательности, сгенерированные в соответствии с частотами AT и GC пар в исходных генетических последовательностях. Было проанализировано по 100 случайных последовательностей для каждого исследуемого гена. Ни в одной из них закономерностей не было найдено.

РЕЗУЛЬТАТЫ АНАЛИЗА ФРАГМЕНТОВ X И Y ХРОМОСОМ ГЕНОМА ВИДА HOMO SAPIENS

Исследования проводились при длине слова от 3 до 7 символов. В результате были выделены наиболее часто встречающиеся закономерности в хромосомах X и Y. Затем в случайной бинарной последовательности определена относительная частота встречаемости для полученных закономерностей.

На Рис. 3 **Рис. 1**и Рис. 4 приведены графики частот встречаемости выявленных закономерностей при $l=5$, $s=6$ и $N=7000$. Отражены частоты появления слов, состоящих из пяти символов (по оси абсцисс средняя строка – пример выявленных закономерностей), с соответствующим окружением (по оси абсцисс первая и третья строки). Нижняя кривая соответствует встречаемости аналогичных комбинаций слов и окружения в рандомизированных последовательностях аналогичной длины.

Рис. 3 Относительная частота встречаемости закономерности в хромосоме Y

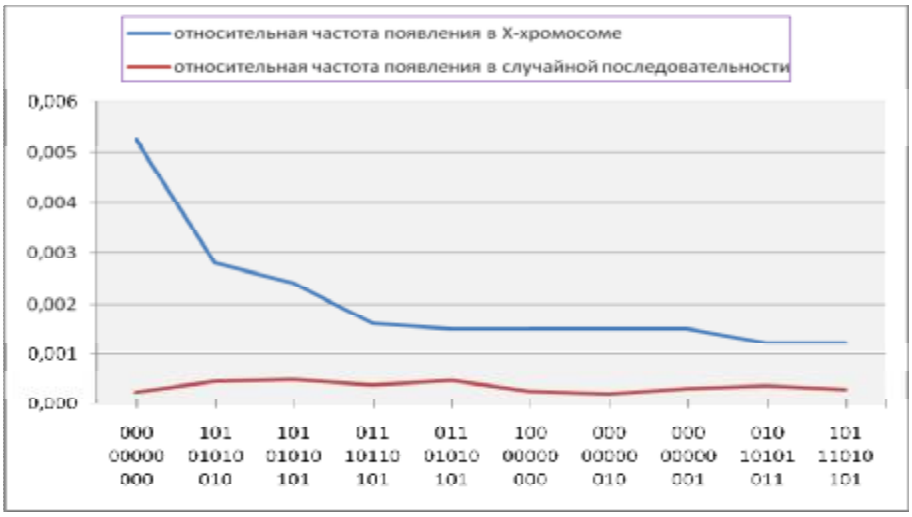

Рис. 4 Относительная частота встречаемости закономерности в хромосоме X

Как видно из графиков, относительные частоты встречаемости закономерностей в хромосомах X и Y человека на порядок отличаются от встречаемости аналогичных комбинаций в случайных последовательностях. Это говорит о том, что полученные пары с

большой частотой встречаемости практически не могут получиться из случайной последовательности. Поэтому в соответствии с гипотезой такие пары являются закономерностями в исходной последовательности.

ЗАКЛЮЧЕНИЕ

Анализ частот встречаемости слов различной длины в последовательностях показал, что частота слов в генетической последовательности значительно отличается от частот слов в случайной последовательности. Это дает основание утверждать, что наблюдаемые в геномах закономерности являются истинными. Следует особо подчеркнуть, что найденные закономерности в кодирующих последовательностях генома E-coli, а также во фрагментах X и Y хромосом человека имеют семантическую природу и непосредственно не связаны с триплетной структурой генома.

ЛИТЕРАТУРА

1. Орлов Ю.Л. Анализ регуляторных геномных последовательностей с помощью компьютерных методов оценок сложности генетических текстов: *Дис. канд. биол. наук*. - Новосибирск. – 2004, 158 С.
2. Abnizova I., Schilstra M., te Boekhorst R., Nehaniv C.L. A statistical approach to distinguish between different DNA functional parts // WSEAS Transactions on Computational Methods. 2003. V. 2. Issue 4. P. 1188–1196.
3. Abe T., Kanaya S., Kinouchi M, Ichiba Y., Kozuki T., Ikemura T. Informatics for unveiling hidden genome signatures // Genome Res. 2003. V. 13(4). P. 693–702.
4. Duplij D., Duplij S. Determinative degree and nucleotide content of DNA strands //***Biophys. Bull***. 2000. V. 497. P. 1-7.
5. Jimenez-Montano M.A., de la Mora-Basanez C.R. & Poschel T. The hypercube structure of the genetic code explains conservative and non-conservative aminoacid substitutions in vivo and in vitro // ***BioSystems.*** 1996. V. 39. P. 117-125.
6. Jimenez-Montano M.A. Protein evolution drives the evolution of the genetic code and vice versa // ***BioSystems***. 1999. V. 54. P. 47-64.
7. Negadi T. Rumer's Transformation in Biology as the Negation in Classic Logic //***Int. Journ. of Quant. Chem***. 2003. V. 94. P. 65-82.
8. Shcherbak V. I. Arithmetic inside the universal genetic code // BioSystems 2003. V. 70. P. 187-209.
9. Карасев В.А. Генетический код: новые горизонты. – Спб.: ТЕССА, 2003.– 116 С.
10. В.А. Гусев, Арифметика и алгебра в структуре генетического кода, логика в структуре генома и биохимическом цикле самовоспроизводства живых систем // Информационный вестник ВОГ и С том 9, № 2. май 2005. С. 153-161.
11. Лбов Г.С., Методы обработки разнотипных экспериментальных данных. Новосибирск: Изд-во Наука, 1981. 160 С.
12. Лбов Г.С., Старцева Н.Г. Логические решающие функции и вопросы статистической устойчивости решений. Новосибирск: Изд-во Ин-та математики, 1999. 212 С.
13. Лбов Г.С., Бериков В.Б. Устойчивость решающих функций в задачах распознавания образов и анализа разнотипной информации. Новосибирск: Изд-во Ин-та математики, 2005. 218 С.
14. Лбов Г.С., Полякова Г.Л. Метод прогнозирования в классе логических решающих функций // Вестник Сибирского государственного аэрокосмического университета имени академика М.Ф. Решетнева. Выпуск 5 (31). 2010. С. 42-45.
15. http://www.ncbi.nlm.nih.gov/(GenBank)