

ОБ ИНФОРМАЦИОННОЙ ЕМКОСТИ СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

М. Г. САДОВСКИЙ

Институт биофизики СО РАН, Красноярск, Россия

e-mail: msad@icm.krasn.ru

The concept of informational capacity is introduced which based on a comparison of the real and expected frequencies of small fragments in a symbol sequence. Main characteristics of the informational capacity are examined. A few examples are presented.

Введение

Символьные последовательности как объект исследования встречаются в самых разных областях знания — от математики и информатики до биологии и лингвистики. Часто символьные последовательности выступают в роли математической модели той среды, которая передает и хранит информацию. Интерпретация той информации, которая представлена в символьной последовательности, зависит от конкретной области знаний, в рамках которой рассматриваются свойства такой последовательности. Вопрос о количественных характеристиках содержания этой информации может быть разрешен в рамках анализа статистических свойств самой последовательности безотносительно к конкретным свойствам природного прототипа.

Ответ на вопрос о количественных показателях информационной емкости может быть сформулирован в точных терминах. Говоря об информационной емкости последовательностей, следует подчеркнуть, что в прикладных задачах, как правило, фигурируют конечные символьные последовательности. Являясь конечным объектом, такие последовательности обладают нулевой информационной емкостью. Для того чтобы содержательно вести речь об информационной емкости, необходимо перейти от рассмотрения конечных символьных последовательностей к рассмотрению ансамблей бесконечных текстов, в том или ином смысле соответствующих исходной конечной последовательности¹. Такой переход совершается путем построения частотного словаря рассматриваемой последовательности. Далее мы не будем различать частотный словарь, построенный по конечной последовательности, и соответствующий ему частотный словарь, построенный по всему ансамблю; специальные указания на различия будут делаться в случае, если это будет необходимо для понимания.

Содержательно информационная емкость является мерой предсказуемости при построении продолжения слова (короткого фрагмента), взятого из последовательности. Формулировка гипотез о продолжениях слов, взятых из последовательности, требует указания того правила, по которому будет оцениваться вероятность того или иного продолжения,

© Институт вычислительных технологий Сибирского отделения Российской академии наук, 2005.

¹Имеющих такой же частотный словарь.

ожидаемого для данного слова. Здесь возможны два варианта. Первый — заменить исходную последовательность на модельную (обычно предполагают, что исходная последовательность является реализацией какого-либо случайного процесса) и оценить вероятность продолжения, опираясь на свойства этого модельного случайного процесса. Вторым вариантом состоит в том, чтобы оценить вероятность продолжений у слов, опираясь лишь на знание частот самих этих слов. Мы будем следовать второму пути.

Вообще говоря, исследования статистических свойств символьных последовательностей весьма популярны, особенно в связи с расшифровкой большого объема генетических данных. Большую часть исследований в этом направлении составляют работы, выполненные в рамках первого подхода, — сравнения с модельными случайными последовательностями; мы не будем подробно на них останавливаться, поскольку настоящая работа выполнена в рамках другого подхода. Изучение статистических свойств нуклеотидных последовательностей имеет давнюю историю.

1. Частотные словари и наиболее ожидаемая частота слов

Рассмотрим какую-либо символьную последовательность длины N символов из конечного алфавита \aleph мощности $C = \|\aleph\|$, не содержащую пробелов²; любую связную подпоследовательность длиной q символов, встречающуюся в ней, будем называть словом (длины q). Совокупность всех слов (длины q), встречающихся в ней, будем называть q -носителем. Сопоставив каждому слову ω из q -носителя его частоту

$$f_{\omega} = \frac{n_{\omega}}{N},$$

получим частотный словарь $W(q)$ (толщины q); здесь n_{ω} — число копий этого слова.

Вопрос об информационной емкости символьной последовательности эквивалентен вопросу об уровне предсказуемости продолжений слов. Ответить на вопрос о предсказуемости продолжений можно двумя способами: первый — высказать гипотезу о свойствах всей последовательности целиком и на ее основе попытаться оценить вероятности возможных продолжений, второй — оставаться в рамках гипотезы о наиболее вероятном продолжении слов данной длины. Мы будем действовать в рамках второго подхода.

Это означает, что, отвечая на вопрос о предсказуемости продолжений, мы ограничимся лишь рассмотрением частотного словаря толщины q . Понятно, что вся информация, содержащаяся в частотном словаре $W(l)$, есть и в частотном словаре $W(q)$, $l < q$. Действительно [1, 2], информационное содержание частотного словаря $W(q)$ — это неопределенность в выборе того или иного конкретного слова из последовательности, по которой составлен этот частотный словарь. Мера такой неопределенности есть энтропия

$$S(W(q)) = - \sum_{\omega} f_{\omega} \ln f_{\omega} \quad (1)$$

соответствующего частотного словаря [3]. Вместе с тем частота любого слова ω' длины l ($l < q$), $\omega' \subset \omega$, может быть однозначно вычислена по частотному словарю $W(q)$ толщины q . Для этого необходимо просуммировать частоты всех слов длины q , отличающихся

²Рассмотрение несвязных последовательностей также возможно, однако ничего содержательно нового это не дает, а возникающие технические трудности лишь затрудняют понимание.

первыми (либо последними) $q - l$ символами. Если частотный словарь строится по конечной последовательности, то эти две суммы не совпадают — возникают краевые эффекты, связанные с конечностью исходной последовательности. Для того чтобы избежать такого рода эффектов, исходную последовательность следует замкнуть в кольцо [4, 5].

Как правило, словарь $W(q)$ толщины q содержит больше информации, чем словарь $W(l)$ толщины l , $q > l$. Однако существует критическая толщина словарей d^* , начиная с которой все словари бóльшей толщины содержат столько же информации, сколько и словарь $W(d^*)$. Отношение

$$r = \frac{d^*}{\ln N},$$

где N — длина последовательности, является мерой избыточности последовательности (некоторые приложения этой меры избыточности к задачам молекулярной биологии см. в [6–9]).

В общем случае информационное содержание частотных словарей возрастает по мере роста их толщины. Именно это обстоятельство и позволяет вести речь об их информационной емкости. Информационная емкость частотного словаря — это мера отличия реальных частот слов длины q от их ожидаемых частот. Тем самым понятие информационной емкости существенно зависит от того, каким именно способом будет определяться ожидаемая частота. Здесь возможны два варианта: делать оценки ожидаемой частоты на основе гипотезы о структуре исходной последовательности либо оценивать ожидаемую частоту исходя из частот комбинаций более коротких слов, которые составляют данное. Первый подход получил определенную популярность в задачах молекулярной биологии [10], однако априорное предположение о существовании определенной структурированности биологических макромолекул (точнее, соответствующих им символьных последовательностей) никак не обосновано.

1.1. Принцип максимума энтропии и ожидаемые частоты

Пусть в распоряжении исследователя имеются два частотных словаря — $W(q)$ и $W(q - 1)$. Реальные частоты слов длины q содержатся в первом из этих двух словарей. Ожидаемые частоты можно оценить, рассматривая все комбинации слов из словаря $W(q - 1)$, которые порождают слова из $W(q)$. Комбинируя слова из $W(q - 1)$, можно получить различные варианты словаря $W'(q)$, $W''(q)$, $W'''(q)$, ..., однако не все из них будут удовлетворять линейному ограничению. Дело в том, что такие словари, скомбинированные из слов меньшей длины, должны порождать вполне конкретный словарь $W(q - 1)$ при суммировании по первому (либо последнему) символу. Такое линейное ограничение выделяет во множестве словарей $W'(q)$, $W''(q)$, $W'''(q)$, ... семейство $\{\widehat{W}(q)\}$, которое, как правило, содержит более одного словаря. Среди всех словарей из указанного семейства найдется тот $\widehat{W}(q)$, для которого энтропия (1) максимальна³. Именно этот словарь реализует принцип наиболее вероятного продолжения слов длины $q - 1$ в слова длины q . Указанный экстремальный принцип позволяет вычислить частоты \tilde{f}_ω , $\tilde{f}_\omega \in \widehat{W}(q)$ слов в словаре $\widehat{W}(q)$ явным образом [1, 4, 5]:

$$\tilde{f}_{i_1 i_2 i_3 \dots i_{q-1} i_q} = \frac{f_{i_1 i_2 i_3 \dots i_{q-2} i_{q-1}} f_{i_2 i_3 i_4 \dots i_{q-1} i_q}}{f_{i_2 i_3 i_4 \dots i_{q-2} i_{q-1}}}. \quad (2)$$

³Существование максимума гарантировано конечностью множества словарей в семействе $\{\widehat{W}(q)\}$.

Формула (2) превращается в

$$\tilde{f}_{i_1 i_2} = f_{i_1} f_{i_2}$$

для случая $q = 2$. Возможен случай восстановления словаря $\widetilde{W}(q)$ по словарю $W(l)$ при $l < q - 1$, однако мы не будем останавливаться на рассмотрении этого случая (подробности см. в [1, 4, 5]).

1.2. Информационная емкость

Построение восстановленного частотного словаря $\widetilde{W}(q)$ позволяет строго ответить на вопрос об информационной емкости частотного словаря $W(q)$. Мерой информационной емкости словаря $W(q)$ является значение условной энтропии $\bar{S}(\widetilde{W}(q)|W(q))$ реального частотного словаря относительно восстановленного. Условная энтропия \bar{S} определяется следующим образом [3]:

$$\bar{S}(\widetilde{W}_q|W_q) = \sum_{\omega} f_{\omega} \ln \left(\frac{f_{\omega}}{\tilde{f}_{\omega}} \right). \quad (3)$$

Подставляя в (3) выражение (2) для ожидаемых частот, получаем

$$\bar{S}(q) = 2S_{q-1} - S_q - S_{q-2}; \quad (4)$$

$$\bar{S}(2) = 2S_1 - S_2, \quad (5)$$

соответственно, для случая $q = 2$. Выражение (3) определено корректно, поскольку восстановленный словарь $\widetilde{W}(q)$ содержит все слова, содержащиеся в $W(q)$, и, быть может, некоторые еще. Формулы (4) и (5) для случая восстановления $\widetilde{W}(q)$ по $W(l)$ при $l < q - 1$ превращаются в

$$\bar{S} = (q - l + 1)S_l - S_q - (q - l)S_{l-1} \quad \text{и} \quad \bar{S} = qS_1 - S_q$$

соответственно.

Поведение величин (4), (5) определяется длиной рассматриваемой последовательности и порядком чередования символов в ней. Для любой конечной последовательности всегда найдется такая критическая длина слов d^* , начиная с которой все словари толщины $q > d^*$ восстанавливаются однозначно [6–9]. Это означает, что такие словари содержат каждое слово в единственном экземпляре и энтропия таких словарей не меняется с увеличением длины слов⁴. Из этого следует, что для конечных последовательностей $\bar{S}_q = 0$, по крайней мере для $q \geq d^*$. С другой стороны, в типичном случае разнообразие частотных словарей по мере роста длины слов, содержащихся в них⁵, нарастает. Соответственно, можно ожидать, что информационная емкость (4) достигнет максимума при некотором $q = q_m$ (рис. 1).

Информационная емкость равна нулю⁶ для последовательностей двух типов: для полностью случайных нескоррелированных последовательностей и для последовательностей, обладающих ярко выраженным порядком и, соответственно, высоким уровнем предсказуемости. Очевидно, что для случайных последовательностей $f_{\omega} = \tilde{f}_{\omega} = f_1 f_2 \dots f_q$ (для

⁴И составляет $S_q = \ln N$, где N — длина последовательности.

⁵Для сравнительно коротких слов.

⁶С точностью до эффектов, связанных с конечностью исходной последовательности, для которой строится частотный словарь.

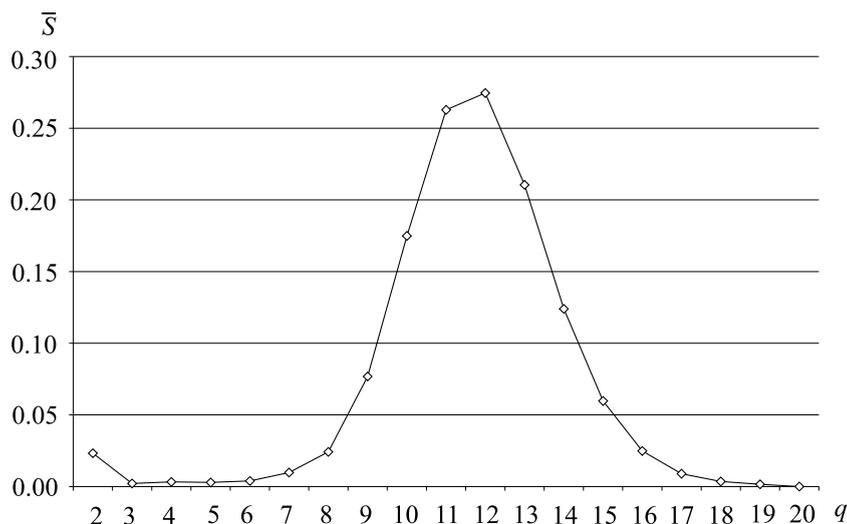


Рис. 1. Информационная емкость генома *Borrelia burgdorferi*.

случая восстановления по словарям толщины $q = 1$; для более толстых все аналогично), и, соответственно, (4), (5) обращаются в нуль. Оценки для высоко скоррелированных и упорядоченных последовательностей сложнее и заметно зависят от структуры самой последовательности, однако факт скоррелированности означает, что продолжения различных слов весьма хорошо предсказуемы, что и делает значение информационной емкости (4), (5) весьма близкой к нулю для них.

Проиллюстрируем изложенный здесь метод определением информационной емкости генома бактерии *Borrelia burgdorferi*. Этот геном представляет собой связную символьную последовательность из четырехбуквенного алфавита $\{A, C, G, T\}$ длиной 910 724 символа. Сама последовательность хранится в EMBL-банке (номер доступа AE000783). Для этой последовательности вычислялись величины (4) и (5) для $2 \leq q \leq 20$. На рис. 1 показано изменение информационной емкости для этого генома в указанном диапазоне длин слов. Для достаточно длинных слов ($q \geq 15$) информационная емкость рассматриваемого генома падает; этот очевидный факт связан с тем, что подавляющее большинство слов такой длины встречается в геноме в единичном экземпляре. Тем самым поведение максимума (4) также оказывается сильно зависящим от длины последовательности (т. е. от структуры частотного словаря соответствующей толщины); значения максимума (4) и его положение имеет смысл сравнивать лишь для близких по длине последовательностей, тогда они выявляют разницу в структурах таких последовательностей.

2. Обсуждение

В работе изложен строгий подход к определению информационной емкости символьной последовательности, точнее, информационной емкости ее частотного словаря. Само понятие информационной емкости становится содержательным лишь при переходе от рассмотрения конечных последовательностей к рассмотрению соответствующих им бесконечных ансамблей (и представляющих их частотных словарей).

Информационная емкость — это мера различия между ожидаемыми продолжениями коротких фрагментов в последовательности в более длинные и реально наблюдаемыми. Смысл величин (4), (5) заключается в том, что они измеряют неопределенность исхо-

да в эксперименте, в котором наугад выбирается некоторое слово из последовательности при условии, что наиболее ожидаемые частоты всех таких слов известны заранее. При этом ожидаемые частоты слов оцениваются по реальным частотам слов меньшей длины, наблюдаемым в той же последовательности. Кроме того, сам по себе способ получения ожидаемых частот также опирается на экстремальный принцип — получая оценки для ожидаемых частот, исследователь не вносит в них никакой априорной или дополнительной информации либо не опирается на знание, не содержащееся в самой последовательности.

Как отмечалось выше, информационная емкость равна нулю для двух типов последовательностей — случайных и высокоупорядоченных. Два этих типа, противоположные по своим статистическим свойствам, оказываются неразличимыми с точки зрения их информационной емкости. Сами такие последовательности тем не менее хорошо различаются по своим статистическим характеристикам: абсолютные значения энтропии S_q частотного словаря (толщины q) для случайной последовательности будут существенно больше, чем для упорядоченной.

Изучая различные символьные последовательности с точки зрения поведения информационной емкости их частотных словарей, важно понимать, в каких пределах может меняться эта величина и как она зависит от толщины рассматриваемого словаря. Максимум величины (4) будет наблюдаться при таком распределении символов, для которого наблюдается наибольшее отклонение реальной частоты от ожидаемой. По-видимому, хорошей оценкой может быть следующая. Рассмотрим последовательность из двухбуквенного алфавита $\aleph = \{0, 1\}$. Максимально равновесный словарь толщины $q = 1$ будет содержать частоты символов $f_0 = f_1 = 1/2$. При этом слов длины $q = 2$ всего возможно четыре; для последовательности

$$\underbrace{0000 \dots 000}_{N/2} \underbrace{111 \dots 1111}_{N/2}$$

получим последовательность, для которой⁷ словарь толщины $q = 2$ будет иметь вид $f_{11} = f_{00} = 1/2$; $f_{01} = f_{10} = 0$. Для такого словаря информационная емкость (5) достигает максимума и составляет

$$\bar{S}_2 = \ln 2.$$

Аналогичную последовательность можно построить и для последовательности из алфавита мощности $M > 2$. Для этого случая информационная емкость (5) составляет $\bar{S}_2 = \ln M$.

Рассмотрим теперь случай словаря толщины $q = 3$ для последовательности из алфавита $\aleph = \{0, 1\}$. По-прежнему мы должны стремиться к тому, чтобы $f_{ij} = 1/4 \forall i, j \in \{0, 1\}$. Последовательность вида $\dots 0011001100110011 \dots$ обеспечивает требуемый словарь толщины $q = 2$ и для слов длины $q = 3$ дает словарь $f_{001} = f_{011} = f_{100} = f_{110} = 1/4$; $f_{000} = f_{010} = f_{101} = f_{111} = 0$. Информационная емкость (4) для этого случая равна $\bar{S}_3 = \ln 2$. В периодической последовательности $\dots (10111000)_n \dots$ частоты всех слов длины $q = 3$ совпадают и равны $f_w = 1/8$; в словаре толщины $q = 4$ ненулевую частоту⁸ также имеют всего восемь слов: 0001, 0010, 0101, 0111, 1000, 1011, 1100 и 1110. Легко видеть, что для такой последовательности $\bar{S}_4 = \ln 2$. Вообще говоря, для алфавита $\{0, 1\}$ всегда можно построить последовательность такого типа и $\bar{S}_q = \ln 2$ для любого q . Аналогично можно показать, что для произвольного алфавита \aleph : $|\aleph| = M$ предельное значение информационной емкости (4) будет близко к $\bar{S}_q \cong (M/2) \ln 2$. Данное обстоятельство делает информационную емкость очень важным инструментом для исследования статистических

⁷В пренебрежении краевыми эффектами.

⁸Частота этих слов равна $1/8$.

и информационных свойств символьных последовательностей, поскольку значения емкости (4) не зависят от толщины словарей q .

Выше говорилось (см. разд. 1.2), что для высокоупорядоченных последовательностей информационная емкость (4), (5) принимает значения, близкие к нулю. Примеры, построенные выше, показывают, что не для всякой упорядоченной последовательности такое утверждение справедливо; очевидно, оно справедливо лишь для последовательностей, аналогичных показанной на предыдущей странице, и словарей достаточной толщины. Тем не менее измерение информационной емкости помогает различать сложные, случайные и высокоупорядоченные последовательности. Для случайных, а также высокоупорядоченных последовательностей с высокой точностью выполняется соотношение $\bar{S}_q \cong 0$, а отклонения от нулевого значения порождены лишь конечностью последовательности.

Сравнение значений \bar{S}_q , наблюдаемое при различных q для одной и той же последовательности, позволяет выявить в ней скрытый порядок. На рис. 2 показано изменение информационной емкости, вычисленное для $q \leq q \leq 8$ для некоторых геномов бактерий. Здесь показаны кривые изменения информационной емкости, вычисленные для шести бактериальных геномов: *Chromobacterium violaceum* (идентификатор AE016825), *Desulfovibrio vulgaris* (идентификатор AE017285), *Geobacter sulfurreducens* (идентификатор AE017180), *Helicobacter hepaticus* (идентификатор AE017125), *Synechocystis sp.* (идентификатор BA000022) и *Thermoanaerobacter tengcongensis* (идентификатор AE008691); все геномы депонированы в EMBL-банке. Информационная емкость определялась для словарей толщины $1 \leq q \leq 9$. Информационная емкость символьной последовательности — еще один инструмент для изучения и анализа ее структурированности. Данные, представленные на рис. 2, показывают, что вычисление информационной емкости для символьных последовательностей, соответствующих естественным объектам (геномам в нашем случае), выявляет в них порядок. При этом данный порядок носит универсальный характер, поскольку определяется по всей последовательности целиком.

Другой возможный вариант выделения порядка в символьных последовательностях — сравнение значений информационной емкости для отдельных фрагментов последовательности, последовательно выделяемых в ней окном фиксированной длины L . Такое срав-

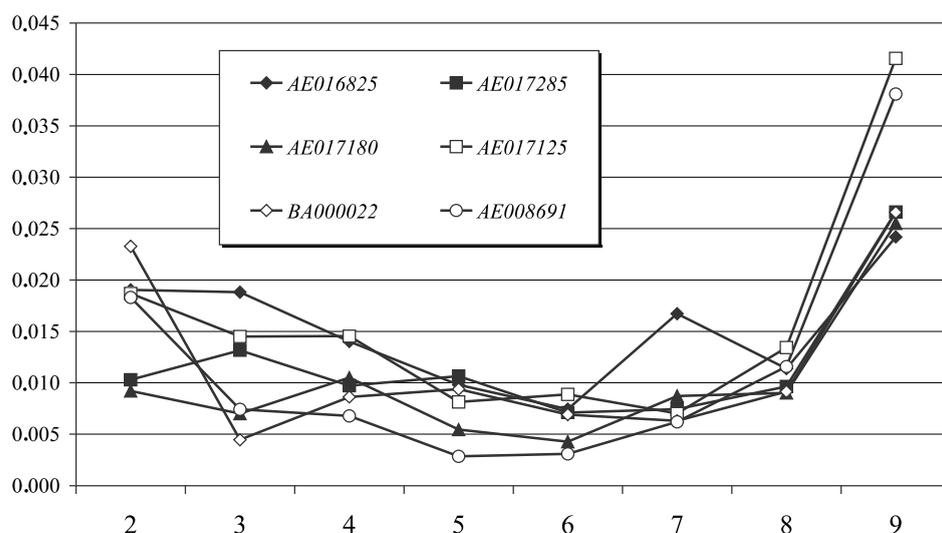


Рис. 2. Немонотонный характер изменения информационной емкости (4), (5) в геномах некоторых бактерий, свидетельствующий о существовании в них порядка.

нение позволяет строить информационный профиль последовательности, являющийся ее структурой.

Обратимся еще раз к определению условной энтропии (3). Очевидно, что наибольший вклад в эту сумму вносят те слова, для которых наблюдается наибольшее отличие реальной и ожидаемой частот. Такие слова следует считать **информационно значимыми**. Изучение таких слов позволяет выявить связь между различными последовательностями (см., например, [1, 2, 11]). Кроме того, большой интерес представляет изучение распределения таких слов вдоль по последовательности. Для биологических макромолекул наблюдается корреляция между их положением и функциональной ролью тех участков, в которые они попадают [12, 13], т. е. возможен переход от статистики к семантике символьных последовательностей. Еще одно продуктивное приложение идей, связанных с изучением распределения информационно значимых слов, возникает в задачах восстановления утраченных данных [14–16], однако подробное обсуждение этих вопросов выходит за рамки настоящей статьи.

Список литературы

- [1] GORBAN A.N., POPOVA T.G., SADOVSKY M.G., WUNSCH D.C. Information content of the frequency dictionaries, reconstruction, transformation and classification of dictionaries and genetic texts // Intelligent Engineering Systems through Artificial Neural Networks. Vol. 11: Smart Engineering System Design. N.Y.: ASME Press, 2001. P. 657–663.
- [2] GORBAN A.N., POPOVA T.G., SADOVSKY M.G. Classification of symbol sequences over thier frequency dictionaries: towards the connection between structure and natural taxonomy // Open Syst.& Inform. Dyn. 2000. Vol. 7, N 1. P. 1–17.
- [3] ГОРБАНЬ А.Н. Обход равновесия. Новосибирск: Наука, 1984. 268 с.
- [4] БУГАЕНКО Н.Н., ГОРБАНЬ А.Н., САДОВСКИЙ М.Г. Информационная емкость нуклеотидных последовательностей и их фрагментов // Биофизика. 1997. Т. 42, № 5. С. 1047–1053.
- [5] BUGAENKO N.N., GORBAN A.N., SADOVSKY M.G. Maximum entropy method in analysis of genetic text and measurement of its information content // Open Syst.& Inform.Dyn. 1998. Vol. 5, N 3. P. 265–278.
- [6] ГОРБАНЬ А.Н., ПОПОВА Т.Г., САДОВСКИЙ М.Г. Избыточность генетических текстов и мозаичная структура генома // Молекул. биология. 1994. Т. 28, № 2. С. 313–322.
- [7] САДОВСКИЙ М.Г. К вопросу об избыточности геномов вирусов и прокариот // Генетика. 2002. Т. 38, № 5. С. 695–701.
- [8] ПОПОВА Т.Г., САДОВСКИЙ М.Г. Избыточность генов уменьшается в результате сплайсинга // Молекуляр. биология. 1995. Т. 29, № 3. С. 500–506.
- [9] ПОПОВА Т.Г., САДОВСКИЙ М.Г. Интроны отличаются от экзонов по своей избыточности // Генетика. 1995. Т. 31, № 10. С. 1365–1369.
- [10] ГЕЛЬФАНД М.С. Компьютерный анализ последовательностей ДНК // Молекуляр. биология. 1998. Т. 32, № 1. С. 103–120.

- [11] ГОРБАНЬ А.Н., ПОПОВА Т.Г., САДОВСКИЙ М.Г. Классификация нуклеотидных последовательностей по частотным словарям обнаруживает связь между их структурой и таксономическим положением организмов // Журн. общей биологии. 2003. Т. 64, № 1. С. 65–77.
- [12] МАМОНОВА М.А., САДОВСКИЙ М.Г. Информационная ценность различных триплетов некоторых генетических систем // Журн. общей биологии. 2003. Т. 64, № 5. С. 421–433.
- [13] МАМОНОВА М.А., САДОВСКИЙ М.Г. К построению статистической семантики геномов // Матер. 11-й Всерос. конф. “Нейроинформатика и ее приложения”, Красноярск 3–5 окт. 2003 г. Красноярск, 2003. С. 105–106.
- [14] НЕМЕНЧИНСКАЯ Е.О., КОНДРАТЕНКО Ю.В., САДОВСКИЙ М.Г. Предварительные результаты в проблеме восстановления утерянных данных с помощью кинетической машины Кирдина // Вычисл. технологии. 2004. Т. 9, № 1. С. 42–57.
- [15] NEMENCHINSKAYA E.O., KONDRATENKO YU.V., SADOVSKY M.G. Entropy based approach to data loss reparation through the indeterminate fine-grained parallel computation / Open Systems & Information Dynamics. 2004. Vol. 11, N 2. P. 161–175.
- [16] GORBUNOVA E.O., KONDRATENKO YU.V., SADOVSKY M.G. Data loss reparation due to indeterminate fine-grained parallel computation // Lecture Notes in Computer Science. Springer-Verlag, Heidelberg. 2003. Vol. 2658. P. 794–801.

*Поступила в редакцию 2 апреля 2004 г.,
в переработанном виде — 11 января 2005 г.*