УДК 577.21

ОБ АСИМПТОТИЧЕСКОМ ПОВЕДЕНИИ ИНФОРМАЦИОННОЙ ЕМКОСТИ ГЕНОМОВ

М.Г.Садовский, А.С.Щепановский*

Рассмотрено поведение максимума информационной емкости генома в зависимости от его длины. Показано, что положение максимума в первую очередь определяется длиной символьной последовательности и мощностью алфавита. Рост длины символьной последовательности приводит к логарифмическому росту положения максимума информационной ценности.

Метод инвариантных многообразий [1–3], первоначально развитый для задач физической кинетики и анализа уравнений Больцмана, дал множество приложений в иных областях. Так, в задачах биофизики биологических макромолекул, геномике, биоинформатике этот метод позволил решить задачу об определении информационного содержания указанных макромолекул без привлечения дополнительных, априорных гипотез о структуре исходной последовательности [4–6]. Информационная емкость определяется как условная энтропия реального частотного словаря относительно восстановленного (т.е. содержащего наиболее ожидаемые продолжения более коротких слов).

Точнее, пусть $\mathbb{T}$ — последовательность длины N символов из алфавита $\aleph = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ (генетический текст, далее ГТ); любую подпоследовательность длины q символов, $1 \le q \le N$, будем называть словом. Список $\mathbb{W}$ всех слов, встречающихся в ГТ, будем называть его носителем. Если каждому элементу $\omega \in \mathbb{W}$ носителя $\mathbb{W}$ (слову) приписать число его копий n_ω в исследуемом ГТ, получится (конечный) словарь W. Заменив число копий n_ω на частоту $\dfrac{n_\omega}{N}$, получим частотный словарь $W(q)$. Здесь «переменная» q указывает на длину слов, которые содержит словарь; будем ее впредь называть толщиной словаря (подробности см. в [4 – 8]).

1. Определение информационной емкости символьной последовательности

Располагая частотным словарём $W(q)$, исследователь располагает и любыми частотными словарями меньшей толщины. Действительно, переход от $W(q)$ к $W(t)$, $t < q$ осуществляется суммированием частот слов, отличающихся первыми (либо последними) $q - t$ символами. Обратный переход существенно неоднозначен: по словарю $W(q)$ можно построить не один словарь толщины $W(q+1)$, а семейство словарей $\{W(q+1)\}$, каждый из которых порождает заданный словарь $W(q)$ при обратном переходе «вниз». Существует критическая толщина словарей d^*, для которой словарь $W(q+1)$ однозначно порождается по словарю $W(q)$, $q \ge d^*$. Определение d^* весьма просто: d^* на два символа длиннее наидлиннейшего повтора, встречающегося в $\mathbb{T}$. Это конструктивный (вычислимый) показатель, его смысл также достаточно прост и прозрачен: отношение d^* к логарифму длины $\mathbb{T}$ есть показатель избыточности последовательности. Некоторые биологические приложения этого показателя см. в [7 – 9].

Пусть $q < d^*$. Тогда порождение словаря $W(q+1)$ по словарю $W(q)$ не однозначно. Возникает задача выбора одного словаря $\widetilde{W}(q+1)$ из семейства $\{W(q+1)\}$, который бы не нес в себе никакой дополнительной, априорной информации. Иными словами, такой словарь должен быть максимально неопределенным; это означает, что его энтропия

$$S = -\sum_{i_1 i_2 \ldots i_{q-1} i_q i_{q+1}} \tilde{f}_{i_1 i_2 \ldots i_{q-1} i_q i_{q+1}} \times \ln \tilde{f}_{i_1 i_2 \ldots i_{q-1} i_q i_{q+1}} \tag{1}$$

должна быть максимальной при очевидных линейных ограничениях:

$$\sum_{i_1} \tilde{f}_{i_1 i_2 \ldots i_{q-1} i_q i_{q+1}} = \sum_{i_1} \tilde{f}_{i_2 i_3 \ldots i_{q-1} i_q i_{q+1} i_1} = f_{i_1 i_2 \ldots i_{q-1} i_q}\,. \tag{2}$$

*© М.Г.Садовский, Красноярский государственный университет, msad@icm.krasn.ru; А.С.Щепановский, Институт вычислительного моделирования СО РАН, suor@g-service.ru, 2006.

Здесь $i_1 i_2 \dots i_{q-1} i_q i_{q+1} = \omega_{q+1}$ — слово длины $q+1$, а i_j — символ, занимающий j-е место в нём; $i_j \in \aleph$. Максимум (1) при линейных ограничениях (2) дает явный вид для частот $\tilde{f}_{i_1 i_2 \dots i_{q-1} i_q i_{q+1}}$:

$$\tilde{f}_{i_1 i_2 \dots i_{q-1} i_q i_{q+1}} = \frac{f_{i_1 i_2 \dots i_{q-1} i_q} \times f_{i_2 i_3 \dots i_{q-1} i_q i_{q+1}}}{f_{i_2 i_3 \dots i_{q-1} i_q}} \,. \tag{3}$$

Выражение (3) зачастую ошибочно принимают за свидетельство того, что исходный ГТ $\mathbb{T}$ является реализацией Марковского процесса; это неверно — выражение (3) получено без каких-либо предположений о структуре $\mathbb{T}$. На самом деле, совпадение этого выражения с соответствующим выражением для Марковского процесса есть лишь отражение того факта, что Марковский процесс соответствующего порядка реализует гипотезу о наиболее вероятном продолжении слова ω_q длины q в слово ω_{q+1} длины $q+1$.

Пусть теперь в распоряжении исследователя имеется частотный словарь $W(q)$; его можно сравнить с восстановленным $\widetilde{W}(q)$ по словарю меньшей толщины $W(t)$, $t < q$. В частности, естественно восстанавливать $\widetilde{W}(q)$ (в силу (1 – 3)) по словарю $W(q-1)$. В этом случае формула (3) превращается в

$$\tilde{f}_{i_1 i_2 \dots i_{q-1} i_q} = \frac{f_{i_1 i_2 \dots i_{q-2} i_{q-1}} \times f_{i_2 i_3 \dots i_{q-1} i_q}}{f_{i_2 i_3 \dots i_{q-2} i_{q-1}}} \,. \tag{4}$$

Сравнение словарей $\widetilde{W}(q)$ и $W(q)$ может быть проведено различными способами (см., напр., [4–8]). Возможны и иные способы определения близости восстановленного и реального словарей. Естественной мерой здесь является условная энтропия $\overline{S}$ двух словарей (мера их взаимной неопределенности):

$$\overline{S} = \sum_{\omega} f_{i_1 i_2 \dots i_{q-1} i_q} \times \ln \left(\frac{f_{i_1 i_2 \dots i_{q-1} i_q}}{\tilde{f}_{i_1 i_2 \dots i_{q-1} i_q}} \right) . \tag{5}$$

Подставляя (4) в (5), получаем

$$\overline{S}_q = 2S_{q-1} - S_q - S_{q-2} \qquad \text{и} \qquad \overline{S}_2 = 2S_1 - S_2 \tag{6}$$

для случаев $q > 2$ и $q = 2$ соответственно. Здесь S_j — абсолютная энтропия частотного словаря $W(j)$ толщины j.

2. О положении максимума информационной емкости

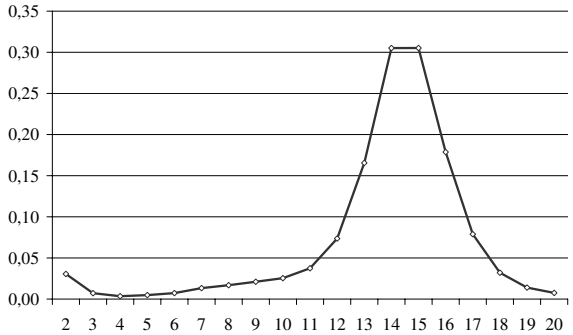

Рис. 1: Кривая изменения информационной емкости XIV хромосомы человека

Для любой конечной последовательности зависимость (6) от толщины словаря q является куполообразной; на рис. 1 показан вид зависимости (6) для хромосомы XIV человека; последовательность депонирована в EMBL банке, номер доступа AL954800. Длина этой хромосомы составляет 87 191 216 пар нуклеотидов. Такой вид зависимости $\overline{S}$ от q всегда обусловлен конечностью исследуемой последовательности. Собственно, задача настоящей работы состоит в определении зависимости положения максимума (6) от длины последовательности N.

Обратимся к формуле (5); наибольший вклад в сумму дают те слагаемые, для которых реальная и ожидаемая частоты отличаются больше всего; такие слова называются информационно значимыми [8 – 10]. Соответственно, нас будет интересовать вопрос о том, как возникают такие слова. Их появление обусловлено двумя факторами. Во-первых, они могут появляться благодаря особенностям структуры изучаемой последовательности; во-вторых, они могут появляться из-за конечности изучаемой последовательности. Действительно, общее число (возможных) слов с ростом толщины словаря растёт как $|\aleph|^q$, где q — длина слов. Число слов с ростом длины последовательности растёт линейно. Соответственно, увеличивается число слов, которые встречаются в единственном экземпляре либо вовсе не встречаются в изучаемой последовательности. Число тех слов, которые однозначно определяются комбинацией слов меньшей длины, также возрастает с ростом q. Найдётся

такая характерная длина слов, для которой число слов, которые наихудшим образом комбинируются из более коротких, будет самым большим. Именно для такой длины слов будет наблюдаться максимум (6).

Ещё один вариант ответа на вопрос о поведении максимума информационной емкости с ростом длины ГТ может быть получен следующим образом. Возьмём ГТ достаточно большой длины; в наших экспериментах бралась XIV хромосома человека. В рассматриваемом ГТ последовательно выделяются участки возрастающей длины; все эти участки начинались в первом нуклеотиде, а их длина увеличивалась в геометрической прогрессии, с показателем e (основанием натурального логарифма). Для каждого из участков вычислялись значения (6) для $2 \le q \le 20$. В табл. 1 приведены положения максимумов и их значения для возрастающих значений длины N_i участков ГТ.

Таблица 1. Положение максимума $L(N_i)$ условной энтропии (6) для участков ГТ различной длины

Длина	$L(N_i)$	$\overline{S}$		Длина	$L(N_i)$	$\overline{S}$
10000	8	0,422130667		1484144	11	0,358202806
27183	8	0,402108819		4034322	12	0,360193452
73891	9	0,410315733		10966424	13	0,353147132
200857	10	0,396716801		29809831	14	0,323291163
545986	11	0,371836355		81031522	14	0,309089936
Случайная последовательность; равные частоты						
10000	8	0,482229060		1484144	11	0,540420848
27183	8	0,529318812		4034322	12	0,535005878
73891	9	0,544634575		10966424	13	0,485369997
200857	10	0,505714103		29809831	13	0,518929479
545986	10	0,492694604		81031522	14	0,544835281
Случайная последовательность; частоты нуклеотидов						
10000	8	0,463117582		1484144	11	0,469917441
27183	8	0,470321576		4034322	12	0,481747327
73891	9	0,489570772		10966424	13	0,461666487
200857	10	0,478586691		29809831	13	0,435859660
545986	11	0,436520379		81031522	14	0,467382010

Для того чтобы определить влияние длины символьной последовательности на положение максимума информационной емкости (6), мы провели два вычислительных эксперимента. Значения информационной емкости (6) вычислялись для двух суррогатных последовательностей: первая была случайной нескоррелированной последовательностью из четырехбуквенного алфавита, в которой все буквы имели одинаковую частоту $f_\nu = 1/4$; вторая представляла собой также случайную нескоррелированную последовательность (также из четырехбуквенного алфавита), у которой вероятности появления отдельных букв совпадали с частотами отдельных нуклеотидов у хромосомы XIV человека. Длина каждой из суррогатных последовательностей также совпадала с длиной хромосомы XIV.

Результаты, приведены в табл. 1, показывают, что для реальной последовательности положение максимума информационной емкости (6) смещается в сторону бóльших значений q, а сами значения максимум уменьшаются. Однако характер изменения функции $\overline{S}_{\max}(q)$, вычисленной для последовательностей разной длины N_i, весьма сложен. Эта сложность обусловлена также тем, что само по себе положение максимума принимает лишь целые значения. Для того чтобы «сгладить» влияние такой дискретности, мы вычислили эффективную толщину словаря $\overline{q}$, на которой наблюдается соответствующий максимум (6).

Очевидно, что для каждого значения $\overline{S}_{\max}(q)$ имеется довольно широкое окно $\{N_k \le N < N_{k+1} \mid L(N) = k\}$ и переход от функции условной энтропии к паре *положение – значение ее максимума* приводит к потере существенной информации. Чтобы избежать этого, мы интерполировали окрестность максимума (6). В табл. 2 приведены положения максимумов (6) для хромосомы XIV человека $\tilde{L}_x(N)$ и случайной последовательности нуклеотидов $\tilde{L}_c(N)$, полученные квадратичной интерполяцией (6) по точкам $L(N)-1, L(N), L(N)+1$:

$$\tilde{L}(N) = q + \frac{\overline{S}_{q+1} - \overline{S}_{q-1}}{2(2\overline{S}_q - \overline{S}_{q+1} - \overline{S}_{q-1})}, \text{ где } q = L(N)\ . \tag{7}$$

Теоретические оценки положения максимума величины (6) представляют собой трудную задачу. Это связано с тем, что для получения таких оценок требуется принять какую-либо гипотезу о структуре исходного ГТ. Один из эффективных методов в таких случаях — вычислительный эксперимент. Рассмотрим семейство геномов возрастающей длины; вычислим для каждого из них положение максимума, а также его значение. Затем найдем отношение $u = t_{\max}/\ln N$ толщины словаря $t_{\max}$, на которой наблюдается максимум (6) к логарифму длины соответствующего генома; на рис. 2 показано изменение значения максимума (6) с увеличением u.

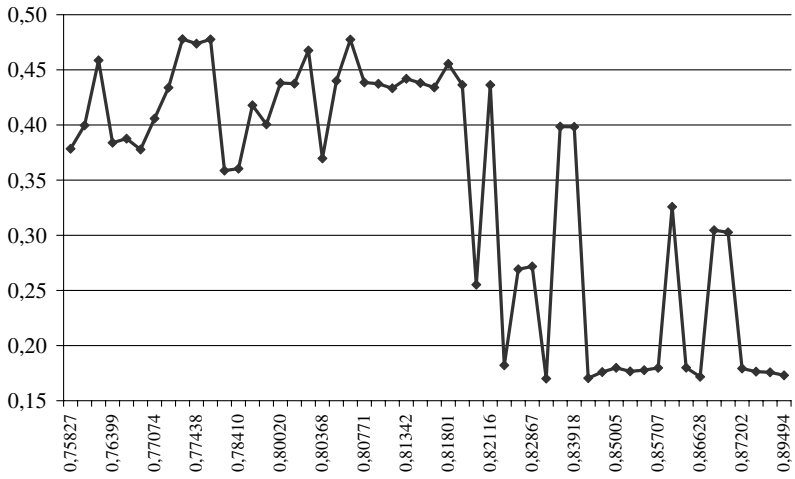

Рис. 2: Изменение максимума информационной емкости генома с длиной. По вертикальной оси отложены значения (6), по горизонтальной — значения u

Таблица 2. Положение максимума (7) для участков хромосомы $\tilde{L}_x$ и случайного текста $\tilde{L}_c$

Длина	$\tilde{L}_x(N)$	$\tilde{L}_c(N)$		Длина	$\tilde{L}_x(N)$	$\tilde{L}_c(N)$
10000	7.876895283	7.644376019		1484144	11.36660791	11.22752789
27183	8.384760818	8.304243174		4034322	12.12997580	12.01357986
73891	9.146163377	9.084036323		10966424	12.99726699	12.78311801
200857	10.02607134	9.851160454		29809831	13.79791461	13.42927356
545986	10.82968316	10.53494272		81031522	14.45156048	14.16417083

Построенное таким образом значение $\tilde{L}(N)$ согласовано с $L(N)$ в том смысле, что $L(N)$ — ближайшее целое к $\tilde{L}(N)$. Кроме того, $\tilde{L}$ непрерывна относительно значений $\overline{S}_q$, в том числе при дискретном изменении L.

Зависимость положения максимума $\overline{S}$ от длины $L(N_i)$ участка ГТ, по которому она определялась, является близкой к линейной. На отклонение этой зависимости от линейной оказывают влияние два существенных фактора: первый — структура самой последовательности, для которой определяется положение максимума (6) в зависимости от q; второй — это дискретный характер изменения q_M, $\overline{S}(q_M) \mapsto \max$. Соответственно, формулировка той или иной гипотезы о виде зависимости

$$\max \overline{S} = \max_q \overline{S}(q)$$

требует тщательного анализа тех данных, которые могут быть получены в ходе описанных выше вычислительных экспериментов.

Обратимся к рис. 2. Очевиден немонотонный характер поведения зависимости максимума информационной ёмкости от длины ГТ. Изменение положения максимума (6) носит дискретный характер; соответственно, ожидать гладкого поведения в изменении положения максимума (6) не приходится. Тем не менее, общая тенденция видна: увеличение длины ГТ приводит к уменьшению значения максимума информационной ёмкости и смещению его в сторону более длинных слов. Этот вывод очевиден: для бесконечно длинных символьных последовательностей (любой природы) максимум (6) будет наблюдаться для малых значений q. При этом полученные данные показывают, что положение максимума (6) растёт логарифмически, а соответствующие значения (6) убывают линейно с ростом $q_{\max}$.

Список литературы

[1] Gorban A.N. *Thermodynamic Equilibria and Extrema: Analysis of Attainability Regions and Partial Equilibria* / A.N.Gorban, B.M.Kaganovich, S.P.Filippov, A.V.Keiko, V.A.Shamansky,

I.A.Shirkalin. – Berlin, Heidelberg, New York: Springer, 2006. – 297 p.

[2] Gorban A.N. *Invariant Manifolds for Physical and Chemical Kinetics* / A.N.Gorban, I.V.Karlin. – Springer, Berlin, Heidelberg: Lect.Notes Phys. – 2005. – V. 660. – 493 p.

[3] Gorban A.N. *Constructive methods of invariant manifolds for kinetic problems* / A.N.Gorban, I.V.Karlin, A.Y.Zinovyev // Phys.Rep. – 2004. – V. 396. – P. 197-403.

[4] Bugaenko N.N. *Maximum entropy method in analysis of genetic text and measurement of its information content* / N.N.Bugaenko, A.N.Gorban, M.G.Sadovsky // Open Systems & Information Dynamics. – 1998. – V. 5, № 3. – P. 265-278.

[5] Gorban A.N. *Information content of the frequency dictionaries, reconstruction, transformation and classification of dictionaries and genetic texts* / A.N.Gorban, T.G.Popova, M.G.Sadovsky, D.C.Wunsch // Intelligent Engineering Systems through Artificial Neural Networks: – Smart Engineering System Design, N.-Y.: ASME Press. – 2001. – V. 11. – P.657-663.

[6] Горбань А.Н. *Информационная ёмкость нуклеотидных последовательностей и их фрагментов* / Н.Н.Бугаенко, А.Н.Горбань, М.Г.Садовский // Биофизика. – 1997. – Т. 42, № 5. – С. 1047-1053.

[7] Садовский М.Г. *К вопросу об избыточности геномов вирусов и прокариот* / М.Г.Садовский // Генетика. – 2002. – Т. 38, № 5. – С. 695-701.

[8] Мамонова М.А. *Информационная ценность различных триплетов некоторых генетических систем* / М.А.Мамонова, М.Г.Садовский // ЖОБ. – 2003. – Т. 64, № 5. – С. 421-433.

[9] Sadovsky M.G. *Comparison of real frequencies of strings vs. the expected ones reveals the information capacity of macromoleculae* / M.G.Sadovsky // Journal of Biological Physics. – 2003. – V. 29, № 1. – P. 23-38.

[10] Sadovsky M.G. *Information capacity of nucleotide sequences and its applications* / M.G.Sadovsky // Bulletin of Mathematical Biology. – 2006. – V. 68, № 2. – P. 156-178.

ON ASYMPTOTIC BEHAVIOUR OF GENOME INFORMATION CAPACITY

M.G.Sadovsky, A.S.Shchepanovsky

An information capacity pattern of a genome is studied. In particular, the location of the maximum of that latter in dependence of the length of a sequence was investigated. A pattern of the dependence is proposed and approved due to some calculations carried out over the real sequences.