Математическая биология и биоинформатика
2016. Т. 11. № 1. С. 114–126 . doi: 10.17537/2016.11.114

========== ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ ==========

УДК 51-76: 57.087

Энтропийный подход к построению меры символьного разнообразия слов и его применение к кластеризации геномов растений

©2016 Сметанин Ю.Г.[1], Ульянов М.В.[2,3*], Пестова А.С.[4]

[1]*ФИЦ «Информатика и управление» РАН, г. Москва*
[2]*МГУ им. М.В. Ломоносова, факультет ВМК, г. Москва*
[3]*ИПУ им В.А. Трапезникова РАН, г. Москва*
[4]*ФКН НИУ ВШЭ, г. Москва*

Аннотация. В статье рассматривается подход к анализу информации, представленной словами конечной длины над конечным алфавитом. Предложен метод построения меры символьного разнообразия слов, основанный на пиковых характеристиках функции энтропии сдвигов. Собственно функция энтропии сдвигов формализована на основе оператора сдвига один и понятия энтропии дискретных распределений. Приводятся модельный пример и результаты применения предложенной меры к кластеризации семейств растений на основе анализа геномов их представителей.

***Ключевые слова*:** энтропия сдвигов, мера символьного разнообразия, кластеризация геномов растений.

ВВЕДЕНИЕ

В данной статье рассмотрен подход к анализу информации, представленной с использованием символьного кодирования, при котором образы исследуемых объектов или процессов представляются словами над некоторым конечным алфавитом. Предлагаемый подход не требует числового кодирования образов объектов, зачастую приводящего к появлению ложной информации, вызванной исключительно особенностями использованного кода. Вместо этого предлагается проводить анализ символьных строк с целью выявления их эффективных характеристик в аспекте решаемых задач, например, в целях их кластеризации. Изучением таких символьных представлений занимается раздел современной дискретной математики – комбинаторика слов [1]. Предельный случай бесконечных последовательностей является предметом изучения в символической динамике [2]. В данной статье методы символической динамики адаптированы к случаю достаточно длинных, но конечных слов. На этой основе предложен метод количественной оценки разнообразия слов, опирающийся на использование информационной энтропии сдвигов.

Энтропия в информатике была впервые введена К. Шенноном для количественной оценки неопределенности имеющихся данных [3, 4]. Ее активное использование задержалось из-за недостаточной строгости изложения и недостаточной обоснованности результатов. Концепция получила развитие в работах А.Н. Колмогорова [5] и

*muljanov@mail.ru

А.Я. Хинчина [6]. Полный обзор результатов и изложение современного состояния дел в области информационной энтропии содержится в [7].

В символической динамике энтропия используется для характеризации разнообразия элементов пространств сдвигов в словах бесконечной длины. Пусть S_F – пространство сдвигов, то есть множество бесконечных последовательностей над конечным алфавитом A, не содержащих в качестве подслов конечных слов из заданного множества F. Энтропия этого пространства сдвигов [2]

$$H(S_F) = \lim_{n\to\infty} \frac{1}{n} \log|B_n|,$$

где B_n – множество подслов длины n, встречающихся в последовательностях из S_F. Значение энтропии позволяет находить отношения между пространствами сдвигов. А именно, если два сдвига S_{F_1} и S_{F_2} сопряжены, то есть переводятся друг в друга с помощью кодов скользящего блока, то их энтропии совпадают; если S_{F_1} сопряжен строгому подпространству S_{F_2}, то $H(S_{F_1}) < H(S_{F_2})$. Этот факт дает достаточное условие различия пространства сдвигов. В частности, для периодических бесконечных последовательностей $H = 0$, поэтому энтропия длинных периодических слов также близка к нулю. Поскольку наименьшим разнообразием подслов среди непериодических последовательностей обладают слова Штурма [1, 2], то, следовательно, слова, энтропия которых меньше энтропии слов Штурма, можно считать квазипериодическими.

Поскольку в данной статье речь идет о словах достаточно большой длины, методы символической динамики оказываются пригодными для получения полезных результатов. Отметим, что предлагаемая мера может быть вычислена и в том случае, когда слово, соответствующее образу, не известно полностью, а задано лишь некоторым списком своих подслов. Задача точной реконструкции слова по полному набору подслов фиксированной длины решена ранее [8]. Если набор неполный, реконструкция может оказаться неоднозначной, а значение энтропии – неточным. Тем не менее, по ее значению можно сделать полезные выводы о принадлежности слова априорно известным классам.

В биоинформатике и геномике методы комбинаторики слов используются достаточно широко. Известны их применения для оценки сложности нуклеотидных текстов. Например, рассматриваются такие оценки, как сложность по Вуттону-Федерхену и др. [9-13]. В целях сравнительного анализа регуляторных геномных последовательностей по возможному разнообразию продуцирования белков такие оценки эффективно использовались в [14]. В качестве еще одного примера использования методов комбинаторики слов в биоинформатике укажем задачу распознавания вторичной структуры белков [15], которая заключается в следующем. Белок можно представлять как одномерную последовательность аминокислот или как одномерную последовательность характерных локальных конфигураций. В настоящее время общепринятым является допущение, что первичная структура белка однозначно определяет вторичную. При этом задача определения вторичной структуры (структуры локальных конфигураций) формулируется как задача преобразования слов в алфавите аминокислот в слова над алфавитом локальных конфигураций с помощью кодов скользящего блока.

Авторы ставили своей целью построение меры, позволяющей не только оценивать сложность текстов в смысле разнообразия слов, но и строить кластерное пространство объектов, представленных словами над конечным алфавитом, полученными методами символьного кодирования. Этот подход ранее был успешно использован в задаче кластеризации временных рядов, закодированных символьными последовательностями [16, 17]. Цель данной статьи – формализовать и унифицировать процесс построения

меры разнообразия слов над конечным алфавитом для его применения в различных задачах при разных типах и качестве информации. В качестве конкретного примера применения в статье приводятся результаты по кластеризации семейств растений на основе анализа их геномов.

Объектом исследования в данной статье являются слова над конечным алфавитом, порожденные символьным кодированием образов исследуемых объектов, а предметом исследования – мера символьного разнообразия, отражающая особенности информационной энтропии.

ТЕРМИНОЛОГИЯ И ОБОЗНАЧЕНИЯ

Далее в тексте авторы будут использовать общепринятые в комбинаторике слов и символической динамике [1, 2] терминологию и обозначения, а так же специальные авторские обозначения, связанные с особенностью данной задачи. В теории формальных грамматик и языков термины «слово» и «строка» исторически считаются равноположенными. Например, известная задача символьного поиска формулируется как «Задача поиска подстроки в строке», но в проблематике комбинаторики слов [1] используется термин «слово», который авторы и будут использовать далее. Введем следующие обозначения:

$\Sigma = \{s_1, s_2, \ldots, s_m\}$ – конечный алфавит, s – произвольный символ алфавита;

Σ^k – k-ая декартова степень множества Σ;

$c \in \Sigma^k$ – k-элементный кортеж, $c = \left(s^{(1)}, s^{(2)}, \ldots, s^{(k)}\right)$, $s^{(i)} \in \Sigma$;

w – слово (над алфавитом) – последовательность символов алфавита;

$WD(c) = w$ – оператор порождения слова w над алфавитом Σ из кортежа c путем последовательной в порядке кортежа конкатенации символов

$$WD\left(\left(s^{(1)}, s^{(2)}, \ldots, s^{(k)}\right)\right) = s^{(1)}s^{(2)}\ldots s^{(k)};$$

$|w| = k$ – длина слова, понимаемая как мощность порождающего кортежа;

$L(\cdot)$ - оператор порождения множества слов, состоящих из символов алфавита Σ, действующий на множество кортежей посредством последовательного применения оператора $WD(\cdot)$ – $L(\cdot): L(C) = W$, где $C \subseteq \Sigma^*$ – множество кортежей, W – множество слов

$$L(C) = W = \{w \mid \forall c \in C\ w = WD(c)\}.$$

Будем говорить, что оператор $L(\cdot)$ порождает некоторый язык L над алфавитом Σ;

L_k – язык всех слов длины k над алфавитом Σ,

$$L_k = L\left(\Sigma^k\right) = \{w \mid |w| = k\};$$

$SW(w, i, l)$ – оператор выделения подслова v длины l в слове w, начиная с символа в позиции i. Пусть $|w| = k$, тогда оператор определен при $i + l - 1 \le k$:

$$SW(s_1 s_2 \ldots s_k, i, l) = v = s_i s_{i+1} \ldots s_{i+l-1};$$

$SH1(w, k)$ – оператор сдвига один по слову w с окном ширины k. Определенный при $|w| > k$ оператор порождает множество подслов длины k, с мощностью $|w| - k + 1$, выполняя сдвиг на единицу окна ширины k по слову w начиная с крайней левой позиции слова w:

$$SH1(w, k) = \left\{v_i \mid v_i = SW(w, i, k),\ i = \overline{1, |w| - k + 1}\right\}.$$

Оператор $SH1(w, k)$, очевидно, может порождать мультимножество

Математическая биология и биоинформатика. 2016. Т. 11. № 1. doi: 10.17537/2016.11.114

$$SH1(w,k)=\left\{v_i^{(c_i)} \mid i=\overline{1,l}\right\}, \sum_{i=1}^{l} c_i = |w|-k+1,$$

где c_i — кратность вхождения подслова v_i в мультимножество.

Например, для алфавита $\Sigma=\{a,b\}$ и слова $bbababa$ при окне ширины 4 имеем:

$$SH1(bbababa,4)=\{bbab, baba, abab, baba\}=\left\{bbab^{(1)}, baba^{(2)}, abab^{(1)}\right\}.$$

ЭНТРОПИЯ ДИСКРЕТНЫХ РАСПРЕДЕЛЕНИЙ С КОНЕЧНЫМ НОСИТЕЛЕМ

Рассмотрим классическую вероятностную модель с конечным носителем $M=\langle\Omega, P(\cdot)\rangle$, в которой вероятностное пространство Ω конечно – $\Omega=\{\omega_1,\omega_2,\ldots,\omega_k\}$, а вероятностная мера $P(\cdot)$ распределяет суммарную единицу вероятности между случайными событиями – $P(\omega_i)=p_i$. Тогда по определению [5] энтропия такого дискретного распределения определяется в виде

$$H_P \overset{def}{=\!=} -\sum_{i=1}^{k} p_i \log p_i . \qquad (1)$$

Индекс P у обозначения энтропии подчеркивает, что энтропия порождена вероятностной мерой $P(\cdot)$ и инвариантна по отношению к элементам вероятностного пространства. По сути, H_P есть функционал (1), действующий в классе дискретных распределений с конечным носителем. Отметим, что для этого класса распределений максимум энтропии достигается для равномерного распределения [7] $P(\omega_i)=1/|\Omega|=1/k$, характеризуя тем самым наибольшую непредсказуемость такой вероятностной модели. Выбрав основание логарифма равное мощности вероятностного пространства мы можем нормировать максимальное значение энтропии конечного дискретного распределения (1) в единицу, то есть при $P(\omega_i)=1/|\Omega|=1/k$ имеем

$$H_p = -\sum_{i=1}^{k} p_i \log_{|\Omega|} p_i = -\sum_{i=1}^{k} \frac{1}{k} \log_k \frac{1}{k} = 1 .$$

Отметим, что при любом другом распределении, отличном от равномерного, значение $0 \le H_P < 1$, а для вырожденного распределения $\exists\omega_i : P(\omega_i)=1,\ \forall\omega_j : j \ne i\ \ P(\omega_j)=0$ энтропия равна нулю.

ФУНКЦИЯ ОЦЕНКИ ЭНТРОПИИ СДВИГОВ

Предлагаемая мера символьного разнообразия слов для исследуемого слова w, $|w| = n$ основана на особенностях функции оценки энтропии сдвигов. Построение этой функции выполняется в два этапа. На первом этапе вычисляется оценка энтропии слова w для подслов фиксированной длины, а на втором этапе эта оценка обобщается на произвольные подслова, длина которых является аргументом необходимой нам функции.

Оценка энтропии слова w для подслов фиксированной длины. На этом этапе мы фиксируем алфавит Σ и длину подслова k. Множество всех слов длины k над алфавитом Σ есть $L_k = L(\Sigma^k)$. Пусть $U = L(\Sigma^k)$. Введем в рассмотрение формальное мультимножество

$$\tilde{U} = \left\{ u_i^{(0)} \mid i = \overline{1, |\Sigma|^k} \right\},$$

элементы которого (все слова длины k над алфавитом Σ) имеют нулевую кратность. Далее применим к исследуемому слову w оператор сдвига один с окном ширины k и получим порожденное оператором $SH1(w, k)$ мультимножество $\tilde{V}$:

$$\tilde{V} = SH1(w,k) = \left\{ v_i^{(c_i)} \mid i = \overline{1, l} \right\}.$$

Пусть m есть число позиций окна ширины k на слове w, отметим, что m равно сумме кратностей элементов мультимножества $\tilde{V}$:

$$m = \sum_{i=1}^{l} c_i = |w| - k + 1 = n - k + 1.$$

Построим объединенное мультимножество $\tilde{V} \cup \tilde{U}$, но, поскольку $\tilde{U}$ содержит все возможные слова длины k, то $\tilde{V} \cup \tilde{U}$ не будет содержать новых по отношению к $\tilde{U}$ элементов. Тем самым объединение мультимножеств приведет только к изменению кратностей некоторых, или, быть может, всех элементов из $\tilde{U}$. На этой основе построим вероятностную модель

$$M_k = \left\langle \Omega_k = \tilde{V} \cup \tilde{U},\ P_k(\cdot) \right\rangle,$$

где мощность $|\Omega_k| = |\Sigma|^k$, а вероятностная мера отражает частоту появления различных слов длины k в слове w на основе кратности элементов мультимножества $\tilde{V}$

$$P_k(\cdot): \begin{cases} p(\omega_i) = \dfrac{c_i}{m}, & \omega_i \in \tilde{V}; \\ p(\omega_i) = 0, & \omega_i \in \tilde{U} \setminus \tilde{V}. \end{cases}$$

Для полученной вероятностной модели с вероятностной мерой $P_k(\cdot)$ мы определяем нормированную энтропию, используя $|\Omega_k| = |\Sigma|^k$ в качестве основания логарифма в (1)

$$H_{P_k} = -\sum_{i=1}^{|\Sigma|^k} p_i \log_{|\Sigma|^k} p_i = -\sum_{i=1}^{m} \left(\frac{c_i}{m} \right) \log_{|\Sigma|^k} \left(\frac{c_i}{m} \right). \qquad (2)$$

Отметим, что значение $H_{P_k} = 0$ означает, что все подслова, порожденные оператором $SH1(w, k)$, одинаковы и, следовательно, состоят из одного и того же символа. Тем самым мы констатируем фундаментальное отсутствие разнообразия в исследуемом слове w. Просто показать, что значение $H_{P_k} = 1$ может быть получено только в случае совпадения множеств $\tilde{V}$ и $\tilde{U}$, при этом все слова в $\tilde{U}$ имеют одинаковую кратность, т.е. при равночастотности всех возможных подслов длины k в исследуемом слове w.

Построение функции оценки энтропии сдвигов. На втором этапе мы используем естественное расширение энтропии H_{P_k} путем введения функции $H(k) = H_{P_k}$, аргументом которой является длина подслова k, с областью определения: $1 \le k \le n$. Значения функции $H(k)$ при фиксированном значении аргумента k вычисляются по формуле (2) на основе вероятностной модели $M_k = \left\langle \Omega_k = \tilde{V} \cup \tilde{U},\ P_k(\cdot) \right\rangle$, полученной в результате применения оператора $SH1(w, k)$ к исходному слову w. Последовательный

Математическая биология и биоинформатика. 2016. Т. 11. № 1. doi: 10.17537/2016.11.114

перебор значений ширины окна k от 1 до n дает нам искомые значения $H(k)$. В соответствии с принятой в символической динамике терминологией [2] будем называть $H(k)$ функцией оценки энтропии сдвигов.

Заметим, что не зависимо от алфавита Σ и конкретного слова w значение $H(n) = 0$, поскольку мы наблюдаем всего одно слово с вероятностью один, при этом функция $H(k)$ не определена при $k > n$, поскольку в этом случае не определен оператор сдвига один. Значение $H(1)$ будет близко к единице, в случае, если в исследуемом слове частотная встречаемость символов алфавита будет приблизительно одинакова. Тем самым в целом функция $H(k)$ будет невозрастающей функцией на области ее определения – мы можем характеризовать ее как «убывающую по совокупности».

КОНЕЧНАЯ РАЗНОСТЬ ФУНКЦИИ ОЦЕНКИ ЭНТРОПИИ СДВИГОВ

Интерес представляет изучение характера убывания значений $H(k)$ с ростом аргумента. Поскольку функция $H(k)$ – убывающая по совокупности, рассмотрим конечную разность функции $H(k)$, взятую с обратным знаком (инверсная конечная разность):

$$\Delta H(k) = -(H(k+1) - H(k)) = H(k) - H(k+1), k = \overline{1, n-1}. \quad (3)$$

Мы используем инверсную конечную разность, с целью обеспечить положительность значений $\Delta H(k)$ для начальных значений аргумента. По определению функции $H(k)$ значения $\Delta H(k)$ ограничены, и $0 \le \Delta H(k) \le 1$, но поведение $\Delta H(k)$ при изменении k от единицы до $n-1$ может быть достаточно сложным. Предположим, что начальное значение $H(1) \approx 1$, т.е. символы алфавита имеют слабо отличающуюся частотную встречаемость. Если и значение $H(2) \approx 1$, то мы констатируем, что все возможные подслова длины два в исследуемом слове также имеют близкую частотную встречаемость и т.д. Тем самым близкие к нулю начальные значения $\Delta H(k)$, характеризует исследуемую символьную последовательность (слово) как последовательность, обладающую достаточно богатым разнообразием в области подслов малой длины.

Однако функция $H(k)$ не может долго «держать единицу». Определим пороговое значение $\hat{k}$, при котором теоретически функция оценка энтропии сдвигов еще может быть равной единице. В сдвигающемся окне ширины k при мощности алфавита кодирования $|\Sigma|$ может наблюдаться максимально $|\Sigma^k| = |\Sigma|^k$ различных подслов. Всего в слове длины n мы имеем $n-k+1$ позиций окна сдвига один. Тогда максимально возможная длина подслова, при которой еще можно наблюдать полное разнообразие подслов, определяется из уравнения $|\Sigma|^{\hat{k}} = n - \hat{k} + 1$, что с учетом целочисленности $\hat{k}$ приводит к уравнению $\hat{k} = \lfloor \log_{|\Sigma|}(n - \hat{k} + 1) \rfloor$. В предположении, что $\hat{k} << n$, пороговое значение $\hat{k} \approx \lfloor \log_{|\Sigma|} n \rfloor$.

В окне ширины $\hat{k}+1$ максимально возможное разнообразие слов в $|\Sigma|$ раз больше полного разнообразия в окне ширины $\hat{k}$. Поэтому мы ожидаем резкого падения значения функции $H(k)$ при $k = \hat{k} + 1$, и, следовательно, резкого скачка $\Delta H(k)$ даже для псевдослучайной последовательности символов в исследуемом слове, обладающего конечной длиной. Таким образом, наличие ярко выраженного максимума у функции $\Delta H(k)$ при $k < \hat{k}$ означает, что начиная с данного значения k разнообразие подслов резко уменьшилось, и исходное слово обладает определенными особенностями, например, регулярностью или периодичностью.

Математическая биология и биоинформатика. 2016. Т. 11. № 1. doi: 10.17537/2016.11.114

МОДЕЛЬНЫЙ ПРИМЕР

В качестве модельного примера приведем (см. табл. 1) значения функции энтропии сдвигов $H(k)$ и ее инверсной конечной разности $\Delta H(k)$, вычисленные по формулам (2) и (3) для периодического слова $w=(abbaab)_4, |w|=24$ в алфавите $\Sigma=\{a,b\}$.

Таблица 1. Значения функции оценки энтропии сдвигов и ее конечной разности для модельного слова

k	$H(k)$	$\Delta H(k)$
1	1.000	0.035
2	0.965	0.107
3	0.858	0.215
4	0.643	0.128
5	0.514	0.085
6	0.429	0.060
7	0.369	0.048
8	0.321	

Соответствующие графики приведены на рисунках 1 и 2. Обе функции являются функциями целочисленного аргумента, но мы показываем их как кусочно-линейные для наглядности отображения тенденций.

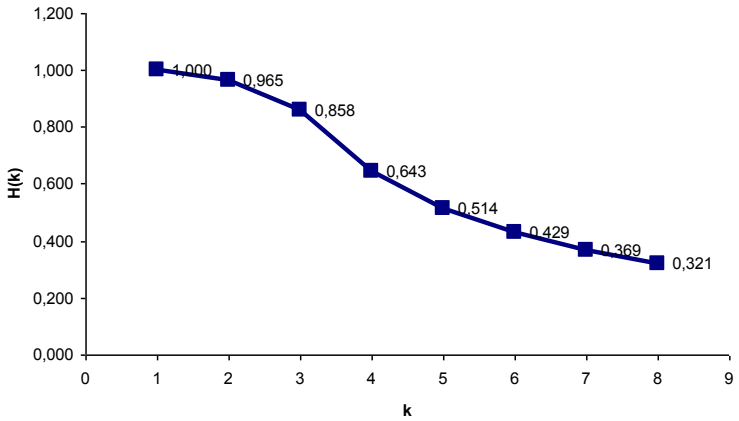

Рис. 1. График функции оценки энтропии сдвигов $H(k)$ для модельного слова.

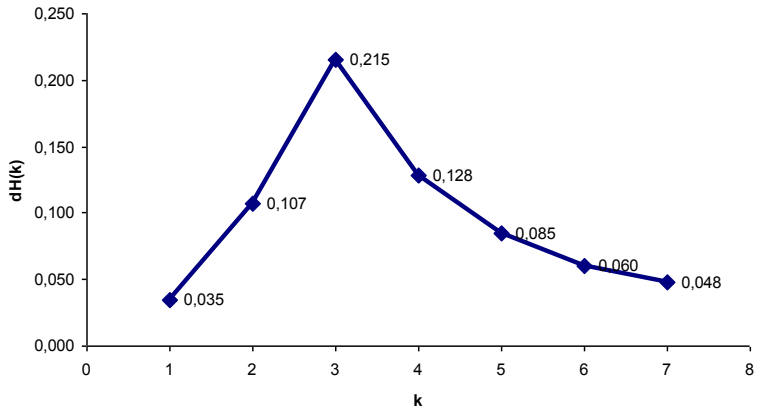

Рис. 2. График функции инверсной конечной разности $\Delta H(k)$ для модельного слова.

Математическая биология и биоинформатика. 2016. Т. 11. № 1. doi: 10.17537/2016.11.114

Заметим, что, поскольку модельное слово имеет период шесть, то, начиная с окна ширины три, мы наблюдаем всего шесть различных подслов в окне. Поскольку $\Sigma = \{a, b\}$, то полное разнообразие подслов увеличивается вдвое при увеличении ширины окна на единицу. Таким образом, при $k = 3$ мы наблюдаем шесть подслов из восьми возможных, а при $k = 4$ – тоже шесть подслов, но уже из 16 возможных, и максимум функции $\Delta H(k)$ фиксирует потерю разнообразия при переходе от окна ширины три к окну ширины четыре. Для этой строки $\hat{k} = \lfloor \log_2 24 \rfloor = 4$, но в силу периодичности в шесть символов мы не наблюдаем даже восьми различных подслов для окна ширины три, а при $k = 4$ значение $H(4)$ уже значительно уменьшается до 0.643, что и обеспечивает максимум $\Delta H(k) = 0.215$ при $k = 3$.

МЕРА СИМВОЛЬНОГО РАЗНООБРАЗИЯ СЛОВ

Очевидно, что помимо значения ширины окна k в операторе сдвига один, при котором достигается максимум функции $\Delta H(k)$, представляет интерес и само пиковое значение этой функции. На основе проведенного исследования функции $\Delta H(k)$ авторы вводят два варианта меры символьного разнообразия слов.

Вариант 1 (абсолютные значения). В этом варианте мы вводим меру символьного разнообразия слов в виде функционала

$$\mu_s(w) : W \to \aleph \times (0,1),$$

отображающего исследуемое слово в пару числовых значений – аргумент максимума $\Delta H(k)$ и само пиковое значение в точке максимума. Формально, пусть

$$k^* = \arg\max_{1 \le k \le n} \Delta H(k),$$

тогда k^* есть аргумент максимального значения $\Delta H(k)$, $\Delta H(k^*)$ – собственно максимальное значение, и

$$\mu_s(w) = \left(k^*, \Delta H\left(k^*\right)\right). \quad (4)$$

Вариант 2 (относительные значения). Может быть рассмотрен также и нормированный вариант, в котором первый компонент меры вычисляется в виде отношения значения k^* к максимально возможной ширине окна, сохраняющей полное разнообразие подслов – $\hat{k}$, тогда

$$\mu_s(w) : W \to (0,1) \times (0,1),\ \mu_s(w) = \left(k^* / \hat{k}, \Delta H\left(k^*\right)\right). \quad (5)$$

Отношение $k^* / \hat{k}$ нормировано в [0;1], а его малые значения свидетельствуют о раннем наступлении потери разнообразия и большей простоте исследуемого слова. Заметим, что в соответствии с (5) для периодических слов с малым периодом с ростом длины слова (числом наблюдаемых периодов) значение $k^* / \hat{k}$ будет уменьшаться, поскольку значение $\hat{k}$ зависит от длины слова. Это соответствует разумному предположению о том, что с ростом числа наблюдаемых периодов для периодического слова символьное разнообразие очевидно уменьшается.

Введенная мера может быть интерпретирована и в качестве координат точки исследуемого объекта в соответствующем кластерном пространстве с осями k^* и $\Delta H\left(k^*\right)$. При больших длинах сравниваемых слов, и не очень больших различиях в их длине, в частности для представлений геномов в алфавите аминокислот, бóльшие значения k^* могут быть интерпретированы, например, как характеристики большей

выживаемости исследуемых объектов или возможности их приспособления к новым условиям.

ПОДХОДЫ К БЫСТРОМУ ВЫЧИСЛЕНИЮ $\mu_s(w)$

На первый взгляд, вычисления по формуле (2) не представляют особых алгоритмических сложностей. Для рассматриваемого значения ширины окна k вводится произвольная нумерация всех возможных подслов $i=\overline{1,|\Sigma|^k}$ и счетчики числа подслов c_i, которые изначально обнуляются. Позиционированное на первом шаге в начало анализируемого слова w длины n, окно ширины k сдвигается каждый раз на один символ. Таким образом, мы имеем $n-m+1$ позиций окна, и для каждого его положения распознается подслово, полученное в окне. Если мы наблюдаем в текущей позиции окна подслово, имеющее номер i в принятой нумерации, то значение счетчика c_i увеличивается на единицу. По полученным значениям c_i по формуле (2) рассчитывается значение функции оценки энтропии сдвигов для данного значения k. В реализации нам понадобится «контейнер» для хранения каждой подстроки и значения соответствующего счетчика. Изменяя значение k в интересующем нас диапазоне значений, мы получаем необходимые для расчета $\mu_s(w)$ значения $H(k)$ для анализируемого слова w.

Но, поскольку для достаточно длинных строк, а длины геномов имеют порядок $n\approx10^6-10^7$, значение $\hat{k}\approx10-12$, то для корректной идентификации значения k^* ширина окна должна быть на несколько единиц больше. В этом случае при $k=14$ и четырехсимвольном алфавите мы получаем $4^{14}=268435456$ различных подстрок. Работа прямым перебором с массивом такой длины очевидно приводит к значительным ресурсным затратам как по памяти, так и по времени, и мы сталкиваемся с проблемой обработки больших данных. Приведем для сравнения экспериментальные данные непосредственной реализации такого подхода по времени расчета $\mu_s(w)$. Машина, на которой проводились тесты, имеет следующие характеристики: Intel® Core™ i5-3317U CPU @ с частотой 1.70 GHz. Тестируемый геном состоит из 155 296 символов (геном *Gordonia bronchialis*). Функция энтропии строк вычислялась для значений ширины окна k от 1 до 12. Время работы программной реализации данного алгоритма на языке C# составило 2.47 минуты, что неприемлемо.

Возможное повышение временной эффективности связано со специальной организацией хранения массива подстрок. Из двух возможных вариантов – использование различных модификаций деревьев или хеш-таблиц, авторами был выбран второй вариант – модель хеширования с коллизиями. В этом алгоритме мы не храним все возможные подстроки, а размещаем только наблюдаемые в окне сдвига один подстроки в хеш-таблице с коллизиями. Наглядное представление хеш-таблицы приведено на рисунке 3. Индексы в хеш-таблице определяются на основе вычисления хеш-функции от наблюдаемой в окне подстроки. Отметим, что выбор значения M – размерности хеш-таблицы, также влияет на ресурсную эффективность алгоритма.

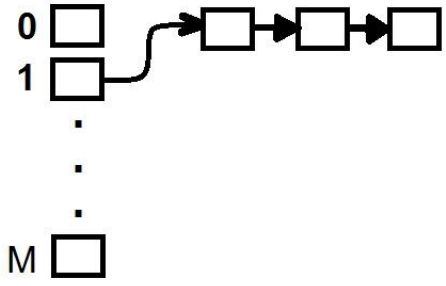

Рис. 3. Представление хеш-таблицы.

Авторами была принята следующая бинарная система кодирования для символов алфавита геномов – A = 00, T = 01, G = 10, C = 11. При этом наблюдаемое в окне ширины k подслово в принятой системе кодирования интерпретируется, как целое неотрицательное число в двоичной системе счисления длиной $2k$ бит. Пусть x – наблюдаемое число, тогда функция хеширования, выбранная авторами на основе ряда экспериментов, имеет вид [18]

$$h(x) = (x^2 + x + 1) \bmod(q).$$

Паллиативное решение, основанное как на оценке коэффициента заполнения хеш-таблицы (в целях разумной экономии памяти), так и на оценке среднего ожидаемого числа коллизий (в целях повышения временной эффективности), привело к выбору модуля $q = k^4 + 1$, что доставляет максимум ожидаемых коллизий порядка 50 при ширине окна в 10 символов для длины строки порядка 500 000. Результат теста – время работы программной реализации алгоритма, использующего хеширование, на языке Java для того же тестового генома (геном *Gordonia bronchialis*, 155296 символов) составило всего 1.827 секунды. Именно программная реализация этого алгоритма и была использована для дальнейших экспериментов с геномами растений.

ПРИМЕНЕНИЕ ПРЕДЛОЖЕННОЙ МЕРЫ ДЛЯ АНАЛИЗА ГЕНОМОВ РАЗЛИЧНЫХ СЕМЕЙСТВ РАСТЕНИЙ

В качестве содержательного примера использования предложенной меры $\mu_s(w)$ приведем данные по кластеризации семейств растений на основе анализа их геномов. Основная гипотеза, которая лежала в основе выбора геномов – обнаружение взаимосвязи введенной пиковой характеристики с биологическими характеристиками самих геномов. Наиболее успешной оказалась идея применения $\mu_s(w)$ к определению принадлежности исследуемых геномов разным семействам растений. В качестве исследуемых семейств были выбраны семейство Пасленовых и семейство Капустных (Крестоцветных). Мы рассмотрели по пять характерных представителей этих семейств. Названия растений и длины исследованных геномов приведены в таблице 2.

Поскольку $\hat{k} \approx \lfloor \log_{|\Sigma|} n \rfloor$, а значение $k^* \le \hat{k}$ то для определения $\mu_s(w)$ достаточно вычислить функцию оценки энтропии сдвигов для значений ширины окна несколько превышающих $\hat{k}$. В данном эксперименте значения $H(k)$ вычислялись на основе хеш-функции с коллизиями в диапазоне от 1 до $2\lfloor \log_4 n \rfloor$, где n – длина анализируемого слова w.

Таблица 2. Данные об использованных геномах растений

Геном	Банк данных - источник генома	Уникальный идентификатор	Длина генома
Solanum nigrum – паслён черный	NCBI	NC_028070	155432
Solanum lycopersicum – томат	NCBI	AC171728	188412
Solanum melongena – баклажан	NCBI	DF357214	448289
Solanum tuberosum – картофель	ENA	JH137862	2045796
Solanum chilense – дикий томат	NCBI	KP117021	155528
Brassica rapa – репа	NCBI	AC155342	151550
Raphanus sativus – редька посевная	NCBI	AB694744	258426
Brassica napus – рапс	ENA	LK032108	686937
Brassica oleracea – капуста огородная	NCBI	AC183495	356505
Brassica juncea – горчица сарептская	NCBI	JF920288	219766

В ходе работы использовались данные из геномных банков данных NCBI [19], ENA [20] и DDBJ [21], более подробная информация приведена в таблице 2.

В качестве примера приведем рассчитанные по формуле (2) значения функции оценки энтропии сдвигов $H(k)$ для генома паслена черного. Соответствующий график показан на рисунке 4.

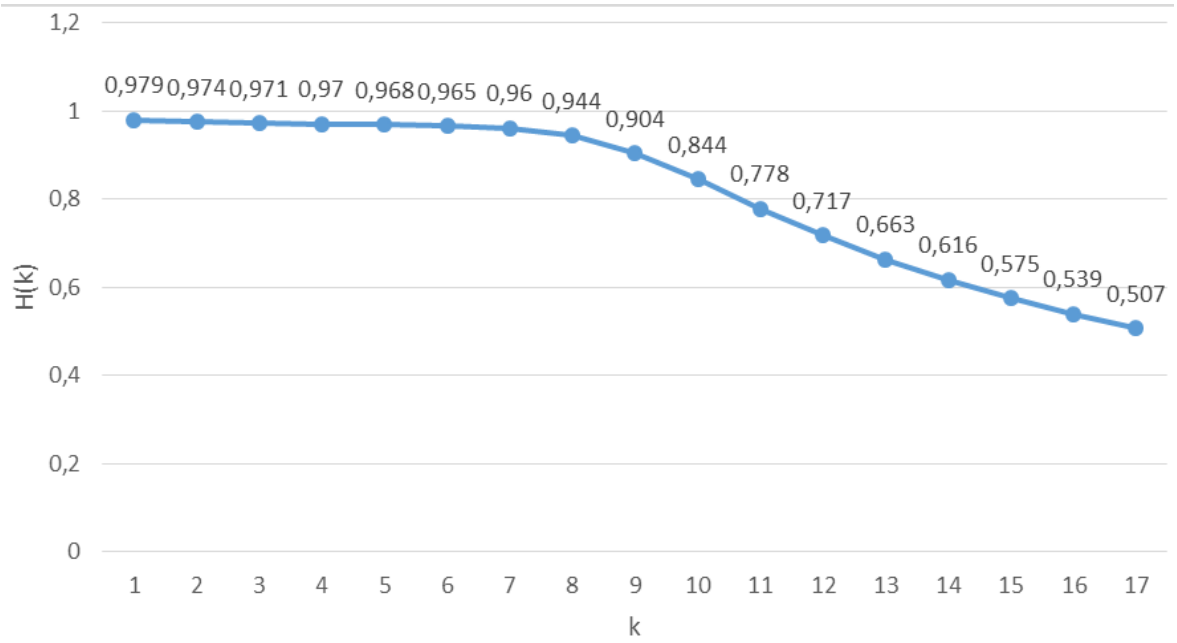

Рис. 4. График функции $H(k)$ для генома паслена черного.

Результаты расчетов значений $\mu_s(w)$ по формуле (4) для всех 10 геномов сведены в таблицу 3, значения $\Delta H\left(k^*\right)$ даны с точностью до шестого знака после запятой.

Таблица 3. Значения меры символьного разнообразия $\mu_s(w)$ для геномов растений

Семейство и вид	$\mu_s(w)$ – компонент k^*	$\mu_s(w)$ – компонент $\Delta H\left(k^*\right)$
Пасленовые		
Solanum nigrum – паслён черный	9	0,065721
Solanum lycopersicum – томат	9	0,065755
Solanum melongena – баклажан	9	0,064565
Solanum tuberosum – картофель	9	0,065727
Solanum chilense – дикий томат	9	0,065742
Капустные		
Brassica rapa – репа	10	0,057147
Raphanus sativus – редька посевная	10	0,059978
Brassica napus – рапс	10	0,057667
Brassica oleracea – капуста огородная	10	0,057370
Brassica juncea – горчица сарептская	10	0,058373

Далее мы интерпретировали значения меры $\mu_s\left(w\right)=\left(k^*,\Delta H\left(k^*\right)\right)$ как координаты точки в пространстве кластеризации. Результаты показаны на рисунке 5. Исследованные геномы растений четко распределились в два кластера,

соответствующие исходным семействам, причем различия наблюдаются как в значениях k^*, так и в значениях $\Delta H\left(k^*\right)$.

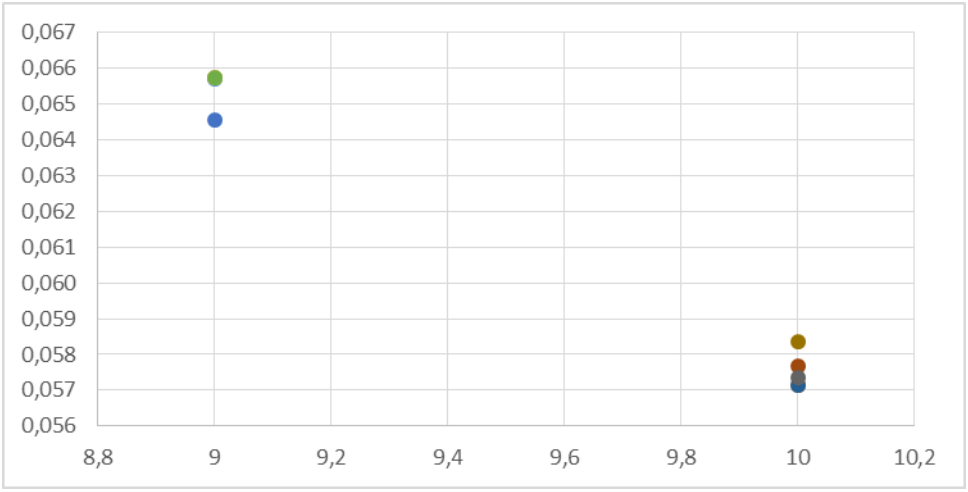

Рис. 5. Кластерное пространство геномов растений из 2-х семейств на основе значений меры символьного разнообразия слов $\mu_s(w)$.

В позиции $k^* = 9$, которая соответствует семейству пасленовых, четыре значения $\Delta H\left(k^*\right)$ настолько близки друг к другу, что отображаются на рис. 5 одной точкой (см. табл. 3).

ЗАКЛЮЧЕНИЕ

В статье предложен метод построения меры символьного разнообразия слов, основанный на пиковых характеристиках функции энтропии сдвигов. Собственно функция энтропии сдвигов формализована на основе оператора сдвига один и понятия энтропии дискретных распределений. Детальное исследование поведения этой функции позволило выявить ее характерную особенность – существование такого значения аргумента, при котором наблюдается максимум ее конечной разности. На этой основе введена мера символьного разнообразия слов над конечным алфавитом. Обсуждены некоторые алгоритмические особенности реализации вычислений предложенной меры с использованием хеш-таблиц. Показано, что предложенная мера символьного разнообразия слов применима к задаче кластеризации семейств растений на основе анализа их геномов, в частности для различения растений из семейств Пасленовых и Капустных (Крестоцветных). Авторы выражают надежду на то, что предложенный аппарат найдет применение для решения задач, связанных с анализом объектов, описываемых словами конечной длины над конечным алфавитом, в частности, в современной биоинформатике для анализа геномов.

Работа выполнена при поддержке гранта РФФИ 15-07-04112А.

СПИСОК ЛИТЕРАТУРЫ

1. Lothaire M. *Algebraic Combinatorics of Words*. Cambridge (UK): Cambridge Univ. Press, 2002. 455 p.
2. Lind D., Marcus B. *An introduction to symbolic dynamics and coding*. Cambridge (UK): Cambridge Univ. Press, 1995. 495 p.
3. Shannon C.E. A mathematical theory of communication. *Bell Syst. Techn. Journ*. 1948. V. XXVII. № 3 P. 379–423.

4. Shannon C.E. A mathematical theory of communication. *Bell Syst. Techn. Journ*. 1948. V. XXVII. № 4. P. 623–656.
5. Колмогоров А.Н. Общая теория динамических систем и классическая механика. В: *Международный математический конгресс в Амстердаме 1954 г*: обзорные доклады. Под ред. Фомина С.В. М.: Изд-во АН СССР, 1961. С. 187–208.
6. Хинчин А.Я. Понятие энтропии в теории вероятностей. *Успехи математических наук*. 1953. Т. 3. № 55. С. 3–20.
7. Мартин Н., Ингленд Дж. *Математическая теория энтропии*. М.: Мир, 1988. 350 с.
8. Smetanin Y.G., Ulyanov M.V. Reconstruction of a Word from a Finite Set of its Subwords under the unit Shift Hypothesis. I. Reconstruction without for Bidden Words. *Cybernetics and Systems Analysis*. 2014. V. 50. No. 1. P. 148–156.
9. Wootton J.C., Federhen S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol*. 1996. V. 266. P. 554–571.
10. Gusev V.D., Kulichkov V.A., Chupakhina O.M.Y. Complexity analysis of genomes. I. Complexity and classification methods of detected structural regularities. *Mol. Biol.* (*Mosk*). 1991. V. 25. No. 3. C. 825-834.
11. Gusev V.D., Kulichkov V.A., Chupakhina O.M. The Lempel-Ziv complexity and local structure analysis of genomes. *Biosystems*. 1993. V. 30. No. 1-3. C. 183-200.
12. Kislyuk O.S., Borovina T.A., Nazipova N.N. Estimation of Redundancy of Genetic Texts by the High Frequency Component of the *l*-Gram Graph. *Biophysics*. 1999. V. 44. No. 4. P. 621-630.
13. Troyanskaya O.G., Arbell O., Koren Y., Landau G. M., Bolshoy A. Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity. *Bioinformatics*. 2002. V. 18. No. 5. P. 679–688.
14. Орлов Ю.Л. *Анализ регуляторных геномных последовательностей с помощью компьютерных методов оценок сложности генетических текстов*: дисс. на соискание уч. ст. канд. биол. наук. Новосибирск, 2004. 148 с.
15. Рудаков К.В., Торшин И.Ю. Об отборе информативных значений признаков на базе критериев разрешимости в задаче распознавания вторичной структуры белка. *ДАН*. 2011. Т. 441. № 1. С. 24–28.
16. Сметанин Ю.Г, Ульянов М.В. Подход к определению характеристик колмогоровской сложности временных рядов на основе символьных описаний. *Бизнес-информатика*. 2013. № 2(24). С. 49–54.
17. Сметанин Ю.Г., Ульянов М.В. Мера символьного разнообразия: подход комбинаторики слов к определению обобщенных характеристик временных рядов. *Бизнес-информатика*. 2014. № 3(29). С. 40–48.
18. Кормен Т., Лейзерсон Ч., Ривест Р., Штайн К. *Алгоритмы: построение и анализ*. М.: Издательский дом «Вильямс», 2005. 1296 с.
19. GenBank. URL: http://www.ncbi.nlm.nih.gov/genbank/ (дата обращения: 20.03.2016).
20. European Nucleotide Archive. URL: http://www.ebi.ac.uk/ena (дата обращения: 20.03.2016).
21. DNA Data Bank of Japan. URL: http://www.ddbj.nig.ac.jp/ (дата обращения: 20.03.2016).

Материал поступил в редакцию 05.04.2016, опубликован 25.05.2016.