ПРЕДВАРИТЕЛЬНЫЕ РЕЗУЛЬТАТЫ В ПРОБЛЕМЕ ВОССТАНОВЛЕНИЯ УТЕРЯННЫХ ДАННЫХ С ПОМОЩЬЮ КИНЕТИЧЕСКОЙ МАШИНЫ КИРДИНА

Е. О. Неменчинская
Санкт-Петербургский государственный политехнический университет
Ю. В. Кондратенко
Красноярский государственный университет, Россия
М. Г. Садовский
Институт биофизики СО РАН, Красноярск, Россия
e-mail: eka_n@mail.ru, uvenal@ktk.ru

Preliminary results of the data loss recovery developed over the frequency dictionary, which are developed over the sequence fragments, are presented. Two criteria are suggested to choose the best recovery. Efficiency of the proposed imitator is proven.

Введение

Проблема восстановления утерянных данных актуальна как для фундаментальных, так и для прикладных областей науки. Результаты восстановления существенно зависят от способа, которым оно производится, и характера самих данных и утраты. Очевидно, что некоторые утраты данных не могут быть восстановлены никакими разумными способами. Рассмотрим какой-нибудь текст, записанный в естественном языке с алфавитной системой письма. Пусть в распоряжении исследователя имеются разрозненные фрагменты текста, в отношении которых можно наверняка утверждать, что они принадлежат одному и тому же тексту, однако алфавит этого языка исследователю неизвестен. Если утерянные части текста содержали какую-то букву алфавита, не встречающуюся во фрагментах, имеющихся в распоряжении исследователя, то почти наверняка можно утверждать, что он (и, соответственно, слова, его содержащие) не будут восстановлены никогда.

Методы восстановления утерянных данных обычно опираются на знание различных свойств данных, которые выходят за пределы тех знаний, которые могут быть извлечены из рассмотрения исключительно имеющихся в распоряжении исследователя фрагментов этих данных. Как правило, восстановление данных носит комплексный характер, и для этого привлекаются самые разные сведения о них, в том числе и информация, не содержащаяся непосредственно в этих данных (например, способы интерпретации или выявления

значения для случая языковых данных) (см. также [1, 2]). Представляет интерес постановка проблемы восстановления утерянных данных в весьма узкой формулировке, которая тем не менее допускает максимальную общность и строгость. Один из возможных вариантов рассмотрен в настоящей статье.

В качестве данных будем рассматривать конечные символьные последовательности. Будем полагать, что алфавит, в котором записаны изучаемые последовательности, заранее известен. Теоретически любые данные можно свести к этой форме; мы не будем подробно останавливаться на обсуждении этой стороны проблемы. Отсутствие части такой последовательности будем рассматривать как потерю данных. В рамках настоящей работы будем полагать, что длина утерянной части последовательности известна, а сама утерянная часть представляет собой связный диапазон. В такой постановке проблема весьма актуальна для самых различных областей знания — от теории передачи данных до молекулярной биологии, где возникает необходимость состыковки больших фрагментов секвенированных генетических текстов, принадлежащих какому-либо организму, причем такие секвенированные участки не пересекаются между собой [3 – 5].

Решение указанной задачи требует развития алгоритмов заполнения пробелов в символьных последовательностях. В нестрогом изложении принцип заполнения заключается в следующем: заполнять пробелы надо таким образом, чтобы последовательность, получающаяся после заполнения (восстановленная), была наиболее похожа на те части последовательности, которые имеются в распоряжении исследователя, однако требуется сделать это так, чтобы восстановленная последовательность несла в себе минимум дополнительной информации [6]. Такой принцип имеет две формулировки: принцип максимума энтропии пополненного частотного словаря (для случая восстановления по опорному частотному словарю) и минимума условной энтропии (для случая восстановления по полному частотному словарю). Перейдем теперь к строгим формулировкам и точным утверждениям. В рамках настоящей работы представлены предварительные данные по восстановлению утерянных данных в рамках первого из двух способов восстановления (принципа максимума энтропии пополненного частотного словаря).

1. Критерии заполнения лакун в последовательности

Рассмотрим конечные последовательности из конечного алфавита Ω; всюду далее в качестве иллюстрации будет рассматриваться алфавит мощности 4. Пусть N — длина всей восстанавливаемой последовательности:

$$N = N_1 + L + N_2,$$

где N_1 и N_2 — длины известных частей последовательности; L — длина участка, который необходимо восстановить. Словом длины q будем называть любую связную последовательность этой длины, составленную из символов алфавита Ω. Опорным частотным словарем W толщины q будем называть список всех слов длины q, встречающихся в тех частях последовательности, которые доступны исследователю, с указанием частот этих слов f_w. Пополненным частотным словарем $\overline{W}$ будем называть частотный словарь (толщины q), составленный по той последовательности, которая возникает в результате заполнения пробела. Наконец, полным частотным словарем W^* будем называть частотный словарь толщины q, содержащий все возможные слова из алфавита Ω. Очевидно соотношение

$$1 \le q \le \min\{N_1, N_2\}.$$

Левой (соответственно, правой) опорой длины t, $0 \le t \le q-1$, будем называть слово этой длины, расположенное сразу слева (соответственно, сразу справа) от лакуны. Тем самым в зависимости от величины t восстанавливаемая часть имеет длину $L+2t$ при том условии, что первые и последние t символов фиксированные. Отметим, что не всякое сочетание L, q и t совместимо.

Построить заполнение лакуны означает найти цепочку из $L-q+t$ слов длины q

$$w_1, w_2, w_3, \ldots, w_{L+2t-q}, w_{L+2t-q+1}\ , \tag{1}$$

образующих слово общей длины $L+2t$, у которого первые t и последние t символов заданы, а для каждой пары соседних слов выполняется условие

$$w_j = i_1\overline{w},\ \overline{w}i_q = w_{j+1},$$

т. е. два соседних слова пересекаются по общему подслову длины $q-1$. Первое (соответственно, последнее) слово в этой цепочке должно начинаться (соответственно, заканчиваться) левой опорой α_l (соответственно, правой опорой α_r). Если цепочка вида (1), составленная из слов опорного словаря, существует и единственна, то задача построения заполнения решена. Если существует несколько цепочек вида (1), составленных из слов опорного словаря, то среди всех возможных следует выбрать ту цепочку, которая обеспечивает максимум энтропии [7]

$$\widetilde{S} = -\sum_{w}\left\{\tilde{f}_w \cdot \ln \tilde{f}_w\right\} \tag{2}$$

пополненного частотного словаря $\overline{W}(q)$, где $\tilde{f}_w$ — частота слов, вычисленная по тексту, полученному в результате заполнения лакуны.

Рассмотрим теперь случай, когда заполнения лакуны словами из опорного словаря не существует. Тогда заполнение следует проводить всеми возможными в данном алфавите словами. Очевидно, что такое заполнение существует всегда и не единственно. Здесь всегда возникает проблема выбора наилучшего заполнения. Принцип выбора наилучшего заполнения в этой ситуации таков: выбрать следует такое заполнение (1), для которого условная энтропия

$$\overline{S} = \sum_{w} f_w \cdot \ln\left(\frac{f_w}{\tilde{f}_w}\right) \tag{3}$$

опорного частотного словаря $W(f_w \in W)$ относительно пополненного $\overline{W}(\tilde{f}_w \in \overline{W})$ достигнет минимума. Здесь f_w — частота слов в опорном частотном словаре, а $\tilde{f}_w$ — частота слов в словаре, построенном по всей последовательности, полученной в результате заполнения лакуны (пополненного); понятно, что для некоторых $w' f_{w'} = 0$, в то время как $\tilde{f}_{w'} > 0$.

Этими двумя принципами — принципом максимума энтропии пополненного словаря и принципом минимума условной энтропии опорного словаря относительно пополненного — не исчерпывается список критериев, которыми можно оценивать качество заполнения лакуны. Возможен, например, способ оценки качества восстановленной последовательности, основанный на подходах, используемых в задачах интерполяции и экстраполяции данных, а также в задачах прогнозирования (см., например, [8, 9]). Следует подчеркнуть, что эти два экстремальных принципа не являются эмпирическими: они представляют собой строгую формулировку требования минимума дополнительной информации, внесенной исследователем при построении заполнения лакуны. Принцип минимума условной энтропии

опорного словаря относительно пополненного носит универсальный характер: он применим всегда. Применимость первого критерия (принцип максимума энтропии пополненного словаря) не гарантирована заранее — он применим только в случае существования заполнения словами из опорного словаря, однако в рамках настоящей статьи мы ограничимся рассмотрением случая заполнения словами из опорного словаря.

Оба экстремальных энтропийных критерия не обеспечивают единственности заполнения. Примером может служить последовательность 0_1 из {0,1}-алфавита, для которой существует два заполнения — 001 и 011, частотные словари которых (толщины $1 \ldots 3$) имеют одинаковые значения энтропии. Такая ситуация является вырожденной: по-видимому, можно утверждать, что вырождение может быть снято заменой отдельных символов в тех фрагментах последовательности, которые имеются в распоряжении исследователя, и влияние этих замен на значение энтропии опорных словарей толщины от 1 до q не будет превосходить некоторой малой величины $\varepsilon = \varepsilon_q$, причем можно ожидать, что по порядку величины $\varepsilon_q \sim N^{-1}$.

Мы не знаем иных способов убедиться в существовании заполнения по опорному словарю, кроме непосредственного перебора всех возможных словосочетаний слов длины L, составленных из подслов длины q, при условии, что первое и последнее из слов имеют фиксированный набор символов в начале (конце, соответственно). Тем не менее можно пытаться строить различные субоптимальные алгоритмы, которые заранее бы исключали самые плохие варианты заполнения лакуны. Другой возможный (и эффективный) подход — использование высокопараллельных и кластерных вычислительных алгоритмов и устройств. Одним из возможных подходов здесь является использование кинетической машины Кирдина (КМК) [10 – 12]. Применение КМК для решения задачи заполнения лакуны в символьных последовательностях не является единственно возможным подходом, а сама КМК может использоваться в других задачах, требующих параллелизма вычислений, не связанных с задачей восстановления утерянных данных (см., например, [10–13]).

2. Кинетическая машина Кирдина в задаче восстановления утерянных данных

Прежде чем излагать процедуру заполнения утерянных данных, кратко рассмотрим понятие кинетической машины Кирдина, являющейся идеальным мелкозернистым параллельным вычислительным формализмом, по степени абстракции подобным машине Тьюринга [7, 10–12]. Известно, что КМК алгоритмически полна [10], т. е. любой мыслимый алгоритм представим в ее терминах. Отличительная черта КМК — мелкозернистый параллелизм. Чтобы описать идеальный вычислитель, необходимо описать объекты, над которыми происходят вычисления, и собственно сам способ вычисления. Пусть Ω — алфавит символов, а Ω^* — множество всех конечных слов или цепочек в этом алфавите. Обрабатываемой единицей является ансамбль слов M из алфавита Ω, который отождествляется с функцией F_M с конечным носителем на Ω^*, принимающей неотрицательные целые значения $F_M\colon \Omega^* \mapsto N \cup \{0\}$. Значение $F_M(w)$ интерпретируется как число экземпляров слова w в ансамбле M.

Обработка ансамблей в КМК состоит в совокупности элементарных событий, происходящих недетерминированно и параллельно. Элементарное событие $S\colon M \to M'$ состоит в том, что из ансамбля M изымается ансамбль K^- (это возможно, если для всех слов из этого второго ансамбля $F_{K^-}(w) \le F_M(w)$) и добавляется ансамбль K^+, т. е. $F_{M'} =$

$F_M - F_{K^-} + F_{K^+}$. Ансамбли K^- и K^+ однозначно задаются правилами или командами, которые объединяются в программу. Команды могут быть только трех видов; рассмотрим их подробнее.

1. Распад. $uvw \to uf + gw$, где u, w — произвольные слова, а v, f и g — фиксированные слова из Ω^*.

2. Синтез. $uk + dw \to usw$, где u, w — произвольные слова, а k, d и s — фиксированные слова из Ω^*.

3. Прямая замена. $uvw \to usw$, где u и w — произвольные слова, а v и s — фиксированные слова из Ω^*.

Неформально КМК можно представить как аналог химического реактора, в котором происходят реакции [14]. Имеется химический реактор идеального смешения, в котором плавают слова. В реактор добавляются правила-катализаторы; одни из них, взаимодействуя со словами, способствуют их распаду, другие, встречая пару подходящих слов, способствуют их синтезу. Наконец, третьи заменяют в словах некоторые подцепочки.

Перейдем теперь к изложению способа заполнения лакуны в последовательности с помощью КМК. Для этого приведем программы для КМК, описывающие алгоритмы построения частотного словаря и заполнения лакун. Пусть у нас имеется текст T, по которому требуется составить частотный словарь W_q. Программа для КМК, реализующая этот процесс, состоит из одной команды и выглядит следующим образом:

$$uf^1v^{q-1}g^1w \to uf^1v^{q-1} + v^{q-1}g^1w,$$

где в качестве M нужно взять ансамбль, состоящий из одного слова T. После того как машина остановится, ансамбль M будет содержать все слова длины q, встречающиеся в исходном тексте, с учетом их кратности.

Программа, реализующая процесс заполнения лакуны, в терминах КМК выглядит следующим образом:

$$\begin{aligned} \alpha_l + \alpha_l v^{q-t} &\to \alpha_l v^{q-t} *, \\ v^{q-t}\alpha_r + \alpha_r &\to *v^{q-t}\alpha_r; \end{aligned} \tag{4}$$

$$\begin{aligned} uv^{q-1} * + v^{q-1}v^1 &\to uv^{q-1}v^1 *; \\ v^1v^{q-1} + *v^{q-1}w &\to *v^1v^{q-1}w; \end{aligned} \tag{5}$$

$$u* + *v \to uv. \tag{6}$$

Первые две строчки программы осуществляют инициализацию “затравок”, т. е. обеспечивает взаимодействие правой (левой, соответственно) опоры с подходящим словом длины q. Третья и четвертая строчки осуществляют рост затравок, в ходе которого, собственно, и строится заполнение лакуны. И, наконец, последняя строка склеивает левые и правые части. Символ $\langle * \rangle$ не принадлежит алфавиту Ω и используется в программе, чтобы помечать те слова, которые успешно прошли стадию инициации.

Исходным ансамблем для этой программы являются некоторое количество копий “затравок” (левая (α_l) и правая (α_r) опоры) и некоторое количество копий словарей, полученных применением к опоре предыдущей программы. В КМК элементарные события происходят недетерминировано и параллельно. Тем не менее программа построена так, что вначале в ансамбле просто нет таких слов, к которым могли быть применены три последние команды. Таким образом, всю программу работы КМК по заполнению лакуны в последовательности можно представить состоящей из трех последовательных этапов.

Первый этап — "Инициализации затравок" (4) — заключается в присоединении к левой (α_l) и правой (α_r) опорам слов из W_q.

Второй этап — "Рост заполнений" (5) — начинается, как только в ансамбле будет достаточное количество слов, помеченных символом $\langle * \rangle$. Следует отметить, что число слов, помеченных символом $\langle * \rangle$, зависит от структуры фрагментов, имеющихся в распоряжении исследователя; может случиться так, что опорный словарь W_q не будет содержать ни одного слова, которое обеспечивало бы инициализацию. Последовательное и многократное применение команд этого этапа позволяет получить слова вида $\alpha_l u*$ и $*v\alpha_r$, длина которых составляет приблизительно $L/2$.

Третий этап (6). Наконец, осталось "склеить" слова вида $u*$ и $*v$ последней командой программы. Слова, полученные применением этой команды, будут составлять финальный словарь для данной программы и вышеописанного исходного ансамбля, так как ни одна из команд программы не будет к ним уже применима. Поскольку КМК функционирует недетерминировано и параллельно, в этом финальном ансамбле будут слова разной длины. Самое короткое из них может иметь длину $q + 2$. Теперь нам нужно выбрать из ансамбля слова длины $L+2t$ и исследовать полученные заполнения лакуны в соответствии с предложенными критериями.

3. Последовательный имитатор КМК в задаче заполнения лакун в последовательности

Кинетическая машина Кирдина является идеальным вычислительным устройством, обеспечивающим высокий уровень распараллеливания вычислений. Тем не менее, поскольку мы работаем на обычных последовательных машинах фон неймановского типа, например на персональных компьютерах (подробнее об этой стороне проблемы см. в [13]), будем строить такой имитатор КМК, который необходим для решения нашей конкретной задачи, а не всех алгоритмов, которые могут быть представимы в КМК. Мы делаем это для исключения нерезультативных шагов КМК и создания более эффективной программы, решающей нашу задачу заполнения лакун.

Предположим, что заполнение по опорному словарю возможно; оно возможно всегда, когда опорный словарь совпадает с полным. Тогда работа последовательного имитатора КМК состоит в продолжении левой и правой опор "навстречу" друг другу, с тем чтобы склеить их по достижении длины $(L+q-1+t)/2$ каждая. Затем такие половинки следует склеить (с помощью последней команды) и среди всех цепочек, получившихся в результате склеивания, выбрать те, для которых удовлетворяется тот или иной критерий — максимум энтропии пополненного частотного словаря либо минимум условной энтропии опорного частотного словаря относительно пополненного.

Работа КМК аналогична кинетике химической реакции. Время достижения результата кинетической машиной определяется аналогично времени достижения равновесия в аналогичной "химической реакции". Иными словами, время надежного вычисления нужной (в нашем случае) цепочки, являющейся заполнением лакуны, существенно зависит от "концентрации" тех слов, которые могут породить подходящие продолжения опоры (точнее, время построения подходящей цепочки определяется произведением таких "концентраций" с точностью до некоторого коэффициента, не зависящего от них). Высокий параллелизм вычислений для заполнения лакуны означает, что мы можем параллельно вычислять любое наперед заданное число продолжений одной и той же опоры.

Эффективность работы имитаторов КМК на последовательной машине фон неймановского типа сильно падает с уменьшением "концентрации" применимых слов. В первом приближении эффективность работы последовательного имитатора КМК может быть оценена по аналогии с оценками скорости химической реакции [7, 14]. Для повышения эффективности заполнения лакуны с помощью последовательного имитатора КМК он был модифицирован.

Имитатор КМК должен до определенной степени отражать работу параллельного вычислительного устройства. Для этого число копий затравок (левых опор, для определенности) бралось достаточно большим; в наших вычислительных экспериментах это число составляло $10^2 \dots 10^4$ копий. Имитатор выполнял параллельную работу кинетической машины одновременно со всеми этими затравками. Поскольку имитатор перебирал все затравки последовательно, лишь имитируя параллельную работу, постольку эффективность работы такого имитатора существенно определялась "концентрацией" тех слов, которые могли осуществить продуктивное взаимодействие с затравкой (опорой либо словом, прошедшим инициализацию). Для повышения эффективности построения заполнений лакун в символьной последовательности последовательный имитатор КМК был модифицирован; в него были внесены три модификации.

Первая модификация. Все затравки росли только в одном направлении — слева направо, для определенности. Как только цепочка достигала нужной длины $L + 2t$, она проверялась на включение в нее правой опоры. Если правая опора в нее входила, то эта цепочка считалась одним из возможных заполнений, в противном случае она отбрасывалась.

Вторая модификация. Модификации подвергся словарь, по которому строилось заполнение. Для реализации этапов, соответствующих работе команд (4), исходный словарь, из которого брались слова, заменялся на модифицированный. Модифицированный словарь содержал только те слова, которые имели начала, соответствующие затравке. Модификация частотного словаря означает построение на ансамбле M новой функции $\overline{F}_M\colon \Omega^* \to N \cup \{0\}$, такой, что $\overline{F}_M = F_M(uv^{q-1}v^1*) + F_M(*v^1v^{q-1})$ для всех v^1, v^{q-1}, для которых выполняются команды подпрограмм (4) и (5), и $\overline{F}_M = 0$ для всех остальных слов. Здесь верхний индекс указывает длину слов. Поскольку в общем случае у одной опоры существует несколько продолжений, постольку из всех мыслимых продолжений случайным образом выбиралось одно (для данной затравки) с вероятностью, пропорциональной доле этого продолжения. Для реализации этапов, соответствующих работе команд (5) КМК, исходный словарь, из которого брались слова, заменялся на тот, который содержал только те слова, которые имели начала, соответствующие слову, прошедшему инициализацию.

Третья модификация. Периодически проводилась селекция всех слов, являющихся продолжениями опор, построенных в силу команд (5) КМК. Среди продолжений слова $uv^{q-1}v^1*$ могут быть такие, которые сами уже не имеют никаких продолжений среди слов из используемого в текущем вычислительном эксперименте частотного словаря. Таким образом, возможна ситуация, в которой для некоторых слов команда (5) не выполняется никогда. С точки зрения повышения эффективности работы последовательного имитатора КМК такие "тупиковые" слова следует удалить. С другой стороны, удаление таких "тупиковых" слов на каждом шаге времени существенно понижает эффективность работы имитатора: приходится сравнивать большое количество слов. Соответственно, селекция (удаление "тупиковых" слов) проводилась не постоянно, а дважды за время роста продолжений. Для этого вся лакуна, которую необходимо было заполнить, разбивалась на три интервала равной длины. Понятно, что некоторые затравки давали такие продолжения, которые обрывались на длине, меньшей длины ее первого фрагмента (см. выше описание

работы КМК). По достижении остальными словами этой пороговой длины (равной трети длины лакуны) “тупиковые” слова из всего множества слов, с которыми работает имитатор КМК, удалялись. Затем те слова, которые достигли этой критической длины, удваивались (либо в общем случае их число увеличивалось в k раз) и процедура построения заполнения в силу команд (5) продолжалась до тех пор, пока эти слова не достигали следующей длины, на которой проводилась селекция. По достижении этой длины (составляющей две трети от длины лакуны) оставшиеся слова опять “размножались”.

4. Результаты

Целью настоящей работы является изучение качества восстановления лакуны малого размера в символьной последовательности с помощью КМК. Изучение качества восстановления проводилось с помощью вычислительных экспериментов по заполнению лакун в символьной последовательности. Для этого использовалась символьная последовательность из четырехбуквенного алфавита (A, C, G и T) длиной 12 310 символов — полный геном вируса диареи быка, штамм Oregon C24V (идентификатор AF091605 в EMBL-банке). Выбор этой последовательности обусловлен тем, что она обладает достаточно простой функциональной структурой [4, 5], а ее статистическая и информационная структура нетривиальна (см. ниже). В исходной последовательности искусственно создавались лакуны длины 6 и 8 символов. Эти лакуны выбирались не произвольным образом, а так, чтобы они соответствовали словам (соответствующей длины), обладающим тем или иным информационным свойством.

Остановимся на выборе места положения лакуны подробнее. Всего было проведено три серии вычислительных экспериментов для трех разных лакун. Слова, в месте вхождения которых создавались лакуны, выбирались следующим образом. Для исходной последовательности составлялись частотные словари толщины 6 и 8; затем в них находились те слова, которые имели наибольшую частоту и наименьшую частоту. Если таких слов было несколько, то конкретное слово выбиралось случайным образом. Лакуны создавались на месте вхождения одного из слов, которое:

1) имеет наибольшую частоту. Если таких слов было несколько, то конкретное слово выбиралось случайным образом;

2) имеет наибольшую информационную значимость, и реальная частота превышает ожидаемую;

3) имеет наибольшую информационную значимость, и ожидаемая частота превышает реальную.

Поясним, как именно определялась информационная ценность слов. Информационно ценными являются слова, для которых наблюдается наибольшее различие между реальной и ожидаемой частотами [15, 16–18]. Содержание этого определения существенно зависит от того, что именно понимать под ожидаемой частотой. В соответствии с [16–19] под ожидаемой частотой будем понимать, ту которая соответствует гипотезе о наиболее вероятном продолжении слова длины q, получаемом из слова длины $q-1$. Эта гипотеза формулируется в виде экстремального принципа; поскольку всюду в рамках настоящей работы рассматривается заполнение лакуны словами длины q по словам длины $q-1$, постольку приведем сразу окончательную формулу для ожидаемой частоты, опустив все промежуточные выкладки:

$$\tilde{f}_{i_1 i_2 i_3 \ldots i_{q-1} i_q} = \frac{f_{i_1 i_2 i_3 \ldots i_{q-2} i_{q-1}} f_{i_2 i_3 i_4 \ldots i_{q-1} i_q}}{f_{i_2 i_3 i_4 \ldots i_{q-2} i_{q-1}}}. \tag{7}$$

Детальное изложение вывода этой формулы приведено в [15, 16–18]. Здесь $f_{i_1 i_2 i_3 \ldots i_{q-1} i_q}$ — реальная частота слова $i_1 i_2 i_3 \ldots i_{q-1} i_q$, а $\tilde{f}_{i_1 i_2 i_3 \ldots i_{q-1} i_q}$ — его ожидаемая частота. Информационно значимыми словами являются такие, для которых не выполняется двойное неравенство

$$\mu^{-1} \le \frac{f_{i_1 i_2 i_3 \ldots i_{q-1} i_q}}{\tilde{f}_{i_1 i_2 i_3 \ldots i_{q-1} i_q}} \le \mu,$$

где $\mu > 1$ — порог информационной значимости. Нас не интересует конкретное значение порога μ, поскольку мы выбираем слова с максимальным и минимальным значениями отношения реальной и ожидаемой частот.

Для данной последовательности были обнаружены следующие слова длины 6 и 8 символов, которые удовлетворяют сформулированным выше критериям: слово `AAAAAT` (24 копии) и слово `AGAAGAAG` (10 копий), и по одному слову с экстремальными значениями информационной значимости. Для слов `CGAGCG` и `CATATA` отношение реальной и ожидаемой частот составляет 19.00000 и 0.17452 соответственно; для слов `GAACAAAG` и `CAGGGACT` длины 8 эти отношения равны 11.00000 и 0.40000 соответственно. Выбор конкретного места положения лакуны в случае исключения из текста наиболее часто встречающегося в нем слова значения не имеет, поскольку восстановление всегда ведется по локальным характеристикам исследуемого текста — по его частотному словарю до толщины 6 либо 8. Сформулированные выше критерии достаточно эффективны: они всегда позволяют выбрать одно слово малой длины в качестве тестового для создания искусственной лакуны. Строго говоря, вполне возможна неединственность тех слов, которые удовлетворяют каждому из критериев. Тем не менее можно с известной уверенностью утверждать, что в этом случае число таких слов невелико, а ситуацию вырождения можно снять малыми мутациями в исходной последовательности. В нашем случае в вычислительных экспериментах лакуны начинались в позициях 1618 (удалено слово `AAAAAT`), 271 (удалено слово `CGAGCG`) и 3548 (удалено слово `CATATA`), а также в позициях 4650 (удалено слово `AGAAGAAG`), 3565 (удалено слово `GAACAAAG`) и 2864 (удалено слово `CAGGGACT`).

Для указанной последовательности с помощью КМК строились заполнения последовательно по словарям толщины от 2 до 8. Параметры эксперимента были выбраны такими, что суммарное число затравок, по которым строилось заполнение, составляло 60 000. Мож-

Таблица 1

Заполнения лакун длины 6 и значения энтропии пополненного словаря

	AAAAAT		CGAGCG		CATATA	
q	ω	$\widetilde{S}$	ω	$\widetilde{S}$	ω	$\widetilde{S}$
2	TGCGCG	2.72077857132	GTCTCC	2.72054223744	ATCGCG	2.72064703059
3	TCTGCG	4.06716077228	GATTTC	4.06672087122	GTCGCG	4.06716662417
4	TTGCGT	5.40525742207	GGAATC	5.40445215830	CTCGAT	5.40504111001
5	TAGGAT	6.71655428452	TTCGGC	6.71604218425	CCCCGT	6.71669328085
6	GCATCT	7.92480972609	CAGTGC	7.92371671175	CAGATC	7.92425582537
7	TGCAAC	8.79709098916	GGGGTC	8.79695251549	—	
8	TGAGGG	9.21280082514	—		—	

Примечание. ω — построенное заполнение, $\widetilde{S}$ — условная энтропия.

Таблица 2

Заполнение лакун длины 8 и значения энтропии $\widetilde{S}$ пополненного частотного словаря

q	ω	$\widetilde{S}$	ω	$\widetilde{S}$	ω	$\widetilde{S}$
2	CGTCCGTT	2.720939760	TTCCGCGC	2.721000178	GTCCGCGC	2.720858333
3	TTGCGCGT	4.067544759	CGACGTCC	4.067541315	CGTCCCGC	4.067491969
4	CTCGCATC	5.405596001	CTGAGCGC	5.405593149	TGCAGCGC	5.405392907
5	CGCTCTCA	6.717693814	AGCTCGGC	6.717526133	CGTCGCTC	6.717531875
6	CCTCCGCG	7.926193567	GTTTGTTA	7.925275907	TCATATCC	7.925287766
7	AAGGCTTG	8.798360548	AAGCATCA	8.796780954	TTTTGCGA	8.797600920
8	AAGAAATG	9.214041201	—		—	

Таблица 3

Значения энтропии частотного словаря всей последовательности (S) и опорного частотного словаря для заполнений по частотным словарям толщины от 2 до 8 лакун длины 6

q	S	AAAAAT	CGAGCG	CATATA
2	2.72038305890055	2.72052439564904	2.72029566841063	2.72040119160596
3	4.06663072991445	4.06681729332964	4.06641771399159	4.06669284470570
4	5.40450313718055	5.40479931012881	5.40419194059492	5.40461526688478
5	6.71624486867158	6.71665928813260	6.71563365391426	6.71635445246348
6	7.92456277097593	7.92512253590089	7.92410798085974	7.92445648293715
7	8.79751586658168	8.79767369650203	8.79699353100612	8.79672553321249
8	9.21347690139651	9.21295968422461	9.21300223510236	9.21300223510236

Таблица 4

Значения энтропии частотного словаря всей последовательности (S) и опорного частотного словаря для заполнений по частотным словарям толщины от 2 до 8 лакун длины 8

q	S	AGAAGAAG	GAACAAAG	CAGGGACT
2	2.72038305890055	2.72057092991200	2.72055186570748	2.72043871398960
3	4.06663072991445	4.06693491606402	4.06689576984170	4.06673700221786
4	5.40450313718055	5.40497600004984	5.40483798998128	5.40467245529098
5	6.71624486867158	6.71694493269837	6.71657993813210	6.71647521962344
6	7.92456277097593	7.92547345269262	7.92496012944630	7.92477978915917
7	8.79751586658168	8.79862122956146	8.79697604776826	8.79794255554618
8	9.21347690139651	9.21424542936787	9.21291913142307	9.21318719460859

но ожидать, что для такого числа затравок имитатор КМК перебирает фактически все мыслимые заполнения лакуны данной длины. В табл. 1 и 2 приведены результаты заполнения лакун длины 6 и 8 в указанной последовательности. Значения энтропии частотных словарей толщины от 2 до 8 для всей последовательности, а также значения энтропии опорных частотных словарей этой же толщины для каждого из трех положений лакун длины 6 и 8 приведены в табл. 3 и 5 соответственно.

Линейный рост размеров лакуны приводит к экспоненциальному росту числа ее возможных заполнений; как следствие, затраты вычислительных ресурсов (в первую очередь — времени вычислений) также растут экспоненциально или близко к экспоненциальному росту. Для того чтобы преодолеть указанный недостаток, последовательный имитатор КМК был модифицирован так, чтобы в ходе вычислений отбрасывать некоторые

Таблица 5

Число удачных заполнений лакуны длины 6 (N_1), 8 (N_2) и 350 (N_3) для заполнений по частотным словарям толщины от 2 до 8

q	N_1	N_2	N_3	S_0	S_1	S_2
2	1039	15596	101763	2.72038305890055	2.72121777683546	2.72390428758459
3	277	5536	31373	4.06663072991445	4.06785558910179	4.07196642562716
4	65	1743	6187	5.40450313718055	5.40597007633796	5.41054190211537
5	13	648	2532	6.71624486867158	6.71679809543189	6.72145961396103
6	1	194	665	7.92456277097593	7.92081327057962	7.92202495678440
7	1	28	91	8.79751586658168	8.78355353634615	8.77846700340517
8	1	3	20	9.21347690139651	9.19005552098473	9.18161322660571

Примечание. S_0 — энтропия оригинала; S_1 — энтропия опоры; S_2 — энтропия пополненного словаря.

заведомо тупиковые продолжения затравок (см. выше).

Работа КМК по заполнению длинных ($10^2 \ldots 10^3$ символов) лакун иллюстрируется табл. 5. В ней представлены результаты работы последовательного имитатора КМК как для шести ранее исследованных лакун, так и для лакуны длиной 350 символов, созданной в той же последовательности генома вируса. По-прежнему, восстановление проводилось для словарей до толщины 8 включительно. Для лакуны этой длины были выбраны следующие параметры вычислительного эксперимента: начало лакуны в 1001 символе, длина лакуны 350 символов, количество точек деления, в которых происходят размножение растущих покрытий и удаление тупиковых, 3, коэффициент размножения 4, а стартовое число затравок составляло 1500.

5. Обсуждение

Цель настоящей статьи — демонстрация возможностей быстрого и эффективного восстановления части утерянной символьной последовательности по фрагментам, имеющимся в распоряжении исследователя. Решение этой задачи и ее сложность существенно зависят от структуры тех фрагментов последовательности, которые имеются в распоряжении исследователя. По-видимому, для некоторого класса исходных последовательностей не существует иного метода построения заполнения лакуны, кроме перебора всех возможных вариантов ее заполнения. Представленные в статье данные свидетельствуют о возможности эффективного разрешения указанной задачи с помощью последовательного имитатора высокопараллельного мелкозернистого бесструктурного вычислительного устройства — кинетической машины Кирдина.

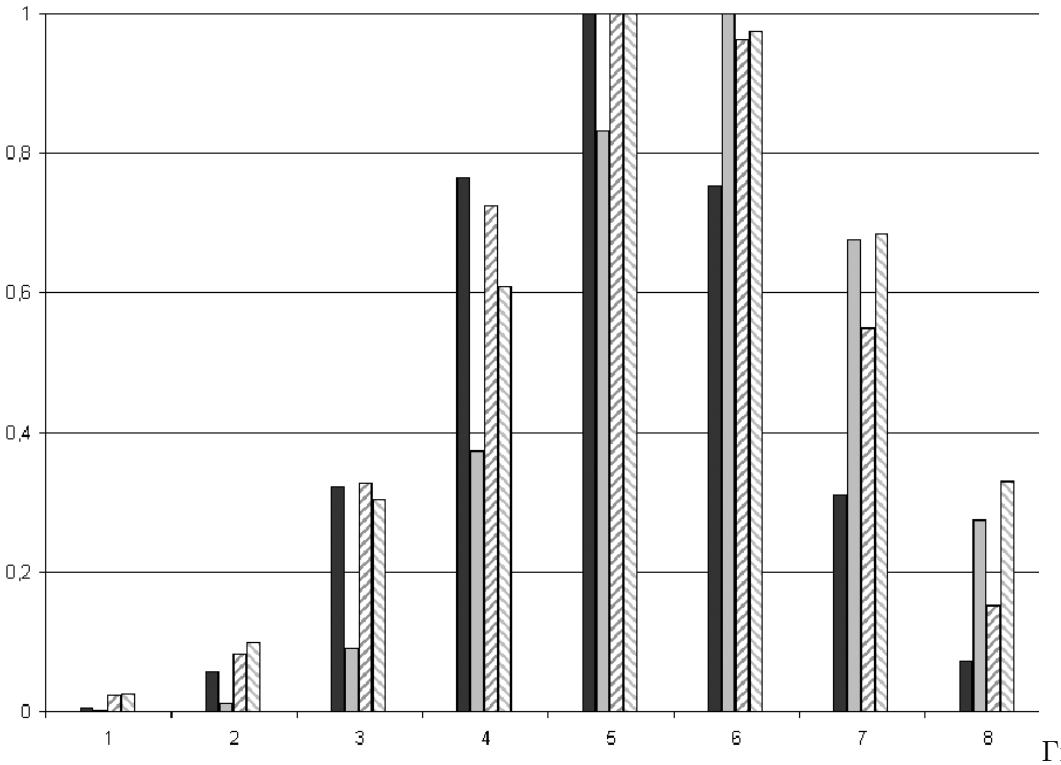
Обратимся к табл. 3 и 5. Первое, что обращает на себя внимание, — это кажущееся аномальным поведение энтропии частотного словаря всей последовательности и опорного частотного словаря. Энтропия опорного частотного словаря превышает энтропию всей последовательности для слов `AAAAAT` и `AGAAGAAG` до толщины $q = 7$ и $q = 8$ включительно соответственно. Никакого противоречия в этом нет. Создание лакуны приводит к удалению одного экземпляра самого часто встречающегося слова, что, очевидно, заметно уменьшает неравновесность опорного частотного словаря. Информационно значимые слова длины 6 и 8 встречаются по одному экземпляру. Тем не менее наблюдаемое превосходство значений энтропии опорных словарей до толщины 5 и 7, соответственно, объясняется теми же

причинами. У слов `CGAGCG` и `GAACAAAG` реальная частота превосходит ожидаемую и значения энтропии всей последовательности самым естественным образом больше значений энтропии опорных словарей для этих слов. У слов `CATATA` и `CAGGGACT`, напротив, ожидаемая частота превосходит реальную; это означает, что частота подслов меньшей длины, включенных в эти слова, весьма велика, что и приводит к тому, что энтропии опорных словарей до толщины 5 и 7 соответственно превосходят энтропии частотного словаря всей последовательности. Отметим, что ожидаемый порядок поведения значений энтропии восстанавливается для этих словарей, как только толщина превосходит 5 и 7.

Параметры вычислительного эксперимента были таковы, что можно с высокой надежностью утверждать, что отсутствие заполнений для лакун, построенных на месте всех информационно значимых слов, означает, что такие заполнения по опорному словарю не существуют. Очевидно, что заполнения по полному словарю существуют, однако задача построения этих заполнений по полному словарю в данной работе не ставилась.

С прикладной точки зрения весьма интересен также случай заполнения длинных лакун. Здесь может наблюдаться уменьшение энтропии пополненного частотного словаря по сравнению со словарем, построенным по всей (целой) последовательности. У этого обстоятельства могут быть две причины: таково свойство последовательности и таково распределение всех возможных заполнений по значениям энтропии пополненного словаря. Поясним сказанное. Вполне может случиться так, что лакуна организована в таком месте последовательности, что оставшиеся части оказываются мало разнообразными и заполнение получается также менее разнообразным, чем оригинал; предельным случаем здесь может служить обсуждавшийся выше пример с утратой символа. Вторая причина заключается в том, что заполнения, получаемые с помощью КМК (точнее, ее последовательного имитатора), по значениям энтропии пополненного частотного словаря распределены неравномерно, а такое распределение смещено в сторону меньших значений энтропии. Обратимся к табл. 5. Хорошо видно, что число подходящих заполнений лакуны падает экспоненциально или почти экспоненциально с ростом толщины словаря, по которому строится заполнение. Тот факт, что для толщины 5 включительно и опорный, и пополненный словари имеют энтропию, большую энтропии словаря, построенного по оригинальной последовательности, скорее всего объясняется свойствами исходной последовательности. Хорошо видно также, что энтропия пополненного словаря оказывается меньше энтропии опорного для заполнений лакуны длиной 350 символов словарями толщины 7 и 8. Здесь можно почти наверняка утверждать, что проявился эффект неравномерного распределения заполнений по значениям энтропии пополненного частотного словаря.

Действительно, общее число мыслимых заполнений (построенных по полному частотному словарю) для лакуны длиной 350 символов составляет 4^{350}; это необозримое число, превосходящее 10^{230}. Пусть далеко не все из этих заполнений могут быть построены по опорному словарю, тем не менее число возможных заполнений существенно превосходит число найденных (см. табл. 5). Из этого следует, что найденные заполнения скорее всего относятся к числу наиболее часто встречающихся (с точки зрения значения энтропии пополненного словаря). Мы не знаем заранее, как именно распределены заполнения; тем не менее первое представление о таком распределении могут дать гистограммы распределения числа заполнений с заданным значением пополненного частотного словаря, построенные для случая заполнения лакуны длиной 350 символов словарями толщины от 4 до 7 включительно (см. рисунок). Гистограммы строились следующим образом: весь диапазон значений энтропии пополненного словаря, от максимального значения до минимального, разбивался на 10 интервалов, и подсчитывалось число заполнений со значением энтропии



Гистограмма числа заполнений по значениям энтропии пополненного словаря.

пополненного словаря, попадающим в каждый интервал. Хорошо видно, что эти гистограммы имеют ярко выраженное смещение пика в сторону меньших значений энтропии.

Резюмируя сказанное, можно сформулировать ряд проблем, имеющих общий характер для задач восстановления утерянных данных с помощью КМК. Первая проблема — выбор тест-объекта. Классическим объектом здесь служит случайная последовательность. Необходимы специальные "инструментальные" исследования на них, они позволят выявить некоторые свойства метода, которые не зависят от структуры последовательности. Есть, однако, и другие варианты тест-объектов. Хорошим тестовым объектом могут быть одномерные фракталы [14, 20] с подходящим числом элементарных ячеек. Возможно использование и других символьных последовательностей, полученных с помощью ясных, четких и сравнительно простых порождающих правил, полностью определяющих структуру такой последовательности, например разложения лиувиллевых либо трансцедентных чисел и т.п.

Поскольку в общем случае точное восстановление невозможно, возникает задача сравнения восстановленной последовательности с оригинальной. Выбор метода сравнения зависит от длины восстанавливаемой лакуны. Для длинных лакун существуют методы сравнения символьных последовательностей по их частотным словарям с помощью их информационных характеристик [21 – 26] помимо обычного евклидового расстояния между словарями [16, 17, 27], для коротких лакун естественным будет сравнение с помощью выравнивания. Еще одно важное направление исследований — переформулировка критерия, по которому выбирается оптимальное заполнение из множества возможных. Так, например, возможен выбор такого заполнения, которое обеспечивает возможно более близкое значение энтропии пополненного частотного словаря к энтропии опорного, а также построение заполнения по словарям одной толщины, а выбор наилучшего заполнения — по

значениям энтропии словарей другой толщины. Важно исследование соотношения между восстановленными последовательностями, полученными в силу принципа максимума абсолютной энтропии и принципа минимума условной энтропии. Восстановление в силу принципа минимума условной энтропии опорного словаря относительно пополненного обеспечивает также подобие всей восстановленной последовательности тем фрагментам, которые есть в распоряжении исследователя.

Наконец, стоит проблема улучшения реализации последовательных имитаторов для кинетической машины Кирдина. В ближайшее время вряд ли появится аппаратная реализация КМК, поэтому последовательные имитаторы таких параллельных вычислительных устройств еще долгое время будут играть существенную роль как в теоретических исследованиях, так и в решении прикладных задач. Такое улучшение будет существенно определяться кругом прикладных задач, однако некоторые вопросы носят общий характер, и мы их здесь укажем. Первый — "самообучение" в выборе точек размножения и выборе коэффициента размножения тех слов, которые остаются после селекции "тупиковых". Нами был выбран вариант, при котором селекция проводилась дважды. Основания такого выбора — дело вкуса; содержательным эвристическим соображением явилось то, что троичная система счисления в определенном смысле наиболее экономична. Аналогичные или близкие соображения могут использоваться и при выборе коэффициента размножения отобранных слов, являющихся заполнениями лакуны. Второй вопрос — построение модифицированного частотного словаря.

Мы благодарим рецензента, замечания которого позволили сделать изложение материала более четким и прозрачным.

Список литературы

[1] Горбань А.Н., Россиев Д.А. Нейронные сети на персональном компьютере. Новосибирск: Наука, 1996.

[2] Зиновьев А.Ю. Визуализация многомерных данных. Красноярск: Изд-во КГТУ, 2000.

[3] Gorban A.N., Rossiev D.A., Wunsch II D.C. Neural Network Modelling of Data with Gaps // Радиоэлектроника. Информатика. Управление. 2000. № 1. С. 47–55.

[4] Konopka A.K. Theoretical Molecular Biology // Molecular Biology and Biotechnology / R.A. Meyers (Ed.) Weinheim: VCH Publ., 1995. P. 888–896.

[5] Oyama S. The Ontogeny of Information: Developmental Systems and Evolution. Cambridge, N. Y.: Cambridge Univ. Press, 1985.

[6] Садовский М.Г. Восстановление пробелов в символьных последовательностях по наблюдаемым информационным характеристикам // Нейроинформатика и ее приложения: Мат. 7-й Всерос. конф. Красноярск, 1–3 октября, 1999. С. 129.

[7] Горбань А.Н. Обход равновесия. Новосибирск: Наука, 1984.

[8] Рябко Б.Я. Прогноз случайных последовательностей и универсальное кодирование // Пробл. передачи информации. 1988. Т. 24, № 2. С. 3–14.

[9] Ryabko B.Ya. The complexity and effectiveness of prediction algorithms // J. Complexity. 1999. Vol. 10, N 3. P. 281–295.

[10] Горбунова Е.О. Формально-кинетическая модель бесструктурного мелкозернистого параллелизма // Сиб. журн. вычисл. математики. 1999. Т. 2, № 3. С. 239–256.

[11] Кирдин А.Н. Модель идеального ансамбля для параллельных вычислений // Нейроинформатика и ее приложения. Красноярск: Изд-во КГТУ, 1997.

[12] Gorban A.N., Gorbunova K.O., Wunsch D.C. Liquid brain: kinetic model of structureless parallelism // Advances in Modelling & Analisis. AMSE. 2000. Vol. 5, N 5. P. 427–453.

[13] Gorban A.N., Gorbunova K.O., Wunsch D.C. Liquid Brain: The proof of algorithmic universality of quasichemical model of fine-grained parallelism // Neural Network World. 2001. N 4. P. 391–412.

[14] Яблонский Г.С., Быков В.И., Горбань А.Н. Кинетические модели каталитических реакций. Новосибирск: Наука, 1983.

[15] Горбань А.Н., Попова Т.Г., Садовский М.Г. Классификация нуклеотидных последовательностей по частотным словарям обнаруживает связь между их структурой и таксономическим положением организмов // ЖОБ. 2003. Т. 64, № 1. С. 51–63.

[16] Gorban A.N., Popova T.G., Sadovsky M.G. Automatic classification of nucleotide sequences and its relation to natural taxonomy and protein function // Proc. of 1st Intern. Conf. on Bioinformatics of Genome Regulation and Structure, Novosibirsk, Aug., 24–27, 1998. Pt II. P. 314–317.

[17] 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 Systems & Information Dynamics. 2000. Vol. 7, N 1. P. 1–17.

[18] 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 11 — Smart Engineering System Design. N. Y.: ASME Press, 2001. P. 657–663.

[19] Бугаенко Н.Н., Горбань А.Н., Карлин И.В. Универсальное разложение трехчастичной функции распределения // Теорет. и мат. физика. 1990. № 28. С. 430–441.

[20] Гуляев В.К. Квазикристаллическая модель для точного задания координат атомов в кристаллах // Докл. РАН. 2001. Т. 381, № 3. С. 325–328.

[21] Горбань А.Н., Попова Т.Г., Садовский М.Г. Корреляционный подход к сравнению нуклеотидных последовательностей // ЖОБ. 1994. Т. 55, № 4/5, С. 420–430.

[22] Кирсанова Е.Н., Садовский М.Г. Метод статистического сравнения объектов // Радиоэлектроника. Информатика. Управление. 2000. № 2. С. 71–82.

[23] Гуляев В.К., Садовский М.Г. Геном как апериодический одномерный кристалл // Нейроинформатика и ее приложения: Мат. 9-й Всерос. конф., Красноярск, 5–7 октября, 2001. С. 50–51.

[24] Садовский М.Г. Сравнение нуклеотидных и аминокислотных последовательностей по их частотным словарям // Математика, компьютер, образование: Мат. 5-й Междунар. конф., Дубна, 29–31 янв. 1998. С. 178.

[25] Popova T.G., Sadovsky M.G. Investigating statistical properties of genetic texts: new method to compare two genes // Modelling, Measurement & Control. Ser. C. AMSE Press. 1994. Vol. 45, N 4. P. 27–36.

[26] Sadovsky M.G. Comparison of Symbol Sequences: No Editing, No Alignment // Open Systems & Information Dynamics. 2002. Vol. 9, N 1. P. 19–36.

[27] Kirsanova E.N., Sadovsky M.G. Entropy approach to a comparison of images. Open Systems & Information Dynamics. 2001. Vol. 8, N 2. P. 183–199.

Поступила в редакцию 4 марта 2003 г.,
в переработанном виде — 27 октября 2003 г.