О СРАВНЕНИИ СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

М. Г. Садовский
Институт биофизики СО РАН, Красноярск, Россия
e-mail: msad@icm.krasn.ru

A new method to compare symbol sequences is proposed in which sequences are compared through their frequency dictionaries. For this comparison calculations are developed which compare the specific entropy of the dictionary with the hybrid dictionary.

Введение

Проблема сравнения символьных последовательностей актуальна для различных областей науки. Основная трудность здесь заключается в том, что в пространстве символьных последовательностей сложно ввести метрику. Формально метрика в таком пространстве существует — это метрика Хэмминга [1]. Однако содержательно такая метризация мало продуктивна: эта метрика позволяет лишь различать полностью совпадающие последовательности и все остальные.

Наиболее распространенным в настоящее время методом сравнения символьных последовательностей является метод выравнивания, или редакционного расстояния [2, 3, 4, 5]. Этот метод заключается в "подгонке" одной последовательности под другую с помощью вставки пробелов и замены (либо удаления) символов так, чтобы эти две последовательности совпали. Каждой вставке и/или замене назначается определенный штраф; наилучшей укладкой считается та, которая дает наименьшее значение суммарного штрафа.

Не обсуждая здесь причин распространенности различных вариаций метода выравнивания, отметим только, что у него есть два принципиальных недостатка, которые не могут быть устранены ни в одной версии метода. Метод выравнивания требует выбора системы штрафных (весовых) функций и выбора опорной последовательности, относительно которой проводится выравнивание. И то и другое выбирается исходя из соображений, лежащих за пределами собственно метода выравнивания.

Во многих исследованиях прикладного характера этот выбор определяется целями прикладного исследования. Тем не менее теоретически обе эти проблемы — выбор опорной последовательности и системы штрафных функций — далеки от решения. Результат сравнения очень зависит от определения штрафов (весовых функций) для вставок отдельных символов (либо пробелов) либо замены (удаления); назначение таких штрафов определяется искусством исследователя.

Кроме того, метод выравнивания чувствителен к длине сравниваемых последовательностей. Точность сравнения падает экспоненциально с ростом их длины (если только они не имеют совпадающих участков, сопоставимых по длине со всей последовательностью).

Фактически невозможно выровнять две последовательности, сильно различающиеся по длине. Выравнивание также имеет свои ограничения и по числу сравниваемых последовательностей. Несмотря на то, что увеличение числа сравниваемых последовательностей ведет, как правило, к росту точности выравнивания [6, 7], общее число выравниваемых последовательностей едва ли может превышать 10^2; содержательно выровнять тысячу (и более) последовательностей едва ли возможно.

В настоящей работе предложен принципиально иной подход к сравнению символьных последовательностей. Он основывается на сравнении словарей этих последовательностей. Словарь W_q — это множество всех слов $\{\omega\}$ (связных цепочек заданной длины q), встречающихся в последовательности, с указанием частоты f_ω каждой такой цепочки (слова) либо числа ее копий n_ω. Каждый словарь представляет собой точку в $\mathbf{C}^q$-мерном пространстве (q — длина слова, а $\mathbf{C}$ — мощность алфавита $\aleph$), в котором можно определить обычное евклидово расстояние [9, 11]. Определяемое так расстояние является метрикой, однако она не всегда является эффективным инструментом сравнения, и тогда возникает задача построения меры близости двух (или нескольких) частотных словарей.

Мера близости двух или нескольких частотных словарей может быть введена различными способами. Настоящая статья посвящена изложению двух методов. В первом случае сравнение группы последовательностей производится через сравнение их носителей. Во втором случае сравниваются частотные словари; сравнение осуществляется через промежуточный объект, также являющийся частотным словарем, — так называемый гибридный словарь. Оба предложенных метода сравнения обладают следующими свойствами:

— не требуют выбора опорной последовательности;

— не требуют назначения штрафных функций;

— позволяют сравнивать последовательности любой длины;

— позволяют сравнивать любое число последовательностей.

1. Сравнение последовательностей по их носителям

Рассмотрим вначале метод сравнения последовательностей по их носителям. Основная идея метода заключается в следующем: две последовательности (T_1 и T_2) считаются тем более близкими одна к другой, чем больше в них общих слов. Пусть необходимо сравнить M последовательностей $\mathsf{T}_1, \mathsf{T}_2, \ldots, \mathsf{T}_M$; обозначим через N_j длину последовательности T_j. Далее, будем полагать, что все последовательности порождены из одного и того же алфавита $\aleph$ мощности $\mathbf{C}$. Данное ограничение не является существенным, однако содержательно сравнение последовательностей в подавляющем большинстве случае ограничивается именно таким случаем. Будем рассматривать лишь связные последовательности; рассмотрение несвязных последовательностей также возможно, однако никаких содержательных результатов оно не дает, а технические трудности заметно возрастают [12, 13, 14, 15]. Рассмотрим какую-либо последовательность из описанного выше семейства; любую связную подпоследовательность длиной q ($1 \le q \le N_j$) символов, встречающуюся в ней, будем называть словом (длины q), а совокупность всех слов (длины q), встречающихся в ней, — q-носителем.

Рассмотрим для начала случай двух последовательностей T_1 и T_2, которые необходимо сравнить. Сопоставим каждой последовательности набор их l-носителей $Q_q^{(1)}$ и $Q_q^{(2)}$ возрастающей толщины $1 \le q \le t$; толщиной носителя будем называть длину слов. Для

носителей толщины q вычислим величину

$$w(q) = \frac{\|Q_q^{(1)} \bigcap Q_q^{(2)}\|}{\|Q_q^{(1)} \bigcup Q_q^{(2)}\|}, \tag{1}$$

здесь $\|\cdot\|$ обозначает мощность множества $\{\cdot\}$. Функция $w(q)$ есть отношение числа одинаковых слов, встречающихся в носителях двух сравниваемых последовательностей к общему числу слов в них. Тогда близость двух последовательностей будет определяться функцией

$$l(s) = \frac{2}{s(s+1)} \sum_{q=1}^{s} qw(q). \tag{2}$$

Тем самым, в качестве меры близости двух последовательностей T_1 и T_2 мы имеем функцию, которая зависит от характерного масштаба, на котором проводится сравнение. Если две последовательности совпадают, то $l(s) = 1$ для любого s. Если в сравниваемых последовательностях нет ни одного общего символа, то $l(s) = 0$ для любого s. Чем больше одинаковых слов встречается в сравниваемых последовательностях, тем сильнее $l(s)$ отличается от нуля. В общем случае найдется такая длина слов t_m, для которой $w(t_m) = 0$, но $w(t_m - 1) > 0$. Иными словами, начиная с определенной длины слов, в носителях двух сравниваемых последовательностей не будет ни одного общего слова. Справедливость данного утверждения следует из того факта, что если две последовательности не совпадают и одна из них не является подпоследовательностью другой, то $w(N_m) = 0$, где N_m — длина меньшей из них. Проверяя наличие общих слов в носителях меньшей толщины, всегда можно найти ту минимальную, для которой указанное свойство выполняется[1]. Соответственно, при $s > t_m$ функция $l(s)$ будет монотонно стремиться к нулю.

Функция $l(s)$ достигает своего максимума[2] при некотором $s = s^*$; такую толщину носителя естественно считать радиусом корреляции двух сравниваемых последовательностей. Резюмируем сказанное. С помощью функций (1), (2) можно сравнивать две последовательности, причем результат сравнения не зависит от порядка сравниваемых последовательностей. Результатом сравнения является мера близости, определяемая для той или иной длины слов; полную картину дает набор таких мер, вычисляемый для словарей толщины $1 \le q \le t_m$.

Метод сравнения последовательностей с помощью функций (1), (2) легко обобщается на случай сравнения произвольного числа последовательностей. Пусть есть набор последовательностей $\{\mathsf{T}_j\}$, $1 \le j \le M$, которые необходимо сравнить. Составим для каждой из них набор носителей $Q_q^{(j)}$ возрастающей толщины $1 \le q \le t$ и вычислим аналог функции (1):

$$\widetilde{w}(q) = \frac{\|\bigcap\limits_{j=1}^{M} Q_q^{(j)}\|}{\|\bigcup\limits_{j=1}^{M} Q_q^{(j)}\|}. \tag{3}$$

Далее вычислим функцию $l(s)$ в силу выражения (2) с функцией $\widetilde{w}(q)$ вместо $w(q)$.

Данный метод сравнения может быть модифицирован для повышения чувствительности в случае очень слабых корреляций между сравниваемыми последовательностями

[1] Поскольку последовательности конечны.

[2] Существование этого максимума также следует из конечности наборов функции $l(s)$.

(если число общих слов очень мало). Для этого функцию (2) можно заменить на

$$l_\alpha(s) = \frac{1}{A}\sum_{q=1}^{s} q^\alpha w(q) \quad \text{либо} \quad l_\alpha(s) = \frac{1}{A}\sum_{q=1}^{s} q^\alpha \widetilde{w}(q)$$

с соответствующим нормировочным множителем A, а $\alpha > 1$.

2. Метод сравнения последовательностей с помощью их частотных словарей

Изложенный в предыдущем разделе метод сравнения не учитывает того обстоятельства, что разные слова в сравниваемых последовательностях могут встречаться несколько раз. Представленный ниже метод сравнения с помощью **гибридного словаря** свободен от этого недостатка. Рассмотрим его в общей постановке. Суть метода заключается в вычислении минимального количества информации, необходимого для того, чтобы один из сравниваемых словарей превратить в другой.

Пусть по-прежнему заданы M последовательностей $\mathsf{T}_1, \mathsf{T}_2, \ldots, \mathsf{T}_M$ и N_j означает длину j-й. Построим для каждой из них свой q-носитель $Q^{(j)}$, припишем каждому элементу (слову) ω, $\omega \in Q^{(j)}$, число n_ω его копий, наблюдаемое в этой последовательности. Полученная конструкция является конечно-частотным словарем соответствующей последовательности T_j. Заменяя число копий n_ω в конечно-частотном словаре на частоту

$$f_\omega = \frac{n_\omega}{N_j}$$

этого слова, получаем частотный словарь W_q толщины q.

Предложенный в настоящей работе метод сравнения опирается на идею сравнения заданного словаря с "равновесным". Каждый частотный словарь может рассматриваться как q-частичная функция распределения. Если в нашем распоряжении имеется равновесное распределение ϕ^*, то относительную энтропию некоторого заданного распределения ϕ относительно этого равновесного можно вычислить всегда. Действительно [19], такая энтропия одного распределения ϕ относительно другого (равновесного) ϕ^* равна

$$\overline{S} = \sum_{\mu\in\Gamma} \phi \cdot \ln\left(\frac{\phi}{\phi^*}\right). \tag{4}$$

Аналогично можно сравнивать и частотные словари. Основную трудность здесь вызывает выбор такого частотного словаря, который бы соответствовал "равновесному" распределению ϕ^*.

Возможны различные варианты определения частотного словаря, являющегося аналогом такого равновесного распределения. Отметим, что непосредственное сравнение двух частотных словарей $W_q^{(1)}$ и $W_q^{(2)}$ в силу формулы (4) возможно далеко не всегда. Для того чтобы формула (4) была применима, необходимо, чтобы носитель одного из словарей полностью содержал носитель другого; понятно, что заранее гарантировать такое включение нельзя. Выход из этой ситуации состоит в том, чтобы сравнивать словари не непосредственно друг с другом, а с некоторым промежуточным объектом. Таким объектом является **гибридный словарь** W_q^G толщины q.

Очевидно, что определение гибридного словаря неоднозначно. Определим гибридный словарь следующим образом: пусть по-прежнему $f_\omega^{(j)} \in W^{(j)}$, $j = 1, \ldots, k$, обозначает частоту слова ω в j-м словаре. Тогда частота слова $\omega \in W^G$ в гибридном словаре определяется как среднее арифметическое частот этого слова в сравниваемых словарях:

$$f_\omega^G = \frac{f_\omega^{(1)} + f_\omega^{(2)} + \cdots + f_\omega^{(k)}}{k}. \tag{5}$$

Такой выбор частот в гибридном словаре гарантирует минимум суммы условной энтропии каждого из сравниваемых словарей относительно гибридного.

Доказательство этого факта весьма просто. Действительно, пусть $\{p^G\}$ — частота слов в некотором "равновесном" словаре. Тогда выражение (4) для суммы по всем словарям будет таким:

$$\overline{S} = \sum_{j=1}^{k} \left[\sum_\omega f_\omega^{(j)} \cdot \ln \left(\frac{f_\omega^{(j)}}{p_\omega^G} \right) \right]. \tag{6}$$

Требуется найти минимум (6) при очевидном ограничении

$$\sum_\omega p_\omega^G = 1.$$

Функция Лагранжа для (6) выглядит так:

$$L = \sum_{j=1}^{k} \left[\sum_\omega f_\omega^{(j)} \cdot \ln \left(\frac{f_\omega^{(j)}}{p_\omega^G} \right) \right] - \lambda \left(\sum_\omega p_\omega - 1 \right), \tag{7}$$

где λ — множитель Лагранжа, а варьируемыми переменными являются p_ω. Дифференцируя (7) по p_ω, получаем очевидное решение (5) при $\lambda = k$.

Собственно мерой близости той или иной последовательности T_j к общему статистическому предку является значение условной энтропии

$$\overline{S}^{(j)} = \sum_\omega \left[f_\omega^{(j)} \cdot \ln \left(\frac{f_\omega^{(j)}}{f_\omega^{(G)}} \right) \right], \tag{8}$$

где $f_\omega^{(j)}$ — частота слова ω в j-м частотном словаре, а $f_\omega^{(G)}$ — частота этого слова в гибридном словаре. Следует подчеркнуть, что мера близости (8) может зависеть от толщины q словарей, для которых проводится сравнение.

Если все последовательности T_1, T_2, …, T_k совпадают, то совпадают и их частотные словари $W_q^{(j)}$, $1 \le j \le k$. Очевидно, что гибридный словарь W_q^G в этом случае также совпадает с любым из группы сравниваемых, а величина (8) равна нулю для всех словарей. Противоположный случай попарно непересекающихся словарей дает значение меры максимально возможного различия словарей в группе и $\overline{S}^{(j)} = \ln k$. В общем случае между значениями абсолютной энтропии $S^{(j)}$ частотного словаря, значением энтропии гибридного словаря S^G и значениями условной энтропии (8) существует простое соотношение

$$\sum_{j=1}^{k} \overline{S}^{(j)} = S^G - \frac{\sum_{j=1}^{k} S^{(j)}}{k}.$$

Результаты сравнения одиннадцати хромосом генома *Encephalitozoon cuniculi*

Хромосома	N	$q=1$	$q=2$	$q=3$	$q=4$	$q=5$	$q=6$
CNS07EGB	209 982	0.000037	0.002831	0.005249	0.002415	0.004704	0.011856
CNS07EGA	197 426	0.000045	0.003328	0.006099	0.000850	0.002880	0.009831
CNS07EG9	194 439	0.000098	0.004041	0.007006	0.001043	0.003133	0.008545
CNS06C8G	218 329	0.000909	0.001627	0.003280	0.000960	0.002717	0.008545
AL590450	262 797	0.000368	0.003259	0.005814	0.006300	0.010336	0.019692
AL590449	238 147	0.000009	0.059943	0.082045	0.002019	0.003924	0.009579
AL590448	226 576	0.000112	0.002747	0.004970	0.000614	0.002172	0.007629
AL590447	220 294	0.000307	0.004132	0.007123	0.001052	0.002592	0.008746
AL590446	211 018	0.000081	0.003376	0.006121	0.001889	0.003660	0.010722
AL590445	251 002	0.000151	0.004013	0.006946	0.000936	0.002648	0.009631
CNS07EGC	267 509	0.000489	0.106007	0.176741	0.001263	0.003254	0.010224

Гибридный частотный словарь является общим статистическим предком для группы сравниваемых частотных словарей. Это означает, что из него можно породить любой частотный словарь из сравниваемой группы, добавив (или изъяв) **минимально необходимое** количество информации. Смысл величины (8) заключается в том, что она определяет количество той самой минимально требуемой информации, с помощью которой можно из гибридного словаря породить заданный. Величина (8) может принимать и отрицательные значения, это происходит в том случае, когда гибридный словарь оказывается более определенным, чем какой-либо из группы. Если величина (8) принимает отрицательные значения, это говорит о том, что для порождения такого словаря требуется изъять минимально необходимое количество информации (сделать его менее определенным); в противном случае эту информацию следует добавить. Проиллюстрируем метод результатами сравнения одиннадцати хромосом полного генома простейшего *Encephalitozoon cuniculi*, паразитирующего как на человеке, так и на других существах. Данный геном представляет собой одиннадцать последовательностей из четырехбуквенного алфавита $\{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}\}$. Все последовательности являются связными; они депонированы в EMBL-банке[3]. В таблице представлены результаты сравнения, проведенного на словарях толщины $1 \le q \le 6$, указана длина каждой из сравниваемых последовательностей. Очевидно, что никакими иными методами сравнить столь обширный генетический материал невозможно.

3. Обсуждение

Предложенные в настоящей работе методы сравнения символьных последовательностей являются новыми и оригинальными. Основное их отличие от наиболее широко распространенного в настоящее время выравнивания заключается в их замкнутости и формализованности: не требуется выбирать систему штрафных (весовых) функций, не требуется выбирать опорную последовательность. Напомним, что выбор этих двух важнейших параметров лежит полностью за пределами самого метода выравнивания. Наконец, оба предложенных метода инвариантны относительно любых перестановок сравниваемых последовательностей.

Тем не менее развитый здесь метод не является полной альтернативой выравниванию: у каждого из них своя область применимости. Несмотря на формальную применимость

[3] *http://www.ebi.ac.uk/genomes*

развитого здесь метода гибридного словаря для сравнения очень коротких ($N \sim \| \aleph \|^2$) последовательностей, содержательных результатов здесь добиться невозможно. Для таких коротких последовательностей более продуктивным является метод выравнивания.

Кроме того, метод выравнивания получил очень широкое распространение, и его использование, по крайней мере, в задачах молекулярной биологии и молекулярной генетики стало повсеместным, а некоторые результаты, полученные с помощью этого метода, носят универсальный характер. Делаются активные попытки адаптировать этот метод для сравнения последовательностей, для которых классические варианты сравнения не являются содержательными [17]. Именно для таких случаев нам представляется удачным комбинированное изучение различных символьных последовательностей с помощью развитого здесь метода и выравнивания.

Поскольку развитый здесь метод является полностью замкнутым и не опирается ни на какую иную информацию, кроме той, которая содержится в комбинациях различных символов, постольку с его помощью можно верифицировать различные системы выбора штрафных (весовых) функций, необходимых для построения выравнивания, а также выбирать ту последовательность, которая может быть использована в качестве опорной. Естественно в качестве опорной выбирать такую последовательность, частотные словари которой в наибольшей степени близки к гибридному частотному словарю. Вопрос о верификации системы штрафных (весовых) функций более сложен. Здесь также есть произвол в выборе той толщины словаря q, которую следует выбрать для такой верификации. Уменьшить такой произвол можно двумя путями: привлекая дополнительные соображения, связанные с непосредственным содержанием той задачи, в рамках которой проводится сравнение символьных последовательностей, либо выбирая такую систему весовых (штрафных) функций, которая на данной длине обеспечивает наилучшее выравнивание, например, в целом для всего семейства сравниваемых последовательностей. Некоторые приложения изложенного выше метода сравнения символьных последовательностей через их гибридный словарь представлены в [21, 22].

Следует также отметить, что развитый в настоящей работе метод может быть обобщен и для сравнения любых дискретных объектов, в частности двумерных изображений на основе решеток [16, 18]. Природа таких объектов (их двумерность) позволяет с помощью изложенного в настоящей работе метода изучать их внутреннюю структуру, например анизотропию двумерных кристаллов. Подробное обсуждение этих вопросов выходит за рамки настоящей статьи.

Список литературы

[1] Hamming R.W. Coding and Information Theory. New Jersey: Prentice-Hall, 1980.

[2] Basseville M. Distance measures for signal processing and pattern recognition // Signal Processing. 1989. Vol. 18, N 4. P. 349–369.

[3] Shapira D., Storer J.A. Large edit distance with multiple block operations SPIRE // LNCS 2857. 2003. P. 369–377.

[4] Давыдов В.А. Коды, исправляющие ошибки в метрике модулей, в метрике Ли и ошибки оператора // Проблемы передачи информации. 1993. Т. 29, № 1. С. 10–20.

[5] Левенштейн В.И. О совершенных кодах в метрике выпадений и вставок // Дискретная математика. 1991. Т. 3, № 1. С. 3–20.

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

[7] Smith T.F., Waterman M.S. Identification of common molecular subsequences // J. Mol. Biol. 1981. Vol. 147. P. 195–197.

[8] 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 Eng. Systems Through Artificial Neural Netwerks. Vol. 11: Smart Eng. System Design. N.Y.: ASME Press, 2001. P. 657–663.

[9] Gorban A.N., Popova T.G., Sadovsky M.G. Classification of symbol sequences over thier frequency dictionaries: towards the connection between structure and natural taxonomy // Open Syst. & Inform. Dyn. 2000. Vol. 7, N 1. P. 1–17.

[10] Popova T.G., Sadovsky M.G. The new measure of relationship between two symbolic sequences // Advances in Modelling & Analisis. Ser. A, AMSE. 1994. Vol. 22, N 2. P. 13–17.

[11] Горбань А.Н., Попова Т.Г., Садовский М.Г. Корреляционный подход к сравнению нуклеотидных последовательностей // Журн. общей биологии. 1994. Т. 55, № 4–5. Р. 420–430.

[12] Gorbunova E.O., Kondratenko Yu.B., Sadovsky M.G. Data loss reparation due to indeterminate fine-grained parallel computation, LNCS 2658 / P. M. A. Sloot et al. (Eds) // Smart Eng. System Design. Berlin, Heidelberg: Springer-Verlag, 2003. P. 794–801.

[13] Неменчинская Е.О., Кондратенко Ю.В., Садовский М.Г. Предварительные результаты в проблеме восстановления утерянных данных с помощью кинетической машины Кирдина // Вычисл. технологии. 2004. Т. 9, № 1. С. 42–57.

[14] Nemenchinskaya E.O., Kondratenko Yu.B., Sadovsky M.G. Entropy based approach to data loss reparation through the indeterminate fine-grained parallel computation // Open Systems & Information Dyn. 2004. Vol. 11, N 2. P. 161–175.

[15] Бурлаков В.П., Неменчинская Е.О., Садовский М.Г. Локальный подход к восстановлению утерянных данных // Матер. 12-й Всерос. конф. “Нейроинформатика и ее приложения”. 2004. С. 99–100.

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

[17] Sunyaev S.R., Bogolepsky G.A., Oleynikova N.V. et al. From analysis of protein structural alignment toward a novel approach to align protein sequences // PROTEINS: Structure, Function, and Bioinformatics. 2004. Vol. 54, N 3. P. 569–582.

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

[19] Горбань А.Н. Обход равновесия. Новосибирск: Наука, 1984. 268 с.

[20] 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.

[21] Sadovsky M.G. Comparison of symbol sequences: no editing, no alignment // Open Systems & Information Dyn. 2002. Vol. 1, N 1. P. 19–36.

[22] Sadovsky M.G. The method of comparison of nucleotide sequences based on the minimum entropy principle // Bull. of Mathem. Biology. 2003. Vol. 65, N 2. P. 309–322.

Поступила в редакцию 30 марта 2004 г.,
в переработанном виде — 11 января 2005 г.