ДОКЛАДЫ АКАДЕМИИ НАУК, 2011, том 441, № 1, с. 24–28

ИНФОРМАТИКА

УДК 519.7

ОБ ОТБОРЕ ИНФОРМАТИВНЫХ ЗНАЧЕНИЙ ПРИЗНАКОВ НА БАЗЕ КРИТЕРИЕВ РАЗРЕШИМОСТИ В ЗАДАЧЕ РАСПОЗНАВАНИЯ ВТОРИЧНОЙ СТРУКТУРЫ БЕЛКА

© 2011 г. Член-корреспондент РАН К. В. Рудаков, И. Ю. Торшин

Поступило 23.06.2011 г.

Разработан проблемно-ориентированный формализм для описания задачи распознавания вторичной структуры белка, рассмотрены ее разрешимость, регулярность и локальность. Получены конструктивные критерии разрешимости задачи, позволяющие исследовать гипотезу о локальном характере зависимости вторичной структуры от первичной. Исследование прецедентов может проводиться как на множествах признаков, так и на множествах значений признаков (так называемых мотивов). Введение порядка на мотивах с помощью эвристических оценок информативности мотивов позволило применить разрабатываемый формализм для анализа реальных множеств прецедентов, при этом анализ разрешимости позволил провести эффективный отбор наиболее информативных значений признаков.

В современной биологии любой белок рассматривается с нескольких точек зрения: 1) как одномерная аминокислотная последовательность, представляющая собой химическую формулу молекулы белка (так называемая “первичная структура”, 1D); 2) как одномерная последовательность характерных локальных конфигураций (“вторичная структура”, 2D); 3) как трехмерный объект (“третичная структура”, “пространственная структура”, 3D) и 4) как особый механизм, выполняющий определенные роли в функционировании клетки. Основной задачей теоретической биологии считается установление закономерностей, определяющих взаимосвязь первичной, вторичной и третичной структур [1, 2].

Противоречивость экспериментальных данных, обусловленная особенностями структурного эксперимента, неоднозначность определения алфавита для описания вторичной структуры и необходимость систематического исследования гипотезы о локальном характере взаимосвязи между первичной и вторичной структурами указывают на целесообразность разработки специализированного формализма для корректной постановки изучаемой проблемы [3, 4]. В настоящей работе данная задача рассматривается как перевод последовательности символов из одного алфавита в другой, причем ее исследование проводится в рамках основных конструкций алгебраического подхода к проблеме синтеза корректных алгоритмов [5–7].

Пусть A – алфавит для описания первичной структуры белка (“верхнего слова”) и B – алфавит для описания вторичной структуры (“нижнего слова”), так что $A = \{a_1, a_2, ..., a_{n(A)}\}$, $n(A) = |A| > 0$ и $B = \{b_1, b_2, ..., b_{n(B)}\}$, $n(B) = |B| > 0$. Алфавит A обычно определяется как $A = \{A, C, D, E, F, G, H, I, K, L, M, N, P, R, S, T, V, W, Y, \Delta\}$, где латинские литеры соответствуют общепринятым однобуквенным обозначениям аминокислот, а Δ обозначает неопределенность. Алфавит B может быть определен существенно различными способами: базовые алфавиты; производные алфавиты на основе последовательностей литер базового алфавита; расширение базового алфавита с учетом сегментов вторичной структуры [3]. Для целей настоящей работы вполне приемлем алфавит $B = \{S, H, L, \Delta\}$, описывающий три принципиально различных вида вторичных структур белков.

Пусть $A^* = \bigcup_{l=1}^{\infty} A^l$ – множество всех исходных слов в алфавите A, а $B^* = \bigcup_{l=1}^{\infty} B^l$ – множество всех слов в алфавите B. Произвольное слово в алфавите A обозначим $V = v_1 v_2 ... v_{n(V)}$, в алфавите B – $W = w_1 w_2 ... w_{n(W)}$, $n(V)$ и $n(W)$ – длины слов. Тогда решение исследуемой задачи распознавания сводится к поиску некоторой функции F: $A^* \to B^*$, причем $|F(V)| = |V|$ ($|V|$ – длина слова V).

О п р е д е л е н и е 1. Пусть $\mathrm{Pr} \subset A^* \times B^*$, $\mathrm{Pr} \neq \phi$. Функция F к о р р е к т н а, если

Вычислительный центр им. А.А. Дородницына Российской Академии наук, Москва

Московский физико-технический институт, Долгопрудный Московской обл.

$$\forall_{\Pr}(V, W)\colon F(V) = W.$$

Теорема 1. *Корректная функция F существует тогда и только тогда, когда*

$$\forall_{\Pr}(V_1, W_1), (V_2, W_2)\colon\ (V_1 = V_2) \Rightarrow (W_1 = W_2). \quad (1)$$

Определение 2. Исследуемую задачу Z, определяемую множеством прецедентов Pr, будем называть разрешимой, если для нее существует корректная функция F.

Определение 3. Разрешимость задачи зависит от множества Pr. Если задача разрешима на некотором множестве Pr прецедентов, то будем называть такое Pr непротиворечивым, в противном случае Pr – противоречиво.

Определение 4. Задача Z регулярна на множестве прецедентов Pr тогда и только тогда, когда выполняется условие регулярности:

$$\forall_{\Pr}(V_i, W_i), (V_j, W_j)\ \ (i \neq j) \Rightarrow (V_i \neq V_j). \quad (2)$$

Предлагаемый формализм разрабатывается с целью тестирования гипотезы о локальном характере взаимосвязи между первичным и вторичным уровнями структуры белка. В рамках формализма локальность означает то, что каждая литера нижнего слова определяется неким подсловом верхнего слова.

Пусть дано слово $U = \{u_1, u_2, \ldots, u_n\}$ длины n. Это может быть верхнее слово (V) или нижнее слово (W). Определим некую ведущую позицию i, $1 \le i \le n$. Дана также "маска" $\hat{m} = \{\mu_1, \mu_2, \ldots, \mu_m\}$, где $\mu_i \in Z$, $\mu_1 < \mu_2 < \ldots < \mu_m$, μ_i – позиции маски, $m = |\hat{m}|$ – размерность маски $\hat{m}$, $\mu_m - \mu_1 + 1 = [\hat{m}]$ – протяженность маски. Оператор выбора подслова $\eta(i, \hat{m}, U)$ выделяет определенную подпоследовательность слова U по маске $\hat{m}$, помещенной на позицию i:

$$\eta(i, \hat{m}, U) = \begin{cases} u_{i+\mu_1}u_{i+\mu_2}\ldots u_{i+\mu_m}, & \text{если } i + \mu_1 \ge 1,\ i + \mu_m \le n, \\ \phi & \text{в противном случае.} \end{cases} \quad (3)$$

Слова в множестве прецедентов Pr имеют конечную длину, поэтому для точности изложения следует описать краевые эффекты с учетом области определения оператора η (3).

Определение 5. Пусть $L, R \in N \cup \{0\}$. Функцию F назовем (L, R)-корректной, если $\forall(V, W) \in \Pr\colon F(V) = W'$, причем $w'_1 = w'_2 = \ldots = w'_L = \Delta$, $w'_{|И|-R+1} = \ldots = w'_{|И|} = \Delta$, $w'_i = w_i$ при $L < i \le |V| - R$.

При заданной системе масок $M = \{\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_{|M|}\}$ есть два существенно различающихся способа определения L и R: как минимальных отступов, при которых применимы все маски из M ($L(M) = \max(-\min\limits_{k=1,\ldots,N} \mu_1^k, 0)$, $R(M) = \max(\max\limits_{k=1,\ldots,N} \mu_{|m_k|}^k, 0)$), и как отступов, при которых применима хотя бы одна маска из M ($l(M) = \max(-\max\limits_{k=1,\ldots,N} \mu_1^k, 0)$, $r(M) = \max(\min\limits_{k=1,\ldots,N} \mu_{|m_k|}^k, 0)$).

Теорема 2. *($l(M)$, $r(M)$)-корректная функция также (L(M), $R(M)$)-корректна.*

Определение 6. Объединенной маской системы масок M назовем $\hat{m}_\Sigma(M) = \bigcup\limits_{k=1}^{|M|} \hat{m}_k$. Гипотеза о локальности исследуемой задачи распознавания формулируется как гипотеза о существовании некоторой локальной функции $f\colon (A^*)^{|\hat{m}_\Sigma(M)|} \to B^*$. Функция f корректна, если для всякого (V, W) из Pr при $1 \le i \le l(M)$ и $n - r(M) + 1 \le i \le n$ выполнено $w_i = \Delta$, а при $l(M) < i \le n - r(M)$ $f(\eta(i, \hat{m}_\Sigma(M), V)) = w_i$.

Теорема 3. *Локальная и корректная функция f существует тогда и только тогда, когда выполняется критерий локальной разрешимости*

$$\forall_{\Pr}(V^1, W^1), (V^2, W^2)\ \forall(i, j)\colon \eta(i, \hat{m}_\Sigma(M), V^1) = \eta(j, \hat{m}_\Sigma(M), V^2) \Rightarrow w_i^1 = w_j^2, \quad (4)$$
$$l(M) < i \le |V^1| - r(M),$$
$$l(M) < j \le |V^2| - r(M), \quad i \neq j.$$

Следствие 1. *Из теорем 2 и 3 следует критерий локальной ($L(M)$, $R(M)$)-разрешимости.*

Следствие 2. *Расширение B-алфавита может приводить к потере разрешимости задачи* Z(Pr, M).

Следствие 3. *При выполнении* (4) *выполним и критерий локальной разрешимости с использованием отдельных масок*:

$$\forall_{\Pr}(V^1, W^1), (V^2, W^2)\ \forall(i, j)\ \left(\forall_{k=1}^{|M|} \hat{m}_k\colon \eta(i, \hat{m}_k, V^1) = \eta(j, \hat{m}_k, V^2)\right) \Rightarrow w_i^1 = w_j^2. \quad (4')$$

При выполнении условия существования локальных функций (4), (4') задача распознавания вторичной структуры локально разрешима, в противном случае – локально неразрешима. В общем случае система масок M, при которой задача

локально разрешима, может быть избыточной в том смысле, что разрешимость сохранится при удалении некоторых масок из M. Рассмотрим возможности варьирования множества масок M при непротиворечивом Pr и L, R = const, используя классические конструкции дискретной математики [8].

Определение 7.1. Систему масок M назовем 0-тупиковой, если условие (4) выполнено для M, но не выполнено для любой $M' \subset M$, такой что $M_\Sigma(M') \subset M_\Sigma(M)$.

Определение 7.2. M – тупиковая, если (4) выполнено для M, но нарушается для любой $M' \subset M$.

Определение 8. Маску $\hat{m}_{i_0}$, $i_0 \in \{1, 2, ..., N\}$, будем называть ядерной, если $\hat{m}_{i_0} \not\subset \not\subset \bigcup\limits_{j = 1, ..., N}^{j \neq i_0} \hat{m}_j$. Ядерными системами масок будем называть M, обладающие свойством ядерности:

$$\forall_{i=1}^{N}\, i\ \exists_{\hat{m}_i} \mu: \forall_{j=1}^{N, i \neq j}\, j(\mu \notin \hat{m}_j). \tag{5}$$

Теорема 4. *M – тупиковая система масок тогда и только тогда, когда M обладает свойством ядерности* (5).

Следствие 4. *Из тупиковости следует* 0*-тупиковость.*

Следствие 5. *Если в некоторой* 0*-тупиковой системе масок M имеется ядерная маска $\hat{m}_{i_0}$, то $\hat{m}_{i_0}$ входит во все тупиковые подсистемы M.*

Следствие 6. *Пусть* в 0*-тупиковой M есть несколько ядерных масок $\hat{m}_{i_1}$, $\hat{m}_{i_2}$, ..., $\hat{m}_{i_L}$ (ядерная подсистема). Если некоторая $\hat{m} \subseteq \bigcup\limits_{j = 1, ..., L} \hat{m}_{i_j}$, то $\hat{m}$ не входит ни в одну тупиковую M.*

Выражения (4) и (4') описывают локальную разрешимость задачи в терминах признаков (“масок”). Для экспериментального тестирования разрешимости и нахождения безызбыточных M целесообразно перейти от исследования разрешимости на множествах признаков к исследованию разрешимости на множествах значений признаков [4].

Определение 9. Элементарными объектами q (далее просто “объектами”) назовем элементы множества $Q = A^{|\hat{m}_\Sigma(M)|} \times B$.

Элементами наблюдаемых множеств объектов $Q(\Pr, M)$ являются пары $q_i^j = (\eta(i, \hat{m}_\Sigma(M), V^j), w_i^j)$, каждая пара есть совокупность подслова, выбранного по $\hat{m}_\Sigma(M)$ в i-й ведущей позиции верхнего слова ($V^j = v_1^j v_2^j ... v_{n(V^j)}^j$) и i-й литеры нижнего слова ($W^j = w_1^j w_2^j ... w_{n(W^j)}^j$) j-го прецедента. Множество объектов $Q^{\text{нп}}$ – непротиворечиво, если $\forall_{Q^{\text{нп}}}(i, j)\ V_i = V_j \Rightarrow w_i = w_j$.

Определение 10. Назовем мотивами к элементы множества $\mathbf{K} = \{(\hat{m}, V) | \ \hat{m} \in M, n(V) = |\hat{m}|\}$.

Будем говорить, что мотив $\kappa = (\hat{m}, V')$ присутствует в объекте $q = (V, w)$, если $\eta(L(M) + 1, \hat{m}, V) = V'$, и обозначим принадлежность мотива объекту q как $\kappa \in^* q$. Мотив к назовем отличающим для произвольной пары объектов q_1 и q_2, если к присутствует в одном из объектов и отсутствует во втором.

Теорема 5. *Условие локальной разрешимости задачи выполнено тогда и только тогда, когда для каждой пары объектов $q_1 = (V_1, w_1)$ и $q_2 = (V_2, w_2)$ при $w_1 \neq w_2$ существует хотя бы один отличающий мотив. Иначе говоря, выполнено условие*

$$\forall_{Q(\Pr, M)}(i, j): w_i \neq w_j \Rightarrow$$

$$\Rightarrow \exists_{K(\Pr, M)} \kappa: (\kappa \in^* V_i) \neq (\kappa \in^* V_j). \tag{6}$$

Определение 11. Множество мотивов назовем тупиковым, если условие (6) выполнено для K, но не выполнено для любого $K' \subset K$.

Редукция множества мотивов $K(\Pr, M)$ и нахождение тупиковых K может рассматриваться как частный случай выделения подкласса “наиболее информативных признаков” во множестве всех значений всех исследуемых признаков. Для этого вводятся эвристические оценки информативности мотивов.

Определение 12. Оценкой информативности мотивов назовем функцию $D: \mathbf{K} \to \mathbf{R}_+$.

Введем несколько эвристических оценок информативности мотивов, основанных на частотах встречаемости мотивов в различных классах объектов. Пусть $K(\Pr, M)$ – множество мотивов для заданных Pr и M, а $Q = Q(\Pr, M)$ – множество объектов. Каждый мотив $\kappa_\alpha \in K(\Pr, M)$ входит в состав $N_\Sigma^\alpha = \sum\limits_{l=1}^{m=|B|} N_l^\alpha$ объектов из Q, где N_l^α соответствует числу объектов $q = (a, b)$, у которых $b = b_l$, $b_l \in B$, так что мотиву κ_α поставлен в соответствие вектор $(\nu_1^\alpha, \nu_2^\alpha, ..., \nu_m^\alpha, N_\Sigma^\alpha)$. Пусть частоты встречаемости литер $b_1 \in B$ во всем множестве объектов

Q составляют вектор $(\nu_1^0, \nu_2^0, ..., \nu_m^0)$. Используя кусочно-линейную функцию, определим информативность α-го мотива по литере b_l как

$$D_l^\alpha = \begin{cases} 1 - \dfrac{\nu_l^\alpha}{\nu_l^0} \ \text{при} \ \nu_l^\alpha \le \nu_l^0, \\ \dfrac{\nu_l^\alpha - \nu_l^0}{1 - \nu_l^0} \ \text{при} \ \nu_l^\alpha > \nu_l^0. \end{cases} \quad (7)$$

Кроме сравнительных оценок распределения объектов между классами на информативность мотива влияет частота его встречаемости среди объектов. Используя введенные обозначения, можно предложить по меньшей мере два способа общей оценки информативности α-го мотива:

$$D_1(\alpha) = \sum_{l=1}^{m} D_l^\alpha,$$

$$D_2(\alpha) = N_\Sigma^\alpha D_1(\alpha) = N_\Sigma^\alpha \sum_{l=1}^{m} D_l^\alpha.$$

Помимо сформулированных выше эвристических оценок информативности мотива могут быть предложены и другие. Интуитивно ясно, что "информативный" мотив должен выделять "достаточно много" объектов l-го класса N_l^α и "достаточно мало" объектов всех остальных классов $N_\Sigma^\alpha - N_l^\alpha$ [4].

Эвристические оценки информативности мотивов необходимы для нахождения тупиковых множеств мотивов с учетом критерия (6) разрешимости задачи. Пусть D – эвристическая оценка информативности мотивов, D: $\mathbf{K} \to \mathbf{R}_+$. Имея упорядоченное множество мотивов, отбор наиболее информативных будем осуществлять как отбор достаточного для разрешимости количества "наиболее информативных" мотивов. Отобранные таким образом мотивы образуют некоторое множество различающих мотивов K^0 с наивысшей информативностью, такое что $K^0 \subseteq K(\Pr, M)$.

Т е о р е м а 6. *Множество K^0 является тупиковым тогда и только тогда, когда для каждого мотива из K^0 в Q существует хотя бы одна пара объектов, для которой данный мотив – единственный различающий.*

С л е д с т в и е 7. *Тупиковость K^0 гарантирована только при постановке задачи в двухклассовой форме.*

С л е д с т в и е 8. *K^0 – тупиково при постановке задачи в двухклассовой форме и определении характеристической функции K^0 в форме*

$$T(\alpha) = \begin{cases} 1 \equiv \underset{Q}{\exists}(i,j)\colon K_f(i,j) = \alpha, \\ 0 \quad \textit{в противном случае}, \end{cases} \quad (8)$$

где $K_f(i,j) = \min\limits_{1,\ldots,|K|} \alpha\colon (\kappa_\alpha \in^* V_i) \ne (\kappa_\alpha \in^* V_j)$.

Для проведения экспериментов по тестированию условия разрешимости (6) и вычисления характеристических функций множеств наиболее информативных мотивов (8) были использованы все 165000 прецедентов, представленные в базе данных PDB [9], на основе которых сформировано непротиворечивое множество из 5 млн объектов. Исследованы выборки этого множества объектов размером 10000, 20000, 30000, 50000, 100000, 200000 объектов, сформированные путем случайного отбора объектов без возвращения. Изученные системы масок получены на основе системы масок с размерностью всех масок, равной $m = 2$ (система M_8^2) и $m = 3$ (M_8^3), в которых нулевая позиция каждой маски соответствовала позиции $\left\lfloor \dfrac{n}{2} \right\rfloor + 1$ в верхнем слове каждого объекта. Произведена частичная редукция систем масок путем удаления сдвиг-эквивалентных масок и для вычисления $\mathrm{T}(\alpha)$ использовалась система масок $\mathrm{M}_8'^3$, $\left|\mathrm{M}_8'^3\right| = 25$. В качестве эвристической оценки информативности использовалась функция $D_2(\alpha)$. Вычисления $T(\alpha)$ показали логарифмический характер зависимости числа отобранных мотивов от $|Q|$, при этом оценки $\left|K^0\right|$, полученные на разных выборках одного размера, отличались не более чем на 5%.

Зависимость числа пар объектов, на которых достигнута разрешимость (максимально $\left|Q^2\right|$), от числа мотивов с максимальной информативностью указала на существование некоторого "ядра" в тупиковых множествах мотивов. Мотивы, входящие в такое "ядро", обеспечивают разрешимость более чем на 95% парах объектов. Полная разрешимость достигается добавлением к "ядру" некоторых низкоинформативных мотивов, каждый из которых обеспечивает разрешимость всего лишь на нескольких парах.

Работа выполнена при поддержке грантов РФФИ 09–07–12098, 09–07–00212-а и 09–07–00211-а и контракта Минобрнауки России № 07.514.11.4001.

СПИСОК ЛИТЕРАТУРЫ

1. *Povolotskaya I.S., Kondrashov F.A.* // Nature. 2001. V. 465. № 7300. P. 922–926.
2. *Torshin I.Y.* Bioinformatics in the Post-Genomic Era: The Role of Biophysics. N.Y.: Nova Biomed. Books, 2006. 256 p.
3. *Рудаков К.В., Торшин И.Ю.* // Информатика и ее применения. 2010. Т. 4. № 2. С. 25–35.
4. *Рудаков К.В., Торшин И.Ю.* Анализ информативности мотивов на основе критерия разрешимости в задаче распознавания вторичной структуры белка. ИОИ-8. Кипр, 2011.
5. *Журавлев Ю.И.* // Проблемы кибернетики. 1978. В. 33. С. 5–68.
6. *Журавлев Ю.И., Рудаков К.В.* В кн.: Проблемы прикладной математики и информатики. М.: Наука, 1987. С. 187–198.
7. *Рудаков К.В.* // ДАН. 1987. Т. 297. № 1. С. 43–46.
8. *Яблонский С.В.* Введение в дискретную математику. М.: Наука, 1979. 272 с.
9. *Berman H.M., Henrick K., Nakamura H., et al.* // Nature Struct. Biol. 2003. V. 10. № 12. P. 980–982.