УДК 51.76

Р-адический анализ первичной структуры белков и реализующий его веб сервис.

Козырев С.В., Козьмин Ю.П., Богатова О.В., Гарковенко А.В., Некрасов А.Н.

Введение

Основная идея, позволившая разработать метод выявления иерархически организованных элементов при анализе первичной структуры белков, была предложена в работе [1] и заключается в том, что в качестве единицы для описания белковой последовательности были использованы короткие фрагменты полипептидной цепи. Такие короткие перекрывающиеся фрагменты были названы «информационными единицами». Теоретическое обоснование такого способа описания было получено в работе [2]. В этой работе были получены зависимости позиционной информационной энтропии как функции расстояния между аминокислотными остатками. Из полученных зависимостей видно, что при маленьких (≤ 5) расстояниях между остатками наблюдается постоянный и низкий уровень информационной энтропии. Эти данные служат обоснованием для нового способа описания полипептидной цепи и в этом случае всю первичную структуру белка можно рассматривать как систему перекрывающихся информационных единиц.

Материалы и методы

Процедура, описывающая АНИС-метод, состоит из нескольких последовательных шагов. Рассмотрим эти шаги. Первичная структура белка есть последовательность аминокислот A_i, $i = 1,...,N$, где аминокислоты могут быть 20-ти видов.

Сопоставим каждой последовательности аминокислот длины $M = 5$ следующую величину. Выберем некоторую базу данных, состоящую из белков. Последовательности $S = S_1...S_M$ из M аминокислот сопоставим частоту $f(S)$ встречаемости в качестве всевозможных подпоследовательностей, стоящих рядом аминокислот во всех белках из рассмотренной базы данных.

Выберем теперь набор последовательностей S' длины M, отличающихся от S заменой одной аминокислоты (таких последовательностей существует 20^M штук). Последовательности S' сопоставим соответствующую частоту $f(S')$. Просуммируем по всевозможным получаемым заменой одной аминокислоты последовательностям S', отвечающим S, и получим функцию

$$F(S) = \sum_{S'} f(S').$$

Теперь для данного рассматриваемого белка $P = \{A_i\}$ длины N мы будем рассматривать всевозможные подпоследовательности $S \subset P$ длины M стоящих рядом аминокислот, таких подпоследовательностей в белке длины N будет $N - M + 1 = N - 4$ штук. Занумеруем эти последовательности $S = S_i$ длины 5 их центрами i, таким образом, $i = 3,...,N-2$, и рассмотрим функцию

$$F(i) = F(S_i),$$

то есть суммарную частоту встречаемости в базе данных всевозможных подпоследовательностей S' длины 5, отвечающих подпоследовательности S с центром в аминокислоте с номером i в данном белке.

Нелинейное сглаживание частоты подпоследовательностей

Ранее для белка из N аминокислот была построена функция $F(i)$, $i = 3,...,N-2$ частоты встречаемости в базе данных подпоследовательности длины 5. Сопоставим этой функции гистограмму, то есть функцию $F(x)$ на отрезке $(2, N-2]$, принимающую для $x \in (i-1, i]$ значение $F(x) = F(i)$

Построим теперь по гистограмме $F(x)$ функцию нелинейного сглаживания $G(a,x)$ по следующему правилу.

Рассмотрим сглаживающую функцию $f(x)$ - непрерывную функцию с носителем на отрезке $[-1/2, 1/2]$, $f(-1/2) = f(1/2) = 0$, $f(0) = 1$, f принимает положительные значения в интервале $(-1/2, 1/2)$, монотонно растёт на $[-1/2, 0]$, монотонно убывает на $[0, 1/2]$, график функции симметричен относительно отображения относительно прямой $x = 0$. Мы также считаем, что функция f гладкая, причём производная не обращается в нуль на отрезках $(-1/2, 0)$ и $(0, 1/2)$.

В качестве сглаживающей функции можно выбрать соответствующим образом сдвинутую и перерастянутую гауссовскую функцию e^{-x^2}, график которой обрезан на половине высоты.

Будем также рассматривать сдвиги и растяжения сглаживающей функции

$$f^{(a,b)}(x) = f\left(\frac{x-b}{a}\right),$$

где $a \geq 1$. Функция $f^{(a,b)}$ имеет носитель в отрезке $[-1/2a + ba, 1/2a + ba]$.

Определим теперь функцию нелинейного сглаживания $G(x,a)$ для функции F по следующей формуле

$$G(b,a) = \sup c, \quad c: \quad cf^{(ab)}(x) \leq F(x), \quad \forall x.$$

Таким образом, $G(b,a)$ есть максимальная высота $\sup c$ сглаживающей функции шириной a с центром носителя в точке b, которую можно вписать в гистограмму F. Параметр a назовём масштабом сглаживания.

Носитель функции $G(x,a)$ имеет следующий вид. Функция $G(x,a)$ может быть отлична от нуля при $a \in [1, N-4]$, $x \in [2 + a/2, N - 2 - a/2]$. Таким образом, функция нелинейного сглаживания имеет носитель, являющийся подмножеством треугольника на плоскости с координатами (x,a) с вершинами $(N/2, N-4)$, $(2 + 1/2, 1)$, $(N - 2 - 1/2, 1)$.

Информационное пространство белка

Пусть $G(x,a)$ есть функция сглаживания, описанная выше. Построим на плоскости (x,a) скелет функции сглаживания $G(x,a)$, то есть нанесём на плоскость все локальные максимумы функции G по x при всевозможных фиксированных a (Рис. 1). Такой скелет будет древовидной структурой (то есть

если мы заменим точки слияния линий локальных максимумов вершинами, а соединяющие их линии локальных максимумов ребрами, то полученный граф будет деревом).

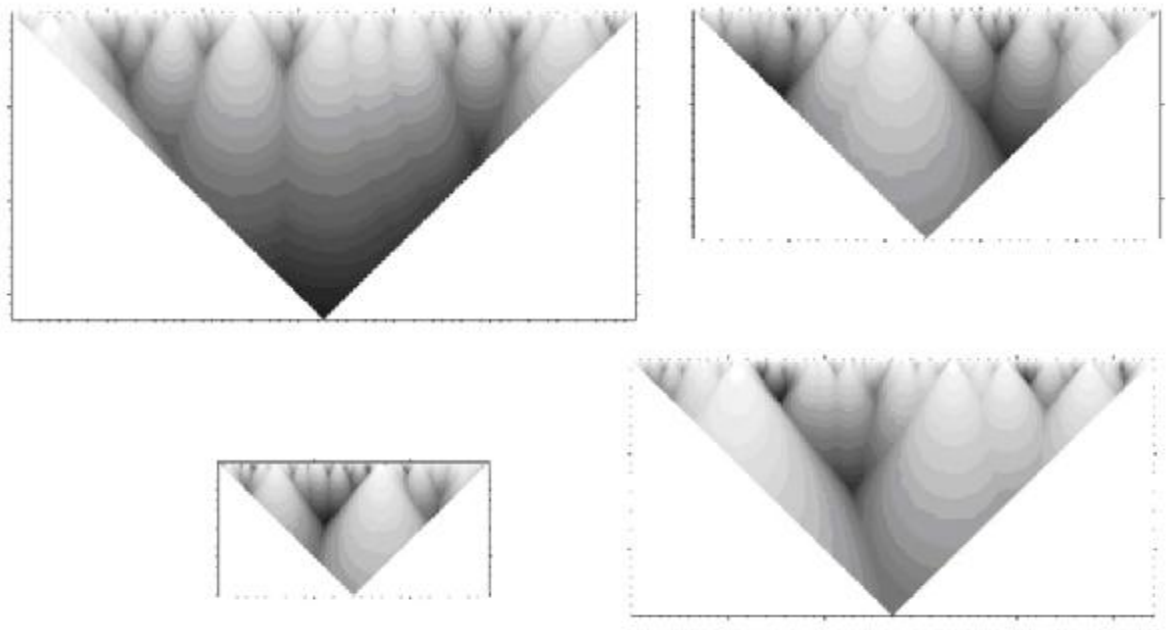

Рис. 1 Примеры расчетов информационных структур различных белков. На рисунках видны иерархически организованные элементы информационной структуры (ЭЛИС).

Точнее, по описанной древовидной структуре можно построить конечное дерево t, вершинами которого будут точки слияния линий древовидной структуры, а также локальные максимумы функции G при минимальной ширине сглаживающей функции $a=1$. Рёбрами дерева t будут линии, соединяющие вершины. Если процедура даёт более одной древовидной структуры (и соответствующих деревьев), то мы добавим к полученному конечному числу деревьев одну вершину, отвечающую максимальному значению $a=N$, и соединим эту вершину рёбрами с каждой из максимальных вершин построенных деревьев.

На дереве t определён естественный частичный порядок, относительно которого соединённые ребром вершины сравнимы, причем из этих вершин больше та, которая отвечает большим масштабам a. На построенном дереве определена функция a (масштаб, отвечающий вершине), растущая относительно введённого частичного порядка.

Полученное дерево t будем называть информационным деревом белка или элементами информационной структуры (ЭЛИС). Набор X локальных максимумов функции G при ширине сглаживающей функции $a=1$ будет границей информационного дерева белка (а также множеством минимальных вершин дерева относительно рассмотренного частичного порядка).

Множество X является ультраметрическим пространством относительно естественной ультраметрики, вводимой следующим образом. Ультраметрика $d(\cdot,\cdot)$ на X принимает для двух точек $A,B\in X$, $A\neq B$ значение

$$d(A,B)=a(\sup(A,B)).$$

Здесь $\sup(A,B)$ есть точная верхняя грань точек A, B в дереве t относительно введённого выше частичного порядка (точка слияния локальных максимумов A и B в древовидной структуре), $a(\sup(A,B))$ есть масштаб сглаживающей функции, при котором сливаются локальные максимумы A и B функции сглаживания.

Для совпадающих точек доопределим

$$d(A,A)=0.$$

Введённое пространство X будем называть информационным пространством белка, ультраметрику $d(\cdot,\cdot)$ информационной метрикой. Всё пространство X имеет конечный диаметр и конечное число точек.

Результаты и обсуждение

На основе описанной выше методики был разработан “ANIS-trees” веб сервис, позволяющий выявлять при анализе аминокислотной последовательности белка ЭЛИС (рис. 2).

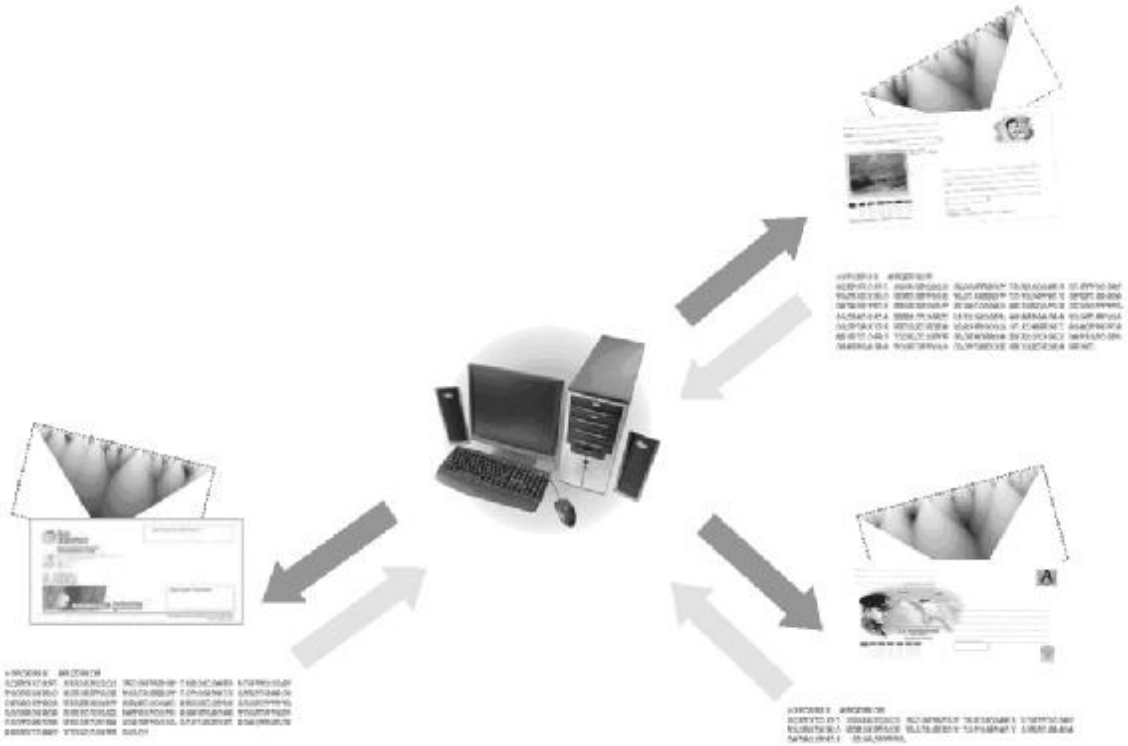

Рис. 2 Схема работы “ANIS-trees” веб сервиса (anis.ibch.ru/trees)

Входной информацией для сервера является последовательность белка, записанная в формате FASTA. Размер последовательности должен превышать 10 аминокислотных остатков. Результатом работы веб-сервиса “ANIS-trees” является файл в формате BMP, который после окончания расчета автоматически высылает по адресу электронной почты указанному при регистрации. На приведенных ниже иллюстрациях по оси абцисс расположена координата вдоль первичной последовательности белка (номер аминокилоты в первичной последовательности), а по оси ординат расположена ширина

сглаживающей функции. Одному пикселю осей соответствует один аминокислотный остаток.

Было показано, что ЭЛИС соответствуют структурно устойчивым элементам пространственной организации белка (в частности, доменам), удаление которых минимально влияет на механизм формирования пространственной структуры модифицируемого белка. Таким образом, структура иерархически организованных ЭЛИС есть обобщение доменной структуры белка. Данные "ANIS-trees" веб сервиса и могут быть использованы в работах по белковой инженерии, что продемонстрировали уже полученные результаты [3-4].

Выводы

В рамках иерархического подхода к анализу первичных структур белков:

1. Предложено описание структуры белка при помощи дерева ЭЛИС.
2. Построен математический аппарат, позволяющий выявлять иерархически организованные элементы ЭЛИС в белковых последовательностях.
3. Разработан программный пакет и веб сервис реализующий поиск ЭЛИС.

Литература

1. Entropy of Protein Sequences: An Integral Approach / A.N. Nekrasov // Journal of Biomolecular Structure & Dynamics. – 2002. – Vol. 20, № 1. – P. 87-92.

2. Analysis of the Information Structure of Protein Sequences: A New Method for Analyzing the Domain Organization of Proteins / A.N. Nekrasov // Journal of Biomolecular Structure & Dynamics. – 2004. – Vol. 21, № 5. – P. 615-623.

3. The Novel Approach to the Protein Design: Active Truncated Forms of Human 1-CYS Peroxiredoxin / A.N. Nekrasov [et al.] // Journal of Biomolecular Structure & Dynamics. – 2007. – Vol. 24, № 5. – P. 455-461.

4. Применение метода анализа информационной структуры для конструирования антагониста интерлейкина-13 / А.Н. Некрасов [и др.] // Биохимия. – 2009. – Т. 74, № 4. – С. 493 – 500.

Козырев Сергей Владимирович, ведущий научный сотрудник, Отдел математической физики, Математический институт им. В.А. Стеклова РАН, докт.физ.-мат.наук, kozyrev@mi.ras.ru

Козьмин Юрий Петрович старший научный сотрудник лаборатории протеомики, Института биоорганической химии имени М.М. Шемякина и Ю.А. Овчинникова РАН, ypko@ibch.ru

Богатова Ольга Викторовна, студентка 6 курса, Московский Физико-Технический Институт (Государственный Университет), Факультет молекулярной и биологической физики, bogatova.olga@gmail.com

Гарковенко Алексей в, младший научный сотрудник, лаборатория белков гормональной регуляции Института биоорганической химии имени М.М. Шемякина и Ю.А. Овчинникова РАН, alexey@garkovenko.ru

Некрасов Алексей Норбертович, старший научный сотрудник лаборатории биотехнологии Института биоорганической химии имени М.М. Шемякина и Ю.А. Овчинникова РАН, канд.физ.-мат.наук, alexei_nekrasov@mail.ru