С. В. Козырев
МИАН им. В.А.Стеклова

2-Мерная 2-адическая модель генетического кода

Конференция по биоматематике
27–28 октября, 2011, ИВМ РАН, Москва

Иерархические методы в теории сложных систем
иерархические системы —
деревья, ультраметрический анализ, p-адические числа

Генетический код – 2-адическая плоскость
(иерархия и многомерность)

p-Адические числа

p-Адическая норма рационального числа: для простого p

$$x = p^{\gamma}\frac{m}{n}, \quad |x|_p = p^{-\gamma}, \qquad |0|_p = 0.$$

Поле $\mathbb{Q}_p$ p-адических чисел —
пополнение $\mathbb{Q}$ по p-адической норме
p-адические числа — ряды (сходящиеся в p-адической норме)

$$x = \sum_{i=\gamma}^{\infty} x_i p^i, \quad x_i = 0, \ldots, p-1.$$

Кольцо $\mathbb{Z}_p$ целых p-адических чисел —
шар диаметра 1: $|x|_p \leq 1$ (ряды по степеням p с $\gamma \geq 0$)

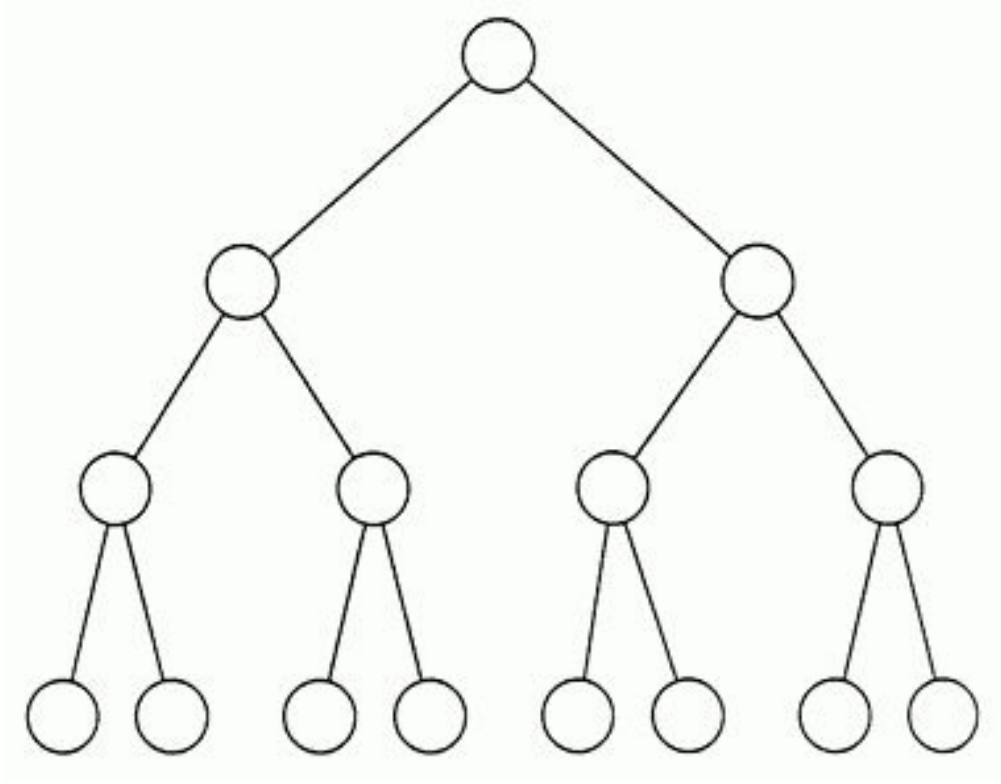
Пример: тождество в $\mathbb{Q}_2$

$$1 + 2 + 4 + \cdots = \sum_{i=0}^{\infty} 2^i = \frac{1}{1-2} = -1.$$

Ультраметричность (сильное неравенство треугольника)

$$|x + y|_p \leq \max(|x|_p, |y|_p)$$

все треугольники равнобедренны
два шара либо не пересекаются, либо один содержит другой
шары вложены друг в друга иерархично
(дерево иерархии шаров, $p = 2$, размерность один)



Общие ультраметрические пространства

метрические пространства, метрика $d(\cdot, \cdot)$
удовлетворяет сильному неравенству треугольника

$$d(x, y) \leq \max(d(x, z), d(y, z)), \qquad \forall x, y, z.$$

Двойственность между ультраметрическими пространствами и деревьями

Множество шаров в ультраметрическом пространстве —
частично упорядоченное дерево (порядок по вложению шаров)
вершины дерева — шары,
шары соединены ребром, если один из них
есть максимальный подшар в другом

граница дерева с соответствующим частичным порядком
(в любом конечном пути существует единственная
максимальная вершина) — ультраметрическое пространство

Многомерные p-адические пространства $\mathbb{Q}_p^d$ с метрикой (стандартной)

$$d(x, y) = \max_{i=1,\ldots,d}(|x_i - y_i|_p),$$

$$x = (x_1, \ldots, x_d), y = (y_1, \ldots, y_d).$$

(расстояние есть максимум от норм разностей координат)
Шар совпадает с кубом в любой размерности

Мера Хаара: инвариантность относительно сдвигов, нормировка:

$$\{x : |x|_p \leq 1\} = \{\sum_{i=0}^{\infty} x_i p^i\}$$

шар с мерой единица ($|\cdot|_p$ — норма, отвечающая стандартной метрике).

Деформированная метрика

$$d_{q_1,\ldots,q_d}(x,y) = \max_{i=1,\ldots,d}(q_i|x_i - y_i|_p), \qquad p^{-1} < q_i \leq 1,$$

$$x = (x_1,\ldots,x_d), y = (y_1,\ldots,y_d).$$

Пример: 2-адический 2-мерный случай,
метрика $d_{1,q}$, $1/2 < q < 1$,
единичный шар $\mathbb{Z}_2^2$ делится на:

два максимальных подшара диаметра q

$$\begin{array}{|c|}\hline 2\mathbb{Z}_2 \times \mathbb{Z}_2 \\ \hline 2\mathbb{Z}_2 \times \mathbb{Z}_2 \\ \hline \end{array},$$

четыре подшара диаметра $1/2$

$$\begin{array}{|c|c|}\hline 2\mathbb{Z}_2 \times 2\mathbb{Z}_2 & 2\mathbb{Z}_2 \times 2\mathbb{Z}_2 \\ \hline 2\mathbb{Z}_2 \times 2\mathbb{Z}_2 & 2\mathbb{Z}_2 \times 2\mathbb{Z}_2 \\ \hline \end{array},$$

восемь подшаров диаметра $q/2$

$4\mathbb{Z}_2 \times 2\mathbb{Z}_2$	$4\mathbb{Z}_2 \times 2\mathbb{Z}_2$
$4\mathbb{Z}_2 \times 2\mathbb{Z}_2$	$4\mathbb{Z}_2 \times 2\mathbb{Z}_2$
$4\mathbb{Z}_2 \times 2\mathbb{Z}_2$	$4\mathbb{Z}_2 \times 2\mathbb{Z}_2$
$4\mathbb{Z}_2 \times 2\mathbb{Z}_2$	$4\mathbb{Z}_2 \times 2\mathbb{Z}_2$

,

16 подшаров диаметра $1/4$

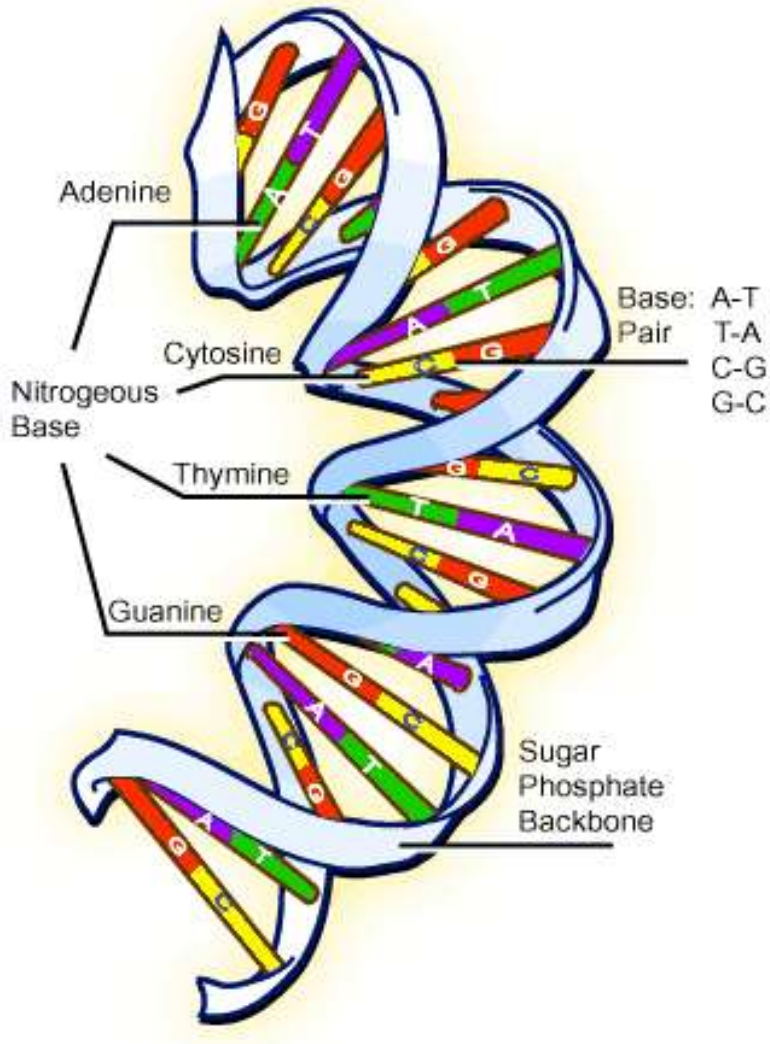
$4\mathbb{Z}_2 \times 4\mathbb{Z}_2$	$4\mathbb{Z}_2 \times 4\mathbb{Z}_2$	$4\mathbb{Z}_2 \times 4\mathbb{Z}_2$	$4\mathbb{Z}_2 \times 4\mathbb{Z}_2$
$4\mathbb{Z}_2 \times 4\mathbb{Z}_2$	$4\mathbb{Z}_2 \times 4\mathbb{Z}_2$	$4\mathbb{Z}_2 \times 4\mathbb{Z}_2$	$4\mathbb{Z}_2 \times 4\mathbb{Z}_2$
$4\mathbb{Z}_2 \times 4\mathbb{Z}_2$	$4\mathbb{Z}_2 \times 4\mathbb{Z}_2$	$4\mathbb{Z}_2 \times 4\mathbb{Z}_2$	$4\mathbb{Z}_2 \times 4\mathbb{Z}_2$
$4\mathbb{Z}_2 \times 4\mathbb{Z}_2$	$4\mathbb{Z}_2 \times 4\mathbb{Z}_2$	$4\mathbb{Z}_2 \times 4\mathbb{Z}_2$	$4\mathbb{Z}_2 \times 4\mathbb{Z}_2$

.

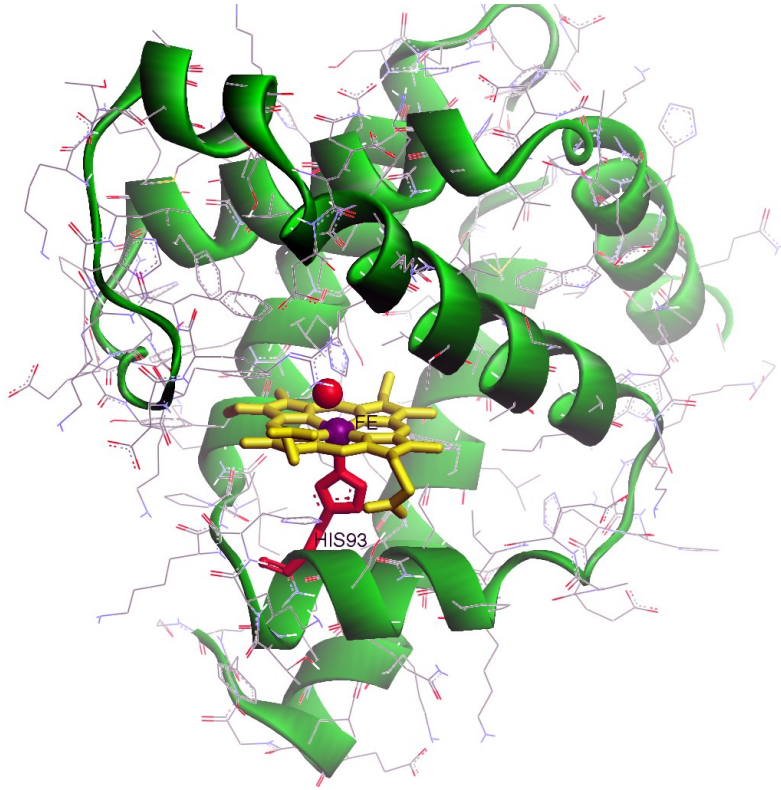
То есть дробление на подшары происходит сначала по одной координате, потом по другой
(для стандартной метрики — по двум координатам одновременно).

Биополимеры — нуклеиновые кислоты и белки

ДНК — цепь нуклеотидов, двойная спираль



Белок — цепь аминокислот, свёрнутая в компактную глобулу (нативное состояние)



Миоглобин

последовательность аминокислот белка кодируется на ДНК при помощи генетического кода

Генетический код

ДНК (РНК) — конечные последовательности
(линейные полимеры) нуклеотидов
белок — конечная последовательность аминокислот

Нуклеотиды: C, A, T (или U), G
Cytosine, Adenine, Thymine (или Uracil), Guanine

Триплетная структура генетического кода (Гамов)
Кодон — тройка $C_1 C_2 C_3$ нуклеотидов
Генетический код — переводит кодоны в аминокислоты

20 аминокислот и 1 символ остановки
4 нуклеотида, $4 \times 4 \times 4 = 64$ кодонов
$64 > 21$ — проблема вырождения генетического кода

AAA Lys AAU Asn AAG Lys AAC Asn	UAA Ter UAU Tyr UAG Ter UAC Tyr	GAA Glu GAU Asp GAG Glu GAC Asp	CAA Gln CAU His CAG Gln CAC His
AUA Met AUU Ile AUG Met AUC Ile	UUA Leu UUU Phe UUG Leu UUC Phe	GUA Val GUU Val GUG Val GUC Val	CUA Leu CUU Leu CUG Leu CUC Leu
AGA Ter AGU Ser AGG Ter AGC Ser	UGA Trp UGU Cys UGG Trp UGC Cys	GGA Gly GGU Gly GGG Gly GGC Gly	CGA Arg CGU Arg CGG Arg CGC Arg
ACA Thr ACU Thr ACG Thr ACC Thr	UCA Ser UCU Ser UCG Ser UCC Ser	GCA Ala GCU Ala GCG Ala GCC Ala	CCA Pro CCU Pro CCG Pro CCC Pro

митохондриальный генетический код

2-Адическая плоскость кодонов

1) Занумеруем нуклеотиды парами 0,1

$$\begin{array}{|c|c|}\hline A & G \\ \hline U & C \\ \hline \end{array} = \begin{array}{|c|c|}\hline 00 & 01 \\ \hline 10 & 11 \\ \hline \end{array} \qquad (A)$$

Химический смысл:
1-ая строка — пурины
2-ая строка — пиримидины
1-ый столбец — слабая H-связь
2-ой столбец — сильная H-связь

2) Порядок нуклеотидов в кодоне

$$2 > 1 > 3 \qquad (B)$$

3) 2-Адическая плоскость — группа $\mathbb{Z}/8\mathbb{Z} \times \mathbb{Z}/8\mathbb{Z}$
с координатами (x, y):

$$x = (x_0 x_1 x_2) = x_0 + 2x_1 + 4x_2, \quad x_i = 0, 1,$$

$$y = (y_0 y_1 y_2) = y_0 + 2y_1 + 4y_2, \quad y_i = 0, 1.$$

с 2-мерной 2-адической метрикой

$$d_{1,q}((x, y), (x', y')) = \max(|x - x'|_2, q|y - y'|_2), \quad 1/2 < q < 1.$$

4) Отображение кодонов

ρ переводит кодон в точку 2-адической плоскости

$$\rho : C_1 C_2 C_3 \mapsto (x, y) = (x_0 x_1 x_2, y_0 y_1 y_2),$$

C_2 определяет пару (x_0, y_0),
C_1 определяет пару (x_1, y_1),
C_3 определяет пару (x_2, y_2).

Нуклеотиды определяют пары цифр по правилу (A), порядок пар задаётся правилом (B).

5) Перенумеровка строк и столбцов

Занумеруем строки и столбцы 2-адической плоскости:

$$\eta : x \mapsto \widetilde{x}, \qquad y \mapsto \widetilde{y};$$

$$\eta : x_0 + 2x_1 + 4x_2 \mapsto 1 + 4x_0 + 2x_1 + x_2;$$

$$\eta : y_0 + 2y_1 + 4y_2 \mapsto 1 + 4y_0 + 2y_1 + y_2.$$

Эквивалентно:

$$\eta : 0, 4, 2, 6, 1, 5, 3, 7 \mapsto 1, 2, 3, 4, 5, 6, 7, 8.$$

(чтобы близкие 2-адически элементы
были расположены в соседних клетках плоскости)

2-Адическая плоскость кодонов

AAA	*AAG*	*GAA*	*GAG*	*AGA*	*AGG*	*GGA*	*GGG*
AAU	*AAC*	*GAU*	*GAC*	*AGU*	*AGC*	*GGU*	*GGC*
UAA	*UAG*	*CAA*	*CAG*	*UGA*	*UGG*	*CGA*	*CGG*
UAU	*UAC*	*CAU*	*CAC*	*UGU*	*UGC*	*CGU*	*CGC*
AUA	*AUG*	*GUA*	*GUG*	*ACA*	*ACG*	*GCA*	*GCG*
AUU	*AUC*	*GUU*	*GUC*	*ACU*	*ACC*	*GCU*	*GCC*
UUA	*UUG*	*CUA*	*CUG*	*UCA*	*UCG*	*CCA*	*CCG*
UUU	*UUC*	*CUU*	*CUC*	*UCU*	*UCC*	*CCU*	*CCC*

с метрикой

$$d_{1,q}(C_1C_2C_3, C_1'C_2'C_3') = \max(|x - x'|_2, q|y - y'|_2), \quad 1/2 < q < 1,$$

$$(x, y) = \rho(C_1C_2C_3), \quad (x', y') = \rho(C_1'C_2'C_3').$$

Применим к плоскости кодонов митохондриальный код:

Lys / Asn	Glu / Asp	Ter / Ser	Gly
Ter / Tyr	Gln / His	Trp / Cys	Arg
Met / Ile	Val	Thr	Ala
Leu / Phe	Leu	Ser	Pro

В частности

AAA	*AAG*	→	Lys	,	*CCA*	*CCG*	→	Pro
AAU	*AAC*		Asn		*CCU*	*CCC*		

Вырождение генетического кода описывается
2-адической 2-мерной метрикой:
кодоны, отображающиеся на одинаковые аминокислоты —
шары относительно метрики $d_{1,q}(\cdot,\cdot)$.

Симметрия Румера

между множествами кодонов

с сильным и слабым вырождением кода

— центральная симметрия 2-адической плоскости

$\frac{*}{*}$	$\frac{*}{*}$	$\frac{*}{*}$	Gly
$\frac{*}{*}$	$\frac{*}{*}$	$\frac{*}{*}$	Arg
$\frac{*}{*}$	Val	Thr	Ala
$\frac{*}{*}$	Leu	Ser	Pro

Физико–химические свойства (гидрофобность, полярность) кластеризуются в 2-адической норме

Гидрофобные аминокислоты

——	——	——	
——	——	$\frac{\text{Trp}}{\text{Cys}}$	
$\frac{\text{Met}}{\text{Ile}}$	Val		
$\frac{\text{Leu}}{\text{Phe}}$	Leu		

гидрофобные аминокислоты кластеризуются в два шара

похожие аминокислоты 2-адически близки

PAM–матрица A (матрица скоростей замен для марковской модели эволюции белков точечными заменами аминокислот) допускает разложение

$$A = A^{(2)} + A^{(\infty)}$$

где $A^{(2)}$ 2-адически регулярна (матричные элементы близки к локально постоянным относительно 2-адической параметризации),
матрица $A^{(\infty)}$ разреженна (мало ненулевых матричных элементов).

Основные ненулевые элементы $A^{(\infty)}$ связаны с переходами между аминокислотами

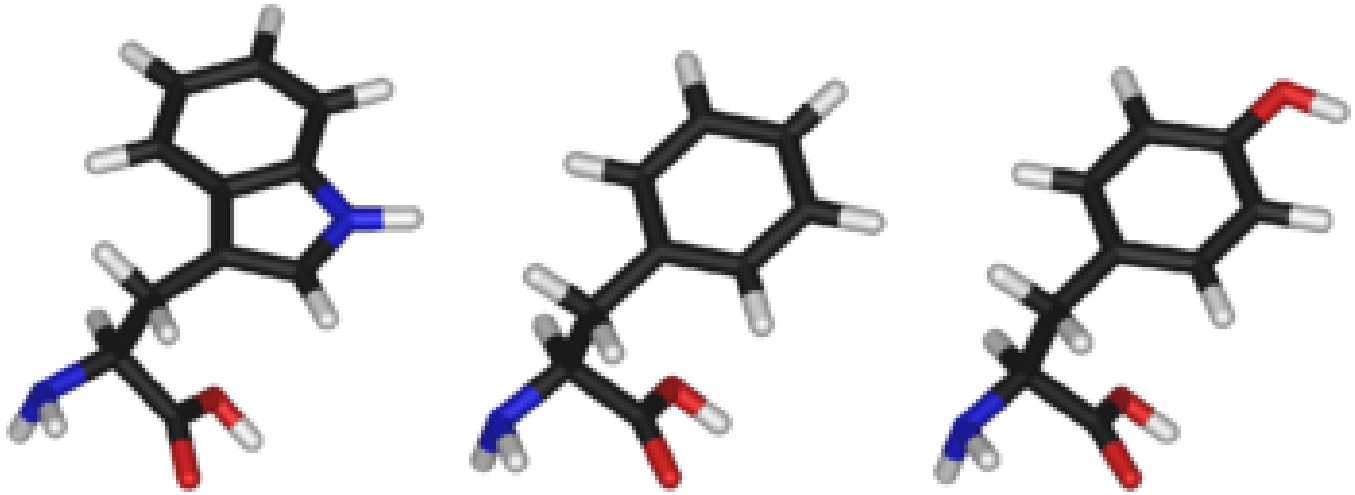


Рис.: tryptophan, phenylalanine, tyrosine

близкими геометрически но разными химически.
Можно выделить вклад в PAM–матрицу от геометрии.

p-Адическая параметризация генетического кода

по существу алгебраическая структура, что ставит под сомнение существенность эволюции для формирования свойств кода (исправление ошибок и т.д.)

близкие химически (гидрофобные либо полярные) аминокислоты 2-адически близки

известные симметрии кода (симметрия Румера) имеют естественное 2-адическое выражение

PAM–матрица 2-адически регулярна (что возможно связано с кластеризацией химически сходных аминокислот)