UDC 519.95

ОБ ОЦЕНКЕ РАЗЛИЧИЙ В ГЕНОМЕ ЧЕЛОВЕКА

Акбаров Б. Х
Университет точных и социальных наук, Национальный университет Узбекистана, Ташкент
bahriddin.akbarov@gmail.com

АННОТАЦИЯ

Предложены методика для поиска различий в геноме человека. Поиск различий производился по 22 хромосомам и последовательностям X, Y, MT. Используется поиск скрытых закономерностей методами вычисления весов и устойчивости признаков. Ограничения на множество допустимых значений устойчивости позволяют эффективно интерпретировать полученные результаты.

Ключевые слова: геном человека, нечёткие множества, устойчивость признака

Введение

Прошло более четверти века со дня сообщения в июне 2000 года в сенате США исследователями Фрэнсисом Коллинзом и Крейг Вентером об расшифровке генома человека. Основная роль в процессе расшифровки отводилась использованию вычислительных методов и компьютерной техники. Ф. Коллинз применял классический иерархический метод. К. Вентер использовал полный метод, реализованный на основе новых алгоритмов:

- выравнивания;
- поиска покрытий;
- вероятностей оценки ошибок;
- оптимизации сборки.

Основные вычислительные подходы связаны с:

- динамическим программированием (например, для сравнения последовательностей по алгоритмам Смит-Уотерман и Нидлман-Вунша);
- графовыми методами;
- статистическими моделями и байесовскими подходами для исправления ошибок.

После построения и выделения хромосом большое внимание было уделено поиску различий между геномами разных людей. В итоге были получены следующие показатели различий.
Наиболее сильные - это SNP + CNV + STR + структурные перестройки.
Самые частые - SNP и STR.
Самые масштабные и “сильные” - CNV и крупные структурные вариации.
Наибольшее медицинское значение имеют InDel и SNP в функциональных областях генома.
Основные показатели различий:

- любые два человека приблизительно отличаются на 0.1% генома (3 миллиона нуклеотидов);
- 99.9% генома у всех людей одинаковый.

Самые сильные различия находятся в:

- участках, отвечающих за иммунную систему (например, HLA-гены);

- генах восприятия запахов;

- участках, связанных с метаболизмом и адаптацией к пище.

Наиболее крупная база данных хранится в [1]. Наличие категорий позволяет расширить возможности для проверки (выявления) скрытых закономерностей из баз данных.

Например, в [2] при анализе структуры ДНК на предмет влияния мутаций в качестве категорий использовались показатели РНК как “нет мутации”, “одна мутация”, “две мутации”. Поиск различий также строился по 4 парам классов (без мутаций, одна и две мутации), (без мутаций, две мутации), (без мутаций, одна мутация), (одна мутация, две мутации). По всем 4 парам определены информативные наборы признаков. Единственность наборов связана с использованием жадных алгоритмов многокритериальной агломеративной группировки [3]. Решения задачи с многими критериями позволило избежать полного перебора всех возможных вариантов.

Реализация методов вычисления обобщённых оценок [4] основана на идее противопоставления описаний объектов двух классов как оппозиции друг другу. При вычислении показателей сравнения с оппозицией используются значения функций принадлежности к нечетким множествам. Проявлением феномена нечёткости является устойчивость признаков по значениям функции принадлежности. Через устойчивость идёт поиск и выражение закономерностей, присущих всем выборкам из генеральной совокупности. Номера хромосом в данном исследовании предложено использовать в качестве категорий генома.

Математическая модель

Имеются две последовательности геномов от двух индивидуумов, представленных как данные по 22 хромосомам, аллосомам (X, Y) и митохондриальным геномом (MT) из [5]. Предметом анализа является наборы из 19 генов в виде комбинаций 4 аминокислот, участвующих в формировании структуры белка. В описании данных есть неизмеренные значения(пропуски). Разбиения на 25 категорий для каждого индивидуума считается отдельным классом. Предлагается сравнение двух индивидуумов по 25 парам категорий.

Введем следующие обозначения по r-ой категории для количества:

K_{dr} - измеренных значений (без пропусков) для d - го индивидуума, $d = 1, 2$;

- p_r градаций по категории;

- l_{dr} различных градаций для d - го индивидуума;

- g_{dr}^t объектов с градацией $t, t = 1, 2, \ldots, p_r$.

Значение межклассового различия и внутриклассового сходства вычисляются как

$$\lambda_r = 1 - \frac{\sum\limits_{t=1}^{p_r} g_{1r}^t g_{2r}^t}{|K_{1r}||K_{2r}|},$$

$$\beta_r = \begin{cases} \dfrac{\sum\limits_{i=1}^{l} g_{1r}^i(g_{1r}^i - 1) + g_{2r}^i(g_{2r}^i - 1)}{D_{1r} + D_{2r}}, & D_{1r} + D_{2r} > 0, \\ 0, & D_{1r} + D_{2r} = 0, \end{cases}$$

где $D_{dr} = \begin{cases} (|K_{dr}| - l_{dr} + 1)(|K_{dr}| - l_{dr}), & p_r > 2, \\ |K_{dr}|(|K_{dr}| - 1), & p_r \le 2. \end{cases}$

Вес признака по (1), (2)

$$w_r = \lambda_r \beta_r.$$

При вычислении функции принадлежности $f_r(\mu)$ к классу K_{1r} по градации $\mu \in \{1, 2, \ldots, p_r\}$ в качестве $d_{1r}(\mu)(d_{2r}(\mu))$ используется число объектов класса $K_{1r}(K_{2r})$ со значением μ

$$f_r(\mu) = \frac{d_{1r}(\mu) / |K_{1r}|}{d_{1r}(\mu) / |K_{1r}| + d_{2r}(\mu) / |K_{2r}|}.$$

Граница между объектами классов по значениям функции принадлежности (4) определяется как $G_r = (q1 + q2)/2$, где $q2 = \max\{f_r(\mu)|0.5 - f_r(\mu) > 0, \mu = 1, ..., p_r\}, q1 = \min\{f_r(\mu)|1 - f_r(\mu) < 0.5, \mu = 1, ..., p_r\}$.

Так как число градаций $p_r, p_r \geq 2$ является постоянной величиной, то устойчивость признака по r-ой категории будет определятся так

$$\varphi(r) = \frac{1}{|K_{1r}| + |K_{2r}|} \sum_{i=1}^{p_r} \begin{cases} f_r(i)\, t_i, f_r(i) \geq 0.5, \\ (1 - f_r(i))\, t_i, f_r(i) < 0.5, \end{cases}$$

где t_i - количество значений номинального признака в $K_{1r} \cup K_{2r}$ с градацией, равной i.

В процессе унификации значениям признака в описании объектов двух классов ставится в соответствие показатель его устойчивости (5) из интервала (0.5;1]. Множество допустимых значений из (0.5;1] формируются по нелинейным преобразованиям данных с использованием функций принадлежности к двум классам. Устойчивость признака является комбинаторной оценкой, вычисляемой на реальных выборках данных.

Показателем информативности признака служит близость значения его устойчивости к 1.0. Для редукции пространства необходимо использовать упорядоченную по устойчивости признаков последовательность. Доказанным фундаментальным свойством устойчивости является сходимость по вероятности к фиксированному значению на выборках из генеральной совокупности. Доказательство фундаментальности в виде теоремы [2] основывается на законе больших чисел применительно к количественным данным и теории нечётких множеств.

Общность интервальных и номинальной шкал измерений выражается в использовании алгоритма метода минимального покрытия значений количественного признака непересекающимся интервалами [6, 7]. Количество интервалов не является фиксированным значением на выборках из генеральной совокупности. Нечёткость выражается через значения функций принадлежности к классам в границах всех интервалов. Как закономерность значение устойчивости признаков используется с целью контроля корректности данных на выборках из генеральной совокупности.

Проблема учёта пропусков при вычислении устойчивости на номинальных (качественных) признаков исследовалась в работе [8]. Случайным образом формировались выборки с числам пропусков от 5% до 35%. Расхождения в значениях устойчивости по всем выборкам были в 5,6 знаках. Целью исследования является подтвердить малые различия между геномами людей с использованием весов и устойчивости признака.

Вычислительный эксперимент

В качестве данных для эксперимента использовались выборки из двух файлов Child1 Genome и Child2 Genome, состоящих соответственно из 601802 и 631983 объектов [5]. Информация по 25 категориям содержится в табл.1. Показатели Child1 Genome идентифицируются как класс K_1 и Child2 Genome как класс K_2. В скобках указано количество объектов по K_1 и K_2.

Таблица 1. Данные по категориям двух индивидуумов

№ хромосомы	Всего объектов	Всего пропусков	Число генов	Значения устойчивости
1	95591 (46656,48935)	1617 (567,1050)	13	0.5586
2	97464 (46127,51337)	1549 (487,1062)	13	0.5676
3	81121 (38516,42605)	1351 (358,993)	13	0.5648
4	73006 (33914,39092)	1357 (409,948)	13	0.5653
5	71111 (34384,36727)	1188 (402,786)	13	0.5704
6	84086 (40383,43703)	1881 (974,907)	13	0.5561
7	67078 (33051,34027)	1302 (464,838)	13	0.5592
8	61694 (30266,31428)	939 (301,638)	13	0.5689
9	52736 (26583,26153)	865 (275,590)	13	0.5622

10	59469 (29211,30258)	965 (332,633)	13	0.5645
11	60007 (29320,30687)	966 (347,619)	13	0.5550
12	57601 (28450,29151)	957 (306,651)	13	0.5555
13	43572 (21652,21920)	628 (237,391)	12	0.5555
14	38473 (18695,19778)	634 (190,444)	13	0.5638
15	37088 (18281,18807)	617 (192,425)	13	0.5546
16	39403 (19198,20205)	729 (258,471)	13	0.5528
17	37885 (18710,19175)	780 (314,466)	13	0.5509
18	34039 (16490,17549)	488 (151,337)	13	0.5642
19	27680 (12989,14691)	782 (303,479)	13	0.5497
20	29146 (14494,14652)	437 (157,280)	12	0.5618
21	16955 (8461,8494)	317 (112,205)	13	0.5624
22	17904 (9096,8808)	393 (154,239)	13	0.5560
X	35621 (19478,16143)	1195 (673,522)	17	0.5209
Y	5803 (2302,3501)	1371 (1004,367)	7	0.5205
MT	9252 (5095,4157)	506 (257,249)	6	0.5098

Близкое к 0.5 значения устойчивости (5) по MT (см. табл.1) объясняется малой разностью частот по 6-ти градациям классов K_1 и K_2 при вычислении функции принадлежности (4).

Другим способом доказательства различия по генотипам является использование показателя (1) (см. табл.2), множество допустимых значений которого принадлежит (0;1]. Чем ближе (1) к нулю, тем меньше различия между классами.

Таблица 2. Результаты анализа различий между классами K_1 и K_2

№ хромосомы	Межклассовое различие (1)	Внутриклассовое сходство (2)	Вес признака (3)
1	0.1735	0.8313	0.1442
2	0.1704	0.8370	0.1426
3	0.1701	0.8365	0.1423
4	0.1711	0.8361	0.1430
5	0.1702	0.8366	0.1423
6	0.1706	0.8346	0.1424
7	0.1687	0.8358	0.1409
8	0.1690	0.8374	0.1415
9	0.1681	0.8361	0.1405
10	0.1716	0.8339	0.1431
11	0.1693	0.8349	0.1413
12	0.1704	0.8335	0.1420
13	0.1610	0.8428	0.1356
14	0.1680	0.8375	0.1407
15	0.1679	0.8359	0.1404
16	0.1720	0.8320	0.1431
17	0.1713	0.8323	0.1425

18	0.1702	0.8359	0.1423
19	0.1834	0.8207	0.1505
20	0.1714	0.8338	0.1429
21	0.1679	0.8366	0.1405
22	0.1776	0.8262	0.1467
X	0.2269	0.7737	0.1755
Y	0.2489	0.7578	0.1886
MT	0.2628	0.7374	0.1938

При вычислении межклассового различия (1) частоты по градациям признака перемножаются. Этим объясняется максимальное значении (1) по MT из 25 пар классов.

Преимущества предложенного подхода

В отличие от классических статистических методов (χ^2-критерий, t-тест, ANOVA), методика вычисления устойчивости обладает рядом преимуществ:

- учитывает как количественные, так и номинальные признаки;
- не требует строгих предположений о законе распределений;
- демонстрирует сходимость по вероятности при увеличении объёма выборки;
- позволяет работать с данными, содержащими пропуски, благодаря использованию функций принадлежности к нечётким множествам.

Эти свойства делают методику удобной для анализа больших геномных баз данных, таких как 1000 Genomes Project или UK Biobank.

Заключение

Проведённые исследования показали, что использование показателя устойчивости и весов признака является эффективным инструментом при сравнительном анализе геномов различных индивидов. Полученные значения устойчивости по каждой хромосоме позволяют:

- выявлять статистически значимые различия между выборками;
- интерпретировать закономерности в распределении генотипов;
- уменьшать размерность исходного пространства признаков без потери информативности.

Результаты вычислительного эксперимента подтверждают применимость методики для практических задач медицинской генетики, таких как поиск мутаций, ассоциированных с наследственными заболеваниями, и построение индивидуальных карт генетических рисков.

ЛИТЕРАТУРА

1. P. Sudmant et al. “An integrated map of structural variation in 2,504 human genomes Nature vol. 526, pp. 75-81, Oct. 2015. doi:10.1038/nature15394.
2. Игнатьев Н.А., Акбаров Б.Х. Оценка близости структур отношений объектов обучающей выборки на многообразиях наборов латентных признаков // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2023. №65. С. 69-78. doi: 10.17223/19988605/65/7
3. Ignatev N. A. and Rahimova M. A. Formation and Analysis of Sets of Informative Features of Objects by Pairs of Classes // Scientific and Technical Information Processing, 2022, Vol. 49, №. 6, pp. 439-445.
4. Ignatev N. A. On Nonlinear Transformations of Features Based on the Functions of Objects Belonging to Classes // Pattern Recognition and Image Analysis. 2021. V. 31. №2. P. 197-204.

5. https://www.kaggle.com/datasets/zusmani/mygenome

6. Мадрахимов Ш.Ф. Системы обнаружения скрытых закономерностей на базе методов вычисления обобщенных оценок : дисс. докт. тех. наук. - Ташкент, 2020.

7. Згуральская Е.Н. Устойчивость разбиения данных на интервалы в задачах распознавания и поиск скрытых закономерностей // Известия Самарского научного центра РАН. - 2018. - Т. 20, №4(3). - С. 451-455.

8. Игнатьев Н.А., Рахимова М.А., Лолаев М.Я. Особенности отбора информативных наборов признаков на данных с пропусками // Проблемы вычислительной и прикладной математики. - 2021. - №6/1(37). - С. 113-122.

Annotatsiya

Inson genomida farqlarni aniqlash uchun metodika taklif qilingan. Farqlarni izlash 22 ta xromosoma hamda X, Y va MT ketma-ketliklari bo‘yicha amalga oshirilgan. Yashirin qonuniyatlarni topishda alomatlar vaznlari va turg’unlik qiymatarini hisoblash usullaridan foydalaniladi. Turg‘unlikning mumkin bo’lgan qiymatlar to‘plamiga qo‘yilgan cheklovlar olingan natijalarni samarali talqin qilish imkonini beradi.

Kalit so‘zlar: Inson genomi, qat’iymas to’plam, alomatlar turg’unligi

Abstract

A methodology for detecting differences in the human genome has been proposed. The search for differences was carried out across 22 chromosomes and the X, Y, and MT sequences. Hidden patterns are identified using methods of calculating feature weights and stability. Restrictions on the set of permissible stability values make it possible to effectively interpret the obtained results.

Key words: Human genome, fuzzy sets, feature stability