24. *Schroeder C., Hurkscat H., Meisel A. et al.* Unusual occurrence of *EcoP1* and *EcoP15* recognition sites and counterselection of type II methylation and restriction sequences in bacteriophage T7 DNA // Gene.— 1986.— **45**, N 1.— P. 77—86.
25. *Sekulic S., Haddad P. R.* Effect of peak tailing on computer optimization procedures for high-performance liquid chromatography // J. Chromatogr.— 1988.— **459**, N 1.— P. 65—77.
26. *Sinhalg R. P., Landes J. P.* High-performance liquid chromatographic analysis of DNA composition and DNA modification by chloroacetaldehyde // Ibid.— 1988.—**458**.— P. 117—128.

Ин-т микробиологии и вирусологии
АН Украины, Киев
ВНИИГенетика, Москва

Получено 18.01.93

УДК 577.112+371.24

А. В. Братусь, С. З. Мальченко, Н. А. Чащин

ПРЕДСКАЗАНИЕ ВТОРИЧНОЙ СТРУКТУРЫ БЕЛКОВ МОДИФИЦИРОВАННЫМ GUHA- МЕТОДОМ *

В работе предложен новый метод предсказания вторичной структуры белка, основанный на известном в области экспертных систем GUHA-методе. Обучение и предсказание базируются на информации о вторичной структуре 108 белков (около 20 000 аминокислотных остатков) с рентгеноструктурным разрешением менее 0,2 нм. Средняя точность предсказания по использованному в работе банку данных составила для α-спирали — 74 %, β-складки — 67 %, нерегулярной структуры — 71 % и общая — 68 %.

Введение. Известно, что предсказание вторичной структуры белка с точностью 75—85 % позволяет составить общее представление о его пространственной структуре, а увеличение точности — получить довольно близкую к реальной пространственную модель белка [1].

Существующие в настоящее время методы не всегда позволяют достичь необходимой точности, поэтому, несмотря на множество способов предсказания вторичной структуры белка, поиск новых подходов в этом направлении не прекращается.

В предлагаемой работе описывается известный в области экспертных систем GUHA-метод [2] и его приложение к предсказанию вторичной структуры белков.

Материалы и методы. В работе использовали банк данных, содержащий информацию о вторичной структуре 108 белков (20 000 аминокислотных остатков) с рентгеноструктурным разрешением менее 0,2 нм. Эти данные были получены из Брухевенского банка данных пространственных структур белков. Вторичная структура классифицирована по трем конформациям: α-спираль (h), β-складка (e) и нерегулярная (c). Таким образом, каждому аминокислотному остатку присваивается одно из трех состояний вторичной структуры — h, e или c.

Для прогнозирования вторичной структуры белка применен модифицированный GUHA-метод. Суть его состоит в следующем.

Пусть исследуемая предметная область отражается эмпирическими данными в виде таблицы. Формально таблица может быть представлена в таком виде:

$$M=\langle Mod,\ f_1, \ldots, f_n\rangle,$$

где Mod — множество объектов;

$f_i: Mod \to V_i$ — унарные отношения, определенные в данной предметной области;

$V_i=\{1, 2, \ldots, K_i, U_n\}$ — множество значений отношения;

f_i, U_n есть символ неопределенной информации.

* Статья представлена членом редколлегии В. И. Даниловым.

Язык описания предметной области состоит из символов:
$F_1, \ldots, F_n$, кодирующих отношения $f_1, \ldots, f_n$;
$(K)F_i$, (K) — символы логических операций;
$\tilde{}a$ — квантор Фишера.

Элементарное высказывание языка имеет вид $(K_i)F_i$, где K_i входит в V_i и называется литералом. Литерал оценивается по формуле $/(K_i)F_i/[o]=1$ (логически истинно), если K_i входит в $f_i(o)$ для данного объекта o из Mod. Конъюнкция литералов образует формулу языка вида

$$P = \bigwedge (K_i) F_i.$$

Каждая формула описывает конкретное состояние предметной области. Пусть даны два состояния предметной области, описываемые формулами $P1$ и $P2$. Исследователя может интересовать, существует ли связь между этими состояниями? Используя таблицу эмпирических данных, можно подсчитать число всех наблюдений, содержащих: оба состояния; одно состояние; другое состояние. Обработав полученные частоты по одному из критериев, можно проверить гипотезу о взаимозависимости данных состояний. Если формула $P1$ описывает цель исследования, а $P2$ пробегает множество состояний предметной области, можно получить множество состояний, связанных с целью исследования. Качество отбираемых гипотез зависит от применяемого критерия.

В данной работе выбран точный критерий проверки на независимость двух признаков, или критерий Фишера. Высказывание «Состояние, описываемое формулой $P1$, взаимосвязано с состоянием, описываемым формулой $P2$» соответствует формальной записи

$$\tilde{}a\,(P1,\ P2) \text{ или } P1\ \tilde{}aP2,$$

где $\tilde{}a$ — квантор Фишера; a — доверительный уровень.

Высказывание $P1$ и $\tilde{}aP2$ оценивается по формуле $/P1\tilde{}\ aP2/(M)=1$ (логически истинно) тогда и только тогда, когда $\mathrm{Fish}(P1,\ P2) = \Sigma\{i=s, \min(r, k)\}g(i, r, k, m) \leqslant a$,
где

$$s = \mathrm{card}\{o:/P1 \,\&\, P2/[o] = 1\};$$

$$r = \mathrm{card}\{o:/P1/[o] = 1\};$$

$$k = \mathrm{card}\{o:/P2/[o] = 1\};$$

$$m = \mathrm{card}\{Mod\};$$

$$g(i, r, k, m) = r!\,k!\,(m-r)!(m-k)!/m!\,i!\,(r-i)!\,(k-i)!\,(m+i-r-k)!;$$

$0 < a \leqslant 0{,}5$ — уровень принятия гипотез;
card — число элементов соответствующего множества.

Точное значение Fish называется критическим уровнем. Критический уровень равен вероятности ложности данной гипотезы, поэтому чем меньше критический уровень, тем вероятнее взаимосвязь между состоянием гипотезы. Этот факт будет использован при выборе вторичной структуры. Осуществлена программная реализация GUHA-метода, программа написана на языке FORTRAN для IBM PC.

Оценку точности предсказания по отдельным конформационным состояниям рассчитывали по формуле

$$Qs = 1/2\,[(Ts^{+}/Ts) + (Ts^{-+}/Ts^{-})],$$

где s — конформационное состояние (h, e, c);

(Ts) — число остатков в структуре s по данным рентгеноструктурного анализа;

(Ts^{+}) — число остатков в структуре s, совпадающих по предсказанию с данными рентгеноструктурного анализа;

(Ts^{-}) — число остатков, не входящих в структуру s по данным рентгеноструктурного анализа;

Результаты работы GUHA-метода на данных по 108 белкам

Наименование белка	Имя файла в БД	Результат, %			
		Qa	Qb	Qo	Q
Acid Proteinase Endothiapepsin	4APE1M	84	80	77	76
Acid Proteinase, Penicillopepsin (Hydrolase : Proteinase)	2APP1E	57	68	66	60
Actinidin (Hydrolase : Sulfhydryl Proteinase)	2ACT1M	58	58	57	42
Lectin (Agglutinin) Wheat Germ	3WGA1M	60	53	59	74
Alpha Lytic Protease (Hydrolase : Serine Proteinase)	2ALP1E	46	62	64	56
Aspartate Transcarbamylase (*E. coli*) Chain 1	4ATC1M	79	69	74	72
Aspartate Transcarbamylase (*E. coli*) Chain 2	4ATC2M	85	64	81	80
Azurin Electron Transport Protein	1AZA1E	68	65	73	62
Calcium-Binding Parvalbumin b (Calcium Binding Proteins)	1CPU1H	76	71	74	71
Calcium Binding Protein Bovine Intestine Vitamin D Dependent	3ICB1H	92	+	92	92
Carbonic Anhydrase Form B Human Erythrocytes	2CAB1E	87	72	76	73
Carboxypeptidase A (C-Terminal Amino Acid Hydrolase)	5CPA1M	69	54	64	58
Catalase Beef Liver	8CAT1M	71	61	66	63
Alpha Chymotrypsin A (Bos Taurus) Chain 1	5CHA1E	—	77	80	74
Alpha Chymotrypsin A (Bos Taurus) Chain 2	5CHA2M	73	77	80	71
Citrate Synthase Pig Heart	2CTS1H	72	62	69	62
Crambin (Plant Seed Protein)	1CRN1M	50	64	60	47
Gamma-Crystallin Calf Eye Lens	1GCR1E	47	58	61	55
Cytochrome C (Oxidized) (Electron Transport)	3CYT1H	72	—	66	66
Cytochrome C Rice Embryos	1CCR1M	89	—	84	85
Cytochrome C Prime (*Rhodospirillum molischianum*)	2CCY1H	83	+	83	79
Cytochrome C Peroxidase (Baker's Yeast)	2CYP1H	63	57	62	56
Ferricytochrome C2 (Electron Transport)	3C2C1H	69	—	70	67
Cytochrome C3 (*Disulfovibrio vulgaris*)	2CDU1M	70	59	62	64
Cytochrome C551 (Oxidized) (Electron Transport)	351C1H	87	+	87	87
Dihydrofolate Reductase (Oxidoreductase : NADPH/DONR)	3DFR1M	71	53	58	53
Elastase Porcine Pancreas	2EST1E	70	74	71	70
Erabutoxin Sea Snake Venom	2EBX1E	+	63	63	67
Hemoglobin (Erythrocruorin, Deoxy) (Oxygen Transport)	1ECD1H	66	—	68	52
Ferredoxin (Electron Transport)	1FDX1M	50	66	52	72
Ferredoxin (Electron Transport)	3FXC1M	98	76	82	90
Flavodoxin (Oxidized) (Electron Transport)	3FXN1M	68	66	60	57
Ferredoxin Azobacter	2FDI1M	73	70	71	77
Glutathione Reductase Bovine Erythrocytes	1GPI1M	67	53	55	57
Hemerythrin (Met) Sipunculid Worm	1HMQ1H	65	—	59	53
Hemoglobin (Human, Deoxy) Chain 1	2HHB1H	75	—	76	64
Hemoglobin (Human, Deoxy) Chain 2	2HHB2H	82	—	84	77
Hemoglobin V (Cyano, Met) Sea Lamprey	2LHB1H	71	—	78	63
Oxidized High Potential Iron Protein (Hipip)	1HIP1M	85	82	80	70
Immunoglobulin Fab Igg (Mouse) Chain 1	1MPC1E	49	66	65	64
Immunoglobulin Fab Igg (Mouse) Chain 2	1MPC2E	—	68	71	67
Immunoglobulin Fab (Human Myeloma) Chain 2	1FB42E	—	59	57	57
Bence-Jones Immunoglobulin Variable Portion (REI)	1REI1E	+	84	84	85
Bence-Jones Protein Lambda Variable Domain (Human)	2RHE1E	—	77	73	73
Kallikrein A (Porcine Pancrease) Chain 1	2PKA1E	+	72	72	75
Kallikrein A (Porcine Pancrease) Chain 2	2PKA2M	99	69	75	76
Lactate Dehydrogenase, Apo Enzyme M4	4LDH1M	66	68	64	65
Leghemoglobin (Acetate, Met) (Oxygen Transport)	1LH11H	68	—	70	69
Lysozyme (Bacteriophage T4)	2LZM1M	80	52	74	67
Lysozyme (Human)	1LZ11M	76	78	72	71
Myoglobin (Oxygen Storage) (Ferric Iron — Metmyoglobin)	1MBN1H	70	—	69	68
Melittin (Hemolytic Polypeptide)	1MLT1H	84	+	84	73
Scorpion Neurotoxin	1SN31M	99	70	81	87

Окончание таблицы

Наименование белка	Имя файла в БД	Результат, %			
		Qa	Gb	Qo	Q
Ovomucoid Third Domain (Proteinase Inhibitor, Kazal)	1OVO1M	66	51	67	57
Papain Sulfhydryl Proteinase (Papaya Fruit Latex)	1PPD1M	70	63	65	64
Phospholipase A2 (Phosphatide Acyl-Hydrolase)	1BP21M	59	46	55	52
Plastocyanin (Electron Transport, Copper Binding)	1PCY1E	44	77	64	64
Prealbumin (Thyroxin, Retinol Transport)	2PAB1E	89	65	74	64
Proteinase A (SGPA) (Hydrolase: Serine Proteinase)	2SGA1E	76	64	62	57
Serine Proteinase (Rat Mast Cell Protease)	3RP21E	88	75	75	75
Ribonuclease A (Bovine Pancrease)	1RN31M	77	68	71	65
Rubredoxin Iron-Sulfur Protein (Clostridium)	4RXN1M	+	79	79	83
Staphylococcal Nuclease	2SNS1M	86	56	73	66
Subtilysin BPN' (Hydrolase: Serine Proteinase)	1SBT1M	90	77	84	82
Cu, Zn Superoxide Dismutase (Oxidoreductase: Superoxide)	2SOD1E	—	69	68	68
Thermolysin (Hydrolase: Neutral Metallo-Proteinase)	3TLN1M	77	66	66	63
Beta Trypsin (Bovine) Orthorombic	1TPO1E	88	75	78	77
Trypsin Inhibitor (Proteinase Inhibitor)	4PTI1M	64	77	70	65
Coat Protein of Satellite Tobacco Necrosis Virus	2STU1E	77	60	56	54
Southern Bean Mosaic Virus Coat Protein	4SBV1E	59	48	52	48
Hydrolase (Aspartic Proteinase)	2APR	86	73	74	71
Calcium Binding Protein	3CLN	82	74	79	75
Hydrolase (Serine Proteinase and Zymogen)	1PSG1	69	75	65	63
Hydrolase (Serine Proteinase and Zymogen)	1PSG2	75	63	60	58
Serine Proteinase	2PRK	75	62	65	64
Complex (Serine Proteinase-Inhibitor)	2SEC1	92	82	83	83
Complex (Serine Proteinase-Inhibitor)	2SEC2	86	66	70	71
Transferase (Phosphotransferase)	3ADK	77	66	71	63
Proteinase Inhibitor (Chymotrypsin)	2CI2	66	80	74	70
Oxidoreductase (Oxygenase)	2CPP	77	69	74	70
Oxidoreductase (Flavoenzyme)	3GRS	80	66	72	67
Transferase (Phosphotransferase)	3PFK	71	77	70	63
Photosynthetic Reaction Center	1PRC1	75	46	75	74
Photosynthetic Reaction Center	1PRC2	71	57	71	62
Photosynthetic Reaction Center	1PRC3	73	54	71	64
Photosynthetic Reaction Center	1PRC4	65	64	66	66
Electron Transfer (Cuproprotein)	2PAZ	76	76	74	73
Contractile System Proteins	5TNC	90	59	86	83
DNA Binding Regulatory Protein	2WRP	71	+	71	70
Chromosomal Protein	1UBG	80	73	74	72
DNA Binding Regulatory Protein	1LRD	79	—	73	74
Glycosidase Inhibitor	1HOE	+	72	72	72
Periplasmic Binding Protein	2LBP	79	70	78	71
Steroid Binding	2UTG	79	+	79	71
Hydrolase (Acid Proteinase)	3HUP	—	67	78	66
Hydrolase (Endoribonuclease)	1RNT	88	64	65	66
Ligase (Synthetase) Chain 1	2TS1	81	67	78	73
Ligase (Synthetase) Chain 2	2TS2	80	99	78	79
Lyase (Carbon — Oxygen) Chain 2	1WSY	80	85	79	80
Lyase (Carbon — Oxygen) Chain 3	1WSY	71	77	70	63
Lyase (Carbon — Oxygen) Chain 4	1WSY	66	80	74	70
Lyase (Carbon — Oxygen) Chain 5	1WSY	80	66	72	67
Oxidoreductase	1PHH	69	75	65	63
Oxidoreductase (Aldehyde (D)-NAD(A))	1GD1	77	68	71	65
Oxidoreductase (NAD(A)—CHOH(D))	8ADH	67	60	66	62
Oxidoreductase (NAD(A) — CHOH(D))	4MDH	76	76	74	73
Oxidoreductase (Oxyfen(A)) Chain 1	1GOX	90	59	86	83
Oxidoreductase (Oxygen(A)) Chain 2	1GOX	71	57	71	62
Усредненный показатель		74	67	71	68

Примечание. Если прогнозируемый белок, по данным рентгеноструктурного анализа, не содержал остатков в конформационном состоянии *s*, то в случае корректного прогнозирования этот факт отмечается в таблице знаком плюс, иначе ставится знак минус.

(Ts^{-+}) — число остатков, не входящих в структуру s и совпадающих по предсказанию с данными рентгеноструктурного анализа.

Общую достоверность оценивали по формуле

$$Q = [Th^{+} (Te^{+}) + (Tc^{+})]/N,$$

где N — общее число остатков.

Результаты и обсуждение. Пусть задан участок белка из m остатков и надо определить, в какую структуру встраивается i-й остаток данного участка. Для этого из банка данных извлекаются все последовательности длиной m, содержащие на i-м месте заданный аминокислотный остаток. Далее в i-е место каждой последовательности вместо остатка заносится значение вторичной структуры этого остатка. Полученные последовательности, записанные одна за одной, образуют таблицу эмпирических данных, i-й столбец которой состоит из значений вторичной структуры определяемого остатка. Остальные столбцы составляют соответствующие остатки (всего в таблице m столбцов).

Если в качестве целевого состояния $P1$ задать состояние $(h)F_i$, то GUHA-метод выдаст все комбинации контекстных остатков, связанные с встраиванием i-го остатка в α-спираль. Аналогично получаются результаты для других значений вторичной структуры. В итоге формируется множество гипотез встраивания данного остатка во вторичную структуру.

Критерий выбора следующий: прогнозируемое значение вторичной структуры то, которое дает наименьшее значение критического уровня во всех полученных гипотезах.

Прогнозирование для белка произвольной длины осуществляется последовательным прогнозированием каждого остатка в контексте $(m - 1)$ соответствующих остатков.

Для практического осуществления метода необходимо определить параметры:

m — длину последовательности остатков;

i — расположение прогнозируемого остатка.

Опытным путем установлены оптимальные значения:

$$m = 5; \; i = 1.$$

Схема предсказания вторичной структуры была следующая: 107 белков составляли обучающее множество, т. е. множество, на котором GUHA-метод формировал гипотезы об окружении аминокислот. 108-й белок брали в качестве тестируемого и предсказывали его вторичную структуру. Такая циклическая процедура проделана для каждого из 108 белков используемой базы данных. Результаты предсказания для каждого белка представлены в таблице. Средняя точность предсказания по всем белкам для всех трех состояний составила 68 % корректного предсказания, что на настоящий момент является одним из лучших результатов по предсказанию вторичной структуры белка [3].

Резюме. У роботі запропоновано новий метод передбачення вторинної структури білка, що базується на відомому в галузі експертних систем GUHA-методі. Навчання та передбачення грунтуються на інформації про вторинну структуру 108 білків (біля 20 000 амінокислотних залишків) з рентгеноструктурним розділенням менше 0,2 нм. Середня точність передбачення з використаного в роботі банку даних складає для α-спіралі — 74 %, β-складки — 67 %, нерегулярної структури — 71 %; загальна — 68 %.

Summary. A new method for protein secondary structure prediction is described in the present article. This method based on GUHA-method has been known in the field of expert systems. Information for secondary structure of 108 [illegible]

X-ray resolution less than 0,2 nm was used for learning and prediction of protein secondary structure. Average accuracy of prediction for helix is 74 %, strand is 67 %, coil is 71 % and for three states simultaneously is 68 % of successful prediction.

СПИСОК ЛИТЕРАТУРЫ

1. *Sternberg M. J. E., Islam S. A.* Local protein sequence similarity does not implpy a structural relationship // Prot. Engen.— 1990.— **4**, N 2.— P. 125—131.
2. *Гаек П., Гавранек Т.* Автоматическое образование гипотез / Пер. с англ.— М. : Наука, 1984.
3. *Мальченко С. З., Чащин Н. А.* Предсказание вторичной структуры белков // Биополимеры и клетка.— 1992.— 8, № 5.— С. 21—31.

Ин-т молекуляр. биологии и генетики
АН Украины, Киев

Получено 26.04.93

УДК 547.963.3+577.323

А. И. Егоренков, В. В. Король

ПАКЕТ ПРИКЛАДНЫХ ПРОГРАММ ДЛЯ ГРАФИЧЕСКОГО ИЗУЧЕНИЯ ТОПОЛОГИИ ПОВЕРХНОСТИ ПОТЕНЦИАЛЬНОЙ ЭНЕРГИИ, СООТВЕТСТВУЮЩЕЙ КОНФОРМАЦИОННОЙ ДИНАМИКЕ МОЛЕКУЛЫ ДНК

В работе описан пакет прикладных программ, предназначенный для визуального (графического) изучения топологии поверхности потенциальной энергии, соответствующей внутренним движениям двойной спирали ДНК. Пакет применен для анализа данных аналитического моделирования торсионной динамики ДНК, описывающей процесс локального раскрытия пар азотистых оснований. Обсуждаются возможности программ для решения задач численного моделирования динамики ДНК методами атом-атомных потенциалов.

Программы имеют удобный пользовательский интерфейс и современные алгоритмы работы с трехмерной графикой (построение карт поверхностей и изополос, динамическое картирование). Программы реализованы для персонального компьютера IBM PC AT/XT в стандартной конфигурации с видеоадаптером CGA/VGA, язык программирования С.

Введение. Молекула ДНК как сложная биологическая система обладает большим разнообразием внутренних движений и их сложной иерархией. При изучении механизмов биологического функционирования ДНК необходимо учитывать динамические возможности молекулы. Классификация разных движений ДНК позволяет выделить следующие их типы, различающиеся по характерным временам и энергиям [1, 2]: малые колебания атомов около положения равновесия; ограниченные движения сахаров, фосфатов, азотистых оснований около положения равновесия; малоамплитудные торсионные и изгибные движения двойной спирали ДНК; движения, связанные с A⟶B⟶Z-переходами; движения большой амплитуды, зависящие от изменения суперспирального состояния ДНК; локальное расплетание двойной спирали. Один из возможных типов внутримолекулярных движений, а именно: локальное раскрытие пар оснований попадает в две сильно отличающиеся по характерным временам группы движений: 10^{-7}—10^{-5} с для раскрытия отдельной пары оснований и 10^{-4}—10^{-2} с для раскрытия пар оснований, связанного с расплетанием двойной спирали ДНК. Данный тип конформационной подвижности важен для понимания предполагаемых механизмов белково-нуклеинового взаимодействия и возможен при взаимодействии ДНК с лигандами разной природы [3], при флуктуационном локальном раскрытии пар азотистых оснований [4], при резонан-