ISSN 0233-7657. Биополимеры и клетка. 1998. Т. 14. № 2

Применение метода LOGIS для предсказания вторичной структуры белка

А. В. Братусь, Н. А. Чащин

Институт молекулярной биологии и генетики НАН Украины
252143, Киев, ул. Академика Заболотного, 150

В работе для предсказания вторичной структуры белка используется новый метод распознавания образов, известный под названием LOGIS. Обучение и предсказание базируется на данных о вторичной структуре 108 белков (около 20000 аминокислотных остатков) с рентгеноструктурным разрешением менее 0,2 нм. Средняя точность предсказания на имеющихся данных составила 71 %.

Введение. Увеличение точности предсказания вторичной структуры белка дает возможность получить его довольно близкую к реальной пространственную модель [1].

Существующие в настоящее время методы не всегда позволяют достичь необходимой точности, поэтому не прекращается поиск новых подходов в этом направлении, в частности, привлекаются новейшие способы распознавания образов. В данной работе для предсказания вторичной структуры белка используется метод распознавания образов LOGIS, предложенный в [2].

Материалы и методы. Обучающую выборку образует банк данных, содержащий информацию о вторичной структуре 108 белков (около 20000 аминокислотных остатков, АА) с рентгеноструктурным разрешением менее 0,2 нм. Данные получены из Брукхейвенского банка данных белков с известной пространственной структурой. Вторичная структура классифицирована по трем конформациям: спираль (h), складка (e), нерегулярная (c). Каждому аминокислотному остатку приписывается одно из трех состояний вторичной структуры.

Основу метода LOGIS составляет точный критерий проверки на независимость двух признаков (критерий Фишера). Суть критерия: пусть в некоторой выборке каждый объект характеризуется двумя признаками (А и В) и пусть

s — число объектов, имеющих и признак А и признак В;

r — число объектов, имеющих признак А;

k — число объектов, имеющих признак В;

m — общее число объектов.

Тогда обладание признаком А взаимосвязано с обладанием признака В, если

$$\mathrm{Fish}(A, B) = \sum_{i=s}^{\min(r,\,k)} q(i, r, k, m) < \alpha,$$

где

$$q(i, r, k, m) = r!\,k!\,(m-r)!\,(m-k)!\ /$$

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

$$0 < a < 0{,}5$$

— уровень принятия гипотезы о зависимости.

Пусть дан участок белка длиной l с неизвестной вторичной структурой (AA(1), ..., AA(l)) и пусть AA(j) есть S. Надо определить, встраивается ли S в спираль в данном контексте или нет? Выберем из банка данных белков с известной вторичной структурой все последовательности длиной l, содержащие на j-м месте аминокислотный остаток S. На этой выборке по критерию Фишера можно проверить, например, такую гипотезу: «Имеется ли связь между тем, что S встроена в спираль и AA(1) есть Q?» Аналогично можно проверить гипотезы о взаимосвязи встраивания S в спираль и значениями AA(1), ..., AA(l), взятых по

одному. Однако интерес представляют не только одиночные АА, но и их комбинации по два, по три и т. д. В этом случае для образования комбинационного признака используются формулировочные возможности логики.

Например, комбинационный признак двух аминокислотных остатков (признак длины два) есть высказывание: «АА($j-2$) есть Q и АА($j+3$) есть P». В итоге для получения всех знаний о встраивании S в спираль генерируются и проверяются все гипотезы с признаком длины один, с признаком длины два и т. д. Все принятые гипотезы имеют свое точное значение критерия. Принятие гипотез ограничено сверху значением, выбранным из интервала (0; 0,5). Множество гипотез о встраивании S в спираль обозначим $S(h)$. Аналогично получается множество гипотез о невстраивании S в спираль (множество $S(\hat{h})$). Далее на основе $S(h)$ и $S(\hat{h})$ надо принять решение о встраивании S в спираль. Необходимо определить некоторую интегральную оценку множеств $S(h)$ и $S(\hat{h})$ и по сравнению оценок сделать выбор.

В математической статистике рассматривается проблема одновременного выполнения множества гипотез, каждая из которых имеет свое значение критерия [3]. Вводится коэффициент ошибок множества гипотез

e. e. r. = число ошибочных гипотез/общее число гипотез.

Доказывается, что математическое ожидание *e. e. r.* не превосходит максимального значения всех критериев этого множества гипотез. Тогда решающее правило для $S(h)$ и $S(\hat{h})$ гласит: «S встроена в спираль, если оценка $S(h)$ меньше оценки $S(\hat{h})$. Иначе S не встроена в спираль». Состав $S(h)$ и $S(\hat{h})$ зависит от выбора граничного значения доверительного уровня из интервала (0; 0,5) и, как следствие, достоверность прогноза зависит от данного выбора.

Метод LOGIS позволяет определить оптимальное значение доверительного уровня, обеспечивающего максимально возможную достоверность прогноза: пусть, как и выше, дан участок белка длиной l с неизвестной вторичной структурой (АА(1), ..., АА(l)) и пусть АА(j) есть S. Надо определить, встраивается ли в данном контексте S в спираль или нет? Выберем из банка данных белков с известной вторичной структурой все последовательности длиной l, содержащие на j-м месте аминокислотный остаток S. Выберем максимально возможное значение доверительного уровня, а именно: 0,5. Пусть для данного граничного значения сформированы множества $S(h)$ и $S(\hat{h})$, состоящие из значений критериев соответствующих гипотез. Расположим элементы множеств $S(h)$ и $S(\hat{h})$ в одном массиве ALFA в порядке возрастания значений элементов. Каждый ALFA(i), взятый в качестве граничного значения, определяет подмножества $S(h)_i$ и $S(\hat{h})_i$ и, следовательно, исход сравнения оценок этих подмножеств. Если оценка $S(h)_i$ не превышает оценки $S(\hat{h})_i$, то S встроена в спираль. Иначе S не встроена в спираль. Таким образом, каждый элемент ALFA определяет либо исход «S встроена в спираль», либо исход «S не встроена в спираль».

Объединив все соседние элементы ALFA с одинаковыми исходами в один интервал, получим систему интервалов известных исходов. Если выбрать любое граничное значение из (0; 0,5), то на основе системы интервалов можно узнать, какой исход соответствует данному выбору. Предположим известно, встраивается S в спираль или нет. Тогда, пометив интервалы, получим систему интервалов истинных и ложных исходов. Такого рода система интервалов есть знание поведения оценки при прогнозе встраивания S в спираль в данном окружении аминокислотных остатков (элементарное знание). Совокупность элементарных знаний, полученных для разных последовательностей длиной l, образует более общее знание о поведении оценки при прогнозе встраивания в спираль j-го аминокислотного остатка.

Каким образом, располагая банком данных белков с известной вторичной структурой, получить такого рода знание по конкретному аминокислотному остатку, например S? Для этого выбирается первая последовательность длиной l, содержащая на j-м месте S, и удаляется из банка данных. Для этой последовательности строится элементарное знание. Затем выбирается другая последовательность, строится элементарное знание и т. д. до получения последнего элементарного знания.

В каждом элементарном знании представлены значения доверительного уровня, дающие истинный либо ложный исход прогноза. Перебрав все значения границ интервалов элементарных знаний, получим значение, дающее истинный прогноз для максимального числа элементарных знаний, и значение, дающее ложный прогноз для максимального числа элементарных знаний. Из двух значений выбирается значение, распознающее максимальное число элементарных знаний. Поскольку осуществляется дихотомический прогноз (либо встроен в спираль, либо не встроен в спираль), интерес представляет и значение, дающее ложный прогноз. Тогда решающее правило гласит: «Принять решение, обратное полученному решению».

Для повышения достоверности прогноза в дан-

Таблица 1

Оптимальные значения доверительного уровня для классификации «c не c»

AA	$j-1$	j	$j+1$
A	0,386	0,392	0,343
C	0,529	0,553	0,506
D	0,325	0,617	0,547
E	0,575	0,591	0,415
F	0,658	0,672	0,639
G	0,617	0,289	0,251
H	0,541	0,520	0,524
I	0,455	0,493	0,657
K	0,278	0,542	0,465
L	0,305	0,236	0,297
M	0,677	0,527	0,660
N	0,554	0,400	0,605
P	0,626	0,522	0,429
Q	0,551	0,356	0,428
R	0,313	0,422	0,590
S	0,318	0,338	0,296
T	0,265	0,519	0,547
U	0,451	0,511	0,416
W	0,342	0,511	0,323
Y	0,309	0,590	0,541

Таблица 2

Оптимальные значени доверительного уровня для классификации «h не h »

AA	$j-1$	j	$j+1$
A	0,550	0,384	0,579
C	0,747	0,739	0,751
D	0,674	0,714	0,664
E	0,556	0,541	0,315
F	0,590	0,602	0,626
G	0,809	0,737	0,676
H	0,541	0,680	0,688
I	0,654	0,702	0,657
K	0,372	0,541	0,414
L	0,453	0,569	0,468
M	0,677	0,726	0,660
N	0,736	0,763	0,729
P	0,788	0,823	0,778
Q	0,551	0,610	0,574
R	0,560	0,578	0,628
S	0,769	0,593	0,400
T	0,580	0,513	0,547
U	0,672	0,474	0,713
W	0,657	0,696	0,584
Y	0,723	0,720	0,748

ной работе предлагается процедура страхования. Суть ее состоит в следующем. Пусть дана последовательность длиной l с неизвестной вторичной структурой, содержащая на j-м месте аминокислотный остаток S. Выше рассматривалась задача: «Встраивается ли S в спираль в данном аминокислотном окружении или нет?» Видоизменим задачу: «Если S стоит на j-м месте, то встраивается ли аминокислотный остаток на $j+1$ месте в спираль или нет»? В этом случае мы исследуем влияние S на соседнее место справа. Эта задача решается точно так же, как и первая, с поправкой на прогнозируемое место.

Аналогично ставится и решается задача влияния S на соседнее место слева, то есть $j-1$. Таким образом, для каждого аминокислотного остатка определяются три значения оптимального доверительного уровня. Первое значение дает максимальную достоверность прогноза встраивания S в спираль. Второе значение дает максимальную достоверность прогноза встраивания в спираль правого соседа S. Третье значение дает максимальную достоверность прогноза встраивания в спираль левого соседа S. Тогда для последовательности длиной l, содержащей, например, с $j-1$ места последовательность аминокислотных остатков ASD, при прогнозировании встраивания S осуществляются три прогноза:

— прогноз встраивания S;

— прогноз встраивания правого соседа A;

— прогноз встраивания левого соседа D.

Последние два прогноза являются страховочными и введены для повышения точности общего прогноза.

Все приведенные выше рассуждения касались прогноза встраивания в спираль (класс h). Они

Таблица 3
Сопоставление вторичной структуры белков, спрогнозированной по методу LOGIS, с фактическими результатами из Брукхейвенского банка данных

Белок	Брукхейвенский банк данных		Прогноз	
	α	β	α	β
Calcium Binding Protein Bovine Intestine	8—17, 26—32, 40—50, 60—64, 99—107	57—59, 96—98	9—17, 25—33, 39—49, 60—65, 80—88, 100—107	57—58, 94—96
Carbonic Ahnydrase Form B Human Erythrocytes	3—14, 25—35, 46—53, 63—74	—	1—14, 25—36, 44—53, 62—69, 71—74	—
Citrate Synthase Pig Heart	17—27, 82—95	8—15, 32—36, 50—55, 57—66, 76—81	17—28, 81—96	8—14, 32—35, 49—54, 56—60, 62—65, 75—80

Таблица 4
Точность предсказания вторичной структуры белков по методу LOGIS

Белок	Имя файла в базе данных	*N*	% *N*
Acid Proteinase Endothiapepsin	4APE1E	330	81
Acid Proteinase, Penicillopepsin (Hydrolase: Proteinase)	2APP1E	323	65
Actinidin (Hydrolase: Sulfhydryl Proteinse)	2ACT1M	218	51
Lectin (Agglutinin) Whet Germ	3WGA1M	170	74
Alpha Lytic Protease (Hydrolase: Serine Proteinase)	2ALP1E	374	64
Aspartate Transcarbamylase (*E. coli*) Chain 1	4ATC1M	198	69
Aspartate Transcarbamylase (*E. coli*) Chain 2	4ATC2M	310	81
Azurin Electron Transport Protein	1AZA1E	153	63
Calcium Binding Parvalbumin b (Calcium Binding Proteins)	1CPV1H	129	64
Calcium Binding Protein Bovine Intestine Vitamin D Dependent	3ICB1H	108	88
Carbonic Anhydrase Form B Human Erythrocytes	2CAB1E	75	91
Carboxypeptidase A (C-Terminal Amino Acid Hydrolase)	5CPA1M	256	63
Catalase Beef Liver	8CAT1M	307	68
Alpha Chymotripsin A (*Bos Taurus*) Chain 1	5CHA1E	498	63
Alpha Chymotripsin A (*Bos Taurus*) Chain 2	5CHA2M	131	61
Citrate Synthase Pig Heart	2CTS1H	97	88
Crambin (Plant Seed Protein)	1CRN1M	437	57
Gamma Crystallin Calf Eye Lens	1GCR1E	46	61
Cytochrome C (Oxidized) (Electron Transport)	3CYT1H	174	70
Cytochrome C Rice Embryos	1CCR1M	103	81
Cytochrome C Prime	2CCY1H	111	83
Cytochrome C Peroxidase (Bayker's Yeast)	2CYP1H	127	60
Ferricytochrome C2 (Electron Transport)	3C2C1H	293	72
Cytochrome C3	2CDV1M	112	69
Cytochrome C551 (Oxidized) (Electron Transport)	351C1H	107	91
Dihydrofolate Reductase (Oxidoreductase: NADPH/DONR)	3DFR1M	82	59
Elastase Porcine Pancrease	2EST1E	162	69

Продолжение табл. 4

Белок	Имя файла в базе данных	N
Erabutoxin Sea Snake Venom	2EBX1E	240
Hemoglobin (Erythrocruorin, Deoxy) (Oxigen Transport)	1ECD1H	62
Ferredoxin (Electron Transport)	1FDX1M	136
Ferredoxin (Electron Transport)	3FXC1M	54
Flavodoxin (Oxidized) (Electron Transport)	3FXN1M	98
Ferredoxin Azotobacter	2FD11M	138
Glutathione Reductase Bovine Erythrocytes	1GP11M	106
Hemerythrin (Met) Sipunculid Worm	1HMQ1H	184
Hemoglobin (Human, Deoxy) Chain 1	2HHB1H	113
Hemoglobin (Human, Deoxy) Chain 2	2HHB2H	141
Hemoglobin V (Cyano, Met) Sea Lamprey	2LHB1H	146
Oxidized High Potential Iron Protein (Hipip)	1HIP1M	149
Immunoglobulin Fab IgG (Muose) Chain 1	1MPC1E	85
Immunoglobulin Fab IgG (Muose) Chain 2	1MPC2E	220
Immunoglobulin Fab IgG (Human Myeloma) Chain 2	1FB42E	222
Bence-Jones Immunoglobulin Variable Portion	1REI1E	229
Bence-Jones Protein Lambda Variable Domain	2RHE1E	107
Kallikrain A (Porcine Pancrease) Chain 1	2PKA1E	114
Kallikrain A (Porcine Pancrease) Chain 2	2PKA2M	80
Lactate Dehydrogenase, Apo Enzyme M4	4LDH1M	152
Leghemoglobin (Acetate, Met) (Oxigen Transport)	1LH11H	329
Lysozyme (Bacteriophage T4)	2LZM1M	153
Lysozyme (Human)	1LZ11M	164
Myoglobin (Oxigen Storage) (Ferric Iron — Metmyoglobin)	1MBN1H	130
Melittin (Hemolytic Polypeptide)	1MLT1H	153
Scorpion Neurotoxin	1SN31M	26
Ovomucoid Third Domain (Proteinase Inhibitor, Kazal)	10V01M	65
Papain Sulfhydryl Proteinase (Papaya Fruit Latex)	1PPD1M	56
Phospholipase A2 (Phosphatide Acyl-Hydrolase)	1BP21M	212
Plastocyanin (Electron Transport, Cooper Binding)	1PCY1E	123
Prealbumin (Thyroxin, Retinol Transport)	2PAB1E	99
Proteinase A (SGPA) (Hydrolase: Serine Proteinase)	2SGA1E	114
Serine Proteinase (Rat Mast Cell Protease)	3RP21E	181
Ribonuclease A (Bovine Pancrease)	1RN31M	224
Rubredoxin Iron-Sulfur Protein (*Clostridium*)	4RXN1M	124
Staphylococcal Nuclease	2SNS1M	54
Subtilysin BPN' (Hydrolase: Serine Proteinase)	1SBT1M	141
Cu, Zn Superoxide Dismutase (Oxidoreductase: Superoxide)	2SOD1E	275
Thermolysin (Hydrolase: Neutral Metallo-Proteinase)	3TLN1M	151
Beta Trypsin (Bovine) Orthorombic	1TPO1E	316
Trypsin Inhibitor (Proteinase Inhibitor)	4PTI1M	58
Coat Protein of Satellite Tobacco Necrosis Virus	2STV1E	184
Southern Bean Mosaic Virus Coat Protein	4SBV1E	222

Окончание табл. 4

Белок	Имя файла в базе данных	N	% N
Hydrolase (Aspartic Proteinase)	2APR	325	75
Calcium Binding Protein	3CLN	143	79
Hydrolase (Serine Proteinase and Zymogen)	1PSG1	200	68
Hydrolase (Serine Proteinase and Zymogen)	1PSG2	165	69
Serine Proteinase	2PRK	279	71
Complex (Serine Proteinase — Inhibitor)	2SEC1	274	88
Complex (Serine Proteinase — Inhibitor)	2SEC2	64	75
Transferase (Phosphotransferase)	3ADK	194	69
Proteinase Inhibitor (Chymotrypsin)	2CI2	65	75
Oxidoreductase (Oxigenase)	2CPP	405	74
Oxidoreductase (Flavoenzyme)	3GRS	461	70
Transferase (Phosphotransferase)	3PFK	319	79
Photosynthetic Reaction Center	1PRC1	332	69
Photosynthetic Reaction Center	1PRC2	273	68
Photosynthetic Reaction Center	1PRC3	323	66
Photosynthetic Reaction Center	1PRC4	258	78
Electron Transfer (Cuproprotein)	2PAZ	123	69
Contractile System Proteins	5TNC	161	76
DNA Binding Regulatory Protein	2WRP	104	78
Chromosomal Protein	1UBQ	76	77
DNA Binding Regulatory Protein	1LRD	87	76
Glycosidase Inhibitor	1HOE	74	78
Periplasmic Binding Protein	2LBP	346	71
Steroid Binding	2UTG	70	74
Hydrolase (Acid Proteinase)	3HVP	66	72
Hydrolase (Endoribonuclease)	1RNT	104	61
Ligase (Synthetase) Chain 1	2TS1	211	64
Ligase (Synthetase) Chain 2	2TS2	106	69
Lyase (Carbon — Oxygen) Chain 2	1WSY1	55	91
Lyase (Carbon — Oxygen) Chain 3	1WSY2	119	73
Lyase (Carbon — Oxygen) Chain 4	1WSY3	74	75
Lyase (Carbon — Oxygen) Chain 5	1WSY4	385	74
Oxidoreductase	1PHH	394	77
Oxidoreductase (Aldehyde (D) — NAD (A))	1GD1	334	79
Oxidoreductase (NAD (A) — CHOH (D))	8ADH	333	73
Oxidoreductase (NAD (A) — CHOH (D))	4MDH	188	88
Oxidoreductase (Oxygen (A)) Chain 1	1GOX1	162	68
Oxidoreductase (Oxygen (A)) Chain 2	1GOX2	156	71
ВСЕГО	—	19559	71

Примечание. N — число аминокислотных остатков в белке; % N — точность предсказания вторичной структуры белка.

остаются верными и для класса *e* (складчатый лист) и для класса *c* (нерегулярная структура).

Результаты и обсуждение. Экспериментальным путем была выбрана следующая последовательность прогнозирования вторичной структуры белка:

1) для каждого аминокислотного остатка осуществить прогноз встраивания в класс *c*;

2) для каждого аминокислотного остатка, не отнесенного к классу *c*, осуществить прогноз встраивания в класс *h*;

3) для каждого аминокислотного остатка, не отнесенного ни к классу *c*, ни к классу *h*, приписать класс *e*.

Эмпирическим путем были определены параметры метода:

1) длина последовательности $l = 9$;

2) расположение прогнозируемого аминокислотного остатка $j = 5$;

В результате применения процедуры настраивания были определены оптимальные значения доверительного уровня для прогнозирования встраивания каждого из 20 аминокислотных остатков, приведенные в табл. 1 и табл. 2.

В итоге было осуществлено прогнозирование вторичной структуры каждого из 108 белков, составляющих обучающую выборку. При этом прогнозируемый белок удаляли из обучающей выборки.

Сравнительные результаты прогнозирования вторичной структуры белков приведены в табл. 3; общие результаты прогнозирования — в табл. 4. Достоверность прогнозирования вторичной структуры белков составила 71 %. Расчет производили по формуле $Q = (T/N) \cdot 100$ %, где T — число верно спрогнозированных остатков; N — общее число остатков белка. По сравнению со средней величиной предсказания вторичной структуры белка (72,1 %) одним из наиболее удачных методов — PHD [5] результаты работы метода LOGIS примерно одинаковы.

Полученные результаты были использованы при написании программной системы предсказания вторичной структуры белка.

О. В. Братусь, М. О. Чащин

Застосування методу LOGIS для передбачення вторинної структури білка

Резюме

У роботі для прогнозування вторинної структури білків запропоновано новий метод розпізнавання образів, відомий як метод LOGIS. Навчання та передбачення грунтуються на даних про вторинну структуру 108 білків (біля 20000 амінокислотних залишків) з рентгеноструктурним розділенням менше 0,2 нм. Середня точність передбачення складає 71 %.

A. V. Bratus, N. A. Chashchin

A method for protein secondary structure prediction

Summary

A new method for protein secondary structure prediction is described in the present article. This method based on LOGIS-method. Information for secondary structure of 108 proteins (20000 AAs) with X-ray resolution less than 0.2 nm was used for learning and prediction of protein secondary structure. Average accuracy of successful prediction is 71 %.

СПИСОК ЛИТЕРАТУРЫ

1. *Sternberg M. J. E., Islam S. A.* Local protein sequence similarity does not imply a structural relationship // Prot. Eng.—1990.—4, N 2.—P. 125—131.
2. *Сергиенко И. В., Гупал А. М., Братусь А. В.* LOGIS-система, реализующая статистический абдуктивный вывод на эмпирических данных // Кибернетика.—1995.—№ 3.—С. 160—173.
3. *Кендал М., Стьюарт А.* Многомерный статистический анализ и временные ряды.—М.: Наука, 1976.—С. 65—68.
4. *Братусь А. В., Мальченко С. З., Чащин Н. А.* Предсказание вторичной структуры белка modGUHA-методом // Биополимеры и клетка.—1993.—№ 5.—С. 61—66.
5. *Rost,* Burkhard, Sander, Chris Combining evolutionary and neural networks to predict protein secondary structure // Proteins.—1994.—19.—P. 55—72.

Поступила в редакцию 05.05.97