ОБ ОДНОМ ПОДХОДЕ К ОПРЕДЕЛЕНИЮ СТАРТ-КОДОНА В ДНК-ПОСЛЕДОВАТЕЛЬНОСТИ ЭУКАРИОТ

В. А. Галинский, А. Н. Гайдук, В. А. Волошко

НИИ прикладных проблем математики и информатики БГУ
Минск, Беларусь
E-mail: GalinskijVA@bsu.by, GaidukAN@bsu.by

Предложен подход к поиску старт-кодона в ДНК-последовательности эукариот, основанный на локальных и глобальных свойствах ДНК-последовательности эукариот.

Ключевые слова: ДНК-последовательность, старт-кодон.

Введение

Одной из основных задач анализа ДНК-последовательности эукариот является задача поиска генов. Задача поиска генов включает задачу поиска старт-кодонов (ATG) и стоп-кодонов (TGA, TAG и TAA). В настоящее время разработаны методы для поиска старт-кодонов на основе локальных свойств ДНК-последовательности [1, 2]. В данной работе для поиска старт-кодонов в ДНК-последовательности эукариот предлагается подход, использующий локальные свойства ДНК-последовательности (свойства старт-кодона, сайтов сплайсинга и промоторов) и глобальные свойства ДНК-последовательности (статистические и периодические свойства кодирующих и некодирующих областей). Подход идейно близок к предложенному в [3] комплексному статистическому подходу для предсказания экзон-интронной структуры генов и отличается используемыми вероятностными моделями.

Вероятностные модели для поиска старт-кодонов

В предлагаемом подходе для учета локальных свойств старт-кодона, сайтов сплайсинга (GT, AG) и промоторов используются модели позиционных весовых матриц [1] и неоднородные цепи Маркова [2]. Для учета глобальных свойств кодирующих и некодирующих областей используются вероятностные распределения n-грамм (n=3, 4, 5, 6) [4], обобщенные скрытые Марковские модели [5], модели, основанные на триплетной периодичности [6]. Приведем краткое описание вероятностных моделей, предлагаемых в настоящей работе для оценивания старт-кодона.

Неоднородные цепи Маркова первого порядка применяются для моделирования вероятностных свойств окрестностей старт-кодона и сайтов сплайсинга [2]. При этом в случае старт-кодона учитывались 7 нуклеотидов в направлении 5' и 10 нуклеотидов в направлении 3' от ATG (схема «7 слева, 10 справа»). Аналогичным образом для GT использовалась схема «3 слева, 4 справа», и для AG — схема «13 слева, 1 справа». При обучении для каждого из трех сигналов (ATG, GT, AG) оценивались параметры двух неоднородных цепей Маркова. Параметры первой неоднородной цепи Маркова оценивались по окрестностям истинных сигналов ДНК-последовательностей, а параметры второй неоднородной цепи Маркова оценивались по окрестностям ATG, GT, AG, которые не являются сигналами ДНК-последовательностей. Решающее правило

для каждого из трех сигналов (ATG, GT, AG) было основано на отношении правдоподобия указанных двух неоднородных цепей Маркова.

При определении старт-кодона используется свойство триплетной периодичности кодирующих областей [6]. Для каждого нуклеотида вычисляются частоты встречаемости в каждой из трех позиций триплета. Для построения решающего правила используется предложенная в [6] статистика отношения максимального значения вычисленных частот к минимальному. Решающее правило основано на сравнении значения данной статистики с некоторым порогом. Порог определяется на этапе обучения.

Модель обобщенной скрытой цепи Маркова [5] позволяет учесть вероятностные свойства кодирующих и некодирующих областей. Рассматривается задача нахождения сайта сплайсинга GT в фрагменте ДНК-последовательности, содержащем первый экзон и первый интрон гена. Для ее решения на основе модели гена [5] строится параметрическое семейство вероятностных распределений, относительно которого методом максимального правдоподобия находится оптимальное разбиение фрагмента ДНК-последовательности на первый экзон и первый интрон. Для построения функции правдоподобия фрагмент ДНК-последовательности разбивается на пять частей, которые описываются независимыми вероятностными моделями. Для описания первой части фрагмента ДНК-последовательности, содержащей старт-кодон, используется индикаторная функция. Для описания вероятностных свойств второй и четвертой частей используются цепи Маркова четвертого порядка. В случае недостаточного числа данных на этапе обучения параметры соответствующих цепей Маркова четвертого порядка определяются методом интерполяции параметров цепей Маркова третьего порядка. Вероятностное распределение третьей части оценивается на этапе обучения. Вероятностное распределение пятой части не используется для нахождения сайта сплайсинга GT.

Для нахождения старт-кодона также применяется различие частот кодонов, кодирующих одну и ту же аминокислоту, для первого экзона, 5' UTR и первого интрона. Оценки частот кодонов, кодирующих одну и ту же аминокислоту, получены по выборке h178 [7] и приведены в табл. 1.

Таблица 1

Частоты кодонов, кодирующих аминокислоты

Аминокислота	Кодон	5' UTR	Первый экзон	Первый интрон
Phe/F	TTT	47,4490	35,9621	63,6957
Phe/F	TTC	52,5510	64,0379	36,3043
Leu/L	TTA	8,5581	3,3333	12,8704
Leu/L	TTG	11,7209	10,4167	14,4764
Leu/L	CTA	8,6512	4,8958	9,5906
Leu/L	CTT	17,7674	8,5938	16,9645
Leu/L	CTC	25,6744	21,6667	20,4705
Leu/L	CTG	27,6279	51,0938	25,6277
Ser/S	TCT	19,2449	17,3785	22,0503
Ser/S	TCC	25,3223	26,0677	22,8631
Ser/S	TCA	16,7587	11,0457	16,9376
Ser/S	TCG	6,3536	7,4374	4,2475
Ser/S	AGT	10,6814	9,1311	15,6529
Ser/S	AGC	21,6390	28,9396	18,2486
Tyr/Y	TAT	60,7407	30,8046	59,2771
Tyr/Y	TAC	39,2593	69,1954	40,7229
Cys/C	TGT	41,2281	38,1766	52,6954

Окончание табл.1

Аминокислота	Кодон	5' UTR	Первый экзон	Первый интрон
Cys/C	TGC	58,7719	61,8234	47,3046
Trp/W	TGG	100,0000	100,0000	100,0000
Pro/P	CCT	25,7220	23,1740	31,8291
Pro/P	CCC	33,9350	39.9459	31.5584
Pro/P	CCA	23,0144	16.8620	27.7677
Pro/P	CCG	17,3285	20.0180	8.8448
Thr/T	ACT	27,7677	23.0516	30.6070
Thr/T	ACC	35,0272	43.3589	27.5206
Thr/T	ACA	25,2269	20.3074	33.8477
Thr/T	ACG	11,9782	13.2821	8.0247
His/H	CAT	35,6667	33.6898	48.8907
His/H	CAC	64,3333	66.3102	51.1093
Gln/Q	CAA	30,7860	18.4615	31.5304
Gln/Q	CAG	69,2140	81,5385	68,4696
Arg/R	CGT	7,9249	9,1470	5,3903
Arg/R	CGC	18,8738	26,0021	10,2905
Arg/R	CGA	6,0480	6,6804	4,5852
Arg/R	CGG	18,8738	19,5272	11,5156
Arg/R	AGA	19,8123	15,9301	29,5765
Arg/R	AGG	28,4672	22,7133	38,6419
Ile/l	ATT	38,4868	22,4756	41,2678
Ile/l	ATC	33,5526	64,3322	28,0078
Ile/l	ATA	27,9605	13,1922	30,7245
Asn/N	AAT	48,8889	36,1953	57,8947
Asn/N	AAC	51,1111	63,8047	42,1053
Lys/K	AAA	48,5849	30,2648	59,5819
Lys/K	AAG	51,4151	69,7352	40,4181
Val/V	GTT	23,1884	11,1001	21,6290
Val/V	GTC	26,0870	28,5431	22,6697
Val/V	GTA	13,4058	6,8385	19,3665
Val/V	GTG	37,3188	53,5183	36,3348
Ala/A	GCT	22,2101	22,4273	29,5051
Ala/A	GCC	37,6368	47,9773	33,5852
Ala/A	GCA	23,3042	15,4720	25,1606
Ala/A	GCG	16,8490	14,1235	11,7491
Gly/G	GGT	16,0156	12,9057	18,0559
Gly/G	GGC	26,3672	43,9245	25,2783
Gly/G	GGA	23,6328	18,6415	23,0519
Gly/G	GGG	33,9844	24,5283	33,6139
Asp/D	GAT	37,8182	32,0809	49,0948
Asp/D	GAC	62,1818	67,9191	50,9052
Glu/E	GAA	39,9577	28,6034	41,7266
Glu/E	GAG	60,0423	71,3966	58,2734

Подход реализован в разрабатываемом пакете прикладных программ для поиска генов в ДНК-последовательности эукариот. Параметры вероятностных моделей оцениваются по обучающей выборке h178 [7]. Пакет допускает использование других обучающих выборок.

Библиографические ссылки

1. *Burge C.* Identification of genes in human genomic DNA. Ph.D. thesis, Stanford, 1997.
2. *Yin M. M., Wang J. T. L.* Effective hidden Markov models for detecting splicing junction sites in DNA sequences // Inf. Sci; 2001. V. 01. P. 139–163.

3. *Gelfand M. S.* 1990. Computer prediction of the exon-in- tron structure of mammalian pre-mRNAs. Nucleic Acids Res. 18, 5865–5869. Gel, A. J. Chemical modification of electrodes // J. Chem. Educ. 1983. V. 60. № 4. P. 302–304.
4. *Haubold B.* Introduction to computational biology an evolutionary approach. T. Wiehe Birkhauser Verlag, Basel – Boston – Berlin, 2006. P. 117–140.
5. *Stanke M.* Gene Prediction with a Hidden-Markov Model. Göttingen, 2003. P. 100.
6. *Fickett J.* Recognition of protein coding regions in DNA sequences // Nucleic Acids Res. 1982 V. 10. № 17. P. 5303–5318.
7. *Burset M., Guigo R.* Evaluation of gene structure prediction programs // Genomics 1996, V. 34. P. 353–367.