АНАЛИЗ РАСПРЕДЕЛЕНИЯ АДЕНОЗИНФОСФАТ-СВЯЗЫВАЮЩИХ САЙТОВ БЕЛКОВ НА ЭКЗОННОЙ СТРУКТУРЕ ГЕНА

И.В. Медведева, П.С. Деменков, В.А. Иванисенко

Учреждение Российской академии наук Институт цитологии и генетики Сибирского отделения РАН, Новосибирск, Россия, e-mail: brukaro@bionet.nsc.ru; salix@bionet.nsc.ru

Проблема сопоставления пространственной структуры белка и кодирующей его последовательности гена возникла с открытием первых методов кристаллографии. Информация о пространственных структурах белков, в частности, позволила с высокой точностью определять их доменную структуру. Многие работы посвящены исследованию взаимосвязи границ экзонов и границ доменов белков, кодируемых соответствующим геном. В данной работе было поставлено целью выявление взаимосвязи между функциональными сайтами белков и экзонной структурой кодирующих их генов на примере сайтов связывания аденозинфосфатов. Показано, что длина экзонов, кодирующих аденозинфосфат-связывающие сайты, предположительно меньше длины экзонов, не участвующих в кодировании этих сайтов. Кроме того, аминокислоты рассматриваемых сайтов связывания удалены друг от друга на первичной структуре белка, несмотря на то что большая часть этих функциональных сайтов кодируется соседними по последовательности соответствующего гена экзонами. Также оказалось, что аденозинфосфат-связывающие сайты формируются не только внутри одного домена белка, но и из других участков последовательности, включая другие домены. Распределение аминокислот функционального сайта относительно положения кодирующего экзона в последовательности показало, что сайты связывания АТФ, АДФ, АМФ в большинстве случаев кодируются первыми экзонами последовательности.

Ключевые слова: функциональные сайты белков, экзонная структура гена, аденозинфосфат-связывающие сайты.

Введение

В 1960-х гг. появились экспериментальные методы определения пространственной структуры белка с помощью кристаллографического анализа и ядерного магнитного резонанса (Richards, 1963; Andreeva, 1964). Экспериментальные методы позволили идентифицировать структурные домены белков, т. е. минимальный по длине фрагмент последовательности, обладающий способностью к самосборке в глобулярную пространственную структуру, идентичную структуре этого фрагмента в составе полноразмерного белка. В настоящее время существует множество свидетельств того, что структурные домены также могут сохранять функциональные свойства, которые им были присущи в составе полноразмерных белков (Kaessmann *et al*., 2002). С появлением данных о пространственной структуре белков возникли вопросы о том, как же соотносятся между собой структурная организация белков и структура кодирующих их генов. Одной из первых задач было выявление соответствия между границами экзонов гена и границами кодируемых этим геном доменов белка. В частности, это могло объяснить молекулярные механизмы возникновения многодоменных белков, некоторые из доменов которых были идентичны у разных организмов, а остальные домены различались.

Гилбертом было выдвинуто предположение, согласно которому один экзон гена кодирует один домен соответствующего белка (Gilbert, 1978). Это предположение давало простое объяснение появлению таких многодоменных белков за счет возможной перетасовки экзонов, при которой происходят встройки экзонов одного гена в экзонно-интронную структуру дру-

гого гена. Явление перетасовки экзонов было открыто следом за работой Гильберта (Maki *et al*., 1980). Позже была показана ключевая роль такой перетасовки в перестройке доменной структуры белков в ходе эволюции (Kaessmann *et al*., 2002). В то же время были показаны случаи, когда один домен может кодироваться несколькими экзонами так же, как и один экзон может кодировать несколько доменов (Graur *et al*., 2000). Эволюционные перестройки доменной структуры таких белков могли происходить за счет множественной перетасовки экзонов, кодирующих один домен белка (Kaessmann *et al*., 2002).

Функция белков в значительной степени определяется их функциональными сайтами, которые могут кодироваться как одним, так и несколькими экзонами. Особенности организации функциональных сайтов белков на уровне экзонной структуры гена могут иметь большое значение для таких эволюционных событий, как формирование доменных структур белков, в том числе и в результате перетасовок экзонов (Ponting *et al*., 2002). Однако работ по анализу особенностей распределения функциональных сайтов белков на экзонной структуре генов пока не проведено. Исследованию этого вопроса на примере сайтов связывания АТФ, АДФ, АМФ (аденозинфосфат-связывающие сайты, АФСС) посвящена данная работа. Метаболические, а также регуляторные процессы, проходящие с участием этих лигандов, распространены во всех организмах. Белки, вовлеченные в эти процессы путем связывания аденозинфосфатов, как правило, имеют длительную эволюционную историю (Oswald *et al*., 2006). Поиск взаимосвязей между структурно-функциональной организацией белков и экзонной структурой генов может иметь большое значение для понимания процессов молекулярной эволюции.

Материалы и методы

Для получения информации о разметке сайтов связывания АТФ, АДФ, АМФ на последовательностях белков использовалась база данных (БД) функциональных сайтов белков PDBSite (Ivanisenko *et al*., 2005). Эта база данных содержит информацию о координатах атомов аминокислотных остатков различных типов функциональных сайтов, а также об их положении в последовательностях белков, представленных в БД пространственных структур белков PDB (Berman *et al*., 2000). Поскольку последовательности одного и того же белка, приведенные в БД PDB и БД SWISS-Prot (Watanabe, Harayama, 2001), могут различаться, то проводилось их выравнивание с использованием программы BLAST (Altschul *et al*., 1997). Различия в последовательностях этих двух баз связаны с тем, что в БД PDB часто приводятся фрагменты белка, которые удалось кристаллизовать, а также мутантные формы белков. Таким образом, с помощью выравнивания нами были установлены позиции аденозинфосфат-связывающих сайтов в последовательностях SWISS-Prot. Дополнительно на этих последовательностях проводилась разметка границ экзонов, которая извлекалась из БД EMBL (Cochrane *et al*., 2008). Для этих целей нуклеотидную последовательность транслировали в аминокислотную, которая затем выравнивалась с последовательностью соответствующего белка из базы данных SWISS-Prot. В данном случае выравнивание осуществлялось с помощью программы ClustalW (Thompson *et al*., 2002).

Анализ проводился на наборе негомологичных белков с уровнем сходства между собой менее 30 %. Рассматривались только многоклеточные организмы и только те гены, для которых была информация об их полной экзон-интронной структуре. Таким образом, для анализа было отобрано 63 белка, содержащих аденозинфосфат-связывающие сайты следующих организмов: человек – 51 белок (81 %); мышь – 3 белка (5 %); нематода – 3 белка (5 %); бык – 2 белка (3 %); крыса – 2 белка (3 %); плодовая мушка – 1 белок (2 %); курица – 1 белок (2 %). В выборке белков были представлены все 6 классов ферментов по номенклатуре ЕС: оксидоредуктазы (4 белка), трансферазы (28 белков), изомеразы (1 белок), лигазы (4 белка), лиазы (2 белка), гидролазы (10 белков). Доля ферментов составила 71 %, доля транспортных белков 8 % (5 белков), доля белков клеточного цикла 5 % (3 белка), доля сократительных белков 3 % (2 белка), доля других белков 13 % (10 белков).

Статистический анализ выборок осуществлялся с помощью математического пакета STATISTICA.

Результаты

Предварительный анализ выборки показал, что в генах в среднем присутствовало 12 экзонов, в то время как среднее число экзонов в генах позвоночных равно 7. При этом 90 % генов содержат 12 и менее экзонов в последовательности (Lewin, 1994). Из них четыре экзона было задействовано в кодировании аденозинфосфат-связывающего сайта. Среднее количество аминокислотных остатков, образующих в белке АФСС сайт, составило 15. Таким образом, сайты связывания АТФ, АДФ и АМФ включают значительное число аминокислот (по сравнению, например, с каталитическими триадами) и кодируются несколькими экзонами.

Для дальнейшего анализа интерес представляла задача сравнения распределений длин экзонов, кодирующих и некодирующих АФСС. Соответствующие распределения приведены на рис. 1. Тест χ^2 показал их статистически значимые различия ($P < 0{,}00001$), при этом значение χ^2 было равным 962,5 при 45 степенях свободы. Средние длины экзонов, кодирующих и некодирующих АФСС, оказались 135 и 147 п.о. соответственно.

Также была выдвинута гипотеза о том, что аминокислоты АФСС кодируются различными экзонами последовательности с равной вероятностью, не зависят от длины экзона и от его положения в последовательности. Для проверки этой гипотезы позиции аминокислот АФСС в каждой последовательности анализируемой выборки менялись случайным образом и фиксировались их новые позиции в рамках границ экзонов, ранее отмеченных на последовательности гена. При этом позиции границ экзонов не менялись. Перемешивание позиций аминокислот проводили 1000 раз для каждой последовательности из выборки, и на основе результатов таких теоретических экспериментов было получено ожидаемое по случайным причинам распределение соответствующего числа аминокислот АФСС, закодированных одним экзоном. Сравнение двух распределений, наблюдаемого и ожидаемого по случайным причинам (рис. 2), показало по критерию χ^2 статистически значимое различие между ними с уровнем значимости, близким к нулю (статистика χ^2 была равна 681 при 25 степенях свободы).

Для того чтобы оценить, насколько удалены друг от друга аминокислоты АФСС в кодирующей экзонной структуре гена, необходимо использование специального оценочного коэффициента. Для этих целей рассчитывался коэффициент разрывности α_1, характеризующий количество экзонов, кодирующих АФСС в заданной последовательности и их положение друг относительно друга в последовательности гена:

$$\alpha_1 = 1 - \frac{N}{p_N^E - p_1^E + 1}, \qquad (1)$$

где, p_1^E – порядковый номер первого экзона в последовательности гена, кодирующего АФСС, p_N^E – порядковый номер последнего экзона в последовательности гена, кодирующего АФСС,

Рис. 1. Плотность распределения экзонов, кодирующих АФСС (I) и некодирующих его (II).

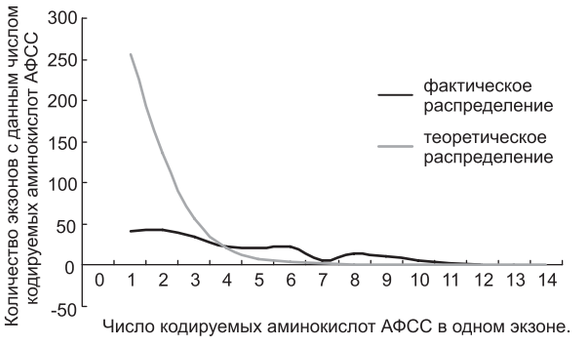

Рис. 2. Наблюдаемое и ожидаемое по случайным причинам распределения числа аминокислот АФСС, располагающихся в границах одного экзона.

N – число экзонов последовательности гена, кодирующих АФСС.

На рис. 3 показано распределение коэффициента разрывности α_1. Чем больше значение коэффициента, тем дальше расположены друг относительно друга экзоны последовательности гена, кодирующие аминокислоты АФСС белка. Исходя из построенного распределения коэффициента разрывности можно заключить, что 28 % АФСС кодируются исключительно соседними экзонами в последовательности.

Коэффициент разрывности α_2 использовался для оценки удаленности друг от друга аминокислот АФСС в первичной структуре белка:

$$\alpha_2 = 1 - \frac{M}{p_M^A - p_1^A + 1},$$

где *M* – количество аминокислот АФСС в последовательности, p_M^A – позиция последней аминокислоты сайта, p_1^A – позиция первой.

Распределение этого коэффициента, полученное для анализируемой выборки АФСС, показано на рис. 4. Из рисунка можно видеть, что в отличие от распределения коэффициента α_1 наблюдается рост числа АФСС при увеличении коэффициента разрывности по аминокислотной последовательности. Большинство АФСС характеризуются α_2, близким к максимальному, в то время когда экзоны, их кодирующие, оказываются сближенными в последовательности гена. В этой связи было интересным узнать, как соотносятся позиции аминокислот АФСС в последовательности белка и границы функциональных доменов в нем. Для определения границ функциональных доменов на анализируемых последовательностях белков использовались данные из БД Pfam-A (Finn *et al*., 2008). В результате оказалось, что лишь 60 % АФСС (38 АФСС) принадлежит одному известному функциональному домену, 5 % (3 АФСС) – двум, а 35 % (22 АФСС) принадлежат полностью или частично той части последовательности белка, в которой не обнаружен известный или гипотетический домен.

Существуют свидетельства того, что функциональные сайты белков с различной частотой присутствуют на N- и C-конце последовательности (Hodgman, 1986). В связи с этим было интересно узнать позиции экзонов в последовательности гена, которые кодируют аминокислоты АФСС. Для этого мы рассматривали только гены, содержащие более 5 экзонов. Графики, отражающие частоту позиции экзона, кодирующего аминокислоты АФСС, приведены на рис. 5. Из рисунка видно, что с наибольшей

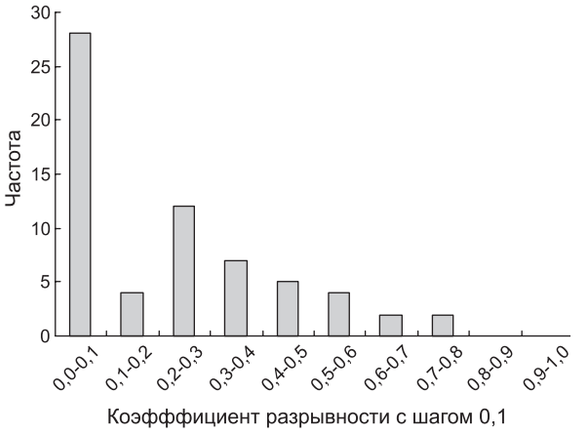

Рис. 3. Распределение коэффициента разрывности α_1.

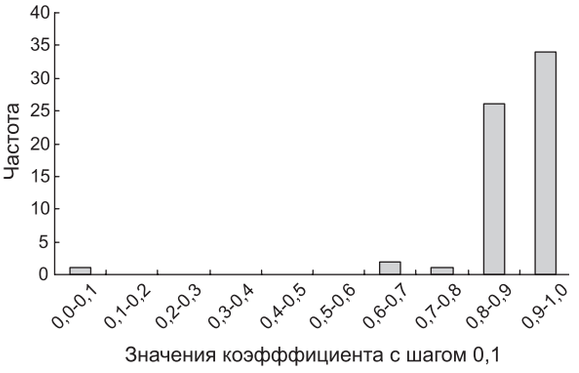

Рис. 4. Распределение коэффициента α_2.

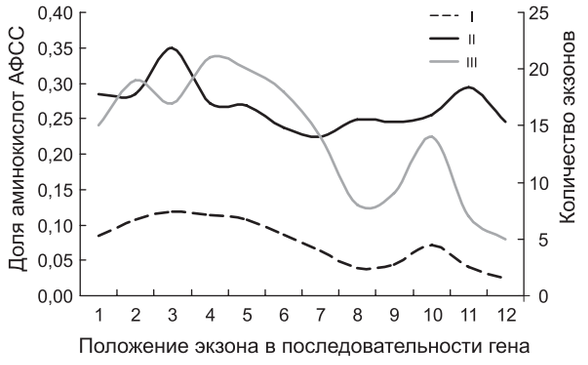

Рис. 5. Частота, с которой аминокислоты АФСС закодированы экзоном в определенной позиции последовательности гена.

I – доля аминокислот АФСС, закодированных в одном экзоне в зависимости от его положения в последовательности гена; II – тот же показатель, что и в случае 1, но рассчитанный только для экзонов, кодирующих АФСС; III – количество экзонов, кодирующих хотя бы одну аминокислоту АФСС в зависимости от их положения в гене.

частотой аминокислоты АФСС кодируются первыми 5 экзонами, а также экзонами, располагающимися на конце последовательности гена.

Обсуждение

Аденозинфосфат-связывающие белки распространены во всех живых организмах, участвуют во многих процессах: клеточное дыхание, фотосинтез, биосинтетические реакции, деление клетки, передача сигналов и т. д.

Анализ выборки генов, кодирующих аденозинфосфат-связывающие белки, продемонстрировал несколько особенностей распределения АФСС на экзонной структуре этих генов. Одним из примечательных свойств можно назвать достоверное различие распределений длин экзонов, кодирующих и некодирующих АФСС. Такое различие показывает существование эволюционного отбора на определенную длину экзонов, кодирующих аминокислоты АФСС. При этом на исследованном наборе генов средняя длина экзона, кодирующего аминокислоты АФСС, оказалась меньше средней длины экзонов, которые не участвуют в кодировании аминокислот этих сайтов. Является ли этот факт следствием вставки интронов в процессе эволюции, может показать дальнейшее исследование ортологичных групп генов. Свидетельством в пользу такого предположения может быть также тенденция кодирования аминокислот АФСС несколькими экзонами, близко расположенными в последовательности гена, несмотря на значительное удаление аминокислот АФСС в первичной структуре белка.

Еще одной особенностью распределения АФСС оказалась более высокая частота их кодирования на 5′-конце последовательности гена, т. е. в первых экзонах последовательности. Для понимания найденной закономерности также необходимо проведение дополнительных исследований.

Более глубокое изучение эволюционных особенностей распределения АФСС на экзонной структуре гена в дальнейшем нами планируется на основе филогенетического анализа последовательностей. Следует принять во внимание, что исследованный набор белков также содержит другие функциональные сайты, такие, как каталитические центры, сайты связывания субстратов, кофакторов и т. д. Наличие таких сайтов может влиять на особенности организации экзонной структуры гена. Поэтому нами будет проведено комплексное изучение распределения также и других типов функциональных сайтов для данного набора белков.

Благодарности

Работа частично была поддержана проектами СО РАН № 115, 10.7, 18.13, государственным контрактом с ФАНИ № 02.514.11.4065, а также НШ-2447.2008.4. Научная школа Н.А. Колчанова «Биоинформатика и системная компьютерная биология».

Литература

Altschul S.F., Madden T.L., Schaffer A.A. *et al.* Gapped BLAST and PSIBLAST: a new generation of protein database search programs // Nucl. Acids Res. 1997. V. 25. P. 3389–3402.

Andreeva N.S. The structure of globular proteins according to x-ray structural crystallography data // Usp. Sovrem. Biol. 1964. V. 58. P. 3–21.

Berman H.M., Westbrook J., Feng Z. *et al.* The Protein Data Bank // Nucl. Acids Res. 2000. V. 28. P. 235–242.

Cochrane G., Akhtar R., Aldebert P. *et al.* Priorities

for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database // Nucl. Acids Res. 2008. V. 36. (Database issue). P. D5–12.
Finn R.D., Tate J., Mistry J. *et al.* // Nucl. Acids Res. 2008. Database Issue 36. P. D281–D288.
Gilbert W. Why genes in pieces // Nature. 1978. V. 271(5645). P. 501.
Graur D., Li W.H. Fundamentals of molecular evolution. Sinauer Associates Inc., Sunderland. Second ed. 2000. P. 481.
Hodgman T.C. The elucidation of protein function from its amino acid sequence // Bioinformatics. 1986. V. 2. № 3. P. 181–187.
Ivanisenko V.A., Pintus S.S., Grigorovich D.A., Kolchanov N.A. PDBSite: a database of the 3D structure of protein functional sites // Nucl. Acids Res. 2005. V. 33. (Database issue). P. D183–187.
Kaessmann H., Zollner S., Nekrutenko A., Li W.H. Signatures of domain shuffling in the human genome // Genome Res. 2002. V. 12(11). P. 1642–1650.
Lewin B. Genes V. N.Y.: Oxford University Press. 1994. P. 1272.
Maki R., Traunecker A., Sakano H. *et al.* Exon shuffling generates an immunoglobulin heavy chain gene // Proc. Natl Acad. Sci. USA. 1980. V. 77(4). P. 2138–2142.
Oswald C., Holland I.B., Schmitt L. The motor domains of ABC-transporters. What can structures tell us? // Naunyn Schmiedebergs Arch. Pharmacol. 2006. V. 372(6). P. 385–399.
Ponting C.P., Russel R.R. The natural history of protein domains // Annu. Rev. Biophys. Biomol. Struct. 2002. V. 31. P. 45–71.
Richards F.M. Structure of proteins // Annu. Rev. Biochem. 1963. V. 32. P. 269–300.
Thompson J.D., Gibson T.J., Higgins D.G. Multiple sequence alignment using ClustalW and ClustalX // Curr. Protoc. Bioinformatics. 2002. Chapter 2. Unit 2.3.
Watanabe K., Harayama S. SWISS-PROT: The curated protein sequence database on Internet // Protein, Nucl. Acid and Enzyme. 2001. V. 46. P. 80–86.

DISTRIBUTION ANALYSIS OF ADENOSINE PHOSPHATE BINDING SITES OF PROTEINS ON EXON STRUCTURE OF GENE

I.V. Medvedeva, P.S. Demenkov, V.A. Ivanisenko

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia,
e-mail: brukaro@bionet.nsc.ru, salix@bionet.nsc.ru

Summary

The problem of comparison of tertiary structure of protein and the structure of the encoding gene appeared after the development of the first methods of x-ray crystallography. A lot of papers were devoted to the investigation of the relationships between the exon borders and the domain borders that belong to the protein encoded by the respective gene. The aim of this work was the investigation of the relationships between protein functional sites and exon structure of the encoding genes for the adenosine phosphate binding sites because the conservative processes that adenosine phosphate ligands are involved. It is shown that the length of the exons encoding adenosine phosphate binding sites differs from the other in nonrandom manner. Besides this, the binding sites discovered to be discontinued on the protein primary structure although the most part of them is encoded by the adjacent exons. Also it has been shown that these functional sites' posses discontinuity not only on primary structure but tertiary structure of protein too. The functional sites aminoacids distribution through the encoding exons positions has shown that the most part of binding sites' of ATP, ADP, AMP are encoded by the first few exons.