Информационные процессы, Том 7, № 3, 2007, стр. 199–213.
© 2007 Русин, Рубанов, Любецкий.

Данные для построения эволюционного дерева высших Metazoa

Л. Русин, Л. Рубанов, В. Любецкий

Институт проблем передачи информации РАН
lyubetsk@iitp.ru, rusin@iitp.ru
Поступила в редколлегию 20.06.2007

Аннотация—Описываются методы, частью оригинальные, которые позволили авторам найти 51 ген у 30-и видов, представляющих 11 типов из царства животных. Эти виды представляют все крупные типы животных с настоящими тканями, высшие Metazoa. На основе этих данных авторами будет представлено эволюционное дерево высших Metazoa и приведен эволюционный анализ основных черт организации животных таких, как особенности раннего эмбриогенеза, способы закладки полости тела, формирование общего плана строения организма и т.д.

1. ВВЕДЕНИЕ

Одна из центральных биологических задач состоит в том, чтобы перенести важные закономерности, полученные на модельных организмах, на другие организмы, имеющие высокое экономическое или медицинское значение, которые однако по той или иной причине трудно исследовать экспериментально. Такой перенос возможен только на основе достоверной информации о родстве исследуемого и модельного организмов. Поэтому фундаментальная проблема состоит в построении филогенетического дерева (*филогении*) живых организмов, и особенно высших Metazoa, животных с настоящими тканями и сложным планом строения, к которым относятся позвоночные, включая человека, и беспозвоночные животные. В частности, надежное филогенетическое дерево играет решающую роль в исследованиях эволюции таких сложных систем, как генетическая регуляция раннего дробления яйца, онтогенеза и закладки общего плана строения организма, что, в свою очередь, важно для создания эффективных систем контроля и управления этими процессами у организмов, для которых соответствующие экспериментальные данные не могут быть легко получены; например, у человека.

В филогенетических исследованиях современное многообразие животных представлено выборочно, однако из этого ясно, что некоторые традиционные представления об эволюции животных нуждаются в существенном пересмотре [1],[2]. Основное многообразие планов строения организмов животных встречается среди так называемых *минорных типов*, к которым относятся редкие и часто микроскопические организмы с неясными родственными отношениями. Молекулярные данные, используемые в настоящее время для установления родственности типов животных, оказались в значительной мере недостаточными: филогенетические построения на основе выравниваний лишь нескольких генов (*филогенетических маркеров*) недостаточно надежны и, главное, выборка сравнительных молекулярных данных существенно *асимметрична*: в базах данных всего несколько видов представлены многими гомологичными генами и всего несколько гомологичных генов секвенированы для широкого круга организмов.

Исследования последних лет положили начало отдельному научному направлению на стыке филогенетики и геномики, которое стали называть филогеномикой. Здесь принципиальная трудность состоит в том, что число многоклеточных организмов, для которых известны

полные геномы, весьма ограниченно [3]. Сравнение многих генов у малого числа таксонов (например, более 100 белков для четырех видов в работе [4] и более 500 белков для шести видов в работе [5]) приводит к получению результатов, достоверность которых опирается на анализ слишком малого числа видов [6]. Поэтому увеличение в таких исследованиях числа видов принципиально важно. Кроме того, при большом числе видов и генов надежность построения дерева видов резко возрастает. Настоящая работа направлена на уменьшение такой асимметрии: нами получены около 50 генов из 30 видов, что позволит нам (1) получить адекватное дерево животных для так называемых *больших типов* и (2) составить конкретный список генов, экспериментальное изучение которых существенно для построения эволюционного дерева у минорных типов; так как получение полных геномов у минорных типов крайне затруднительно.

2. МЕТОД И РЕЗУЛЬТАТЫ

2.1. Метод

Опишем наш подход как последовательность задач, которые мы решали.

Задача 1. Составление набора генетических маркеров для основных типов животных.

Здесь цель состояла в нахождении родственных (гомологичных) маркеров (белков и/или белковых доменов) для некоторых представителей типов животных.

Подзадача 1. По базе данных ортологичных групп белков эукариотов (euKaryotic Orthology Groups, KOGs, NCBI, [7]) были составлены два списка. Первый из них включал белки, входящие в структуру рибосомы (79 семейств белков). Второй включал белки, которые представлены ровно одним геном у эукариотических организмов, входящих в эту базу данных (это — представители растений, животных, грибов и микроспоридий). В результате были отобраны белковые семейства с важными клеточными функциями, находящиеся под существенным давлением стабилизирующего отбора. Эти белки, в силу консервативности первичной структуры, предположительно сохраняют молекулярные признаки древних эволюционных событий. Отсутствие во втором списке мультигенных белковых семейств обеспечивает с достаточной надежностью отсутствие паралогичных генов у этих белков. Паралогичные гены непригодны для филогенетического анализа, так как они разделились в результате дупликации, а не видообразования. Объединение этих двух списков дало 178 белковых семейств. При этом были просмотрены все 4852 ортологичных групп белков эукариотов.

Подзадача 2. Гены человека из указанных 178-ми белковых семейств были использованы для поиска гомологичных белков в аннотированных геномах базы RefSeq из банка GenBank, NCBI [8] у представителей типов животных с полным известным геномом, которые перечислены в Таблице 1. При этом для запроса к базе данных использовались алгоритм поиска контекстного сходства (гомологии) аминокислотных последовательностей blastp из семейства BLAST [9] и программа blastcl3 v. 2.2.14 [10] с параметрами командной строки:
-T F -p blastp -d “nr” -e 0.0000000001 -F T -g T -M BLOSUM62 -m 9 -u “phylum_name[orgn]”,
где ключ -u определяет название типа животных, в геноме которого проводится поиск. В результате найдены белки, представляющие эти 178 семейств, для 5 видов из 4 типов животных с полным геномом, Таблица 1.

Таблица 1. Таксономический перечень видов с указанием используемой базы данных. Пометка + означает, что использовалась кДНК, и пометка −, что использовался полный геном. Сокращения dbEST и RefSeq означают dbEST, NCBI или RefSeq, NCBI. Для классов и типов видов кроме латинского по возможности указано русское название.

№	Вид — представитель класса и типа	кДНК или полный геном	Используемая база данных	Класс вида	Тип вида
1	*Mnemiopsis leidyi*	+	dbEST, NCBI	Cyclocoela	Ctenophora (гребневики)
2	*Hydra magnipapillata*	+	dbEST, NCBI	Hydrozoa (гидроидные полипы)	Cnidaria (стрекающие)
3	*Podocoryne carnea*	+	dbEST, NCBI		
4	*Nematostella vectensis*	+	dbEST, NCBI	Anthozoa (кораллы)	
5	*Acropora palmata*	+	dbEST, NCBI		
6	*Priapulus caudatus*	+	dbEST, NCBI	Priapulomorpha	Priapulida
7	*Hypsibius dujardini*	+	TardiBASE, PartiGene	Eutardigrada	Tardigrada (тихоходки)
8	*Schistosoma mansoni*	+	dbEST, NCBI	Trematoda	Platyhelminthes (плоские черви)
9	*Schmidtea mediterranea*	+	dbEST, NCBI	Turbellaria	
10	*Convoluta roscoffiensis*	+	LophDB, PartiGene		
11	*Xiphinema index*	+	NEMBASE2	Dorylaimia	Nematoda (круглые черви)
12	*Caenorhabditis elegans*	−	RefSeq, NCBI	Rhabditia	
13	*Ascaris suum*	+	NEMBASE2		
14	*Boophilus microplus*	+ +	NEMBASE2 NEMBASE2	Chelicerata (хелицеровые)	Arthropoda (членистоногие)
15	*Penaeus sp.*	+	dbEST, NCBI	Crustacea (ракообразные)	
16	*Daphnia pulex*	+	dbEST, NCBI		
17	*Drosophila melanogaster*	−	RefSeq, NCBI	Insecta (насекомые)	
18	*Apis mellifera*	−	RefSeq, NCBI		
19	*Lumbricus rubellus*	+	LumbriBASE, PartiGene	Oligochaeta (малощетинковые черви)	Annelida (кольчатые черви)
20	*Euprymna scolopes*	+	dbEST, NCBI	Cephalopoda (головоногие)	Mollusca (моллюски)
21	*Lymnaea stagnalis*	+	MolluscDB, PatiGene	Gastropoda (брюхоногие)	
22	*Crassostrea gigas*	+	MolluscDB, PatiGene	Bivalvia (двустворчатые)	
23	*Asterina sp.*	+	dbEST, NCBI	Asterozoa (морские звёзды)	Echinodermata (иглокожие)
24	*Strongylocentrotus purpuratus*	−	RefSeq, NCBI	Echinozoa (морские ежи)	
25	*Branchiostoma floridae*	+	dbEST, NCBI	Cephalochordata (головохордовые)	Chordata (хордовые)
26	*Ciona intestinalis*	+	dbEST, NCBI	Tunicata (оболочники)	
27	*Molgula sp.*	+	dbEST, NCBI		
28	*Petromyzon marinus*	+	dbEST, NCBI	Pisces (рыбы)	
29	*Danio rerio*	−	RefSeq, NCBI		
30	*Homo sapiens*	−	RefSeq, NCBI	Mammalia (млекопитающие)	

Подзадача 3. Аналогично 178 генов человека использовались для поиска гомологичных генов в базе dbEST неаннотированных последовательностей кДНК банка GenBank и также в базах неаннотированных последовательностей кДНК web-ресурса Nematode and Neglected Genomics [11],[12] для представителей типов животных соответствующим образом отмеченных в Таблице 1.

При обращении к базе данных dbEST использовался алгоритм tblastn поиска контекстного сходства аминокислотной структуры запроса и нуклеотидной последовательности и программа blastcl3 с параметрами командной строки:
-T F -p tblastn -d "est_others" -e 0.0000000001 -F "T V" -g T -M BLOSUM62 -m 9 -u "phylum_name[orgn]", где ключ -u определяет название типа животных, в геноме которого проводится поиск. Запрос был проведен для тех же 178 семейств.

Базы кДНК на web-ресурсе [11] запрашивались только в случаях, когда гомологичные маркеры не были найдены в базе dbEST или в последней отсутствовали типы животных из Таблицы 1. Использовался web-интерфейс ресурса с аналогичными параметрами tblastn. В результате были получены наборы частично перекрывающихся нуклеотидных последовательностей генов-представителей исходных 178-ми семейств у большинства из видов животных, для которых доступны данные по кДНК, Таблица 2.

Таблица 2. Перечень найденных белковых семейств для видов, перечисленных в Таблице 1; числа от 1 до 30 соответствуют нумерации видов в Таблице 1. Присутствие белка из данного семейства (строка) в данном виде (столбец) отмечается знаком +. Функциональная группа белкового семейства указана по классификации базы данных euKaryotic Orthology Groups, KOGs, NCBI: A — модификация и процессинг РНК, D — контроль клеточного цикла, деление клетки, разделение хромосом, F — транспорт и метаболизм нуклеотидов, I — транспорт и метаболизм липидов, J — трансляционный аппарат, структура рибосомы, K — транскрипция, L — репликация, рекомбинация, O — посттрансляционная модификация, метаболизм белков, шапероны, P — метаболизм и транспорт неорганических ионов, T — механизмы сигнальных каскадов, U — внутриклеточный транспорт, секреция, везикулярный транспорт, Z — цитоскелет. Для семейства указан идентификатор по базе KOGs.

№	Белковое семейство (ген)	Функциональная группа	Идентификатор по базе KOGs	Вид														
				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	Small nuclear ribonucleoprotein Sm D3	A	KOG3172	+	+	+	+	+	+		+	+	+	+	+	+	+	
2	Splicing factor 3b, subunit 1	A	KOG0213	+	+			+	+	+	+	+	+	+	+		+	+
3	rRNA processing protein Rrp5	A	KOG1070	+	+	+		+	+	+	+	+	+		+		+	+
4	Structural maintenance of chromosome protein 3 (sister chromatid cohesion complex Cohesin, subunit SMC3)	D	KOG0964		+	+	+	+		+	+	+	+	+	+	+	+	+
5	Predicted nucleotide kinase/nuclear protein involved oxidative stress response	F	KOG3347	+		+	+	+		+	+	+	+	+	+	+		+
6	Mevalonate kinase MVK/ERG12	I	KOG1511		+	+		+	+	+	+	+		+	+	+	+	+

Таблица 2 (продолжение)

№	Белковое семейство (ген)	Функциональная группа	Идентификатор по базе KOGs	Вид														
				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
7	40S ribosomal protein SA (P40)/Laminin receptor 1	J	KOG0830	+	+	+	+		+	+	+	+	+		+	+	+	
8	40S ribosomal protein S2/30S ribosomal protein S5	J	KOG0877		+	+	+			+	+	+	+	+	+	+	+	
9	40S ribosomal protein S4	J	KOG0378	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
10	40S ribosomal protein S7	J	KOG3320	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
11	40S ribosomal protein S8	J	KOG3283	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
12	40S ribosomal protein S17	J	KOG0187	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
13	40S ribosomal protein S11	J	KOG1728	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
14	40S ribosomal protein S19	J	KOG3411	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
15	40S ribosomal protein S13	J	KOG0400	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
16	40S ribosomal protein S23	J	KOG1749	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
17	40S ribosomal protein S25	J	KOG1767	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
18	40S ribosomal protein S27	J	KOG1779	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
19	Ubiquitin-like/40S ribosomal S30 protein fusion	J	KOG0009	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
20	60s acidic ribosomal protein P1	J	KOG1762	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
21	60s ribosomal protein L2/L8	J	KOG2309	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
22	60s ribosomal protein L6	J	KOG1694	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
23	60s ribosomal protein L7	J	KOG3184	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
24	60s ribosomal protein L9	J	KOG3255	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
25	60s ribosomal protein L10	J	KOG0857	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
26	60s ribosomal protein L5	J	KOG0875	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
27	60s ribosomal protein L13	J	KOG3295	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
28	60s ribosomal protein L15	J	KOG1678	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
29	60s ribosomal protein L18	J	KOG1714	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
30	60s ribosomal protein L18A	J	KOG0829	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Таблица 2 (продолжение)

№	Белковое семейство (ген)	Функциональная группа	Идентификатор по базе KOGs	Вид														
				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
31	60s ribosomal protein L21	J	KOG1732	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
32	60s ribosomal protein L24	J	KOG1722	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
33	60s ribosomal protein L28	J	KOG3412	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
34	60s ribosomal protein L29	J	KOG3504	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
35	60s ribosomal protein L37	J	KOG0402	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
36	Ubiquitin/60s ribosomal protein L40 fusion	J	KOG0003	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
37	RNA polymerase III, large subunit	K	KOG0261	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
38	RNA polymerase I, large subunit	K	KOG0262	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
39	DNA polymerase alpha, catalytic subunit	L	KOG0970	+	+	+	+		+	+	+	+	+		+	+	+	
40	Ribonuclease HI	L	KOG2299		+	+	+			+	+	+	+	+	+	+	+	
41	Molecular chaperone Prefoldin, subunit 3	O	KOG3313	+		+	+	+	+	+		+	+	+	+	+	+	
42	20S proteasome, regulatory subunit beta type PSMB1/PRE7	O	KOG0179		+	+	+	+	+		+	+	+	+	+	+	+	+
43	Beta-tubulin folding cofactor D	O	KOG1943	+		+		+	+	+	+		+	+	+	+	+	+
44	P-type ATPase	P	KOG0209	+	+		+	+	+	+		+	+	+	+		+	+
45	Nuclear porin	P	KOG2196	+	+	+		+	+	+		+	+	+	+	+	+	+
46	Glycosylphosphatidy-linositol anchor synthesis protein	T	KOG2126	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
47	GTP-binding protein DRG2 (ODN superfamily)	T	KOG1486	+	+	+		+	+	+	+	+	+		+	+	+	+
48	Transport protein particle (TRAPP) complex subunit	U	KOG3315	+		+	+	+	+	+		+	+	+	+	+	+	+
49	Vacuolar sorting protein VPS45/Stt10 (Sec1 family)	U	KOG1299	+	+	+		+	+	+	+	+	+	+	+	+	+	+
50	Membrane component of ER protein translocation complex	U	KOG2927		+	+	+	+	+	+	+	+	+	+	+	+	+	+
51	Actin-related protein Arp2/3 complex, subunit ARPC4	Z	KOG1876	+	+	+	+	+	+		+	+	+	+	+	+	+	+

Таблица 2 (продолжение по видам)

№	Белковое семейство (ген)	Функциональная группа	Идентификатор по базе KOGs	Вид														
				16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	Small nuclear ribonucleoprotein Sm D3	A	KOG3172		+	+	+		+	+	+	+	+	+	+	+	+	+
2	Splicing factor 3b, subunit 1	A	KOG0213	+	+	+	+	+		+	+	+		+	+		+	+
3	rRNA processing protein Rrp5	A	KOG1070	+	+	+	+	+		+	+	+		+	+	+	+	+
4	Structural maintenance of chromosome protein 3 (sister chromatid cohesion complex Cohesin, subunit SMC3)	D	KOG0964	+	+	+	+		+	+	+	+	+	+	+	+	+	+
5	Predicted nucleotide kinase/nuclear protein involved oxidative stress response	F	KOG3347	+	+	+	+	+	+		+	+	+	+	+	+	+	+
6	Mevalonate kinase MVK/ERG12	I	KOG1511	+	+		+	+	+	+	+	+		+	+	+	+	+
7	40S ribosomal protein SA (P40)/Laminin receptor 1	J	KOG0830	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
8	40S ribosomal protein S2/30S ribosomal protein S5	J	KOG0877	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
9	40S ribosomal protein S4	J	KOG0378	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
10	40S ribosomal protein S7	J	KOG3320	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
11	40S ribosomal protein S8	J	KOG3283	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
12	40S ribosomal protein S17	J	KOG0187	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
13	40S ribosomal protein S11	J	KOG1728	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
14	40S ribosomal protein S19	J	KOG3411	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
15	40S ribosomal protein S13	J	KOG0400	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
16	40S ribosomal protein S23	J	KOG1749	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
17	40S ribosomal protein S25	J	KOG1767	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
18	40S ribosomal protein S27	J	KOG1779	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
19	Ubiquitin-like/40S ribosomal S30 protein fusion	J	KOG0009	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
20	60s acidic ribosomal protein P1	J	KOG1762	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Таблица 2 (продолжение)

№	Белковое семейство (ген)	Функциональная группа	Идентификатор по базе KOGs	Вид														
				16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
21	60s ribosomal protein L2/L8	J	KOG2309	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
22	60s ribosomal protein L6	J	KOG1694	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
23	60s ribosomal protein L7	J	KOG3184	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
24	60s ribosomal protein L9	J	KOG3255	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
25	60s ribosomal protein L10	J	KOG0857	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
26	60s ribosomal protein L5	J	KOG0875	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
27	60s ribosomal protein L13	J	KOG3295	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
28	60s ribosomal protein L15	J	KOG1678	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
29	60s ribosomal protein L18	J	KOG1714	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
30	60s ribosomal protein L18A	J	KOG0829	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
31	60s ribosomal protein L21	J	KOG1732	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
32	60s ribosomal protein L24	J	KOG1722	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
33	60s ribosomal protein L28	J	KOG3412	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
34	60s ribosomal protein L29	J	KOG3504	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
35	60s ribosomal protein L37	J	KOG0402	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
36	Ubiquitin/60s ribosomal protein L40 fusion	J	KOG0003	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
37	RNA polymerase III, large subunit	K	KOG0261	+	+	+		+	+	+	+	+	+		+	+	+	+
38	RNA polymerase I, large subunit	K	KOG0262	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
39	DNA polymerase alpha, catalytic subunit	L	KOG0970	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
40	Ribonuclease HI	L	KOG2299	+	+	+	+	+		+	+	+	+		+	+	+	+
41	Molecular chaperone Prefoldin, subunit 3	O	KOG3313	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
42	20S proteasome, regulatory subunit beta type PSMB1/PRE7	O	KOG0179		+	+		+	+	+	+	+	+	+	+	+	+	+
43	Beta-tubulin folding cofactor D	O	KOG1943	+	+	+	+	+	+	+	+	+	+	+	+		+	+
44	P-type ATPase	P	KOG0209	+	+	+	+		+	+	+	+	+		+	+	+	+
45	Nuclear porin	P	KOG2196	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Таблица 2 (продолжение)

№	Белковое семейство (ген)	Функциональная группа	Идентификатор по базе KOGs	Вид														
				16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
46	Glycosylphosphatidylinositol anchor synthesis protein	T	KOG2126	+	+	+		+	+		+	+	+	+	+	+	+	+
47	GTP-binding protein DRG2 (ODN superfamily)	T	KOG1486		+	+	+	+	+	+	+	+	+	+		+	+	+
48	Transport protein particle (TRAPP) complex subunit	U	KOG3315	+	+	+	+	+	+	+	+	+		+	+		+	+
49	Vacuolar sorting protein VPS45/Stt10 (Sec1 family)	U	KOG1299		+	+		+	+	+	+	+	+	+		+	+	+
50	Membrane component of ER protein translocation complex	U	KOG2927	+	+	+	+	+		+	+	+	+	+	+	+	+	+
51	Actin-related protein Arp2/3 complex, subunit ARPC4	Z	KOG1876	+	+	+		+	+	+	+	+	+	+		+	+	+

Подзадача 3 имеет следующую особенность. Быстрота получения и секвенирования одноцепочечных кДНК достигается за счёт малой длины и часто низкого качества последовательностей, для которых кроме того неизвестна принадлежность к “+” или “−”-цепи ДНК. Для сборки полного гена (контига) из частичных последовательностей последние сравниваются в шести рамках считывания и находятся области их перекрывания. Для этой цели нами адаптирован Perl-скрипт tgicl Института геномных исследований TIGR, [13]. Сборка контигов проводилась для каждого гена и каждого представителя животных с данными из базы кДНК. Иногда такое перекрывание имеет место только для нескольких последовательностей, что из-за низкого качества индивидуальных кДНК, зачастую приводит к сбоям рамки считывания или к появлению неопределённых нуклеотидных оснований. В этом случае для формирования контига и его транслирования использовались методы, основанные на вычислениях частот кодонов в каждом из геномов, ESTScan [14] и DECODER, [15]. На последующих этапах работы использовались транслированные версии кДНК.

В результате для 25 видов из 11 типов животных были получены белки, представляющие большинство из исходных 178-ми семейств белков, Таблица 2.

Два списка белков, полученных в результате решения задачи 1, объединились в общий список.

Задача 2. Определение ортологичных маркеров.

Каждое семейство белков из последнего списка проверялось на присутствие в нём паралогов с использованием метода “наилучшего возвратного соответствия” (BLAST score, [16]). За опорный геном был принят геном человека как один из наиболее полно изученных и аннотированных. Для каждого гена-представителя белкового семейства из генома человека выполнялся запрос к геному интересующего нас организма, и результат использовался для обратного запроса к геному человека. Найденный ген считался ортологичным, если и только если результат обратного запроса совпадал с исходным геном в геноме человека. В случае базы кДНК, когда присутствие всех генов у рассматриваемого генома не может быть гарантировано и, следовательно, настоящее родство (ортологичность) при обратном запросе не имеет строгого обоснования, в процедуру обратного запроса был введён второй опорный геном (нематода

C. elegans). Ген из базы кДНК считался ортологичным, если и только если результаты запросов "человек⟶база⟶человек" и "человек⟶база⟶*C. elegans*⟶человек" совпадали. Если в каком-то семействе более половины белков являлись паралогами, то это семейство удалялось из списка. В результате был получен меньший список семейств белков.

Для каждого так полученного семейства белков были построены множественные выравнивания (МВ) с использованием программы PROBCONS v 1.1 [17] с параметрами по умолчанию. Это выравнивание было улучшено в результате решения следующей Задачи 3.

Задача 3. Улучшение качества множественных выравниваний.

Поскольку реконструкция филогении с использованием геномных данных требует сравнения большого числа филогенетических деревьев отдельных белковых семейств, качество филогении напрямую зависит от качества МВ. Эволюция биологических макромолекул приводят к мутационному насыщению некоторых их участков, где имеют место высокие скорости замещения нуклеотидов и/или аминокислот. Это приводит к сокращению числа гомологичных признаков и к потере филогенетического сигнала в высоко вариабельных областях. Здесь использовался разработанный авторами алгоритм удаления информационного шума в МВ, т.е. удаления отдельных столбцов из МВ, которые плохо согласуются с набором надежных клад (имеют высокую условную энтропию), — такое удаление выполняется до достижения первого локального минимума некоторой специальной функции, характеризующей наилучшую степень "очистки филогенетического сигнала" в МВ. Филогенетическое дерево белкового семейства строилось по таким образом улучшенному МВ [18]. Этот алгоритм применялся ко всем МВ, полученным в результате решения Задачи 2.

Задача 4. Выявление информативных генов.

Цель задачи состоит в выявлении маркеров из исходного списка, которые обеспечивают "хорошую" реконструкции родственных отношений высокого таксономического уровня. Для определения количества "древнего" филогенетического сигнала использовался подход, основанный на анализе распределения длин большого числа случайных деревьев, генерированных для последовательностей из любого данного МВ. Здесь длина дерева равна суммарному количеству молекулярных замещений (мутаций) по всем его ветвям. По критерию максимальной экономии, чем меньше длина дерева, тем выше достоверность того, что оно правильно описывает преемственность эволюционных событий. Форма кривой распределения длин большого числа случайных деревьев несет информацию о доле более коротких деревьев и, таким образом, измеряет филогенетическую информативность данного МВ [19]. Для выявления такого сигнала на уровне расхождения крупных таксономических групп животных, нами был разработан алгоритм, позволяющий генерировать бинарные случайные деревья, разрешающие все политомии данного дерева, см. пункт 3 ниже. Алгоритм генерирует бинарные деревья, в каждом из них случайным образом соединяя заданный список узлов (клад), которые объединяют представителей типов животных, когда их общность происхождения (монофилетичность) надёжно установлена. Эксцесс формы кривой распределения длин деревьев используется для определения пороговой доли коротких деревьев, при которой МВ считается информативным [19]. МВ, не удовлетворяющие такому требованию, и соответствующие им белковые семейства исключались. В результате получен окончательный список из 51 белка для 30 видов из 11 типов животных.

2.2. Результаты

Составлен набор генетических маркеров для определения родственных отношений типов животных. В него вошёл 51 ген для большинства из 30-ти видов животных-представителей 11-ти типов, существенно характеризующих состав высших Metazoa, Таблица 2. Гены из этого списка удовлетворяют требованиям, сформулированным в Задачах 2 и 4. В современных ис-

следованиях по филогеномике, впервые получен набор из десятков генов для представителей более чем четырёх типов животных. На этой основе сейчас завершается работа по построению филогенетического дерева, которое устанавливает родственные отношения этих 11-ти типов животных на основе сравнения десятков генов. Это дерево позволит решить ряд актуальных вопросов, касающихся, в частности, примитивности черт организации гребневиков, общности закладки вторичной полости тела (целома) в пределах билатерально-симметричных животных и её судьбы у типов, считающихся ацеломическими, таких, как нематоды и плоские черви.

3. АЛГОРИТМ ПОРОЖДЕНИЯ СЛУЧАЙНЫХ БИНАРНЫХ ДЕРЕВЬЕВ, СОГЛАСОВАННЫХ С ДАННЫМ ПОЛИТОМИЧЕСКИМ ДЕРЕВОМ

Для построения случайного набора бинарных деревьев, являющихся вариантами бинарного разрешения исходного бескорневого политомического дерева, в котором значительная часть вершин имеет степень 3, а остальные (за исключением листьев) — больше трёх, нами был разработан алгоритм, который последовательно разрешает все политомические (степени > 3) вершины путем разбиения множества всех вершин, инцидентных такой вершине, на три непустых подмножества. Затем, если какое-то из подмножеств содержит два или более элементов, то вводится вспомогательная вершина, по построению инцидентная всему этому подмножеству вершин и исходной политомической вершине, тогда как вершины подмножества перестают быть инцидентными исходной вершине. В итоге исходная политомическая вершина приобретает степень 3, но к дереву добавляется от 1 до 3 дополнительных вершин более низкой степени, чем у исходной вершины. Описанный процесс повторяется до исчерпания всех политомических вершин, т.е. построения бинарного дерева. Очевидно, этот алгоритм конечен, поскольку если в качестве исходной вершины всегда выбирать вершину наибольшей степени во всем дереве, то эта наибольшая степень на каждом шаге будет строго монотонно уменьшаться, и неизбежно достигнет значения 3, т.е. дерево станет бинарным, что и является условием окончания разбиения. Поскольку такие разбиения осуществляются во всех политомических вершинах независимо (если их сравнительно немного), то общее число возможных различных вариантов бинарного разрешения дерева с s политомическими вершинами, оценивается сверху как произведение s значений, каждое из которых есть число различных вариантов разложения одной из политомических вершин в предположении, что степень остальных вершин не больше 3:

$$R = \prod_{k=1}^{s} N(n_k), \tag{1}$$

где n_k — степень k-й политомической вершины. Для $N(n)$ в [20] предлагается формула

$$N(n) = \frac{(2n-5)!}{2^{n-3} \cdot (n-3)!}, \tag{2}$$

поэтому произведение (1) может достигать больших величин, что позволяет проводить статистический анализ на множестве предполагаемых бинарных деревьев.

Чтобы статистические оценки, полученные с использованием построенного набора бинарных деревьев, были достоверными, желательно обеспечить неповторяемость деревьев в этом наборе. Очевидный способ достичь этого может состоять просто в запоминании всех построенных деревьев и сравнении каждого вновь построенного со всеми предыдущими. Однако он, очевидно, слишком трудоемкий. Нами для этой цели использовалась хэш-функция, задающая инъективное отображение множества всех деревьев на множество целых чисел от 0 до некоторого $K-1$. Тогда, если вести массив из K битовых элементов, помечая в нем значения хэш-функции для уже построенных деревьев, можно сделать трудоемкость проверки неповторяемости константой, не зависящей от числа уже найденных вариантов разрешения, и притом

малой, поскольку трудоемкость вычисления значения хэш-функции для строки, как правило, линейно зависит от ее длины. Поскольку отображение инъективно, такой способ теоретически позволяет построить K различных деревьев, но, к сожалению, лишь теоретически, поскольку не все значения хэш-функции реально достигаются. Это заставляет выбирать K с большим запасом (скажем, K^2 вместо K), что повышает затраты памяти. Однако главный недостаток связан с вырожденностью хэш-функции: реально возможны так называемые коллизии, когда двум различным деревьям соответствует одинаковое значение хэш-функции. Это значит, что в данном методе принципиально невозможно осуществить порождение всех вариантов разрешения политомического дерева. Какая доля деревьев останется не выявленной, зависит от применяемой хэш-функции. Известны и широко используются в криптографии специально разработанные хэш-функции (например, MD5) с весьма низкой вероятностью коллизий, однако их применение затруднительно, поскольку они реализуют отображение строки символов на целочисленное множество с числом элементов $K = 2^{128} \dots 2^{160}$.

При любом способе сравнения деревьев отдельная трудность вытекает из неоднозначности представления дерева в скобочном формате, так как возможен разный порядок перечисления рёбер, исходящих из одной вершины. Чтобы результаты сравнения были достоверными, после построения каждого нового дерева оно нормализуется путём лексикографического упорядочивания внутри каждой группы вершин, инцидентных одной и той же вершине. При формировании скобочной записи за корень бескорневого бинарного дерева для единообразия принимается родительская вершина того из листьев дерева, который стоит первым в лексикографическом порядке.

Поскольку заранее было неясно, какой из описанных способов окажется более эффективным на практике, в текущей версии 1.8 разработанной нами программы GenTree, реализующей этот алгоритм, предусмотрены все три варианта порождения деревьев: без контроля повторений, с контролем путём непосредственного сравнения со всеми ранее построенными деревьями, с контролем на основе хэш-функции, отображающей дерево в скобочном формате на множество с числом элементов $K = 2^{32}$. Начальное представление о результатах работы программы дает таблица 3, где приведены результаты работы на тестовых примерах, в которых исходное дерево получалось из чисто бинарного дерева T_3 путем частичного склеивания вершин, так чтобы появлялась ровно одна политомическая вершина со степенью от 4 до 10. Кроме того, из T_3 были получены ещё дерево $T_{4,4}$ с двумя политомическими вершинами степени 4 и дерево $T_{8,4,6}$ с тремя политомическими вершинами со степенями 8, 6 и 4 соответственно. Приведем соответствующие деревья в нормализованном скобочном формате:

T_3: (((A,B),((((C,D),E),I),(F,(G,H)))),((J,M),(K,L)),(((N,O),P),Q));
T_4: (((A,B),(((C,D,E),I),(F,(G,H)))),((J,M),(K,L)),(((N,O),P),Q));
T_5: (((A,B),((C,D,E,I),(F,(G,H)))),((J,M),(K,L)),(((N,O),P),Q));
T_6: (((A,B),(C,D,E,(F,(G,H)),I)),((J,M),(K,L)),(((N,O),P),Q));
T_7: ((A,B,(((C,D),E),I),F,G,H),((J,M),(K,L)),(((N,O),P),Q));
T_8: (((A,B),(C,D,E,F,G,H,I)),((J,M),(K,L)),(((N,O),P),Q));
T_9: (A,B,C,D,E,F,(G,H),I,(((J,M),(K,L)),(((N,O),P),Q)));
T_{10}: ((A,B,C,D,E,F,G,H,I),((J,M),(K,L)),(((N,O),P),Q));
$T_{4,4}$: (((A,B),(((C,D,E),I),(F,G,H))),((J,M),(K,L)),(((N,O),P),Q));
$T_{8,4,6}$: (A,B,C,D,E,I,(F,G,H),(((J,M),(K,L)),N,O,P,Q));

В программе реализованы три независимых варианта условия остановки алгоритма: (а) по числу построенных вариантов бинарного дерева (что в сочетании с проверкой уникальности рискованно, т.к. априори неизвестно, существует ли вообще заданное число деревьев и удастся ли их получить за разумное время); (б) по общему времени работы (“длительность”); (в) по времени, прошедшему с момента обнаружения последнего уникального дерева (“задержка”). В таблице 3 приведены типичные результаты работы программы для случая без контроля

повторений за время $t = 10$ с. (вариант (б)), а также для двух способов контроля повторений с остановкой алгоритма по любому из критериев (б) и (в) при максимальной длительности $t = 3600$ с. и задержке $d = 10$ с.

Таблица 3. Результаты работы программы GenTree. Обозначения: n — степень политомической вершины; R — теоретически возможное число уникальных бинарных деревьев, согласно (1), (2); N_0 — общее число построенных деревьев без проведения проверки их уникальности; N_c — общее число построенных деревьев при проверке их уникальности путем непосредственного сравнения с ранее построенными; R_c — число найденных уникальных деревьев; t_c — время счета в секундах; N_h, R_h, t_h — аналогичные величины для случая, когда проверка уникальности выполняется с помощью хэш-функции.

n	R	Без контроля ($t = 10$ с.)	Контроль сравнением ($d = 10$ с.)			Контроль с хэш-функцией ($d = 10$ с.)		
		N_0	N_c	R_c	t_c	N_h	R_h	t_h
4	3	400704	500736	3	10	491708	3	10
5	15	403405	487792	15	10	468962	15	10
6	105	350254	448823	105	10	431375	105	10
7	945	360216	271964	945	10	396449	944	10
8	10395	349301	144536	10391	107	288638	7863	13
9	135135	345433	425539	118782	3600	3917020	121136	129
10	2027025	333676	272197	248749	3600	47361222	878834	1684
4;4	9	400280	473287	9	10	621127	9	10
8;4;6	3274425	307060	266118	252444	3600	47363564	1043863	1188

Как видно из таблицы, производительность собственно порождения деревьев без контроля их уникальности слабо зависит от степени политомической вершины и числа таких вершин (скорость определяется в основном общими размерами дерева). При контроле повторяемости деревьев методом прямого сравнения производительность программы заметно падает с ростом степени политомической вершины, ввиду увеличения общего числа несовпадающих деревьев, с которыми необходимо сравнивать каждое построенное дерево. Контроль уникальности деревьев с помощью хэш-функции осуществляется значительно быстрее, однако уже при степени вершины 8 удается найти только около 70% от всех уникальных деревьев, хотя и на порядок быстрее. При дальнейшем увеличении степени вершины трудоемкость возрастает настолько, что за выбранное время счета 1 час непосредственное сравнение дает даже меньше уникальных деревьев, чем при использовании хэш-функции.

Динамику порождения неповторяющихся деревьев при двух способах проверки уникальности для случая T_9 наглядно демонстрирует рисунок 1 (приведена зависимость числа найденных уникальных вариантов разрешения от времени счета в секундах).

Как видно из графиков, способ с использованием хэш-функции позволяет быстрее получить основную часть неповторяющихся деревьев, но затем частота появления новых вариантов резко падает, и перебрать все оказывается невозможно. В случае прямого сравнения вновь генерируемых деревьев с каждым из уже найденных уникальных алгоритм оказывается значительно медленнее, что может свести на нет его потенциальную способность сгенерировать все возможные варианты разрешения исходного дерева. Правильный выбор варианта подскажут результаты дальнейших экспериментов с реальными деревьями, но более перспективным все же представляется вариант с хэш-функцией. Можно будет использовать специальные виды хэш-функции для уменьшения частоты возникновения коллизий, что позволит полнее исчерпывать предельное значение R.

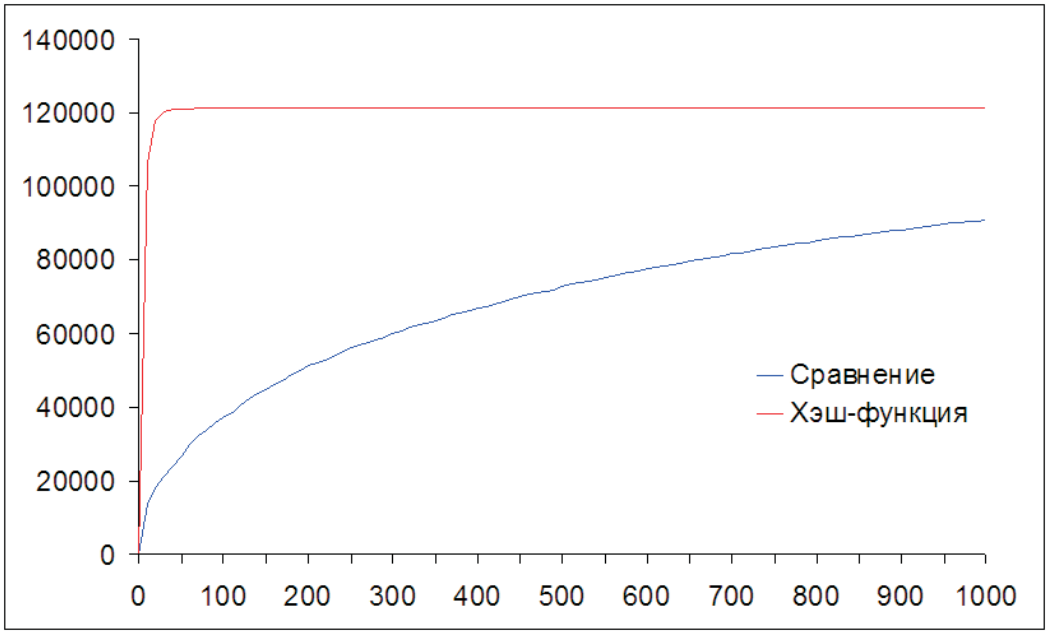

Рис. 1. Нахождение неповторяющихся вариантов разрешения дерева T_9

Авторы глубоко благодарны К. Горбунову за ценное обсуждение и помощь в подготовке статьи. Работа проведена при поддержке РФФИ, грант 05-04-49705.

СПИСОК ЛИТЕРАТУРЫ

1. Giribet G. Current advances in the phylogenetic reconstruction of metazoan evolution. A new paradigm for the Cambrian explosion? *Mol. Phylogenet. Evol.*, 2002, vol. 24, pp. 345–357.

2. Schmidt-Rhaesa A. Old trees, new trees — is there any progress? *Zoology*, 2003, vol. 106, pp. 291–301.

3. Dopazo H., Santoyo J., Dopazo J. Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryote model species. *Bioinformatics*, 2004, vol. 20, pp. i116–i121.

4. Blair J.E., Ikeo K., Gojobori T., Hedges S.B. The evolutionary position of nematodes. *BMC Evol. Biol.*, 2002, vol. 2, p. 7.

5. Wolf Y.I., Rogozin I.B., Koonin E.V. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome Res.*, 2004, vol. 14, pp. 29–36.

6. Philippe H., Lartillot N., Brinkmann H. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol. Biol. Evol.*, 2005, vol. 22, pp. 1246–1253.

7. Koonin E.V., Fedorova N.D., Jackson J.D., Jacobs A.R., Krylov D.M., Makarova K.S., Mazumder1 R., Mekhedov S.L., Nikolskaya A.N., Rao B.S., Rogozin I.B., Smirnov S., Sorokin A.V., Sverdlov A.V., Vasudevan S., Wolf Y.I., Yin J.J., Natale D.A. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.*, 2004, vol. 5, pp. R7.

8. Pruitt K.D., Tatusova T., Maglott D.R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, 2005, vol. 33, pp. D501–D504.

9. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. *J. Mol. Biol.*, 1990, vol. 215, pp. 403–410.

10. Tao T. Netblast (blastcl3). www.ncbi.nlm.nih.gov/blast/docs/netblast.html.

11. Parkinson J., Anthony A., Wasmuth J, Schmid R, Hedley A., Blaxter M. PartiGene: constructing partial genomes. *Bioinformatics*, 2004, vol. 20, pp. 1398–1404.

12. Parkinson J., Whitton C., Schmid R., Thomson M., Blaxter M. NEMBASE: a resource for parasitic nematode ESTs. *Nucleic Acids Res.*, 2004, vol. 32, pp. D427–D430.

13. TIGR Gene Indices Clustering Tools (TGICL). www.tigr.org/tdb/tgi/sofware.

14. Lottaz C., Iseli C., Jongeneel C.V., Bucher P. Modeling sequencing errors by combining Hidden Markov models. *Bioinformatics*, 2003, vol. 19, pp. 103–112.

15. Fukunishi Y., Hayashizaki Y. Amino acid translation program for full-length cDNA sequences with frameshift errors. *Physiol. Genomics*, 2001, vol. 5, pp. 81–87.

16. Tatusov R.L., Koonin E.V., Lipman D.J. A genomic perspective on protein families. *Science*, 1997, vol. 278, pp. 631–637.

17. Do C.B., Mahabhashyam M.S.P., Brudno M., Batzoglou S. PROBCONS: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, 2005, vol. 15, pp. 330–340.

18. Любецкий В.А., Горбунов К.Ю., Вьюгин В.В., Русин Л.Ю. Удаление шума в множественном выравнивании белковых последовательностей. *Информационные процессы*, 2005, том 5, №5, стр. 380–391.

19. Hillis D.M., Huelsenbeck J.P. Signal, noise, and reliability in molecular phylogenetic analyses. *J. Hered.*, 1992, vol. 83, pp. 189–195.

20. Edwards A.W.F. and Cavalli-Sforza L.L. Reconstruction of evolutionary trees. In: *Phenetic and Phylogenetic Classification* (Systematics Association Publication, No. 6), pp. 67–76, V.E. Heywood and J. McNeill (Editors). Systematics Association, London, 1964.