Санкт-Петербургский Государственный Политехнический Университет
Физико-Механический факультет
Кафедра Биофизики

А.В. ВАСИН

ВВЕДЕНИЕ В МОЛЕКУЛЯРНУЮ ЭВОЛЮЦИЮ

Санкт-Петербург
2010

Васин А.В. **Введение в молекулярную вирусологию**. 2010 г. 149 с.

Представлен курс лекций, прочитанный студентам 5 курса кафедры биофизики физико-механического факультета Санкт-Петербургского Государственного Политехнического Университета в 2003–2006 гг. Дисциплина «Введение в молекулярную эволюцию» входит в число дисциплин, завершающих магистерскую подготовку в рамках магистерской программы «Биофизика» и имеет существенное значение для успешной адаптации обучающегося на месте выполнения магистерской работы а также для дальнейшей научной или педагогической деятельности.

Целью изучения дисциплины «Введение в молекулярную эволюцию» является ознакомление студентов с современными методами исследования эволюции молекул ДНК (РНК) и белков. В курсе рассматриваются молекулярные механизмы, лежащие в основе эволюции генетических макромолекул, а также математические модели, описывающие процессы эволюции. В частности, описаны различные методы оценки эволюционной дистанции между нуклеотидными и аминокислотными последовательностями, синонимических и несинонимических замен, методы постороения филогенетических деревьев (дистанционные методы, методы максимальной парсимонии и максимального правдоподобия). Обсуждаются механизмы генетической вариабельности, нейтральной теории эволюции и дупликации генов.

Дисциплина направлена на увеличение научно-методического кругозора будущих исследователей, показывает примеры применения знаний из области математики и статистики для изучения эволюции биологических молекул

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Молекулярная эволюция – это раздел молекулярной биологии, занимающийся изучением закономерностей эволюционных изменений генетических макромолекул в живых организмах. В основе молекулярной эволюции лежит эволюционная теория – комплекс знаний об общих закономерностях и движущих силах исторического развития живой природы. Эволюционной теория, в свою очередь, базируется на утверждении о том, что все ныне существующие организмы произошли от ранее существовавших путем длительного их изменения под воздействием внешних и внутренних факторов. Основными задачами молекулярной эволюции являются изучение закономерностей эволюции генетических макромолекул, а также реконструкция эволюционной истории генов и организмов. При решении этих задач используются результаты исследований в других научных дисциплинах: палеонтологии, генетике, молекулярной биологии, биофизике, математике и информатике.

Развитие теории молекулярной эволюции тесно связано с развитием биологии. В конце XIX века Мендель впервые сформулировал понятие гена как единицы наследственности. В начале XX века концепция гена была развита в работах де Фриза, Вейсмана, Моргана и др. В 30-х годах XX в. в работах математиков Фишера, Райта и Холдейна были сформулированы основы популяционной генетики – науки о генетической структуре популяций.

К середине XX века были установлены структуры генетических макромолекул – ДНК и белков. Разработка методов секвенирования ДНК позволила проводить сравнение последовательностей ДНК разных видов и организмов и исследовать эволюционную изменчивость видов на молекулярном уровне, то есть на уровне ДНК и белков. Анализ этой изменчивости привел в 60-х годах XX века японского эволюциониста

М.Кимуру к формулированию нейтральной теории молекулярной эволюции Успехи экспериментальной молекулярной биологии к концу XX в. позволили решить задачу расшифровки последовательностей ДНК полных геномов ряда живых организмов. На современном этапе развития молекулярной биологии можно выделить несколько ключевых направлений развития биологии. Это секвенирование полных геномов большого количества (в том числе и высших) организмов, проведение их полномасштабного сравнительного анализа, развитие структурной биологии.

1. МОЛЕКУЛЯРНЫЕ ОСНОВЫ ЭВОЛЮЦИИ

1.1. Структура и функции генов

Основной причиной эволюции живых организмов являются постоянные мутационные изменения, происходящие в геноме. Мутации в гене или последовательности ДНК, вызванные нуклеотидными заменами, инсерциями/делециями, рекомбинацией, конверсией гена и т.д. могут распространиться по всей популяции посредством генетического дрейфа и/или естественного отбора и, в конечном счете, зафиксироваться в данном виде. Далее этот мутантный ген будет наследоваться всеми потомственными видами до тех пор, пока вновь не подвергнется мутации. Таким образом, если построить филогенетическое дерево для группы видов, можно определить линию видов, в которых посредством мутаций появилась любая специфическая черта.

Рассмотрим базовую структуру генов эукариот. С точки зрения выполняемых функций, гены могут быть классифицированы на две группы: **белок-кодирующие** гены и **РНК-кодирующие** гены. Белок-кодирующие гены транскрибируются в матричные РНК (**мРНК**[1]), которые в свою очередь транслируется в аминокислотные последовательности белка. Продуктами РНК-кодирующих генов являются транспортные РНК (**тРНК**[2]), рибосомальные РНК (**рРНК**[3]), малые ядерные РНК (**snРНК**[4]) и т.д. Эти немессенджерные РНК являются конечными продуктами РНК-кодирующих генов. Рибосомные РНК являются компонентами рибосом, ядра машинерии белкового синтеза, в то время как транспортные РНК важны для переноса генетической информации от мРНК к белку. Малые ядерные РНК связаны с

[1] messenger RNA
[2] transfer RNA
[3] ribosomal RNA
[4] small nuclear RNA

ядром, и некоторые из них вовлечены в сплайсинг интронов, а также в другие реакции процессирования РНК.

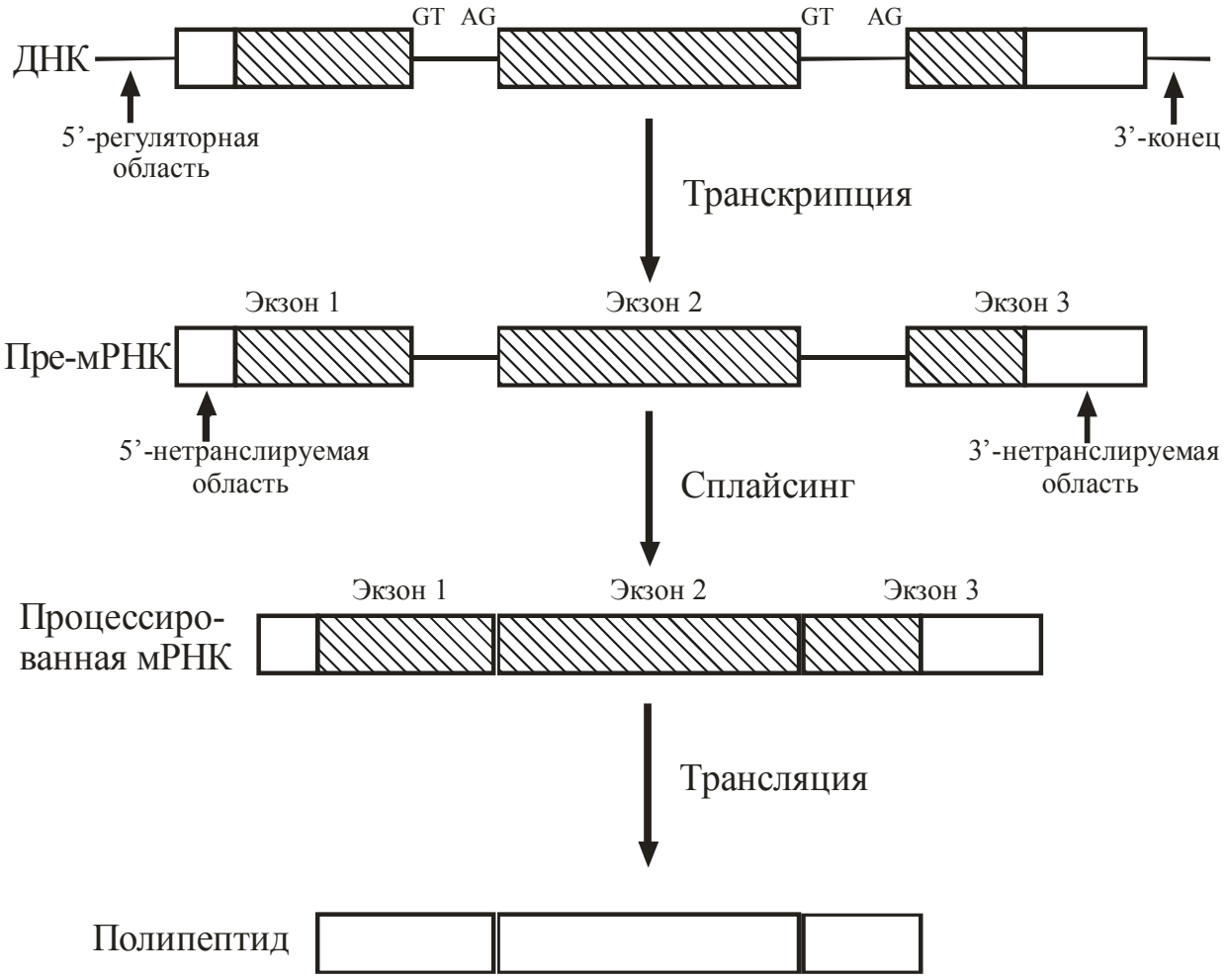

Рис. 1.1. Базовая структура белок-кодирующего гена эукариот.

Базовая структура белок-кодирующего гена эукариот представлена на рис. 1.1. Ген представляет собой линейную последовательность четырех нуклеотидов A (аденин), T (тимин), C (цитозин) и G (гуанин), и состоит из транскрибирующейся части ДНК и 5‘-регуляторной области, которая важна для контроля транскрипции и процессирования **пре-мРНК**. Пре-мРНК состоит из кодирующих и некодирующих областей. Кодирующие области содержат информацию, определяющую аминокислотную последовательности белка, продукта данного гена, в то время как

некодирующие области (5`-, 3`-UTR[5]) содержат информацию, необходимую для образования полипептида. Некоторые сегменты некодирующих областей выщепляются в процессе образования **зрелой мРНК[6]**. Эти сегменты называются **интронами**, а остающиеся области называются **экзонами** (рис. 1.1). Число экзонов в разных генов варьируется. Так, гены прокариот интронов не содержат, а некоторые гены эукариот (например, ген мышечной дистрофии) содержат в своем составе до 78 интронов. Обычно интрон начинается с динуклеотида GT и заканчивается динуклеотидом AG. Эти динуклеотиды обеспечивают корректный **сплайсинг** интронов. Генетическая информация, заложенная в нуклеотидной последовательности гена, переводится в информацию, заложенную в нуклеотидной последовательности РНК зрелой матричной РНК в процессе **транскрипции**. В свою очередь генетическая информация мРНК определяет аминокислотную последовательность белка, образующегося в процессе **трансляции**. Нуклеотиды мРНК считываются последовательно по три за раз, каждый такой триплет, или **кодон**, по правилам генетического кода соответствует одному аминокислотному остатку белка.

Генетический код для ядерных генов является универсальным для всех прокариотических и эукариотических организмов, за редким исключением. Тот же генетический код (**универсальный** или **стандартный генетический код**) используется и в генах хлоропластов, однако для генов митохондрий используется слегка измененный генетический код. Универсальный генетический код приведен в табл.1.1, аминокислотны представлены в трехбуквенных обозначениях. Существует $4^3 = 64$ возможных кодона для четырех нуклеотидов, урацила (U), цитозина (C), аденина (A) и гуанина (G). Три из них (UAA, UAG, UGA) являются **терминирующими** или **стоп кодонами** и не кодируют ни одну аминокислоту. Каждый из оставшегося 61 кодона (**sense кодоны**) кодирует определенную аминокислоту, но так как

[5] 5`-, 3`-untranslated regions
[6] mature mRNA

существует только 20 аминокислот, используемых при синтезе белков, то многим кодонам соответствует более чем одна аминокислота. Это свойство называется **вырожденностью генетического кода**. Кодоны, кодирующие одну и ту же аминокислоту, называются **синонимическими кодонами**. В таблице генетического кода, триплет AUG кодирует метионин, но этот же кодон также используется как **инициирующий кодон**. Метионин, кодируемый инициирующим кодоном, представлен в модифицированной форме, и позже удаляется из полипептида. CUG и UUG также используются в качестве инициирующих кодонов в некоторых ядерных генах. Эти инициирующие кодоны должны быть исключены из исследования эволюции последовательностей ДНК, так как они в большинстве случаев не изменяются. Терминирующие кодоны также должны быть исключены из рассмотрения.

Табл. 1.1. Универсальный генетический код

Кодон		Кодон		Кодон		Кодон	
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

В табл. 1.2 показан генетический код для митохондриальных генов позвоночных. Существует несколько различий между этим генетическим кодом и стандартным ядерным генетическим кодом. В митохондриальном генетическом коде кодон UGA является не терминирующим, а кодоном для триптофана, а кодоны AGA и AGG являются терминирующими, а не кодонами для аргинина. AUA, кодирующий изолейцин в ядерном коде, в митохондриальном коде используется для метионина. Примеры отклонений от универсального генетического кода приведены в табл. 1.3.

Табл. 1.2. Генетический код для митохондриальных генов позвоночных

Кодон		Кодон		Кодон		Кодон	
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Trp
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Met	ACA	Thr	AAA	Lys	AGA	Ter
AUG	Met	ACG	Thr	AA G	Lys	AGG	Ter
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

В митохондриальных генах растений, кодон CGG транслируется в триптофан не напрямую: нуклеотид C в этом кодоне превращается в U после образования мРНК, и этот измененный кодон UGG кодирует триптофан, как и в универсальном генетическом коде. Такой процесс называется **редактированием РНК**. При сравнении нуклеотидных и аминокислотных последовательностей разных видов растений следует иметь в виду, что кодон

CGG рассматривается как триптофановый. Редактирование РНК происходит в некоторых митохондриальных генах других эукариотических царств, поэтому при транслировании последовательностей ДНК в аминокислотные последовательности нужно учитывать эти особенности.

Табл. 1.3. Примеры отклонений от универсального генетического кода

	Кодоны						
	UGA	AUA	AAA	AGR	CUN	CGG	UAR
Стандартный генетический код	Ter	Ile	Lys	Arg	Leu	Arg	Ter
Митохондриальный код							
Позвоночные	Trp	Met	-	Ter	-	-	-
Асцидии	Trp	Met	-	Gly	-	-	-
Иглокожие	Trp	-	Asn	Ser	-	-	-
Drosophila	Trp	Met	-	Ser	-	-	-
Дрожжи	Trp	Met	-	-	Thr	-	-
Простейшие	Trp	-	-	-	-	-	-
Плесень	Trp	-	-	-	-	-	-
Кишечнополостные	Trp	-	-	-	-	-	-
Ядерный код							
Tetrahymena	-	-	-	-	-	-	Gln
Mycoplasta	Trp	-	-	-	-	-	-
Euplotid	Cys	-	-	-	-	-	-

1.2. Мутационные изменения последовательностей ДНК.

Так как все морфологические и физиологические черты организма определяются генетической информацией, заложенной в ДНК, любые их изменения имеют в основе некоторые мутации в молекуле ДНК. Существует 4 основных типа изменений в ДНК:

1. **Замена** одного нуклеотида на другой (рис. 1.2.А)
2. **Делеция** нуклеотида (рис. 1.2.Б)
3. **Инсерция** нуклеотида (рис. 1.2.В)
4. **Инверсия** нуклеотидов (рис. 1.2.Г)

Инсерции, делеции и инверсии могут происходить как с одним нуклеотидом, так и с несколькими сразу. Если инсерции и делеции происходят в белок-кодирующем гене, то они могут привести к сдвигу рамки считывания нуклеотидной последовательности. Такие инсерции и делеции называются мутациями, вызывающими **сдвиг рамки считывания**[7].

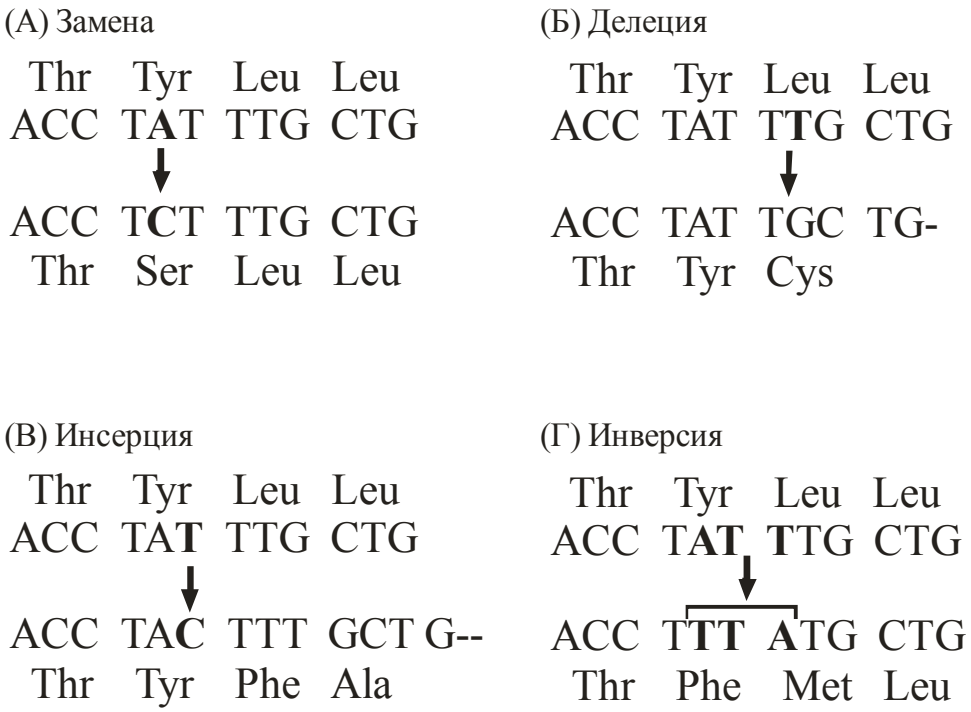

Рис. 1.2. Основные типы мутационных изменений в нуклеотидных последовательностях

Нуклеотидные замены могут быть разделены на два класса: **транзиции** и **трансверсии**. Транзиция – это замена **пурина** (аденин и гуанин) на другой пурин или замена **пиримидина** (тимин или цитозин) на другой пиримидин (рис. 1.3). Остальные типы замен называются трансверсиями. В большинстве сегментов ДНК нуклеотидные замены по типу транзиции происходят чаще, чем замены по типу трансверсии. В случае белок-кодирующих генов нуклеотидные замены в кодоне, не приводящие к замене соответствующей аминокислоты, называются **синонимическими,** или молчащими заменами, в

[7] frameshift mutations

то время как замены в кодоне, приводящие к замене соответствующей аминокислоты, называются **несинонимическими,** или приводящими к изменению аминокислоты заменами. Кроме того, существуют **нонсенс** (nonsense) мутации, то есть мутации, приводящие к образованию стоп-кодона. Синонимические замены происходят в основном в третьем, а также иногда в первом положениях кодона (4% замен первого и 70% замен третьего нуклеотидов в кодоне являются синонимическими). Все нуклеотидные замены во втором положении кодона являются несинонимическими, либо нонсенс мутациями. Если допустить, что все кодоны встречаются в геноме с равной вероятностью, и вероятность замен одинакова для всех пар нуклеотидов, то соотношение синонимических, несинонимических и нонсенс мутаций было бы равно 25, 71 и 4%, соответственно. Хотя на практике такое допущение является мало реалистичным, это процентное соотношение дает примерную картину относительных частот различных мутаций на уровне нуклеотидов.

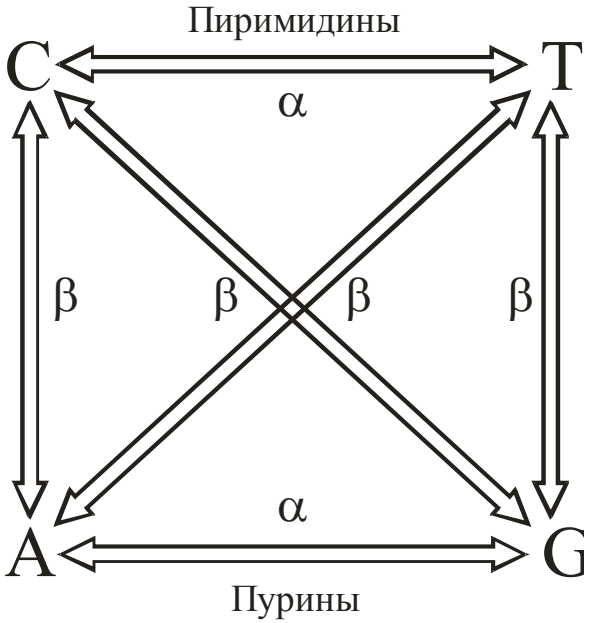

Рис. 1.3. Нуклеотидные замены по типу транзиций и трансверсий.
α и β – скорости транзиций и трансверсий, соответственно

Инсерции и делеции встречаются довольно часто, особенно в некодирующих областях ДНК. Число нуклеотидов, вовлеченных в делеции и

инсерции, варьируется от одного нуклеотида до целых блоков ДНК. Короткие делеции и инсерции в основном вызваны ошибками репликации ДНК. Появление длинных делеций и инсерций в основном вызвано неравным кроссинговером или транспозициями ДНК. **Транспозиция ДНК** – это передвижение сегментов ДНК из одной области хромосомы в другую посредством транспозонов или транспозирующих элементов. Другой возможный механизм инсерции гена – это горизонтальный перенос генов между видами, вызванный транспозирующими элементами. Неравный кроссинговер играет важную роль в эволюции, так как вызывает уменьшение или увеличение содержания ДНК. Возможная роль неравного кроссинговера в увеличении числа генов в геноме была предложена еще в 30-х годах XX века, однако только после начала молекулярных исследований ДНК была осознана важность этого механизма в процессе увеличения или уменьшения содержания ДНК в процессе эволюции. Генетическим событием, связанным с неравным кроссинговером, является конверсия гена. Конверсия гена – это изменение сегмента ДНК, которое делает этот сегмент идентичным другому сегменту ДНК. Это событие, видимо, происходит путем репарации несовпадающих оснований в гетеродуплексной ДНК и способно гомогенизировать членов мультигенного семейства, но оно не меняет числа копий гена.

1.3. Использование кодонов.

Если бы нуклеотидные замены происходили с одинаковой вероятностью для каждого нуклеотидного сайта, то каждый нуклеотидный сайт должен был бы иметь один из четырех нуклеотидов A, T, C и G с равной вероятностью. Поэтому, если нет давления отбора и мутационного смещения[8], можно ожидать, что кодоны, кодирующие одну и ту же аминокислоту, в среднем будут иметь одинаковые частоты в белок-

[8] mutational bias

кодирующих областях ДНК. Например, аминокислота валин (Val) кодируется четырьмя кодонами: GUU, GUC, GUA и GUG. Так, если рассмотреть большое число кодонов для валина в гене, то относительные частоты встречаемости GUU, GUC, GUA и GUG должны быть равны примерно 25%.

На практике частоты разных кодонов для одной и той же аминокислоты обычно отличаются, и одни кодоны используются чаще, чем другие. На рис. 1.4 показаны частоты использования кодонов в РНК-полимеразе бактерии *Escherichia coli*. Для валина все четыре кодона используются примерно одинаково, хотя GUU и используется в два раза чаще, чем GUC. Для аргинина же преимущественно используются кодоны CGU и CGC, а кодоны CGA, CGG, AGA и AGG остаются почти не задействованными. Такой тип **смещения в использовании кодонов**[9] наблюдается как в прокариотических, так и в эукариотических генах.

Phe	UUU	15	(0.51)	Ser	UCU	32	(1.86)	Tyr	UAU	18	(0.64)
	UUC	44	(1.49)		UCC	38	(2.21)		*UAC*	38	(1.36)
Leu	UUA	2	(0.07)		UCA	2	(0.12)	Stop	UAA		
	UUG	8	(0.27)		UCG	5	(0.29)	Stop	UAG		
Leu	CUU	11	(0.36)	Pro	CCU	9	(0.48)	His	CAU	5	(0.36)
	CUC	18	(0.60)		CCC	0	(0.00)		CAC	23	(1.64)
	CUA	1	(0.03)		CCA	11	(0.59)	Gln	CAA	15	(0.34)
	CUG	141	(4.67)		*CCG*	55	(2.93)		*CAG*	73	(1.66)
Ile	AUU	29	(0.69)	Thr	ACU	19	(0.78)	Asn	AAU	4	(0.11)
	AUC	98	(2.31)		ACC	63	(2.57)		AAC	66	(1.89)
	AUA	0	(0.00)		ACA	3	(0.12)	Lys	AAA	77	(1.35)
Met	AUG	60	(1.00)		ACG	13	(0.53)		AAG	37	(0.65)
Val	*GUU*	55	(1.53)	Ala	GCU	30	(0.94)	Asp	GAU	60	(0.83)
	GUC	21	(0.58)		GCC	19	(0.59)		GAC	85	(1.17)
	GUA	34	(0.94)		GCA	30	(0.94)	Glu	GAA	147	(1.52)
	GUG	34	(0.94)		GCG	49	(1.53)		GAG	46	(0.48)

Cys	UGU	5	(1.00)	Ser	AGU	3	(0.17)	Gly	GGU	78	(2.40)
	UGC	5	(1.00)		AGC	23	(1.34)		GGC	47	(1.45)
Stop	UGA			Arg	AGA	0	(0.00)		GGA	0	(0.00)
Trp	UGG	8	(1.00)		AGG	0	(0.00)		GGG	5	(0.15)
Arg	CGU	89	(3.93)								
	CGC	46	(2.03)								
	CGA	1	(0.04)								
	CGG	0	(0.00)								

Рис. 1.4. Частоты использования кодонов в генах РНК-полимеразы (rpoB и rpoD) Escherichia coli *(Ikemura, 1985).*
*Оптимальные для трансляции кодоны выделены курсивом. Относительные частоты использования синонимических кодонов **RSCU**, рассчитанные по формуле 1.1 приведены в скобках.*

[9] codon usage bias

Что же вызывает смещение в использовании кодонов? Точного объяснения этому явлению нет, однако, можно выделить ряд причин, которые могут играть роль в этом процессе.

Во-первых, было замечено, что частота использования кодона в активно экспрессирующихся генах коррелирует с относительным содержанием соответствующих изоакцепторных тРНК в клетке. Таким образом, можно предположить, что появляющиеся в результате мутаций кодоны, не соответствующие превалирующим в клетке тРНК, подвергаются **очищающему отбору**[10] в активно экспрессирующихся генах, так как они не эффективны при синтезе белка. В «неактивно» экспрессирующихся генах давление отбора, видимо, так мало, что эффективно могут использоваться все кодоны. Такой отбор, действительно, наблюдается во многих одноклеточных организмах, и даже в *Drosophila melanogaster*, однако данное правило не применимо для генов человека.

Во-вторых, хотя относительная совокупность изоакцепторных тРНК является важным фактором, существует еще один фактор, влияющий на использование кодонов: **смещенное мутационное давление** (biased mutation pressure). В бактериях относительная частота нуклеотидов G и C (**GC-содержание**) в геноме варьируется от 25 до 75%, и это колебание, как считают, происходит из-за различий между скоростями прямых и обратных мутаций GC и AT пар в нуклеотидных последовательностях. В некоторых видах бактерий (например, *Mycoplasma capricolum*) мутационное давление от GC к AT парам настолько велико, что нуклеотиды в молчащем третьем положении кодона почти всегда являются A или T. В некоторых других бактериях (например, *Micrococcus luteus*), мутационное давление действует в обратном направлении (AT→GC), так что наиболее часто используемыми нуклеотидами в третьем положении являются G или C (рис. 1.5).

Конечно, для поддержания функции белка, GC-содержание даже в третьем положении кодона будет отличаться от равновесной частоты,

[10] purifying selection

обусловленной только мутационным давлением, так как некоторые нуклеотидные замены в третьем положении все же приводят к замене аминокислоты и, таким образом, они подвергаются действию очищающего отбора. Все замены во втором положении кодона являются несинонимическими, поэтому они в первую очередь контролируются функциональными ограничениями, а не мутационным давлением. В первом положении кодона небольшая часть замен является синонимической, поэтому влияние мутационного давления должно занимать промежуточное значение между влияниями на первое и третье положение в кодоне.

На рис. 1.5 показана зависимость между GC-содержанием в первом, втором и третьем положениях кодонов в генах и общим содержанием GC пар по всему геному для 11 различных видов бактерий, GC-содержание в которых варьируется в широких пределах. В третьем положении кодона содержание GC пар в генах примерно равно содержанию GC пар в геноме, что говорит о сильном влиянии мутационного давления. Во втором положении, однако, наклон линейной функции, описывающей эту зависимость от геномного содержания GC пар, гораздо ниже, чем в третьем положении кодона. Это позволяет предположить, что влияние мутационного давления менее важно во втором положении кодона, где содержание GC пар во многом определяется очищающим отбором, возникающим из-за функциональных ограничений, накладываемых на ген, как упоминалось ранее. Наклон графика зависимости для первого положения, как и ожидалось, занимает промежуточное положение, что поддерживает предположение о контроле за использованием кодонов как мутационным давлением, так и очищающим отбором

Тот факт, что содержание GC пар сильно варьируется среди разных видов бактерий указывает на то, что паттерн нуклеотидных замен не одинаков для разных групп бактерий. Этот факт вносит определенные сложности в изучение филогенетических отношений между этими организмами. Различные группы рассмотренных здесь бактерий, по всей

видимости, дивергировали более миллиарда лет назад, поэтому может показаться, что при изучении эволюции высших организмов эта проблема не столь важна, однако, есть основания полагать, что паттерн нуклеотидных замен изменяется даже в гораздо меньшие промежутки эволюционного времени.

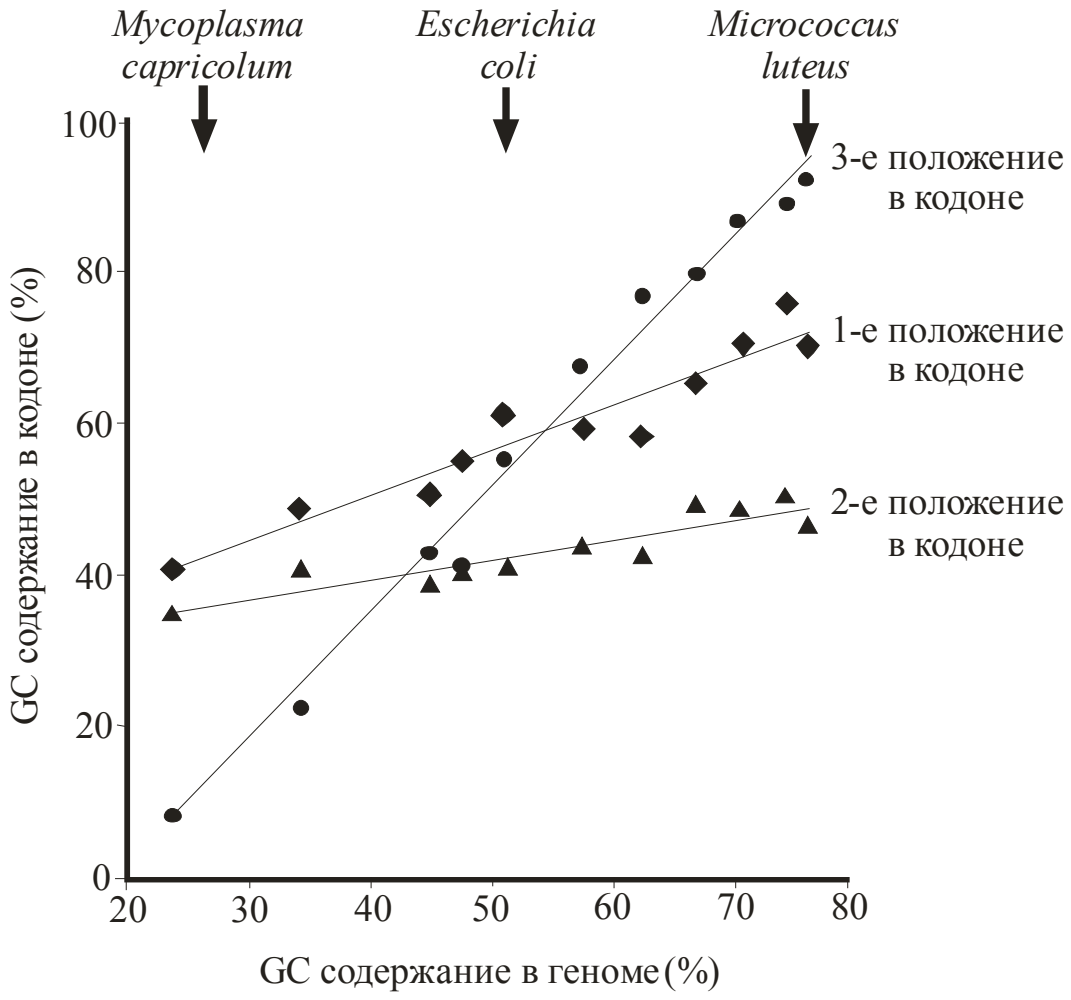

Рис. 1.5. Зависимость GC-содержания в целом геноме от GC-содержания в первом, втором и третьем положениях кодона для генов из 11 видов бактерий.

В отличие от одноклеточных организмов, у растений и животных наблюдается небольшое колебание содержания GC пар в пределах всего генома. В частности, GC-содержание в геномах позвоночных составляет 40-45%. Однако, смещение в использовании кодонов наблюдается и во многих

генах высших организмов. У некоторых беспозвоночных, например, у дрозофилы, смещение довольно сильное, и оно, вероятно, вызвано относительным содержанием изоакцепторных тРНК, как и в случае микроорганизмов.

У позвоночных этот вопрос еще более запутан, потому что генная экспрессия является тканеспецифической, и геном является гетерогенным с точки зрения содержания GC пар. Было показано, что геномы позвоночных мозаично построены из GC-богатых и GC-обедненных областей, и содержание GC пар в GC-богатых областях равно примерно 60%, а в GC-обедненных областях – примерно 30%. Такие GC-богатые и GC-обедненные области могут иметь длину до 300 т.п.н. и содержать функциональные гены. Эти GC-богатые и GC-обедненные области называются **изохорами**. Интересно, что содержание GC пар в третьем положении кодона генов внутри изохоры обычно близко к содержанию GC пар в целой изохоре. У теплокровных позвоночных, таких как млекопитающие и птицы, существует 4 основные группы изохор (две GC-богатые и две GC-обедненные изохоры), а у хладнокровных позвоночных GC-богатые изохоры встречаются редко или отсутствуют вовсе. Граница между GC-богатыми и две GC-обедненными изохорами довольно узкая.

Происхождение изохор у позвоночных является дискуссионным вопросом, однозначного ответа на который в настоящее время нет. Однако, важно отметить, что гены, расположенные в разных изохорах, видимо, обладают разными паттернами смещения в использовании кодонов, а так как смещение в использовании кодонов влияет на скорость нуклеотидных замен, то эти гены будут эволюционировать с разными скоростями.

1.4. Статистическая мера смещения использования кодонов.

Так как абсолютные частоты использования кодонов не удобны при сравнении смещений у разных генов или организмов из-за разницы в общем

числе исследуемых кодонов, то вводят меру, называемую **относительным использованием синонимических кодонов**[11] (***RSCU***). Она определяется как

$$\boldsymbol{RSCU} = \frac{\boldsymbol{X}_i}{\overline{\boldsymbol{X}}} \qquad (1.1)$$

где $\boldsymbol{X}_i$ – это наблюдаемая частота *i*-го кодона для определенной аминокислоты, а $\overline{\boldsymbol{X}}$ – это среднее $\boldsymbol{X}_i$ по всем кодонам, то есть $\overline{\boldsymbol{X}} = \sum_i \boldsymbol{X}_i / \boldsymbol{m}$, где $\boldsymbol{m}$ - это полное число кодонов для данной аминокислоты. На рис. 1.4 приведены значения ***RSCU*** для каждого кодона генов РНК-полимеразы (rpoB и rpoD) *E. coli*, рассчитанные по формуле 1.1.

[11] relative synonymous codon usage

2. ЭВОЛЮЦИОННЫЕ ИЗМЕНЕНИЯ АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ.

До изобретения быстрых методов секвенирования ДНК по Максаму и Гилберту и по Сенгеру в 1977 году, большинство исследований по молекулярной эволюции проводилось с использованием данных по аминокислотным последовательностям. В настоящее время секвенировать молекулы ДНК гораздо проще, чем белки, поэтому аминокислотные последовательности вычисляют эмперически из нуклеотидных по правилам генетического кода. В тоже время, в некоторых случаях аминокислотные последовательности являются чрезвычайно полезными при эволюционных исследованиях, так как они обладают большей степенью консервативности, чем нуклеотидные последовательности, и, следовательно, предоставляют достоверную информацию по долгосрочной эволюции генов или видов. Также белковые последовательности играют важную роль при выравнивании последовательностей ДНК кодирующих их генов. Кроме того, математические модели оценки эволюционных изменений аминокислотных последовательностей гораздо проще, чем для нуклеотидных последовательностей. Поэтому сначала мы будем рассматривать эволюционные изменения аминокислотных последовательностей.

Основной целью этой главы является описание статистических моделей для измерения **эволюционной дистанции** между двумя аминокислотными последовательностями. Эволюционная дистанция является фундаментальным понятием молекулярной эволюции и используется при построении филогенетических деревьев и оценке времени дивергенции. В случае аминокислотных последовательностей дистанция обычно измеряется как число аминокислотных замен, но на практике применяется несколько более сложных способов ее вычисления, определяемых используемыми допущениями.

2.1. Различия в аминокислотах и соотношение различающихся аминокислот.

Анализ эволюционных изменений в белках и полипептидах начинается со сравнения двух или более аминокислотных последовательностей из разных организмов. На рис. 2.1 показаны аминокислотные последовательности α-цепей гемоглобина человека, лошади, коровы, кенгуру, тритона и карпа. Все аминокислоты представлены в однобуквенных обозначениях. Существует несколько подходов к измерению степени эволюционной дивергенции между этими последовательностями.

```
Человек   V-LSPADKTN  VKAAWGKVGA  HAGEYGAEAL  ERMFLSFPTT   KTYFPHF-DL  SHGSAQVKGH   60
Лошадь    .-..A.....  .....S...G  ..........  .....G....  .......-..  ........A.
Корова    .-..A...G.  .........G  ..A.......  ..........  .......-..  ........A.
Кенгуру   .-..A...GH  ...I.....G  .....A..G.  ..T.H.....  .......-..  ......IQA.
Тритон    MK..AE..H.  ..TT.DHIKG  .EEAL.....  F...T.L.A.  R....AK-..  .E..SFLHS.
Карп      S-..DK..AA  ..I..A.ISP  K.DDI.....  G..LTVY.Q.  ....A.WA..  .P..GP..-.

Человек   GKKVA-DALT  NAVAHVDDMP  NALSALSDLH  AHKLRVDPVN  FKLLSHCLLV  TLAAHLPAEF  120
Лошадь    .....-.G..  L..G.L..L.  G...D..N..  ..........  .........S  ...V...ND.
Корова    .A...-A...  K..E.L..L.  G...E.....  ..........  ......S...  ...S...SD.
Кенгуру   ...I.-...G  Q..E.I..L.  GT..K.....  ..........  ..........  .F....GDA.
Тритон    ....M-G..S  .....I..ID  A..CK...K.  .QD.M...A.  .PK.A.NI..  VMGI..K.HL
Карп      ....IMG.VG  D..SKI..LV  GG.AS..E..  .S......A.  ..I.ANHIV.  GIMFY..GD.

Человек   TPAVHASLDK  FLASVSTVLT  SKYR  114
Лошадь    ..........  ..S.......  ....
Корова    ..........  ...N......  ....
Кенгуру   ..E.......  ...A......  ....
Тритон    .YP..C.V..  ..DV.GH...  ....
Карп      P.E..M.V..  .FQNLALA.S  E...
```

Рис. 2.1. Аминокислотные последовательности α-цепей гемоглобина 6 видов позвоночных

(-) соответствует положениям инсерций или делеций, (.) соответствует идентичности аминокислоты последовательности человека.

Простейшей мерой оценки дивергенции белков является число аминокислотных различий n_d между двумя последовательностями. Если число аминокислот n одинаково во всех сравниваемых последовательностях, то можно использовать n_d для попарного сравнения степеней дивергенции между ними. На практике, сравниваемые аминокислотные

последовательности часто содержат **инсерции** и **делеции** (инделы[12]). В этом случае все индылы, или гэпы[13], должны быть элиминированы перед расчетом n_d. В противном случае сравнение n_d между разными парами последовательностей будет неоднозначным.

Таблица 2.1. Число аминокислотных отличий (выше диагонали) и относительное число аминокислотных отличий (ниже диагонали) между α-цепями гемоглобина разных видов позвоночных.

	Человек	Лошадь	Корова	Кенгуру	Тритон	Карп
Человек		17	17	26	61	68
Лошадь	0.121		17	29	66	67
Корова	0.121	0.121		25	63	65
Кенгуру	0.186	0.207	0.179		66	71
Тритон	0.436	0.471	0.450	0.471		74
Карп	0.486	0.479	0.464	0.507	0.529	

Примечание: Делеции и инсерции были исключены при расчете, полное число исследованных аминокислотных остатков в каждой последовательности равно 140.

Более удобной мерой оценки степени эволюционной дивергенции между белками является доля аминокислотных различий между двумя последовательностями. Это соотношение $\hat{p}$ может быть использовано для сравнения степеней дивергенции последовательностей даже при разных n. Оно выражается как

$$\hat{p} = \frac{n_d}{n} \qquad (2.1)$$

Это соотношение иногда называют ***p*-дистанцией**. Если все аминокислотные сайты подвергаются заменам с равной вероятностью, то n_d будет подчиняться биноминальному распределению. Поэтому дисперсия $\hat{p}$ рассчитывается по формуле

[12] indels (образовано из первых частей слов *in*sertion и *del*etion)
[13] gap

$$V(\hat{p}) = \frac{p(1-p)}{n} \qquad (2.2)$$

В фактических расчетах дисперсии ***p***-дистанции ***p*** заменяется на $\hat{p}$.

В примере, приведенном на рис. 2.1, полное число аминокислотных сайтов после элиминирования всех гэпов равно 140, то есть, в данном случае, $n = 140$. Значения n_d представлены в табл. 2.1 выше диагонали, по ним можно легко сосчитать значения $\hat{p}$ (расположены ниже диагонали). Из таблицы видно, что $\hat{p}$ имеет большее значение у более удаленных видов (например, человек и карп), чем у менее удаленных (например, человек и лошадь). Можно предположить, что число аминокислотных замен возрастает с увеличением времени прошедшего после дивергенции двух видов. Однако, как будет показано далее, ***p*** не строго пропорционально времени дивергенции ***t*** (рис. 2.2).

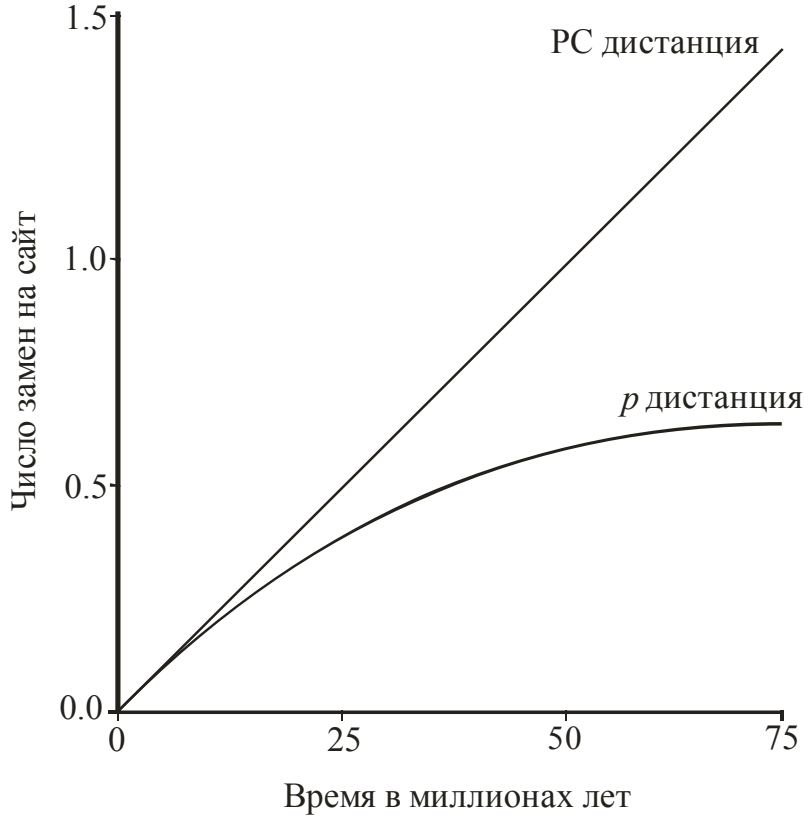

Рис. 2.2. Зависимость значений p-дистанции и дистанции с учетом коррекции Пуассона (PC) от времени

*Скорость аминокислотных замен **r** полагалась равной 10^{-8} на сайт в год.*

2.2. Коррекция Пуассона и гамма-дистанции

Одной из причин нелинейной зависимости p от t является тот факт, что в одних и тех же сайтах может многократно происходить несколько аминокислотных замен, и несоответствие между наблюдаемым числом n_d и действительным числом аминокислотных замен будет постепенно увеличиваться с течением времени. Одним из путей более аккуратной оценки числа замен является использование распределения Пуассона. Пусть r будет скоростью аминокислотных замен в год в определенном аминокислотном сайте. Предположим для простоты, что она одинакова для всех сайтов. Это допущение редко применимо в реальности, однако, ошибка, получаемая при нем мала до тех пор, пока величина p мала. Тогда среднее число аминокислотных замен на сайт в течение периода из t лет будет равно rt, и вероятность появления k аминокислотных замен в данном сайте (k = 0, 1, 2, 3,…) будет задаваться распределением Пуассона

$$P(k;t) = \frac{e^{-rt}(rt)^k}{k!} \qquad (2.3)$$

Вероятность того, что в данном аминокислотном сайте не произойдет ни одной замены, равна

$$P(0;t) = \frac{e^{-rt}(rt)^0}{0!} = e^{-rt} \qquad (2.4)$$

Если число аминокислот в полипептиде равно n, то предполагаемое число неизмененных аминокислот будет равно ne^{-rt}.

В реальности мы не знаем аминокислотных последовательностей предковых видов, поэтому уравнение 2.3 неприменимо. По этой причине число аминокислотных замен оценивается исходя из сравнения двух гомологичных последовательностей, дивергировавших t лет назад. Так как вероятность того, что за t лет в сайте не произошло ни одной аминокислотной замены, равна e^{-rt}, то вероятность q того, что ни один из гомологичных сайтов двух последовательностей не претерпел замены, равна

$$q = (e^{-rt})^2 = e^{-2rt} \qquad (2.5)$$

Эта вероятность может быть вычислена как $1 - \hat{p}$, так как $q = 1 - p$. Уравнение $q = e^{-2rt}$ является приближенным, так как в нем не учитываются обратные и параллельные мутации, то есть одинаковые мутации, происходящие в гомологичных аминокислотных сайтах двух разных эволюционных линий. Однако, при больших $\hat{p}$ (скажем, при $\hat{p} > 0.3$) влияние этих мутаций в основном очень мало.

Если использовать уравнение 2.5, то полное число аминокислотных замен на сайт для двух последовательностей ($d = 2rt$) задается уравнением

$$d = -\ln(1 - p) \qquad (2.6)$$

Оценка $\hat{d}$ значения d может быть получена заменой p на $\hat{p}$, и дисперсия большой выборки[14] $\hat{d}$ дается выражением

$$V(\hat{d}) = \frac{p}{(1-p)n} \qquad (2.7)$$

В фактических вычислениях $V(\hat{d})$ p опять заменяется на $\hat{p}$. Это справедливо для всех дисперсионных формул для $\hat{d}$ или других оценок, приведенных в этой книге. В дальнейшем, мы будем называть $\hat{d}$ дистанцией с поправкой Пуассона (**PC- дистанция**[15]).

При изучении молекулярной эволюции часто важно знать скорость аминокислотных замен r. Она может быть оценена как

$$\hat{r} = \frac{\hat{d}}{2t} \qquad (2.8)$$

если мы знаем время дивергенции двух последовательностей из других источников биологической информации. Заметим, что $\hat{d}$ делится на $2t$, а не на t, потому что скорость замен соответствует одной эволюционной линии. Дисперсия $\hat{r}$ вычисляется как $\frac{V(\hat{d})}{(2t)^2}$. С другой стороны, если мы знаем

[14] large-sample variance
[15] Poisson correction distance

скорость r из предыдущих исследований, но не знаем эволюционное время, то t может быть оценено, как

$$\hat{t} = \frac{\hat{d}}{2r} \tag{2.9}$$

с дисперсией $V(\hat{d})/(2r)^2$.

В вышеприведенных рассуждениях, мы предполагали, что скорость аминокислотных замен одинакова для всех аминокислотных сайтов. Это допущение обычно не выполняется для реальных данных, так как скорость замен обычно выше в функционально менее значимых сайтах, чем в функционально более значимых сайтах. В действительности, было показано, что распределение числа аминокислотных замен на сайт k имеет дисперсию большую, чем дисперсия Пуассона, и что число замен приблизительно подчиняется отрицательному биноминальному распределению. Известно, что когда скорость аминокислотных замен r варьируется от сайта к сайту в соответствии с гамма-распределением, наблюдаемое число замен на сайт k будет распределяться как отрицательное биноминальное распределение. Следовательно, скорость замен варьируется от сайта к сайту в соответствии с гамма-распределением, которое задается формулой

$$f(r) = \frac{b^a}{\Gamma(a)} e^{-br} r^{a-1} \tag{2.10}$$

где $a = \frac{\bar{r}^2}{V(r)}$ и $b = \frac{\bar{r}}{V(r)}$, а $\bar{r}$ и $V(r)$ – это среднее и дисперсия r, соответственно. Здесь $\Gamma(a)$ – это гамма функция, определяемая, как

$$\Gamma(a) = \int_0^\infty e^{-t} t^{a-1} dt \tag{2.11}$$

Форма распределения $f(r)$ определяется значением параметра a, который часто называют образом[16] или **гамма-параметром**, в то время как b – это коэффициент масштабирования[17].

[16] shape
[17] scaling factor

Гамма распределение является очень гибким, и оно может принимать разные формы в зависимости от значения гамма параметра ***a*** (рис. 2.3). Когда ***a*** = ∞, ***r*** одинакова для всех сайтов. Когда ***a*** = 1, ***r*** следует экспоненциальному распределению, что указывает на то, что ***r*** значительно изменяется от одного аминокислотного сайта к другому. Когда ***a*** < 1, распределение ***r*** еще более ассиметрично, и существенная доля сайтов имеет значение ***r*** близкое к нулю, то есть сайты остаются практически неизменными.

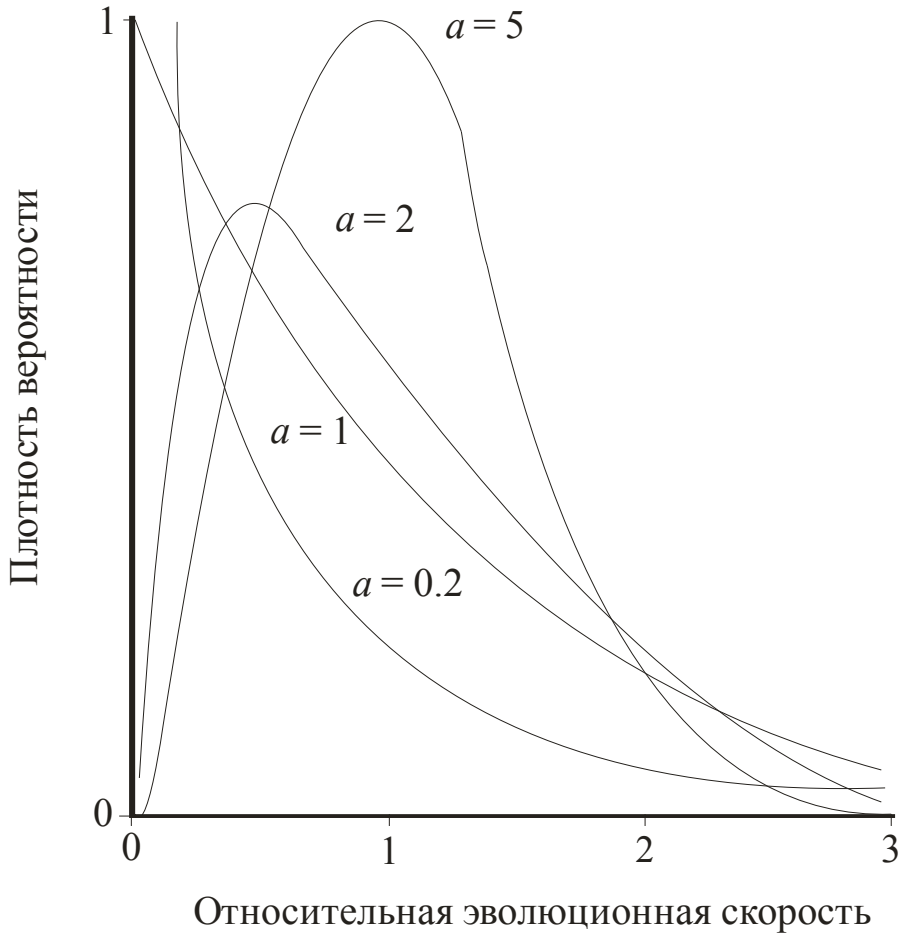

Рис. 2.3. Вид гамма распределений скоростей замен по сайтам при разных значениях гамма параметров a.

Используя анализ методом парсимонии[18] было оценино, что для последовательностей цитохрома *c* позвоночных ***a*** = 2. Это указывает на то,

[18] parsimony

что *r* значительно изменяется в зависимости от аминокислотного сайта (рис. 2.3). Оценка значений ***a*** для 51 ядерного и 13 митохондриальных белков из разных видов позвоночных показала, что ***a*** колеблется в пределах от 0.2 до 3.5. Это указывает на то, что в некоторых генах вариация ***r*** среди сайтов очень велика.

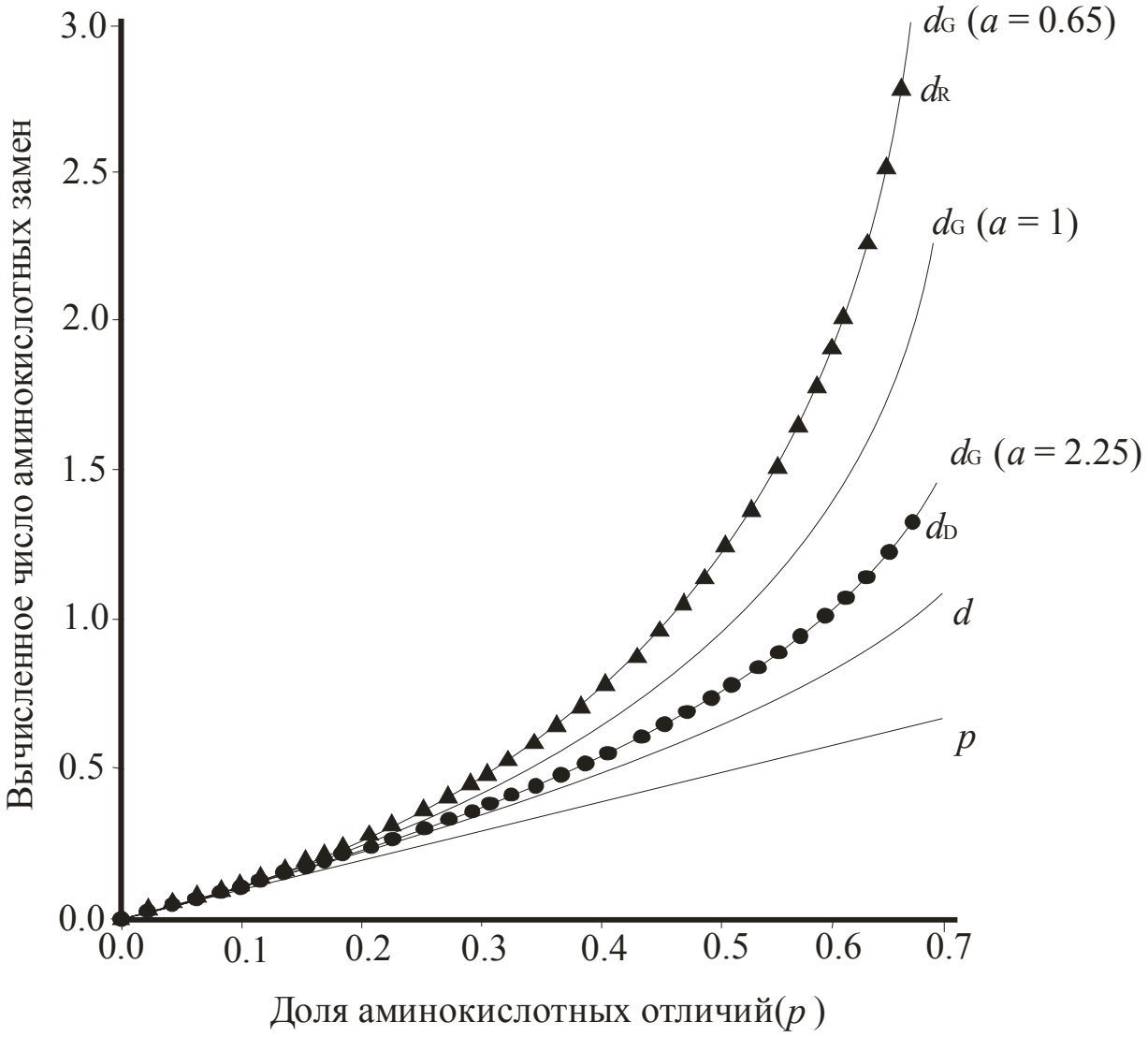

Рис. 2.4. Зависимости различных эволюционных дистанций от доли аминокислотных отличий p.
Точками обозначены значения дистанции Дэйхофф $\boldsymbol{d_D}$*, треугольники соответствуют дистанции Гришина* $\boldsymbol{d_R}$.

Когда ***r*** изменяется в соответствии с гамма-распределением, можно оценить число аминокислотных замен на сайт. Для этого рассматриваем

вероятность идентичности аминокислот в данном сайте между двумя последовательностями в момент времени t, которая задается уравнением 2.4. Среднее значение q по всем сайтам вычисляется как

$$\bar{q} = \int_0^{\infty} qf(r)dr = \left(\frac{a}{a + 2\bar{r}t}\right)^a \qquad (2.12)$$

Если заметить, что полное число аминокислотных замен на сайт (d_G) равно $2\bar{r}t$, а $\bar{q}$ выражается как $1 - p$, то мы получаем следующее уравнение для d_G

$$d_G = a[(1-p)^{-\frac{1}{a}} - 1] \qquad (2.13)$$

Оценка $\hat{d}_G$ дистанции d_G получается заменой p на $\hat{p}$. Дисперсия большой выборки $\hat{d}_G$ вычисляется по формуле

$$V(\hat{d}_G) = \frac{p(1-p)^{-\left(1+\frac{2}{a}\right)}}{n} \qquad (2.14)$$

В дальнейшем мы будем для простоты называть $\hat{d}_G$ **гамма-дистанцией** (**Г-дистанция**). На рис. 2.4 показана зависимость между p и d_G для разных значений a. Влияние вариации r на оценку числа замен большое только при $p > 0{,}2$ и $a < 0{,}65$. Поэтому при $p < 0{,}2$ нет необходимости использовать гамма дистанцию d_G.

В вышеприведенных вычислениях мы учитывали различия в скорости замен по разным аминокислотным сайтам. В реальности, однако, скорость замен изменяется не только по аминокислотным сайтам, но также и по аминокислотным парам. Например, аргинин и лизин являются основными аминокислотами и заменяются друг на друга гораздо чаще, чем на другие аминокислоты в процессе эволюции. Если принять этот факт во внимание, также как и вариацию r, то можно использовать следующую формулу для оценки числа аминокислотных замен на сайт d_R **(дистанция Гришина)**.

$$q = \frac{\ln(1 + 2d_R)}{2d_R} \qquad (2.15)$$

Здесь d_R оценивается численно путем решения вышеприведенного уравнения для заданной величины q. Один из путей решения

вышеприведенного уравнения – это применение метода итераций Ньютона. Однако, зависимость между q и d_R может быть также выражена уравнением 2.13 с параметром a = 0.65. По сути, из рис. 2.4 видно, что дистанция Гришина d_R может быть оценена как

$$\hat{d}_R = 0{,}65[(1-p)^{-\frac{1}{0{,}65}} - 1] \qquad (2.16)$$

до тех пор пока $\hat{d}_G \leq 3{,}0$.

Феng и соавт. использовал дистанцию Гришина для оценки времени дивергенции между эубактериямя и эукариотами, в то время как Гогартен и соавт. использовал гамма-дистанцию с a = 0,7. Так как дистанция Гришина представляет собой практически гамма-дистанцию с a = 0,65, неудивительно, что две группы авторов получили очень сходные результаты (около 3-4 миллиардов лет назад). Однако значения a могут изменяться в зависимости от набора данных, поэтому важно оценивать a исходя из рассматриваемого набора данных. Для этого можно использовать метод Гу-Жанга, являющийся простым, но дающим вполне точные результаты.

2.3. Оценка эволюционных дистанций и скорости аминокислотных замен в α цепях гемоглобина.

В табл. 2.1 приведены доли отличающихся аминокислот $\hat{p}$, полученные при попарном сравнении α цепей гемоглобина из шести видов позвоночных. Исходя из этих значений, мы можем оценить РС-дистанцию d и Г-дистанцию d_G. Например, $\hat{p}$ для пары человек/корова составляет 0.121. Если подставить это значение в уравнение 2.6, получаем $\hat{d}_G$ = 0.129. Дисперсия и среднеквадратическая погрешность $\hat{d}_G$ будут равны $V(\hat{d})$ = 0/000961 и $s(\hat{d}) \equiv \left[V(\hat{d})\right]^{1/2}$ = 0.031, соответственно. Значения $\hat{d}_G$ и $s(\hat{d})$ для других видов представлены, соответственно, в таблицах 2.2 и 2.3. Для оценки скорости аминокислотных замен r помимо эволюционной дистанции необходимо знать время дивергенции t (уравнение 2.8). Для пары

человек/корова ***t*** примерно равно 90 миллионам лет, а $\hat{d}_G$ = 0.129. Поэтому ***r*** = $0.129/(2\times90\times10^6) = 0.717 \times 10^{-9}$ на сайт в год.

Для пары человек/корова $\hat{p}$ и $\hat{d}_G$ отличаются не сильно, так как $\hat{p}$ в этом случае мало, и, следовательно, мало число замен, приходящихся на один сайт. Однако, при увеличении $\hat{p}$ разность между $\hat{p}$ и $\hat{d}_G$ также увеличивается (см. рис. 2.4). Это четко видно из данных по сравнению последовательностей гемоглобина в паре карп/человек, для которой разница между $\hat{p}$ (=0.486) и $\hat{d}_G$ (=0.665) значительно больше, чем для пары человек/корова (табл. 2.1 и 2.2).

Таблица 2.2. PC-дистанции (выше диагонали) и Г-дистанции (ниже диагонали) при a = 2 для разных пар α цепей гемоглобина шести видов позвоночных.

	Человек	Лошадь	Корова	Кенгуру	Тритон	Карп
Человек		0.129	0.129	0.205	0.572	0.665
Лошадь	0,134		0.129	0.232	0.638	0.651
Корова	0,134	0,134		0.197	0.598	0.624
Кенгуру	0,216	0,246	0,207		0.638	0.708
Тритон	0,662	0,751	0,697	0,751		0.752
Карп	0,789	0,770	0,733	0,849	0,913	

PC дистанция ***d*** вычисляется с учетом допущения о том, что все аминокислотные сайты эволюционируют с одинаковой скоростью. Если это допущение неприменимо, то PC дистанция будет давать переоценку числа аминокислотных замен на сайт. В это случае, необходимо использовать Г-дистанцию, для расчета которой необходимо знать гамма параметр ***a***. В данном случае, мы полагаем ***a*** = 2 и вычисляем d_G из уравнения 2.13. Для пары человек/корова $\hat{p}$ = 0.121 и $\hat{d}_G$ = 0.134 (табл. 2.2). Дисперсия и среднеквадратическая погрешность $\hat{d}_G$ равны 0.0011225 и 0.034,

соответственно (табл. 2.3). $\hat{d}_G$ лишь немного выше, чем $\hat{d}$, потому что $\hat{p}$ не очень большое. Для пары человек/карп, однако, получаем $\hat{d}$ = 0.665 ± 0.082 и $\hat{d}_G$ = 0.789 ± 0.115. Таким образом, отличие между $\hat{d}$ и $\hat{d}_G$ увеличивается с возрастанием $\hat{p}$. Среднеквадратическая погрешность для $\hat{d}$ меньше, чем для $\hat{d}_G$. Это происходит потому, что $\hat{d}$ основывается на модели с одним параметром ***r***, в то время как $\hat{d}_G$ основывается на модели с двумя параметрами (***r***, ***a***). В целом, чем больше параметров в модели, тем больше вариация и среднеквадратическая погрешность.

Таблица 2.3. Оценки среднеквадратической погрешности PC-дистанций, вычисленные аналитическим (выше диагонали) и бутстрэп (ниже диагонали) методами.

	Человек	Лошадь	Корова	Кенгуру	Тритон	Карп
Человек		0.031	0.031	0.039	0.078	0.083
Лошадь	0.031		0.030	0.043	0.083	0.081
Корова	0.031	0.031		0.038	0.080	0.079
Кенгуру	0.040	0.043	0,039		0.081	0.084
Тритон	0.074	0.080	0,076	0,080		0.090
Карп	0.082	0.081	0,079	0,086	0,089	

2.4. Матрица аминокислотных замен.

Эмпирические исследования аминокислотных замен показали, что замены происходят чаще между аминокислотами, сходными по своим биохимическим свойствам, таким как полярность или масса. Другими словами, аминокислотные замены обычно происходят не случайным образом; обратные и параллельные мутации могут достаточно часто иметь место для схожих аминокислот. Некоторые аминокислоты, такие как цистеин, глицин и триптофан заменяются редко. Неравные уровни замен в

разных аминокислотных сайтах также будут давать вклад в неточность оценок, полученных с помощью PC-дистанции. Для того чтобы учесть эти факторы М.Дэйхофф и соавт. предложили другой метод оценки эволюционных дистанций. В этом методе рассматривается матрица аминокислотных замен для относительно малого периода времени, и зависимость между долей идентичных аминокислот и числом аминокислотных замен выводится эмпирически. Матрица аминокислотных замен, которую использовала М.Дэйхофф, была выведена из эмпирических данных для многих белков, таких как гемоглобин, цитохром *c* и фибринопептиды. Относительные частоты замен между разными аминокислотными остатками выводились на основе предварительно построенных эволюционных деревьев для близкородственных аминокислотных последовательностей. На основе полученных данных была составлена эмпирическая матрица аминокислотных замен ***M*** для 20 аминокислот (рис. 2.5).

Аминокислоты после замен

Исходные аминокислоты	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	L eu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	9867	1	4	6	1	3	10	21	1	2	3	2	1	1	13	28	22	0	1	13
Arg	2	9913	1	0	1	9	0	1	8	2	1	37	1	1	5	11	2	2	0	2
Asn	9	1	9822	42	0	4	7	12	18	3	3	25	0	1	2	34	13	0	3	1
Asp	10	0	36	9859	0	5	56	11	3	1	0	6	0	0	1	7	4	0	0	1
Cys	3	1	0	0	9973	0	0	1	1	2	0	0	0	0	1	11	1	0	3	3
Gln	8	10	4	6	0	9876	35	3	20	1	6	12	2	0	8	4	3	0	0	2
Glu	17	0	6	53	0	27	9865	7	1	2	1	7	0	0	3	6	2	0	1	2
Gly	21	0	6	6	0	1	4	9935	0	0	1	2	0	1	2	16	2	0	0	3
His	2	10	21	4	1	23	2	1	9912	0	4	2	0	2	5	2	1	0	4	3
Ile	6	3	3	1	1	1	3	0	0	9872	22	4	5	8	1	2	11	0	1	57
Leu	4	1	1	0	0	3	1	1	1	9	9947	1	8	6	2	1	2	0	1	11
Lys	2	19	13	3	0	6	4	2	1	2	2	9926	4	0	2	7	8	0	0	1
Met	6	4	0	0	0	4	1	1	0	12	45	20	9874	4	1	4	6	0	0	17
Phe	2	1	1	0	0	0	0	1	2	7	13	0	1	9946	1	3	1	1	21	1
Pro	22	4	2	1	1	6	3	3	3	0	3	3	0	0	9926	17	5	0	0	3
Ser	35	6	20	5	5	2	4	21	1	1	1	8	1	2	12	9840	32	1	1	2
Thr	32	1	9	3	1	2	2	3	1	7	3	11	2	1	4	38	9871	0	1	10
Trp	0	8	1	0	0	0	0	0	1	0	4	0	0	3	0	5	0	9976	2	0
Tyr	2	0	4	0	3	0	1	0	4	1	2	1	0	28	0	2	2	1	9945	2
Val	18	1	1	1	2	1	2	5	1	33	15	1	4	0	2	2	9	0	1	9901

Рис. 2.5. Матрица аминокислотных замен М.Дэйхофф для эволюционной дистанции, равной 1 PAM .
Все значения умножены на 10000.

Элемент m_{ij} этой матрицы представляет собой вероятность замены аминокислоты в ряде i на аминокислоту в колонке j в течение одной эволюционной единицы времени. За единицу времени, используемую в матрице, берется время, в течение которого в среднем происходит одна аминокислотная замена на 100 аминокислотных сайтов. Число аминокислотных замен измеряется в единицах принятых точечных мутаций[19] (**PAM**). Одна единица **PAM** представляет собой одну аминокислотную замену на 100 аминокислотных сайтов. Матрица замен на рис. 2.5 показывает вероятность аминокислотных замен для одной **PAM**.

Матрица аминокислотных замен может быть использована для предсказания аминокислотных замен для любого эволюционного времени, если мы знаем исходные частоты аминокислот. Пусть g_o будет вектор-строкой относительных частот 20 аминокислот в полипептиде в момент времени 0. Аминокислотные частоты в момент времени t или для t **PAM** будут равны

$$g_t = g_o M_t \qquad (2.17)$$

где $M_t = M^t$.

Здесь мы замечаем, что элемент $m_{t(ij)}$ матрицы M_t дает вероятность того, что аминокислота в ряде i в момент времени 0 изменится на аминокислоту в колонке j в момент времени t. В частности $m_{t(ii)}$ предсталяет собой вероятность того, что i-ая аминокислота в момент времени t останется такой же, как и исходная. Эта вероятность может быть использована для установки отношения разных аминокислот между гомологичными последовательностями p к числу аминокислотных замен на сайт d_D при допущении, что аминокислотные частоты уравновешены и остаются одинаковыми в течение эволюционного времени. В этом случае p дается формулой

$$p = 1 - \sum_i g_i m_{2t(ii)} \qquad (2.18)$$

[19] accepted point mutations

где g_i – это уравновешенная частота ***i***-ой аминокислты в исследуемой последовательности. Здесь, мы используем $m_{2t(ii)}$ вместо $m_{t(ii)}$, потому что мы рассматриваем пару последовательностей, дивергировавших **t** временных единиц назад. Так как $d_D = 0{,}01 \times t$ ($= 0{,}01$**PAM**) и $m_{2t(ii)}$ может быть получена из M_{2t}, p может быть соотнесена с d_D.

На практике, аминокислотные частоты g могут варьироваться от белка к белку, поэтому используются аминокислотные частоты, усредненные по многим разным белкам. Такой подход не учитывает специфичность каждого белка, но за то несомненно делает метод применимым для многих разных белков. Более того, если заметить, что разные белки имеют довольно схожие аминокислотные частоты, эта процедура оказывается приемлемой для получения оценок числа аминокислотных замен.

Используя вышеописанный метод, Дэйхофф и соавт. вывели зависимость между p и d_D (рис. 2.4). Этот рисунок также содержит значения $d = -ln(1 - p)$ из уравнения 2.6 и d_G из уравнения 2.13 с $a = 2{,}25$. Как и предполагалось, разница между d_D и d постепенно увеличивается с увеличением **p**. Значение d_G с $a = 2{,}25$ очень близко к $\mathbf{d_D}$ практически для всех значений p, что указывает на то, что d_D может быть аппроксимировано d_G с $a = 2{,}25$.

Хотя матрица аминокислотных замен М.Дэйхофф все еще используется, Джоунс и соавт. составил новую матрицу, основываясь на большем количестве данных по заменам для различных белков. Адачи и Хасегава составили матрицу замен для 13 митохондриальных белков позвоночных. Теоретически для разных белков должны быть разные матрицы, поэтому желательно создавать матрицы замен для каждой группы белков по мере накопления данных по аминокислотным последовательностям. Тем не менее, для измерения эволюционной дистанции между парой последовательностей проще оценить гамма параметр a и использовать гамма дистанцию из уравнения 2.13. Для матрицы Джоунса,

зависимость между p и числом аминокислотных замен приблизительно дается уравнением 2.13 с $a = 2{,}4$.

В последние годы матрица аминокислотных замен использовалась для реконструкции филогенетических деревьев методом наибольшего правдоподобия и для получения аминокислотных последовательностей предковых белков.

2.5. Скорость мутаций и скорость замен.

До этого времени мы рассматривали замены, как если бы любое произошедшее мутационное изменение нуклеотида или аминокислоты закреплялось в рассматриваемых последовательностях. На практике же это не так, потому что каждый вид представляет собой популяцию, состоящую из многих особей, и новые мутации, происходящие у отдельной особи, могут исчезнуть из популяции случайным образом или посредством очищающего отбора. Мутация закрепляется в геноме данного вида, только когда она распространяется по всей популяции. Это событие называется **фиксацией мутации в популяции**. Вероятность фиксации мутантного аллеля зависит от: 1) изначальной частоты аллеля, 2) селективного преимущества и недостатка аллеля, 3) эффективного размера популяции.

Когда мутация нейтральна и не влияет на приспособленность генотипа, относительная частота x мутантного аллеля может увеличиться или уменьшиться случайным образом в популяции. Для простоты рассмотрим организмы с дискретными поколениями, такие как однолетние растения или некоторые виды насекомых, и пусть N будет числом взрослых индивидуумов. Частота мутантного аллеля исходно равна $x = 1/(2N)$. Судьба этого аллеля определяется только случайным образом, и x может, как увеличиваться, так и уменьшаться. Это процесс будет продолжаться до тех пор, пока аллель не зафиксируется или, наоборот, не потеряется в популяции.

Так как начальная частота равна 1/(2***N***), то вероятность фиксации ***u*** равна $\boldsymbol{u} = \frac{1}{2\boldsymbol{N}}$, в то время как вероятность потери аллеля равна $1 - \boldsymbol{u} = 1 - \frac{1}{2\boldsymbol{N}}$.

А теперь предположим, что нейтральные мутации происходят со скоростью ***v*** на локус на поколение, так что полное число мутаций, появляющихся в целой популяции, равно 2***Nv*** на локус на поколение. Так как доля новых нейтральных мутаций, которые закрепятся в популяции, равна 1/(2***N***), то скорость замены гена[20] в локусе за поколение (***a***), равна

$$\boldsymbol{a} = 2\boldsymbol{N}\boldsymbol{v} \times \frac{1}{2\boldsymbol{N}} = \boldsymbol{v} \qquad (2.19)$$

Другими словами, скорость замены гена в локусе равна скорости мутаций.

В случае аминокислот мы обычно рассматриваем замены на аминокислотный сайт в год. Однако, если мы переопределим скорость мутаций на скорость ***μ*** мутаций на аминокислотный сайт в год, то вышеописанное правило будет применимо к нейтральным мутациям. Следовательно, скорость аминокислотных замен на сайт в год (***r***) равна скорости мутаций ***μ***. Она равна

$$\boldsymbol{r} = 2\boldsymbol{N}\boldsymbol{\mu} \times \frac{1}{2\boldsymbol{N}} = \boldsymbol{\mu} \qquad (2.20)$$

В качестве примера рассмотрим гипотетический полипептид из 100 аминокислотных остатков и предположим, что мутация, приводящая к замене одной аминокислоты на другую, происходит со скоростью 10^{-9} на аминокислотный сайт в год, или со скоростью 10^{-7} на весь полипептид в год. Если эффективный размер популяции 10^5, то полное число мутаций в этом полипептиде будет равно $2\times10^5\times10^{-7} = 0.02$ в год. Однако, так как они фиксируются с вероятностью $1/(2\times10^5)$, доля аминокислотных замен на сайт ***r*** $= 0{,}02/(2\times10^5\times100) = 10^{-9}$, что равно скорости замен на сайт в год.

А что будет, если мутация благоприятная? Для простоты предположим, что относительная приспособленность[21], выраженная в числе произведенных

[20] rate of gene substitution
[21] fitness

потомков, равна 1, 1 + ***s***, и 1 + 2***s*** для для генотипов A_1A_1, A_1A_2 и A_2A_2. Здесь ***s*** называется коэффициентом селекции (селективным преимуществом), и он положительный для благоприятных мутаций и отрицательный для неблагоприятных мутаций. Вывод вероятности фиксации благоприятных мутаций довольно сложен, но в больших популяциях ($Ns \gg 1$), вероятность примерно равна 2***s*** и она независима от ***N***. Следовательно, если скорость благоприятных мутаций на аминокислотный сайт в год обозначить как μ, то доля аминокислотных замен, появляющихся под действием положительного отбора, равна

$$r = 2N\mu \times 2S = 4Ns\mu \qquad (2.21)$$

Предположим, что $\mu = 10^{-9}$, $N = 10^5$ и $s = 0{,}01$. Тогда ***r*** становится равным 4×10^{-6}, что в 4000 раз выше, чем для нейтральных мутаций. Это указывает на то, что скорость аминокислотных замен весьма сильно увеличивается при селективном преимуществе. Конечно, этот пример довольно искуственен. Если функция белка установилась в процессе эволюции, большинство мутаций будет вредоносными[22] или нейтральными, и только небольшое число мутаций смогут увеличить активность белка. Следовательно, скорость благоприятных мутаций должна быть гораздо меньше, чем нейтральных, и различные статистические исследования эволюции белков позволили предположить, что доля аминокислотных замен, появляющихся под действием положительного отбора, довольно мала.

[22] deleterious

3. ЭВОЛЮЦИОННЫЕ ИЗМЕНЕНИЯ НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ.

3.1. Нуклеотидные отличия между последовательностями.

Процесс эволюционных изменений в последовательностях ДНК является более сложными, чем в белках, так как существует несколько типов областей ДНК, таких как белок-кодирующие области, некодирующие области, экзоны, интроны, фланкирующие области, повторы ДНК и инсерционные последовательности. Потому важно знать тип исследуемой области ДНК и выполняемую ее функцию. Более того, даже если мы рассматриваем только белок-кодирующую область, паттерн нуклеотидных замен в первом, втором и третьем положениях кодона неодинаковый. Некоторые области подвергаются влиянию естественного отбора чаще других, и это также вносит вклад в вариацию эволюционных паттернов среди разных областей ДНК.

Мы в основном затронем белок- и РНК-кодирующие области, так как они наиболее важны для понимания основных аспектов эволюции. Нашей целью будет ввести статистические методы изучения эволюции этих областей и показать, как проводить их анализ.

Когда две последовательности ДНК дивергируют от общей предковой последовательности, они постепенно изменяются путем накопления нуклеотидных замен. Простейшей оценкой степени дивергенции последовательностей является относительная частота замен, по которым две последовательности отличаются. Она вычисляется по формуле

$$\hat{p} = \frac{n_d}{n} \qquad (3.1)$$

где n_d и n – это число различающихся нуклеотидов между двумя последовательностями и полное число исследуемых нуклеотидов, соответственно. В дальнейшем, мы будем называть величину p, как **p-дистанция** для нуклеотидных последовательностей.

Хотя полное нуклеотидное отличие вычисляется из формулы 3.1, часто полезно знать частоты разных нуклеотидных пар между двумя последовательностями X и Y. Так как молекулы ДНК состоят всего из четырех типов нуклеотидов (A, G, T и C), то при сравнении двух последовательностей существует 16 типов нуклеотидных пар (табл. 2.1). Существует 4 пары идентичных нуклеотидов (AA, TT, CC и GG), 4 пары типа транзиции (AG, GA, TC и CT) и 8 пар типа трансверсии (все остальные пары). Обозначим относительные частоты идентичных пар как $\boldsymbol{O}$, пар типа транзиции как $\boldsymbol{P}$ и пар типа трансверсии как $\boldsymbol{Q}$. Очевидно, что $\boldsymbol{p} = \boldsymbol{P} + \boldsymbol{Q}$.

Таблица 3.1. 16 возможных типов нуклеотидных пар между двумя последовательностями

Типы нуклеотидных пар	***Нуклеотидная пара***				
Идентичные нуклеотиды	AA	TT	CC	GG	*Общая частота:*
Частота	$\boldsymbol{O_1}$	$\boldsymbol{O_2}$	$\boldsymbol{O_3}$	$\boldsymbol{O_4}$	$\boldsymbol{O}$
Пары типы транзиция	AG	GA	TC	CT	*Общая частота:*
Частота	$\boldsymbol{P_{11}}$	$\boldsymbol{P_{12}}$	$\boldsymbol{P_{21}}$	$\boldsymbol{P_{22}}$	$\boldsymbol{P}$
Пары типа трансверсия	AT	TA	AC	GA	
Частота	$\boldsymbol{Q_{11}}$	$\boldsymbol{Q_{12}}$	$\boldsymbol{Q_{21}}$	$\boldsymbol{Q_{22}}$	
	TG	GT	CG	GC	*Общая частота:*
Частота	$\boldsymbol{Q_{31}}$	$\boldsymbol{Q_{32}}$	$\boldsymbol{Q_{41}}$	$\boldsymbol{Q_{42}}$	$\boldsymbol{Q}$

Если замены происходят случайным образом по всем четырем нуклеотидам, то $\boldsymbol{Q}$ будет в два раза выше, чем $\boldsymbol{P}$ при малых значениях $\boldsymbol{p}$. В реальности же транзиции встречаются гораздо чаще, чем трансверсии, следовательно, $\boldsymbol{P}$ может быть больше, чем $\boldsymbol{Q}$. Когда степень дивергенции последовательностей невелика, **отношение транзиций к трансверсиям**[23] $\boldsymbol{R}$ может быть вычислено как

[23] transition/transversion ratio

$$\hat{R} = \frac{\hat{P}}{\hat{Q}} \qquad (3.2)$$

где $\hat{P}$ и $\hat{Q}$ – наблюдаемые значения P и Q, соответственно. Для многих ядерных генов R обычно находится в пределах от 0.5 до 2, однако в митохондриальной ДНК значение R может достигать 15.

Как мы заметим ниже, оценка числа нуклеотидных замен часто зависит от допущения, что нуклеотидные частоты в каждой последовательности уравновешены и не изменяются с течением времени. Тогда можно ожидать, что $P_{11} = P_{12}$, $P_{21} = P_{22}$, $Q_{11} = Q_{12}$, $Q_{21} = Q_{22}$, $Q_{31} = Q_{32}$, $Q_{41} = Q_{42}$ в табл. 3.1.

3.2. Оценка числа нуклеотидных замен

Как и для случая аминокислотных замен, p-дистанция для нуклеотидных последовательностей дает хорошую оценку числа нуклеотидных замен на сайт только для близкородственных последовательностей. Однако, когда значения p велики, происходит недооценка числа замен, так как не учитываются обратные и параллельные мутации. Эта проблема более серьезна для нуклеотидных последовательностей, чем для аминокислотных, потому что в нуклеотидных последовательностях есть только четыре составляющих их нуклеотида.

Для оценки числа нуклеотидных замен необходимо использовать специальные математические модели для нуклеотидных замен. Некоторые из них представлены в табл. 3.2 в виде матриц вероятностей замен.

3.2.1. Метод Джукса-Кантора[24]

Одной из самых простых моделей нуклеотидных замен является модель, предложенная Джуксом и Кантором в 1969 году. В этой модели

[24] Jukes and Cantor method

Таблица 3.2. Модели нуклеотидных замен.

*Элемент e_{ij} в вышеприведенных матрицах замен соответствует уровню замен нуклеотида в **i**-ом ряду на нуклеотид в **j**-ой колонке. g_A, g_T, g_C и g_G – это частоты нуклеотидов. $\theta_1 = g_C + g_G$, $\theta_2 = g_A + g_T$.*

(А) Модель Джукса-Кантора

	A	T	C	G
A	–	α	α	α
T	α	–	α	α
C	α	α	–	α
G	α	α	α	–

(Б) Модель Кимуры

	A	T	C	G
A	–	β	β	α
T	β	–	α	β
C	β	α	–	β
G	α	β	β	–

(В) Модель equal-input

	A	T	C	G
A	–	αg_T	αg_C	αg_G
T	αg_A	–	αg_C	αg_G
C	αg_A	αg_T	–	αg_G
G	αg_A	αg_T	αg_C	–

(Б) Модель Тамуры

	A	T	C	G
A	–	$\beta\theta_2$	$\beta\theta_1$	$\alpha\theta_1$
T	$\beta\theta_2$	–	$\alpha\theta_1$	$\beta\theta_1$
C	$\beta\theta_2$	$\alpha\theta_2$	–	$\beta\theta_1$
G	$\alpha\theta_2$	$\beta\theta_2$	$\beta\theta_1$	–

(Д) Модель HKY

	A	T	C	G
A	–	βg_T	βg_C	αg_G
T	βg_A	–	αg_C	βg_G
C	βg_A	αg_T	–	βg_G
G	αg_A	βg_T	βg_C	–

(Е) Модель Тамуры-Нея

	A	T	C	G
A	–	βg_T	βg_C	$\alpha_1 g_G$
T	βg_A	–	$\alpha_2 g_C$	βg_G
C	βg_A	$\alpha_2 g_T$	–	βg_G
G	$\alpha_1 g_A$	βg_T	βg_C	–

(Ж) Модель general reversible

	A	T	C	G
A	–	$a g_T$	$b g_C$	$c g_G$
T	$a g_A$	–	$d g_C$	$e g_G$
C	$b g_A$	$d g_T$	–	$f g_G$
G	$c g_A$	$e g_T$	$f g_C$	–

(З) Модель unrestricted

	A	T	C	G
A	–	a_{12}	a_{13}	a_{14}
T	a_{21}	–	a_{23}	a_{24}
C	a_{31}	a_{32}	–	a_{34}
G	a_{41}	a_{42}	a_{43}	–

предполагается, что нуклеотидные замены происходят в каждом нуклеотидном сайте с одинаковой частотой, и что в каждом сайте нуклеотид заменяется на один из трех оставшихся нуклеотидов с вероятностью α в год (или в любой другой промежуток времени) (табл. 3.2А). Таким образом, вероятность замены нуклеотида на любой из трех других нуклеотидов $r = 3\alpha$. Вероятность r равна скорости нуклеотидных замен на сайт в год. Рассмотрим две нуклеотидные последовательности, X и Y, дивергировавшие от общей предковой последовательности t лет назад. Обозначим через q_t долю идентичных нуклеотидов между X и Y, а через p_t (= 1 – q_t) долю различающихся нуклеотидов. Доля идентичных нуклеотидов q_{t+1} в момент времени t + 1 (измеренное в годах) может быть получена следующим образом. Во-первых, мы замечаем, что сайт, в котором содержится один и тот же нуклеотид в X и Y в момент времени t, останется тем же в момент времени t + 1 с вероятностью $(1 - r)^2$, что примерно равно $1 - 2r$, так как r достаточно мало и слагаемым r^2 можно пренебречь. Во-вторых, сайт, в котором содержатся разные нуклеотиды в X и Y в момент времени t, останется тем же в момент времени t +1 с вероятностью $2r/3$. Эта вероятность получается, если заметить, что когда последовательности X и Y содержат нуклеотиды i и j, соответственно, в момент времени t, они становятся одинаковыми, если i в X меняется на j, а j в Y остается тем же, или, если j в Y меняется на i, а i в X остается тем же. Вероятность первого события равна $\alpha(1 - r) = r(1 - r)/3$, так как i в X должен замениться на j, а не на оставшиеся два нуклеотида, а j в Y должен остаться неизменным. Вероятность второго события также равна $r(1 - r)/3$. Поэтому полная вероятность равна $2r(1 - r)/3$, что примерно равно $2r/3$, если пренебречь слагаемым r^2. Таким образом, мы получаем следующее уравнение

$$q_{t+1} = (1-2r)q_t + \frac{2}{3}r(1-q_t) \qquad (3.3)$$

что может быть записано, как

$$q_{t+1} - q_t = \frac{2r}{3} - \frac{8r}{3} q_t \qquad (3.4)$$

Перейдем к непрерывной модели, заменяя $q_{t+1} - q_t$ на dq/dt и отбрасывая индекс t в q_t. Тогда мы получаем следующее дифференциальное уравнение

$$\frac{dq}{dt} = \frac{2r}{3} - \frac{8r}{3} q \qquad (3.5)$$

Решение этого уравнения при начальных условиях $q = 1$ и $t = 0$ такое

$$q = 1 - \frac{3}{4}\left(1 - e^{-8rt/3}\right) \qquad (3.6)$$

В этой модели ожидаемое число нуклеотидных замен на сайт d для двух последовательностей равно $2rt$. Тогда d рассчитывается по формуле

$$d = -\frac{3}{4} \ln\left(1 - \frac{4}{3} p\right) \qquad (3.7)$$

где $p = 1 - q$ является долей различающихся аминокислот между X и Y.

Оценка $\hat{d}$ дистанции d может быть получена заменой p на наблюдаемую величину $\hat{p}$ и дисперсия большой выборки $\hat{d}$ задается формулой

$$V(\hat{d}) = \frac{9p(1-p)}{(3-4p)^2 n} \qquad (3.8)$$

В вышеприведенной модели, мы предполагали, что скорость нуклеотидных замен одинакова для кадой пары последовательностей, так что предполагаемые последовательности A, T, C и G будут в конечном счете равны 0.25. Тем не менее, раз мы не делали допущений об исходных частотах, уравнение 3.8 является независимым от исходных частот. Другими словами, нет необходимости принимать стационарность нуклеотидных частот для того чтобы уравнение 3.8 было применимо.

3.2.2. Двухпараметрический метод Кимуры[25]

Как отмечалось ранее, скорость транзиций часто выше скорости трансверсий в реальных данных. Кимура предложил метод оценки числа

[25] Kimura's two-parameter method

нуклеотидных замен на сайт, учитывающий эти наблюдения. В этой модели скорость транзиций на сайт в год α предполагается отличной от скорости трансверсий 2β (рис. 1.3 и табл. 2.2Б). Полная скорость замен в год r, таким образом, равна $\alpha + 2\beta$. Используя эту модель Кимура показал, что P и Q в табл. 3.2 определяются выражениями

$$P = \frac{1}{4}\left(1 - 2e^{-4(\alpha+\beta)t} + e^{-8\beta t}\right) \tag{3.9}$$

$$Q = \frac{1}{2}\left(1 - e^{-8\beta t}\right) \tag{3.10}$$

где t – это время, прошедшее после дивергенции двух последовательностей X и Y. Тогда ожидаемое число нуклеотидных замен на сайт между X и Y дается выражениями

$$d \equiv 2rt = 2\alpha t + 4\beta t = -\frac{1}{2}\ln(1 - 2P - Q) - \frac{1}{4}\ln(1 - 2Q) \tag{3.11}$$

а оценка ($\hat{d}$) d может быть получена заменой P и Q на их наблюдаемые величины. Вариация $\hat{d}$ дается формулой

$$V(\hat{d}) = \frac{1}{n}[c_1^2 P + c_3^2 Q - (c_1 P + c_3 Q)^2] \tag{3.12}$$

где

$$c_1 = \frac{1}{1 - 2P - Q}, \qquad c_2 = \frac{1}{1 - 2Q}, \qquad c_3 = \frac{c_1 + c_2}{2}$$

В данной модели возможно оценить число транзиций ($s = 2\alpha t$) и трансверсий ($v = 4\beta t$) на сайт. Формулы для s и v следующие

$$s = -\frac{1}{2}\ln(1 - 2P - Q) + \frac{1}{4}\ln(1 - 2Q) \tag{3.13}$$

$$v = -\frac{1}{2}\ln(1 - 2Q) \tag{3.14}$$

а вариации для оценок ($\hat{s}$ и $\hat{v}$) s и v следующие

$$V(\hat{s}) = \frac{1}{n}[c_1^2 P + c_4^2 Q - (c_1 P + c_4 Q)^2] \tag{3.15}$$

$$V(\hat{v}) = \frac{1}{n}[c_2^2 Q(1 - Q)] \tag{3.16}$$

где $c_4 = \frac{c_1 - c_2}{2}$. Следовательно, отношение транзиций к трансверсиям оценивается как

$$\hat{R} = \frac{\hat{s}}{\hat{v}} \qquad (3.17)$$

В модели Кимуры уравновешенная частота каждого нуклеотида равна 0.25. Однако, вышеприведенные формулы применимы в независимости от исходных нуклеотидных частот, и в этом отношении модель сходна с моделью Джукса-Кантора. Это свойство позволяет применять эти две модели при более широких условиях, чем многие другие модели.

Большинство биологов используют R, определенное по уравнениям 3.2 или 3.17. Однако, определение R варьируется в зависимости от математической модели и может быть достаточно сложным, когда используется сложная модель. Теоретики также склоняются к использованию скорости отношения транзиций к трансверсиям[26] k вместо R. В случае модели Кимуры R определяется как $\alpha/(2\beta)$, в то время как k равна α/β. Разные компьютерные программы часто используют разные определения отношения транзиций к трансверсиям, поэтому следует с осторожностью смотреть на введенное в программу отношение транзиций к трансверсиям.

3.2.3. Метод Таджимы-Нея[27]

Таджима и Ней разработали в 1984 году метод оценки числа замен, обладающий малой чувствительностью к разным возмущающим факторам. Он частично основан на модели «equal-input» (табл. 3.2B), независимо предложенный Фельсенштейном в 1981 году и Таджимой и Неем в 1982 году. В этом методе необходимо допустить стационарность частот нуклеотидов для оценки числа нуклеотидных замен d, и d дается выражением

$$d = -b\ln\left(1 - p/b\right) \qquad (3.18)$$

[26] transition/transversion rate ratio
[27] Tajima and Nei's method

где

$$\boldsymbol{b} = 1/2\left[1 - \sum_{i=1}^{4} \boldsymbol{g}_i^2 + \boldsymbol{p}^2/\boldsymbol{c}\right] \qquad (3.19)$$

Здесь *c* задается выражением

$$\boldsymbol{c} = \sum_{i=1}^{3} \sum_{j=i+1}^{4} \frac{\boldsymbol{x}_{ij}^2}{2\boldsymbol{g}_i \boldsymbol{g}_j} \qquad (3.20)$$

где $\boldsymbol{x}_{ij}(\boldsymbol{i} < \boldsymbol{j})$ – относительная частота нуклеотидной пары $\boldsymbol{i}$ и $\boldsymbol{j}$, когда сравниваются две последовательности ДНК. Нуклеотидные частоты $\boldsymbol{g}_i$ оцениваются из сравнения двух последовательностей. Оценка $\hat{\boldsymbol{d}}$ дистанции $\boldsymbol{d}$ получается заменой $\boldsymbol{p}$ на $\hat{\boldsymbol{p}}$ в уравнении 3.18, и вариация $\hat{\boldsymbol{d}}$ равна

$$\boldsymbol{V}(\hat{\boldsymbol{d}}) = \frac{\boldsymbol{b}^2 \boldsymbol{p}(1-\boldsymbol{p})}{(\boldsymbol{b}-\boldsymbol{p})^2 \boldsymbol{n}} \qquad (3.21)$$

Заметим, что когда скорость нуклеотидных замен одинакова для всех нуклеотидных пар, $\boldsymbol{b}$ ожидается равной 3/4 в равновесии, и уравнения 3.18 и 3.21 упрощаются до уравнений 3.7 и 3.8, соответственно. На практике $\boldsymbol{b}$ обычно меньше, чем 3/4 из-за неравности скоростей нуклеотидных замен, и в этом случае уравнение 3.18 дает большее значение, чем формула Джукса-Кантора.

3.2.4. Метод Тамуры

В модели Кимуры частоты четырех нуклеотидов, в конечном счете, становятся равными 0.25, как отмечалось ранее. В случае реальных данных, однако, частоты нуклеотидов редко равны друг другу, и GC содержание часто сильно отличается от 0.5. Например, в митохондриальной ДНК *Drosophila* GC содержание равно примерно 0.1.

Учитывая эти факты, Тамура разработал в 1992 году метод оценки $\boldsymbol{d}$ с моделью замен, приведенной в табл. 3.2Г. Эта модель является

продолжением 2-параметрической модели Кимуры для случая низкого или высокого GC содержания, и d вычисляется как

$$d = -h\ln\left(1 - P/h - Q\right) - 1/2\left(1 - h\right)\ln\left(1 - 2Q\right) \qquad (3.22)$$

где $h = 2\theta\left(1-\theta\right)$ и θ – GC содержание.

Как и в случае модели Кимуры мы можем рассчитать $V(\hat{d})$, $\hat{s}$, $V(\hat{s})$, $\hat{v}$, $V(\hat{v})$, $\hat{R}$ и $V(\hat{R})$, но из-за сложности формул, они не приводятся.

3.2.5. Метод Тамуры-Нея.

Одной из моделей филогенетической оценки методом наибольшего правдоподобия является HKY модель, которая представляет собой гибрид 2-параметрической модели Кимуры и модели «equal input», и учитывает и разницу между транзициями и трансверсиями и GC-содержание (табл. 3.2E). Формулы для $\hat{d}$ в этой модели достаточно сложны, поэтому мы рассмотрим модель Тамуры и Нея (табл. 3.2E), который включает в себя HKY модель в качестве частного случая и позволяет провести аналитический расчет d. Согласно этой модели, формула для *d* будет следующей:

$$d = -\frac{2g_Ag_G}{g_R}\ln\left[1 - \frac{g_R}{2g_Ag_G}P_1 - \frac{1}{2g_R}Q\right] - \frac{2g_Tg_C}{g_Y}\ln\left[1 - \frac{g_Y}{2g_Tg_C}P_2 - \frac{1}{2g_Y}Q\right]$$
$$- 2\left[g_Rg_Y - \frac{g_Ag_Gg_Y}{g_R} - \frac{g_Tg_Cg_R}{g_Y}\right]\ln\left[1 - \frac{1}{2g_Rg_Y}Q\right] \qquad (3.23)$$

где P_1 и P_2 – соотношения транзиционных отличий между A и G и между T и C, соответственно, а Q – соотношение трансверсионных отличий.

3.3. Сравнение дистанционных методов

Рассмотрим теоретические взаимоотношения разных дистанционных оценок числа нуклеотидных замен, полагая $n = \infty$. На рис. 2.1 показано число нуклеотидных замен, оцененных разными дистанциями, когда

действительная замена аминокислот подчиняется модели Тамуры-Нея. Здесь $g_A = 0.3$, $g_T = 0.4$, $g_C = 0.2$ и $g_G = 0.1$, $\alpha_1/\beta = 4$ и $\alpha_2/\beta = 8$. Очевидно, что оценка, полученная методом Тамуры-Нея равна вероятному числу замен d, а все остальные дистанции дают недооценку при увеличении вероятного числа замен. Различные дистанции дают существенно отличающиеся друг от друга результаты при $d \geq 0.6$.Дистанция Тамуры практически идентична дистанции Тамуры-Нея вплоть до $d = 0.5$, в то время как дистанции Тамуры, Кимуры и Джукса-Кантора, практически сходны с дистанцией Тамуры-Нея при $d \leq 0.25$. Даже p-дистанция становится схожей с остальными дистанциями при $p \leq 0.1$. Поэтому при изучении довольно близких последовательностей, нет необходимости использовать сложные дистанционные методы. В этом случае лучше использовать более простые, так как они дают меньшую дисперсию.

Стоит также заметить, что для постройки филогенетических деревьев по оцененным дистанциям, сложные дистанции вовсе необязательно более эффективны для получения правильной топологии, чем более простые, даже если математическая модель более точно передает данные. Для оценки длин ветвей дерева, однако, дистанция, лучше описывающая данные, обычно дает более правдоподобные результаты.

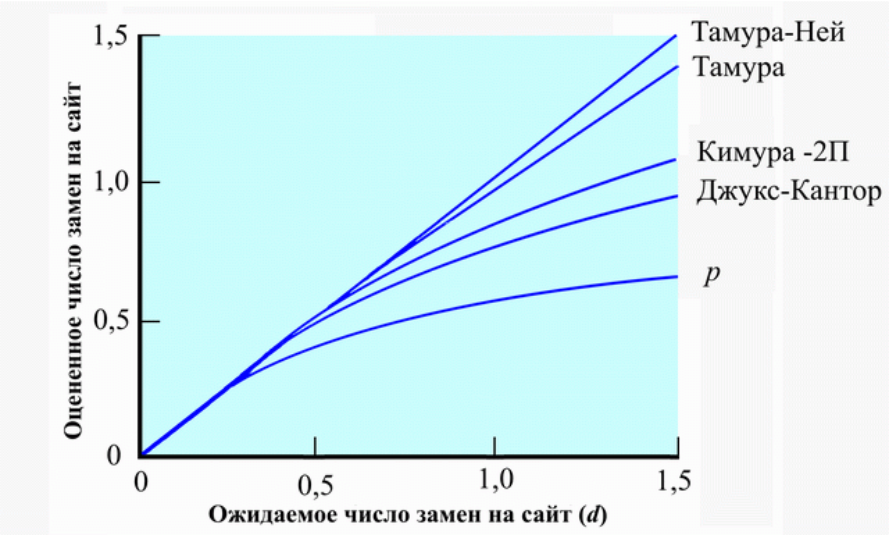

Рис. 3.1. Оценка числа нуклеотидных замен, полученная разными дистанционными методами, в случае, если действительные нуклеотидные замены подчиняются модели Тамуры-Нея.

3.4. Гамма-дистанции

В вышерассмотренных эволюционных дистанциях, скорость нуклеотидных замен считалась одинаковой. На самом деле это допущение выполняется достаточно редко, и скорость варьируется от сайта к сайту. В случае белок-кодирующих генов, это очевидно, так как первое, второе и третьи положения в кодонах имеют разные скорости замен. Функциональное давление аминокислот, составляющих активные центры белков, также вносит вклад в вариацию скорости замен в нуклеотидных сайтах. Колебания скорости замен также наблюдается и в РНК-кодирующих генах, так как РНК обладает функциональными константами и обычно образует вторичную структуру из петель и стеблей, обладающих разными скоростями замен. Статистический анализ показал, что вариация скорости нуклеотидных замен примерно подчиняется гамма распределению. По этой причине были разработаны гамма-дистанции, учитывающие вариацию скоростей нуклеотидных замен. Г-дистанции могут быть выведены теми же математическими методами, которые использовались для выведения Г-дистанций для аминокислотных последовательностей.

3.4.1. Гамма-дистанция для модели Джукса-Кантора

Когда нуклеотидные замены подчиняются модели Джукса-Кантора, но скорость нуклеотидных замен r варьируется согласно гамма-распределению, Г-дистанция равна

$$\boldsymbol{d} = \frac{3}{4}\boldsymbol{a}\left[\left(1-\frac{4}{3}\boldsymbol{p}\right)^{-\frac{1}{a}}-1\right] \qquad (3.24)$$

Дисперсия оценки $\hat{\boldsymbol{d}}$ дистанции $\boldsymbol{d}$ дается формулой

$$V(\hat{d}) = \left[\frac{p(1-p)}{n}\right]\left[\left(1-\frac{4}{3}p\right)^{-2(\frac{1}{a}+1)}\right] \quad (3.25)$$

Здесь a – это гамма параметр, определенный в главе 2.

3.4.2. Гамма-дистанция для модели Кимуры

Для этой модели гамма дистанция d и вариация оценки $\hat{d}$ дистанции d вычисляется следующим образом

$$d = \frac{a}{2}\left[(1-2P-Q)^{-1/a} + \frac{1}{2}(1-2Q)^{-1/a} - \frac{3}{2}\right] \quad (3.26)$$

$$V(\hat{d}) = \frac{1}{n}\left[c_1^2 P + c_3^2 Q - (c_1 P + c_3 Q)^2\right] \quad (3.27)$$

где $c_1 = (1-2P-Q)^{-1/a+1}$, $c_2 = (1-2Q)^{-1/a+1}$, $c_3 = (c_1 + c_2)/2$, где P и Q рассчитываются так же как для модели Кимуры.

Как и для случая дистанции Кимуры мы можем сосчитать число транзиций s и трансверсий v на сайт:

$$s = \frac{a}{2}\left[(1-2P-Q)^{-1/a} - \frac{1}{2}(1-2Q)^{-1/a} - \frac{1}{2}\right] \quad (3.28)$$

$$v = \frac{a}{2}\left[(1-2Q)^{-1/a} - 1\right] \quad (3.29)$$

3.4.3. Гамма-дистанция для модели Тамуры-Нея.

Известно, что в контрольной области митохондриальной ДНК млекопитающих скорость нуклеотидных замен сильно варьируется от сайта к сайту и существует сильный разброс транзиций/трансверсий. Гамма дистанция для модели Тамуры-Нея изначально была разработана для последовательностей данной области. Существует две гипервариабельных области (5`- и 3`-сегменты в этой области) и высоко консервативная промежуточная область. Используя данные по последовательности митохондриальной ДНК человека, были получены оценки значений a: a =

0.11 для всей контрольной области и $\boldsymbol{a}$ = 0.47 для 5`-гипервариабельной области. Гамма дистанция для модели Тамуры-Нея будет равна

$$d = 2a\left[\frac{g_A g_G}{g_R}\left(1-\frac{g_R}{2g_A g_G}P_1-\frac{1}{2g_R}Q\right)^{-1/a}+\frac{g_T g_C}{g_Y}\left(1-\frac{g_Y}{2g_T g_C}P_2-\frac{1}{2g_Y}Q\right)^{-1/a}\right.$$
$$\left.+\left(g_R g_Y-\frac{g_A g_G g_Y}{g_R}-\frac{g_T g_C g_R}{g_Y}\right)\left(1-\frac{1}{2g_R g_Y}Q\right)^{-1/a}-g_A g_G-g_T g_C-g_R g_Y\right] \quad (3.30)$$

Мы предполагали, что гамма параметр $\boldsymbol{a}$ известен, но на практике нужно оценить $\boldsymbol{a}$ используя одновременно множество последовательностей. Существует несколько методов оценки $\boldsymbol{a}$, но наиболее предпочтительным является метод наибольшего правдоподобия.

Гамма дистанции являются более реалистичными, чем не гамма дистанции, но у них и большая дисперсия. Поэтому они не обязательно дают лучшие результаты в филогенетических оценках до тех пор пока число рассматриваемых нуклеотидов велико. Для оценки длины ветвей дерева гамма дистанции обычно дают лучшие результаты.

3.5. Численные оценки эволюционных дистанций

Большинство из рассмотренных аналитических формул для расчета эволюционных дистанций основывается на относительно простых математических моделях нуклеотидных замен, и оценки схожи с оценками максимального правдоподобия при определенных допущениях. Однако, для оценки дистанций, основывающихся на сложных моделях, таких как, например, гамма Тамура-Ней, зачастую удобней производить численный расчет дистанций, используя метод наибольшего правдоподобия. Этот метод особенно полезен для оценок дистанций для многих пар последовательностей. Например, аналитическая формула для расчета гамма дистанции Тамуры-Нея требует информации о частотах нуклеотидов и гамма параметре $\boldsymbol{a}$. Частоты нуклеотидов обычно оцениваются при сравнении двух

последовательностей, в то время как значение ***a*** вычисляется из данных по дополнительным последовательностям.

Если использовать численный метод, тем не менее, возможно оценить частоты нуклеотидов и значение ***a*** из данных по исследуемым последовательностям в том случае, если их имеется в достаточном количестве. Другими словами, можно оценить все параметры замен, как и значения дистанций, максимизируя правдоподобие для данного набора данных и для данной топологии. Такой подход дает оценки дистанций для всех попарных сравнений последовательностей одновременно. Также возможно выбрать наиболее подходящую модель замен, используя тест отношения правдоподобия[28].

Теоретически этот метод дает хорошие оценки дистанций для последовательностей ДНК при большом числе исследуемых нуклеотидов (***n***) и более или менее корректной используемой топологии. Это верно для оценки длин ветвей или для оценки времен эволюции. Когда ***n*** относительно мало, оценки дистанций, полученные данным методом, не обязательно дают хорошие оценки истинной топологии дерева. Оценка топологии дерева является достаточно сложной статистической задачей, и простые методы измерения дистанции часто дают лучшие результаты.

[28] likelihood ratio test

4. ВЫРАВНИВАНИЕ НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ.

При оценке филогенетической дистанций мы предполагали, что две сравниваемые нуклеотидные последовательности не имеют ни инсерций, ни делеций, и могут сравниваться напрямую. На практике же число нуклеотидов в сравниваемых последовательностях отличается, и мы должны определить положение инсерций и делеций, чтобы выравнять две последовательности. И делеции, и инсерции вносят гэпы в выравнивание последовательностей ДНК. Если дивергенция последовательностей мала, то положение гэпов можно с относительной легкостью определить при визуальном сравнении последовательностей, особенно их белок-кодирующих областей. Однако, если степень дивергенции высока, или требуется одновременное выравнивание сразу нескольких последовательностей, то выравнивание является не такой простой задачей и требует применения компьютерных расчетов. Разработано несколько методов для выравнивания нуклеотидных и аминокислотных последовательностей.

4.1. Выравнивание двух последовательностей.

Рассмотрим две нуклеотидные последовательности

(1) ATGCGTCGTT (A1)

(2) ATCCGCGAT

Последовательность (1) состоит из десяти нуклеотидов, а последовательность (2) из девяти нуклеотидов. Таким образом, по крайней мере, один гэп должен быть введен в выравнивание этих последовательностей. Простейший метод выравнивания состоит в двумерном сравнении, приведенном на рис. 4.1. В этом сравнении (метод точечной

матрицы[29]) точки ставятся, когда нуклеотиды в последовательностях (1) и (2) идентичны. Если две последовательности идентичны, то точки выстроятся в диагональную прямую линию, если же последовательности идентичны за исключением одного гэпа в одной из последовательностей, диагональная линия сместится вниз или вверх в середине линии. Таким образом, можно идентифицировать гэп. На практике, обычно кроме гэпов наблюдаются несколько отличий между последовательностями, что затрудняет идентификацию гэпа. По этой причине было разработано несколько математических методов для получения разумного выравнивания.

Один из самых популярных методов выравнивания последовательностей был предложен Нидлманом и Вунщем в 1970 году. В этом методе сходство между двумя последовательностями измеряется посредством **индекса сходства**[30] и выбирается такое выравнивание двух последовательностей, при котором индекс сходства будет максимальным. Четырьмя годами позже Селлерс предложил другой метод, в котором измеряется коэффициент дистанции между двумя последовательностями (**дистанция выравнивания**[31]) и выбирается выравнивание, которое минимизирует эту дистанцию. Однако, оказалось, что эти два метода практически одинаковы и дают в большинстве случаев одинаковые результаты.

Рассмотрим две последовательности А и Б длиной ***m*** и ***n***, соответственно. **Выравнивание между последовательностями** А и Б определяется как упорядоченная последовательность нуклеотидных пар, каждая из которых содержит по одному нуклеотиду из каждой последовательности, или же один нуклеотид из любой из последовательностей и нулевой элемент в таком порядке, чтобы сохранялись исходные последовательности. Делеции или инсерции (гэпы) обозначаются

[29] dot matrix
[30] similarity index
[31] alignment distance

прочерком (-) в парах, содержащих нулевой элемент. Например, нижеприведенное выравнивание

ATGC-GTCGTT (A2)

AT-CCG-CGAT

содержит три гэпа из одного элемента (длиной единица), семь пар совпадающих друг с другом элементов[32] и одну пару несовпадающих элементов[33].

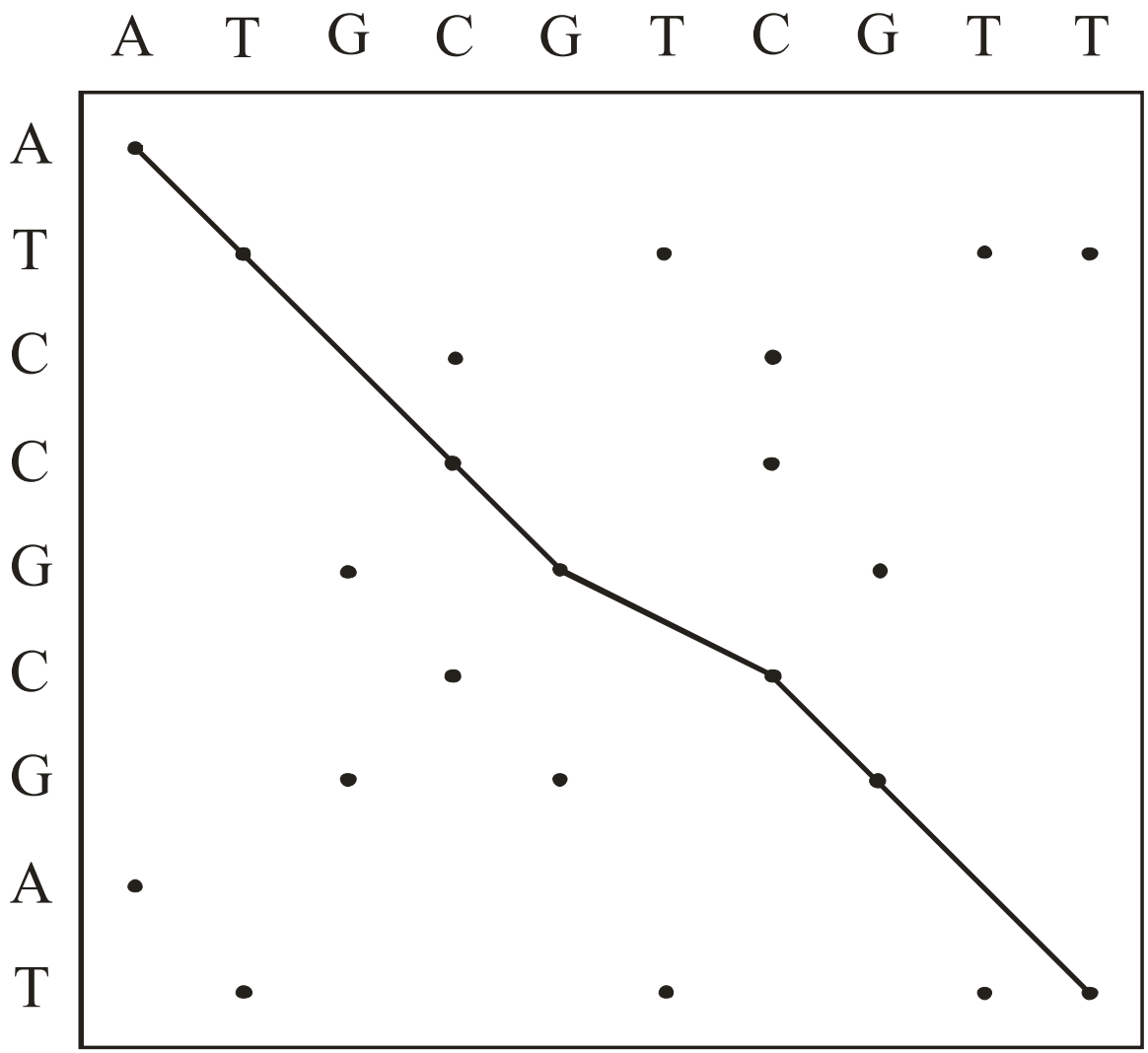

Рис. 4.1. Выравнивание двух последовательностей методом точечной матрицы.

[32] matched elements
[33] mismatched elements

Для измерения дистанции между двумя последовательностями, обозначим число пар совпадающих элементов через α, число пар несовпадающих элементов через β, а число гэпов, независимо от их длины, через γ. Дистанция между последовательностями может быть измерена как

$$E = Min(w_1\beta + w_2\gamma) \qquad (4.1)$$

где $Min(\bullet)$ соответствует наименьшему значению $w_1\beta + w_2\gamma$ среди всех возможных выравниваний. Здесь w_1 и w_2 представляют собой штраф[34] за несовпадение или гэп, соответственно. Однако предположение, что штраф за гэп будет одинаковым в независимости от его длины, представляется маловероятным. Поэтому разумно предположить, что штраф за гэп является функцией его длины. Аналогично несоответствия нуклеотидов могут быть разделены на несоответствия транзиций и трансверсий, и им можно приписать разные штрафы. Для аминокислотных последовательностей разным аминокислотным несоответствиям приписываются разные штрафы, основывающиеся на матрицах аминокислотных замен, составленных М.Дейхофф. Расчет E является довольно сложным процессом, однако были разработаны различные алгоритмы для быстрого компьютерного расчета.

Рассмотрим, как рассчитать дистанцию выравнивания с $w_1 = 1$ и $w_2 = 4$ в уравнении 4.1. Для выравнивания (А2) получаем $E = 1\times1 + 4\times3 = 13$, так как $\beta = 1$ и $\gamma = 3$. Для выравнивания вида

ATGCGTCGTT (A3)
ATCCG-CGAT

$\beta = 2$ и $\gamma = 1$, следовательно $E = 1\times2 + 4\times1 = 6$. Это значение меньше, чем значение E для выравнивания (А2), и, таким образом, выравнивание (А3) считается лучше, чем выравнивание (А2). Выбор наилучшего выравнивания во многом зависит от относительных значений w_1 и w_2. Если мы будем использовать маленькое значение w_2 относительно w_1, то получим выравнивание с большим количеством гэпов и небольшим количеством

[34] penalty

несовпадающих нуклеотидов. Однако, так как делеции и инсерции происходят гораздо менее часто, чем нуклеотидные замены, то такое выравнивание будет нереалестическим. Поэтому советуется использовать значения w_2 большие, чем значения w_1.

4.2. Выравнивание нескольких последовательностей.

Когда надо выравнять несколько последовательностей обычно используют **прогрессивное множественное выравнивание**[35]. В этом алгоритме сначала выравниваются пары последовательностей с маленькими расстояниями, и выравнивание наиболее удаленных последовательностей делается постепенно для больших и больших групп. На первом шаге, все последовательности попарно выравниваются по вышеописанному механизму. На следующем этапе неободимо выравнять группы последовательностей друг с другом. Это делается по алгоритму **профильного выравнивания**[36], который схож с выравниванием двух последовательностей, за исключением того, что средние дитанции теперь считаются рассматривая все нуклеотиды (или аминокислоты) в каждом положениидвух групп последовательностей. Если имеется гэп, то он вставляется во все последовательности этой группы.

В алгоритме прогрессивного множественного выравнивания порядок, в котором последовательности подвергаются выравниванию, является ключевым. В подходе, разработанном Хиггинсом в 1996 году, этот порядок определяется выводом дерево-подобного отношения последовательностей, основанного на матрице счетов попарных дистанций[37] между последовательностями. Для этого сначала оцениваются счета дистанций E и строится филогенетическое дерево по методу связывания ближайших

[35] progressive multiple alignment
[36] profile alignment algorithm
[37] pairwise distance scores

соседей[38]. Алгоритм прогрессивного множественного выравнивания, как правило, дает быстрые и разумные результаты, однако все равно нет гарантии того, что полученное выравнивание будет наилучшим.

Для белок-кодирующих последовательностей ДНК выравнивание на уровне аминокислот обычно является более правдоподобным, чем на уровне нуклеотидов, так как аминокислотные последовательности эволюционируют гораздо медленнее. В этом случае нуклеотидные последовательности можно выравнивать после выравнивания соответствующих аминокислотных последовательностей. Информация о вторичной структуре и более высоких уровнях организации белков или РНК также полезна при множественном выравнивании. Например, вторичная структура рибосомальных РНК часто используется для выравнивания последовательностей из филогенетически удаленных организмов, таких как животные, растения, грибы и бактерии.

4.3. Трактовка гэпов при оценке эволюционных дистанций.

Присутствие гэпов в последовательностях при выравнивании вносит некоторые сложности при оценке эволюционных дистанций. Более того, в последовательностях могут присутствовать сайты, последовательность которых неизвестна (например, неотсеквенированные участки), и они создают такие же проблемы как и гэпы. Эти сайты обычно выбрасываются из рассмотрения при оценке дистанций, но это можно сделать двумя способами. Первый способ – это просто исключить эти сайты из анализа. Такой вариант называется опция полного удаления гэпов[39]. Обычно лучше всего использовать его, так как разные области ДНК или белка часто эволюционируют по разному. Однако, если число нуклеотидов в гэпах мало, и гэпы распределены более или менее случайным образом, то можно рассчитать дистанцию между каждой парой последовательностей игнорируя

[38] neighbour-joining
[39] complete-deletion option

только гэпы, имеющиеся в двух сравниваемых последовательностях. Этот вариант называется опцией попарного удаления гэпов[40].

Чтобы проиллюстрировать эти два способа расчета дистанций, рассмотрим три следующие последовательности.

A-AC-GGAT-AGGA-ATAAA

AT-CC?GATAA?GAAAAC-A

ATTCC-GA?TACGATA-AGA

Здесь гэпы обозначены дефисами, а сайты с неизвестной последовательностью обозначены восклицательными знаками. В табл. 4.1 приведены результаты расчетов с использованием обоих вариантов учета гэпов. При втором варианте все гэпы и сайты с неизвестной последовательностью удалены, и мы сравниваем всего 10 сайтов, тогда ***p***-дистанции между последовательностями 1 и 2, 1 и 3, и 2 и 3 становятся равными 0.1, 0.0 и 0.1, соответственно. При первом варианте число сравниваемых нуклеотидов варьируется у разных пар последовательностей, и ***p***-дистанция также варьируются по парам последовательностей.

Таблица 4.1. Полное и попарное удаление гэпов.

Вариант удаления гэпов	Последовательности	Отличия/Число сравниваемых нуклеотидов (1, 2)	(1, 3)	(2, 3)
Полное удаление		1/10	0/10	1/10
	(1) A C GA A GA A A A			
	(2) A C GA A GA A C A			
	(3) A C GA A GA A A A			
Попарное удаление		2/12	3/13	3/14
	(1) A-AC-GGAT-AGGA-ATAAA			
	(2) AT-CC?GATAA?GAAAAC-A			
	(3) ATTCC-GA?TACGATA-AGA			

[40] pairwise-deletion option

5. СИНОНИМИЧНЫЕ И НЕСИНОНИМИЧНЫЕ НУКЛЕОТИДНЫЕ ЗАМЕНЫ

В главе 3 мы видели, что скорость нуклеотидных замен в третьем положении кодона значительно выше, чем в первом и втором положениях. Это вызвано тем, что большинство замен по третьему положению кодона являются молчащими и не приводят к замене аминокислоты. Однако, не все замены в третьем положении кодона молчащие. Более того, некоторые молчащие замены могут происходить и в первом положении кодона. Поэтому представляется интересным знать скорости отдельно синонимических и несинонимических нуклеотидных замен. Так как синонимические замены, видимо, свободны от естественного отбора, их скорость часто приравнивается к скорости нейтральных нуклеотидных замен. В самом деле, скорость синонимических замен является одинаковой для многих генов до тех пор, пока она не подвергается влиянию смещению использования кодонов или другим факторам. Скорость несинонимических замен, наоборот, во многом ниже, чем синонимических и сильно варьирует от гена к гену, что объясняется действием очищающего отбора.

Однако, существуют некоторые гены, в которых несинонимические замены происходят с более высокой скоростью, чем синонимические. Такие несинонимические замены, видимо, вызваны действием положительного Дарвиновского отбора, потому что при нейтральной эволюции ожидается равенство скоростей синонимических и несинонимических нуклеотидных замен. По этим причинам, оценка скоростей синонимических и несинонимических нуклеотидных замен играет важную роль при изучении молекулярной эволюции.

Алгоритм оценки скоростей синонимических и несинонимических замен является более сложным, чем для оценки общего числа нуклеотидных замен. В большинстве нуклеотидных последовательностей имеется больше число сайтов, в которых потенциально происходят несинонимические

мутации, чем синонимические, при чем оно варьирует от гена к гену. По этой причине скорости синонимических и несинонимических нуклеотидных замен определяются, соответственно, как число синонимических замен на синонимический сайт r_S и число несинонимических замен на несинонимический сайт r_N в год или на поколение. На практике мы обычно не знаем время дивергенции между двумя сравниваемыми последовательностями ДНК, поэтому, считают число синонимических замен на синонимический сайт $d_S = 2r_S t$ и число несинонимических замен на несинонимический сайт $d_N = 2r_N t$ для пары последовательностей.

Существует три основные группы методов оценки d_S и d_N: (1) методы эволюционных путей, (2) методы, основанные на 2-параметрической модели Кимуры, и (3) методы наибольшего правдоподобия с моделями замены кодонов. Эти методы основаны на различных допущениях, и, следовательно, они не обязательно будут давать одинаковые результаты. В этой главе мы детально рассмотрим первые две группы методов, так как они наиболее часто используются в литературе. Все рассуждения будут относиться к стандартному генетическому коду, но они могут быть легко применены для любого из известных генетических кодов.

5.1. Методы эволюционных путей

В методе эволюционных путей рассматриваются все возможные эволюционные пути между каждой парой гомологичных кодонов двух последовательностей ДНК и разработали метод оценки d_S и d_N. Однако предложенный метод является достаточно сложным, так как каждой нуклеотидной замене придается вес, который определяется вероятностью появления замены, учитывающей схожесть кодируемых аминокислот. Однако, на основе компьютерной имитации было показано, что приписывание веса разным путям не является необходимым и обычная

версия без учета весов дает практически те же результаты. Поэтому мы рассмотрим невзвешенный метод эволюционных путей, предложенный Нэем и Гожобори, и его модификации.

5.1.2. Метод Нея-Гожобори[41]

В этом методе d_S и d_N оцениваются путем подсчета числа синонимических и несинонимических замен и числа потенциально синонимических и несинонимических сайтов. Сначала рассмотрим число потенциально синонимических и несинонимических сайтов. В этом методе эти числа подсчитываются для каждого кодона при допущении равной вероятности всех нуклеотидных замен. Обозначим через f_i долю синонимических замен (отношение числа синонимических замен к сумме синонимических и несинонимических замен, включая нонсенс мутации) в i-ом положении кодона (i = 1, 2, 3). Числа потенциально синонимических **s** и несинонимических **n** сайтов для этого кодона даются выражениями $s = \sum_{i=1}^{3} f_i$ и $n = 3 - s$, соответственно. Например, в случае кодона для фенилаланина TTT, **s** становится равным

$$s = 0 + 0 + \frac{1}{3} \qquad (5.1)$$

потому что ни одна из замен в первом или во втором положениях не приводит к появлению синонимического кодона, а в третьем положении одна из трех возможных замен приводит к появлению синонимического кодона (TTC). Так как все остальные замены несинонимические, n равно 3 – 1/3 = 8/3. Если какая-нибудь из замен приводит к появлению терминирующего кодона, то такой заменой пренебрегают. Например, нуклеотидные замены в третьем положении цистеинового кодона TGT приводят к появлению терминирующего кодона при замене T на A, синонимического кодона при

[41] Nei and Gojobori method

замене T на C и несинонимического кодона (Trp) при замене T на G. В этом случае $f_3 = 1/2$, а так как $f_1 = f_2 = 1/2$ для этого кодона, мы получаем $s = 0.5$ и $n = 2.5$.

Для того чтобы получить полное число синонимических S и несинонимических N сайтов для всей последовательности, используются формулы $S = \sum_{j=1}^{C} s_j$ и $N = 3C - S$, где s_j – это значение s для j-го кодона, а C – это полное число кодонов. На практике обычно сравниваются две последовательности, поэтому в непосредственных вычислениях используются средние значения S и N для двух последовательностей. Заметим, сто $S + N = 3C$ равно общему числу сравниваемых нуклеотидов.

Теперь рассмотрим число синонимических и несинонимических нуклеотидных различий между парой гомологичных последовательностей. Для этого сравниваются две последовательности, кодон за кодоном, и считается число нуклеотидных отличий для каждой пары сравниваемых кодонов. Обозначим числа синонимических и несинонимических различий на кодон как s_d и n_d, соответственно. Если имеется только одна нуклеотидная замена, можно сразу решить, является ли она синонимической или несинонимической. Например, между кодонами GTT (Val) и GTA (Val), имеется одна синонимическая замена, тогда для них $s_d = 1$ и $n_d = 0$. В случае двух нуклеотидных различий между сравниваемыми кодонами, существует два возможных пути их оценки. Например, для кодонов TTT и GTA имеется два пути минимального числа замен между кодонами:

(1) TTT (Phe) ↔ GTT (Val) ↔ GTA (Val)

(2) TTT (Phe) ↔ TTA (Leu) ↔ GTA (Val)

Путь (1) включает в себя одну синонимическую и одну несинонимическую замены, тогда как путь (2) включает в себя две несинонимические замены. Мы полагаем, что пути (1) и (2) происходят с равной вероятностью. Тогда число синонимических и несинонимических различий становится равным $s_d = 0.5$ и $n_d = 1.5$, соответственно. Если при сравнении кодонов

встречаются пути, в которые вовлечены стоп-кодоны, то такие пути выбрасываются из расчетов.

В случае трех нуклеотидных отличий между сравниваемыми кодонами существует шесть различных возможных путей замен между ними, при этом в каждый путь содержит по три мутационных шага. Учитывая все эти пути и мутационные шаги, также как и для случая двух замен можно рассчитать число синонимических и несинонимических отличий. Например, если два сравниваемых кодона, это TTG и AGA, то имеется шесть следующих путей замен:

(1) TTG (Leu) ↔ ATG (Met) ↔ AGG (Arg) ↔ AGA (Arg)

(2) TTG (Leu) ↔ ATG (Met) ↔ ATA (Ile) ↔ AGA (Arg)

(3) TTG (Leu) ↔ TGG (Trp) ↔ AGG (Arg) ↔ AGA (Arg)

(4) TTG (Leu) ↔ TGG (Trp) ↔ TGA (Stop) ↔ AGA (Arg)

(5) TTG (Leu) ↔ TTA (Leu) ↔ ATA (Ile) ↔ AGA (Arg)

(6) TTG (Leu) ↔ TTA (Leu) ↔ TGA (Stop)↔ AGA (Arg)

Пути (4) и (6) содержат стоп-кодоны, поэтому они не рассматриваются. Число синонимических замен в путях (1), (2), (3) и (5) равно 1, 0, 1 и 1, соответственно, а число несинонимических замен равно 2, 3, 2 и 2, соответственно. Так как мы полагаем равную вероятность всех четырех путей, получаем $s_d = 3/4$ и $n_d = 9/4$.

Полное число синонимических и несинонимических отличий для сравниваемых последовательностей может быть получено суммированием этих значений по всем кодонам:

$$S_d = \sum_{j=1}^{C} s_{dj} \text{ и } N_d = \sum_{j=1}^{C} n_{dj} \qquad (5.2)$$

где s_{dj} и n_{dj} – это число синонимических и несинонимических отличий для j-го кодона, а C – это число сравниваемых кодонов. Заметим, что $S_d + N_d$ равно полному числу нуклеотидных отличий между двумя сравниваемыми последовательностями ДНК.

Таким образом, можно оценить соотношение синонимических p_S и несинонимических p_N отличий следующими уравнениями:

$$\hat{p}_S = S_d / S,\ \hat{p}_N = N_d / N \qquad (5.3)$$

где S и N – это среднее число синонимических и несинонимических сайтов для двух сравниваемых последовательностей. Для оценки числа синонимических $\hat{d}_S$ и несинонимических $\hat{d}_N$ замен на сайт используется метод Джукса и Кантора (уравнение 3.7), в котором p заменяется на $\hat{p}_S$ или $\hat{p}_N$. Этот метод, конечно, дает только приблизительные оценки d_S и d_N, так как нуклеотидные замены в синонимических и несинонимических сайтах, на самом деле, не следуют модели Джукса и Кантора. Несмотря на эту теоретическую проблему, компьютерные симуляции показали, что уравнение 3.7 дает хорошие оценки синонимических и несинонимических замен при условии, что частоты нуклеотидов A, T, C и G примерно одинаковы и нет значительного смещения транзиций/трансверсий[42].

Примерные дисперсии большой выборки[43] $\hat{d}_S$ и $\hat{d}_N$ могут быть посчитаны по уравнению 3.8, если заменить $\hat{p}$ на $\hat{p}_S$ или $\hat{p}_N$ и n на S или N. Теоретически более аккуратные дисперсии большой выборки [$V(\hat{d}_S)$ и $V(\hat{d}_N)$] $\hat{d}_S$ и $\hat{d}_N$ даются формулами

$$V(\hat{d}_S) = V(\hat{p}_S) \Big/ \left(1 - \frac{4}{3} p_S\right)^2,\ V(\hat{d}_N) = V(\hat{p}_N) \Big/ \left(1 - \frac{4}{3} p_N\right)^2 \qquad (5.4)$$

где

$$V(\hat{p}_S) = \sum_{i=1}^{C} (s_{di} - p_S s_i)^2 / S^2,\ V(\hat{p}_N) = \sum_{i=1}^{C} (n_{di} - p_N n_i)^2 / N^2 \qquad (5.5)$$

Однако имитация с помощью компьютера показала, что вышеприведенные формулы дают примерно те же результаты, что и уравнение 3.7.

---

[42] transition/transversion bias
[43] large-sample variance

5.1.3. Модифицированный метод Нея-Гожобори

В методе Нея-Гожобори предполагается, что замены нуклеотидов происходят случайным образом. На практике это предположение не всегда выполняется, и скорость замен по типу транзиции обычно выше, чем по типу трансверсии. В этом случае число (***S***) потенциальных сайтов, в которых могут произойти синонимические замены, ожидается быть большим, чем при расчете методом Нея-Гожобори, так как большинство транзиций в третьем положении кодонаявыляются синонимическими. По этой причине метод Нея-Гожобори может давать переоценку p_S и d_S и недооценку p_N и d_N.

Поэтому Ина в 1995 году предложил метод оценки d_S и d_N с использованием 2-параметрической модели Кимуры. В модели Кимуры скорости транзиций и трансверсий определяются как α и β, соответственно, но, так как любой нуклеотид может претерпеть два разных изменения по типу трансверсии, доля транзиций среди всех изменений будет даваться выражением

$$\frac{\alpha}{\alpha+2\beta}=\frac{R}{1+R} \qquad (5.6)$$

где $R=\frac{\alpha}{2\beta}$ – это соотношение числа транзиций к числу трансверсий, и R становится равным 0.5, если нет систематических отклонений. Ина показал, что ожидаемое число синонимических замен на кодон можно выразить посредством R для всех кодонов. Например, для кодона TTT это значение определяется как

$$s=0+0+\frac{\alpha}{\alpha+2\beta}=\frac{R}{1+R} \qquad (5.7)$$

потому что в этом случае только третье положение в кодоне может вызвать синонимическую замену и только одна ($T\rightarrow C$) из трех возможных замен является синонимической. В другом случае, например, для кодона CTA (Leu)

$s = \frac{R}{1+R} + 1$, потому что замены в первом, втором и третьем положениях в кодоне могут быть синонимическими с вероятностями $\frac{R}{1+R}$, 0 и 1, соответственно. В этих расчетах нонсенс мутации не учитываются, как и ранее.

Очевидно, что, зная R, можно рассчитать s для всех кодонов, а затем оценить S и N ($= 3C - S$). Но как оценить R для реального набора данных? R можно оценить, например, из уравнений 3.2 или 3.17, или же исходя из другой используемой информации. В рассматриваемом методе ожидается, что S будет увеличиваться, а N – уменьшаться по сравнению со значениями, полученными в исходном методе Нея-Гожобори. Обозначим новые S и N через S_R и N_R, соответственно. В отличие от S и N систематическое отклонение транзиции/трансверсии не сильно влияет на число синонимических (S_d) и несинонимических (N_d) отличий, потому что S_d и N_d основываются на действительном числе наблюдаемых замен. Поэтому отношение синонимических ($\hat{p}_S$) и несинонимических ($\hat{p}_N$) отличий дается теперь выражением

$$\hat{p}_S = \frac{S_d}{S_R},\ \hat{p}_N = \frac{N_d}{N_R} \qquad (5.8)$$

в то время как оценки ($\hat{d}_S$ и $\hat{d}_N$) d_S и d_N вновь примерно даются формулой Джукса-Кантора. Теоретически, существует лучший путь оценки d_S и d_N, , но практически нет большой разницы между оценками, полученными этими двумя методами, если d_S и d_N не очень высоки (Когда $d_S > 1.0$ и $d_N > 1.0$, достоверность $\hat{d}_S$ и $\hat{d}_N$ достаточно низкая, потому что реальный процесс синонимических и несинонимических замен очень сложный). Более того, настоящие методы дают меньшие вариации $\hat{p}_S$, $\hat{p}_N$, $\hat{d}_S$ и $\hat{d}_N$, чем при методе Ина.

Хотя модифицированный метод Нея-Гожобори теоретически лучше, чем его оригинальная версия, при применении модели Кимуры с большим значением ***R***, следует учитывать, что при недостоверной оценке ***R***, он может привести к ложным выводам. Особенно, при использовании переоцененного ***R*** модифицированная версия может привести к значительному превосходству $\hat{d}_N$ над $\hat{d}_S$, даже когда это абсолютно не так. Надо учитывать, что действительный паттерн нуклеотидных замен гораздо сложнее, чем в модели Кимуры, и при определенных условиях модифицированный метод Нея-Гожобори даст переоценку ***S*** и недооценку ***N***. Поэтому всегда лучше использовать оба метода для определения положительного отбора. Если оригинальный метод укажет положительный Дарвиновский отбор, выводы будут более безопасными.

5.2. Методы, основанные на 2-параметрической модели Кимуры.

5.2.1. Метод Ли-Ву-Луо[44]

Ли и соавт. разработали в 1985 году другой метод, основанный на 2-параметрической моделе Кимуры. Они заметили, что при учете вырожденности генетического кода, нуклеотидные сайты кодонов, за редким исключением (например, кодоны изолейцина), могут быть классифицированы на 4-хкратно вырожденные, 2-хкратно вырожденные и 0-кратно вырожденные (невырожденные). Сайт называется 4-хкратно вырожденным, если все возможные изменения в нем синонимические, 2-хкратно вырожденным, если одно из трех возможных изменений синонимическое, и невырожденным, если все изменения являются несинонимическими или нонсенс мутациями. Например, третьи положения в

[44] Li-Wu-Luo method

кодонах валина являются 4-хкратно вырожденными сайтами, а вторые положения всех кодонов – 0-кратно вырожденными сайтами. Третьи положения трех кодонов изолейцина являются 3-хкратно вырожденными сайтами, но они принимаются за 2-хкратно вырожденные для упрощения расчетов.

Таким образом, можно рассчитать количество трех типов сайтов для каждой из двух последовательностей и обозначить среднее число 0-кратно, 2-хкратно и 4-хкратно вырожденных сайтов для двух сравниваемых последовательностей как L_0, L_2 и L_4, соответственно. Затем две последовательности сравниваются, кодон за кодоном, и каждое нуклеотидное отличие классифицируется как транзиция или трансверсия. Если обозначить доли транзиций и трансверсий в i-ом классе нуклеотидных сайтов через P_i и Q_i (i = 0, 2 или 4), то можно оценить число транзиций A_i и трансверсий B_i на сайт для каждого из трех классов нуклеотидных сайтов:

$$A_i = \frac{1}{2}\ln(a_i) - \frac{1}{4}\ln(b_i) \qquad B_i = \frac{1}{2}\ln(b_i) \tag{5.9}$$

где $a_i = 1/(1-2P_i-Q_i)$ и $b_i = 1/(1-2Q_i)$.

Можно заметить, что все замены в 4-хкратно вырожденных сайтах являются синонимическими, а все замены в 0-кратно вырожденных сайтах – несинонимическими. В 2-хкратно вырожденных сайтах транзиции в основном синонимические, а трансверсии в основном несинонимические. Допуская одинаковые частоты нуклеотидных замен для четырех нуклеотидов A, T, C и G, можно считать, что треть всех 2-хкратно вырожденных сайтов являются потенциально синонимическими, а две трети – несинонимическими. С учетом этого допущения можно оценить d_S и d_N как

$$\hat{d}_S = \frac{3[L_2A_2 + L_4(A_4 + B_4)]}{L_2 + 3L_4} \tag{5.10}$$

$$\hat{d}_N = \frac{3[L_0(A_0 + B_0) + L_2B_2]}{3L_0 + 2L_2} \tag{5.11}$$

Эти формулы зависят от ряда допущений, которые не всегда выполняются для исследуемого набора данных. Во-первых, типы гомологичных нуклеотидных сайтов могут отличаться в двух сравниваемых последовательностях. Такое случается достаточно часто в случае высокой степени дивергенции последовательностей. В этом случае, половина сайта считается одного класса, а вторая половина – другого класса. Во-вторых, нонсенс мутации считаются как несинонимические изменения. Например, замена в третьем положении тирозинового кодона TAT может привести к появлению синонимического кодона (TAC) и двух нонсенс кодонов (TAA и TAG), и две последние замены рассматриваются как несинонимические. Так как нонсенс мутации появляются с вероятностью около 4 % (глава 1), этот метод будет давать переоценку $\boldsymbol{d}_N$. В-третьих, транзиции в первом положении четырех 2-хкратно вырожденных кодонов для аргинина (CGA, CGG, AGA и AGG) не являются синонимическими, они все несинонимические с одним исключением (CGA), приводящим к появлению нонсенс кодона. В третьем положении трех кодонов изолейцина, являющихся 3-хкратно вырожденными, некоторые трансверсии являются синонимическими. Несмотря на эти проблемы, метод Ли-Ву-Луо дает результаты, схожие с таковыми, полученными методом Нея-Гожобори, когда число кодонов велико, а дивергенция последовательностей мала. Однако, при малом числе исследуемых кодонов (скажем < 100) метод Ли-Ву-Луо может дать отрицательные оценки, так как $\boldsymbol{a}_i$ и $\boldsymbol{b}_i$ в уравнении 5.9 подвергаются большим ошибкам выборки.

5.2.2. Метод Памило-Биянчи-Ли[45].

Другой проблемой метода Ли-Ву-Луо является влияние систематического отклонения транзиций/трансверсиий, и вызываемая этим отклонением ошибка может быть существенной при больших $\boldsymbol{R}$, как в случае

[45] Pamilo-Bianchi-Li method

метода Нея и Гожобори. Поэтому в 1993 году Памило и Биянчи, и Ли независимо расширили метод Ли-Ву-Луо.

Учитывая, что синонимические замены по типу транзиции происходят только в 2-хкратно и 4-хкратно вырожденных сайтах в модели Ли-Ву-Луо, они предложили, что полное число этих изменений может быть оценено взвешенным средним $(L_2A_2 + L_4A_4)/(L_2 + L_4)$. Так как трансверсии в 4-хкратно вырожденных сайтах также синонимические, полное число синонимических замен на синонимический сайт теперь оценивается как.

$$\hat{d}_S = (L_2A_2 + L_4A_4)/(L_2 + L_4) + B_4 \qquad (5.12)$$

Аналогично, $\hat{d}_N$ может быть оценена как

$$\hat{d}_N = A_0 + (L_0B_0 + L_2B_2)/(L_0 + L_2) \qquad (5.13)$$

5.2.3. Метод Комерона-Кумара[46].

Как отмечалось ранее, трактовка кодонов аргинина и изолейцина в методе Ли-Ву-Луо является неаккуратной. Это же применимо и для метода Памило-Биянчи-Ли. Поэтому возникает проблема, когда в последовательности имеется много таких аминокислот. Например, в протамине P1 млекопитающих около 50 % аминокислот являются аргининами. Комерон пытался решить эту проблему путем разделения 2-хкратно вырожденных сайтов на две группы: 2S-кратно и 2V-кратно вырожденные сайты. Первыми являются сайты, в которых транзиция является синонимической, а две трансверсии – несинонимические, а вторыми – сайты, где транзиция является несинонимической, а две трансверсии – синонимические. Такая классификация позволяет скорректировать нечеткости классификации синонимических и несинонимических сайтов (например, кодоны метионина) в методе Ли-Ву-Луо.

[46] Comeron and Kumar method

Однако, это не решает всех проблем. Например, мутация в первом положении кодона аргинина CGG приводит к образованию кодонов TGG (Trp), AGG (Arg) и GGG (Gly). В этом случае транзиция (C→T) приводит к несинонимической замене, в то время как одна трансверсия (C→A) приводит к синонимической замене, а другая трансверсия (C→G) является несинонимической. Поэтому этот нуклеотидный сайт не является ни 2S-кратно, ни и 2V-кратно вырожденным. Аналогично, первое положение трех кодонов аргинина (CGU, CGC и CGA) и третье положение двух кодонов изолейцина (ATT и ATC) не могут быть отнесены ни к одной из категорий в модели Comeron.

Учитывая эти проблемы, С.Кумар разработал другой вариант метода Памило-Биянчи-Ли. В нем нуклеотидные сайты классифицируются на 0-кратно, 2-хкратно и 4-хкратно вырожденные, а 4-хкратно вырожденные и 2-хкратно вырожденные сайты далее разделяются на просто 2-хкратно и сложно 2-хкратно вырожденные сайты. Просто 2-хкратно вырожденными сайтами являются те сайты, в которых транзиция приводит к синонимической замене, а две трансверсии вызывают несинонимические замены или нонсенс мутации. Все остальные 2-хкратно вырожденные сайты, включая три вышеописанных кодона изолейцина, относятся к сложным 2-хкратно вырожденным сайтам. Используя эту классификацию, С.Кумар разработал новый метод оценки d_S и d_N.

5.2.4. Метод Ина[47].

В 1995 году Ина разработал еще один метод оценки d_S и d_N, объединяющий некоторые черты оригинального метода Нея-Гожобори и метода Памило-Биянчи-Ли. Он предложил 2 метода: метод I и метод II. В методе I отношение скоростей транзиций к трансверсиям $k = \alpha/\beta$ оценивается уравнением 3.17 с использованием только данных по третьему

[47] Ina's method

положению кодона. Это основывается на допущении, что нуклеотидные замены в третьем положении кодона в основном нейтральны. В этом случае S и N оцениваются с использованием процедуры модифицированного метода Нея-Гожобори, тогда как S_d и N_d рассчитываются по методу Нея-Гожобори. Однако Ина разделил S_d на синонимические транзиции S_{Ts} и синонимические трансверсии S_{Tv}, а N_d – на несинонимические транзиции N_{Ts} и синонимические трансверсии N_{Tv}. Он затем оценил $\hat{d}_S$ и $\hat{d}_N$ по формулам, аналогичным уравнениям 5.12 и 5.13. В методе II S и N оцениваются исходя из данных по всем трем положениям в кодоне, но α и β оцениваются только с использованием синонимических замен, чтобы отразить скорости мутаций перед отбором.

Компьютерные симуляции показали, что метод II дает чуть более аккуратные оценки d_S и d_N, чем метод I при большом числе нуклеотидов. Однако, различия в $\hat{d}_S$ и $\hat{d}_N$ между двумя методами или между методами Ина и модифицированным Нея-Гожобори обычно малы. Более того, когда число рассматриваемых нуклеотидов мало и дивергенция последовательностей мала, метод Ина может быть неприменим, потому что транзиций или трансверсий может не быть вовсе, а это делает оценки α/β равными 0 или ∞. Поэтому необходимо проявлять осторожность при использовании метода Ина.

5.3. Нуклеотидные замены в разных положениях кодона.

При сравнении относительно близких видов ожидается, что число синонимических замен будет увеличиваться практически линейно со временем, так как они практически свободны от отбора. Однако с увеличение числа замен аккуратность оценок будет уменьшаться, потому что допущения, используемые для оценки числа синонимических замен, скорее всего не будут выполняться в течение долгого времени. Как отмечалось ранее,

синонимические и несинонимические сайты не фиксированы, а варьируются со временем. Поэтому некоторые авторы предпочитают использовать число нуклеотидных замен в третьем положение кодона для оценки эволюционных времен. В этих сайтах определенная доля замен несинонимическая, но нуклеотидные сайты отчетливо определяются и не изменяются со временем. Поэтому число замен в третьем положении может линейно зависеть от эволюционного времени.

На практике число синонимических замен на ген обычно больше, чем число замен в третьем положении. Так, была изучена зависимость между числом синонимических замен $\hat{d}_S$, полученным методом Нея-Гожобори, и числом замен в третьем положении кодона $\hat{d}_3$ для последовательностей гена алкоголь дегидрогеназы (*Adh*) из 14 разных видов *Drosophila*. Значения $\hat{d}_3$ были получены методом Таджимы и Нея, так как частоты нуклеотидов в третьем положении кодона существенно отличаются от 0.25. Результаты показывают, что $\hat{d}_S$, как правило, немного больше, чем $\hat{d}_3$, как и ожидалось, но для $\hat{d}_S < 0.8$ существует примерно линейная зависимость между $\hat{d}_S$ и $\hat{d}_3$. Поэтому в данном случае ни $\hat{d}_3$, ни $\hat{d}_S$ не могут быть использованы для оценки времени дивергенции при $\hat{d}_S < 0.8$.

Зависимость между числом несинонимических замен $\hat{d}_N$ и дистанциями Джукса-Кантора для первого и второго положения в кодоне $\hat{d}_{12}$ для тех же последовательностей гена *Adh*, показывает, что в данном случае значения $\hat{d}_N$ и $\hat{d}_{12}$ гораздо меньше, чем значения $\hat{d}_S$ и $\hat{d}_3$, но $\hat{d}_N$ и $\hat{d}_{12}$ примерно равны друг другу для всех сравниваемых последовательностей. Это указывает на возможность использования и $\hat{d}_N$, и $\hat{d}_{12}$ для оценки времени дивергенции. Ранее мы замечали, что число аминокислотных замен часто дает хорошую оценку времени дивергенции. Зависимость между дистанцией с коррекцией Пуассона $\hat{d}$ для аминокислотных

последовательностей и $\hat{d}_{12}$ опять же является линейной, хотя $\hat{d}$ больше, чем $\hat{d}_{12}$, как и ожидалось.

5.4. Методы правдоподобия с моделями замен в кодоне.

Голдман и Янг в 1994 году разработали метод правдоподобия для оценки скоростей синонимических и несинонимическх нуклеотидных замен, учитывающий модель нуклеотидных замен для 61 sense кодона. Их модель в чем-то похожа на модель HGY[48] (табл. 3.2E) для нуклеотидных замен. Рассмотрим пару последовательностей из C гомологичных кодонов и обозначим относительную частоту j-го кодона через π_j. Голдман и Янг предположили, что мгновенная скорость q_{ij} замены кодона i на кодон j ($i \neq j$) дается следующими уравнениями:

$$q_{ij} = \begin{cases} 0, & \textit{замена нуклеотида происходит в двух и более положениях кодона} \\ \pi_j, & \textit{для синонимических трансверсий} \\ k\pi_j, & \textit{для синонимических транзиций} \\ \omega\pi_j, & \textit{для несинонимических трансверсий} \\ \omega k\pi_j & \textit{для несинонимических транзиций} \end{cases}$$

где k – это отношение скоростей транзиций к трансверсиям, а ω - это отношение скоростей несинонимических к синонимическим заменам. k может быть записана как α/β, если скорости транзиций и трансверсий, соответственно, α и β. аналогично, ω может быть записано как r_N/r_S, если скорости синонимических и несинонимических изменений, соответственно, r_S и r_N. Поэтому, если ω одинаково для всех пар кодонов, как полагалось, возможно соотнести r_N/r_S к d_N/d_S.

Существует 61 параметр для π_j, но, если допустить, что частоты кодонов равновесны, то их можно оценить через наблюдаемые частоты кодонов, когда число используемых кодонов C велико. Поэтому, единственными параметрами, которые надо оценить, являются k и ω, и они

[48] Hasegawa-Kishino-Yano model

могут быть оценены методом наибольшего правдоподобия. Когда C относительно мало, однако, этот метод не дает правдоподобной оценки π_j, потому что π_j обычно очень мало, и, таким образом, ошибка выборки оценки π_j будет велика. В этом случае π_j можно оценить через продукт наблюдаемых нуклеотидных частот. При таком подходе $\omega < 1$, $\omega = 1$ и $\omega > 1$ представляют очищающий отбор, нейтральную эволюцию и положительный отбор, соответственно. Поэтому, если оценка ($\hat{\omega}$) ω, полученная из имеющихся данных, значительно больше, чем 1, предполагается действие положительного отбора. Теоретически, этот тест может быть проведен с использованием теста соотношения правдоподобия.

Пусть $\ln L_2$ будет *log* значения **наибольшего правдоподобия**[49] (ML), когда ω оценивается из данных, а $\ln L_1$ будет значением **ML**, когда $\omega = 1$ (нуль гипотеза). *Log* соотношения правдоподобия тогда будет равен

$$LR = 2(\ln L_2 - \ln L_1) \qquad (5.13)$$

Когда число синонимических и несинонимических замен достаточно большое и используется подходящая модель, LR примерно следует распределению χ^2 с одной степенью свободы. Поэтому, если $\hat{\omega} > 1$ и $LR \geq 3.84$, можно заключить, что скорость несинонимических замен значительно выше, чем синонимических замен на уровне 5%, и что это происходит из-за действия положительного отбора.

Одним из преимуществ такого подхода является то, что и k, и ω можно оценить одновременно, если модель, представленная в уравнении 5.12, выполняется. Поэтому, нет необходимости знать R (= $2k$) для оценки d_S и d_N, как в случае модифицированного метода Нея-Гожобори.

Однако существует и несколько проблем при таком подходе. Во-первых, оценки π_j, основанные на наблюдаемых частотах, не будут достоверными при малых C, как говорилось выше. Оценка π_j через

[49] maximum likelihood

продукты нуклеотидных частот также будет недостоверной, когда существует смещение использования кодонов. Во-вторых, допущение, что ω одинакова для всех положений кодонов, мало реалистично, что ясно видно из паттерна аминокислотных замен, обсужденного в главе 1. Поэтому $\hat{\omega}$ будет существенно отличаться от $\hat{d}_N / \hat{d}_S$, потому что среднее отношения r_N / r_S не будет равно отношению средних r_N и r_S. В-третьих, допущение независимости k и ω для каждой пары кодонов не будет выполняться для реальных данных. Поэтому необходимо более аккуратное изучение влияния нарушения допущений на $\hat{\omega}$.

Мьюс в 1996 году разработал похожий метод правдоподобия, основанный на другой модели замен в кодонах. В этом методе частоты кодонов оцениваются через продукты нуклеотидных частот, а смещение транзиции/трансверсии не учитывается. Поэтому число подлежащих оценке параметров меньше, чем в модели Голдмана-Янга. Этот метод будет давать $\hat{d}_S$ и $\hat{d}_N$ схожие с таковыми, полученными методом Нея-Гожобори, когда смещение использования кодонов мало. Когда это смещение и смещение транзиции/трансверсии велики, ожидается, что метод Мьюса даст смещенные оценки.

По мере развития компьютерных технологий становится возможным использовать все более сложные математические модели и проводить статистический анализ, основанный на этих моделях. Однако, по мере усложнения математической модели, требуется все больше параметров, и лежащие в основе допущения, вероятно, не будут выполняться. Сложная модель, поэтому, может давать смещенные оценки параметров. Напротив, методы эволюционных путей, описанные ранее, основываются на концепции анализа парсимоний и в основном свободны от моделей. Адаптивные аминокислотные замены обычно происходят в некоторых определенных сайтах из-за функциональных причин, и паттерн замен, вероятно, будет отличаться от общего паттерна аминокислотных замен. Особенно при

больших d_S и d_N (скажем, $d_S, d_N > 0.4$) эти методы будут давать менее правдоподобные оценки, чем простые методы эволюционных путей, так как существует множество возмущающих факторов, которые влияют на оценки d_S и d_N.

Другой проблемой подхода с использованием правдоподобия является достоверность теста отношения правдоподобий. Этот тест требует удовлетворения допущений математической модели реальным данным. В тесте эволюционных гипотез это требование часто не выполняется, и в этом случае тест может быть или слишком либеральным, или слишком консервативным в зависимости от ситуации. Следует заметить, что тест отношения правдоподобий является тестом больших выборок, поэтому он может давать ошибочные выводы, когда число синонимических и несинонимических замен мало. Поэтому этот тест необходимо применять с особым вниманием.

6. ФИЛОГЕНЕТИЧЕСКИЕ ДЕРЕВЬЯ.

Молекулярная филогения – изучение эволюционных взаимоотношений между организмами методами молекулярной биологии (“Phylon” – племя, в переводе с греческого). Филогенетический анализ последовательностей ДНК или белка стал важным инструментом изучения эволюционной истории организмов от бактерий до человека. Так как скорости эволюции последовательностей варьируются в зависимости от гена или сегмента ДНК, можно изучать эволюционные взаимоотношения практически всех уровней классификации организмов (царства, типы, семейства, роды, виды и внутривидовые популяции), используя разные гены и сегменты ДНК.

В основе молекулярной филогении лежит утверждение о том, что все живые организмы происходят от общего предка(ков), живших около 4 миллиардов лет назад. Целью молекулярной филогении является установление правильных генеалогических связей между организмами и времени дивергенции между организмами, то есть время, когда жил их последний общий предок.

6.1.Типы филогенетических деревьев

6.1.1. Укорененные и неукорененные филогенетические деревья.

Филогенетические взаимоотношения генов или организмов обычно представляются в форме деревьев, которые могут, как содержать (рис. 5.1А), так и не содержать (рис. 5.1Б), корень. Первое дерево называется **укорененное дерево**, а второе – **неукорененное дерево**. Ветвящийся паттерн дерева, укорененного или неукорененного, называется **топологией**. Существует множество возможных укорененных и неукорененных деревьевных топологий для счетного числа таксонов (таксон - систематическая группа любой категории), которыми могут быть, например,

семейство, вид, популяция, последовательность ДНК и т.п. Так, если число таксонов (***m***) равно четырем, то существует 15 возможных топологий укорененного дерева и три возможных топологии неукорененного дерева, которые изображены на рис. 6.1. Число возможных топологий резко возрастает с увеличением ***m***. В целом возможное число топологий бифуркационного укорененного дерева из m таксонов вычисляется по следующему уравнению:

$$1 \cdot 3 \cdot 5 \cdots (2m-3) = \frac{(2m-3)!}{2^{m-2}(m-2)!} \qquad (6.1)$$

для $m \geq 2$. Значения всех возможных топологий для разных ***m*** приведены в табл. 6.1.

Таблица 6.1. Возможное число топологий для разного количества таксонов.

Число таксонов	N_R	N_U
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
…	…	…
10	34459425	2027025

Так, при ***m*** = 10 число возможных топологий достигает значения 34459425, и только одна из этих топологий является истинной. Число возможных топологий для бифуркационного неукорененного дерева для ***m*** таксонов вычисляется путем замены ***m*** на ***m*** – 1 в формуле 6.1. Так при ***m*** = 10 число возможных топологий достигает значения 2027025. Во многих случаях большинство возможных топологий может быть исключено из рассмотрения, так как они представляют собой очевидно невероятные эволюционные

картины, или же исходя из другой биологической информации. Но все равно довольно сложно определить истинную топологию дерева при больших ***m***.

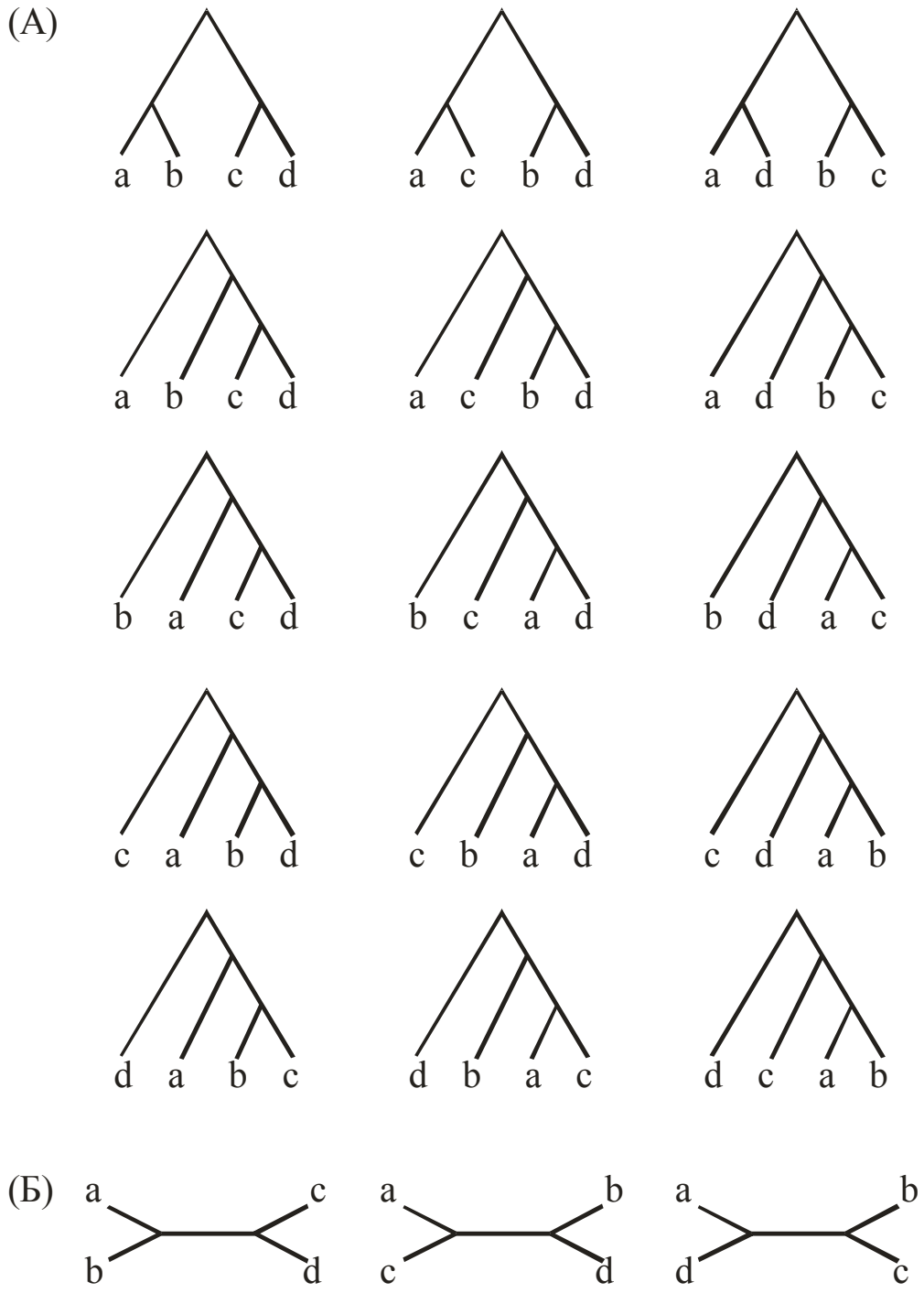

Рис. 6.1. (А) Пятнадцать возможных укорененных деревьев и (Б) три возможных неукорененных дерева для случая четырех таксонов.

В неукорененном бифуркационном дереве из ***m*** таксонов существует 2***m*** – 3 **ветвей**. Так как существует m внешних ветвей, ведущих к m внешним таксонам, то число внутренних ветвей равно ***m*** – 3. Число внутренних узлов

равно $m - 2$. В укорененном дереве число внутренних ветвей и внутренних узлов равно $m - 2$ и $m - 1$, соответственно, а полное число ветвей равно $2m - 2$ (рис. 6.2).

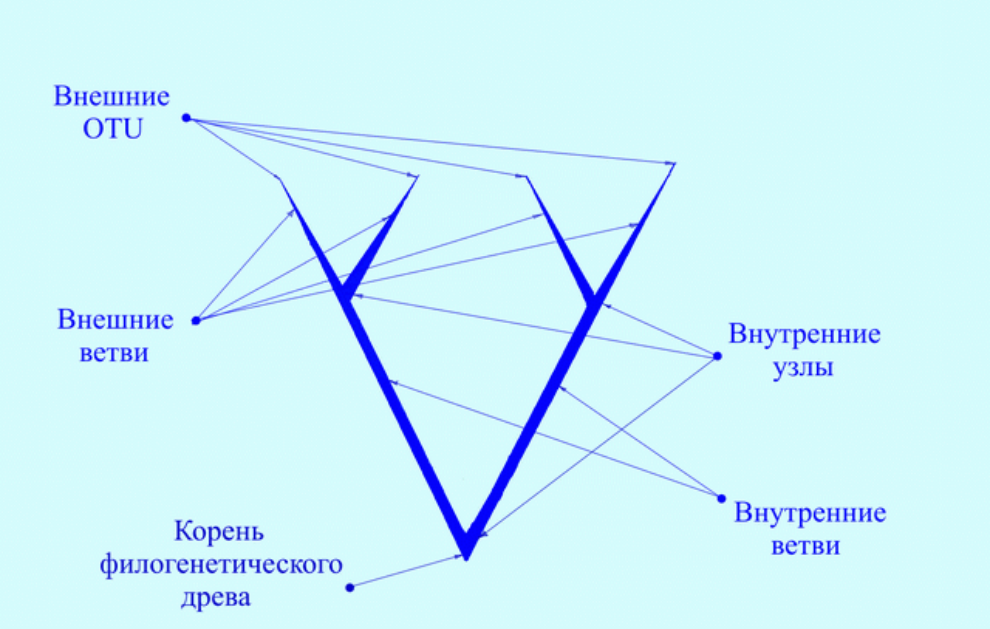

Рис. 6.2. Характеристики филогенетического дерева

Теоретически, последовательность ДНК расходится в две потомственные последовательности после дупликации или видообразования. Поэтому филогенетические деревья обычно **бифуркационные**. Однако когда рассматривается относительно короткая последовательность, между некоторыми внутренними ветвями может не быть ни одной замены, таким образом, может появиться мультифуркационный узел. Такое дерево будет называться **мультифуркационным**. Большинство методов построения деревьев предназначены для построения бифуркациооных деревьев, но полученное дерево может быть сведено к мультифуркационному, если исключить все ветви с нулевой длиной. Также возможно, что даже если истинное дерево бифуркациооное, построенное дерево окажется мультифуркационным из-за статистических ошибок. На практике эти два случая отличить довольно сложно.

6.1.2. Генные и видовые деревья.

Эволюционисты часто рассматривают филогенетические деревья, которые представляют эволюционную историю группы видов (или популяций). Такой тип деревьев называется **видовым** (или **популяционным**) **деревом**. В нем время дивергенции между двумя видами соответствует времени, когда два вида стали репродуктивно изолированными друг от друга. Однако, когда филогенетическое дерево строится по данным одного гена из каждого вида, полученное дерево не обязательно будет согласовываться с видовым деревом. В присутствии полиморфных аллелей в локусе, время дивергенции генов, взятых из разных видов, предполагается, будет больше, чем время дивергенции видов (рис. 6.3). Топология дерева, построенного по данным гена, также может отличаться от видового дерева. Поэтому такое дерево называют **генным деревом**. На рис. 6.4. показано три различных типа взаимоотношений между видовыми и генными деревьями для случая трех видов. На рис. 6.4А и Б топологии видового и генного деревьев совпадают, а на рисунке 6.4В не совпадают. Если использовать теорию генной генеалогии из популяционной генетики, то можно посчитать вероятность появления событий А, Б и В. Вероятность события, изображенного на рисунке 6.4В достаточно велика, когда временной интервал между расхождением первого и второго вида, выраженный в числе поколений T, короткий, и эффективный размер популяции N большой.

Предположим, что эффективный размер популяции N равен 10000, как в случае некоторых млекопитающих, и интервал между двумя событиями расхождения видов равен одному миллиону лет. Если время жизни одного поколения 5 лет, то T становится равным 200000 поколений. В этом случае вероятность [$P(B)$] события В равна:

$$[P(B)] = \frac{2}{3} e^{-\frac{T}{2N}} = 0{,}00003 \qquad (6.2)$$

Это значение практически равно 0. Если ***N*** большое и равно 100000, а время жизни одного поколения равно 1 год, как в случае некоторых беспозвоночных организмов, ***P(B)*** становится равным 0.004, что опять пренебрежительно мало. Таким образом, если мы рассматриваем группу организмов, в которой интервал между двумя событиями расхождения видов равен один или два миллиона лет, то вероятность того, что генное дерево будет отличаться от видового очень мала.

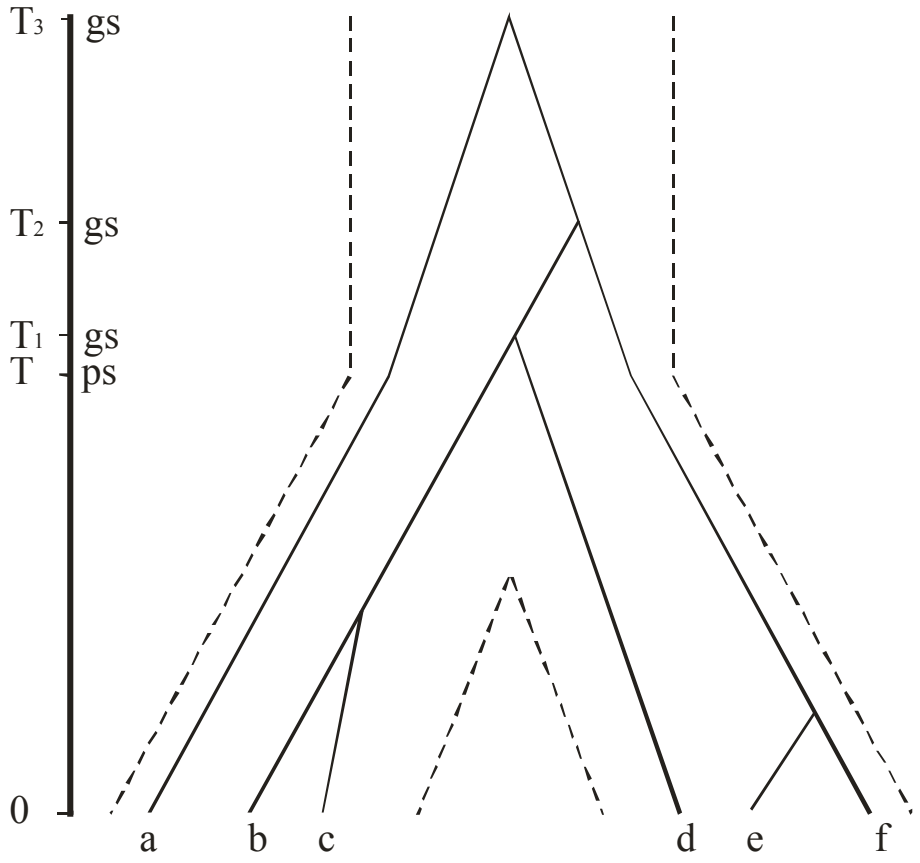

Рис. 6.3. Рисунок, показывающий, что расхождения генов (gs) обычно происходит раньше, чем расхождение популяций (ps), когда присутствует полиморфизм.

При ***N*** = 10000, ***T*** = 100000, и времени смены поколения 5 лет мы получаем ***P(B)*** = 0.245, что уже является существенным значением. Поэтому для группы близко родственных видов или внутривидовых популяций вероятность того, что генное дерево не согласуется с видовым или

популяционным, достаточно велика. Для получения правдоподобного дерева внутривидовых популяций или близкородственных видов необходимо использовать большое число генов, взятых из независимо эволюционирующих (непохожих) локусов.

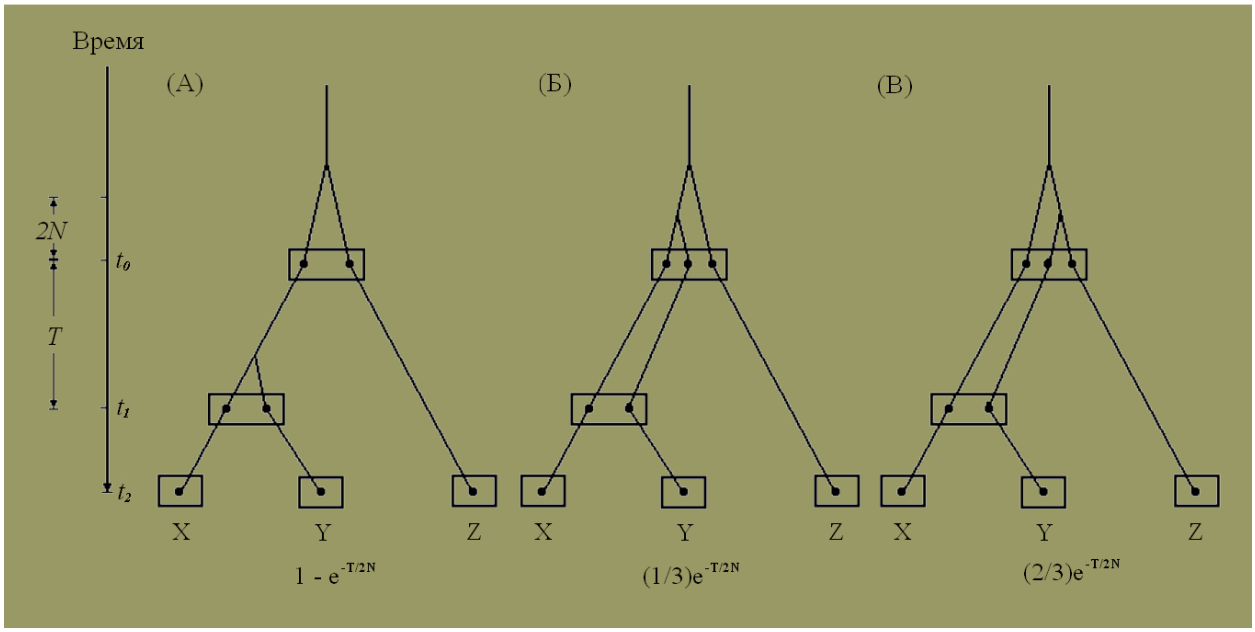

Рис. 6.4. Варианты взаимоотношений видовых и генных деревьев в случае трех видов при наличии полиморфизма.
Время первого и второго расхождений видов равны t_0 *и* t_1*, соответственно. Вероятности данных события приведены под рисунками каждого дерева.* $T = t_1 - t_0$*, а* N *– это эффективный размер популяции.*

Следует также заметить, что даже если действительный паттерн расхождения генов согласуется с паттерном расхождения видов, ветвящийся паттерн построенного генного дерева может не согласоваться с видовым деревом, если число исследуемых нуклеотидов или аминокислот мало. Это происходит потому, что нуклеотидные или аминокислотные замены происходят стохастически, и число замен в линии Z на рисунке 6.4Б может быть меньше, чем в линиях X и Y. Во избежание такого типа ошибок мы должны исследовать большое число нуклеотидов или аминокислот.

Когда изучаемый ген принадлежит мультигенному семейству, возникает другая проблема. Предположим, что два близкородственных вида,

виды 1 и 2, имеют два дуплицированных гена a_1 и b_1 и a_2 и b_2, соответственно, и что дуплицированные гены возникли в результате генной дупликации, произошедшей до дивергенции двух видов (рис. 6.5). В этом случае гены a_1 и a_2 или b_1и b_2 из разных видов называются **ортологичными** генами, в то время как пары генов a_1 и b_1, a_2 и b_2, a_1 и b_2 , a_2 и b_1 называются **паралогичными** генами. Для построения филогенетического дерева разных видов мы должны использовать ортологичные гены, а не паралогичные, потому что только ортологичные гены представляют события расхождения. На практике, однако, различить ортологичные и паралогичные гены не всегда легко, особенно, когда в геноме существует множество копий дуплицированных генов. Таким образом, нужно быть очень внимательным при построении видовых деревьев по результатам генных деревьев.

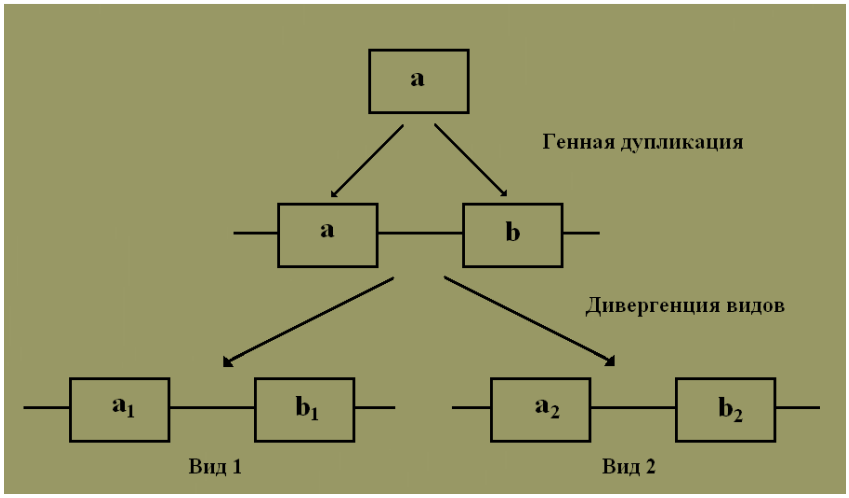

Рис. 6.5. Дуплицированные гены из двух разных видов.
Гены a_1 и a_2 или b_1 и b_2 называются ортологичными, а пары генов a_1 и b_1, a_2 и b_2, a_1 и b_2 , a_2 и b_1 называются паралогичными.

Конечно же, генные деревья не всегда строятся только для выяснения видовых деревьев. При изучении эволюции мультигенных семейств важно знать эволюционную историю генов этого семейства и процесса генной дупликации. В этом случае мы должны изучать генные деревья.

6.2. Ожидаемые и реализованные деревьева.

В теории филогенетического анализа часто предполагают, что исследуемые последовательности ДНК или белка достаточно длинные (теоретически бесконечно длинные), так как такое допущение упрощает статистический анализ, но очень часто необходимо исследовать эволюционную историю коротких последовательностей. Например, если исследователь хочет изучить эволюционную историю генов домашнего хозяйства[50], то ему придется рассматривать последовательности, содержащие примерно 60 кодонов, входящие в состав высококонсервативного «хоумбокс» домена.

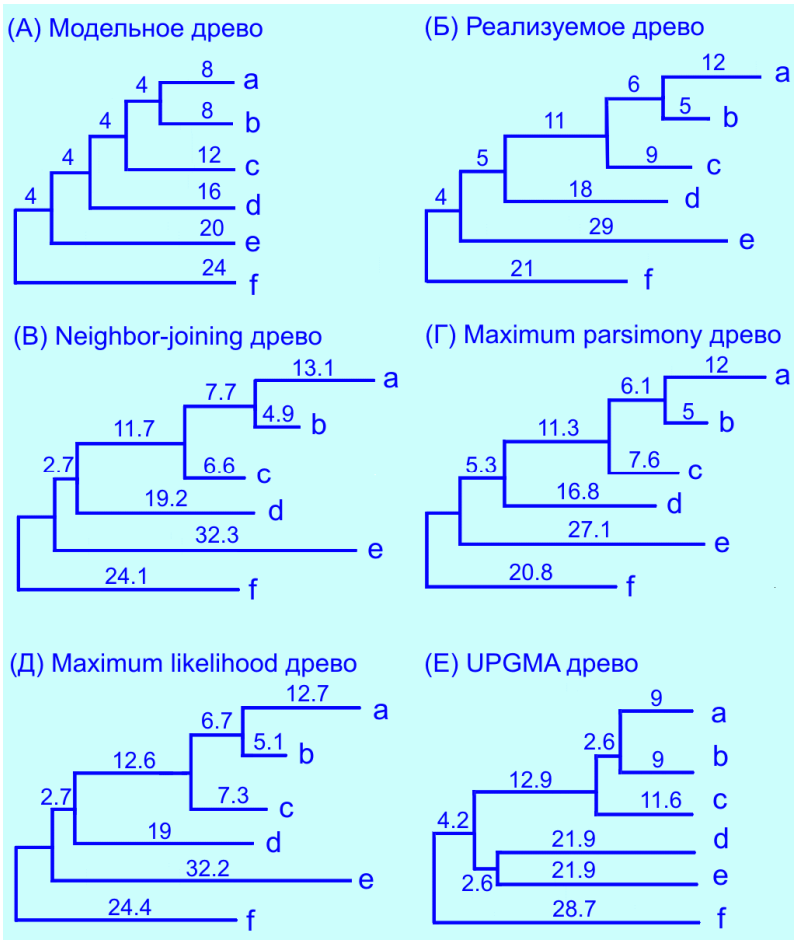

Рис. 6.6. (А) Модельное дерево, (Б) реализованное дерево и (В-Е) воссозданные деревья.

[50] homebox genes

Если мы рассматриваем короткий ген или сегмент ДНК, число нуклеотидных или аминокислотных замен подвергается большим стохастическим ошибкам. Поэтому, даже если ожидаемое число замен увеличивается со временем линейным образом, филогенетическое дерево, описывающее действительное число замен может сильно отличаться от дерева, которое исследователь может ожидать интуитивно. В этом случае топология дерева может сильно отличаться даже от дерева, построенного для длинных последовательностей ДНК. Дерево, которое может быть построено с использованием бесконечно длинных последовательностей или ожидаемого числа замен для каждой ветви, называется **ожидаемым** деревом[51], в то время как дерево, основанное на действительном числе замен, называется **реализованным** деревом[52]. Заметим, что оба, ожидаемое и реализованное, деревья часто отличаются от деревьев, построенных на основе наблюдаемых данных (**построенных**[53] или **оцененных**[54] деревьев). На рис. 6.6 приведен пример отличий между ожидаемым, реализованным и построенным деревьями в случае допущения молекулярных часов. Дерево 6.6А – ожидаемое дерево, у которого длины ветвей равны ожидаемому числу нуклеотидных замен. В этом случае ожидаемое число замен от корня до конечного узла равно 0.12 на нуклеотидный сайт. Поэтому, если рассматривается последовательность из 200 нуклеотидов, то ожидаемое число замен на последовательность будет равно 24. Дерево 6.6Б – это реализованное дерево, полученное при компьютерной симуляции при допущении, что число нуклеотидов равно 200, а нуклеотидные замены происходят по модели Джукса-Кантора. Значение, присвоенное каждой ветви на рис. 6.6Б, представляет собой число замен, которые реально произошли в этой ветви. Эти значения координально отличаются от ожидаемых значений дерева 6.6А из-за стохастических ошибок нуклеотидных замен.

---

[51] expected tree
[52] realized tree
[53] reconstructed tree
[54] inferred tree

Какое же дерево воссоздается методами построения филогенетических деревьев, оджидаемое или реализованное? Ответ зависит от методов построения деревьев, однако большинство методов воссоздает реализованные деревья. На рисунках 6.6В, Г и Д показаны деревья, построенные методами связывания ближайших соседей, максимальной парсимонии и наибольшего правдоподобия, соответственно. Их топология сходна с топологиями и ожидаемого (модельного), и реализованного дерева, однако длины ветвей все же ближе к длинам ветвей реализованнго дерева, таким образом эти три метода являются методами оценки реализованного дерева.

Топология же дерева Е, полученная методом невзвешенных парных групп со средним арифметическим[55] (UPGMA), отличается от обоих модельных и реализованных деревьев. Так как дерево, построенное методом UPGMA, имеет неправильную топологию, сравнение его длин ветвей с длинами ветвей реализованного и ожидаемого деревьев не слишком достоверно[56], однако длины ветвей правильной части (последовательности a, b и c) топологии ближе к модельному дереву. В этом примере ожидаемое число нуклеотидных замен (4) для каждой внутренней ветви мало, поэтому метод UPGMA не может произвести корректную топологию из-за стохастических ошибок. Однако, если увеличить длину исследуемых последовательностей хотя бы в 2 раза, методом UPGMA можно получить корректную топологию с высокой вероятностью, и в этом случае длины ветвей будут ближе к модельному дереву. Другими словами, методом UPGMA можно оценивать модельное или видовое дерево, но, к сожалению, UPGMA дерево подвержено стохастическим ошибкам и другим факторам гораздо сильнее, чем деревья, построенные другими методами.

[55] unweighted pair-group method with arithmetic mean
[56] meaningful

6.3. Символическое представление топологий дерева.

Хотя филогенетическое дерево может быть в целом нарисовано на плоскости, часто бывает более удобно использовать символические выражения для представления разных топологий дерева. Любое бифуркациооное или мультифуркационное дерево может быть выражено через простое символьное выражение. Например, топология деревьев А - Д, изображенных на рис. 6.6, может быть представлена как (f(e(d (c(b;a))))), а топология дерева Е – как (e((d;f)(c(b;a)))). Мультифуркационное дерево также можно представить в символическом виде. Предположим, что в деревьях А – Д таксоны a, b и c дивергировали из одного трифуркационного узла, а не из двух бифуркационных узлов. Тогда топологию дерева можно записать как (f(e(d(c;b;a)))).

В случае неукорененных деревьев существует несколько различных способов описания их топологии. Один простой способ состоит в разделении всех таксонов на три подгруппы, которые соединяются во внутреннем узле, а затем перекомпоновать каждую подгруппу, состоящую из трех или более таксонов в последующие подгруппы таксонов. Например, в случае дерева А, изображенного на рис. 6.6, сначала можно рассмотреть 3 подгруппы таксонов (1;2), 3 и (4;5;6;7;8). Такой вид формирует только одну топологию, но можно далее перекомбинировать подгруппу (4;5;6;7;8) и записать топологию всего дерева в виде ((1;2)3(4((5;6);(7;8)))). В случае мультифуркационных узлов используют немного отличающееся выражение. Предположим, что таксоны 5, 6, 7 и 8 соединены через один мультифуркационный узел в дереве А. Тогда топология может быть записана как ((1;2)3(4(5;6;7;8))) или (((1;2)3)4(5;6;7;8)).

Если использовать такие символические представления деревьев, то можно отличить все топологии друг от друга. Такой метод нахождения отличий важен при исследовании большого числа различных топологий для поиска наиболее правдоподобного дерева.

7. МЕТОДЫ ПОСТРОЕНИЯ ДЕРЕВЬЕВ

Существует множество статистических методов, которые можно использовать для реконструкции филогенетических деревьев на основе молекулярных данных. Наиболее часто используемые методы можно разделить на три группы:

1. Дистанционные методы
2. Методы парсимонии
3. Методы правдоподобия

Кроме того сравнительно недавно были предложены метод ближайшего дерева и метод нейронных сетей для построения филогенетических деревьев, однако практическая целесообразность этих методов пока не выяснена.

7.1. Дистанционные методы[57]

В дистанционных методах, или методах построения филогенетических деревьев на основе **матрицы дистанций**, эволюционные расстояния считаются для всех пар таксонов и филогенетическое дерево строится с учетом отношений между этими значениями дистанций. Существует 4 основных метода построения деревьев на основе матрицы дистанций:

- Методом невзвешенных парных групп со средним арифметическим
- Метод наименьших квадратов
- Метод минимальной эволюции
- Метод связывания ближайших соседей

7.1.1. UPGMA

Самым простым методом в данной категории является **метод невзвешенных парных групп со средним арифметическим** (UPGMA).

[57] distance methods

Дерево, построенное данным методом, иногда называют фенограммой, потому что этот метод изначально использовался для представления степени фенотипического сходства в группе видов в количественной таксономии. Однако метод можно использовать и для воссоздания молекулярной филогении, если скорость генных замен примерно постоянная, то есть соотношение между эволюционным расстоянием и временем дивергенции изменяется примерно линейно. UPGMA предполагается использовать для построения видовых деревьев, однако часто он приводит к топологическим ошибкам, если уровень генных замен не постоянен, или если число генов или нуклеотидов мало.

Рассмотрим алгоритм этого метода. В UPGMA значения эволюционной дистанции рассчитываются для всех пар таксонов или последовательностей и представляются в виде матрицы:

		Таксон		
		A	**B**	**C**
таксон	**B**	d_{AB}		
	C	d_{AC}	d_{BC}	
	D	d_{AD}	d_{BD}	d_{CD}

Здесь $\boldsymbol{d_{ij}}$ – это дистация между ***i***-ой и ***j***-ой taxa. Среди всех таксонов выявляются 2 таксона с наибольшим сходством (наименьшей дистанцией), и они рассматриваются как один новый таксон. Рассчитав расстояние между новым таксоном и остальными, выбираем (создаем) новый таксон и так далее, пока не останется 2 последних таксона.

Пусть $\boldsymbol{d_{AB}}$ – наименьшее эволюционное расстояние. Здесь мы полагаем, что длины ветвей, ведущих от этой точки ветвления к A и B равны, поэтому точка ветвления от них будет на расстоянии $\frac{\boldsymbol{d_{AB}}}{2}$. Таксоны A и B

объединяются в общий таксон, или кластер, (AB), и дистанция между ним и оставшимися таксонами C и D будет равна $\boldsymbol{d}_{(AB)C} = \frac{\boldsymbol{d}_{AC} + \boldsymbol{d}_{BC}}{2}$ и $\boldsymbol{d}_{(AB)D} = \frac{\boldsymbol{d}_{AD} + \boldsymbol{d}_{BD}}{2}$, соответственно. Таким образом, мы получаем следующую новую матрицу:

		O T U	
		(AB)	**C**
O T U	**C**	$d_{(AB)C}$	
	D	$d_{(AB)D}$	d_{BC}

Если $\boldsymbol{d}_{(AB)C}$ окажется наименьшим расстоянием в новой матрице, то таксон C объединяется с кластером (AB), исходящими из узла ветвления, находящемся на расстоянии $\frac{\boldsymbol{d}_{(AB)C}}{2}$. Тогда остается простой таксон D и новый кластер ABC. Дистанция между ABC и D будет равна $\boldsymbol{d}_{(ABC)D} = \frac{(\boldsymbol{d}_{AD} + \boldsymbol{d}_{BD} + \boldsymbol{d}_{CD})/3}{2}$.

В случае составных таксонов расстояние между двумя составными таксонами рассчитывается как среднее арифметическое между простыми таксонами, входящими в их состав. Если мы имеем два кластера (ij) и (mn), то расстояние между ними считается по формуле: $\boldsymbol{d}_{(ij)(mn)} = \frac{(\boldsymbol{d}_{im} + \boldsymbol{d}_{in} + \boldsymbol{d}_{jm} + \boldsymbol{d}_{jn})}{4}$. Если мы имеем два кластера (ijk) и (mn), то расстояние между ними будет равно $\boldsymbol{d}_{(ijk)(mn)} = \frac{(\boldsymbol{d}_{im} + \boldsymbol{d}_{in} + \boldsymbol{d}_{jm} + \boldsymbol{d}_{jn} + \boldsymbol{d}_{km} + \boldsymbol{d}_{kn})}{6}$ и т.д.

Таким образом, очевидно, что расстояние между двумя кластерами (A и B) дается формулой

$$d_{AB} = \sum_{ij} \frac{d_{ij}}{rs} \qquad \text{с}$$

где r и s – это число таксонов в кластерах А и В, соответственно, а d_{ij} это дистанция между таксоном i в кластере А и таксоном j в кластере В. Точка ветвления между двумя кластерами равна $\frac{d_{AB}}{2}$. Для компьютерного программирования этого метода вышеприведенные уравнения не очень удобны, поэтому в программах используются несколько другие, более быстрые, алгоритмы расчета d_{AB}.

Дерево, полученное методом UPGMA, обычно представляется в виде укорененного, потому что достаточно легко вывести корень дерева при допущении постоянной скорости эволюции. Однако, UPGMA – это метод для оценки топологии и длин ветвей, сходный с другими методами, и не обязательно приписывать корень UPGMA дереву. В других методах филогенетической оценки обычно строятся неукорененные деревья, потому что достаточно трудно определить корень, когда скорость эволюции варьируется от ветви к ветви. Мы можем использовать такой же подход и построить неукорененное UPGMA дерево, не учитывая корень, обычно приписываемый UPGMA дереву. Когда мы сравниваем UPGMA дерево с деревьями, построенными другими методами, мы должны использовать нукорененное UPGMA дерево, потому что укоренение может вводить дополнительные ошибки в построение дерева. Неукорененные деревья также полезны при тестировании правдоподобности дерева, полученного бутстрэп[58] или другими методами.

Так как филогенетическое дерево обычно строится на основе ограниченного количества данных, важно исследовать правдоподобность полученного дерева. Существует два основных метода тестирования правдоподобности топологии дерева, полученного дистанционными

[58] bootstrap

методами: тест внутренних ветвей[59] и бутстрэп тест. Оба теста исследуют правдоподобность каждой внутренней ветви дерева. Если каждая внутренняя ветвь оказывается положительной, то дерево рассматривается как правдоподобное со статистической точки зрения. Однако этот метод становится запутанным при большом количестве исследуемых таксонов. Более простой метод тестирования положительности внутренней ветви использовать бутстрэп тест для неукорененных UPGMA деревьев. В этом тесте принято считать количественный эквивалент вероятности достоверности (1 – ошибка I типа), а не уровень значимости. Это значение называется **бутстрэп значением**[60]. Если это значение больше, чем 95% (или 99% в зависимости от уровня значимости, который исследователь хочет получить), то внутренняя ветвь считается статистически значимой.

7.1.2. Метод наименьших квадратов[61].

В случае, когда скорость нуклеотидных замен варьируется от одной эволюционной ветви к другой, UPGMA часто дает некорректную топологию. В этом случае нужно использовать методы, которые позволяют приписывать разные скорости нуклеотидных замен для разных ветвей. Одна из групп таких методов – это **методы наименьших квадратов** (LS). Существует несколько типов LS методов, но чаще всего используются простой LS метод и взвешенный LS метод.

7.1.2.1. Построение топологии.

В простом LS методе филогенетической оценки рассматриваем следующую сумму квадратов

[59] interior branch test
[60] bootstrap value
[61] least squares method

$$R_S = \sum_{i<j} \left(d_{ij} - e_{ij} \right)^2 \qquad (7.2)$$

где d_{ij} и e_{ij} – это наблюдаемая и патристическая дистанции[62] между таксонами i и j, соответственно. Патристическая дистанция между taxa i и j – это сумма оценок длин всех ветвей, соединяющих два таксона в дереве. В стандартном LS методе R_S считается для всех правдоподобных топологий, и для конечного дерева выбирается топология с наименьшим значением R_S.

Фитч и Марголиаш в 1967 году использовали следующее значение R_S для выбора конечной топологии

$$R_S = \sum_{i<j} \left[\frac{\left(d_{ij} - e_{ij} \right)^2}{d_{ij}} \right] \qquad (7.3)$$

Эта процедура называется **взвешенным LS методом**[63]. На практике оба значения R_S обычно дают одну и ту же или очень схожую топологии.

Теоретически, лучше использовать **генерализованный LS метод**[64] расчета R_S, в которых принимаются во внимание и вариация, и ковариация d_{ij}. Однако этот метод довольно времязатратный. Более того, когда d_{ij} достигает значения 0, матрица вариации-ковариации становится сингулярной, и поэтому этот метод не может давать правдоподобные филогенетические деревья.

Было показано, что вероятность получения корректной топологии дерева LS методами, зачастую ниже, чем при использовании других дистанционных методов. Одной из причин этого является то, что эти методы иногда дают отрицательные оценки длин ветвей, что является нереалистическим. Для увеличения эффективности этого метода используют LS метод с ограничением неотрицательных ветвей, что значительно повышает вероятность получения корректной топологии. Оценка длин ветвей с ограничвением неотрицательных ветвей требует итерационного

[62] patristic distance
[63] weighted LS method
[64] generalized LS method

вычисления оценок длин ветвей. Также известно, что в случае четырех таксонов этот метод дает ту же топологию, что и при использовании метода связывания ближайших соседей.

7.1.2.2. Оценка длин ветвей

7.1.2.2.1. Метод Фитча-Марголиаша[65]

Для расчета остаточной суммы квадратов $\boldsymbol{R_S}$ мы сначала должны оценить длины ветвей и $\boldsymbol{e_{ij}}$ для каждой топологии. Простой путь оценки длин ветвей – использовать **метод Фитча-Марголиаша**. Хотя оценки, полученные этим методом, не всегда совпадают с оценками, полученными LS методом, отличия обычно очень малы, поэтому метод Фитча-Марголиаша используется до сих пор. Этот метод использует преимущество свойства, заключающегося в том, что, когда существует только три таксона, оценки длин ветвей для всех трех таксонов могут быть однозначно определены.

Рассмотрим три таксона 1, 2 и 3, эволюционные взаимоотношения которых показаны на рисунке 7.1А. Эволюционные дистанции между таксонами 1 и 2, 1 и 3 и 2 и 3 равны

$$\begin{aligned} d_{12} &= x + y \\ d_{13} &= x + z \\ d_{23} &= y + z \end{aligned} \qquad (7.4)$$

где x, y и z – длины ветвей для таксонов 1, 2 и 3, соответственно. Решая эту одновременную систему уравнений, получаем

$$\begin{aligned} x &= (d_{12} + d_{13} - d_{23})/2 \\ y &= (d_{12} - d_{13} + d_{23})/2 \\ z &= (-d_{12} + d_{13} + d_{23})/2 \end{aligned} \qquad (7.5)$$

Это LS оценки.

---

[65] Fitch –Margoliash method

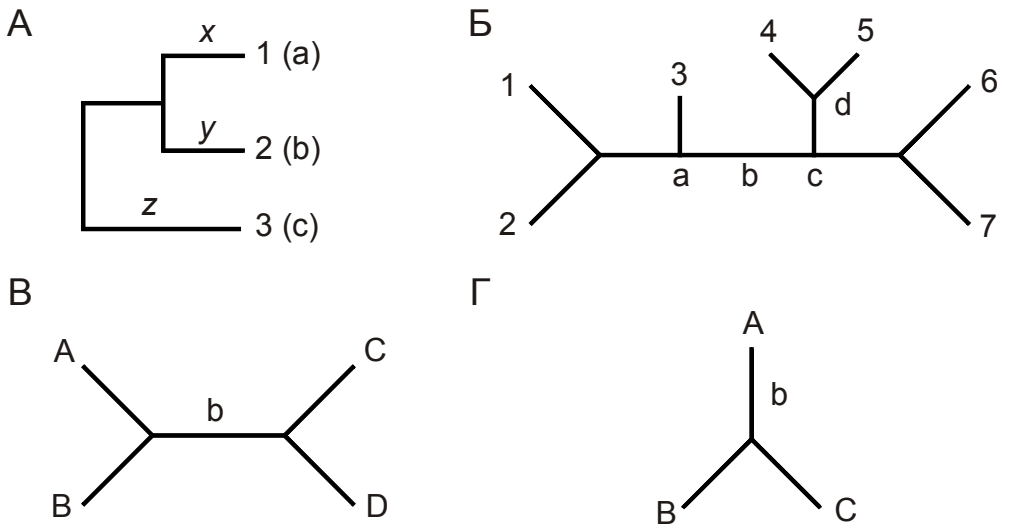

Рис. 7.1. Оценка длин ветвей.

В случае четырех и более таксонов мы сначала выбираем два таксона с наименьшей дистанцией и обозначаем их как А и В. Все оставшиеся таксоны объединяем в один составной таксон, обозначаемый как С. Дистанция между таксонами А и В такая же как оригинальная дистанция (d_{12}), но дистанция между таксонами А и С будет представлена простым средним значением дистанций между А и всеми таксонами в С. Аналогично дистанция между таксонами В и С является средним значением дистанций между В и всеми таксонами в С. Например, в матрице дистанций (табл. 7.1) наименьшая дистанция наблюдается между человеком и шимпннзе. Поэтому мы обозначаем человека как А, шимпанзе как В, а все оставшиеся виды как С. Из оценок дистанций, приведенных в табл. 7.1, получаем $d_{AA} = 0.095$, $d_{AC} = (0.113 + 0.183 + 0.212)/3 = 0.169$ и $d_{BC} = (0.118 + 0.201 + 0.225)/3 = 0.181$. Значения x, y и z следовательно становятся равными 0.042, 0.054 и 0.124, соответственно, из уравнений 7.5. Здесь x и y представляют собой число оцененных нуклеотидных замен (а и b) для линий человека и шимпанзе, соответственно, а z – это дистанция между составным таксоном С и точкой ветвления между человеком и шимпанзе.

Таблица 7.1. Дистанции Кимуры, вычисленные для 896 п.н. фрагментов митохондриальной ДНК гоминидов.

	Человек	Шимпанзе	Горилла	Орангутан
Человек	0,095±0,011			
Шимпанзе	0,113±0,012	0,118±0,013		
Горилла	0,183±0,016	0,201±0,018	0,195±0,017	
Орангутан	0,212±0,018	0,225±0,0129	0,225±0,019	0,222±0,018

Теперь мы объединяем таксоны 1 и 2 и переобозначаем таксон как (AB). После этого пересчитываем дистанции между этим составным таксоном (AB) и всеми остальными таксонами и выбираем два таксога, которые имеют наименьшее значение среди всех дистанций, включая те, которые не вовлекают (AB). Эти два таксона снова обозначают как A и B, в то время как C представляет собой составной таксон, состоящий из всех остальных таксонов. Новые значения x, y и z подсчитываются по той же процедуре. В случае данных для гоминидов дистанции между (AB) и другими таксонами (гориллы, орангутаны и гиббоны) уже были рассчитаны (0.115, 0.192 и 0.218, соответственно) при построении UPGMA дерева, и наименьшее расстояние в новой матрице между (AB) и гориллами. Поэтому (AB) и гориллы переобозначаются как новые A и B, соответственно, а C представляет собой орангутанов и гиббонов. Теперь мы получаем $d_{AB} = 0.115$, $d_{AC} = (0.183 + 0.201 + 0.212 + 0.225)/4 = 0.205$ и $d_{BC} = (0.195 + 0.225)/2 = 0.210$. Таким образом, мы получаем $x = 0.055$, $y = 0.060$ и $z = 0.150$ из уравнений 7.5. Длины ветвей c и d дерева оцениваются, используя следующие соотношения:

$$d_{AB} = (a + b)/2 + c + d$$
$$d_{AC} = (a + b)/2 + c + z$$
$$d_{BC} = d + z$$

Мы знаем, что $(a+b)/2 = 0.048$ и $z = 0.150$. Таким образом, получаем, что ***c*** = 0.008 и ***d*** = 0.060. Вышеприведенная процедура повторяется до тех пор пока все длины ветвей (e, f и g) не будут оценены.

Теперь можно сосчитать e_{ij} для всех пар таксонов а затем и значения R_S в уравнениях 7.2 и 7.3. R_S становятся равными 0.000047 и 0.002264, ссответственно. Для нахождения LS дерева, однако, мы должны рассмотреть все возможные и все праввдоподобные деревья. На практике, число топологий обычно очень велико, поэтому для расчета R_S используется только малая часть возможных топологий. В методе **Фитча-Марголиаша** первая топология строится по вышеописанному алгоритму. После того как такая топология получается, остальные топологии исследуются различными алгоритмами обмена ветвей[66]. Эти алгоритмы важны в отношении построения деревьев максимальной парсимонии.

Раз окончательная топология дерева получается минимизацией R_S, то лучшие оценки длин ветвей окончательного дерева могут быть получены LS методом, который будет описан далее. Математически LS оценки более правдоподобны, чем оценки, полученные методом **Фитча-Марголиаша**, но на практике отличия между ними очень малы, когда используются последовательности ДНК и белка.

7.1.2.2.2. Метод наименьших квадратов[67]

Стандартный метод оценки длин ветвей дерева – использование LS метода. Ржетски и Ней разработали быстрый алгоритм получения LS оценок длин ветвей для любой данной топологии. Представим себе гипотетическое дерево для пяти последовательностей, изображенных на рис. 7.2A и используем простой LS метод для оценки длин ветвей, обозначаемых b_1,

[66] branch-swapping algorithm
[67] Least squares method

b_2,…, и b_7. Обозначим оценку эволюционной дистанции между последовательностями i и j как d_{ij}. Тогда мы можем записать d_{ij} как

$$
\begin{array}{lllllllllllllllll}
d_{12} & = & b_1 & + & b_2 & & & & & & & & b_6 & & & + & \varepsilon_{12} \\
d_{13} & = & b_1 & & & + & b_3 & & & & & + & b_6 & & & + & \varepsilon_{13} \\
d_{14} & = & b_1 & & & & & + & b_4 & & & + & b_6 & + & b_7 & + & \varepsilon_{14} \\
d_{15} & = & b_1 & & & & & & & + & b_5 & + & b_6 & + & b_7 & + & \varepsilon_{15} \\
d_{23} & = & & & b_2 & + & b_3 & & & & & + & b_6 & + & b_7 & + & \varepsilon_{23} \\
d_{24} & = & & & b_2 & & & + & b_4 & & & + & b_6 & + & b_7 & + & \varepsilon_{24} \\
d_{25} & = & & & b_2 & & & & & + & b_5 & + & b_6 & + & b_7 & + & \varepsilon_{25} \\
d_{34} & = & & & & & b_3 & + & b_4 & & & & & + & b_7 & + & \varepsilon_{34} \\
d_{35} & = & & & & & b_3 & & & + & b_5 & & & + & b_7 & + & \varepsilon_{35} \\
d_{45} & = & & & & & & & b_4 & + & b_5 & & & & & + & \varepsilon_{45}
\end{array}
$$

где ε_{ij} – это sampling errors. Мы полагаем, что ε_{ij} распределены со средним 0 и вариацией $V(d_{ij})$. Если использовать матричную алгебру, то вышеприведенную систему уравнений можно записать в виде

$$d = Ab + \varepsilon \qquad (7.6)$$

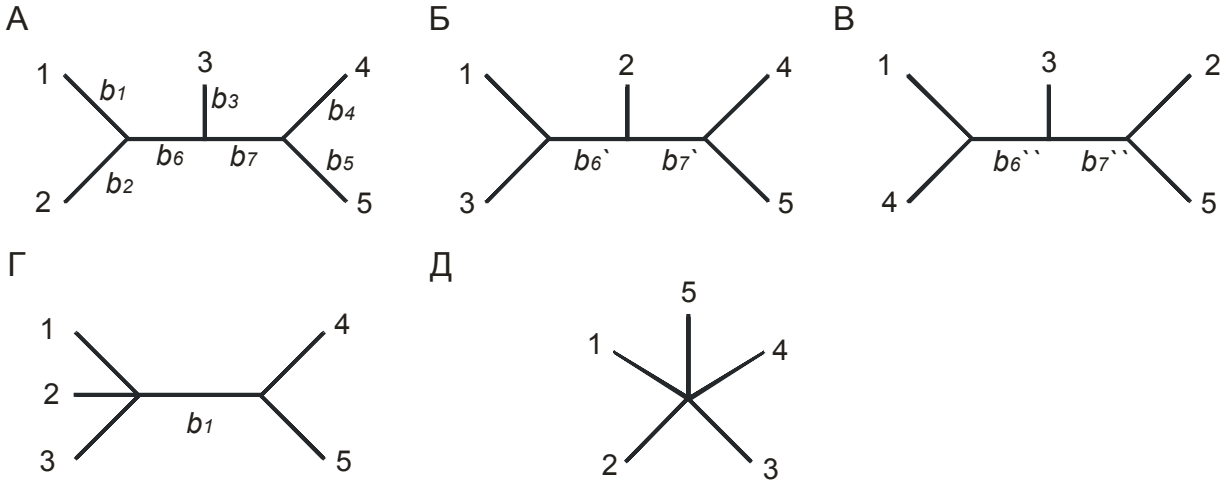

Рис. 7.2. Три топологии деревьев для пяти таксонов и два «ноль дерева» для тестирования топологических отличий.

H_0: T(А) = T(Б); $b_6 = 0$ и H_0: T(А) = T(В); $b_6 = b_7 = 0$

где $\boldsymbol{d}$, $\boldsymbol{b}$ и $\boldsymbol{\varepsilon}$ это вектор-столбцы из $\boldsymbol{d_{ij}}$, $\boldsymbol{b_i}$ и $\boldsymbol{\varepsilon_{ij}}$, соответственно; а именно $\boldsymbol{d^t} = (\boldsymbol{d_{12}}, \boldsymbol{d_{13}}, ..., \boldsymbol{d_{45}})$, $\boldsymbol{b^t} = (\boldsymbol{b_1}, \boldsymbol{b_2}, ..., \boldsymbol{b_7})$ и $\boldsymbol{\varepsilon^t} = (\boldsymbol{\varepsilon_{12}}, \boldsymbol{\varepsilon_{13}}, ..., \boldsymbol{\varepsilon_{45}})$. Здесь $\boldsymbol{t}$ означает операцию транзпонирования вектора или матрицы. Заметим, что векторы $\boldsymbol{d}$ и $\boldsymbol{\varepsilon}$ состоят из $\boldsymbol{r \equiv m(m-1)/2}$ элементов и $\boldsymbol{b}$ состоит из $\boldsymbol{T \equiv 2m-3}$ элементов, где m это число последовательностей. $\boldsymbol{A}$ – это матрица, представляющая топологию, и в данном случае (топология [A] на рис. 7.2) имеет следующий вид

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix} \tag{7.7}$$

Элементы этой матрицы равны 1, если существует соответствующая ветвь и 0 в противном случае (см. уравнения для $\boldsymbol{d_{ij}}$). LS оценка $\boldsymbol{b}$ тогда дается уравнением:

$$\hat{\boldsymbol{b}} = (\boldsymbol{A^t A})^{-1} \boldsymbol{A^t d} = \boldsymbol{Ld} \tag{7.8}$$

где $\boldsymbol{L} = (\boldsymbol{A^t A})^{-1} \boldsymbol{A^t}$. Очевидно, что оценка длины *i*-ой ветви равна

$$\hat{\boldsymbol{b}}_i = \boldsymbol{L}_i \boldsymbol{d} \tag{7.9}$$

где $\boldsymbol{L}_i$ – это *i*-ый ряд матрицы $\boldsymbol{L}$. Если использовать эту формулу для топологии A рис. 7.2 мы получаем

$$\hat{b}_1 = \frac{1}{2}d_{12} + \frac{1}{6}(d_{13} - d_{23} + d_{14} - d_{24} + d_{15} - d_{25})$$

$$\hat{b}_2 = \frac{1}{2}d_{12} - \frac{1}{6}(d_{13} - d_{23} + d_{14} - d_{24} + d_{15} - d_{25})$$

$$\hat{b}_3 = \frac{1}{4}(d_{13} + d_{23} + d_{34} + d_{35}) - \frac{1}{8}(d_{14} + d_{24} + d_{15} + d_{25})$$

$$\hat{b}_4 = \frac{1}{2}d_{45} + \frac{1}{6}(d_{14} - d_{15} + d_{24} - d_{25} + d_{34} - d_{35}) \qquad (7.10)$$

$$\hat{b}_5 = \frac{1}{2}d_{45} - \frac{1}{6}(d_{14} - d_{15} + d_{24} - d_{25} + d_{34} - d_{35})$$

$$\hat{b}_6 = -\frac{1}{2}d_{12} + \frac{1}{4}(d_{13} + d_{23} - d_{34} - d_{35}) + \frac{1}{8}(d_{14} + d_{24} + d_{15} + d_{25})$$

$$\hat{b}_7 = \frac{1}{4}(d_{34} + d_{35} - d_{13} - d_{23}) + \frac{1}{8}(d_{14} + d_{24} + d_{15} + d_{25}) - \frac{1}{2}d_{45}$$

Сходные выражения могут быть получены для любой другой топологии, таких как, например, топологии Б или В, изображенных на рис. 7.2 или для любого числа последовательностей ($\boldsymbol{m}$).

Кроме того, существует еще один способ оценки длин ветвей без использования матричной алгебры, требующий меньшее количество временных затрат. В качестве примера рассмотрим дерево Б, изображенное на рис. 7.1. Если мы выберем одну внутреннюю ветвь этого дерева, то оно может быть представлено в виде дерева В, где A, B, C и D представляют собой кластер последовательностей. Например, для внутренней ветви $\boldsymbol{b}$ дерева Б на рис. 7.1 A, B, C и D представляют собой кластеры (3), (1,2), (4,5) и (6,7), соответственно. В этом случае длина ветви b в древе В может быть оценена из следующего уравнения

$$\hat{b} = \frac{1}{2}\left\{\gamma\left[\frac{d_{AC}}{m_A m_C} + \frac{d_{BD}}{m_B m_D}\right] + (1-\gamma)\left[\frac{d_{BC}}{m_B m_C} + \frac{d_{AD}}{m_A m_D}\right] - \frac{d_{AB}}{m_A m_B} - \frac{d_{CD}}{m_C m_D}\right\} \qquad (7.11)$$

где

$$\gamma = \frac{(m_B m_C + m_A m_D)}{(m_A + m_B)(m_C + m_D)}$$

Здесь m_A, m_B, m_C и m_D – это числа последовательностей в кластерах A, B, C и D, соответственно, а d_{AC} – это сумма парных дистанций между кластером A (последовательность 3) и кластером C (последовательности 4 и 5). Дистанции d_{BD}, d_{BC}, d_{AD}, d_{AB} и d_{CD} определяются таким же образом. Наоборот, LS оценка длины (b) внешней ветви дерева D рис. 7.1 дается выражением

$$\hat{b} = \frac{1}{2}\left(\frac{d_{AB}}{m_B} + \frac{d_{AC}}{m_C} - \frac{d_{BC}}{m_B m_C}\right) \qquad (7.12)$$

где d_{AB} – это сумма всех парных дистанций между последовательностью A (представляющей одну внешнюю ветвь) и всеми последовательностями, принадлежащими к кластеру B, d_{AC} – это сумма дистанций между и всеми последовательностями, принадлежащими кластеру C, d_{BC} – это сумма всех парных дистанций между последовательностями в кластерах B и C, а m_B и m_C – это количество последовательностей в кластерах B и C, соответственно.

Вышеприведенные уравнения значительно упрощают расчет оценок длин ветвей. Например, $\hat{b}_1$ в уравнении 7.10 может быть получена, используя уравнение 7.12. В этом случае дерево представлено на рис. 7.2A, а последовательности в кластерах A, B и C это 1, 2 и (3, 4, 5), соответственно. Поэтому, $d_{AB} = d_{12}$, $d_{AC} = d_{13} + d_{14} + d_{15}$, $d_{BC} = d_{23} + d_{24} + d_{25}$, $m_B = 1$ и $m_C = 3$, и мы получаем $\hat{b}_1 = \frac{1}{2}\left(d_{12} + \frac{d_{13} + d_{14} + d_{15}}{3} - \frac{d_{23} + d_{24} + d_{25}}{3}\right)$, что идентично с $\hat{b}_1$ в уравнении 7.10. аналогично все другие оценки длин ветвей могут быть получены уравнением 7.11, либо 7.12. Если все $\hat{b}_i$ оценены, то все e_{ij} в уравнениях 7.2 и 7.3 могут легко быть получены суммированием $\hat{b}_i$ для всех ветвей, соединяющих последовательности i и j, а затем могут быть сосчитаны и R_S.

В 1997 году был разработан быстрый алгоритм для расчета $\hat{\boldsymbol{b}}$ с использованием уравнений 7.11 и 7.12. Этот алгоритм используется в пакетах программ PAUP* и MEGA.

7.1.3. Дистанции, используемые при построении филогенетических деревьев

Ранее мы рассматривали разные дистанционные меры оценки числа нуклеотидных и аминокислотных замен $\boldsymbol{d}$, используя разные математические модели. В целом, дистанционные измерения, основанные на сложных математических моделях, требуют оценки большого числа параметров, что увеличивает дисперсию оценки $\boldsymbol{d}$. Теоретически можно выбрать математическую модель, наиболее подходящую для конкретных исследуемых последовательностей, используя определенные статистические критерии. Однако, признанная критерием, как наиболее подходящая, мера дистанции может таковой и не являться для правильной реконструкции филогенетического дерева, однако такие тесты обычно оказываются полезными при оценки длин ветвей.

В настоящее время не существует общего статистического метода выбора подходящего способа измерения дистанции для построения топологии дерева. Однако, компьютерное моделирование и эмпирические исследования позволяют дать следующие рекомендации для воссоздания правильной топологии дерева:

1. В случае, когда оценка числа нуклеотидных замен на сайт ($\boldsymbol{d}$), полученная по модели Джукса-Кантора, равна 0.05 или менее ($\boldsymbol{d} \leq 5$), используйте p или дистанцию Джукса-Кантора, независимо от того, есть или нет отклонение транзиции,трансверсии или варьируется скорость замен ($\boldsymbol{r}$) с нуклеотидным сайтом или нет. В этом случае дистанция Кимуры и другие более сложные дистанции дают преимущественно те же значения, что и p-дистанция (рис. 3.1), но при этом дисперсии этих оценок больше, чем для p-

дистанции. ***p***-дистанция дает хорошие результаты, особенно при малом числе нуклеотидов или аминокислот.

2. Когда $0{,}05 < d < 1{,}0$ и число исследуемых нуклеотидов велико, используйте дистанцию Джукса-Кантора при небольшом значении transition/transversion отношения ***R***, скажем не более пяти, $R > 5$. При большом значении отношения ***R*** и большом числе исследуемых нуклеотидов ***n*** используйте дистанцию Кимуры или гамма-дистанцию. Однако, при большом числе последовательностей и относительно малом значении ***n***, ***p***-дистанция зачастую дает лучшие результаты пока скорость эволюции сильно не варьируется по эволюционной линии. Если число нуклеотидов очень велико (>10000) и скорость нуклеотидных замен варьируется существенно с эволюционной линией, то оправдано применение сложных методов оценки дистанций (например, HKY гамма-дистанция).

3. Когда $d > 1$ для многих пар последовательностей, построенное филогенетическое дерево будет недостоверно по целому ряду причин (например, большие дисперсии $\hat{d}$ или ошибки выравнивания). В этом случае лучше всего избегать такого типа данных для анализа. Для этого из рассмотрения выбрасываются быстро эволюционирующие участки гена и рассматриваются только оставшиеся более консервативные области (например, так обычно поступают с генами вариабельных областей иммуноглобулинов).

4. Многие дистанции для оценки числа нуклеотидных замен на сайт ***d*** становятся не пригодными для применения, когда дистанция очень велика или n мало. Этого возникает по причине того, что математические формулы для оценки дистанции обычно содержат логарифмические функции, и аргументы логарифма часто становятся отрицательными. Теоретически эта проблема может быть решена путем разложения логарифма в бесконечный ряд, однако вариация такой дистанции будет довольна велика. Поэтому для построения топологии лучше не использовать сильно дивергировавшие последовательности. В этом случае ***p***-дистанция часто более эффективна для

получения достоверной топологии, так как она всегда применима и имеет меньшую дисперсию.

5. Когда филогенетическое дерево строится для кодирующих областей гена, может быть полезна разница между синонимическими d_S и несинонимическими d_N заменами, так как скорость синонимических замен обычно выше скорости несинонимических замен. Когда изучаются относительно близкие виды и рассматривается большое число кодонов и $d_S < 0{,}5$, то d_S можно использовать для построения деревьев. Такая процедура должна снижать эффект от вариации скорости замен по разным сайтам, потому что синонимические замены подвергаются отбору менее часто, чем несинонимические. Однако для относительно удаленных видов d_N и аминокислотные дистанции будут более подходящими.

6. Как общее правило, если две оценки дистанций дают схожие значения для одного и того же набора данных, то лучше использовать более простую оценку, так как для нее буде меньше дисперсия. Когда скорость нуклеотидных замен примерно одинакова для всех эволюционных линий и нет сильного смещения транзиции/трансверсии, то ***p***-дистанция дает корректные деревьева даже при большой степени дивергенции последовательностей чаще, чем все остальные методы. Когда скорость замен варьируется по эволюционным линиям, то это не всегда так. Важно не доверять построенным компьютером деревьевам без тщательного рассмотрения паттерна нуклеотидных и аминокислотных замен, различий в частотах нуклеотидов в первом, втором и третьем положениях кодонов, temporal изменений частот нуклеотидов и т.д. При анализе реальных данных существует огромное количество неизвестных факторов, поэтому нужно с большой осторожностью и здравым смыслом интерпретировать полученные филогенетические деревья.

7.2. Методы наибольшей парсимонии

Методы наибольшей парсимонии (MP)[68] изначально были разработаны для анализа морфологических характеристик. Мы будем рассматривать только методы, полезные для анализа молекулярных данных. Эк и Дэйхофф в 1969 году первыми использовали MP метод для построения филогенетических деревьев по данным аминокислотных последовательностей. Позже Фитч и Хартиган применили более совершенный алгоритм MP для данных нуклеотидных последовательностей. В этих алгоритмах MP рассматривались четыре и более выравненных последовательностей ($\boldsymbol{m} \geq 4$)

7.2.1. Оценка минимального числа замен.

Рассмотрим, как рассчитать минимальное число замен для данной топологии. Рассмотрим топологию укорененного дерева для шести последовательностей ДНК (1, 2, 3, 4, 5 и 6), изображенного на рис. 7.3A, и предположим, что нуклеотиды в данном сайте существующих в настоящее время последовательностей представлены во внешних узлах дерева. Это один C, три T и два A. По данным этих нуклеотидов, мы можем предсказать нуклеотиды в пяти предковых таксонах (узлы) a, b, c, d и e. Нуклеотид в узле a должен быть либо C, либо T, если учитывать минимально возможное число замен. Нуклеотид в узле b должен быть T, а в узле c должны быть A или T. Узел d должен быть T, потому что его непосредственные потомственные узлы (a и c) оба содержат T. И наконец, в узле e должно быть A или T. Очевидно, что минимальное число нуклеотидных замен для такого набора данных может быть получено в случае, если все предковые узлы будут T. Число замен будет равно трем. Однако, такой набор нуклеотидов в

[68] maximum parsimony method

предковых узлах (путях) не единственно возможный для объяснения эволюционных изменений нуклеотидов.

Если предположить, что в узлах *a*, *c*, *d* и *e* находится нуклеотид А, а в узле *b* нуклеотид Т, то число требующихся замен опять будет равно трем (см. рис. 7.1Б). На самом деле существует еще три возможных пути с минимальным числом замен: (*a* – Т, *b* – Т, *c* – А, *d* – А, *e* – А), (*a* – С, *b* – Т, *c* – А, *d* – А, *e* – А) и (*a* – Т, *b* – Т, *c* – Т, *d* – Т, *e* – А). Эти результаты показывают, что нуклеотиды в предковых узлах не всегда могут быть определены однозначно, и все нуклеотиды, представленные на рисунке 7.1Б являются наиболее экономичными. Однако возможно сосчитать минимальное число требующихся замен. Оно равно трем для всех вышеописанных случаев.

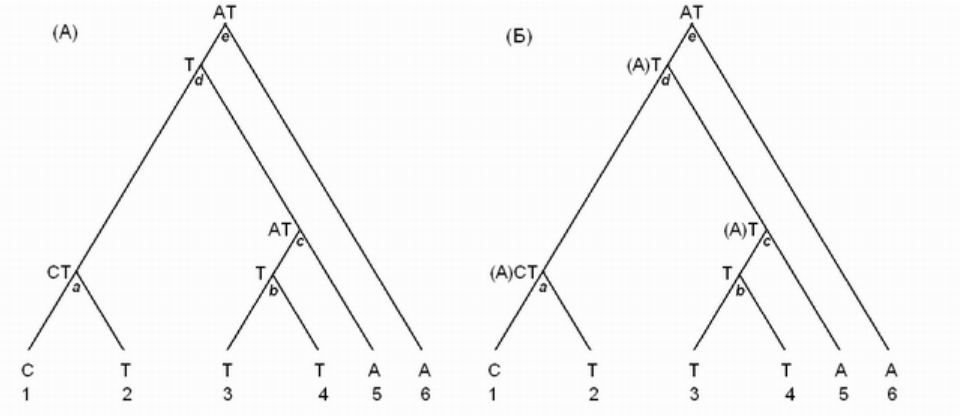

Рис. 7.3. Нуклеотиды в шести внешних последовательностях и возможные нуклеотиды в пяти предковых последовательностях.

В вышеприведенном примере мы рассматривали укорененное дерево, однако дерево может быть трансформировано в неукорененное, если элиминировать верхний узел *e*. Исключение этого узла не изменяет минимальное число замен, однако число возможных путей сокращается. Для данного примера два возможных пути (*a* – Т, *b* – Т, *c* – Т, *d* – Т, *e* – Т) и (*a* – А, *b* – Т, *c* – Т, *d* – Т, *e* – Т) становятся более не различимыми, потому что узел *e* может быть как Т, так и А. В данном случае полное число возможных путей для неукорененного дерева равно четырем. Так как методы MP обычно

не позволяют определить корень дерева, то обычно рассматриваются неукорененные деревьева.

В рассмотренном примере минимальное число замен было равно трем, и существовало четыре экономных пути для неукорененного дерева. Подсчет этих значений был относительно прост, но с ростом числа таксонов он становится все более громоздким. Поэтому все эти расчеты проводятся на компьютере.

7.2.2. Длины дерева

В вышеприведенном примере мы рассматривали только одну топологию, но в реальности мы должны рассматривать все потенциально корректные топологии и из них выбрать ту, которая требует наименьшего минимального числа замен. Рассмотрим деревья, изображенные на рис. 7.4, Они состоят из шести таксонов, однако топологии у этих деревьев отличаются друг от друга. Мы вновь рассматриваем один определенный нуклеотидный сайт и считаем минимальное число замен. Для топологии А, минимальное число замен равно двум, для топологии Б, в которой таксоны 3 и 4 поменяны местами, минимальное число замен равно трем. Топологии В и Г также требуют минимум три замены. В случае шести таксонов существует 105 различных топологий, и мы должны рассчитать минимальное число замен для всех этих топологий. После того как такие расчеты проведены для всех сайтов всех топологий, мы можем сосчитать сумму минимального числа замен для всех сайтов каждой топологии. Эта сумма (***L*** или ***TL***) называется **длиной дерева**[69]. Дерево с максимальной парсимонией (MP) – это топология с наименьшей длиной дерева. Может получиться такая ситуация, при которой существует несколько топологий с одинаковым минимальным числом замен. В этом случае мы не можем определить единственную окончательную топологию, и все они рассматриваются как потенциальное

[69] tree length

корректные. Можно построить составное дерево, учитывающее все полученные MP топологии. Такое дерево называется **консенсусным**.

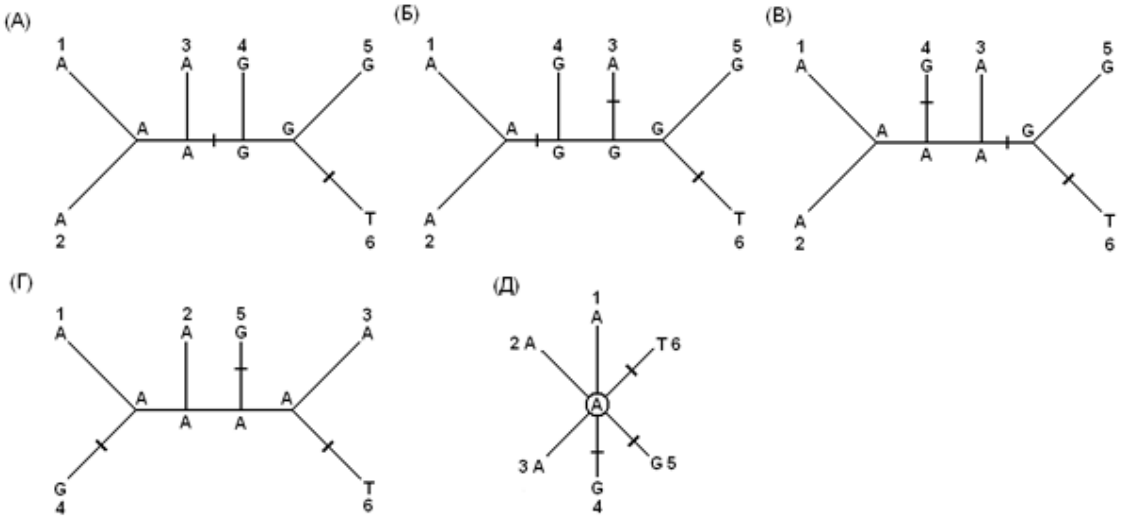

Рис. 7.4. Распределение мутаций в разных ветвях в информативном сайте.

7.2.3. Информативные сайты и гомоплазия.

При поиске MP дерева нуклеотидные (или аминокислотные) сайты, в которых содержатся одинаковые нуклеотиды (или аминокислоты) для всех таксонов (**невариабельные** сайты) не включаются в анализ, а используются только **вариабельные** сайты. Однако не все вариабельные сайты могут оказаться полезными для нахождения топологии MP дерева. Любой нуклеотидный сайт, в котором присутствуют только неповторяющиеся нуклеотиды (синглет[70]), является не информативным, потому что изменение нуклеотида в сайте всегда может быть объяснено одним и тем же числом замен во всех топологиях. Такой сайт называется **синглетным**. Для того чтобы сайт был информативным для построения MP дерева, должно быть как минимум два разных типа нуклеотидов, каждый из которых представлен как минимум два раза. Эти сайты называются **информативными** сайтами. В

[70] singleton

деревьях А, Б, В и Г рис. 7.4 нуклеотидный сайт удовлетворяет этим условиям, поэтому его можно использовать для нахождения топологии с минимальным числом замен.

При построении MP деревьев достаточно рассматривать только информативные сайты, но для построения корректной топологии важно рассматривать большое количество информативных сайтов. Однако когда велика степень **гомоплазии** (обратные и параллельные мутации), полученные MP деревьева не будут правдоподобными даже при наличии большого числа информативных сайтов.

MP методы часто используются для построения деревьевной топологии без учета длин ветвей, однако при некоторых допущениях возможно произвести оценку длин ветвей MP дерева.

7.3. Метод наибольшего правдоподобия

Идея использовать метод наибольшего правдоподобия (ML) для филогенетических оценок впервые была предложена в 1967 году для данных по частотам генов, однако, они столкнулись с некоторыми сложностями в его применении. Позже Фельсенштейн разработал алгоритм построения филогенетических деревьев для нуклеотидных последовательностей методом ML. Кишино и соавт. в 1990 году расширили этот метод для белковых последовательностей с использованием матриц Дэйхофф. В ML методах правдоподобие наблюдать данный набор данных для последовательностей для определенной модели замен максимизируется для каждой топологии, и в качестве финального дерева выбирается топология с наибольшим максимальным правдоподобием. Рассматриваемые параметры – не топологии, а длины ветвей для каждой топологии, и правдоподобие максимизируется для оценки длин ветвей. Далее мы рассмотрим основу этих ML методов и обсудим некоторые аспекты теоретического обоснования методов.

7.3.1. Расчетная процедура методов наибольшего правдоподобия.

7.3.1.1. Расчет значений правдоподобия.

Вначале объясним, как рассчитать значения правдоподобия для данного дерева с использованием данных по последовательностям ДНК. Рассмотрим простое дерево для четырех таксонов (рис. 7.5А) и допустим, что последовательности ДНК n нуклеотидов длиной и выравниваются без инсерций/делеций. Теперь рассмотрим нуклеотиды из последовательностей 1, 2, 3 и 4 в определенном сайте (k-ый сайт) и обозначим их как x_1, x_2, x_3 и x_4, соответственно. Мы не знаем нуклеотидов в узлах 0, 5 и 6, но, допустим, что они x_0, x_5 и x_6, соответственно. Здесь x_i может быть одним из четырех нуклеотидов A, T, C или G.

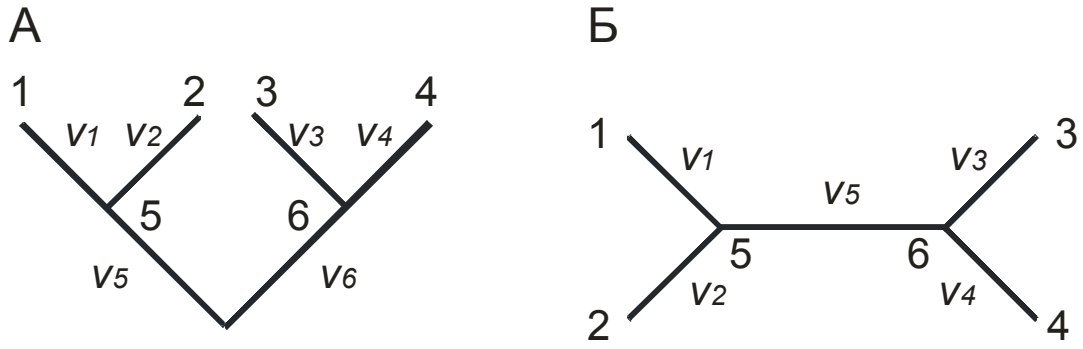

Рис. 7.5. Укорененное и неукорененное филогенетические деревья для четырех таксонов, объясняющие метод максимального правдоподобия. *$v_i = r_i t_i$, где r_i – это скорость нуклеотидных замен, t_i – это эволюциооное время для ветви. В дереве Б v_5 представляет из себя сумму v_5 и v_6 в дереве А.*

Рассмотрим нуклеотидный сайт и допустим, что $P_{ij}(t)$ – это вероятность того, что нуклеотид i в момент времени 0 становится нуклеотидом j в момент времени t в данном сайте. Здесь i и j могут быть любым нуклеотидом из A, T, C или G. В ML методе скорость замен r может варьировать от ветви к ветви, поэтому удобно измерять эволюционное время

в единицах ожидаемого числа замен $v = rt$. В дальнейшем, мы назовем ожидаемое число замен для i-ой ветви как $v_i \equiv r_i t_i$. В ML методе длины ветвей v_i рассматриваются в качестве параметров, они оцениваются путем максимизации функции правдоподобия для данного набора наблюдаемых нуклеотидов, как было отмечено выше. Функция правдоподобия для нуклеотидного сайта (k-ый сайт) дается выражением

$$I_k = g_{x_0} P_{x_0 x_5}(v_5) P_{x_5 x_1}(v_1) P_{x_5 x_2}(v_2) P_{x_0 x_6}(v_6) P_{x_6 x_3}(v_3) P_{x_6 x_4}(v_4) \qquad (7.13)$$

где g_{x_0} – это априорная вероятность того, что в узле 0 находится нуклеотид x_0. g_{x_0} часто приравнивается к относительной частоте нуклеотида x_0 в целом наборе последовательностей, но он может быть оценен ML методом.

Для того чтобы знать $P_{ij}(v)$ в явном виде, необходимо использовать специфическую модель замен. Фельсенштейн использовал equal-input модель. В ней $P_{ii}(v)$ и $P_{ij}(v)$ $(i \neq j)$ равны

$$P_{ii}(v) = g_i + (1 - g_i)e^{-v} \qquad (7.14а)$$

$$P_{ij}(v) = g_j(1 - e^{-v}) \qquad (7.14б)$$

где g_i – это относительная частота i-го нуклеотида. Когда $g_i = 1/4$ и $v = 4rt$, вышеприведенные уравнения становятся идентичными уравнениям в модели Джукса-Кантора.

В этих рассуждениях мы рассматривали укорененное дерево. Однако, если использовать обратимую модель нуклеотидных замен для определения $P_{ij}(v)$, не обязательно учитывать корень (рис. 7.5Б). Обратимая модель означает, что процесс нуклеотидных замен между временем 0 и t остается одинаковым, не зависимо от того, рассматриваем ли мы эволюционный процесс в прямом или обратном направлении по времени. Математически условие обратимости обозначается как

$$g_i P_{ij}(v) = g_j P_{ji}(v) \qquad (7.15)$$

для всех i и j. Уравнения 7.14 удовлетворяют этому условию.

При использовании обратимой модели число нуклеотидных замен $(v_5 + v_6)$ между узлами 5 и 6 дерева А остается одинаковой независимо от положения корня 0. Поэтому мы переобозначаем $v_5 + v_6$ в дереве А на v_5 в дереве Б и рассчитываем I_k, учитывая, что эволюционное изменение начинается в некоторой точке дерева, например, в узле 5 (уточнение точки значительно упрощает расчет). Тогда уравнение 7.13 примет вид

$$I_k = g_{x_5} P_{x_5 x_1}(v_1) P_{x_5 x_2}(v_2) P_{x_5 x_6}(v_5) P_{x_6 x_3}(v_3) P_{x_6 x_4}(v_4) \qquad (7.16)$$

На практике, конечно же, мы не знаем ни x_5, ни x_6, поэтому правдоподобие будет суммой вышеприведенной величины по всем возможным нуклеотидам в узлах 5 и 6:

$$L_k = \sum_{x_5}\sum_{x_6} g_{x_5} P_{x_5 x_1}(v_1) P_{x_5 x_2}(v_2) P_{x_5 x_6}(v_5) P_{x_6 x_3}(v_3) P_{x_6 x_4}(v_4) =$$
$$\sum_{x_5} g_{x_5} \left[P_{x_5 x_1}(v_1) P_{x_5 x_2}(v_2)\right]\left[\sum_{x_6} P_{x_5 x_6}(v_5) P_{x_6 x_3}(v_3) P_{x_6 x_4}(v_4)\right] \qquad (7.17)$$

До этого момента мы рассматривали только один нуклеотидный сайт. На практике же мы должны рассматривать все нуклеотидные сайты, включая невариабельные. Так как правдоподобие (L) для целой последовательности является продуктом L_k для всех сайтов, *log* правдоподобия целого дерева становится равным

$$\ln L = \sum_{k=1}^{n} \ln L_k \qquad (7.18)$$

Теперь можно максимизировать $\ln L$, изменяя параметры v_i. Это вычисление производится численно по методу Ньютона или используя другие численные методы. Максимизация дает ML оценки длин ветвей v_i для этой топологии, но нас также интересует значение максимального правдоподобия для этой топологии. Теперь рассматриваем две оставшиеся топологии, возможных для четырех последовательностей, и рассчитываем значения ML для них. ML дерево представляет собой топологию с наибольшим значением ML. Длины ветвей для этой топологии, конечно же, даются ML оценками v_i, полученными для данной топологии.

Из этого примера ясно, что построение ML дерева является очень длительным процессом, потому что мы должны рассматривать все возможные нуклеотиды в каждом внутреннем узле. Число нуклеотидных комбинаций, подлежащих исследованию, для дерева из m таксонов составляет $4^{(m-2)}$, потому что в нем $m-2$ внутренних узлов. Если, например, $m=10$, то необходимо исследовать 65536 разных комбинаций нуклеотидов. Однако, если переписать уравнение 7.6 во вторую форму, то можно значительно сократить объем вычислений. Эта процедура называется подрезкой ветвей (pruning) (Felsenstein 1981). Однако, даже при применении такого рода алгоритмов объем вычислений достаточно велик, особенно при высокой степени дивергенции последовательностей. Кроме того число исследуемых топологий быстро увеличивается по мере роста m, как отмечалось ранее. Для $m=10$ это число равно 2027025. По этим причинам построить истинное ML дерево непросто при больших m. Поэтому были разработаны различные эвристические алгоритмы поиска, как в случае с MP методами.

В приведенных рассуждениях мы рассматривали простую модель нуклеотидных замен. Однако, практически такая же формулировка может быть использована практически для любой модели замен в случае обратимых во времени моделей. В общем случае, функция правдоподобия L для топологии может быть записана как

$$L = f(\mathbf{x};\theta) \qquad (7.19)$$

где $\mathbf{x}$ – это набор наблюдаемых нуклеотидных последовательностей, а θ – это набор параметров, таких как длины ветвей, частоты нуклеотидов и параметры замен в используемых математических моделях. В модели equal-input параметр скорости замен r_i сочетается со временем t_i ($v_i = r_i t_i$) и v_i используются в качестве параметров длин ветвей. Поэтому число параметров, подлежащих оценке, равно $2m-2$, если нуклеотидные частоты оцениваются из наблюдаемых частот. Если нуклеотидные частоты оцениваются методом ML, необходимы три дополнительных параметра с

условием $g_A + g_T + g_C + g_G = 1$. Если мы используем модель HKY, необходимо оценить один дополнительный свободный параметр (α/β). Все эти параметры могут быть оценены путем максимизации L для данного набора наблюдаемых данных.

Как отмечено ранее, максимизация $\ln L$ проводится численно, и поэтому полученное действительное значение ML зависит от используемого численного метода и желаемой степени погрешности. Поэтому разные компьютерные программы могут давать разные значения ML для одного и того же набора данных. Но все же относительные значения ML для разных топологий обычно остаются одинаковыми при небольшом числе последовательностей. При большом числе исследуемых последовательностей разница в значениях ML между разными топологиями может быть невелика и поэтому погрешность расчетного метода для ML становится важной. Было показано, что даже для одной и той же топологии для четырех последовательностей на поверхности правдоподобия может возникнуть два пика, и заметил, что это может стать настоящей проблемой при нахождении ML деревьев. Однако необходимы дополнительные теоретические исследования по этой проблеме для более определенных выводов.

7.3.1.2. Стратегии поиска деревьев наибольшего правдоподобия.

Так как поиск ML деревьев является очень трудоемким процессом, были предложены различные эвристические методы нахождения ML деревьев. Большинство из них схожи с методами для получения ME или MP деревьев, поэтому нет необходимости повторять их здесь. Однако, эффективности этих алгоритмов при получении корректных топологий не обязательно такие же, как при ME, MP и ML методах.

Например, метод разбиения звезды[71] (SD) концептуально схож с методом связывания ближайших соседей и начинается со звездообразного дерева, представленного на рис. 7.2Д. Звездообразное дерево раскладывается в бифуркационное дерево шаг за шагом как в случае алгоритма связывания ближайших соседей путем расчета ML значения на каждой стадии образования пар таксонов и выбирая пару соседей, которые дают наибольшее значение ML. Таким образом, можно получить SD ML дерево. Однако при анализе правдоподобия SD метод не является таким эффективным для получения корректной топологии как некоторые другие эвристические методы. Поэтому Nei et al. (1998) предложил, что применение CNI[72] поиска к SD ML дереву с несколькими циклами итераций может найти истинную топологию также часто, как и исчерпывающий алгоритм поиска ML дерева. Как и ME и MP методы, ML методы имеют тенденцию к получению некорректных топологий при больших m и малых n. Поэтому неблагоразумно тратить чрезмерное компьютерное время для поиска ML дерева. Что действительно важно, так это найти истинное дерево или дерево, близкое к нему, а не ML дерево.

7.3.2. Модели нуклеотидных замен.

7.3.2.1. Часто используемые модели.

Модель замен, представленная в уравнении 7.14, является простой и не учитывает различные составляющие факторы, такие как смещение транзиций/трансверсий. В связи с этим была предложена модель HKY, представленная в виде матрицы E в таблице 3.2. Элемент e_{ij} этой матрицы представляет собой мгновенную скорость замен от нуклеотида i (i-ый ряд) к j (j-ая колонка) ($i, j = \mathrm{A,T,C,G}$). Все элементы в каждом ряду суммируются

[71] star-decomposition method
[72] close neighbor interchange algorithm

до 0, так что диагональные элемент $e_{ii} = -\sum_j e_{ii}(i \neq j)$, хотя это и не представлено. Скорости транзиций и трансверсий от нуклеотида i к j равны αg_j и βg_j, соответственно. Поэтому отношение транзиций к трансверсиям дается выражением $R = \alpha/(2\beta)$. Эта модель становится идентичной модели equl-input при $\alpha = \beta$ и модели Кимуры при $g_A = g_T = g_C = g_G = 1/4$.

Модель Фельсенштейна (1984 год) может быть представлена в следующем виде:

	A	T	C	G
A		βg_T	βg_C	$(\delta/g_R + \beta)g_A$
T	βg_A		$(\delta/g_Y + \beta)g_C$	βg_G
C	βg_A	$(\delta/g_Y + \beta)g_T$		βg_G
G	$(\delta/g_R + \beta)g_A$	βg_T	βg_C	

Здесь, g_Y и g_R – это относительные частоты пиримидинов (T и C) и пуринов (A и G), то есть $g_Y = g_T + g_C$ и $g_R = g_A + g_G$. Параметр β обозначает скорость трансверсий, а δ/g_Y и δ/g_R – это параметры, оценивающие число транзиционных изменений, которые превосходят β. В этой модели отношение транзиций к трансверсиям дается выражением

$$R = (a_1 \delta/\beta + a_2)/a_3 \qquad (7.20)$$

где $a_1 = g_T g_C/g_Y + g_A g_G/g_R$, $a_2 = g_T g_C + g_A g_G$ и $a_3 = g_Y g_R$. Заметим, что эта модель сводится к модели Кимуры при $g_i = 0.25$ и $R = (\alpha/\beta + 0.5) = \alpha/2\beta$.

В программе DNAML пакета программ PHYLIP g_i оценгивается по наблюдаемым частотам в целых последовательностях, а R задается произвольно. Поэтому, единственными оцениваемыми путем максимизации правдоподобия параметрами являются длины ветвей. Такие компьютерные программы, как PAML и PAUP* включают в себя в добавок к двум вышеописанным разные модели замен, такие как модели Джукса-Кантора, Кимуры и Тамуры-Нея. Также они содержат общую обратимую модель

(REV)[73] (табл. 3.2G). Это наиболее общая модель, удовлетворяющая условиям обратимости, и включающая в себя 8 независимых параметров. Три из них относятся к частотам нуклеотидов (g_i с условием $\sum g_i = 1$), но эти параметры зачастую оцениваются на основе наблюдаемых частот. Пять параметров a, b, c, d и e должны быть оценены максимизацией правдоподобия. (f может быть положена равной 1).

7.3.2.2. Сравнение разных моделей.

Действительный паттерн нуклеотидных замен очевидно является довольно сложным, поэтому может показаться, что математическая модель с большим числом параметров лучше, чем модель с меньшим числом параметров для построения филогенетических деревьев. В реальности это не всегда так. Модель с многими параметрами лучше описывает данные, чем более простая модель, но статистическое предсказание (или оценка топологии), основанное на сложной модели подвержено большему количеству ошибок. Поэтому предпочтительней использовать простые модели, пока модель относительно хорошо представляет паттерн замен.

В случае ML методов качество соответствия модели наблюдаемым данным может быть исследовано с помощью теста соотношения правдоподобия или информационного критерия Акаике (AIC)[74]. Если существует две модели, модели 1 и 2, и модель 1 является особым случаем модели 2, говорят, что модель 1 вложена в модель 2. Когда корректная топология известна и модель 1 вложена в модель 2, можно сосчитать **log** соотношения правдоподобия по уравнению 5.13, где $\ln L_1$ и $\ln L_2$ – это значения ML для моделей 1 и 2, соответственно. Таким образом, мы можем протестировать, является ли модель 2 значительно лучше модели 1 или нет. Например, модель Кимуры является частным случаем HKY модели, первая

[73] general reversible
[74] Akaike's information criterion

имеет один свободный параметр, а последняя – четыре (заметим, что $g_A + g_T + g_C + g_G = 1$. Поэтому разница в соответствии между двумя моделями может быть оценена с помощью LR или χ^2 теста с тремя степенями свободы.

В общем случае, тест соотношения правдоподобия не может быть использован, если сравниваемы модели не являются вложенными. Однако, можно сравнить две невложенные модели используя AIC, если рассматриваемая топология остается неизменной. AIC определяется как

$$AIC = -2 \ln L + 2p \qquad (7.20)$$

где $\ln L$ – это значение **log** правдоподобия для данной модели, а p – число свободных параметров, подлежащих оценке. Предполагается, что статистическая предсказуемость модели тем выше, чем ниже AIC. Уравнение 8.9 показывает, что даже при малых $-\ln L$ AIC может быть высоким при большом числе свободных параметров и что модель с высоким значением ML и малым числом параметров лучше.

7.3.3. Методы правдоподобия для белковых последовательностей.

Когда последовательности ДНК относительно близкие друг к другу, методы правдоподобия ДНК хорошо работают, особенно, когда используются данные по первому и второму положениям кодонов. Однако, если последовательности являются эволюционно отдаленными белок-кодирующими генами, появляются сложности, так как скорость синонимических замен в целом намного выше, чем несинонимических. Относительные частоты четырех нуклеотидов в третьих положениях кодонов могут также сильно варьироваться от вида к виду, поэтому стационарная модель нуклеотидных замен может не обязательно соответствовать действительности. Напротив, эволюционные изменения белковых последовательностей не подвержены вышеприведенным проблемам, и

поэтому белковые последовательности имеют ряд преимуществ перед ДНК в случае относительно высокой дивергенции последовательностей. Кишино и соавт. предложили метод правдоподобия для белков, в котором используется эмпирическая транзиционная матрица Дэйхофф для 20 разных аминокислот. Позже, Адачи и Хасегава использовали разные транзиционные матрицы, включая модель Пуассона, эмпирическую транзиционную матрицу для ядерных белков Джоунса и собственную матрицу для митохондриальных белков. Они применили эти методы для разных последовательностей и получили достаточно хорошие деревья для нескольких групп организмов позвоночных.

8. СКОРОСТИ И ПАТТЕРНЫ НУКЛЕОТИДНЫХ ЗАМЕН.

8.1. Генетическая вариабельность.

Когда биологи хотят воссоздать эволюционную историю видов, то они смотрят на генетические сходства и отличия между изучаемыми видами. Однако, хотя каждый вид имеет единую генетическую сущность, существует генетическая изменчивость, или **вариация**, (genetic variation) среди отдельных особей одного вида. Так, одни люди имеют голубой цвет глаз, а другие коричневый, одни люди высокого роста, а другие низкого и т.д. Эти характеристики наследуются от родителей к детям. Генетическая изменчивость внутри вида представляет собой материал эволюции путем естественного отбора. Некоторые изменения могут обеспечить адаптивные преимущества особи, ими обладающей. Такие особи будут иметь большую приспособленность в смысле эволюции по Дарвину и оставят больше потомства, чем те, которые не несут данное преимущество. Например, особи некоторого вида хищников, которые более эффективно охотятся за своей добычей, будут иметь селективное преимущество. Аналогично, особи вида, являющегося добычей для хищников, наиболее эффективно избегающие схваток с хищником, будут поддерживаться естественным отбором. Со временем такие черты, несущие какие-либо преимущества, могут стать доминирующими в виде, так же как и гены, лежащие в основе этих фенотипических проявлений.

Большинство видов существует в виде множества в разной степени географически разделенных популяций, а не в виде одной единой популяции. В результате в каждой отдельной популяции одного и того же вида часто развиваются разные генетические изменения. Такие изменения могут распространиться в другие популяции одного вида только в случае, если две особи из разных популяций оставят потомство. Если же одна популяция будет полностью изолирована от другой, например, отделена речной или

горной системой, то между ними может накопиться существенное генетическое различие, что может вылиться в образование подвида, или даже нового отдельного вида.

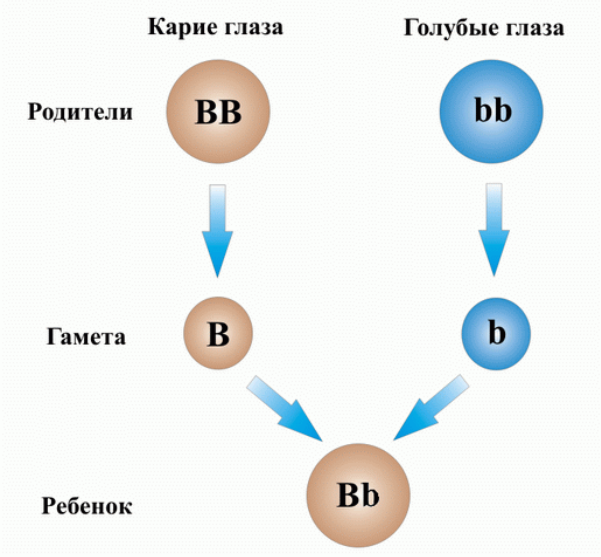

Рис. 8.1. Пример вариации, основанной на генетическом различии в цвете глаз

С момента появления популяционной генетики как научной дисциплины в начале 20 века, ее целью было описание и объяснение генетической вариабельности внутри и между популяций. Эта цель остается такой же и по сей день, но с некоторыми изменениями. Начиная с 1960-х годов, у популяционных генетиков появилась возможность измерять генетическую вариабельность напрямую. Один из примеров генетической вариабельности является цвет глаз (рис. 8.1). Различные варианты гена, кодирующего пигмент радужной оболочки глаза, определяют 2 варианта цвета глаз: голубой или карий. Такие варианты гена называются аллели, и для каждого гена мы наследуем два аллеля, один от одного родителя, другой от другого. Любая наследуемая черта является результатом каждой такой пары аллелей, каждый из которых может быть доминантным или рецессивным. Например, карие глаза могут быть как из пары “карих”

аллелей, так и из одного “карего” и одного “голубого”, потому что “карий” аллель является доминантным, а “голубой” – рецессивным. Таким образом, для того чтобы глаза были голубыми, необходимо чтобы оба аллеля были “голубыми”. Эти и другие аллели присутствуют с разными частотами в разных популяциях: так, голубые глаза часто встречаются, например, в северной Европе, и довольно редко встречаются среди населения Восточной Азии.

С развитием методов молекулярной биологии стали рассматривать не только полиморфизм морфологических признаков, но и полиморфизм белков и ДНК, и тут возник вопрос о механизмах появления полиморфизма. Существовало две точки зрения:

1. **Селекционисты**. Полиморфизм – продукт естественного отбора, представляет собой активное накопление вариантов, являющихся адаптивно преимущественными при разных условиях окружающей среды, в которых организм был беззащитным. Другими словами, значительная часть генетических мутаций фиксируется в популяции, так как они полезны. Вредоносные мутации элиминируются, так как несущая их особь обладает меньшей приспособленностью.

2. **Нейтралисты**. Генетическая вариабельность – просто пассивное накопление случайных мутаций, приводящих к появлению новых аллелей, большинство их которых адаптивно нейтральны – они не усиливают, но и не уменьшают приспособленность организма. С этой точки зрения принципиальная роль естественного отбора – удаление редких вредоносных аллелей.

Для селекционистов, поэтому, большинство мутаций либо благоприятные, либо губительные; благоприятные мутации сохраняются в популяции, создавая широкую вариабельность, в то время как вредоносные мутации элиминируются. Для нейтралистов большинство мутаций адаптивно нейтральны, и они закрепляются (фиксируются) в популяции, так как их присутствие не приносит вреда; обширная генетическая вариабельность

является результатом. Противопоставление этих двух точек зрения известно как дебаты нейтралистов и селекционистов **(neutralist-selectionist debate**).

8.2. Нейтральная теория эволюции

В 1965 году Цукеркандл и Паулинг опубликовали данные об эволюции молекул гемоглобина, полученные путем сравнения аминокислотных последовательностей гемоглобинов из нескольких видов. Было показано, что аминокислотные замены происходили с постоянной скоростью, и эта скорость была довольна высока. Эти и последующие данные по сравнению аминокислотных последовательностей белков натолкнули японского генетика М.Кимуру на создание **нейтральной теории эволюции**, которая была опубликована в 1968 году. Природу генетической вариабельности Кимура объяснил тем, что альтернативные аллели, быстро накапливающиеся в популяции в большинстве своем селективно нейтральны, или точнее селективно эквивалентны. Поэтому они не влияют на функционирование белка, и таким образом, они являются невидимыми для естественного отбора, так не несут преимущества друг над другом. Даже когда в популяции существует несколько эквивалентных аллелей, тем не менее, один из них через какое-то время станет более распространенным, чем другие, или вовсе их вытеснит. Селективно эквивалентные аллели, таким образом, зафиксируются в популяции путем случайных процессов, а не отбора. Этот процесс случайного распространения генетических мутаций в популяции называется **генетическим дрейфом**[75].

Представим такой эксперимент, иллюстрирующий данный процесс. Представим себе мешок, в котором находятся 100 шариков, 50 синих и 50 красных. Теперь вслепую достаем из этого мешка 50 любых шаров. Допустим, что наугад мы достали 25 синих и 25 красных шаров, однако мы могли бы достать любое число шаров, скажем 30 синих и 20 красных и т.д.

[75] genetic drift

Теперь представим себе, что оставшиеся шары реплицируют (воспроизводят) сами себя, и, в конце концов, получаем опять 100 шаров (это эквивалентно популяции организмов с постоянной численностью). Если удалить 30 синих шаров, новая популяция шаров будет состоять из 40 синих и 60 красных (это эквивалентно тому, что случайным образом “красные” особи произведут больше потомства, чем “синие” особи). Если повторять этот процесс много раз, то относительная частота синих и красных шаров будет флуктуировать из-за случайности данного процесса. Вскоре, однако, существует высокая вероятность того, что популяция будет образована из шаров только одного цвета. Такие же процессы происходят и в естественных популяциях (рис. 8.2).

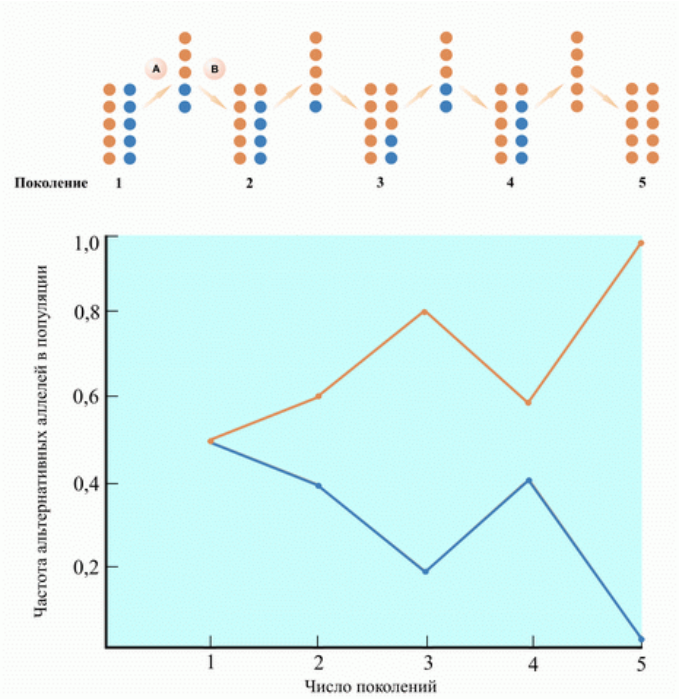

Рис. 8.2. Генетический дрейф.

Теория нейтральности, однако, не отрицает роль отбора, как источника вариабельности, она просто сводит эту роль к минимуму. Так, М.Кимура писал: “В соответствии с теорией, большинство эволюционных мутационных замен на молекулярном уровне вызваны случайной фиксацией, посредством выборочного дрейфа, селективно нейтральных (т.е. селективно

эквивалентных) мутантов под беспрерывным мутационным давлением. Это находится в остром противоречии с традиционной неодарвинистской теорией эволюции, которая провозглашает, что распространение мутантов внутри вида в процессе эволюции может происходить только с помощью положительного естественного отбора".

Из-за математической простоты, теория нейтральности позволяет сделать количественные предположения о вариабельности, а именно, касающиеся ожидаемой степени вариабельности в популяции, скорости, с которой вариабельность накапливается, и условий, при которых эта скорость может быть максимальной.

По теории нейтральности степень вариабельности, ожидаемая в популяции, является простой функцией от скорости мутаций и эффективного размера популяции. Когда это уравнение используется для расчета вариабельности, или гетерогенности, в большинстве популяций результат оказывается обычно выше, чем для вариабельности, наблюдаемой в действительности. Нейтралисты объясняют это двумя причинами: 1) Аллели нужно рассматривать не как строго нейтральные, а как примерно нейтральные, или слегка вредоносными (slightly deleterious), и отбор будет элиминировать эти слегка вредоносные аллели, таким образом, снижая гетерогенность популяции, и 2) Эффективный размер популяции обычно переоценивается, так как размер популяций обычно флуктуирует в течение времени, а иногда и довольно сильно по причине эпидемий, резких изменений условий среды и т.п., что не учитывается при оценке эффективного размера популяции. Однако, вторая причина является сложно доказуемой.

Понятие скорости накопления генетической вариабельности является центральным в теории нейтральности. Как будет показано дальше, оно ведет напрямую к идее молекулярных эволюционных часов. По теории эволюции **скорость накопления вариабельности** просто определяется скоростью мутаций (то есть фиксации нейтральных аллелей) для определенной

молекулы. (Идея, что вариабельность накапливается с постоянной скоростью, зависящей только от скорости мутаций, является прямо противоположенной Дарвинистской точке зрения, согласно которой отбор является основным арбитром, сохраняющим или отвергающим исключительно нейтральные продукты мутации).

Молекулы, отличаются по степени модификации, которую они могут себе позволить: например, гистоны (белки, участвующие в организации молекул ДНК в хромосоме) обладают малой толерантностью по отношению к структурной вариабельности, и поэтому обладают низкой скоростью фиксации мутаций; гемоглобин, наоборот, может претерпевать значительные изменения (по крайней мере, в некоторых частях молекулы) и поэтому обладает большей скоростью мутаций. В основе различной толерантности к мутациям лежат две предпосылки теории нейтральности: 1) Разные генетические локусы в одном и том же организме будут накапливать мутации с разными скоростями; 2) Один и тот же ген в разных видах будет обладать одинаковой скоростью фиксации мутации (однако, в реальности скорости замен в одном и том же белке в разных видах зачастую отличаются, и эта предпосылка является слабым местом теории нейтральности); 3) Молекулярная эволюция является достаточно консервативной, поэтому функционально важные молекулы или части молекул будут изменяться менее быстро, чем функционально более важные.

8.3. Примеры, подтверждающие нейтральную теорию эволюции.

1. Замена нуклеотида может приводить к разным последствиям в зависимости от положения в кодоне. Так, нуклеотидная замена в первом или во втором положении кодона почти всегда приводит к замене аминокислоты (несинонимическая замена), замена же в третьем положении кодона обычно не изменяет аминокислоту, кодируемую данным кодоном (синонимическая

замена). Известно, что мутации, приводящие к синонимическим заменам нуклеотидов, происходят со скоростью примерно в 2 раза большей, чем замены, приводящие к несинонимическим заменам, как и предсказывает теория нейтральности.

2. У эукариотических организмов гены имеют мозаичное строение и состоят из белок-кодирующих областей (экзонов) и некодирующих областей (интронов), которые не вносят информацию в конечный белковый продукт. Мутации в интронах (а точнее в тех участках интронов, структура которых не несет функциональной нагрузки) накапливаются быстрее, чем в экзонах, и при некоторых обстоятельствах они накапливаются со скоростью, сравнимой со скоростью синонимических замен в кодонах.

3. Другой пример представляют собой **псевдогены**. Это участки ДНК, происходящие из функционирующих генов путем дупликаций или реакции обратной транскрипции. Впервые исследования замен в псевдогене проводились на певдогене глобина в начале 1980-х годов. Оказалось, что скорости замен в псевдогене по сравнению с действующим геном глобина были в 5 раз выше. Более того, не наблюдалось отличий в скоростях между тремя позициями в кодонах псевдогена.

4. Интересный пример представляют собой гены, освободившиеся от функциональных ограничений. В частности в конце 1980-х годов был изучен ген альфаА-кристаллина ближневосточного крота *Spalax ehrenbergi*. Этот ген кодирует белок хрусталика глаза. Кроты большинство времени своей жизни проводят под землей и практически не сталкиваются с дневным светом. Их глаза развились практически до рудементного состояния, и животные практически слепые. По теории нейтральности, если генный продукт не нужен, то ген альфаА-кристаллина будет накапливать мутации с большей скоростью, чем эквивалентные гены из видов, обладающих зрением. Оказалось, что это действительно так, скорость была в 4 раза выше. Хотя эта скорость и не так велика, как в псевдогенах, но в отличие от псевдогенов ген альфаА-кристаллина экспрессируется (т.е. транскрибируется и транслируется

в полноценный белковый продукт). Ген альфаА-кристаллина слепых кротов является частью координированной развивающейся системы генов, которые кодируют определенную структуру – глаз, хотя и рудементальный.

Основоположник теории нейтральной эволюции Кимура говорил, что нейтральная эволюция справедлива, когда мы рассматриваем эволюцию на молекулярном уровне. Но на более высоких уровнях – организма и популяций – отбор, очевидно, играет большую роль в эволюции. Поэтому относительный вклад отбора и нейтральности отличается на разных уровнях эволюции.

8.4. Гипотеза молекулярных часов.

Гипотеза **молекулярных часов** гласит, что скорость аминокислотных или нуклеотидных замен *примерно* постоянна в течение времени эволюции, хотя действительное число замен подвержено стохастическим ошибкам. Строго говоря, ни один ген или белок не будет эволюционировать с постоянной скоростью в течение длительного времени эволюции, потому что функция гена, вероятно, изменится со временем, особенно, когда число генов в геноме увеличивается от простых к более сложным организмам, либо когда изменяются условия окружающей среды. Механизмы повреждений ДНК и их репарации также могут варьироваться среди разных групп организмов.

По вышеперечисленным причинам, бесполезно пытаться найти гены, в которых реализуются идеальные молекулярные часы. Однако, молекулярные часы и не должны быть универсальными. Даже если они работают для определенной группы организмов, то они все равно полезны для изучения эволюционных отношений между организмами или для оценки времен дивергенции организмов внутри этой группы.

После того как виды дивергировали от общего предка, они накапливали мутации с постоянной скоростью, постепенно становясь генетически удаленнее друг от друга. Сравнивая генетические отличия

между двумя родственными видами, тем самым, в принципе можно сосчитать время, истекшее с тех пор, как они отделились от общего предка. Молекулярные часы, таким образом, придают временную размерность филогенетическим деревьевам, которые показывают связи между видами.

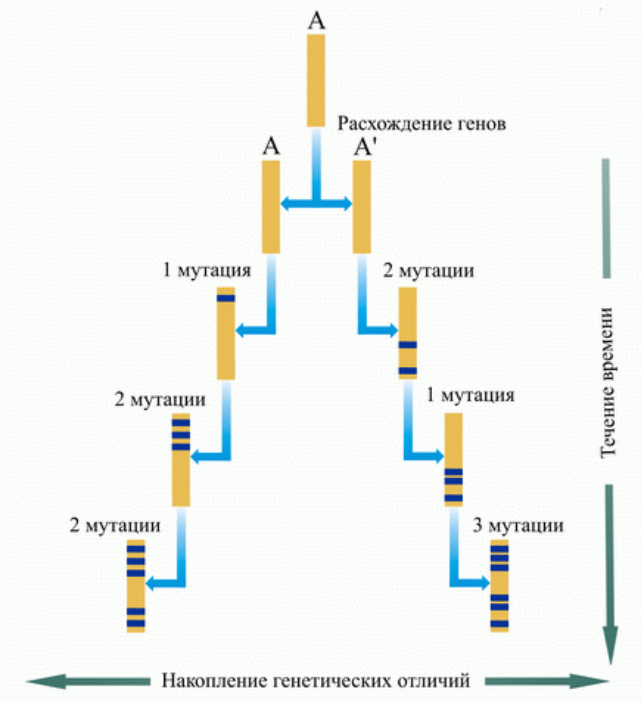

Рис. 8.3. Постоянство средней скорости мутаций.
После события расхождения видов гены будут накапливать мутации независимо в каждой из линий. Здесь рассматриваются мутации в генах A и A' в течение времени. Хотя скорость мутаций не является постоянной и равной в двух линиях, средняя скорость будет примерно одинаковой; в данном случае она равна 5.5 заменам в представленный период времени, полная дивергенция между последовательностями A и A' составляет 11 мутаций.

Однако следует отметить, что даже если мутации происходят строго нейтрально, молекулярные эволюционные часы не будут регулярными, тикая регулярно год за годом (рис. 8.3). Наоборот, это будут стохастические часы, следующие за вероятностью мутации в определенной молекуле год за годом (рис. 8.4). Усредненные по времени, стохастичекие часы такого типа, тем не менее, могут быть чрезвычайно точными. Регулярные часы ходят регулярно: допустим, что один раз в тысячу лет, тогда по прошествии 5 миллионов лет

произойдет 5000 ударов часов через одинаковые промежутки времени. Стохастические часы не регулярны, по крайней мере, за короткий промежуток времени. В этом примере они могут сделать всего лишь 500 ударов за первый миллион лет, 1500 за второй, 1000 за третий, 300 за четвертый и 1700 за пятый. Но к концу этого периода, однако, стохастические часы в общей сложности сделают такое же количество ударов (то есть эволюционных изменений в виде мутаций), что и регулярные. Таким образом, среднее число изменений будет равно за 5 миллионов лет. Важным является то, что стохастические часы становятся точнее по мере увеличения измеряемого временного периода. Причина этого такая же, как и в случае с подбрасыванием монеты. После четырех подкидываний, вполне вероятно, что, например, 2 раза выпадет “решка” и 4 раза выпадет “орел”, то есть в 33% случаев выпадает “решка и в 66% – “орел”. Однако после тысячи подбрасываний монеты распределение будет 50:50, что является статистической вероятностью. С увеличением времени (или бросков в случае с монетой) несимметричные шансы выравниваются. Логично, что при малых временных промежутках, стохастические часы будут неточны.

Даже если строгая нейтральность в накоплении мутаций не применима – если, наоборот, отбор играет важную, хотя и флуктуирующую, роль – применение молекулярных часов все равно возможно. Усредненные по времени, даже флуктуирующие скорости мутаций, могут дать полезную оценку эволюционного времени. Если требуется 99% временной точности, то применение молекулярных часов не будет правомочно, однако, если достаточно и 80 % точности – как часто случается в биологии, где не существует никаких других средств измерения этого времени – то применение таких часов, пусть и дающих большую погрешность, может дать удовлетворительные результаты.

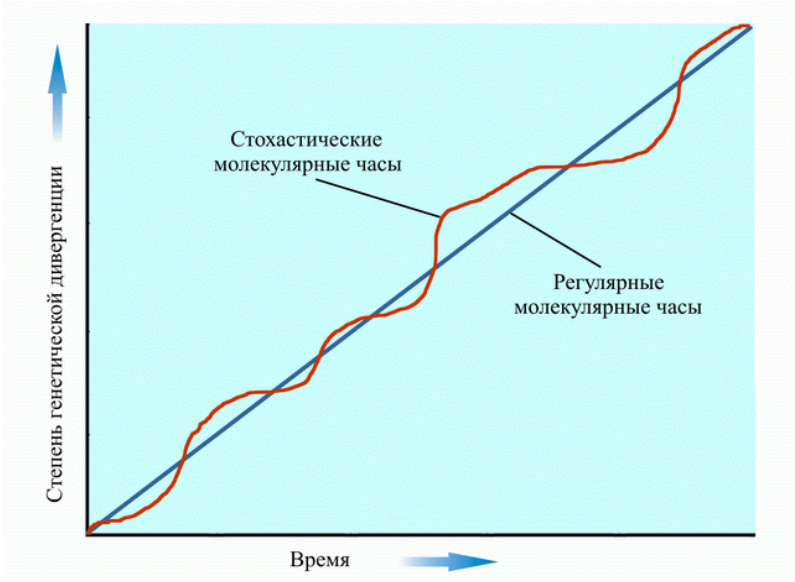

Рис. 8.4. Стохастичность молекулярных часов.

Конечно, разные участки одного гена (1-е, 2-е и 3-е положения в кодоне, активные центры белка и менее важные области этого белка), разные гены одного организма (например, гистоны и рибосомальные гены с одной стороны и псевдогены с другой) и разные геномы в одном организме (ДНК митохондрий, как правило, эволюционирует в 10 раз быстрее ядерной ДНК, которая, в свою очередь, эволюционирует в несколько раз быстрее ДНК хлоропластов) могут подвергаться мутациям с абсолютно разными скоростями (рис. 8.5). Эти отличия в большинстве случаев могут быть объяснены в рамках теории нейтральности. Более того, различные скорости, с которыми накапливаются мутации в разных типах ДНК, могут быть использованы для решения задач при разных временных шкалах. Исследователю нужно просто выбрать молекулярные часы с подходящей шкалой. Так, например, если изучаемый филогенетический вопрос требует проведения сравнения данных на протяжении миллионов лет, то наиболее подходящими будут очень медленно идущие часы – рибосомальные гены, например. И так далее.

Главным вопросом теории нейтральности является не то, мутируют ли разные гены с разными скоростями, а то, мутируют ли одни и те же гены с постоянной скоростью в разных линиях или нет. В настоящее время ясно, что

скорости мутаций действительно варьируются среди линий, но не так сильно, как можно было бы ожидать.

Рис. 8.5. Разница в скорости накопления мутаций в разных областях ДНК

Например, сравнение 20 разных видов путем ДНК-ДНК гибридизации, проведенное Р.Бриттеном, показало пятикратную разницу в скорости мутаций. Самые низкие скорости мутаций оказались у высших приматов (особенно у приматов и человека, или гоминидов), а также в некоторых линиях птиц. Скорости замен значительно выше у грызунов, морских ежей и *Drosophila*. Аналогичные данные были получены Бриттеном и при исследовании молчащих замен в 25 генах из 29 видов. Эти данные можно считать типичными. Почему скорости варьируются? Существует несколько возможностей для ответа на этот вопрос:

1) Группы с низкой скоростью мутаций имеют более эффективные механизмы репарации ДНК, которые избавляют от ошибок, происходящих при репликации.

2) Виды с коротким временем жизни одного поколения – как, например, мыши в сравнении с человеком – будут иметь более высокие скорости мутаций, так у них существует больше возможностей для совершения ошибок при переходе генов от поколения к поколению. Действительно, скорость мутаций у мышей в 5 раз больше, чем у человека, однако это намного меньше почти стократной разницы во времени жизни одного поколения. Эту разницу можно объяснить, тем, что она является последствием непрерывного оборота гамет (в особенности мужских гамет, или сперматозоидов), который всегда имеет место в независимости от

времени жизни поколения. Сперматозоиды производятся постоянно в обоих видах, что позволяет накапливаться ошибкам постоянно, а не только в момент передачи[76].

3) Скорость метаболизма в организме также может влиять на скорость мутаций. Так при анализе определенных последовательностей митохондриальной ДНК у некоторых видов акул было показано, что скорость замен в 7 – 8 раз ниже, чем у приматов и копытных. Время одного поколения у всех этих видов примерно одинаково, но скорость метаболизма у акул в 5 – 10 раз ниже, чем у млекопитающих примерно такого же размера.

Было признано, что скорость мутаций в данном гене неминуемо замедлится при достаточном эволюционном времени. Причина заключается в том, что мутации, являющиеся стохастическим процессом, обладают определенной вероятностью повторного появления в одном и том же локусе; это называется многократным попаданием[77]. С течением времени число многократных попаданиий неминуемо увеличивается, уменьшая тем самым наблюдаемую скорость мутаций. Расчеты с использованием молекулярных часов в плановом порядке принимают эту проблему в расчет статистической корректировкой. В реальности, однако, все несколько сложнее, так как не все локусы в гене не одинаково уязвимы для мутаций, и сама уязвимость может меняться в зависимости от мутаций в локусе. Многократные попадания будут происходить более часто в гене, в котором есть особенно уязвимые для мутаций локусы, чем в гене, который одинаково уязвим, что опять уменьшает наблюдаемую скорость мутаций. Все эти возможности и условия должны учитывать при применении идеи молекулярных часов для анализа каких-либо данных.

В настоящее время можно сказать, что молекулярные эволюционные часы оказались не настолько хороши, как многие надеялись, но и не настолько плохи, как многие опасались. Очевидно, что существует несколько

[76] transmission
[77] multiple hit

“глобальных часов” – часов, которые действуют при больших временных шкалах и большом наборе данных. Кроме того, очевидно, что можно находить и “локальные часы”, которые можно с достаточной степенью достоверности использовать для ограниченных временных масштабов и линий. Достоверность можно проверить с помощью теста относительной скорости[78].

Противоречивость идеи молекулярных часов:

1. Постоянная скорость эволюции была неприемлема для классических эволюционистов, изучающих эволюцию морфологических признаков. Во времена расцвета синтетической теории эволюции или неодарвинизма, считалось, что скорость эволюции зависит от изменений окружающей среды и естественного отбора, и поэтому не может быть постоянной.

2. Механизмы, лежащие в основе постоянства скорости аминокислотных замен, были не ясны. Было показано, что, если большинство аминокислотных или нуклеотидных замен происходят в результате нейтральных мутаций и генетического дрейфа, а скорость нейтральных мутаций постоянна в год, молекулярные часы могут быть объяснены. Однако, такое объяснение было неприемлемо для большинства генетиков и эволюционистов, которые считали, что скорость замен скорее постоянна за одно поколение, чем за один год.

3. Большинство мутаций, рассматриваемых в классической генетике, являются результатом делеций или мутаций, приводящих к сдвигу рамки считывания, происходящих при мейозе, а негубительные мутации (например, мутации, приводящие к устойчивости к фагам у бактерий) происходят с постоянной скоростью за единицу хронологического времени, а не за время клеточного деления. Однако данных по негубительным мутациям было слишком мало, для того чтобы убедить большинство генетиков и эволюционистов того времени.

---

[78] relative rate test

4. По мере накопления данных по аминокислотным заменам число случаев, в которых молекулярные часы были неприменимы, возрастало.

5. Времена дивергенции, полученные по данным палеонтологических исследований, зачастую ошибочны, и такая недостоверность вносит ошибки в изучение молекулярных часов. Позже был предложен тест относительной скорости замен[79] с использованием трех последовательностей, одна из которых является внешней группой[80]. В этом тесте оценки геологических времен не требуется, что позволяет избавиться от ошибок палеонтологических данных.

[79] relative rate test
[80] outgroup

9. ЭВОЛЮЦИЯ ПОСРЕДСТВОМ ДУПЛИКАЦИИ ГЕНОВ.

9.1. Дупликации

В 1970 году Оно постулировал, что единственным способом образования новых генов является дупликация уже существующих генов. Важность эволюции путем генных дупликаций, впервые убедительно доказанная Оно, сейчас принята повсеместно. Больше трети типичного эукариотического генома состоит из дуплицированных генов и генных семейств. Таким образом, генные дупликации играют ключевую роль в эволюции геномов. После дупликации, эволюционное давление на дубликат гена ослабевает. Это обеспечивает функциональную диверсификацию (разнообразие) дубликатов и биохимические нововведения путем мутаций и рекомбинации.

Существует пять основных типов дупликаций:

1) Частичная дупликация гена
2) Дупликация всего гена
3) Частичная дупликация хромомомы
4) Дупликация всей хромосомы (анеуплодия)
5) Дупликация всего генома (полиплодия)

Первые четыре типа дупликаций являются региональными, так как не затрагивают всего гаплоидного набора хромосом.

Рассмотрим вопрос о скорости потери гена после дупликации. Во-первых, необходимо различать два процесса: потерю гена после дупликации целого генома (полиплоидизация) и потерю гена после дупликации единичного гена. Были проведены исследования потери гена после полиплоидизации. В частности рассматривали, сколько дуплицированных генов долгое время (более 50 миллионов лет) оставались в геноме после его дупликации. Исследования проводились с применением широкого круга методов, от изучения электрофоретического разделения изоферментов в

дуплицированном локусе до анализа последовательностей всего генома. Разброс ответов был широким. От 50% до 92% всех дуплицированных генов, в конечном счете, были потеряны, в зависимости от методов изучения. При доступности последовательностей всего генома, оценки потери существенно выше, чем 50%. Например, у дрожжей порядка 92% всех генов могли быть потеряны после дупликации генома, произошедшей 100 миллионов лет назад.

Что касается дупликаций одного гена, то большинство исследований приводили к выводу, что большинство генных дубликатов теряются. Так как оба дубликата имеют идентичные функции после дупликации, один их них может свободно деградировать через мутации, приводящие к потере функции. Вопрос только в том, когда это произойдет? Существующие модели могут переоценивать скорость потери гена из-за ограниченной роли, придаваемой в них функциям гена. Ген либо функционирует правильно, либо приобретает вредоносную мутацию. Однако, гены могут иметь несколько функций, на каждую из которых мутации могут действовать независимо, и при этом эти мутации не обязательно будут вредоносными, так как доступны две копии гена. Результатом таких частично деградирующих мутаций в одной из двух копий гена является то, что дубликат гена развивает перекрывающиеся функции от изначально идентичных функций. И в противовес потери гена с полностью избыточной функцией, потеря гена с перекрывающимися функциями не может пройти так просто. Поэтому простая модель функционирования гена по принципу «все-или-ничего» может сильно переоценить скорость потери гена.

Важную роль в процессе эволюции играют частичные дупликации генов, в частности дупликации экзонов, которые могут проявляться в удвоении доменов в белке. **Белковый домен** – хорошо определяемый участок молекулы, который либо осуществляет определенную функцию (например, связывание субстрата), либо составляет стабильную компактную структурную единицу в пределах молекулы, хорошо отличаемую от прочих

частей. Существует два типа доменов: **функциональные** и **структурные**. Экзоны структурного гена не всегда строго соответствуют структурным доменам, не говоря уже о функциональных доменах.

Одни и те же аминокислоты, которые входят в состав разных структурных доменов, могут входить в состав одного и того же функционального домена. Если есть удвоение домена, то это в большинстве случаев свидетельство удвоения экзона. Дупликация доменов в ходе эволюции происходила часто. Удлинение генов, сопровождающих дупликацию доменов, – один из важнейших шагов эволюции сложных генов из простых. Ген может удлиниться заменой стоп-кодона на кодирующий кодон, транспозицией, сплайсингом. Но такие варианты удлинения, скорее всего, нарушат функции гена или его продукта. Фенотипически обнаруживается у больных с наследственной патологией. В ходе эволюции удлинение генов в основном зависело от дупликации экзонов. Все гены, принадлежащие к определенно группе повторяющихся последовательностей, относят к одному **семейству** генов (мультигенному семейству). Члены одного семейства часто, но не всегда, располагаются на одной хромосоме. Если дуплицированные гены в ходе эволюции приобретают слишком много различий, то вводят надсемейства для демаркации родственных и близкородственных групп. Белки, имеющие сходство аминокислотных последовательностей 50% и выше относят к одному семейству, а белки (гомологичные) со сходством менее 50% считаются продуктами генов, относимых к надсемейству. Например α и β глобины относят к двум разным семействам, но вместе с миоглобином они образуют надсемейство.

9.2. Гомология между генами.

Когда сравниваются одинаковые гены из разных видов, то между ними будет наблюдаться определенная степень сходства: скажем, 50 процентов последовательности идентична. Изначально молекулярные биологи,

описывая такие гены, говорили, что они имеют 50 процентов гомологии друг с другом, в то время как они имели в виду 50 процентов идентичности. Два гена могут быть не гомологичными вовсе, так как они не имеют общего происхождения. Термин “**гомология**” должен употребляться в случае общего происхождения, а термин “**идентичность**” должен употребляться в случае сходства последовательностей.

Поиск подлинной гомологии между генными последовательностями гораздо более сложный процесс, чем это может показаться. В двух относительно недавно дивергировавших видах, может вовсе не быть ни одного отличия в последовательностях, или всего лишь несколько. Выравнивание этих последовательностей даст очень высокий уровень идентичности последовательностей, который может быть принят за показание гомологии. Но по прошествии эволюционного времени две гомологичные последовательности будут независимо накапливать мутации, и уровень их идентичности будет снижаться. Теоретически возможно, что при достаточном времени уровень идентичности не будет превышать уровень, который может появиться случайно по нескольким интересным причинам.

На молекулярном уровне функционально важным элементом белков является их третичная структура, которая определяет их взаимодействие с другими молекулами, такими как ДНК, РНК, другие белки, углеводы и жиры. Белковые молекулы с очень сходными третичными структурами могут быть построены из совершенно отличных друг от друга аминокислот (которые, соответственно, кодируются разными последовательностями ДНК генов). С точки зрения эволюции, таким образом, один и тот же ген в двух дивергировавших видах может претерпеть существенные мутации, и при этом поддерживать функциональную целостность кодируемых белков. Такие гены являются гомологичными, несмотря на неидентичность их последовательностей ДНК.

Следующая трудность возникает, потому что мутации заключаются не только в замене одного нуклеотида на другой в специфических сайтах, но так

же и в делециях и вставках. Выравнивание таких последовательностей требует корректной вставки гэпов.

Геномы высших организмов организованы далеко не так просто, как, скажем, геномы бактерий. Например, у простейших организмов гены представлены единственной копией, тогда как у высших организмов это далеко не так, что делает понятие гомологии более сложным. Многие гены высших организмов представлены в геноме несколькими копиями, и такие гены образуют генные семейства. Иногда, члены генного семейства идентичны друг другу, и клетка может продуцировать большое количество одинакового продукта за короткий период времени (например, рибосомальные гены, кодирующие белок-кодирующую машинерию). Однако часто члены одного семейства отличаются друг от друга и могут выполнять слегка отличающиеся функции. Например, глобин приматов (белок, переносящий кислород в эритроцитах) существует в виде семейства из полдюжины генов, являющихся слегка отличающими вариантами друг друга, каждый из которых функционирует в разные периоды развития организма. Такие семейства появляются в процессе эволюции путем дупликации существующего гена. Исходно последовательности оригинального гена и его копии будут идентичными, но через некоторое время они накопят некоторые отличия. Некоторые генные семейства состоят из двух членов, некоторые из десятка.

Для распознавания генных семейств было введено два новых термина. Для генов, дупликация которых не произошла (то есть для генов, представленных единственными копиями) был введен термин ортология для проведения сравнений между родственными видами; **ортология** функционально эквивалентна традиционной морфологической гомологии. Термин **паралогия** был введен для описания связей между членами генного семейства внутри видов.

Существование паралогий может приводить к потенциальным ошибкам в молекулярной филогенетики, которые приводят к ошибочным выводам об

эволюционной истории родственных организмов. Рассмотрим гипотетический пример. Представим себе предковые виды, несущие ген X. Теперь предположим, что ген дуплицировался 10 миллионов лет назад, и теперь виды несут генное семейство, X1 и X2. Изначально идентичные по последовательности X1 и X2 постепенно будут накапливать независимые мутации внутри вида в течение 10 миллионов лет. Далее предположим, что виды разошлись 5 миллионов лет назад и образовались два дочерних вида, каждый из которых содержит генное семейство их двух генов. В конце концов, предположим, что молекулярные филогенетики хотят узнать эволюционную историю дочерних видов, исследуя последовательность гена X, не зная о том, что ген представлен двумя вариантами. Если, случайно, ген X1 будет изолирован из одного дочернего вида и ген X2 из второго вида, то предсказанная эволюционная история будет искаженной. Так, сравнение двух последовательностей, используя их отличия, как индикатор времени с момента эволюционного расхождения, покажет, что виды дивергировали 10 миллионов лет назад – но это время дупликации гена, а не дивергенции видов. Такой вывод будет отображать генное дерево, а не видовое.

Термины ортология и паралогия, таким образом, относится к гомологии при рассмотрении истории генов в родственных видах или внутри одного вида. Третья форма гомологии, также применимая на молекулярном уровне, называется **ксенология**. (Греческий корень *xeno-* означает “инородный”, “чуждый”). Хотя большинство генетических изменений подвержены вертикальному наследованию через последующие поколения, остающиеся изолированными от генов других видов, но бывают и исключения. Так, гены могут переноситься горизонтально между видами, часто посредством вирусов, и встраиваться в хозяйский геном (вирус встраивается в хозяйский геном для репродуктирования, случайным образом вирусная репликация может захватить гены хозяина, или оставить в геноме свои гены или гены, захваченные от предыдущих хозяев). Это частый процесс у микроорганизмов, но редко встречается у высших организмов.

Именно для таких генов вводится термин ксенология, то есть для генов из филогенетически удаленных видов с удивительно высокой степенью идентичности последовательности, в то время как все остальные гены этих видов сильно отличаются друг для друга.

Так как гены высших организмов имеют экзон/интронную структуру, то в процессе эволюции некоторые гены могут быть образованы из экзонов других генов (перетасовка экзонов[81]). Такой процесс приводит к еще одной форме гомологии, невозможной в точки зрения морфологической гомологии, частичной гомологии. Например, ген, кодирующий активатор тканевого плазминогена у млекопитающих, составлен из экзонов из генов, кодирующих другие белки: плазминоген, фибронектин и ростовой фактор эпидермия. Это означает, что ген активатора плазминогена частично гомологичен (точнее паралогичен) каждому из этих генов.

Следует также учитывать существование псевдогенов, так как они вносят дополнительные потенциальные сложности в филогенетический анализ. Предположим, что ген А в предковом виде дуплицировался, и образовались паралогичные гены А1 и А2. Как и ранее, предполагаем, что из предкового вида сформировались два дочерних вида, каждый из которых содержит оба гена А1 и А2. Теперь предположим, что в одном дочернем виде ген А1 стал псевдогеном. Молекулярные филогенетики, ничего не знающие об истории этих генов, невольно будут сравнивать последовательности этих генов как ортологические, а не паралогические, какими они являются на самом деле. Таким образом, исследователь построит генное дерево (на основе события дупликации, в результате которого образовалось генное семейство из двух генов), а будет полагать, что построил видовое дерево (на основе расхождения предкового организма на дочерние виды).

[81] exon shuffling

ЛИТЕРАТУРА

1. M. Nei, S. Kumar, Molecular Evolution and Phylogenetics, Oxford University Press, 2000
2. R. Lewin, Patterns in Evolution, Scientific American Library, 1999
3. С. Оно, Генетические механизмы прогрессивной эволюции, Мир, 1973
4. М. Кимура, Молекулярная эволюция: тория нейтральности, Мир, 1985