На правах рукописи

Науменко Сергей Анатольевич

Динамика однолокусного мультиаллельного адаптивного ландшафта в молекулярной эволюции белок-кодирующих последовательностей ДНК

03.01.09 – математическая биология, биоинформатика

Автореферат
диссертации на соискание учёной степени
кандидата биологических наук

Москва – 2014

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН).

Научный руководитель: Базыкин Георгий Александрович,
заведующий сектором молекулярной эволюции
ИППИ РАН, кандидат биологических наук

Официальные оппоненты: Животовский Лев Анатольевич,
заведующий лабораторией генетических проблем
идентификации Института общей генетики
им. Н.И. Вавилова Российской академии наук,
доктор биологических наук,
кандидат физико-математических наук;

Гунбин Константин Владимирович,
старший научный сотрудник лаборатории
эволюционной биоинформатики и теоретической
генетики Института цитологии и генетики
Сибирского отделения Российской академии наук,
кандидат биологических наук

Ведущая организация: Федеральное государственное бюджетное учреждение
науки Институт прикладной математики
им. М.В. Келдыша Российской академии наук

Защита диссертации состоится «16» октября 2014 г. в 16 часов 00 минут на заседании диссертационного совета Д 002.077.04 при Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации им. А.А. Харкевича РАН, расположенном по адресу: 127994, г. Москва, ГСП-4, Большой Каретный пер., д. 19, стр. 1.

С текстом автореферата и диссертации можно ознакомиться в библиотеке ИППИ РАН, а также на сайте ИППИ РАН по адресу http://www.iitp.ru/ru/dissertation/1163.htm.

Автореферат разослан «____» _____ 2014 г.
Отзывы и замечания по автореферату в двух экземплярах, заверенные печатью, просьба высылать по вышеуказанному адресу на имя учёного секретаря диссертационного совета.

Учёный секретарь
диссертационного совета Д 002.077.04
доктор биологических наук, профессор Рожкова Г.И.

Общая характеристика работы

Объект исследования

Последовательность белка определяется последовательностью белок-кодирующего гена в соответствии с генетическим кодом. Несинонимические замены в белок-кодирующем гене в ходе эволюции приводят к изменению последовательности белка, которая определяет то, как он выполняет свою функцию в клетке, что, в свою очередь, влияет на приспособленность организма. Таким образом, определенный аллель в определенном сайте белок-кодирующего гена дает некоторый вклад в приспособленность организма. Назовем этот вклад приспособленностью аллеля.

Теоретически в аминокислотном сайте допустимы 20 аллелей, соответствующие 20 аминокислотам. В природе в каждом сайте некоторые аминокислоты, как правило, недопустимы, так что каждый сайт обычно ограничен их подмножеством. Кроме того, можно выделить классы аминокислот, сходных по физико-химическим свойствам; если разные аминокислоты в пределах класса являются функционально эквивалентными, то в качестве аллеля может выступать класс аминокислот.

Аллели различаются по абсолютной приспособленности w_a; эта величина может принимать значение от 0 (летальный аллель, наличие которого приводит к гибели организма) до некоторой положительной величины, характеризующей работу белка (скажем, скорость реакции, которую он катализирует).

На практике удобно определить абсолютную приспособленность аллеля в контексте популяции: приспособленность аллеля – это средняя приспособленность популяции, в которой этот аллель зафиксирован, а частоты аллелей во всех остальных локусах сохранены. Очевидно, что приспособленность одного и того же аллеля в разных популяциях может различаться.

Относительная приспособленность аллеля w определяется по отношению к аллелю, обладающему максимальной приспособленностью:

$$w = \frac{w_a}{w_{max}} . \qquad (1)$$

Эта величина изменяется в промежутке [0,1].

Относительная приспособленность аллеля зависит от генетического состава популяции: один и тот же аллель может иметь разную приспособленность на разном генетическом фоне.

В каждый момент времени аллели x_1, x_2, ... ,x_n, допустимые в данном сайте, обладают приспособленностями $w(x_1)$, $w(x_2)$, ... ,$w(x_n)$. В популяции могут расщепляться несколько аллелей (внутрипопуляционный полиморфизм), однако при рассмотрении молекулярной эволюции на уровне видов мы считаем, что в каждый момент времени данный вид характеризуется только одним аллелем в данном сайте.

Приспособленность аллеля, как присутствующего в сайте, так и отсутствующего, может меняться со временем, $w = w(t)$.

Адаптивный ландшафт — это функция, сопоставляющая генотипу значение приспособленности. Генотип определяется набором аллелей в соответствующих локусах.

Адаптивный ландшафт для двух локусов *L* и *M* с некоторым набором аллелей $\{l_1,...l_n\}$, $\{m_1,...,m_k\}$ в каждом из них можно изобразить в виде трехмерного графика (Рисунок 1), на котором для каждого из сочетаний аллелей l_i и m_j определена величина $z(l_i,m_j)$ — приспособленность, которую дает данное сочетание признаков. Области высокой приспособленности называются адаптивными пиками, области низкой приспособленности — адаптивными долинами.

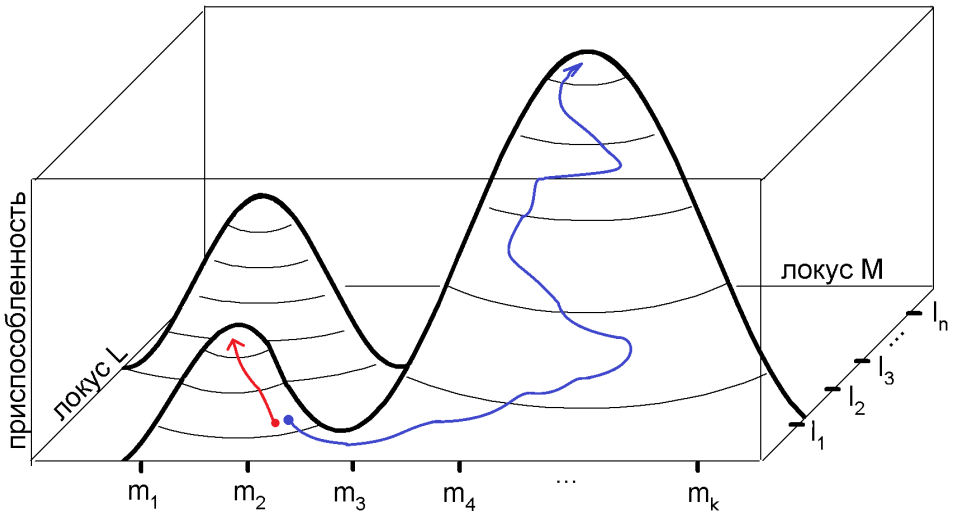


Рисунок 1. Адаптивный ландшафт в пространстве двух локусов. Изображение ландшафта в виде непрерывной поверхности условно; ландшафт является дискретным

Если рассматривать эволюцию на масштабе генов или на масштабе полных геномов, размерность адаптивного ландшафта возрастает; он также может становиться изломанным — с множеством локальных пиков.

Генотипы особей одного вида в популяции различаются; так, для человека среднее попарное нуклеотидное различие между двумя генотипами составляет 0,1%. Таким образом, на адаптивном ландшафте популяция особей будет отображаться как облако близко расположенных точек.

Эволюционный процесс можно представить себе как движение популяций по адаптивному ландшафту под действием основных факторов эволюции — отбора, мутаций, генетического дрейфа (Рисунок 1).

В данной работе исследуется адаптивный ландшафт аминокислотного сайта белок-кодирующего гена; соответственно, адаптивным ландшафтом называется вектор-функция $\vec{f}$, для каждого момента времени *t* задающая вектор значений приспособленности $\vec{W}=\left(w\left(x_1\right),w\left(x_2\right),\ldots,w\left(x_n\right)\right)$ для всех аллелей $\{x_1,..x_n\}$ в данном сайте:

$$\vec{W}=\vec{f}\left(t\right). \qquad (2)$$

Генетический код формально можно представить как отображение имеющего определенную внутреннюю структуру множества 64 нуклеотидных триплетов (кодонов) на множество аминокислот. Триплеты делятся на смысловые, кодирующие аминокислоты, и стоп-кодоны (TAA, TAG, TGA). Назовем разметкой генетического

кода способ деления кодонов на смысловые и стоп-кодоны. Генетические коды с различной разметкой могут обладать различной приспособленностью.

В данной работе исследуется адаптивный ландшафт на пространстве разметок генетических кодов.

Предмет исследования

Хотя нуклеотидную или аминокислотную последовательности можно рассматривать как простую линейную структуру, в клетке белок свернут в сложную структуру, в которой удаленные по линейному порядку аминокислоты оказываются рядом, и соотношение их свойств определяет выполнение белком своей функции. Кроме того, аминокислоты в различных положениях могут взаимодействовать посредством влияния на стабильность белка. Таким образом, замена одной аминокислоты влияет на приспособленность других аминокислот. Такие взаимодействия называются эпистатическими, в отличие от случая, когда аминокислоты в сайтах влияют на приспособленность белка независимо.

Адаптивный ландшафт аминокислотного сайта не является статическим: даже если в сайте не происходят замены, приспособленность аллелей в нем может меняться из-за замен в других сайтах за счет эпистатических взаимодействий.

В диссертации изучается динамика адаптивного ландшафта после замены одной аминокислоты (A) на другую (B). Рассматриваются отличия между видами, т.е. прямая замена происходит на макроэволюционных временах. Будем обозначать такую замену A ─ >B. Изучается вопрос о том, что происходит с приспособленностью аллеля A после такой замены с течением времени (Рисунок 2).

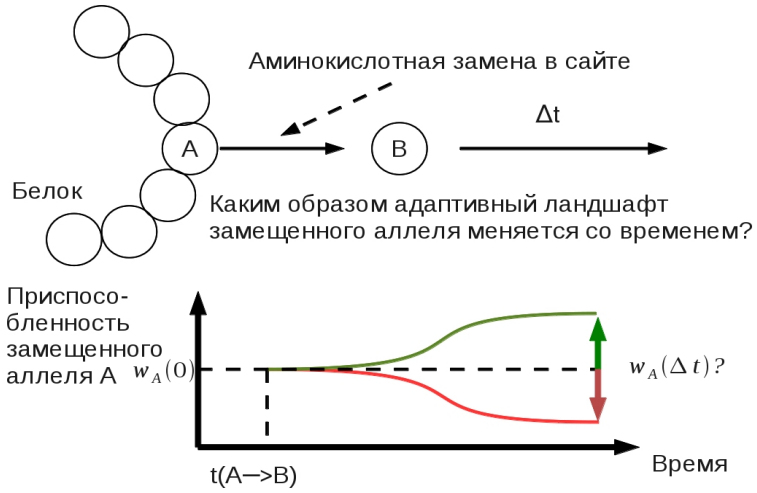


Рисунок 2. Локальный адаптивный ландшафт

Расположение стоп-кодонов в генетическом коде влияет на частоту точечных нонсенс-мутаций (мутаций, переводящих смысловой кодон в стоп-кодон) и нонсенс-мутаций при сдвиге рамки считывания. Таким образом, альтернативные генетические коды могут различаться по устойчивости к нонсенс-мутациям. Изучается устойчивость разметки стандартного генетического кода по сравнению с альтернативными разметками.

Методы исследования

Непосредственное изучение адаптивного ландшафта возможно для отдельно взятых белков, когда путем замены отдельных аминокислот о приспособленности судят, например, по результатам моделирования пространственной структуры, или для генно-модифицированных модельных организмов (о приспособленности судят по скорости размножения организмов, несущих различные аллели).

Изучение адаптивных ландшафтов возможно также по данным популяционной генетики: аллели, которые наблюдаются чаще в популяции (имеют большую частоту), обладают большей приспособленностью, а наблюдаемые реже — меньшей.

В данной работе изучение динамики адаптивного ландшафта на больших эволюционных временах, для клады позвоночных и насекомых, проводится с использованием данных о внутрипопуляционной изменчивости, а также при помощи сравнительной геномики: рассчитывая частоты замен в молекулярной эволюции можно делать выводы о динамике ландшафта.

Актуальность темы

Адаптивный ландшафт является одной из центральных концепций современной эволюционной биологии. Процесс эволюции рассматривается как движение популяций по адаптивному ландшафту в результате изменения генотипов. Математический аппарат эволюционной биологии — популяционная генетика — складывался как изучение равновесных состояний на фоне факторов эволюции (мутаций, отбора, дрейфа, миграции), действующих с постоянной силой. Однако, как становится ясно, все эти факторы меняются во времени, что вызывает необходимость обобщения классических представлений на подвижные адаптивные ландшафты.

В диссертации затрагивается тема динамики адаптивного ландшафта для одного из простейших случаев. Выбор масштаба объекта исследования позволяет полностью проанализировать динамику адаптивного ландшафта аналитически. С другой стороны, накопление молекулярных данных в масштабах полных геномов позволяет провести анализ динамики адаптивного ландшафта при помощи методов биоинформатики. Таким образом, удается сравнить предсказания теоретических моделей с эволюцией молекулярных последовательностей.

Эволюция ранней жизни на Земле, в частности, вопрос о возникновении аппарата трансляции, остается сравнительно малоизученной темой. С открытием альтернативных генетических кодов актуализировался вопрос о возникновении и эволюции генетического кода. В диссертации изучается одна из первичных адаптаций аппарата трансляции — выделение стоп-кодонов из множества смысловых.

Цели и задачи исследования

Целью работы является изучение динамики однолокусного адаптивного ландшафта в эволюции белок-кодирующих генов позвоночных и насекомых. В ходе работы были решены следующие задачи:

- подготовлены множественные выравнивания белок-кодирующих генов для филогений позвоночных и насекомых, включающие информацию о полиморфизме у *Drosophila melanogaster* и *Homo sapiens*;
- восстановлены предковые состояния во внутренних узлах филогений;

- разработано программное обеспечение, позволяющее отбирать сайты из полногеномных выравниваний, соответствующие заданным шаблонам;
- рассчитаны частоты обратных замен в эволюции позвоночных и насекомых для различных эволюционных расстояний;
- проведен теоретический анализ динамики однолокусного адаптивного ландшафта с тремя аллелями;
- разработан ABC-тест, позволяющий оценить относительную силу эффекта роста приспособленности производного аллеля и эффекта уменьшения приспособленности предкового аллеля;
- проведена оценка эффекта гетерогенности сайтов.

В ходе изучения первичных адаптаций аппарата трансляции было определено положение стандартного генетического кода на ландшафте всех возможных кодов для двух критериев приспособленности:

- вероятность прерывания трансляции мРНК со сдвигом рамки;
- вероятность точечной нонсенс-мутации.

Научная новизна и практическая значимость

В диссертации был проведен теоретический и сравнительно-геномный филогенетический анализ динамики ландшафта приспособленности сайта белок-кодирующего гена.

Показано, что в молекулярной эволюции происходит эффект “забывания”, который состоит в том, что вклад в приспособленность ранее присутствовавших в сайте аллелей падает со временем до фонового значения. По-видимому, это подтверждает гипотезу о наличии множества эпистатических связей между сайтами белок-кодирующего гена.

Результаты работы важны для понимания общих закономерностей эволюционного процесса и могут быть использованы при изучении адаптивных ландшафтов в молекулярной эволюции, в том числе для экспериментального изучения адаптивных ландшафтов в эволюции вирусов и бактерий.

Апробация работы

По результатам работы были сделаны доклады на следующих конференциях: 13-й конференции “Математика. Компьютер. Образование.” (г. Дубна, 23-28 января 2006 г.), 32-й и 33-й конференциях “Информационные технологии и системы” (ИТиС'09: пос. Бекасово, 15-18 декабря 2009 г., ИТиС'10: г. Геленджик, 20-24 сентября 2010 г.), ежегодных конференциях сообщества по молекулярной биологии и эволюции (SMBE'10: г.Лион, 4-8 июля 2010 г., SMBE'12: г.Дублин, 23-26 июня,), 2-й и 3-й конференциях “Математическая биология и биоинформатика” (г. Пущино, 7-13 сентября 2008 г., 10-15 октября 2010 г.), 3-й и 5-й конференциях молодых ученых “Биология: от молекулы до биосферы” (г. Харьков, 18-21 ноября 2008 г., 22-25 ноября 2010 г.), Московской конференции по вычислительной молекулярной биологии (MCCMB'11, г. Москва, 21-24 июля 2011 г.), RECOMB Satellite Conference on Bioinformatics Education (г. Санкт-Петербург, 26 августа 2012 г.).

Структура и объем диссертации

Работа состоит из введения, пяти глав, заключения и списка литературы. Глава 1 содержит обзор литературы по теме диссертации. В главе 2 дается теоретический анализ динамики адаптивного ландшафта. В главах 3 и 4 изложены результаты анализа динамики адаптивного ландшафта и анализа взаимосвязи ширины спектра допустимых аминокислот и скорости молекулярной эволюции для белок-кодирующих генов позвоночных и насекомых. В главе 5 содержится анализ адаптивного ландшафта разметок генетических кодов.

Работа изложена на 97 страницах, содержит 7 таблиц и 21 рисунок. Список литературы содержит 80 наименований.

Содержание работы

Экспериментальное изучение динамики адаптивного ландшафта затруднено. Однако его изменения могут приводить к паттернам во внутривидовом полиморфизме или межвидовой дивергенции, которые можно наблюдать методами сравнительной геномики, и такие данные позволяют сопоставлять различные модели динамики адаптивного ландшафта.

Прямые, обратные и боковые замены в триаллельном сайте

Пусть в некотором сайте белок-кодирующего гена в эволюционной истории произошла замена A —> B. Спустя время t в этом же сайте может произойти обратная замена (реверсия) B —> A, боковая замена B —> C, $C \neq A,B$, либо сайт может остаться инвариантным.

Анализ динамики адаптивного ландшафта в эволюции позвоночных и насекомых

Частоты событий замен B —> A, B —> C, а также частоты соответствующих полиморфизмов в популяциях *Homo sapiens* и *Drosophila melanogaster*, произошедших через некоторое время после замены A —> B, были измерены для клад позвоночных и насекомых (Рисунок 3). Каждый вид был представлен своим полным геномом в выравнивании по геному *Homo sapiens* (*Drosophila melanogaster*). Множественные выравнивания геномов были получены из базы данных UCSC Genome Browser. В соответствии с аннотацией генома *Homo sapiens* и генома *Drosophila melanogaster*, из множественных выравниваний были выделены участки, которые являются белок-кодирующими генами. Таким образом, были получены ортологичные последовательности для геномов позвоночных и насекомых (Таблица 1). К этим данным были добавлены данные о полиморфизме в популяции *Homo sapiens* (1094 особи из проекта “1000 геномов”) и популяции *Drosophila melanogaster* (162 особи из проекта “Drosophila Genetic Reference Panel”). Филогении указанных видов были загружены с сайта UCSC Genome Browser. По геномным последовательностям и филогении в каждом сайте при помощи программы codeml пакета PAML были восстановлены состояния во внутренних узлах филогенетических деревьев.

По горизонтальной оси на Рисунке 3 отложен эволюционный возраст аллеля B, который возник в результате замены A─>B, в предположении, что замены происходят в середине сегмента. Прямоугольниками отмечены сегменты, в которых произошла прямая замена A─>B. Черными прямоугольниками под прямой отмечены сегменты, для которых проводился анализ полиморфизма в популяции *Homo sapiens* или *Drosophila melanogaster*, белыми прямоугольниками отмечены сегменты, для которых проводился анализ дивергенции между линиями *Homo sapiens* и общего предка *Homo sapiens* и *Mus musculus* или *Drosophila melanogaster* и общего предка *Drosophila melanogaster* и *Drosophila sechellia*. В случае полиморфизма были проанализированы 6(5) эволюционных расстояний для дерева позвоночных (насекомых), а в случае дивергенции — 5(4). В сайтах, где произошла прямая замена A─>B, измерялся уровень полиморфизма в популяции *Homo sapiens* (*Drosophila melanogaster*), а также частоты замен на линии *Homo sapiens* после её расхождения с линией *Mus musculus* (на линии *Drosophila melanogaster* после её расхождения с *Drosophila sechellia*).

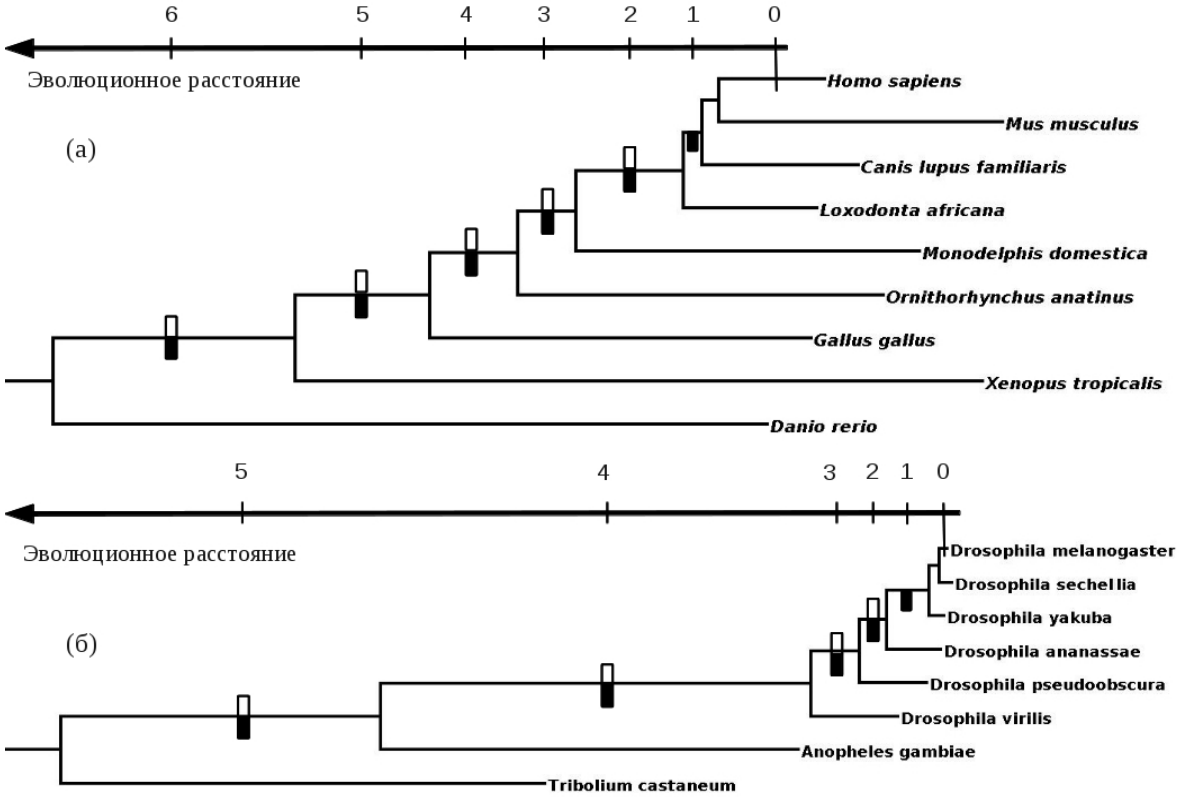


Рисунок 3. Филогенетические деревья видов, которые использовались в анализе: позвоночных (а) и насекомых (б)

Таблица 1. Количество проанализированных генов и кодонов

	Позвоночные	Насекомые
Видов	9	8
Генов	7 967	8 477
Кодонов	10 441 107	8 838 651

Измерение динамики уровня полиморфизма и частот обратных замен позволяет установить действие очищающего (отрицательного) отбора. Наряду с двумя независимыми наборами данных, две измеряемые статистики — уровень полиморфизма и частота обратных замен — увеличивают количество повторностей и значимость получаемой зависимости.

Данные по полиморфизмам (Рисунок 4а-в) свидетельствуют о том, что внутрипопуляционное предпочтение предкового аллеля A, полученного в результате обратной мутации, уменьшается с возрастом текущего фиксированного аллеля B. Эта же величина для боковых аллелей C либо независима от возраста B, либо уменьшается с этим возрастом (у *Drosophila*, Рисунок 4б). Однако предпочтение предкового аллеля A уменьшается сильнее, чем боковых аллелей C. В результате, отношение частот полиморфизмов B─>A к B─>C падает в ~3 раза за то время, пока возраст B увеличивается с очень молодого до такого, что ~1 нуклеотидная замена произошла в синонимичном сайте, с момента замены A─>B. Данные по аминокислотным заменам дают практически те же зависимости.

Наблюдаемые зависимости сохранялись качественно, когда рассматривались только те аминокислотные замены A─>B, которым соответствовали нуклеотидные замены слабого основания (AT) на сильное (CG), что исключает влияние эффекта смещенной генной конверсии.

Наблюдаемые зависимости также сохранялись, если замены A─>B были биохимически консервативными или, напротив, радикальными. Радикальность аминокислотной замены измерялась как расстояние между аминокислотами по шкале Мияты. В случае консервативных замен A─>B реверсии были более частыми, а боковые полиморфизмы или замены были менее частыми; однако во всех случаях отношение частот замен B─>A/B─>C уменьшалось со временем как для полиморфизмов, так и для межвидовых отличий.

Прямая замена A─>B может быть вредной (*w(A) >> w(B)*), околонейтральной (*w(A) ~ w(B)*) или полезной (*w(A) << w(B)*). Наблюдаемое уменьшение отношения частот обратных полиморфизмов и замен B─>A после прямой замены можно объяснить либо увеличением w(B), либо уменьшением w(A) со временем (Рисунок 5).

Чтобы разделить эти два эффекта, проанализируем, каким образом каждый из них влияет на динамику отношения частот замен и полиморфизмов B─>A / B─>C со временем.

Пусть аллели *A, B, C* имеют относительные приспособленности $w(A) = 1- s_1(t)$, $w(B) = 1$ и $w(C) = 1 - s_2(t)$, где *t* соответствует времени после исходной замены A─>B, а $s_1, s_2 > 0$ — коэффициенты отбора. Тогда частоты слабовредных замен $f(B{\rightarrow}A)$, $f(B{\rightarrow}C)$ относительно нейтральной замены определяются следующими уравнениями:

$$f(B \rightarrow A) = \frac{S_1}{e^{S_1} - 1}$$
$$f(B \rightarrow C) = \frac{S_2}{e^{S_2} - 1}, \qquad (3)$$

где $S_1 = 4N_e s_1$, $S_2 = 4N_e s_2$, и N_e – эффективная численность популяции.

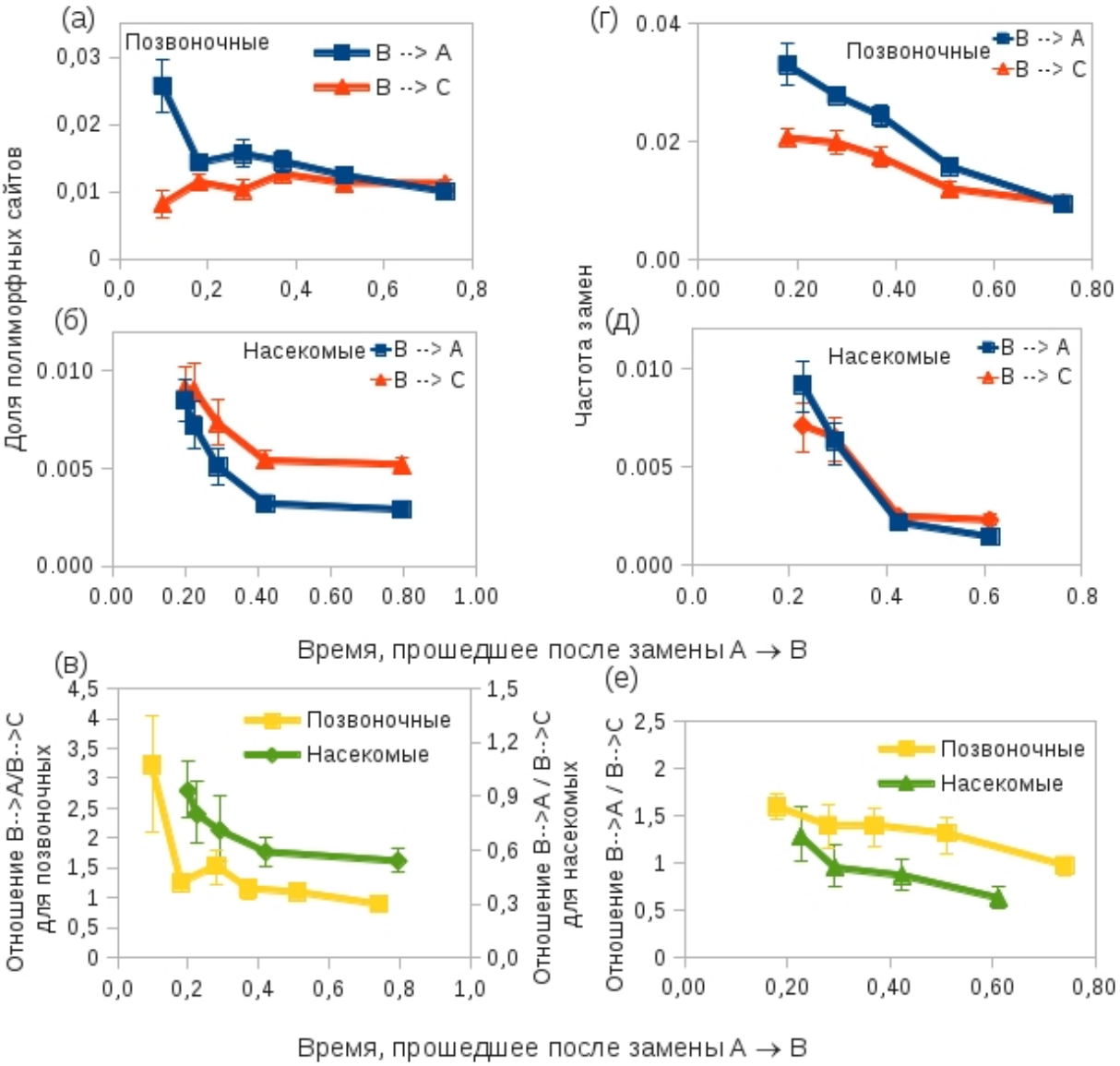


Рисунок 4. Зависимость уровня полиморфизма (а-в) и частот замен (г-е) от возраста предпочтительного в данный момент времени аллеля В, который зафиксировался ранее в результате замены А─>В. По оси абсцисс отложено время, прошедшее после замены А ─ >В, измеренное в количестве нуклеотидных замен на четырехкратно вырожденный синонимический сайт (а,б,г,д). Доля сайтов среди сайтов с заменой А ─ >В, которые несут аллель А (синяя линия) или один из двух аллелей С (красная линия) (а,б), или в которых произошла замена В─>А (синяя линия) или В─>С (красная линия) на терминальной ветви (г,д). (в,е) отношение количества полиморфизмов В─>А к В─>С (в), и замен (е): позвоночные (желтая линия), насекомые (зеленая линия). (а,г) – позвоночные, (б,д) – насекомые

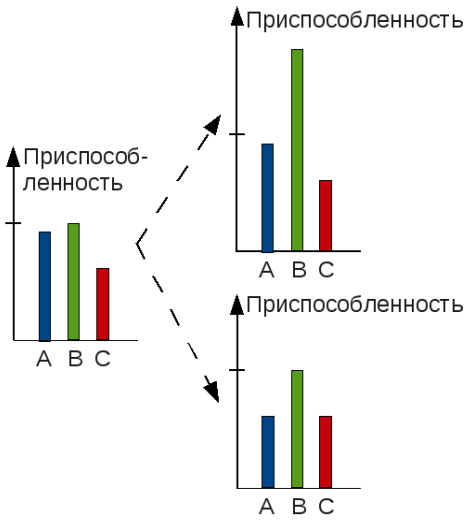


Рисунок 5. Два возможных типа динамики адаптивного ландшафта

Предположим, что после замены A─>B абсолютные значения приспособленности w(A) и w(C) остаются постоянными, а w(B) возрастает со временем (Рисунок 5). Это соответствует уменьшению относительной приспособленности w(A) и w(C), и, одновременно, увеличению коэффициентов отбора s_1 и s_2:

$$s_1 = s_1(0) + g(t);\ \ g(t) > 0;\ \ dg/dt > 0 \quad s_2 = s_2(0) + g(t) \quad s_1(0) < s_2(0) \Leftrightarrow w(A,0) > w(C,0) \tag{4}$$

Отношение *R(t)* частот замен B─>A и B─>C будет равно

$$R(t) = \frac{f(B \to A)}{f(B \to C)} = \frac{\dfrac{S_1}{e^{S_1} - 1}}{\dfrac{S_2}{e^{S_2} - 1}} = \frac{4N_e\left(s_1(0) + g(t)\right)}{4N_e\left(s_2(0) + g(t)\right)} \cdot \frac{e^{4N_e\left(s_2(0) + g(t)\right)} - 1}{e^{4N_e\left(s_1(0) + g(t)\right)} - 1}\ . \tag{5}$$

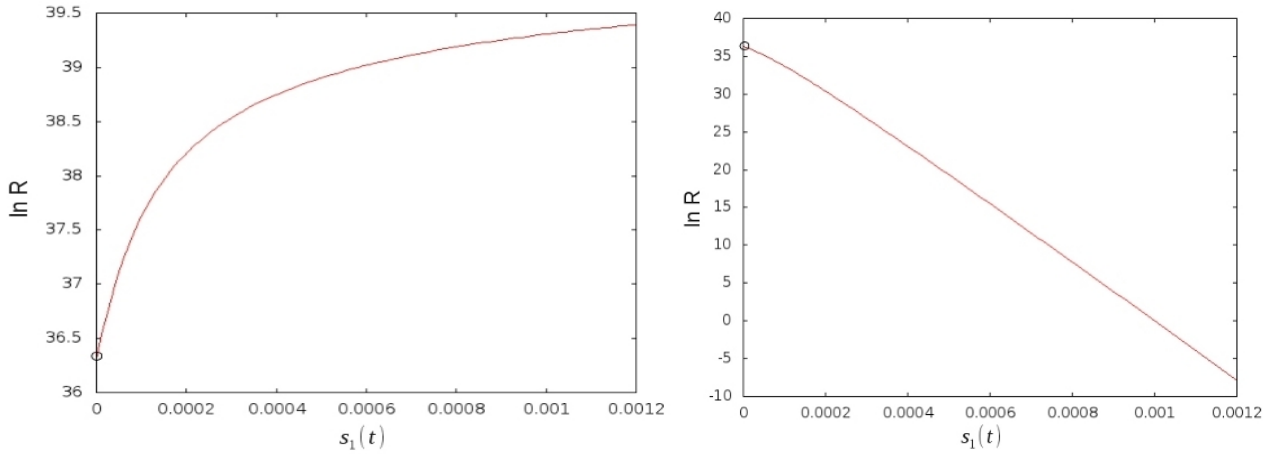
Численное решение и анализ этой функции показывают, что *R(t)* возрастает со временем (Рисунок 6а). Таким образом, в рассмотренном случае увеличения приспособленности аллеля B отношение частот (вероятностей) обратных замен B─>A и B ─>C после прямой замены (A─>B) увеличивается со временем.

Предположим, что после прямой замены A─>B w(A) уменьшается с течением времени, а w(B) и w(C) остаются неизменными (Рисунок 5). Это соответствует уменьшению относительной приспособленности w(A), т.е. коэффициент отбора s_1 увеличивается, а s_2 — остается постоянным:

$$s_1 = s_1(0) + g(t); s_1 > 0;\ \ g(t) > 0;\ \ dg/dt > 0; \qquad s_2 = s_2(0); s_2 > 0; \qquad s_1(0) < s_2(0) \Leftrightarrow w(A,0) > w(C,0)$$
$$S_1 = 4N_e s_1 \qquad S_2 = 4N_e s_2\ . \tag{6}$$

Тогда отношение *R(t)* частот замен B─>A и B─>C будет равно

$$R(t) = \frac{f(B \to A)}{f(B \to C)} = \frac{\dfrac{S_1}{e^{S_1} - 1}}{\dfrac{S_2}{e^{S_2} - 1}} = \frac{e^{4N_e s_2(0)} - 1}{4N_e s_2(0)} \cdot \frac{4N_e\left(s_1(0) + g(t)\right)}{e^{4N_e\left(s_1(0) + g(t)\right)} - 1} = C_4 \cdot \frac{4N_e\left(s_1(0) + g(t)\right)}{e^{4N_e\left(s_1(0) + g(t)\right)} - 1}, C_4 > 0\ . \tag{7}$$



(а) (б)

Рисунок 6. Отношение частот замен A─>B и A─>C как функция силы отбора против замены B─>A, в предположении, (а) что приспособленность аллеля B увеличивается, а приспособленности аллелей A и C остаются постоянными. N_e = 10 000; s_2 — s_1 = 0,001; (б) что приспособленность аллеля A уменьшается, а приспособленности аллелей B и C остаются постоянными. N_e = 10 000; $s_2(0)$ = 0,001

Тогда отношение *R(t)* частот замен B─>A и B─>C будет равно

$$R(t)=\frac{f(B\to A)}{f(B\to C)}=\frac{\frac{S_1}{e^{S_1}-1}}{\frac{S_2}{e^{S_2}-1}}=\frac{e^{4N_e s_2(0)}-1}{4N_e s_2(0)}\cdot\frac{4N_e\left(s_1(0)+g(t)\right)}{e^{4N_e\left(s_1(0)+g(t)\right)}-1}=C_4\cdot\frac{4N_e\left(s_1(0)+g(t)\right)}{e^{4N_e\left(s_1(0)+g(t)\right)}-1},C_4>0\ . \quad (8)$$

Численный расчет и анализ функции показывает, что *R(t)* уменьшается со временем (Рисунок 6б). Таким образом, в рассмотренном случае замены A─>B с последующим уменьшением приспособленности аллеля A отношение частот (вероятностей) замен B─>A и B─>C уменьшается монотонно со временем.

Уменьшение частоты полиморфизмов B─>C и замен в зависимости от эволюционного времени, прошедшего после фиксации аллеля B, указывает на то, что приспособленность текущего аллеля B увеличивается со временем. Однако если бы это увеличение было единственным изменением ландшафта приспособленности после замены A─>B, а приспособленность A и C оставалась неизменной, отношение частот замен B─>A к B─>C и полиморфизмов должно было бы увеличиваться, а не уменьшаться со временем, в предположении, что в момент замены предковой аминокислоте A соответствовало большее значение приспособленности, чем аминокислотам C, которые никогда ранее не наблюдались. Напротив, уменьшение приспособленности A приводит к наблюдаемой динамике, т.е. падению отношения частот B─>A/B─>C. Поэтому наблюдаемая динамика отношения B─>A/B─>C подтверждает, что, хотя оба эффекта – и увеличение приспособленности B, и уменьшение приспособленности A – происходят после замены A ─ >B, второй эффект сильнее, и именно он должен быть основной причиной уменьшения уровня ревертирующих полиморфизмов и реверсий с возрастом прямой замены (Рисунок 4).

Спустя время, соответствующее ~1 нуклеотидной замене на синонимичный сайт после замены A─>B, приспособленность замененной аминокислоты A достигает нового, стабильно низкого значения.

Изменение локального адаптивного ландшафта – медленный процесс, и, вероятно, он происходит в большой степени из-за замен в других сайтах белка или в других белках.

Поскольку в среднем отношение количества несинонимичных замен и замен в нефункциональных частях генома составляет ~0,1, белок “забывает” о замещенной аминокислоте после того, как ~20% аминокислот в геноме заменились. Конечно, эти второстепенные замены не влияют напрямую на приспособленность замененной аминокислоты A. Их отрицательное влияние на приспособленность A должно происходить из-за преобладания отрицательного эпистаза: замена в произвольном сайте генома скорее снижает, чем увеличивает, приспособленность случайно выбранного генотипа с большой приспособленностью.

Оценка эффекта гетерогенности сайтов

При расчете частот эволюционных событий по файлам множественных выравниваний все сайты перечислены в линейном порядке, информация о пространственной структуре белка утрачена, следовательно, все сайты равноправно учитываются при подсчете числа событий. В действительности сайты не одинаковы по своим свойствам: в зависимости от положения сайта в пространственной структуре белка в нем может быть предпочтительна та или иная аминокислота.

Пусть каждый сайт белок-кодирующего гена принадлежит одному из двух классов с разными, но статическими, адаптивными ландшафтами (Рисунок 7). Ландшафт сайтов класса 1 более плоский, и в нем замена A─>B близка к нейтральной, а в ландшафте сайтов класса 2 замена A─>B более радикальна.

В силу статичности адаптивного ландшафта обратная замена B─>A будет более вероятной (будет происходить с большей скоростью) в сайтах класса 1, чем в сайтах класса 2. Следовательно, в среднем в сайтах класса 1 эволюционное расстояние между прямой и обратной заменой будет меньше, чем в сайтах класса 2.

Таким образом, наблюдаемое падение частоты обратных замен (полиморфизмов) с увеличением расстояния между прямой и обратной заменой (полиморфизмом) может быть следствием подразделения сайтов на классы при статическом адаптивном ландшафте, т.е. следствием зависимости адаптивного ландшафта от линейной пространственной координаты в выравнивании белок-кодирующего гена, а не следствием изменения адаптивного ландшафта со временем.

Назовем описанную гипотезу гипотезой гетерогенности сайтов. В действительности должны наблюдаться оба эффекта, влияющие на наблюдаемые паттерны — и изменчивость сайтов по форме адаптивного ландшафта, и зависимость ландшафта от времени. Вопрос в том, имеет ли временной эффект значимую амплитуду на фоне эффекта разнообразия сайтов.

Для проверки гипотезы гетерогенности сайтов был использован ABC-тест. Исходим из того, что мы уже наблюдаем падение частоты замен B─>A с увеличением эволюционного расстояния между прямой заменой A─>B и обратной B─>A. Будем наблюдать за отношением количества обратных замен из аминокислоты B в

аминокислоту A к количеству боковых замен B─>C. Это отношение также убывает с увеличением эволюционного расстояния (Рисунок 4).

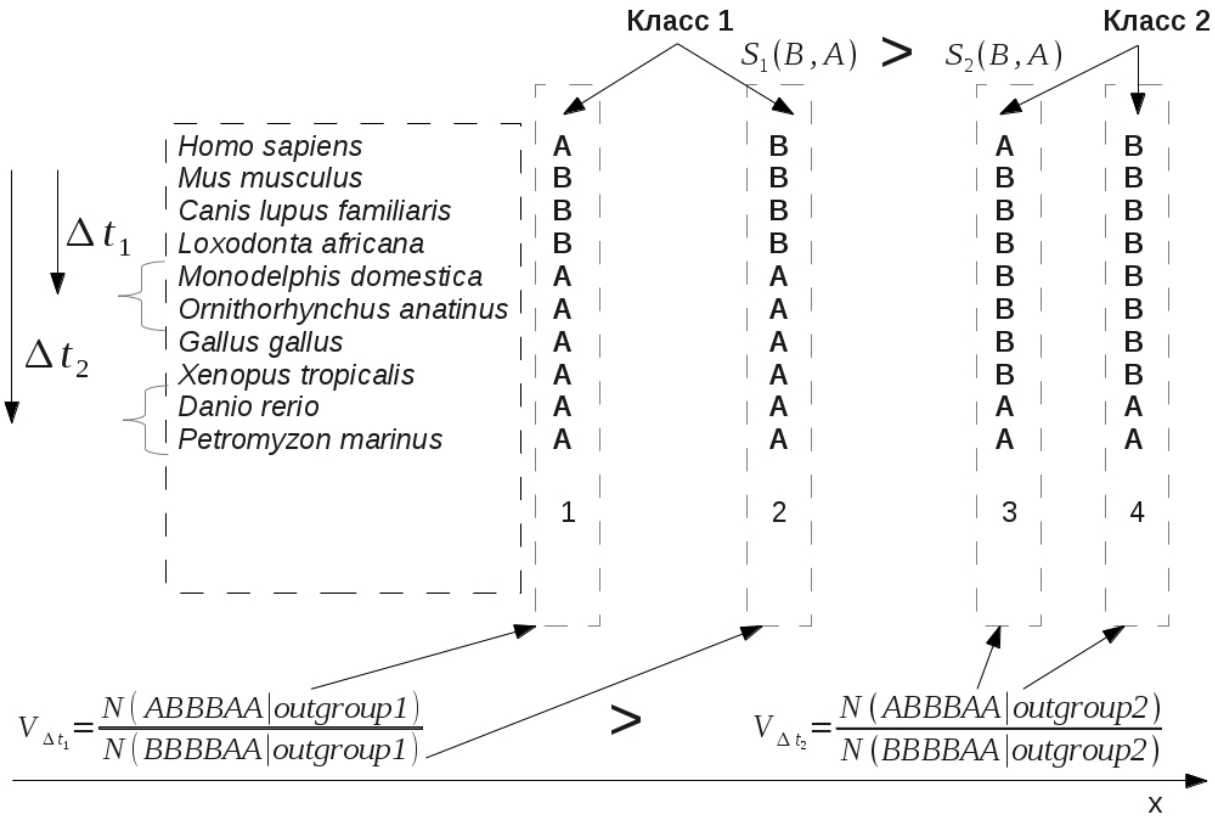


Рисунок 7. Классы сайтов в выравнивании белок-кодирующей последовательности

Пусть эффект гетерогенности сайтов преобладает; это значит, что большая частота замен B─>A и B─>C на меньших эволюционных расстояниях объясняется преимущественным попаданием сайтов в класс 1, в котором предпочтительнее обратные замены B─>A, а меньшая частота замен B─>A и B ─>C объясняется преимущественным попаданием сайтов в класс 2, в котором предпочтительнее боковые замены B─>C. В таком случае в сайтах класса 2 должно наблюдаться повышенное количество боковых замен B ─ >C по сравнению с обратными заменами B ─>A.

Количество боковых замен связано с количеством аминокислот, разрешенных в данном сайте — с шириной спектра аминокислот. Чем шире спектр возможных аминокислот, тем больше замен из аминокислоты B в аминокислоту, отличную от A и B (именно таково определение варианта C). Следовательно, в сайтах класса 1 должен быть более узкий спектр разрешенных аминокислот; поэтому там чаще происходят обратные замены A ─>B, по сравнению с сайтами класса 2.

В сайтах класса 1 мы наблюдаем большую частоту замен (скорость эволюции), чем в сайтах класса 2 (Рисунок 4в,е). Таким образом, в случае преобладания эффекта гетерогенности сайтов в сайтах с более узким спектром разрешенных аминокислот скорость эволюции должна быть выше, чем в сайтах с более широким спектром.

Однако ширина спектра должна быть положительно скоррелирована со скоростью молекулярной эволюции: чем больше разрешенных состояний в сайте, тем большее количество замен в нем должно происходить.

Протестируем это предположение в молекулярной эволюции плацентарных млекопитающих и насекомых.

Пусть на филогении из N видов (не считая внешнего вида (аутгрупп), по которому определяется предковое состояние) (Рисунок 8) в сайте произошло n независимых замен предковой аминокислоты на одну из k производных аминокислот. Замена каждого типа (замена на одну и ту же производную аминокислоту) встречается x_i раз, $\sum_{i=1}^{k} x_i = n$, или с частотой $p_i = \frac{x_i}{n}$.

Мы характеризуем скорость эволюции в сайте величиной n – количеством независимых замен предковой аминокислоты. В случае N видов n лежит в промежутке от 0 до N.

Ширину спектра допустимых аминокислот в сайте можно охарактеризовать величиной B, аналогичной несмещенной выборочной оценке ожидаемой гетерозиготности, которая широко используется в популяционной генетике:

$$B = \frac{n}{n-1}\left(1 - \sum_{i=1}^{k} p_i^2\right). \quad (9)$$

Эта формула дает несмещенную оценку количества допустимых аминокислот в сайте.

Величину B можно также интерпретировать как долю пар замен предковой аминокислоты, таких, что производные аминокислоты отличны друг от друга.

Для n замен существует $C_n^2 = \frac{n(n-1)}{2}$ пар (сочетаний по 2). Если среди n замен есть замены с одинаковым производным состоянием, то количество пар таких замен составит $\sum_{i=1}^{k} \frac{x_i(x_i-1)}{2}$ (количество сочетаний по 2 для каждого из множеств x_i) . Тогда долю пар замен с различными производными состояниями можно выразить как

$$\frac{\frac{n(n-1)}{2} - \sum_{i=1}^{k} \frac{x_i(x_i-1)}{2}}{\frac{n(n-1)}{2}} = \frac{\left(n^2 - n - \sum_{i=1}^{k}\left(x_i^2 - x_i\right)\right)}{n(n-1)} = \frac{n^2 - n - \sum_{i=1}^{k} x_i^2 + \sum_{i=1}^{k} x_i}{n(n-1)} = \frac{n^2 - \sum_{i=1}^{k} x_i^2}{n(n-1)}. \quad (10)$$

Легко показать, что выражение (10) равно B. Преобразуем B:

$$B = \frac{n}{n-1}\left(1 - \sum_{i=1}^{k} p_i^2\right) = \frac{n\left(1 - \sum_{i=1}^{k} \frac{x_i^2}{n^2}\right)}{n-1} = \frac{n^2 - \sum_{i=1}^{k} x_i^2}{n(n-1)}. \quad (11)$$

B может меняться в пределах от 0 (все наблюдаемые замены ведут к одной и той же аминокислоте, что означает, что лишь одна не-предковая аминокислота является допустимой) до 1 (все наблюдаемые замены являются заменами на разные аминокислоты, что означает максимальное разнообразие допустимых замен). Ясно, что B может быть определено только для сайтов с количеством замен не меньше двух.

Для анализа связи ширины спектра и скорости эволюции использовались два независимых набора данных: 13 полных геномов плацентарных млекопитающих с

опоссумом *Monodelphis domestica* в качестве внешнего вида и 11 геномов плодовых мушек из рода *Drosophila* с малярийным комаром *Anopheles gambiae* в качестве внешнего вида (Рисунок 8).

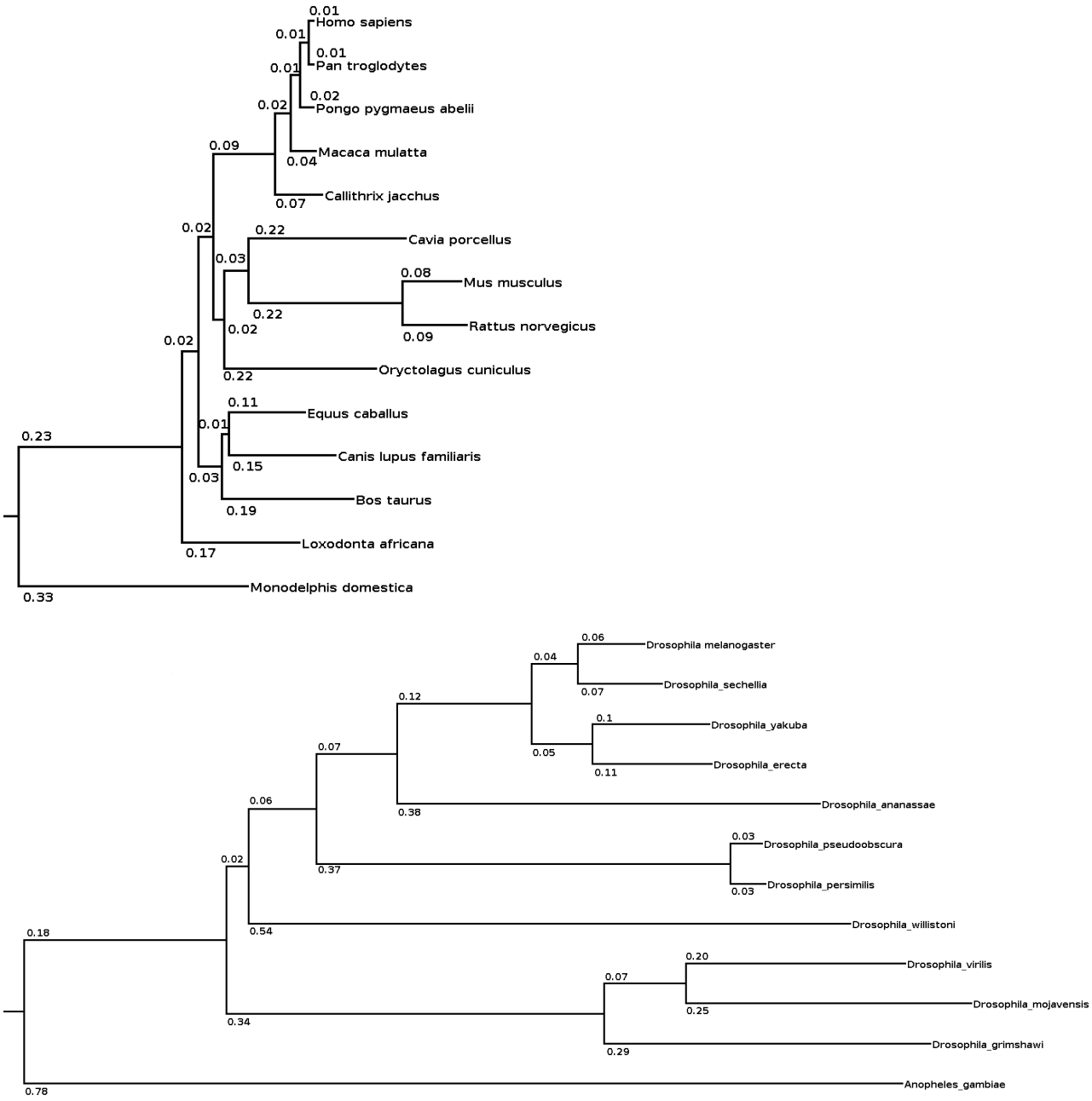


Рисунок 8. Филогенетические деревья плацентарных млекопитающих и дрозофил, использованные в анализе. Показаны длины ветвей, измеренные в единицах Ks (синонимических замен на синонимический сайт)

На Рисунке 9 показаны корреляции между *n* и *B* на уровне отдельных белков. Для каждого белка расчитано арифметическое среднее величин, характеризующих скорость эволюции (*n*) и ширину спектра (*B*) для всех сайтов, в которых произошло как минимум две замены предковой аминокислоты. Количество сайтов, в которых не произошло замен, либо произошла одна замена приведено в Таблице 2.

В Таблице 3 приведены коэффициенты корреляции Пирсона и Спирмена между скоростью эволюции и шириной спектра, расчитанные на основе полученных данных. Все коэффициенты значимо отличаются от 0 (p-value < $2{,}2 \cdot 10^{-16}$).

Таблица 2. Сайты с 0 и 1 заменой

	Количество сайтов (% от общего числа)	
	Инвариантных	С одной заменой
Drosophila	2 886 411 (78%)	324 330 (9%)
Плацентарные млекопитающие	5 086 661 (40%)	353 544 (3%)

Таблица 3. Коэффициенты корреляции между скоростью эволюции и шириной эволюционного спектра

Drosophila		Плацентарные млекопитающие	
R Пирсона	Rho Спирмена	R Пирсона	Rho Спирмена
0,29	0,26	0,34	0,39

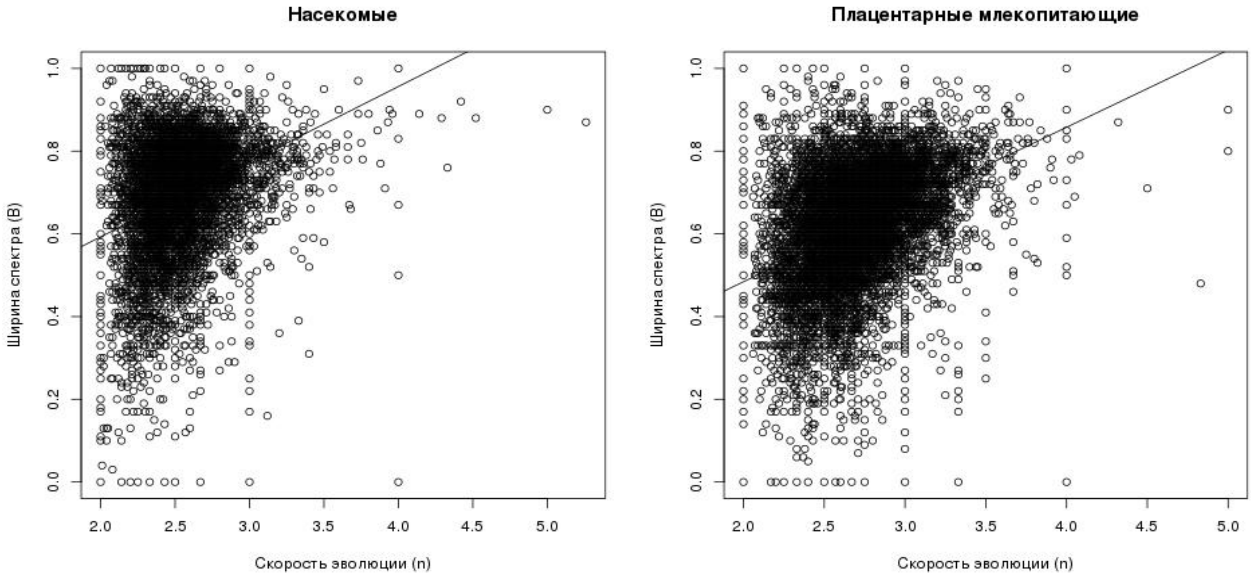


Рисунок 9. Скорость эволюции и ширина спектра, усредненные по белкам для *Drosophila* (слева) и плацентарных млекопитающих (справа)

Полученные данные показывают слабую положительную корреляцию между скоростью эволюции и шириной спектра как для млекопитающих, так и для насекомых. В сайтах, которые эволюционируют быстрее, разнообразие допустимых аминокислот также немного выше. Простейшее объяснение этой положительной корреляции в том, что большинство аминокислотных замен происходит в селективно нейтральных, или практически нейтральных, сайтах. Эти сайты эволюционируют со скоростью выше средней вследствие отсутствия селективного ограничения, и в них допустимы многие аминокислоты, что приводит к положительной корреляции между скоростью и шириной спектра.

Размещение стоп-кодонов уменьшает количество нонсенс-мутаций

Всего существует $C_{64}^{3}=41\ 664$ вариантов разметок с тремя стоп-кодонами.

Признаки начала и конца нуклеотидного триплета в мРНК отсутствуют. Трансляция происходит последовательно, триплет за триплетом. С одной и той же мРНК можно считать три различные белковые последовательности: в рамках 0, +1, -1. В подавляющем большинстве случаев функциональный белок синтезируется только в одной рамке считывания. Трансляция со сдвигом рамки часто приводит к преждевременной терминации белковой цепи.

Пару смысловых кодонов, которая при трансляции со сдвигом дает терминирующий кодон, будем называть запретной. Такова, например, пара А**TA–G**CA, поскольку она даёт стоп-кодон TAG при чтении в рамке +1.

Смысловые кодоны, соседствующие с терминирующими и вследствие этого подверженные точечным нонсенс-мутациям, будем называть уязвимыми кодонами.

Разметка генетического кода влияет на вероятность возникновения стоп-кодонов при чтении со сдвигом рамки (при условии, что триплетный состав последовательности однороден) и на количество кодонов, подверженных точечной нонсенс-мутации, а также на общее количество таких мутаций.

Критерий 1. Сдвиг рамки считывания – это нелокальная ошибка, полностью меняющая аминокислотную последовательность белка. По всей видимости, преимущество будет иметь такой аппарат трансляции, который возможно раньше прерывает чтение за счет появления стоп-кодона в новой рамке. Следовательно, оптимальной является разметка, максимизирующая вероятность появления стоп-кодона, или, иными словами, разметка с максимальным количеством запретных пар кодонов.

Стандартная разметка удовлетворяет этому условию, т.е. является оптимальной, давая 192 запретные пары кодонов в случае обоих сдвигов рамки. Полный перебор вариантов показывает, что всего существует 2 432 (5,8%) таких разметок. Распределение разметок по количеству запретных пар кодонов для сдвига -1 показано на Рисунке 10. Распределение для сдвига +1 выглядит точно так же.

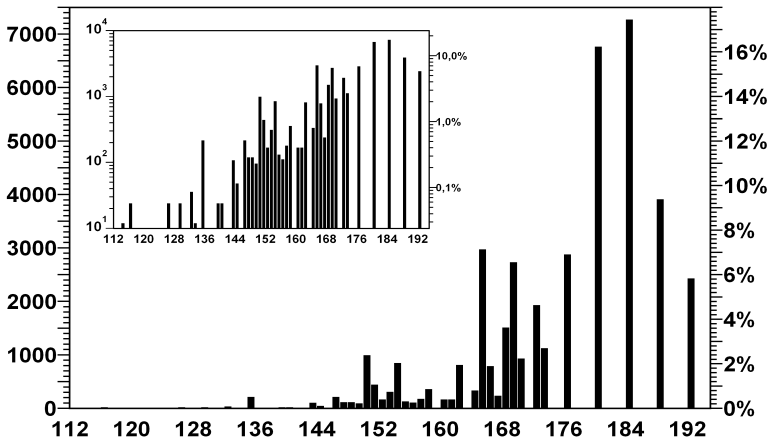


Рисунок 10. Распределение разметок по числу запретных пар кодонов. По оси абсцисс отложено количество запретных пар кодонов, по осям ординат — количество разметок с данным значением количества запретных пар и их доля. На врезке – те же данные с логарифмическим представлением по оси ординат

Критерий 2. Нонсенс-мутация приводит к преждевременному прекращению трансляции. Частично сформированная последовательность белка может привести к ошибкам образования вторичной, третичной структур белка и потере функциональности. Согласно этому критерию оптимальна разметка, минимизирующая вероятность нонсенс-мутаций. Критерий 2 распадается на 2 случая. Возможна минимизация суммарного (по всему коду) числа нонсенс-мутаций, либо минимизация количества уязвимых кодонов.

С точки зрения устойчивости к точечным мутациям разметки можно разбить на 4 группы, определяемые взаимным соседством стоп-кодонов:

В группе 1 каждый стоп-кодон является соседним по отношению к двум другим. Такие разметки имеют по 19 уязвимых кодонов и 21 нонсенс-мутацию. Последнее значение является минимально возможным.

В группе 2 один стоп-кодон соседствует с двумя другими, но те, в свою очередь, друг другу соседями не являются. Именно такова стандартная разметка. На смысловую часть кода приходится 23 нонсенс-мутации. Разметки имеют по 18 уязвимых кодонов, что является минимально возможным значением. Всего группа содержит 1 728 разметок, или 4,15% от общего числа.

В группе 3 два стоп-кодона являются соседними, а третий не соседствует ни с одним из них. Это соответствует 25 нонсенс-мутациям и количеству уязвимых кодонов от 20 до 23.

В группе 4 соседних кодонов среди стоп-кодонов нет. 27 нонсенс-мутаций приходятся на смысловую часть кода. Число уязвимых кодонов — от 21 до 27.

Стандартная разметка генетического кода обеспечивает максимальную вероятность появления стоп-кодона при чтении со сдвигом рамки на однородной по кодонному составу последовательности, а также минимальное количество кодонов, подверженных точечным нонсенс-мутациям. Всего существует 528 (1,3%) разметок с аналогичными свойствами.

Оптимальность стандартного генетического кода предполагает эволюционный сценарий его происхождения. Вероятно, в критерии приспособленности ранних организмов входила устойчивость аппарата трансляции к ошибкам. Фиксация стоп-кодонов, минимизирующая ошибки, связанные со сдвигом рамки считывания, могла повысить устойчивость раннего аппарата трансляции.

Стандартная разметка обеспечивает минимальное количество уязвимых для точечных нонсенс-мутаций кодонов, а не общее количество нонсенс-мутаций по коду. Вероятно, критерием приспособленности раннего аппарата трансляции могло быть количество кодонов, расположенных на расстоянии более одной мутации от стоп-кодонов. Достижение генетического кода, оптимального по этому критерию, могло способствовать дальнейшей специализации неуязвимых кодонов.

Основные результаты и выводы

1. В молекулярной эволюции белок-кодирующих последовательностей позвоночных и насекомых после аминокислотной замены происходит изменение адаптивного ландшафта аминокислотного сайта: приспособленность предкового аллеля уменьшается со временем до фонового значения.
2. Вследствие изменения адаптивного ландшафта со временем, прошедшим после аминокислотной замены, наблюдается уменьшение частот обратных полиморфизмов и замен. Эта динамика может быть также обусловлена изменением адаптивного ландшафта во времени, состоящем в увеличении приспособленности производного аллеля, или неравномерностью адаптивного ландшафта по соотношению приспособленностей аллелей в разных аминокислотных сайтах (эффект гетерогенности сайтов).
3. Увеличение приспособленности производного аллеля должно приводить к увеличению отношения частоты обратных замен к частоте замен на аминокислоты, отличные от предковой и производной (боковых замен) со временем, прошедшим после прямой замены.
4. Уменьшение приспособленности предкового аллеля должно приводить к уменьшению отношения частоты обратных замен к частоте боковых замен со временем, прошедшим после прямой замены.
5. В эволюции позвоночных и насекомых наблюдается уменьшение отношения частоты прямых замен к частоте боковых замен, из чего следует преобладание эффекта уменьшения приспособленности предкового аллеля по сравнению с эффектом увеличения приспособленности производного аллеля.
6. В эволюции позвоночных и насекомых наблюдается слабая положительная корреляция ширины спектра допустимых аминокислот и скорости молекулярной эволюции, из чего следует преобладание эффекта временной динамики адаптивного ландшафта над эффектом гетерогенности сайтов.
7. Стоп-кодоны в генетическом коде расположены таким образом, что обеспечивают устойчивость к мутациям сдвига рамки считывания и нонсенс-мутациям.

Список публикаций по теме диссертации

Статьи в научных журналах

1. Naumenko S.A., Kondrashov A.S., Bazykin G.A. Fitness conferred by replaced amino acids declines with time. // Biology Letters. - 2012. - V. 8. - N 5 - P.825-828.
2. Naumenko S.A., Kondrashov A.S. Rate and breadth of protein evolution are only weakly correlated. // Biology Direct. - 2012. - V. 7. - N 8. - P.1-12.
3. Малинецкий Г.Г., Науменко С.А., Подлазов А.В. Об экстремальных свойствах разметки генетического кода. //Доклады академии наук (биохимия, биофизика, молекулярная биология). - 2007. - Т. 414. - N 6. - С.831-835.

Тезисы конференций

1. Науменко С.А., Подлазов А.В. Об экстремальных свойствах разметки генетического кода. // “Математика. Компьютер. Образование.” Сборник трудов XIII международной конференции. Под общей редакцией Г.Ю. Ризниченко. - Ижевск: Научно-издательский центр “Регулярная и хаотическая динамика”. - 2006. - Т.2. - С.404.-413.
2. Науменко С.А. Роль нонсенс-кодонов в обеспечении оптимальности генетического кода. // Материалы докладов XV Международной конференции студентов, аспирантов и молодых ученых “Ломоносов” / Отв. ред. И.А. Алешковский, П.Н. Костылев. [Электронный ресурс] — М.: Издательство МГУ; СП МЫСЛЬ, 2008. — 1 электрон. опт. диск (CD-ROM); 12 см. - Систем. требования: ПК с процессором 486 +; Windows 95; дисковод CD-ROM; Adobe Acrobat Reader.
3. Науменко С.А., Кондрашов А.С., Базыкин Г.А. О высоких частотах реверсий в эволюции позвоночных и насекомых. // “Биология: от молекулы до биосферы”. Материалы IV Международной конференции молодых ученых (17-21 ноября 2009 г., г. Харьков, Украина). - Харьков: ЧПИ “Новое слово”. - 2009. - С.149- 150.
4. Naumenko S., Kondrashov A., Bazykin G.A. High Frequency of reversals in evolution of vertebrates and insects // Информационные технологии и системы (ИТиС'09): сборник трудов конференции. [Электронный ресурс] - М.: ИППИ РАН. - 2009. - 463с. - 1 электрон. опт. диск (CD-ROM) - С.367- 368.
5. Naumenko S., Kondrashov A., Bazykin G. High frequency of reversals in evolution of vertebrates and insects. // Annual meeting of the society for molecular biology and evolution “SMBE-2010”, Lyon, France, July 4-8.
6. Naumenko S., Kondrashov A., Bazykin G. Frequency of reversals in evolution of vertebrates and insects decreases on increased phylogenetic distance between substitutions // Информационные технологии и системы (ИТиС'10): сборник трудов конференции. - [Электронный ресурс] — М.: ИППИ РАН. - 2010. - С.361- 362.
7. Науменко С.А., Кондрашов А.С., Базыкин Г.А. Исследование эволюционных ландшафтов приспособленности аминокислотных сайтов при помощи реверсий // Математическая биология и биоинформатика: III Международная конференция., г. Пущино, 10-15 октября 2010 г.: Доклады / Под. ред. В.Д. Лахно. - С.105-106.

8. Науменко С.А., Базыкин Г.А. Вклад в приспособленность недавно замененной аминокислоты уменьшается со временем. // “Биология: от молекулы до биосферы”. Материалы V Международной конференции молодых ученых (22-25 ноября 2010 г., г.Харьков, Украина). - Х.: Оперативная полиграфия. - 2010. - С.9-10.
9. Naumenko S.A., Kondrashov A.S., Bazykin G.A. The fitness conferred by recently replaced amino acids rapidly declines with time. // Proceedings of the International Moscow Conference on Computational Molecular Biology (MCCMB'11). - 2011. - P.243.
10. Naumenko S.A., Kondrashov A.S. Rate and breadth of protein evolution are only weakly correlated. // Annual meeting of the society for molecular biology and evolution “SMBE-2012”, Dublin, Ireland, June 23-26.
11. Naumenko S.A. The number of reversing subsitutions. // The program of Forth RECOMB Satellite Conference on Bioinformatics Education (RECOMB-BE 2012). - 2012. - P.11.