ГЛАВА 3
МЕТОДЫ ИЗУЧЕНИЯ СИНОНИМИЧНЫХ И НЕСИНОНИМИЧНЫХ НУКЛЕОТИДНЫХ ЗАМЕЩЕНИЙ

В исследованиях эволюционных различий нуклеотидных последовательностей молекул РНК или ДНК часто требуется определить не только общее число и скорость нуклеотидных замен, но и отдельно установить долю и скорости синонимичных и несинонимичных замещений.

Синонимичной или молчащей заменой принято считать замену нуклеотида, не приводящую к замене аминокислоты. Несинонимичной заменой называют замену, приводящую к изменению кодируемой аминокислоты. Таблица генетического кода указывает, что все замены во втором положении нуклеотида в кодоне являются несинонимичными, в то время как часть замен нуклеотидов в первом и третьем положениях синонимичны. Согласно гипотезам о равных частотах нуклеотидов и случайных заменах, эта часть приблизительно составляет 5% для первого положения и 72% для третьего положения.

Так как синонимичные замены не подвержены действию естественного отбора на уровне белка, то скорость синонимичных замен соответствует скорости нейтральных нуклеотидных замен. Установлено, что скорость синонимичных замен приблизительно одинакова для многих генов, если не дифференцировать ее в соответствии с использованием кодонов и другими факторами. Скорость несинонимичных замен, напротив, обычно гораздо ниже таковой для синонимичных и сильно варьирует для разных генов. Однако важно отметить, что существуют гены, в процессе эволюции которых несинонимичные замены встречаются чаще, чем синонимичные. Эти несинонимичные замены, по-видимому, вызваны положительным дарвиновским отбором, потому что согласно представлениям о нейтральной селекции скорости синонимичных и несинонимичных замен должны быть равны. По этой причине определение скоростей синонимичных и несинонимичных замен стало важным предметом исследований в молекулярной эволюции.

Определение скоростей синонимичных и несинонимичных замен гораздо сложнее определения общей скорости нуклеотидных замен. В большинстве нуклеотидных последовательностей содержится больше нуклеотидных сайтов, в которых могут произойти несинонимичные мутации, чем сайтов, в которых возможны синонимичные мутации, при том в разных генах число синонимичных и несинонимичных сайтов различно. Скорость

синонимичных и несинонимичных замен необходимо определять как число синонимичных замен на синонимичный сайт (r_S) и как число несинонимичных замен на несинонимичный сайт (r_N) в год или иную единицу времени (t), соответственно. Однако время дивергенции двух сравниваемых последовательностей ДНК обычно не известно и возможно лишь рассматривать число синонимичных замен на синонимичный сайт ($d_S=2r_St$) и число несинонимичных замен на несинонимичный сайт ($d_N=2r_Nt$) для пары последовательностей.

Все существующие методы определения d_S и d_N можно разделить на три группы:

1. Методы, основанные на эволюционных путях.
2. Методы, основанные на двухпараметрической модели Кимуры.
3. Методы максимального сходства с моделями замещений кодонов.

Эти методы основаны на разных гипотезах и именно поэтому они не всегда дают одинаковые результаты.

3.1. Методы, основанные на эволюционных путях

Родоначальником этой группы методов является статистический метод Перлера, предположенный в 1980 году и предназначенный для определения количества синонимичных замен. В последующем Мията-Ясунага и Ли разработали более сложные методы. Основное различие метода Перлера от двух последующих методов заключается в том, что двум и более эволюционным путям, возможным между парой кодонов, придается равное значение. В то время как в последующих двух методах большее значение придается эволюционному пути, включающему синонимичные замены, чем пути, включающему несинонимичные замены. Однако всем этим методам присущ один технический недостаток – они столь сложны, что исследователи часто отказываются от их использования.

Метод Ней-Годжобори. При выполнении компьютерного моделирования Ней и Годжобори установили, что учет различных эволюционных путей не является необходимым и что равнозначная версия метода Мията-Ясунага дает приблизительно те же результаты, что и оригинальная версия.

Метод Ней и Годжобори основан на определении количества синонимичных и несинонимичных замен, а также числа потенциально синонимичных и несинонимичных сайтов. Рассмотрим методику вычисления количества потенциально синонимичных и несинонимичных сайтов. Эти величины рассчитываются для каждого кодона на основании гипотезы о

равных частотах всех нуклеотидных замен. Обозначим как f_i долю синонимичных замен (отношение количества синонимичных замен к общему числу синонимичных и несинонимичных замен, исключая нонсенс-мутации) в i-положении данного кодона (i = 1, 2, 3). Тогда количество потенциально синонимичных (s) и несинонимичных сайтов (n) для этого кодона вычисляются по формулам:

$$s = \sum_{i=1}^{3} f_i,$$
$$n = (3-s).$$

Рассмотрим методику расчета на примере кодона УУА, кодирующего лейцин. По первому положению, как и по двум другим возможны три замещения. Если в первом положении кодона произойдет замена У на А или Г, то это приведет к изменению кодируемой аминокислоты (Лей→Иле, Лей→Вал, соответственно). При замене У на Ц кодируемая аминокислота останется той же, то есть замена является синонимичной. Тогда доля возможных синонимичных замен по этому положению равна $f_1 = ⅓$. Все замены во втором положении указанного кодона приведут к изменению аминокислоты, следовательно, $f_2 = 0$. И наконец, в третьем положении замены А на У и Ц являются несинонимичными (Лей→Фен), а замена А на Г – синонимичной, что дает $f_3 = ⅓$. Таким образом, для этого кодона $s = ⅓ + 0 + ⅓ = ⅔$ и $n = 3 - s = 2⅓$.

Если нуклеотидная замена приводит к появлению терминального кодона, то она не учитывается. Например, замена У на А в третьем положении кодона УГУ, кодирующего цистеин, приводит к появлению терминального кодона. Замена У на Ц в этом положении кодона является синонимичной, а замена У на Г – несинонимичной (Цис→Три), что дает $f_3 = ½$. В первом и втором положениях этого кодона все замены являются несинонимичными ($f_1 = 0$, $f_2 = 0$). Тогда $s = 0 + 0 + ½ = 0,5$ и $n = 3 – 0,5 = 2,5$.

Вычисление общего количества синонимичных и несинонимичных сайтов производится по формулам:

$$S = \sum_{j=1}^{C} S_i,$$
$$N = (3C-S),$$

где s_i – это значение s для i-кодона, а C – общее количество кодонов нуклеотидной последовательности.

При сравнении двух последовательностей используются средние значения S и N для этих двух последовательностей. При этом N + S должно быть равно 3C (общему количеству сравниваемых нуклеотидов).

Перейдем к вычислению синонимичных и несинонимичных нуклеотидных различий между парой сравниваемых последовательностей. Для этого сравним соответствующие кодоны этих двух последовательностей и

вычислим значение синонимичных и несинонимичных нуклеотидных различий для каждой пары сравниваемых кодонов. Когда имеется только одно нуклеотидное различие, можно сразу определить, является ли замена синонимичной или несинонимичной. Например, при сравнении пары кодонов - ГУУ (Вал) и ГУА (Вал) наблюдается одно синонимичное различие. Обозначим количество синонимичных и несинонимичных различий в кодоне как s_d и n_d, соответственно. В данном случае $s_d = 1$ и $n_d = 0$.

Когда существуют 2 нуклеотидных различия между парой сравниваемых кодонов, то возможны два пути их возникновения. Например, при сравнении кодонов УУУ и ГУА эти два пути таковы:

1) УУУ (Фен) ↔ ГУУ (Вал) ↔ ГУА (Вал)

2) УУУ (Фен) ↔ УУА (Лей) ↔ ГУА (Вал)

Первый путь включает одну синонимичную и одну несинонимичную замену, в то время как второй путь включает две несинонимичные замены. Предположим, что оба пути имеют равную вероятность. Тогда $s_d = 0,5$ и $n_d = 1,5$. При сравнении некоторых пар кодонов эволюционные пути, объясняющие возникновение нуклеотидных различий, могут включать терминальные кодоны. При наличии таких путей их не следует учитывать.

Когда имеются три нуклеотидных различия между сравниваемыми кодонами, то возможны уже 6 различных путей возникновения различий и каждый из них состоит из трех мутационных шагов. Рассматривая все эти пути и мутационные шаги, необходимо определять s_d и n_d таким же образом, как в описанном выше случае двух нуклеотидных различий. Например, при сравнении кодонов УУГ и АГА эти шесть путей таковы:

1) УУГ (Лей) ↔ АУГ (Мет) ↔ АГГ (Арг) ↔ АГА (Арг)

2) УУГ (Лей) ↔ АУГ (Мет) ↔ АУА (Иле) ↔ АГА (Арг)

3) УУГ (Лей) ↔ УГГ (Три) ↔ АГГ (Арг) ↔ АГА (Арг)

4) УУГ (Лей) ↔ УГГ (Три) ↔ УГА (Тер) ↔ АГА (Арг)

5) УУГ (Лей) ↔ УУА (Лей) ↔ АУА (Иле) ↔АГА (Арг)

6) УУГ (Лей) ↔ УУА (Лей) ↔ УГА (Тер) ↔ АГА (Арг)

Четвертый и шестой пути включают терминальные кодоны, поэтому их необходимо исключить. Количество синонимичных замен в путях (1), (2), (3) и (5) составляют 1, 0, 1 и 1, соответственно. В то время как количество несинонимичных замен равно 2, 3, 2 и 2, соответственно. В соответствие с равной вероятностью всех эволюционных путей $s_d = ¾$ и $n_d = 2¼$.

Общее количество синонимичных и несинонимичных различий для сравниваемых последовательностей может быть получено путем суммирования этих значений для всех кодонов:

$$S_d = \sum_{j=1}^{C} s_{dj},$$
$$N_d = \sum_{j=1}^{C} n_{dj},$$

где s_{dj} и n_{dj} – это s_d и n_d для j-кодона, соответственно, а C – это количество сравниваемых кодонов. Сумма S_d и N_d равна общему количеству нуклеотидных различий между двумя сравниваемыми нуклеотидными последовательностями.

Долю синонимичных (p_S) и несинонимичных (p_N) различий, также часто называемую p-дистанцией) можно рассчитать по следующим уравнениям:

$$p_S = S_d / S,$$
$$p_N = N_d / N,$$

где S и N – среднее количество синонимичных и несинонимичных сайтов в двух сравниваемых последовательностях. Варианси этих показателей определяются формулами:

$$V(p_S) = p_S (1 - p_S) / S,$$
$$V(p_N) = p_N (1 – p_N) / N$$

Для определения числа синонимичных (синонимичная дистанция, d_S) и несинонимичных замен (несинонимичная дистанция, d_N) на сайт воспользуемся формулой Джукса-Кантора:

$$d = - ¾ \ln (1 - 4/3p),$$

подставив вместо p - p_S или p_N. Конечно, этот метод дает только приблизительные значения d_S и d_N, потому что формула Джукса-Кантора не применяется для некоторых синонимичных сайтов (двух и трехкратно вырожденных сайтов). Однако с помощью компьютерного моделирования установлено, что формула Джукса-Кантора позволяет получить корректные значения по синонимичным и несинонимичным заменам, когда частоты всех нуклеотидов, а также транзиций и трансверсий приблизительно одинаковы. Варианси могут быть рассчитаны по формулам:

$$V(d_S) = V(p_S) / (1 – 4/3\ p_S)^2$$
$$V(d_N) = V(p_N) / (1 – 4/3\ p_N)^2,$$

или с помощью метода бутстрэп, описанного ниже.

Метод Нея-Годжобори следует использовать для определения приблизительных значений S и N, а также вычисляемых на основе их показателей, а также когда расчетное значение соотношения транзиций и трансверсий составляет приблизительно 0,5. Еще одним показанием к использованию этого метода являются селекционные тесты.

Модифицированный метод Ней-Годжобори. Оригинальный метод, предложенный Ней и Годжобори, основывается на предположении о случайных равновероятных заменах нуклеотидов и исходя из этого рассчитывает число синонимичных и несинонимичных сайтов. В действительности замены нуклеотидов не являются равновероятными, так как частота транзиций обычно выше частоты трансверсий. Тогда число потенциально синонимичных сайтов (S) оказывается больше, чем таковое, вычисленное методом Ней-Годжобори. Это объясняется тем, что большинство транзиций по третьему положению приводят к синонимичным заменам. Это объясняет завышение значений N и соответственно занижение значений p_N и d_N, а также занижение значений S и соответственно завышение значений p_S и d_S при использовании метода Ней-Годжобори.

Для получения корректных значений оригинальный метод Ней-Годжобори был модифицирован с помощью метода Ина, основанного на двухпараметрической модели Кимуры. В этой модели частота транзиций обозначается α, а частота трансверсий – β. Так как по каждому нуклеотиду может произойти одна транзиция и две различных трансверсии, то пропорция транзиций к общему количеству замен выглядит так:

$$\alpha / (\alpha + 2\beta) = R / (1 + R),$$

где R – это расчетное соотношение транзиций и трансверсий. Если частоты транзиций и трансверсий равны, то R = 0,5. Расчетное соотношение транзиций и трансверсий (R) не следует путать с соотношением наблюдаемых транзиций и трансверсий $k = \alpha / \beta$.

Ина установил, что ожидаемое число синонимичных замен в кодоне может определятся формулой $R = \alpha / 2\beta$ для всех кодонов. Например, для кодона УУУ эта величина вычисляется так:

$$s = 0 + 0 + \alpha / (\alpha + 2\beta) = R / (1 + R),$$

потому что в этом случае синонимичные замещения возможны лишь по третьему положению кодона и только одна (У на Ц) из трех возможных замен синонимична. Для другого примера, кодона ЦУА (Лей) ожидаемое значение $s = R / (1 + R) + 1$, потому что синонимичные замены возможны по первому, второму и третьему положению нуклеотида в кодоне с вероятностью $R / (1 + R)$, 0 и 1, соответственно. В этих расчетах не следует учитывать нонсенс-мутации.

Для ядерных генов характерно значение R = 0,5 – 2, что делает метод применимым в большинстве случаев.

Если известно R, то можно рассчитать s для всех кодонов и затем вычислить S и N ($N = 3C - S$).

При использовании модифицированного метода Нея-Годжобори значения S будут выше, а значения N – ниже таковых, полученных оригинальным методом. Обозначим новые S и N как S_R и N_R, соответственно. В отличие от S и N значения S_d и N_d не зависят от величины R, так как они получаются путем подсчета наблюдаемых различий. Тогда доля синонимичных и несинонимичных различий выглядит так:

$$p_S=S_d/S_R,$$

$$p_N=N_d/N_R,$$

а приблизительные значения d_S и d_N определяются по формуле Джукса-Кантора. Вариансы d_S, d_N, p_S и p_N можно рассчитать по указанным выше формулам или с помощью бутстрэп-метода.

Теоретически доказано, что лучше определять d_S и d_N по методу Ина, но на практике значения, полученные этими двумя методами, примерно одинаковы, если d_S и d_N не очень велики. (Когда d_S >1,0 и d_N >1,0, то их значения не надежны, потому что процесс подсчета синонимичных и несинонимичных замен очень сложен.)

При высоких значениях R модифицированный метод, использующий модель Кимуры, теоретически лучше оригинального метода Ней-Годжобори. Однако следует отметить, что когда значения R не надежны, они могут приводить к ошибочным выводам. В частности, когда используются завышенные значения R, с помощью модифицированной версии можно ошибочно получить, что d_N значительно выше, чем d_S.

Модифицированный метод Ней-Годжобори имеет преимущества по сравнению с другими методами, так как он никогда не дает отрицательные значения d_S и d_N, когда различия между последовательностями малы и число неприменимых случаев гораздо меньше такового для других методов, когда имеются сильные различия между последовательностями. Использование p_S и p_N позволяет полностью избежать неприменимых случаев. Кроме того, оригинальный и модифицированный методы Ней-Годжобори дают меньшие вариансы p_S, p_N, d_S и d_N, чем другие методы.

Этот метод следует использовать при малых различиях между нуклеотидными последовательностями, в случае неприменимости других методов при сильных различиях между последовательностями, при значениях d_S <1,0 и d_N <1,0, а также при значениях R = 0,5 – 2. Для получения корректных значений S, N, p_S, p_N, d_S и d_N особое внимание следует уделять определению расчетного соотношения транзиций и трансверсий. Значения, полученные модифицированным методом Ней-Годжобори, используются и в селекционных тестах (см. главу 4).

Метод бутстрэп. Метод бутстрэп (bootstrap) используется в молекулярной эволюции при изучении синонимичных и несинонимичных нуклеотидных замещений как альтернативный метод для вычисления значений варианс дистанций (d_S, d_N, p_S и p_N).

Основа метода бутстрэп заключается в создании с помощью метода повторяющихся моделей (resampling method) пары последовательностей, содержащих количество случайных кодонов, равное таковому в сравниваемых последовательностях. Наиболее оптимально установить количество повторов (B) автоматически (чаще 500 или 1000). При каждом повторе рассчитывается промежуточная величина, а затем варианса.

В случае больших значений S_d, S_n, S и N этот метод позволяет получить более точные значения варианс дистанций, чем указанные выше формулы. Это связано с тем, что метод бутстрэп не основывается на гипотезе о том, что s_{di} и n_{di} определяются через $p_S s_i$ и $p_N n_i$.

Метод бутстрэп следует использовать для расчета варианс при работе с большими значениями S_d, S_n, S и N, а также в селекционных тестах.

3.2. Методы, основанные на двухпараметрической модели Кимуры

К методам, основанным на двухпараметрической модели Кимуры, относятся методы Ли-Ву-Ло, Памило-Бьянчи-Ли, Камерона, Камерона-Кумара и Ина. Рассмотрим наиболее часто используемые из них.

Метод Ли-Ву-Ло. В 1985 году Ли и соавторы предположили качественно новый метод изучения синонимичных и несинонимичных замещений нуклеотидов. Они обнаружили, что генетический код вырожден не только на уровне триплетов, но и на уровне сайтов. Так, нуклеотидные сайты кодонов можно разделить на "0"-кратно (невырожденные), "2"-кратно, и "4"-кратно вырожденные, исключая кодоны, соответствующие изолейцину. Невырожденным называется сайт, любая замена в котором является несинонимичной. Двукратно вырожденным считается сайт, в котором одна из трех замен является синонимичной. К четырехкратно вырожденным относят сайты, в которых все замены синонимичны. Например, третье положение кодона ГУА, кодирующего валин, является "4"-кратно вырожденным, так как все три замены являются синонимичными. Следует сказать, что второе положение всех кодонов является невырожденным, потому что все замены по этому положению являются несинонимичными. Третье положение трех кодонов, кодирующих изолейцин, следует считать "3"-кратно вырожденным, так как по ним две замены из трех синонимичны. Однако, по причине небольшого количества "3"-кратно вырожденных сайтов и для упрощения

вычислений Ли предложил подсчитывать их количество вместе с "2"-кратно вырожденными сайтами.

Подсчитаем количество трех вышеназванных типов сайтов для каждой из двух сравниваемых последовательностей. Затем обозначим среднее количество "0"-кратно, "2"-кратно и "4"-кратно вырожденных сайтов L_0, L_2 и L_4, соответственно.

Потом последовательно сравним все кодоны двух последовательностей и определим каждое нуклеотидное замещение как транзицию или трансверсию. Обозначим частоту транзиций и трансверсий в i-типе (i = 0, 2 и 4) нуклеотидного сайта как P_i и Q_i. Следует сказать, что транзиции, наблюдаемые в двукратно вырожденных сайтах, преимущественно синонимичны, в то время как большинство трансверсий - преимущественно несинонимичны. При вычислении P_i и Q_i этот метод учитывает все возможные эволюционные пути с учетом вероятностей аминокислотных замен.

Значения транзиций (A_i) и трансверсий (B_i) на сайт для каждого из трех типов нуклеотидных сайтов установим по формулам:

$$A_i = ½ \ln (a_i) - ¼ \ln (b_i),$$

$$B_i = ½ \ln (b_i),$$

где $a_i = 1 / (1 – 2P_i – Q_i)$ и $b_i = 1 (1 – 2Q_i)$. Основываясь на гипотезах о случайных и равновероятных заменах нуклеотидов, Ли предположил, что одна треть "2"-кратно вырожденных сайтов потенциально синонимичны, а две трети – потенциально несинонимичны. Исходя из этого, они предложили вычислять значения d_S и d_N и их варианс по формулам:

$$d_S = 3(L_2A_2 + L_4 (A_4 + B_4)) / L_2 + 3L_4,$$

$$d_N = 3(L_2B_2 + L_0 (A_0 + B_0)) / 2L_2 + 3L_0,$$

$$V(d_S) = 9(L_2^2V(A_2) + L_4^2 V (A_4 + B_4)) / (L_2 + 3L_4)^2,$$

$$V(d_S) = 9(L_2^2V(B_2) + L_0^2V (A_0 + B_0)) + / (2L_2 + 3L_0)^2,$$

где $V(A_i) = (a_i^2P_i + c_i^2Q_i – (a_iP_i + c_iQ_i)^2)/L_i$, $V(B_i) = b_i^2Q_i (1 – Q_i) / L_i$ и $c_i = (a_i – b_i)/2$.

Значения, полученные по этим формулам, не всегда соответствуют истинным значениям из-за наличия у метода Ли-Ву-Ло нескольких недостатков.

Во-первых, в сравниваемых последовательностях может отличаться тип гомологичных нуклеотидных сайтов. Например, по третьему положению один из гомологичных сайтов может быть двухкратно вырожденным, а второй – четырехкратно вырожденным. Тогда половина сайта будет учитываться как "2"-кратно вырожденная, а вторая половина как "4"-кратно вырожденная. Этот феномен часто наблюдается при сильных различиях сравниваемых последовательностей.

Во-вторых, нонсенс мутации учитываются как несинонимичные мутации. Например, в результате замены нуклеотида по третьему положению кодона УАУ, кодирующего тирозин, может произойти одна синонимичная (УАУ→УАЦ) и две терминальных мутации (УАУ→УАА и УАУ→УАГ). Две последние замены буду учтены как несинонимичные. Известно, что нонсенс мутации встречаются с частотой 4%, следовательно, значения d_N, полученные по методу Ли-Ву-Ло, будут завышены на эту величину.

В-третьих, транзиции по первому положению нуклеотида в трех из четырех "2"-кратно вырожденных аргининовых кодонах (ЦГГ, АГА и АГГ) являются несинонимичными, а в четвертом кодоне (ЦГА) приводят к возникновению нонсенс кодона. В то же время некоторые транзиции по третьему положению трех "3"-кратно вырожденных кодонов, кодирующих изолейцин, являются синонимичными.

В-четвертых, метод основан на некорректном соотношении транзиций и трансверсий, что объясняет неточные результаты особенно при высоких значениях R.

Несмотря на эти недостатки, при большом количестве кодонов в сравниваемых последовательностях и незначительных различиях метод Ли-Ву-Ло дает результаты, сходные с таковыми, полученными методом Нея-Годжобори. Если количество сравниваемых кодонов мало (<100), то метод Ли-Ву-Ло может давать отрицательные результаты из-за ошибочно вычисленных a_i и b_i.

В отличие от методов первой группы метод Ли-Ву-Ло позволяет рассчитать количество замен по невырожденным (d_0) и "4"-кратно (d_4) вырожденным сайтам и их вариансы ($V(d_0)$ и $V(d_4)$):

$$d_0 = A_0 + B_0,$$

$$d_4 = A_4 + B_4,$$

$$V(d_0) = (a_0^2P_0 + k_0^2Q_0 - (a_0P_0 + k_0^2Q_0)^2) / L,$$

$$V(d_4) = (a_4^2P_4 + k_4^2Q_4 - (a_4P_4 + k_4^2Q_4)^2) / L,$$

где $k_i = (a_i + b_i) / 2$. Значения d_0 применяются для определения скорости эволюции аминокислот, значения d_4 – для определения скорости нейтральной эволюции. В методе Ли-Ву-Ло вариансы можно рассчитывать как по формулам, так и методом бутстрэп.

Метод Ли-Ву-Ло применяется при большом количестве кодонов в сравниваемых последовательностях и незначительных различиях между ними, в селекционных тестах и при изучении скорости эволюции вырожденных сайтов.

Метод Памило-Бьянчи-Ли. В 1993 году Памило и Бьянчи и Ли независимо друг от друга модифицировали метод Ли-Ву-Ло, устранив его четвертый недостаток (некорректное соотношение транзиций и трансверсий).

Заметив, что согласно модели Ли и соавт. синонимичные транзиции встречаются только в двукратно и четырехкратно вырожденных сайтах, они предложили определять их общее число как $(L_2A_2 + L_4A_4) / L_2 + L_4$. Так как все трансверсии по "4"-кратно вырожденным сайтам (B_4) также синонимичны, то общее число синонимичных замен на синонимичный сайт определяется как:

$$d_S = (L_2A_2 + L_4A_4) / (L_2 + L_4) + B_4.$$

Используя те же аргументы, число несинонимичных замен на несинонимичный сайт определяется как:

$$d_N = (L_0B_0 + L_2B_2) / (L_0 + L_2) + A_0.$$

Вариансы синонимичной и несинонимичной дистанций вычисляются по формулам:

$$\mathbf{V(d_S) = V(B_4)} + \frac{[L_2^2V(A_2) + L_4^2V(A_4)]}{(L_2 + L_4)^2} - \frac{b_4Q_4[2a_4P_4 - c_4(1-Q_4)]}{(L_2 + L_4)}$$

$$\mathbf{V(d_N) = V(A_0)} + \frac{[L_0^2V(B_0) + L_2^2V(B_2)]}{(L_0 + L_2)^2} - \frac{b_0Q_0[2a_0P_0 - c_0(1-Q_0)]}{(L_0 + L_2)}$$

Значения $V(A_i)$, $V(B_i)$, c_i, k_i, d_0, d_4, $V(d_0)$ и $V(d_4)$ определяются так же, как и в методе Ли-Ву-Ло.

Метод Памило и соавт. устраняет лишь четвертый недостаток метода Ли-Ву-Ло и сохраняет три остальных недостатка.

Показания к использованию метода Памило-Бьянчи-Ли совпадают с таковыми метода Ли-Ву-Ло. Однако метод Памило-Бьянчи-Ли позволит получить более точные значения синонимичной и несинонимичной дистанций и их вариансс, особенно при высоком значении R.

Метод Камерона. Третьим недостатком метода Ли-Ву-Ло и Памило-Бьянчи-Ли является некорректная оценка кодонов, кодирующих аргинин и изолейцин. Это становится особенно важным, когда высока частота встречаемости этих аминокислот (для протамина P1 млекопитающих частота встречаемости аргинина составляет приблизительно 50%).

Камерон предпринял попытку устранить третий недостаток метода Ли-Ву-Ло, разделив "2"-кратно вырожденные сайты на две группы: 2S и 2V. К первой группе он отнес сайты, в которых две транзиции синонимичны, а трансверсия несинонимична. Ко второй группе Камерон отнес сайты, в которых транзиция несинонимична, а две трансверсии синонимичны. Это разделение

позволило частично устранить неточность классификации Ли по синонимичным и несинонимичным сайтам (например, кодоны, кодирующие метионин).

Однако это не устранило всю проблему. Так, замены нуклеотидов в первом положении кодона ЦГГ, кодирующего аргинин, выглядят так: ЦГГ→УГГ (Три), ЦГГ→АГГ (Арг) и ЦГГ→ГГГ (Гли). В этом случае транзиция Ц→У приводит к несинонимичной замене, трансверсия Ц→А – к синонимичной, а трансверсия Ц→Г – к несинонимичной. Таким образом, этот нуклеотидный сайт не может быть классифицирован как 2S или 2V. Первое положение нуклеотида еще трех аргининовых кодонов (ЦГУ, ЦГЦ и ЦГА) и третье положение нуклеотида двух изолейциновых кодонов (АУУ и АУЦ) также не могут быть отнесены ни к одной из групп по классификации Камерона.

Ввиду редкого использования метода Камерона нецелесообразно рассматривать формулы, по которым производится вычисление основных показателей.

Метод Камерона-Кумара. Еще одну модификацию для полного устранения третьего недостатка метода Ли-Ву-Ло предложил Кумар. Так как эта модификация является логичным продолжением модификации Камерона, то ее часто называют методом Камерона-Кумара.

Этот метод основан на более сложной классификации вырожденных сайтов, заключающейся в делении "2"-кратно вырожденных сайтов на простые и сложные (табл.1).

Таблица 3.1

Классификация и обозначения вырожденных сайтов по Кумару

Вырожденность сайта	"0"-кратно	"2"-кратно			"4"-кратно
		простые	сложные		
Количество сайтов	L_0	L_2S	L_2C		L_4
			синонимичные	несинонимичные	
Транзиции (s)	s_0	s_2	s_2S	s_2N	s_4
Трансверсии (v)	v_0	v	v_2S	v_2N	v_4

К простым двукратно вырожденным сайтам Кумар относит сайты, в которых транзиция приводит к синонимичной замене, а две трансверсии – к нонсенс или несинонимичным заменам. Все остальные "2"-кратно

вырожденные сайты, включая три кодона, кодирующих изолейцин, относятся к сложным. На основе этой классификации Кумар предложил новый метод расчета d_S и d_N. Вычислим частоты транзиций и трансверсий в "0"-кратно, "2"-кратно и "4"-кратно вырожденных сайтах как:

$$P_0 = (s_0 + s_2N) / (L_0 + L_2C),$$
$$P_2 = (s_0 + s_2S) / (L_2S + L_2C),$$
$$P_4 = s_4 / L_4,$$
$$Q_0 = v_0 / L_0,$$
$$Q_2 = (v_0 + v_2N) / (L_2S + L_2C),$$
$$Q_4 = (v_4 + v_2S) / (L_4 + L_2C).$$

Затем рассчитаем A_i и B_i так же, как в методе Ли-Ву-Ло. Для получения d_S, d_N и их варианс подставим в формулу Памило-Бьянчи-Ли $L_2 = L_2C + L_2S$. Для вычисления d_0, d_4 и их варианс метод Кумара использует формулу Ли-Ву-Ло.

Учитывая вышесказанное, метод Кумара является наиболее корректным методом, основанным на двухпараметрической модели Кимуры, что обусловливает его использование для более точного вычисления синонимичных и несинонимичных дистанций, в селекционных тестах и при изучении скорости эволюции вырожденных сайтов.

Методы Ина. Ина предложил другой подход к определению d_S и d_N, основанный на комбинации оригинального метода Ней-Годжобори и метода Памило-Бьянчи-Ли. Он разработал два метода: метод I и метод II.

В первом методе используется соотношение наблюдаемых в третьем положении кодона транзиций и трансверсий, определяемое как $k = \alpha/\beta$. Это основано на гипотезе о том, что замены нуклеотидов в третьем положении кодона в большинстве случаев нейтральны. Значения S и N вычисляются по модифицированному методу Ней-Годжобори, а значения S_d и N_d - по оригинальному методу Ней-Годжобори. Однако Ина разделяет S_d на синонимичные транзиции (S_{Ts}) и синонимичные трансверсии (S_{Tv}), а также N_d на несинонимичные транзиции (N_{Ts}) и несинонимичные трансверсии (N_{Tv}).

При малом количестве нуклеотидов и незначительных различиях сравниваемых последовательностей этот метод не может быть применим, поскольку частота транзиций или трансверсий может быть равна 0, что даст $k = 0$ или ∞.

Во втором методе значения S и N определяются для всех положений нуклеотида в кодоне, но α и β вычисляются только для синонимичных

замещений для исключения произошедших до отбора мутаций, что усложняет процесс вычисления.

Компьютерное моделирование показало сходство значений d_S и d_N, полученных методом Ина I и II с таковыми, полученными модифицированным методом Ней-Годжобори. Однако практическое применение методов Ина ограничивается их сложностью.

3.3. Методы максимального сходства с моделями замещений кодонов

К методам максимального сходства с моделями замещений кодонов относятся методы Голдмана-Янга и Мьюза.

Метод Голдмана-Янга. Голдман и Янг разработали модель нуклеотидных замещений для 61 смыслового кодона, прототипом которой является модель Хасегава-Кишино-Яно. Посчитаем относительную частоту j-кодона (π_j) для пары последовательностей, содержащих C гомологичных кодонов.

Голдман и Янг предположили, что частота неспонтанных замещений кодона i на кодон j ($i \neq j$), обозначаемая как q_{ij}, определяется следующими уравнениями:

$q_{ij} = 0$ (если замена нуклеотида происходит по двум и более положениям),

$q_{ij} = \pi_j$ (для синонимичных трансверсий),

$q_{ij} = k\pi_j$ (для синонимичных транзиций),

$q_{ij} = \omega\pi_j$ (для несинонимичных трансверсий),

$q_{ij} = \omega k\pi_j$ (для несинонимичных транзиций),

где k – это соотношение наблюдаемых транзиций и трансверсий ($k = \alpha / \beta$), а ω - соотношение несинонимичных (r_N) и синонимичных (r_S) замен ($\omega = r_N / r_S$).

Каждому из смысловых кодонов соответствует свое значение π_j. Предположим, что частоты кодонов равны. Тогда значения π_j могут быть определены исходя из наблюдаемых частот использования кодонов, если число кодонов (C) велико. Значения k и ω можно определить, используя метод максимального сходства Голдмана и Янга.

В случае если C относительно мало, этот подход не позволит получить точные значения π_j, потому что значение π_j будет мало, а его стандартная ошибка велика. В таком случае можно определить π_j как произведение наблюдаемых частот нуклеотидов.

Одним из преимуществ такого подхода является одновременное определение значений k и ω, если удовлетворяется модель Голдмана –Янга.

Так, нет необходимости определять R (= 2k) для вычисления d_S и d_N, как в модифицированном методе Ней-Годжобори.

Однако этот метод имеет ряд недостатков. Во-первых, значения π_j, основанные на наблюдаемых частотах, не будут корректными, когда С мало. Определение π_j как произведения частот нуклеотидов, также даст некорректные результаты, особенно при неравновероятном использовании кодонов.

Во-вторых, предположение о равных значениях ω для всех положений нуклеотида в кодоне является ошибочным. Это приведет к тому, что значение ω будет резко отличаться от величины d_S/d_N, так как среднее соотношение r_N/r_S для двух последовательностей не совпадает с соотношением средних r_N и r_S.

В-третьих, предположение о том, что значения k и ω для каждой пары кодонов не зависят друг от друга, также является ошибочным. Поэтому необходимы дальнейшие исследования, посвященные устранению названных выше недостатков.

Компьютерное моделирование показало, что в большинстве случаев значения синонимичной и несинонимичной дистанций, полученные методом Голдмана-Янга, резко отличаются от таковых, полученных с помощью других методов. Таким образом, данный метод не может быть применен в практических исследованиях и нуждается в доработке.

Метод Мьюза. В 1996 году Мьюз разработал похожий метод сходства, основанный на разных моделях замещений нуклеотидов. В этом методе частоты кодонов определяются как произведения частот нуклеотидов, а соотношение транзиций и трансверсий не учитывается. Следует отметить, что количество показателей, определяемых в этой модели, меньше, чем в модели Голдмана-Янга. При малых различиях в частоте использования кодонов этот метод позволяет получить значения d_S и d_N, сходные с оригинальным методом Ней-Годжобори.

Метод Мьюза дает некорректные значения дистанций при высоком соотношении транзиций и трансверсий и значительных различиях в частотах использования кодонов. По этой причине использование метода Мьюза в молекулярной эволюции и эволюционной биохимии ограничено.

3.4. Тенденции и перспективы развития методов изучения синонимичных и несинонимичных нуклеотидных замещений

С развитием компьютерных технологий стало возможным использовать более сложные математические модели и, основываясь на них, выполнять статистический анализ. В качестве примера такой модели можно назвать модель Нилсена-Янга. Однако, требуя большее количество величин, эти модели

становятся нелогичными, что приводит к нарушению основных гипотез. Вследствие этого модель дает неточные результаты.

Совместное использование модифицированного метода Ней-Годжобори и метода Камерона-Кумара позволит получить наиболее корректные, хотя и не совсем точные результаты. Это связано с тем, что полученные модифицированным методом Ней-Годжобори результаты зависят от методики вычисления R, а метод Камерона-Кумара наследует два недостатка (первый и второй) методов Ли-Ву-Ло и Памилло-Бьянчи-Ли.

Следовательно, основными перспективами развития методов изучения синонимичных и несинонимичных нуклеотидных замещений являются коррекция методики определения R, усовершенствование метода Камерона-Кумара и создание новых методов на основе математических моделей и максимального сходства с моделями замещений нуклеотидов.

3.5. Сравнительный анализ методов изучения синонимичных и несинонимичных нуклеотидных замен в последовательностях мРНК алкогольдегидрогеназ

Оценим результаты, полученные наиболее часто используемыми методами, на примере нуклеотидных последовательностей мРНК алкогольдегидрогеназы класса III земноводных и человека, которые были предварительно проанализированы и выровнены с помощью программы Clustal W.

Для определения картины замещений в сравниваемых последовательностях определен индекс несоответствия (index disparity - ID) и проведен ID-тест для определения вероятности (P) отклонения нулевой гипотезы о гомогенной картине замещений нуклеотидов на 5%-ном уровне. Если величина P больше 0,05, то картину замещений принято считать гомогенной, в обратном случае – гетерогенной. Полученное значение ID = 0,0000 и значение P = 1,0000 свидетельствую о гомогенной картине замещений.

Расчетное соотношение транзиций и трансверсий R для гомогенной картины замещений в сравниваемых последовательностях рассчитанное по методу Кимура составило 1,3654±0,1882 (вариансы вычислялись по аналитическим формулам), по методу Тамура – 1,3670±0,1884 и по методу Тамура-Ней - 1,3871±0,1930, что дает среднее R = 1,3732 ±0,0086.

Значение расчетного соотношения транзиций и трансверсий необходимо для вычисления дистанций модифицированным методом Нея-Годжобори. Полученные значения d_S и d_N, а также их вариансы представлены в табл. 3.2.

Таблица 3.2

Значения синонимичных и несинонимичных дистанций, а также их варианс, полученные различными методами, для последовательностей мРНК алкогольдегидрогеназы класса III земноводных и человека

Метод	d_S	V (d_S)	d_N	V (d_N)
Ней-Годжобори	2,6648	0,9508	0,0902	0,0108
Модифицированный Ней-Годжобори	1,4019	0,1994	0,0948	0,0130
Ли-Ву-Ло	1,8306	0,2768	0,0883	0,0106
Памило-Бьянчи-Ли	1,6032	0,2100	0,0905	0,0110
Камерона-Кумара	1,3399	0,1571	0,0888	0,0110

Примечание. Значения варианс вычислены по аналитическим формулам.

Из приведенных данных видно, что значения d_S, вычисленные разными методами, сильно варьируют. Так, наибольшее значение синонимичной дистанции получено методом Ней-Годжобори (2,6648±0,9508), а наименьшее – методом Камерона-Кумара (1,3399±0,1571).

Такая большая разбежка значений связана с тем, что соотношение транзиций и трансверсий в сравниваемых последовательностях значительно выше такового, при котором оригинальный метод Ней-Годжобори дает корректные результаты (R = 0,5). Еще одной причиной, объясняющей сильную вариацию значений синонимичной дистанциии, является большое количество синонимичных замещений нуклеотидов.

Значения несинонимичной дистанции, рассчитанные разными методами, напротив, незначительно отличаются друг от друга.

Наибольшее значение d_N получено модифицированным методом Ней-Годжобори (0,0948±0,0130), а наименьшее – методом Ли-Ву-Ло (0,0883±0,0106). Это можно объяснить незначительным количеством несинонимичных нуклеотидных замен.

Учитывая методику расчета значений d_S и d_N, наиболее точные результаты следует ожидать от модифицированного метода Ней-Годжобори и метода Камерона-Кумара.

Так, полученные ими значения d_S сходны и составляют 1,4019±0,1994 и 1,3399±0,1571, соответственно. Значения несинонимичной дистанции также сходны (0,0948±0,0130 и 0,0888±0,0110, соответственно).

Значения d_0 и d_4, а также их вариансы представлены в табл. 3.3.

Таблица 3.3

Значения дистанций по невырожденным и "4"-кратно вырожденным сайтам, а также их варианс, полученные различными методами, для последовательностей мРНК алкогольдегидрогеназы класса III земноводных и человека.

Метод	d_0	V (d_0)	d_4	V (d_4)
Ли-Ву-Ло	0,0786	0,0108	1,5125	0,2900
Памило-Бьянчи-Ли	0,0786	0,0108	1,5125	0,2900
Камерона-Кумара	0,0779	0,0107	1,2975	0,2231

Примечание. Значения варианс вычислены по аналитическим формулам.

Из приведенных данных видно, что методы Ли-Ву-Ло и Памило-Бьянчи-Ли дают идентичные результаты значений d_0 и d_4, поскольку используют одинаковые формулы. Сравним значения полученные методами Ли-Ву-Ло и Памило-Бьянчи-Ли с таковыми, вычисленными методом Камерона-Кумара.

Значение дистанции по "4"-кратно вырожденным сайтам при использовании методов Ли-Ву-Ло и Памило-Бьянчи-Ли выше (1,5125±0,2900), чем при использовании метода Камерона-Кумара (1,2975±0,2231), что связано с большим количеством замещений нуклеотидов в "4"-кратно вырожденных сайтах. Значение дистанции по невырожденным сайтам, полученное методами Ли-Ву-Ло и Памило-Бьянчи-Ли, незначительно выше (0,0786±0,0108) такового, полученного методом Камерона-Кумара (0,0779±0,0107).Учитывая теоретические основы этих методов можно утверждать, что метод Камерона-Кумара даст более точные результаты.