Sequence Analysis

MV

Module II

- **Sequence analysis** is the process of subjecting a DNA, RNA or peptide sequence to any of a wide range of analytical methods to understand its features, function, structure, or evolution.

- Methodologies used include sequence alignment, searches against biological databases, and others. Since the development of methods of high-throughput production of gene and protein sequences, the rate of addition of new sequences to the databases increased exponentially. Such a collection of sequences does not, by itself, increase the scientist's understanding of the biology of organisms. However, comparing these new sequences to those with known functions is a key way of understanding the biology of an organism from which the new sequence comes.

- Thus, sequence analysis can be used to assign function to genes and proteins by the study of the similarities between the compared sequences.

Sequence Similarity Search

- Sequence analysis is used to compare two or more sequences

- Comparison of protein & DNA sequences to find similarities/differences - chief task in bioinformatics

- The process of comparing two or more sequences to find out similarity between them is called sequence alignment

- By sequence comparison it is possible to find out relationship in structure, function and evolution from a common ancestor

- Similarity - identical (similar) residues occur at identical (similar) positions

- No. of such matches indicates the degree of similarity

  ATCGTA   4/6 = 66%

  ATGCTA

- Similarity occurs by chance, evolutionary convergence or evolutionary divergence

Sequence Similarity

Homologous sequences - similarity by evolutionary divergence from a common ancestor

- Sequence with common origin
- All homologous sequences are similar, but all similar sequences are not necessarily homologous
- Eg: flight muscle protein in bat, crow and fly

Paralogous sequences - homologous sequences that exist in the same organism, but have different functions

- Eg: myoglobin and haemoglobin

Orthologous sequences - homologous origin, same function in different species

- Eg: Haemoglobin in horse and zebra

Xenologous sequences - homologues resulted from lateral or horizontal gene transfer

- Similar sequence in completely unrelated species
- Xenologous are orthologous
- Eg: Bacteriophage and bacteria

Types of Sequence Alignment

I.   According to no. of sequences being compared

1.   Pair wise sequence alignment - only 2 sequences are compared with each other to find the region of similarity

2.   Multiple sequence alignment - more than 2 sequences are compared with each other to find the region of similarity

II. According to the length of sequences being compared

1.   Global sequence alignment - sequences are compared along their entire length to include as many matching characters as possible

2.   Local sequence alignment - sequences are aligned to find local region of higher similarity

Types of sequence alignment

- There are two main types of sequence alignment. Pair-wise sequence alignment (PSA) only compares two sequences at a time and multiple sequence alignment (MSA) compares many sequences in one go.

- Two important algorithms for aligning pairs of sequences are the Needleman-Wunsch algorithm and the Smith-Waterman algorithm.

- A common use for pairwise sequence alignment is to take a sequence of interest (Query sequence) and compare it to all known sequences in a database to identify homologous sequences.

- A **multiple sequence alignment** (**MSA**) of three or more biological sequences, generally protein, DNA, or RNA. In many cases, the input set of query sequences are assumed to have an evolutionary relationship by which they share a lineage and are descended from a common ancestor. From the resulting MSA, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences' shared evolutionary origins.

- Multiple sequence alignment is often used to assess sequence conservation of protein domains, tertiary and secondary structures, and even individual amino acids or nucleotides.

Popular tools for sequence alignment include:

- Pair-wise alignment - BLAST
- Multiple alignment - ClustalW, PROBCONS, MUSCLE, MAFFT, and T-Coffee.

Tools for Sequence Alignment

1. BLAST - Basic Local Alignment Search Tool

- Developed by Altschul *et al.,* 1990
- Fast and sensitive database search to find similar sequences
- BLAST is available in NCBI website - 5 types

- 1. BLAST N - for Nucleotide search
- 2. BLAST P - for Protein search
- 3. BLAST X - search protein db for Nucleotide translated to Protein
- 4. TBLASTN - search nucleotide translated to protein db for Protein
- 5. TBLASTX - search nucleotide translated to protein db for nucleotide translated to protein

Global Alignment v/s Local Alignment

Local Alignment

Target Sequence

5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
         ||||| ||||||| |||||||||||||||||||
Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'

Global Alignment

Target Sequence

5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   ||||||||||||| |||||||| ||||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'

Query Sequence

Species/Abbrv
1. Rhizobium_leguminosarum_bv._viciae_3841_g115254414
2. Sinorhizobium_medicae_WSM419_g150026743
3. Agrobacterium_fabrum_str._C58_g159139455
4. Agrobacterium_fabrum_str._C58_g159140696
5. Rhizobium_etli_CIAT_652_g190694918
6. Rhizobium_leguminosarum_bv._trifolii_WSM2304_g209533368
7. Agrobacterium_radiobacter_K84_g221721649
8. Agrobacterium_vitis_S4_g221737306
9. Sinorhizobium_fredii_NGR234_g227339586
10. Rhizobium_leguminosarum_bv._trifolii_WSM1325_g240856645
11. Sinorhizobium_meliloti_1021_g30407155
12. Candidatus_Liberibacter_solanacearum_CLso-ZC1_g313495152
13. Agrobacterium_sp._H13-3_g325062059

Scoring Matrices


```
T G C G T A T G A T A
| |   |   |     |   |       |
T G C G G A T - - T C
2+2+2+2+0+2+2-1-1+2+0 = 12
```

Distance matrices from alignment scores

- The distance matrix can come from a number of different sources, including measured distance or morphometric analysis, various pairwise distance formulae applied to discrete morphological characters, or genetic distance from sequence. For phylogenetic character data, raw distance values can be calculated by simply counting the number of pairwise differences in character states (Hamming distance).

TABLE 27.6. Distance matrix

OTUs	A	B	C	D	E	F
A	0	2	4	6	6	8
B	2	0	4	6	6	8
C	4	4	0	6	6	8
D	6	6	6	0	4	8
E	6	6	6	4	0	8
F	8	8	8	8	8	0

Genetic Distance

ID	01-155344-BillB	12-163235-BillB-L	23-228581-Betty-Hal	27-268192-JBR	28-235333-Cindy-Justin	15-132792-Rob	09-116075-Lynda	04-093103-Dr.Bob	22-088443-Diane	18-N31075-John	11-086737-Geoff	06-164024-Camilla	13-111309-Sandra	24-199275-Scott	08-121393-Donna	25-170166-Joe-James	07-057662-LouDean	05-172807-Donald	19-180265-Rev.Rod
01-155344-BillB.........	111	5	9	11	10	9	9	7	6	8	7	7	8	10	9	9	12	10	11
12-163235-BillB-L......	5	111	8	8	7	6	6	6	5	5	4	4	6	7	6	7	9	7	8
23-228581-Betty-Hal....	9	8	111	10	8	8	6	6	5	7	6	6	5	7	6	4	9	7	8
27-268192-JBR..........	11	8	10	111	8	5	6	6	5	5	4	4	6	7	6	9	10	7	8
28-235333-Cindy-Justin.	10	7	8	8	111	6	6	6	5	5	4	4	6	7	6	9	10	7	8
15-132792-Rob..........	9	6	8	5	6	111	4	4	3	3	2	2	4	5	4	7	8	5	6
09-116075-Lynda........	9	6	6	6	6	4	111	4	3	3	2	2	4	4	4	5	8	5	6
04-093103-Dr.Bob.......	7	6	6	6	6	4	4	111	1	3	2	2	3	5	4	6	8	5	6
22-088443-Diane........	6	5	5	5	5	3	3	1	111	2	1	1	2	4	3	5	7	4	5
18-N31075-John.........	8	5	7	5	5	3	3	3	2	111	1	1	3	4	3	6	7	4	5
11-086737-Geoff........	7	4	6	4	4	2	2	2	1	1	111	0	2	3	2	5	6	3	4
06-164024-Camilla......	7	4	6	4	4	2	2	2	1	1	0	111	2	3	2	5	6	3	4
13-111309-Sandra.......	8	6	5	6	6	4	4	3	2	3	2	2	111	3	2	3	6	3	4
24-199275-Scott........	10	7	7	7	7	5	4	5	4	4	3	3	3	111	3	6	7	4	5
08-121393-Donna........	9	6	6	6	6	4	4	4	3	3	2	2	2	3	111	3	4	3	4
25-170166-Joe-James....	9	7	4	9	9	7	5	6	5	6	5	5	3	6	3	109	6	6	7
07-057662-LouDean......	12	9	9	10	10	8	8	8	7	7	6	6	6	7	4	6	111	7	8
05-172807-Donald.......	10	7	7	7	7	5	5	5	4	4	3	3	3	4	3	6	7	111	2
19-180265-Rev.Rod......	11	8	8	8	8	6	6	6	5	5	4	4	4	5	4	7	8	2	111

Related | Probably Related | Possibly Related

- Distance-matrix methods of phylogenetic analysis explicitly rely on a measure of "genetic distance" between the sequences being classified, and therefore they require an MSA (multiple sequence alignment) as an input. Distance is often defined as the fraction of mismatches at aligned positions, with gaps either ignored or counted as mismatches. Distance methods attempt to construct an all-to-all matrix from the sequence query set describing the distance between each sequence pair.

- Distance-matrix methods may produce either rooted or unrooted trees, depending on the algorithm used to calculate them.

scoring schemes

- **Basic scoring schemes**
- The simplest scoring schemes simply give a value for each match, mismatch and indel. For this system the alignment score will represent the edit distance between the two strings. Different scoring systems can be devised for different situations, for example if gaps are considered very bad for your alignment you may use a scoring system that penalises gaps heavily, such as:
- Match = 0
- Mismatch = 1
- Indel = 10

Similarity Matrix

- More complicated scoring systems attribute values not only for the type of alteration, but also for the letters that are involved. For example, a match between A and A may be given 1, but a match between T and T may be given 4. Here (assuming the first scoring system) more importance is given to the Ts matching than the As, i.e. we think the Ts matching is more significant to our alignment. This weighting based on letters also applies to mismatches. In order to represent all the possible combinations of letters and their resulting scores we use a similarity matrix.

Gaps

- GCATG-CU
- G-ATTACA
- Gaps in a DNA sequence result from either insertions or deletions in the sequence, sometimes referred to as indels. Insertions or deletions occur due to single mutations, unbalanced crossover in meiosis, slipped strand mispairing in the replication process and chromosomal translocation. In alignments gaps are represented as contiguous dashes on a protein/DNA sequence alignment. Gaps are missing data in an alignment matrix

gap penalties

- The **gap penalty** is a scoring system used in bioinformatics for aligning a small portion of genetic code, more accurately, fragmented genetic sequence.

- The transcription and translation or DNA replication can produce errors resulting in mutations in the final nucleic acid sequence. Therefore, in order to make more accurate decisions in aligning sequences, mutations are annotated as gaps.

- Gaps are penalised via various Gap Penalty scoring methods. The scoring that occurs in Gap Penalty allows for the optimisation of sequence alignment in order to obtain the best alignment possible based on the information available. The three main types of gap penalties are constant, linear and affine gap penalty

1. Constant

- This is the simplest type of gap penalty: a fixed negative score is given to every gap, regardless of its length

2. Linear

- Compared to the constant gap penalty, the linear gap penalty takes into account the length (L) of each insertion/deletion in the gap. Therefore, if the penalty for each inserted/deleted element is B and the length of the gap L; the total gap penalty would be the product of the two BL. This method favors shorter gaps, with total score decreasing with each additional gap.

3. Affine

- The most widely used gap penalty function is the affine gap penalty. The affine gap penalty combines the components in both the constant and linear gap penalty, taking the form A+(B·L). This introduces new terms, A is known as the gap opening penalty, B the gap extension penalty and L the length of the gap. Gap opening refers to the cost required to open a gap of any length, and gap extension the cost to extend the length of an existing gap by 1.

Working of BLAST

- 3 steps

1. Constructing a list of words:

    query sequence is broken into short fragments of x length - known as neighbor words

ATGCGTAGT - ATGCGTA, TGCGTAG, GCGTAGT

2. Searching db for hits:

    each neighbor word of query is compared with each word of db,

    if any word of query sequence is identical to the word of db, a hit is recorded

3. Hit extension:

    All the hits generated in the previous steps are extended without gaps in both directions to determine a larger segment of similarity

    A threshold score is maintained to make extension process fast

    Extension terminated when the score of the segment pair falls below the threshold score

    All the scores equal to or better than threshold score are termed as High Scoring Segment Pairs (HSPs)

    The highest scoring pair is termed as Maximal Segment Pair (MSP)

2. FASTA

- Developed by Pearson & Lipman, 1988
- Designed to find pair wise similarity between sequences
- In the FASTA method, the user defines a value k to use as the word length with which to search the database. The method is slower but more sensitive at lower values of k, which are also preferred for searches involving a very short query sequence.

Steps:

1. Identification of 10 best similarity regions between query sequence and db
2. Restoring 10 best regions using substitution matrix

    PAM - Percent Accepted Mutation

    BLOSUM - Block Amino acid Substitution Matrix

- Allows conservative replacement and sub-alignment
- The highest scoring sub-alignments are determined - whose score is called init 1 score
- init 1 score is used to rank matches for further analysis

3. Selection of longer region of similarity

- init n score - score of fragments from a longer region that can be joined

4. Alignment of highest scoring db sequences

- Sequences with init n score greater than the threshold score is aligned with query sequence
- Score of final alignment is called opt score - it is used to rank db matches

3. CLUSTAL

- Multiple sequence alignment program for DNA or proteins
- Clustal W - command line interface
- Clustal X - graphical interface
- Produce multiple sequence alignments of divergent sequences
- Calculate the best match for the selected sequences - so similarities and difference can be seen
- Evolutionary relationship can be seen in phylogenetic trees

Steps:

1. Pair wise alignment
2. Construction of phylogenetic trees

- Clustal is used to align sequences that are related to each other over their entire length
- Not used
  1. Sequences without common ancestry
  2. Sequences have large variable C & N subunits
  3. Sequences are partially related
  4. Sequences include short overlapping fragments

Weights

- • Helices
- • Sheets
- • Active Sites or other functional regions
- • Disulfide Bonds

Mutations are the raw materials …
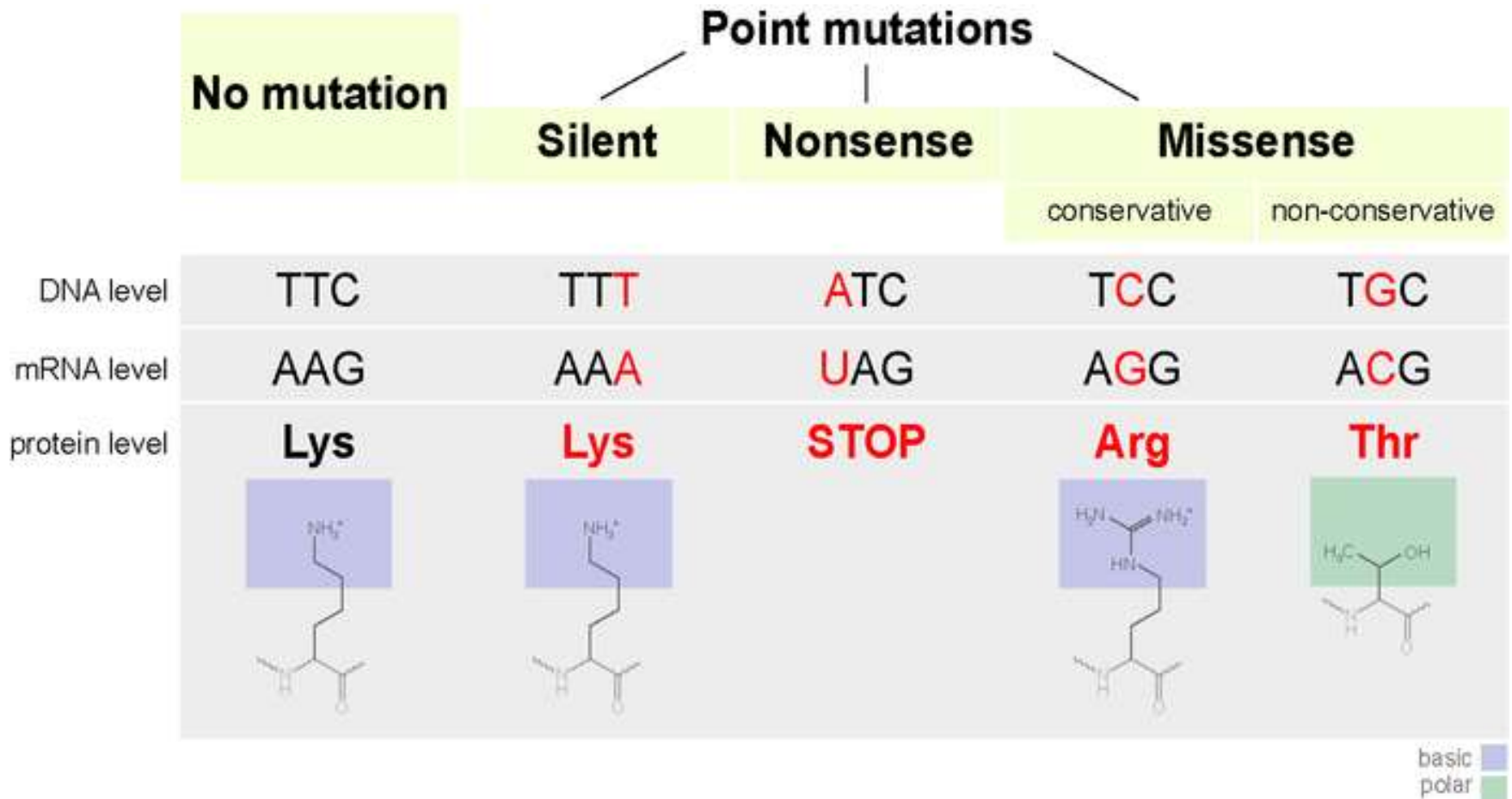
Types of Mutations

- These mutations are easily detected by multiple alignment
- • Base Substitutions
- • Indel - Insertions & Deletions
- These mutations are not easily detected by multiple alignments
- • Transposition
- • Exon (domain) Shuffling

Nucleotides contain more data?

Sequence 1 | CTA GCT AGA GGA AGC CCA ACA GTA

Sequence 2 | TTG GCG CGT GGG TCT CCG ACC GTT

LARGSPT

The protein could be under evolutionary pressure. If you only use protein data, you won't see all the mutational events going on in the background.

Substitution Model

- Jukes - Cantor - No bias. Substitutions occur randomly. Equal probability of mutation for all nucleotides.

- Kimura (2 parameter)- Transitions (C->T or A->G) more frequent than transversions (A ->T, C->G)

Tree calculation method

- Distance Methods –

- Evolutionary distances are used to construct trees (UPGMA & Neighbor Joining).

- Fast, easy to handle large numbers of sequences.

- Character Methods –

- Parsimony Methods - trees are created to minimize the number of changes that are needed to explain the data.

- Maximum Likelihood - Using a model for sequence evolution, create a tree that gives the highest likelihood of occurring with the given data.

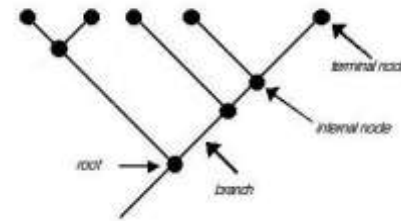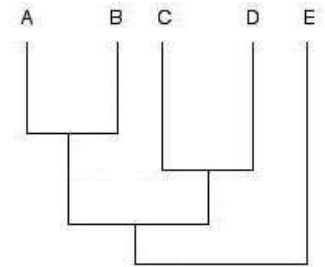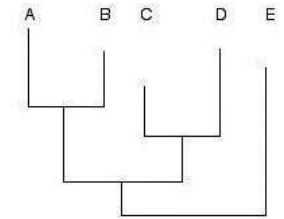# Construction of Phylogenetic Trees

- Phylogram - Branch Lengths Proportional To Distance

- Cladogram - All Branch Lengths Equal

- Outgrouping (Outgroup - An OTU that is the least related to the group of taxa studying). Defining an outgroup is one way of rooting a tree. (Rooted Tree-A common ancestor is defined)

- Unrooted Tree-No common ancestor
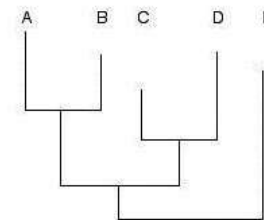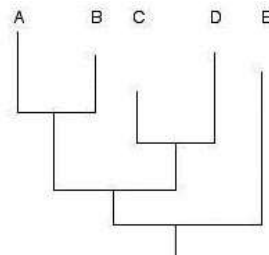
Phylogenetic trees

- Phylogeny - evolutionary relationship among species
- Phylogram - branching diagram with different branch length
  to estimate phylogeny

  Branch length - amount of evolutionary change
- Cladogram - phylogram with equal branch length

  No indication of evolutionary time separating taxa
- Phylogenetic trees are constructed by nodes and branches
- Every node represents a taxonomic unit

- Trees are 2 types
- Rooted tree - single node is designated as a common ancestor
- Unrooted tree - say nothing about the direction of evolution or origin

- There are two basic methods to construct a phylogenic tree from genetic distance matrix
- UPGMA
- Neighbour Joining

- Unweighted Pair Gap Method with Arithmetic Mean (UPGMA)

- Simplest Method for Tree Construction

- Sequential clustering method - Start with one pair of OTUs (Operational Taxonomic Unit) and sequentially add other OTUs

- The most parsimonious tree, or shortest tree is one that requires the fewest total evolutionary events (for example, substitutions).

- Heuristic Search - Fastest method of finding trees. Uses a variation of Neighbor-Joining
- Not guaranteed to find the optimal tree

Statistical Methods to Evaluate Trees

- Bootstrapping - commonly used for estimating statistics when the distribution is difficult to derive analytically.

- Method - resample and reanalyze single row of characters. Look for groupings that appear frequently as a measure of confidence in a particular tree.

PAUP

- PAUP - Phylogenetic Analysis Using Parsimony
- Common software program for analyzing sequences using parsimony.
- Can also create trees using distance methods.

Arctic fox — 100 / 100
Kit fox — 100
Corsac fox — 100 / 100
Ruppell's fox — 52 / 98
Red fox — 81 / 100
Cape fox — 67 / 99
Blanford's fox — 96 / 100
Fennec fox
Raccoon dog — 75 / 100
Bat-eared fox

Short-eared dog
Crab-eating fox — 100 / 100
Sechuran fox — 99
Culpeo fox — 81 / 100
Pampas fox — 99
Chilla — 99
Darwin's fox — 93 / 100
Hoary fox
Maned wolf — 96 / 100
Bush dog

Side-striped jackal — 100 / 100
Black-backed jackal
Golden jackal — 100 / 100
Dog — 93 / 100
Grey wolf — 78 / 100
Coyote — 99 / 100 (2)
Ethiopian wolf — 92 / 100
Dhole — 78 / 100
African wild dog

Grey fox — 100 / 100
Island fox

Black bear — 100 / 100
Giant panda
Northern elephant seal — 100 / 100
Walrus

80 / –
100 / 100
75 / 100
9–10 Myr
6–7.4 Myr
3–4 Myr
100 / 100
(9)
75 / 100
100 / 100 (4)
100 / 100