

Bioinformatics CSM17 Week 7: Molecular Analysis

- Sequence comparison
- Molecular characters
- Homoplasy and convergence
- Multiple Sequence Alignment
- Cladograms from Molecular Data

Molecular data

A T G C A T G C Sense Strand
(Partner)

| | | | | | | |

A U G C A U G C Messenger RNA

| | | | | | | |

T A C G T A C G Antisense
(Template)

Sequence Comparison

Simple Alignment (see also Skelton & Smith [2002], Sect. 2.2 p29)

match score: 1

mismatch score 0

A A T C T A T A

A A G A T A

4 + 0 = 4 (best)

A A T C T A T A

A A G A T A

1 + 0 = 1 (worst)

A A T C T A T A

A A G A T A

3 + 0 = 3

Sequence Comparison

Simple Alignment with gap penalties

match score: 1 *mismatch score 0* *gap penalty -1*

A A T C T A T A

A A G - A T - A

$$3 + 0 - 2 = 1 \quad (\text{worst})$$

A A T C T A T A

A A - G - A T A

$$5 + 0 - 2 = 3 \quad (\text{equal$$

best)

A A T C T A T A

A A - - G A T A

$$5 + 0 - 2 = 3 \quad (\text{equal$$

best)

A A T C T A T A

- A A G A T A -

$$1 + 0 - 2 = -1 \quad (\text{worst})$$

Sequence Comparison

Simple Alignment with origination and length penalties

match score: 1 *mismatch score 0*

origination penalty: -2 *length penalty -1*

A A T C T A T A
A A - G - A T A 5 + 0 - 4 - 2 = -1 (worst)

A A T C T A T A
A A - - G A T A 5 + 0 - 2 - 2 = 1 (best)

Origination penalty is applied for starting a series of gaps

Length penalty is also applied for each gap

Mutation (and copying errors)



Changes of nucleotide base sequences

- caused by
 - ionizing radiation, mutagenic chemicals, errors
- Mutations are usually harmful (damaging)
- may be
 - single base (changing one amino acid)
 - frameshift (more serious – indels in Open Reading Frames)

Transitions (most common)

- Purine to Purine
 - A changed to G
 - G changed to A
- Pyrimidine to Pyrimidine
 - C changed to T
 - T changed to C

Transversions (less common)

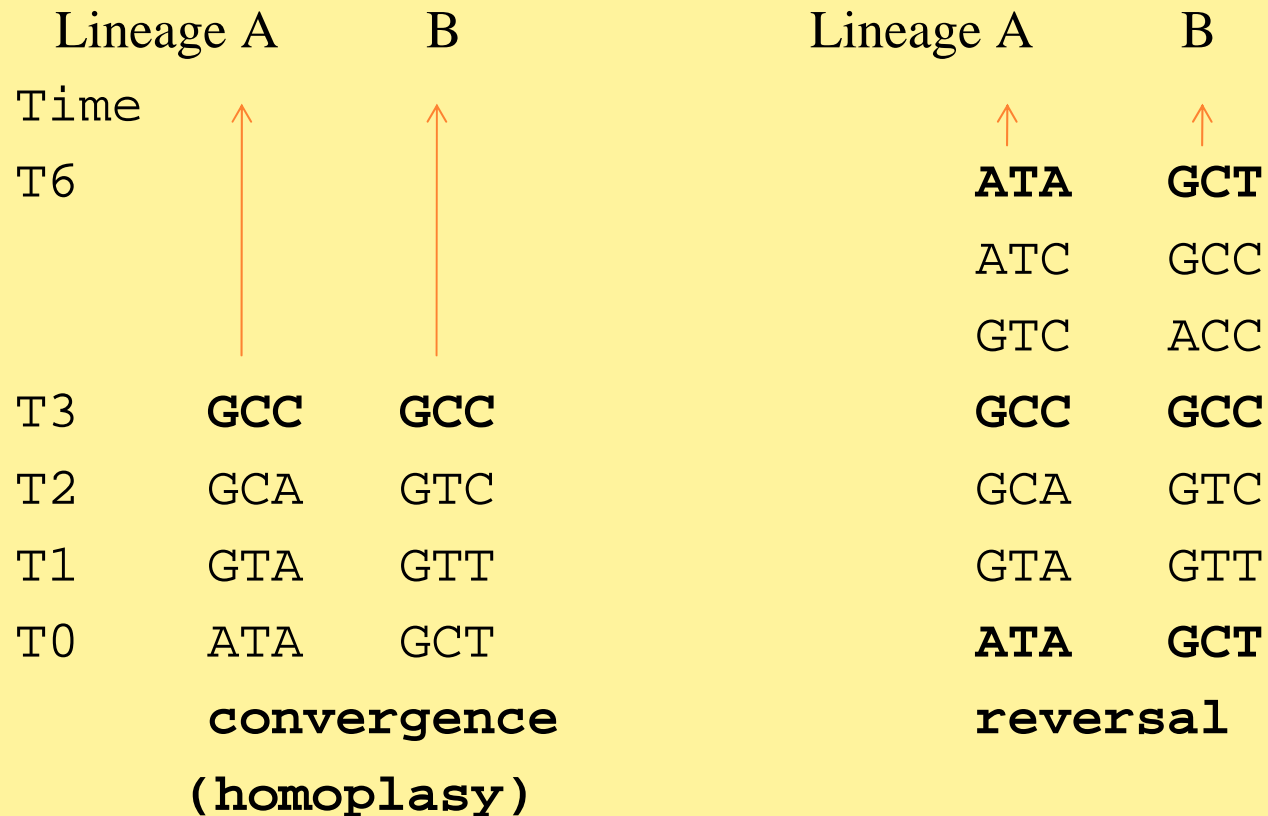
- Purine to Pyrimidine
 - A changed to C or T
 - G changed to C or T
- Pyrimidine to Purine
 - C changed to A or G
 - T changed to A or G

Molecular Character Definitions

See also Skelton & Smith [2002], Sect. 2.3 p33)

- **Uninformative Sites**
 - invariant sites (all bases the same)
 - phylogenetically uninformative
- **Informative Sites**
 - cause some trees to be more parsimonious

Homoplasy and convergence



Adapted from Skelton & Smith (2002)

Multiple Sequence Alignment

- ... to enable production of cladogram
- Clustal W
- Using BioEdit (for Windows)
- Or MacClade (Mac OS X)
- Save alignment ...

BioEdit Sequence Alignment Editor

File Edit Sequence Alignment View World Wide Web Accessory Application RNA Options Window Help

C:\BioEditDev\Bacterial_P_proteins.gb

Courier New 10 B 20 total sequences

Mode: Slide Residues Selection: 0 Position: 29

I D I D G +

E.coli MVKLAFFPRELRLRLTPSQFTFVFQOPQORAGTPQI
 P.mirabali MVKLAFFPRELRLRLTPKHFNFVFQOPQORASSPEV
 H.influenz MLKVVKVYLHNHNSQFLVVKLNFSAELRLRLTP
 P.putida MSQDFPREKRLRLTPRHFKAVFDSPTGKVPQKNL
 B.aphidico MLNYFFKKKSKLLKSTNFQYVFSNPKCNKNTFHI
 C.burnetti MEKGFSSVGRIRTTAEFRRIYAAQRRIIGRYYL
 S.bikinien MLPTE--NRLRRREDFATAVRRGRRAGRPLLVVHR
 S.coelicol MLPTE--NRLRRREDFATAVRRGRRVGRSTLVVHL
 M.luteus MLPRDRRVRTPAEFRHLGRTGTTRAGRRTVVVSV
 M.tubercul MLRAR--NRMRSSADFETTVMKHMRTVRS--DMVVYV
 M.leprae MLSAC--NRMRSSSEFDATVKFGLRAVQSDVVIHV
 B.subtilis MSHLKKR--NRLKKNEDFQKVFKHGTSVANRQFV

Untitled

Courier New 10 B 20 total sequences

Mode: Slide Residues Selection: 0 Position: 3: H.influenz

I D I D G +

E.coli -----MVKLAFFPRE--LRLLT
 P.mirabali -----MVKLAFFPRE--LRLLT
 H.influenz MLKVVKVYLHNHNSQFLVVKLNFSAELRLRLTP
 P.putida -----MSQDFPRE--KRLLT
 B.aphidico -----MLNYFFKKK--SKLLK
 C.burnetti -----MEKGFSSVGRIRTTAEFRRIYAAQRRIIG--RYVLLYYRENEIKH
 S.bikinien -----MLPTE--NRLRRREDFATAVRRGRRAGRPLLVVHRLSGATDPHAP
 S.coelicol -----MLPTE--NRLRRREDFATAVRRGRRVGRS--TLVHRLSGATDPHAP
 M.luteus -----MLPRD--RRVRTPAEFRHLGRTGTTRAGR--TVVSVATDPDQTRST
 M.tubercul -----MLRAR--NRMRSSADFETTVMKHMRTVRS--DMVVYVWRGSGG
 M.leprae -----MLSAC--NRMRSSSEFDATVKFGLRAVQSDVVIHVWRGCMRDETK
 B.subtilis -----MSHLKKR--NRLKKNEDFQKVFKHGTSVANRQFVLTLDQPENDE

ClustalW Options

ClustalW Multiple alignment

Reference:
 Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994)
 CLUSTAL W: improving the sensitivity of progressive multiple
 sequence alignment through sequence weighting, position specific
 gap penalties and weight matrix choice.
 Nucleic Acids Research, submitted, June 1994.

Full Multiple alignment
 Calculate NJ Tree
 FAST algorithm for guide tree
 Bootstrap NJ Tree

Number of bootstraps: 1000

Other Parameters:

Note: enter additional parameters as a single line.

Additional Parameters for ClustalW:
 ****General settings:****
 /QUICKTREE :use FAST algorithm for the alignment guide tree

Run ClustalW View ClustalW Doc Cancel

90 100
 FVVVAKKGVADLDNRALSEALEKLW
 FVVLVRRKVAELDNHQLTEVLGKLW
 LVRESFRLSQHRLPAYDFVFAKNG
 IVIVARKGLGEIENPELHQHFGLKW
 FVVIARKNIVYLMNKKIVNILEYIW
 VVVAKASSVEADNKELYECINKLFT
 PPGSLVVVRALPGAGDADHAQLARD
 PPGSLVVVRALPGAGDADHAQLARD
 PLRDLFVLVQVRALPAAAEADYALL
 ALPSSRHVSSARLEQQLRCGLRRAV
 VVIRALPSSRNVSAAWLAQQLRNL
 TIARKPASQLTYEETKKSLOHLFRK

90 100
 PRIGLTVAKKNVRAHEMRNIKRL
 PRIGLTIAKKNVRAHEMRNIKRL
 PRLGLTVAKKHLKRAHEMRNIKRL
 PRLGLVIGKKSVKLAVQRNRLKRL
 LGLSISRKNIKHAYRRNKIKRL
 LGVVASKRNVKAVWRNRVRRV
 3: H.influenz
 GESAPPTRAGFVVSKA--VGGAVVRNQVKRR
 GESAPRTRAGFVVSKA--VGVAVVRNKVKRR
 SPSAPRPRAGFVVSKA--VGNVAVTRNRVKRR
 GPRVGLIIAKS--VGSVERHRVARR
 APHVGLIIAKT--VGSVERHRVARR
 LRVGLSVSKK--IGNAVMRNRKRL

Cladograms from Molecular Data

- Using PAUP (Phylogenetic Analysis Using Parsimony)
- ... import alignment file
- Generate cladogram
- View Cladogram with TreeView

Useful Websites

- NCBI Genbank

www.ncbi.nlm.nih.gov/Genbank/index.html

- PAUP

<http://paup.csit.fsu.edu/>

- European Molecular Biology Laboratory

www.embl.org

- BioEdit

www.mbio.ncsu.edu/BioEdit/bioedit.html

References & Bibliography

- Skelton, P. & Smith, A (2002). *Cladistics – a practical primer on CD-ROM*. Cambridge University Press, UK. ISBN 0-521-52341 (hardback + CD-ROM)
- Kitching, I. J. *et al.* (1998) *Cladistics - the theory and practice of parsimony analysis*. Systematics Association Publication No. 11. Oxford University Press, UK. ISBN 0-19-850138 (paperback)
- Gibas, C. & Jambeck, P. (2001). *Developing bioinformatics computer skills*. O'Reilly, USA. Chapter 8, p191-214 ISBN 1-56592-664-1 (paperback)
- Page, R.D.M. & Holmes, E.C. (1998). *Molecular Evolution – A Phylogenetic Approach*, Blackwell Publishing, Malden, MA, USA. ISBN 978-0-86542-889-8 (softback)