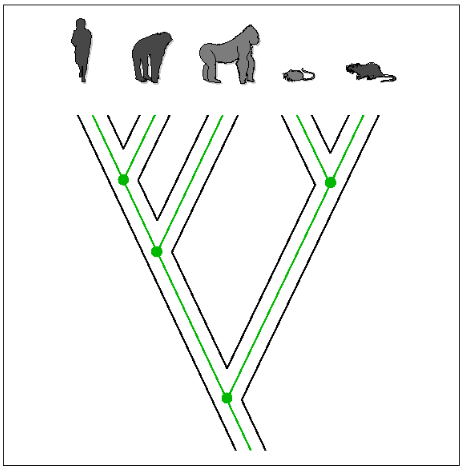
CIBIV MFPL
Center for Integrative Bioinformatics Vienna

One Step Mutation (OSM) matrices

joint work with

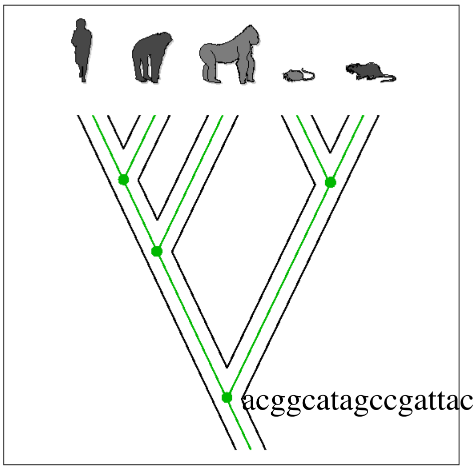
Sequence Evolution

CIBIV MFPL
Center for Integrative Bioinformatics Vienna



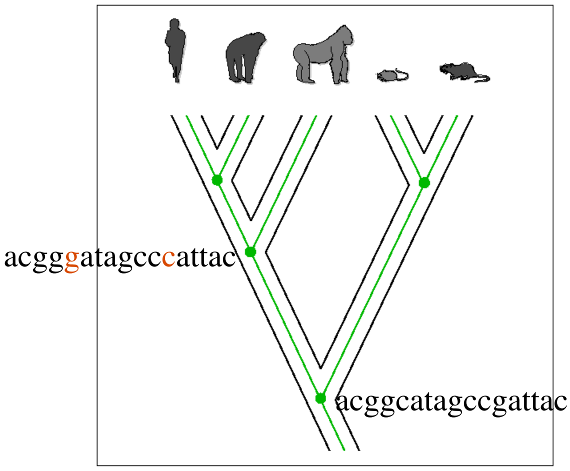
Sequence Evolution

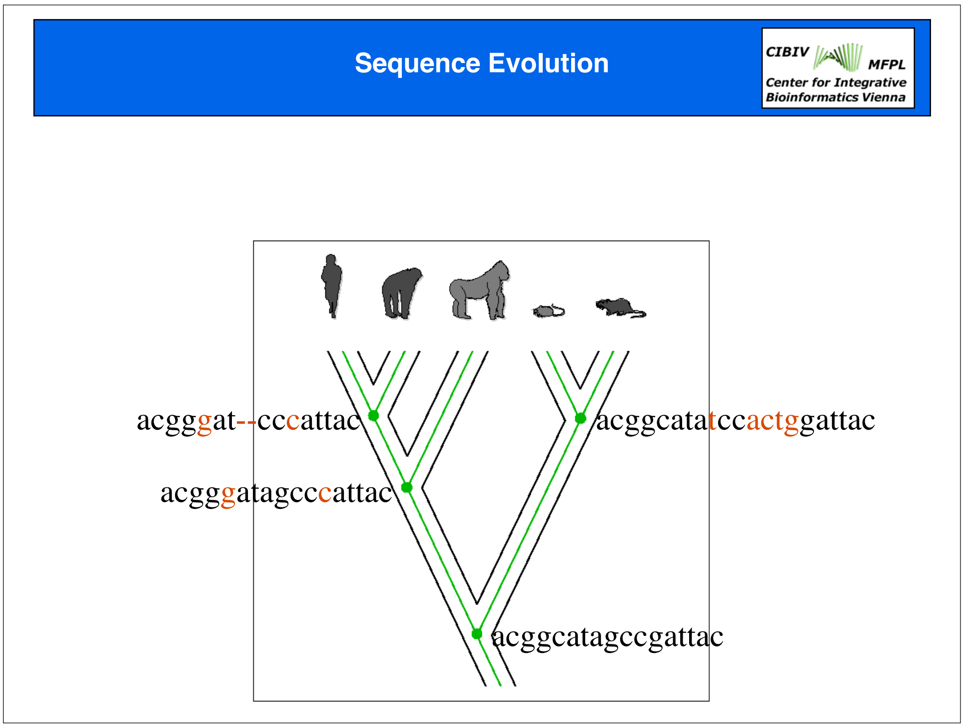
CIBIV MFPL Center for Integrative Bioinformatics Vienna

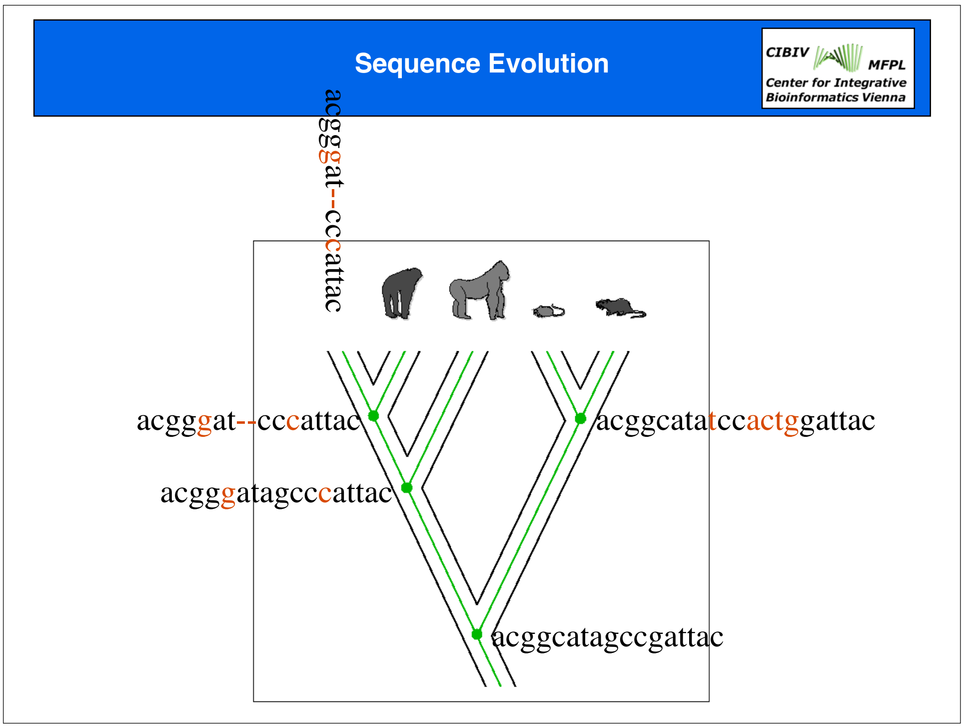


Sequence Evolution

CIBIV MFPL Center for Integrative Bioinformatics Vienna

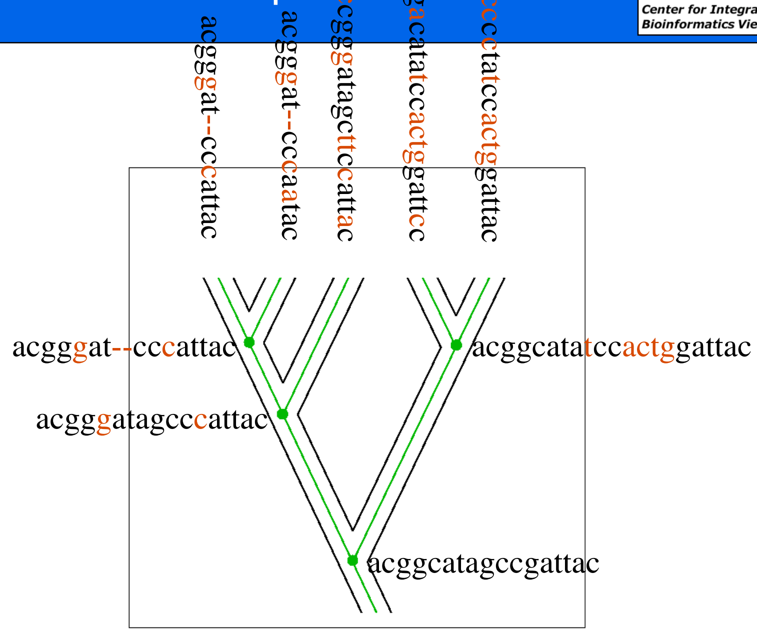



Sequence Evolution
CIBIV
MFPL
Center for Integrative Bioinformatics Vienna
acgggat--cccattac
acgggatagcccattac
acggcatatccactggattac
acggcatagccgattac


Sequence Evolution
CIBIV
MFPL
Center for Integrative Bioinformatics Vienna
acgggat--cccattac
acgggat--cccattac
acgggatagcccattac
acggcatatccactggattac
acggcatagccgattac

Sequence Evolution

CIBIV MFPL Center for Integrative Bioinformatics Vienna



Multiple Sequence Alignment (MSA)

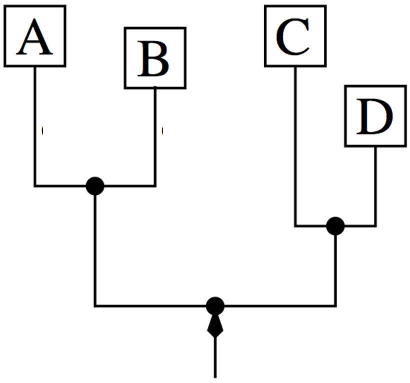
CIBIV MFPL Center for Integrative Bioinformatics Vienna

seq 1	a	g	c	t	t	a	c	c	t	g	t	t	a	c	t
seq 2	c	g	t	a	a	a	t	t	t	c	c	c	g	a	t
seq 3	c	g	c	a	a	g	t	t	t	c	c	c	g	a	t
seq 4	c	a	c	t	t	a	t	t	a	g	t	c	a	a	c

Alignment column or alignment pattern

Example: Binary Alphabet {R, Y}

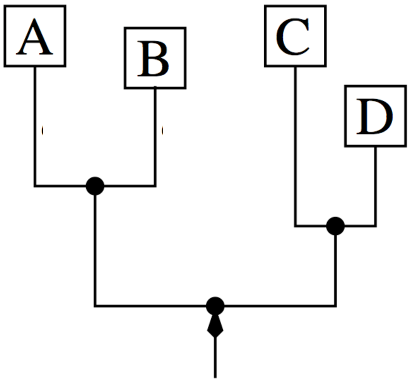
CIBIV MFPL Center for Integrative Bioinformatics Vienna



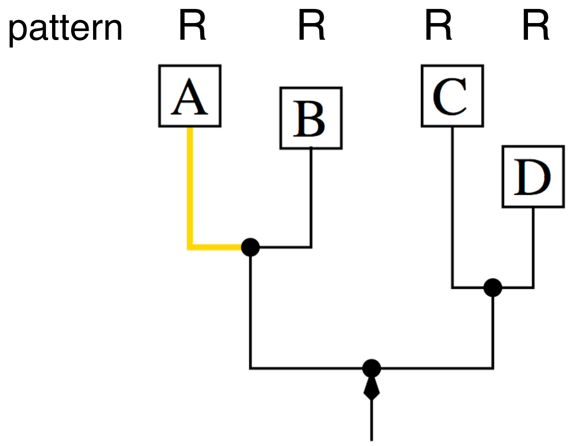
Binary Alphabet {R, Y}

CIBIV MFPL Center for Integrative Bioinformatics Vienna

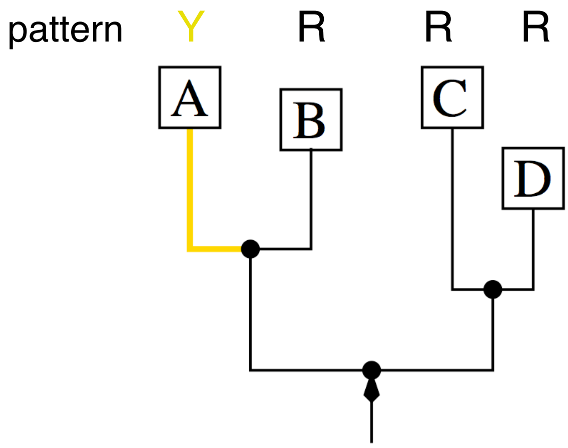
pattern R R R R



Binary Alphabet {R, Y}



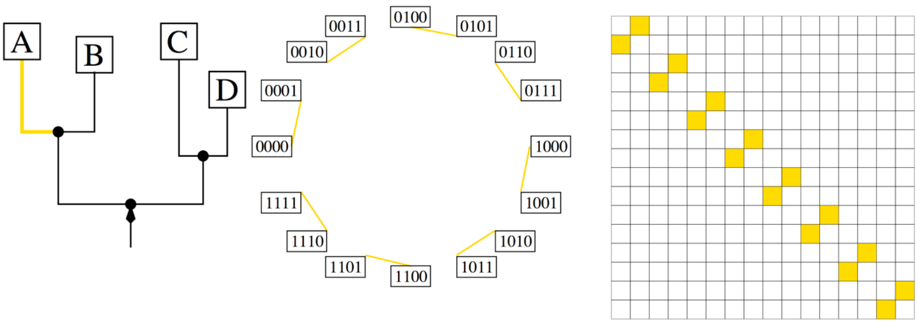
Binary Alphabet {R, Y}



Binary Alphabet {R, Y}

CIBIV MFPL Center for Integrative Bioinformatics Vienna

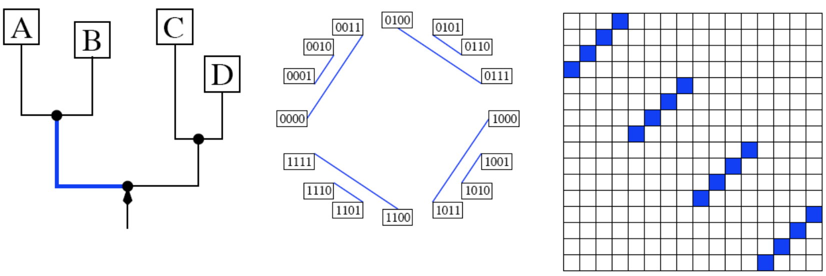
permutation matrix σ_A



Binary Alphabet {R, Y}

CIBIV MFPL Center for Integrative Bioinformatics Vienna

permutation matrix σ_{AB}

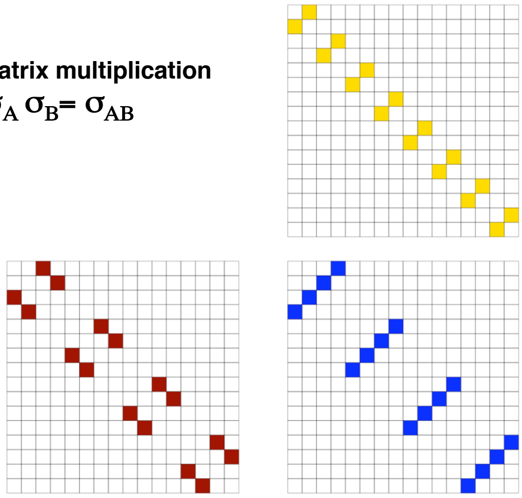


Internal branches

CIBIV MFPL
Center for Integrative Bioinformatics Vienna

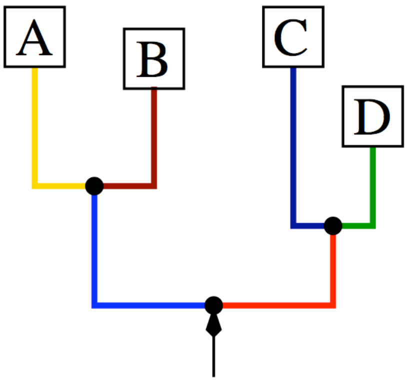
matrix multiplication

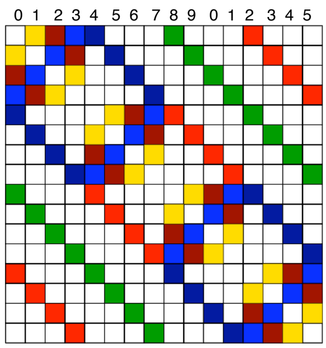
$\sigma_A \sigma_B = \sigma_{AB}$



One Step Mutation Matrix

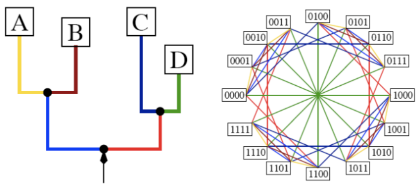
CIBIV MFPL
Center for Integrative Bioinformatics Vienna

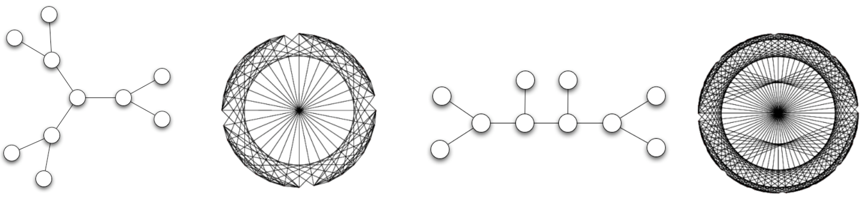




Examples of OSM-Graphs

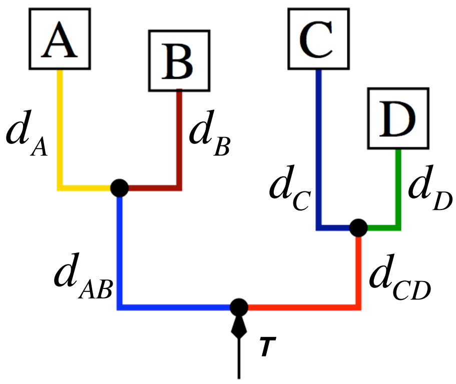
CIBIV MFPL
Center for Integrative Bioinformatics Vienna





Branch Lengths:

CIBIV MFPL
Center for Integrative Bioinformatics Vienna



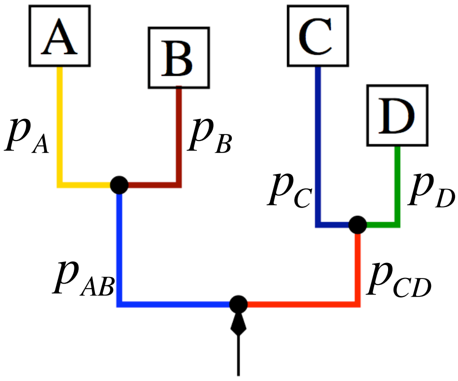
Total branch length

$$\Delta = d_A + d_B + d_C + d_D + d_{AB} + d_{CD}$$

relative edge length

$$p_{edge} = \frac{d_{edge}}{\Delta}$$

Some Formalisms:



relative branch lengths

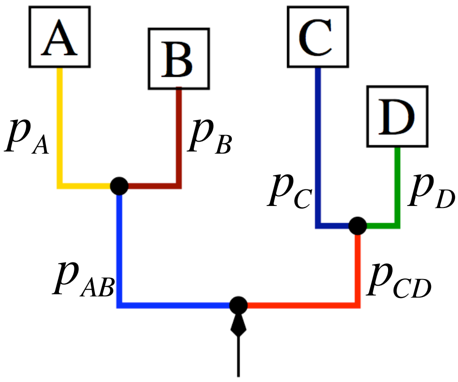
$$p_A + p_B + p_C + p_D + p_{AB} + p_{CD} = 1$$

used to assign mutation probabilities

$$p_{edge} \cdot \sigma_{edge}$$

general permutation matrix

Constructing the OSM:



$$\mathbf{M}_T = \sum_{edge} p_{edge} \cdot \sigma_{edge}$$

Many Substitutions:

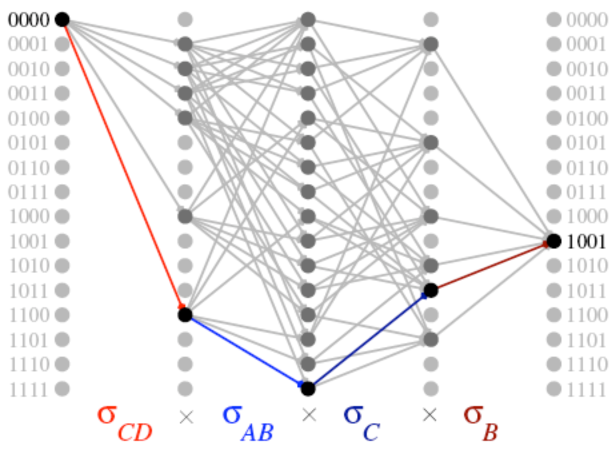
$$\mathbf{M}_T = \sum_{edge} p_{edge} \cdot \sigma_{edge}$$

One substitution

$$\mathbf{M}_T^k = \left(\sum_{edge} p_{edge} \cdot \sigma_{edge} \right)^k$$

***k* substitutions**

Many Substitutions: Random walk



Maximum Parsimony (MP)

CIBIV MFPL Center for Integrative Bioinformatics Vienna

$$\underset{k}{\mathrm{Min}}\left\{\mathbf{M}_T^k(i,j) > 0 \,\middle|\, k \in N\right\}$$

describes the minimal number of mutations to move from pattern *i* to *j*

MP: For a tree *T* and pattern *j* compute:

$$\underset{k}{\mathrm{Min}}\left\{\mathbf{M}_T^k(R...R,j) > 0 \text{ or } \mathbf{M}_T^k(Y...Y,j) > 0 \,\middle|\, k \in N\right\}$$

Maximum Likelihood

CIBIV MFPL Center for Integrative Bioinformatics Vienna

We assume that the number of substitutions is Poisson distributed with parameter Δ. Then we compute, the expected OSM as

$$\overline{\mathbf{M}_T} = \sum_{k=0}^{\infty} \frac{\exp(-\Delta)\Delta^k (\mathbf{M}_T)^k}{k!}$$

$$\overline{\mathbf{M}_T} = \exp(-\Delta) \cdot \exp(\Delta \cdot \mathbf{M}_T)$$

$$\overline{\mathbf{M}_T} = \exp(-\Delta) \cdot \mathbf{H}_{2^n} \cdot \exp(\Delta \cdot \mathbf{D}_T) \cdot \mathbf{H}_{2^n}$$

where $\mathbf{D}_T = \mathbf{H}_{2^n} \cdot \mathbf{M}_T \cdot \mathbf{H}_{2^n}$

and $\mathbf{H}_{2^n} = \underbrace{\mathbf{H}_2 \otimes \mathbf{H}_2 \otimes \ldots \otimes \mathbf{H}_2}_{n \text{ times}}$ $\qquad \mathbf{H}_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$

Maximum Likelihood

CIBIV MFPL Center for Integrative Bioinformatics Vienna

$$\overline{\mathbf{M}_T} = \exp(-\Delta) \cdot \mathbf{H}_{2^n} \cdot \exp(\Delta \cdot \mathbf{D}_T) \cdot \mathbf{H}_{2^n}$$

The likelihood of a tree *T* with branch length Δ, given an alignment of length *L* is then

$$\Pr(T,\Delta) = \prod_{i=1}^{L} \overline{\mathbf{M}_T}(\{R...R,Y...Y\}, pattern(i))$$

Another View at the Mutations

CIBIV MFPL Center for Integrative Bioinformatics Vienna

$$\overline{\mathbf{M}_T} = \exp(-\Delta) \cdot \mathbf{H}_{2^n} \cdot \exp(\Delta \cdot \mathbf{D}_T) \cdot \mathbf{H}_{2^n}$$

From the above formula, we can **analytically** compute the posterior probability of the number of mutations that have occurred on a fixed tree.

$$\Pr(k \text{ mutations} \mid \text{pattern}) = \frac{\exp(-\Delta)\Delta^k}{k!} \frac{(\mathbf{M}_T(R,\ldots,R,\text{pattern}))^k}{\overline{\mathbf{M}_T}(R\ldots,R,\text{pattern})}$$

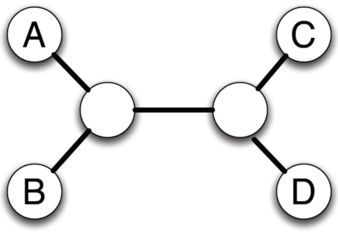
similar work by Rasmus Nielsen, John Huelsenbeck, Jonathan Bollback (2002, 2003, 2005)

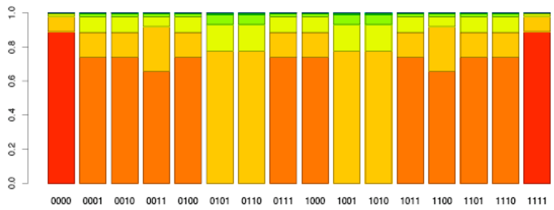
Posterior probabilities:clock-like tree

CIBIV MFPL Center for Integrative Bioinformatics Vienna

$$\mathrm{ppd}[k \mid \mathbf{a}] = \frac{\exp[-\Delta]\Delta^k\left(\pi_0 \mathbf{M}_T^k(\mathbf{0},\mathbf{a}) + \pi_1 \mathbf{M}_T^k(\mathbf{1},\mathbf{a})\right)}{\pi_0 \overline{\mathbf{M}_T}(\mathbf{0},\mathbf{a}) + \pi_1 \overline{\mathbf{M}_T}(\mathbf{1},\mathbf{a})}$$

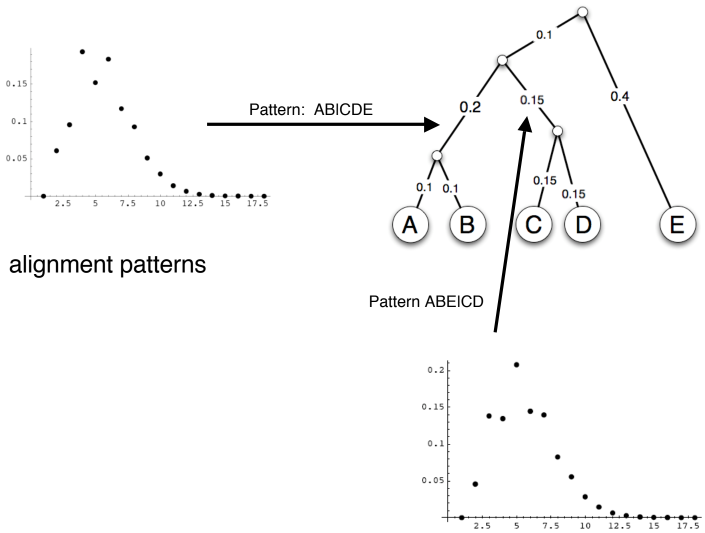
Δ=1.0





Posterior probabilities: five Taxa Tree

CIBIV MFPL Center for Integrative Bioinformatics Vienna



Summary and Outlook

CIBIV MFPL Center for Integrative Bioinformatics Vienna

Developed an evolutionary model that describes the action of a single substitution on an alignment pattern.

This leads to a tree-topology mediated random walk on the space of words of length *n*.

Maximum Parsimony and Maximum Likelihood are “extreme” cases within this framework.

Practical Aspect: Analytical formula for the posterior probabilities of the number of substitutions for a pattern.

Open Questions:

•Connection between OSM and Hadamard transform (Hendy, Penny 1989) and its generalization, the Fourier calculus on evolutionary trees (Szekely, Steel, Erdös 1993).

•Other type of substitution distributions?

•Computational issues

The real stuff

CIBIV MFPL Center for Integrative Bioinformatics Vienna

```
..GUCAUAGAGGGUGAGAAUCCCGUG..
..GCCGGAGAGGGUGACAGCCCCAUC..
..CCCGUGGACGGUGUGAGGCCGGUA..
..GUGAUACAGGGUGACAACCCCGUA..
..ACCAGAGAAGGUGAAAGUCCUGUA..
..GCGAUACAGGGUGACAGCCCCGUA..
```

The real stuff

CIBIV MFPL Center for Integrative Bioinformatics Vienna

```
..GUCAUAGAGGGUGAGAAUCCCGUG..
..GCCGGAGAGGGUGACAGCCCCAUC..
..CCCGUGGACGGUGUGAGGCCGGUA..
..GUGAUACAGGGUGACAACCCCGUA..
..ACCAGAGAAGGUGAAAGUCCUGUA..
..GCGAUACAGGGUGACAGCCCCGUA..
```

Observed pattern count

$$\mathbf{O}\left(d_1,\ldots,d_{4^n}\right)$$

The real stuff

CIBIV MFPL Center for Integrative Bioinformatics Vienna

```
..GUCAUAGAGGGUGAGAAUCCCGUG..
..GCCGGAGAGGGUGACAGCCCCAUC..
..CCCGUGGACGGUGUGAGGCCGGUA..
..GUGAUACAGGGUGACAACCCCGUA..
..ACCAGAGAAGGUGAAAGUCCUGUA..
..GCGAUACAGGGUGACAGCCCCGUA..
```

Observed pattern count

$$\mathbf{O}\left(d_1,\ldots,d_{4^n}\right)$$

Maximum likelihood etc.

The real stuff

CIBIV MFPL Center for Integrative Bioinformatics Vienna

```
..GUCAUAGAGGGUGAGAAUCCCGUG..
..GCCGGAGAGGGUGACAGCCCCAUC..
..CCCGUGGACGGUGUGAGGCCGGUA..
..GUGAUACAGGGUGACAACCCCGUA..
..ACCAGAGAAGGUGAAAGUCCUGUA..
..GCGAUACAGGGUGACAGCCCCGUA..
```

→ Observed pattern count

$$\mathrm{O}\left(d_1,\ldots,d_{4^n}\right)$$

↓ Maximum likelihood etc.

$$\mathrm{E}\left(p_1,\ldots,p_{4^n}\right)$$

$$\hat{T}$$

The real stuff

CIBIV MFPL Center for Integrative Bioinformatics Vienna

```
..GUCAUAGAGGGUGAGAAUCCCGUG..
..GCCGGAGAGGGUGACAGCCCCAUC..
..CCCGUGGACGGUGUGAGGCCGGUA..
..GUGAUACAGGGUGACAACCCCGUA..
..ACCAGAGAAGGUGAAAGUCCUGUA..
..GCGAUACAGGGUGACAGCCCCGUA..
```

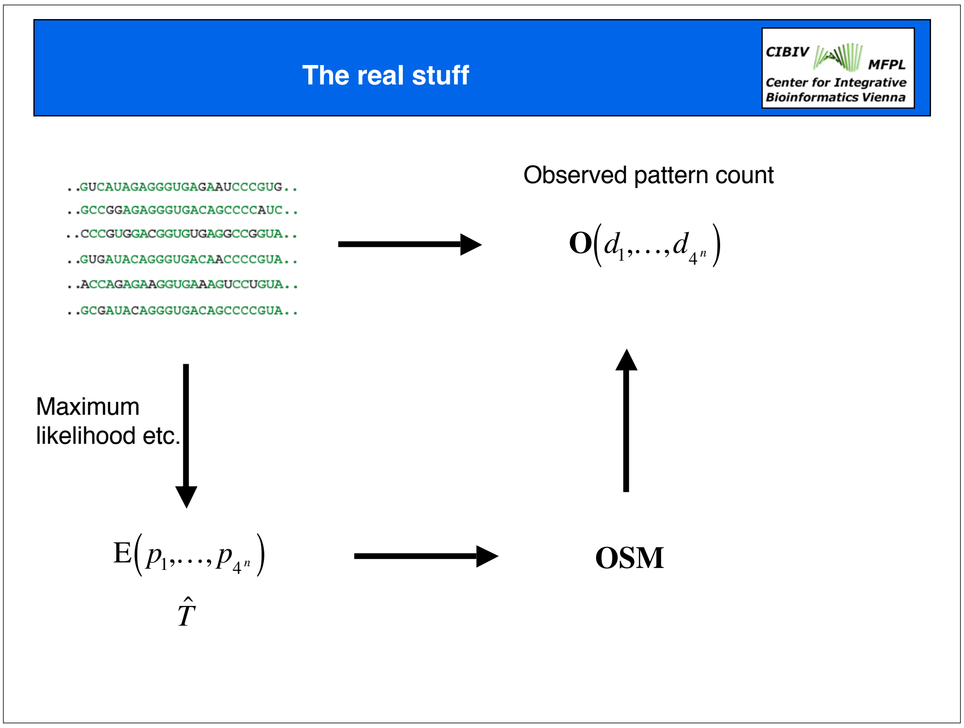
→ Observed pattern count

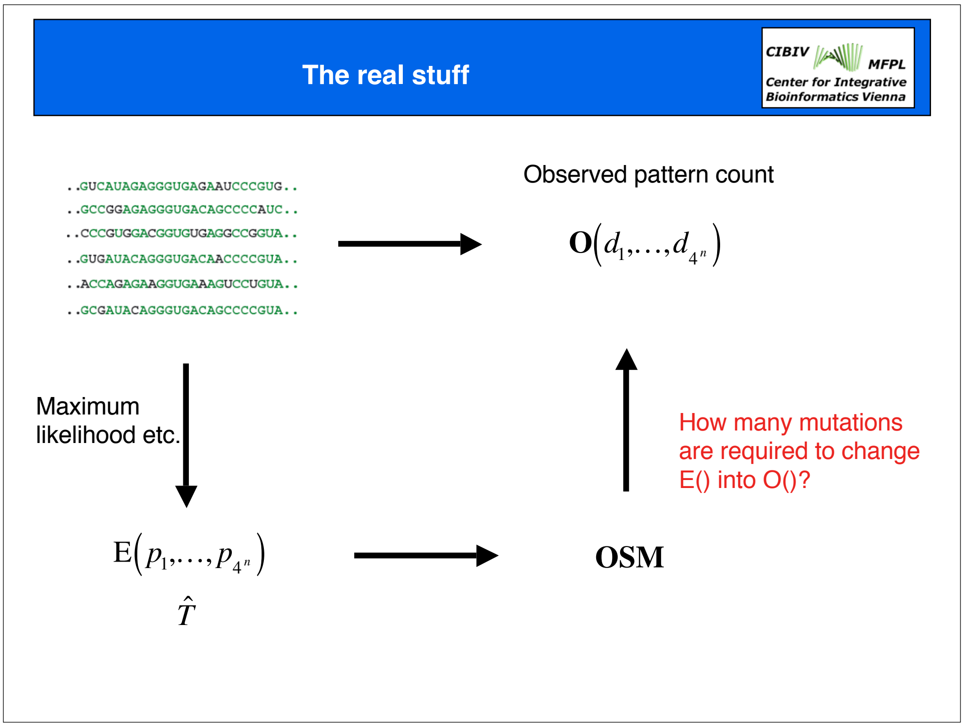
$$\mathrm{O}\left(d_1,\ldots,d_{4^n}\right)$$

↓ Maximum likelihood etc.

$$\mathrm{E}\left(p_1,\ldots,p_{4^n}\right) \rightarrow \textbf{OSM}$$

$$\hat{T}$$


The real stuff
CIBIV
MFPL
Center for Integrative Bioinformatics Vienna
..GUCAUAGAGGGUGAGAAUCCCGUG..
..GCCGGAGAGGGUGACAGCCCCAUC..
..CCCGUGGACGGUGUGAGGCCGGUA..
..GUGAUACAGGGUGACAACCCCGUA..
..ACCAGAGAAGGUGAAAGUCCUGUA..
..GCGAUACAGGGUGACAGCCCCGUA..
Observed pattern count
$O(d_1,\ldots,d_{4^n})$
Maximum likelihood etc.
$E(p_1,\ldots,p_{4^n})$
$\hat{T}$
OSM


The real stuff
CIBIV
MFPL
Center for Integrative Bioinformatics Vienna
..GUCAUAGAGGGUGAGAAUCCCGUG..
..GCCGGAGAGGGUGACAGCCCCAUC..
..CCCGUGGACGGUGUGAGGCCGGUA..
..GUGAUACAGGGUGACAACCCCGUA..
..ACCAGAGAAGGUGAAAGUCCUGUA..
..GCGAUACAGGGUGACAGCCCCGUA..
Observed pattern count
$O(d_1,\ldots,d_{4^n})$
Maximum likelihood etc.
How many mutations are required to change E() into O()?
$E(p_1,\ldots,p_{4^n})$
$\hat{T}$
OSM