

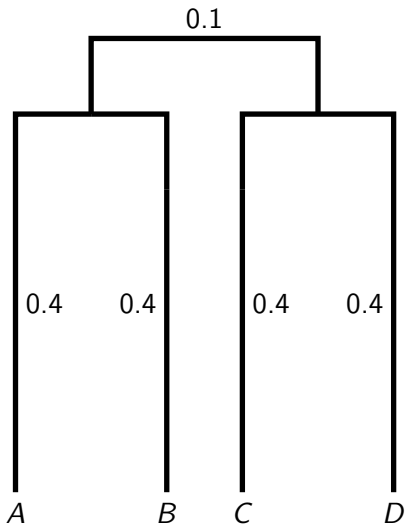
Model Misspecification due to Site Specific Rate Heterogeneity: how is tree inference affected?

Stephen Crotty

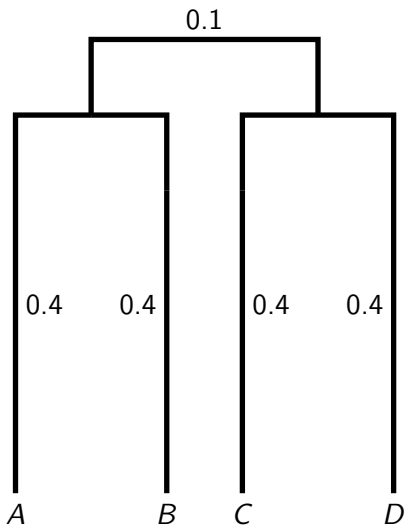
School of Mathematical Sciences, University of Adelaide

October, 2013

What is Site Specific Rate Heterogeneity (SSRH)?



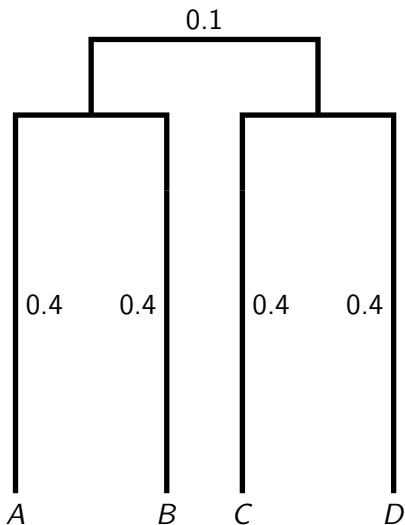
What is Site Specific Rate Heterogeneity (SSRH)?



The model contains 3 site types:

- Invariable sites

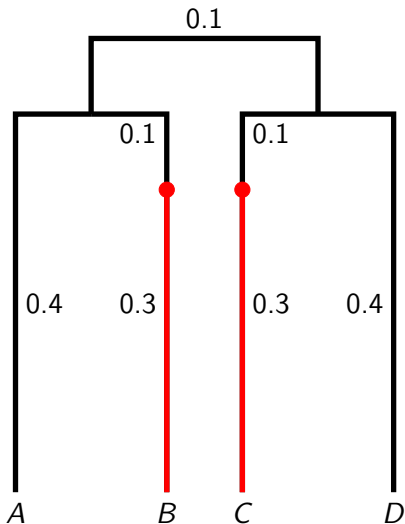
What is Site Specific Rate Heterogeneity (SSRH)?



The model contains 3 site types:

- Invariable sites
- Variable sites

What is Site Specific Rate Heterogeneity (SSRH)?



The model contains 3 site types:

- Invariable sites
- Variable sites
- Switching sites

Why should we care about SSRH?

Why should we care about SSRH?



Why should we care about SSRH?



Tasmanian Pygmy Possum



Tasmanian Native Hen

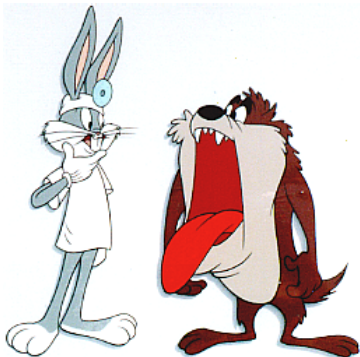
Why should we care about SSRH?



Tasmanian Devil

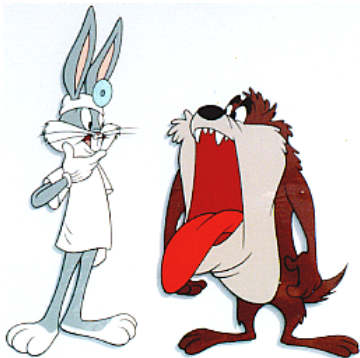
Why should we care about SSRH?

What's up Doc?



Why should we care about SSRH?

What's up Doc?

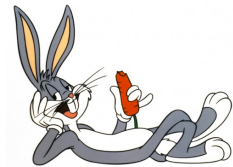


Devil Facial Tumour Syndrome

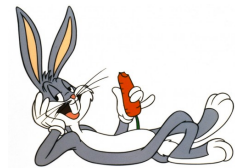
Why should we care about SSRH?



Why should we care about SSRH?



Why should we care about SSRH?



Experimental Procedure

- 1 Data was simulated using the program LineageSpecificSeqgen¹

¹Source: L. Shavit Grievink, D. Penny, M. D. Hendy, and B. R. Holland.
BMC Evolutionary Biology, 8:317, 2008.

²<http://evolution.genetics.washington.edu/phylip/>

Experimental Procedure

- ① Data was simulated using the program LineageSpecificSeqgen¹
- ② The Phylip² software package was used to perform tree inference using the maximum parsimony (MP), neighbour joining (NJ) and maximum likelihood (ML) methods.

¹Source: L. Shavit Grievink, D. Penny, M. D. Hendy, and B. R. Holland.
BMC Evolutionary Biology, 8:317, 2008.

²<http://evolution.genetics.washington.edu/phylip/>

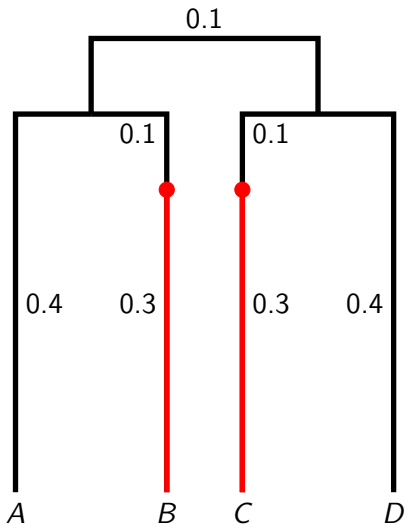
Experimental Procedure

- ① Data was simulated using the program LineageSpecificSeqgen¹
- ② The Phylip² software package was used to perform tree inference using the maximum parsimony (MP), neighbour joining (NJ) and maximum likelihood (ML) methods.
- ③ A theoretical analysis of each method was carried out in an effort to understand their performance.

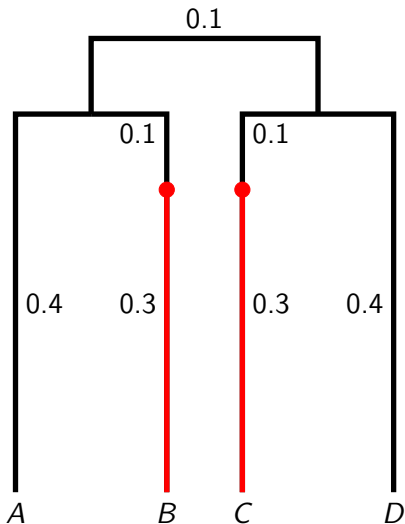
¹Source: L. Shavit Grievink, D. Penny, M. D. Hendy, and B. R. Holland.
BMC Evolutionary Biology, 8:317, 2008.

²<http://evolution.genetics.washington.edu/phylip/>

Simulation Parameters



Simulation Parameters

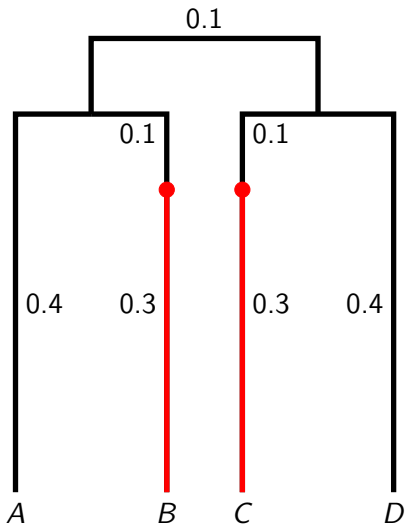


$$p_{inv} = 80\%$$

$$p_{var} = 20\%$$

$$p_{switch} = 0, 1, 2, \dots, 100\%$$

Simulation Parameters



$$p_{inv} = 80\%$$

$$p_{var} = 20\%$$

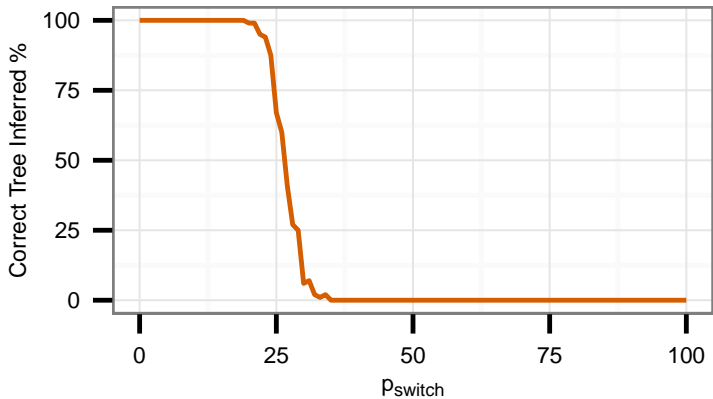
$$p_{switch} = 0, 1, 2, \dots, 100\%$$

100000 base pairs

Jukes Cantor substitution model

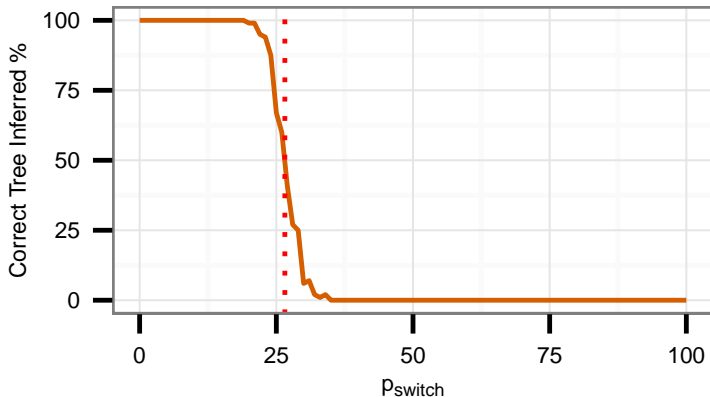
100 replications

Maximum Parsimony

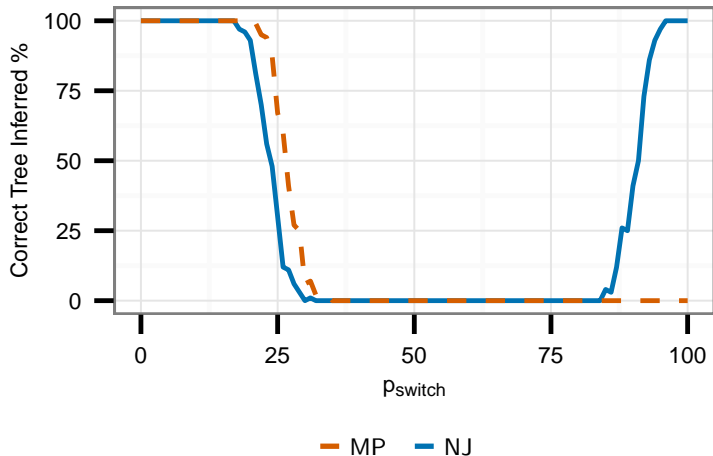


Maximum Parsimony

- Site pattern analysis predicts the asymptotic failure point of MP to be 26.56%.



Neighbour Joining



The neighbour joining algorithm

r = number of taxa.

D_{ij} = JC distance between taxa i and j .

$$Q_{ij} = (r - 2)D_{ij} - \sum_{k=1}^r D_{ik} - \sum_{k=1}^r D_{jk}$$

Q is the matrix used by the NJ algorithm: the pair of taxa with the smallest Q_{ij} are joined together and the process is repeated.

The Q matrix for a 4-taxa tree

$$\begin{aligned}Q_{AB} &= (4 - 2)D_{AB} - \sum_{k \in \{B, C, D\}} D_{Ak} - \sum_{k \in \{A, C, D\}} D_{Bk} \\ &= -(D_{AC} + D_{AD} + D_{BC} + D_{BD})\end{aligned}$$

Similarly,

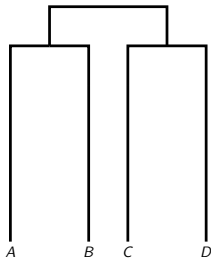
$$Q_{AD} = -(D_{AB} + D_{AC} + D_{BD} + D_{CD})$$

and,

$$Q_{AC} = -(D_{AB} + D_{AD} + D_{BC} + D_{CD})$$

Digression - what tree might we infer?

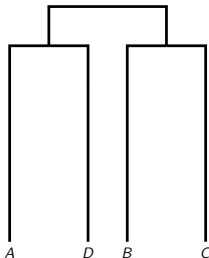
AB|CD



$$\min(Q_{AB}, Q_{AD}, Q_{AC}) = Q_{AB}$$

⇒ ✓

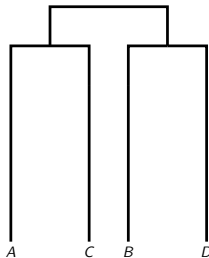
AD|BC



$$\min(Q_{AB}, Q_{AD}, Q_{AC}) = Q_{AD}$$

⇒ ✗

AC|BD

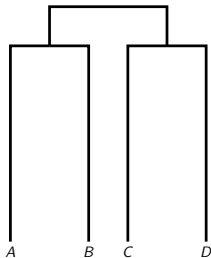


$$\min(Q_{AB}, Q_{AD}, Q_{AC}) = Q_{AC}$$

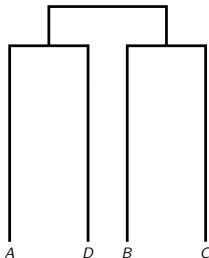
⇒ ✗

Digression - what tree might we infer?

AB|CD



AD|BC



$$Q_{AB} < Q_{AD}$$

⇒ ✓

$$Q_{AD} < Q_{AB}$$

⇒ ✗

The Q matrix for a 4-taxa tree

The correct tree (AB|CD) will be inferred given the condition:

$$\begin{aligned} & Q_{AB} < Q_{AD} \\ \implies & 0 < Q_{AD} - Q_{AB} \\ \implies & 0 < D_{AD} + D_{BC} - D_{AB} - D_{CD} \end{aligned}$$

We now define

$$C = D_{AD} + D_{BC} - D_{AB} - D_{CD}$$

so that the correct tree will be inferred when $C > 0$.

Deriving the expected value of C

T = the tree topology

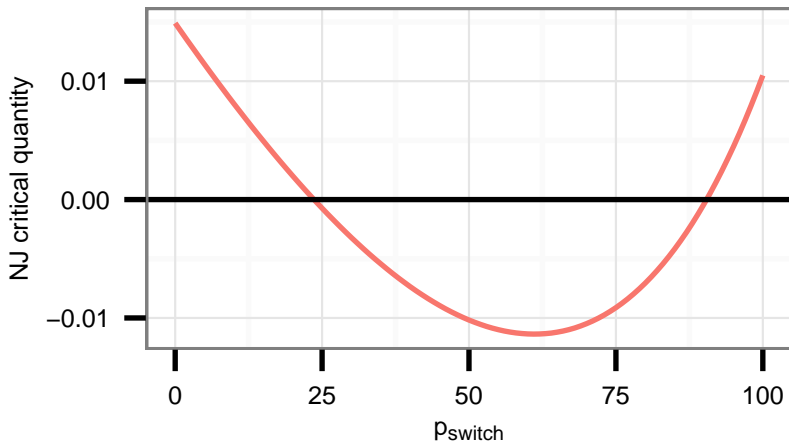
P_{ij} = the proportion of differing sites between taxa i and j

$$E[P_{ij}] = f(p_{\text{switch}}, T)$$

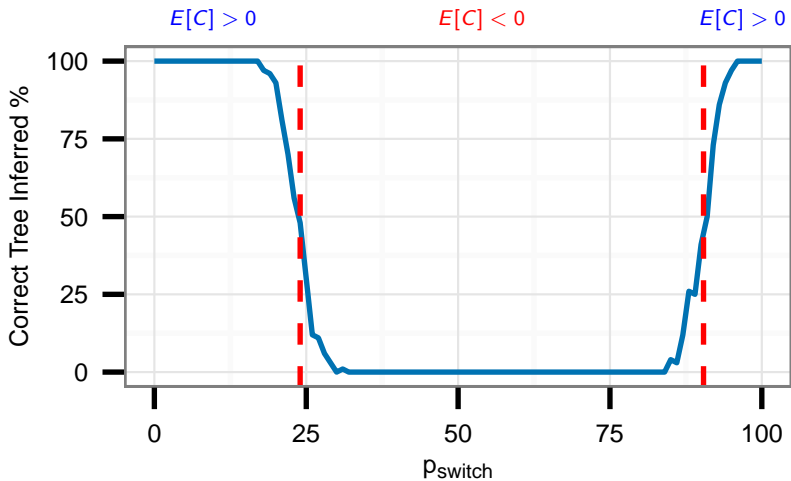
$$E[D_{ij}] = -\frac{3}{4} \ln\left(1 - \frac{4}{3} E[P_{ij}]\right)$$

$$E[C] = E[D_{AD}] + E[D_{BC}] - E[D_{AB}] - E[D_{CD}]$$

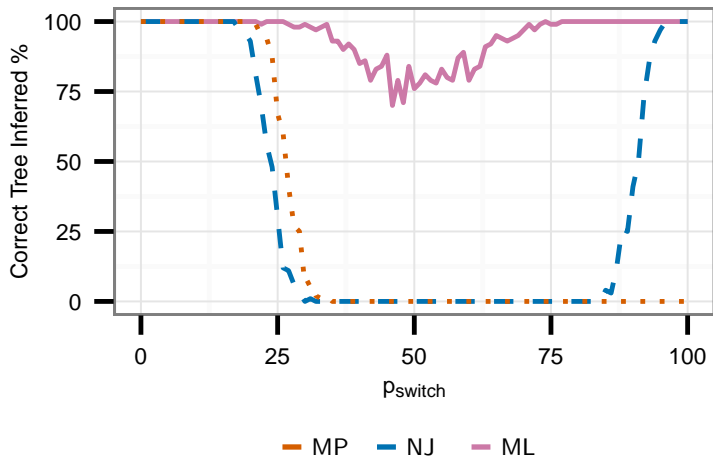
Expected value of C



Neighbour Joining



Maximum Likelihood



Why is this important?

- Traditional methods of phylogenetic inference may be compromised by SSRH.

Why is this important?

- Traditional methods of phylogenetic inference may be compromised by SSRH.
- Diagnostic tools need to be developed to help identify the presence and extent of SSRH in sequence data.

Why is this important?

- Traditional methods of phylogenetic inference may be compromised by SSRH.
- Diagnostic tools need to be developed to help identify the presence and extent of SSRH in sequence data.
- Data driven model checking will be the focus of my PhD going forward.

Acknowledgements

I would like to thank my supervisory team for their input and guidance:

- Prof. Nigel Bean - University of Adelaide
- Dr Lars Jermiin - CSIRO
- Dr Barbara Holland - University of Tasmania
- Dr Jono Tuke - University of Adelaide

That's all folks!



Questions?

Why is the AC|BD tree never inferred? I'm glad you asked!

$$Q_{AB} - Q_{AC} = D_{AB} + D_{CD} - D_{AC} - D_{BD}$$

Why is the AC|BD tree never inferred? I'm glad you asked!

$$Q_{AB} - Q_{AC} = D_{AB} + D_{CD} - D_{AC} - D_{BD}$$

$=$

Why is the AC|BD tree never inferred? I'm glad you asked!

$$Q_{AB} - Q_{AC} = D_{AB} + D_{CD} - D_{AC} - D_{BD}$$

$$=$$

The diagram illustrates the four terms in the equation using phylogenetic trees with four taxa: A, B, C, and D. The trees are arranged horizontally and separated by plus and minus signs. The first tree, representing D_{AB} , shows a clade of (A, B, C) with D as the outgroup. The second tree, representing D_{CD} , shows a clade of (A, B) with (C, D) as the outgroup. The third tree, representing $-D_{AC}$, shows a clade of (A, B, C) with D as the outgroup, but the branch leading to C is highlighted in red. The fourth tree, representing $-D_{BD}$, shows a clade of (A, B) with (C, D) as the outgroup, but the branch leading to D is highlighted in red.

Why is the AC|BD tree never inferred? I'm glad you asked!

$$Q_{AB} - Q_{AC} = D_{AB} + D_{CD} - D_{AC} - D_{BD}$$

$$= \begin{array}{c} \text{---} \\ | \quad | \\ A \quad B \quad C \quad D \end{array} + \begin{array}{c} \text{---} \\ | \quad | \\ A \quad B \quad C \quad D \end{array} - \begin{array}{c} \text{---} \\ | \quad | \\ A \quad B \quad C \quad D \end{array} - \begin{array}{c} \text{---} \\ | \quad | \\ A \quad B \quad C \quad D \end{array}$$

The diagram illustrates the decomposition of the difference in likelihoods between two tree topologies. The first two trees, representing D_{AB} and D_{CD} , have a horizontal line connecting the branches leading to nodes (A, B) and (C, D) respectively. The last two trees, representing $-D_{AC}$ and $-D_{BD}$, have a horizontal line connecting the branches leading to nodes (A, C) and (B, D) respectively. The lines in the last two trees are highlighted in red.

Why is the AC|BD tree never inferred? I'm glad you asked!

$$\begin{aligned}
 Q_{AB} - Q_{AC} &= D_{AB} + D_{CD} - D_{AC} - D_{BD} \\
 &= \begin{array}{c} \text{[Tree 1: Root splits into (A,B) and (C,D)]} \\ + \\ \text{[Tree 2: Root splits into (A,B) and (C,D)]} \\ - \\ \text{[Tree 3: Root splits into (A,C) and (B,D)]} \\ - \\ \text{[Tree 4: Root splits into (A,C) and (B,D)]} \end{array} \\
 &= - \begin{array}{c} \text{[Red bracket over (A,B)]} \\ - \\ \text{[Red bracket over (C,D)]} \end{array} \\
 &\leq 0
 \end{aligned}$$

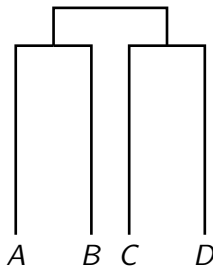
How was the MP crash point derived? I'm glad you asked!

Site Pattern	Correct Tree	Incorrect Tree
xxxx	0	0
xxxy	1	1
xxyx	1	1
xyxx	1	1
yxxx	1	1
xyyy	1	2
xyyx	2	1
xyxy	2	2
xxyz	2	2
xyzx	2	2
xyxz	2	2
yxxz	2	2
yxzx	2	2
yzxx	2	2
wxyz	3	3

How was the MP crash point derived? I'm glad you asked!

Site Pattern	Correct Tree	Incorrect Tree
xxxx	0	0
xxxy	1	1
xxyx	1	1
xyxx	1	1
yxxx	1	1
xxyy	1	2
xyyx	2	1
xyxy	2	2
xxyz	2	2
xyzx	2	2
xyxz	2	2
yxxz	2	2
yxzx	2	2
yzxx	2	2
wxyz	3	3

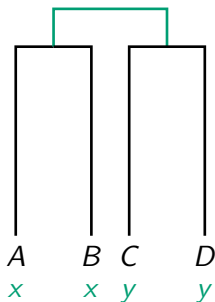
Consider site pattern *xyyy*:



How was the MP crash point derived? I'm glad you asked!

Site Pattern	Correct Tree	Incorrect Tree
xxxx	0	0
xxxy	1	1
xxyx	1	1
xyxx	1	1
yxxx	1	1
xxyy	1	2
xyyx	2	1
xyxy	2	2
xxyz	2	2
xyzx	2	2
xyxz	2	2
yxxz	2	2
yxzx	2	2
yzxx	2	2
wxyz	3	3

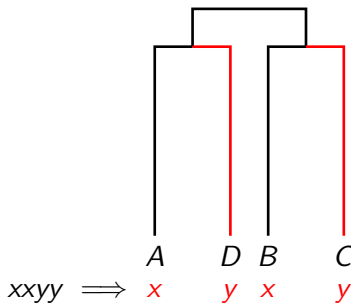
Consider site pattern *xyyy*:



How was the MP crash point derived? I'm glad you asked!

Site Pattern	Correct Tree	Incorrect Tree
xxxx	0	0
xxxy	1	1
xxyx	1	1
xyxx	1	1
yxxx	1	1
xxyy	1	2
xyyx	2	1
xyxy	2	2
xxyz	2	2
xyzx	2	2
xyxz	2	2
yxxz	2	2
yxzx	2	2
yzxx	2	2
wxyz	3	3

Consider site pattern $xyyy$:



How was the MP crash point derived? I'm glad you asked!

Site Pattern	Correct Tree	Incorrect Tree
xxxx	0	0
xxxy	1	1
xxyx	1	1
xyxx	1	1
yxxx	1	1
xyyy	1	2
xyyx	2	1
xyxy	2	2
xxyz	2	2
xyzx	2	2
xyxz	2	2
yxxz	2	2
yxzx	2	2
yzxx	2	2
wxyz	3	3

$$P(xxyy) = f(T)$$

$$P(xyyx) = g(p_{switch}, T)$$

How was the MP crash point derived? I'm glad you asked!

Site Pattern	Correct Tree	Incorrect Tree
xxxx	0	0
xxxy	1	1
xxyx	1	1
xyxx	1	1
yxxx	1	1
xyyy	1	2
xyyx	2	1
xyxy	2	2
xxyz	2	2
xyzx	2	2
xyxz	2	2
yxxz	2	2
yxzx	2	2
yzxx	2	2
wxyz	3	3

$$P(xxyy) = f(T)$$

$$P(xyyx) = g(p_{switch}, T)$$

The failure point of MP is given

by finding p_{switch} such that:

$$P(xxyy) = P(xyyx)$$

How was the MP crash point derived? I'm glad you asked!

