Non-Reversible Models for Phylogenetics Using Nucleotide or Amino Acid Data

Sarah Parks and Nick Goldman

What is reversibility?

$$\pi_i q_{ij} = \pi_j q_{ji} \qquad Q = \begin{pmatrix} - & \pi_C a & \pi_G b & \pi_T c \\ \pi_A a & - & \pi_G d & \pi_T e \\ \pi_A b & \pi_C d & - & \pi_T f \\ \pi_A c & \pi_C e & \pi_G f & - \end{pmatrix}$$

- Can't find a root – "pulley principle"
- Biologically unrealistic

Questions

- Should we be using non-reversible models?
  - Nucleotides?
  - Amino Acids?
- How do we quantify reversibility?

Quantifying Reversibility

From data: estimate Q^{UN}, non–reversible model, and Q^{GTR}, reversible model

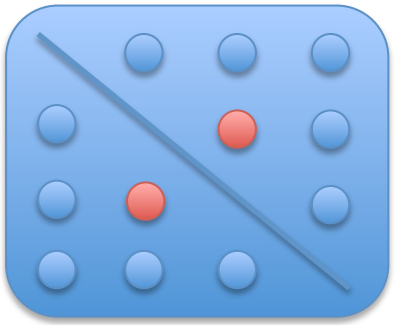
A. Likelihood Ratio Test statistic between Q^{UN} and Q^{GTR}

Quantifying Reversibility

From data: estimate Q^{UN}, non–reversible model, and Q^{GTR}, reversible model

A. Likelihood Ratio Test statistic between Q^{UN} and Q^{GTR}

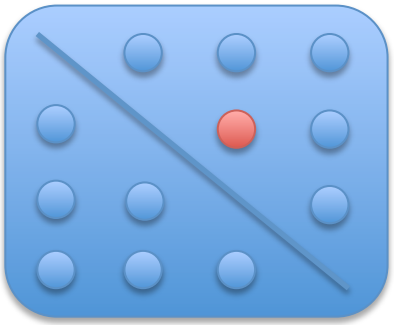
B. Deviation from the detailed balance equation
$\Sigma_{i,j} |\pi_i Q^{UN}_{i,j} - \pi_j Q^{UN}_{j,I}|$

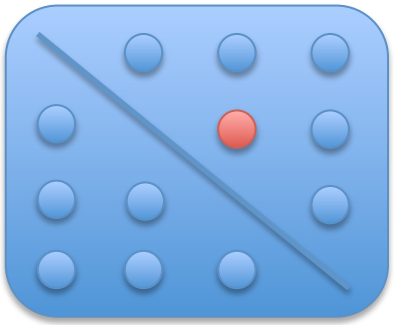


Quantifying Reversibility

From data: estimate Q^{UN}, non–reversible model, and Q^{GTR}, reversible model

C. Distance between Q^{UN} and Q^{GTR}
   i. $\Sigma_{i,j} |Q^{UN}_{i,j} - Q^{GTR}_{i,j}|$
   ii. $\Sigma_{i,j} |\pi_i Q^{UN}_{i,j} - \pi_i Q^{GTR}_{i,j}|$

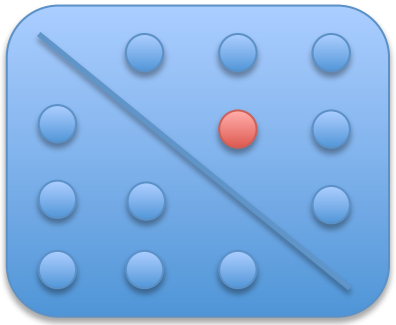


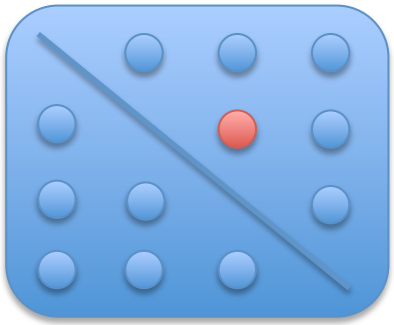


Quantifying Reversibility

From data: estimate Q^{UN}, non–reversible model, and Q^{GTR}, reversible model

D. Distance between Q^{UN} and closest rev model Q^{REV}
   $\min \Sigma_{i,j} |\pi_i Q^{UN}_{i,j} - \pi^{REV}_i Q^{REV}_{i,j}|$





Quantifying Reversibility

From data: estimate Q^{UN}, non–reversible model, and Q^{GTR}, reversible model

A. Likelihood Ratio Test statistic between Q^{UN} and Q^{GTR}

B. Deviation from the detailed balance equation
   $\Sigma_{i,j} |\pi_i Q^{UN}_{i,j} - \pi_j Q^{UN}_{j,I}|$

C. Distance between Q^{UN} and Q^{GTR}
   i. $\Sigma_{i,j} |Q^{UN}_{i,j} - Q^{GTR}_{i,j}|$
   ii. $\Sigma_{i,j} |\pi_i Q^{UN}_{i,j} - \pi^{GTR}_i Q^{GTR}_{i,j}|$

D. Distance between Q^{UN} and closest rev model Q^{REV}
   $\min \Sigma_{i,j} |\pi_i Q^{UN}_{i,j} - \pi^{REV}_i Q^{REV}_{i,j}|$

Quantifying Reversibility

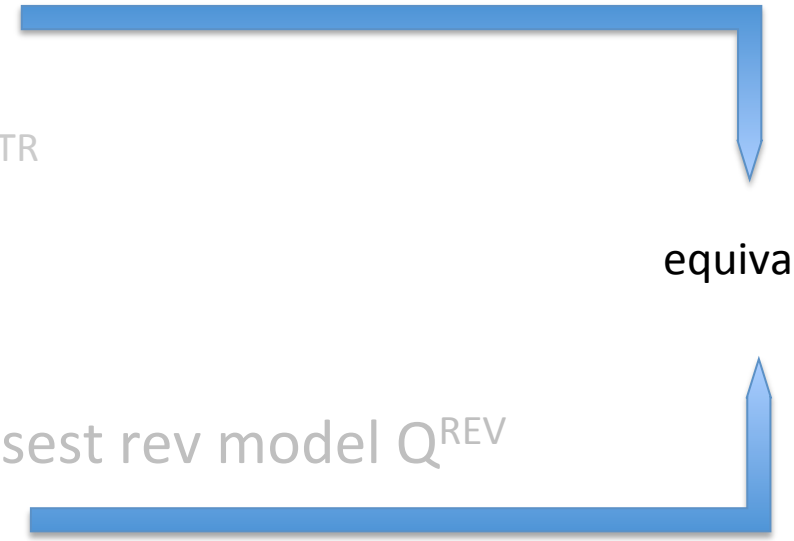
From data: estimate Q^{UN}, non–reversible model, and Q^{GTR}, reversible model

A. Likelihood Ratio Test statistic between Q^{UN} and Q^{GTR}

B. Deviation from the detailed balance equation
   $\Sigma_{i,j} |\pi_i Q^{UN}_{i,j} - \pi_j Q^{UN}_{j,I}|$

C. Distance between Q^{UN} and Q^{GTR}
   i. $\Sigma_{i,j} |Q^{UN}_{i,j} - Q^{GTR}_{i,j}|$
   ii. $\Sigma_{i,j} |\pi_i Q^{UN}_{i,j} - \pi^{GTR}_i Q^{GTR}_{i,j}|$

D. Distance between Q^{UN} and closest rev model Q^{REV}
   $\min \Sigma_{i,j} |\pi_i Q^{UN}_{i,j} - \pi^{REV}_i Q^{REV}_{i,j}|$



Quantifying Reversibility

From data: estimate Q^{UN}, non–reversible model, and Q^{GTR}, reversible model

A. Likelihood Ratio Test statistic between Q^{UN} and Q^{GTR}

B. Deviation from the detailed balance equation

$$\Sigma_{i,j} \left| \pi_i Q^{UN}_{i,j} - \pi_j Q^{UN}_{j,I} \right|$$

C. Distance between Q^{UN} and Q^{GTR}

i. $\Sigma_{i,j} \left| Q^{UN}_{i,j} - Q^{GTR}_{i,j} \right|$

ii. $\Sigma_{i,j} \left| \pi_i Q^{UN}_{i,j} - \pi^{GTR}_i Q^{GTR}_{i,j} \right|$

equivalent

D. Distance between Q^{UN} and closest rev model Q^{REV}

$$\min \Sigma_{i,j} \left| \pi_i Q^{UN}_{i,j} - \pi^{REV}_i Q^{REV}_{i,j} \right|$$

Quantifying Reversibility

From data: estimate Q^{UN}, non–reversible model, and Q^{GTR}, reversible model

A. Likelihood Ratio Test statistic between Q^{UN} and Q^{GTR}

B. Deviation from the detailed balance equation
   $\Sigma_{i,j} |\pi_i Q^{UN}_{i,j} - \pi_j Q^{UN}_{j,I}|$

C. Distance between Q^{UN} and Q^{GTR}
   i. $\Sigma_{i,j} |Q^{UN}_{i,j} - Q^{GTR}_{i,j}|$
   ii. $\Sigma_{i,j} |\pi_i Q^{UN}_{i,j} - \pi^{GTR}_i Q^{GTR}_{i,j}|$

D. Distance between Q^{UN} and closest rev model Q^{REV}
   $\min \Sigma_{i,j} |\pi_i Q^{UN}_{i,j} - \pi^{REV}_i Q^{REV}_{i,j}|$

Practicalities

- Datasets
  - Pandit - http://www.ebi.ac.uk/goldman-srv/pandit/
  - 38 mammals
- ML models estimated in HyPhy

Results – Nucleotides

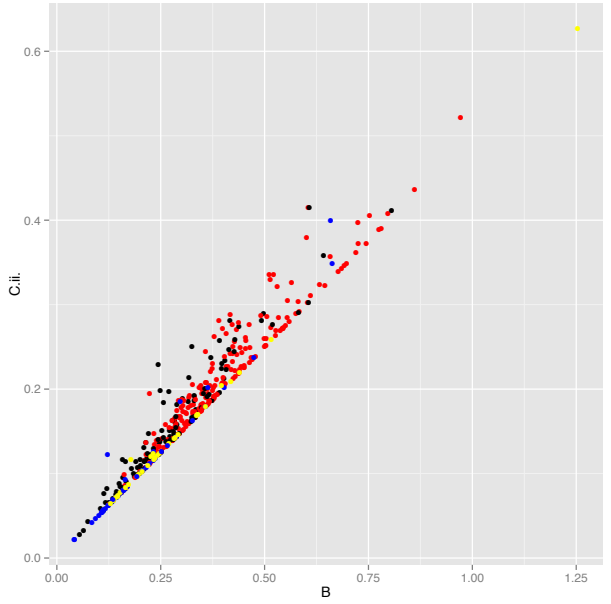
Measures

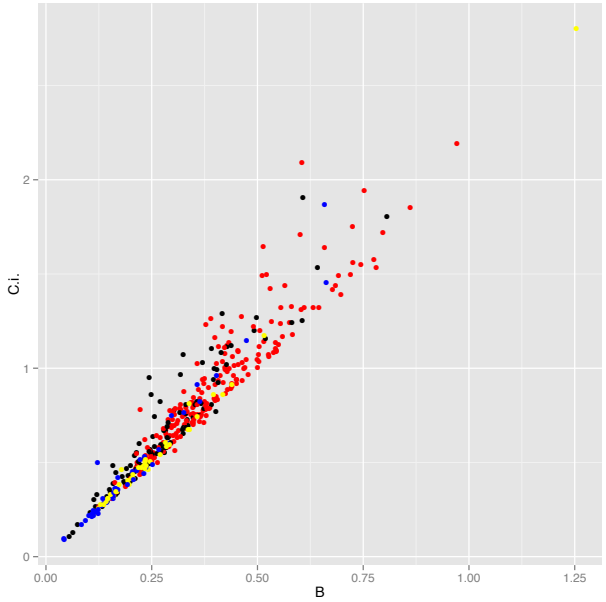
A. $-2(L(Q^{GTR})-L(Q^{UN}))/n$

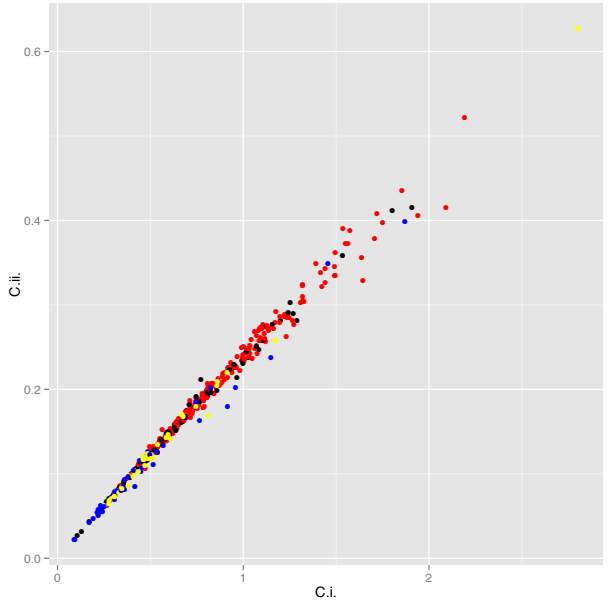
B. $\Sigma_{i,j} |\pi_i Q^{UN}_{i,j} - \pi_j Q^{UN}_{j,I}|$

C. i. $\Sigma_{i,j} |Q^{UN}_{i,j} - Q^{GTR}_{i,j}|$

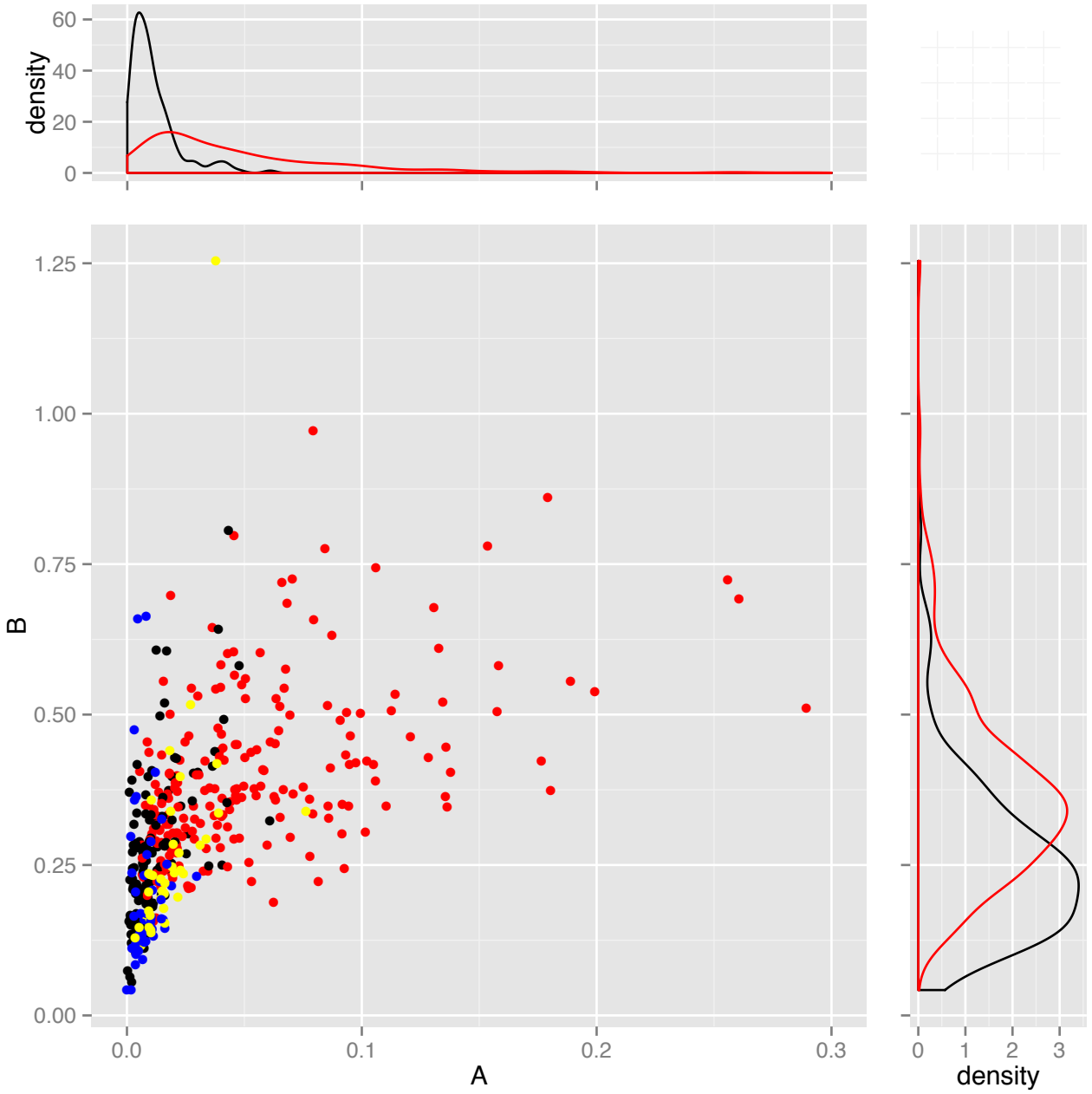
ii. $\Sigma_{i,j} |\pi_i Q^{UN}_{i,j} - \pi^{GTR}_i Q^{GTR}_{i,j}|$







Results – Nucleotides



Results – Amino Acids

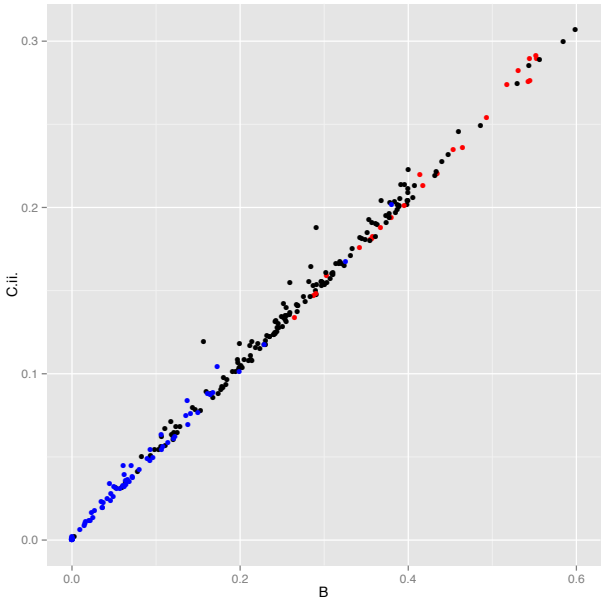
Measures

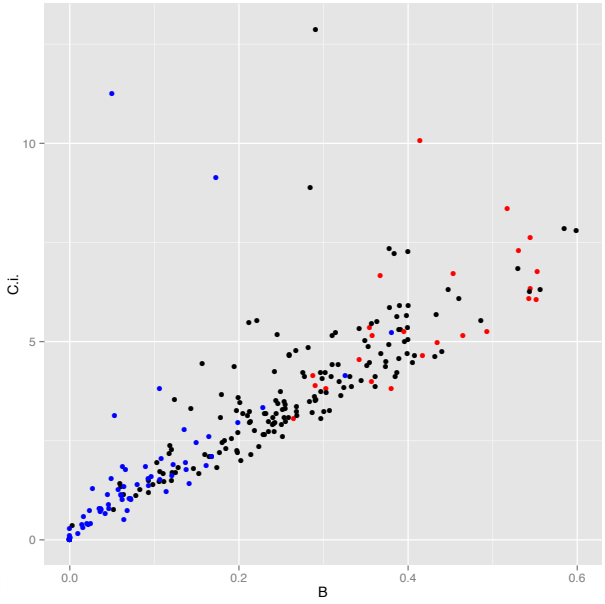
A. $-2(L(Q^{GTR})-L(Q^{UN}))/n$

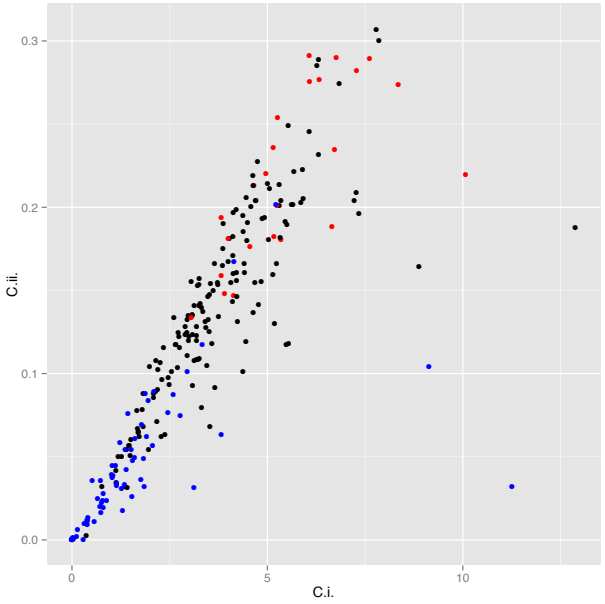
B. $\Sigma_{i,j} |\pi_i Q^{UN}_{i,j} - \pi_j Q^{UN}_{j,I}|$

C. i. $\Sigma_{i,j} |Q^{UN}_{i,j} - Q^{GTR}_{i,j}|$

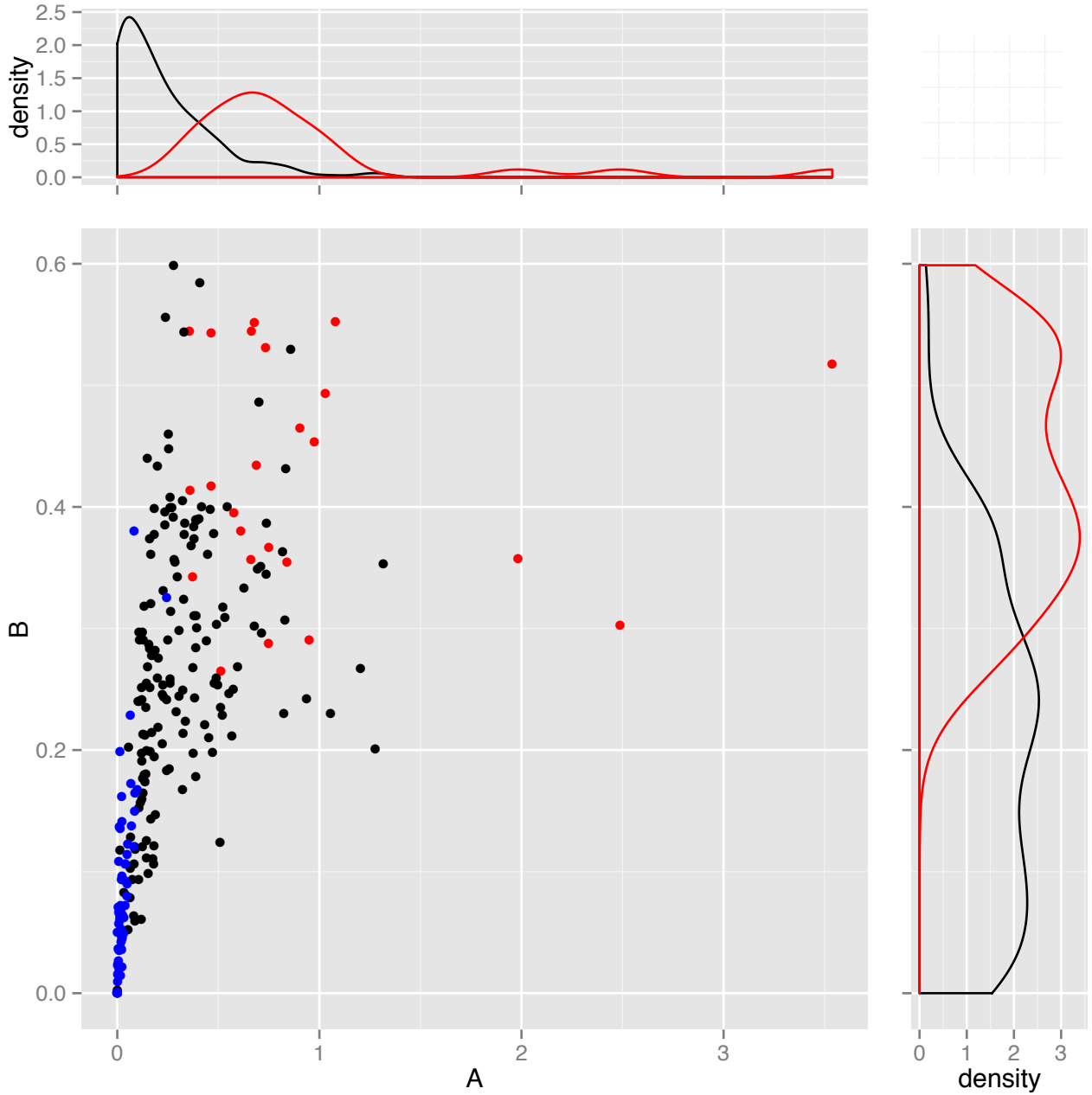
ii. $\Sigma_{i,j} |\pi_i Q^{UN}_{i,j} - \pi^{GTR}_i Q^{GTR}_{i,j}|$







Results – Amino Acids



Conclusions

- Non-reversible models often give a better fit for nucleotide data
- The branch lengths of the trees are not significantly changed by using a non-reversible model
- Even for small gene datasets non-reversible amino-acid models can be better
- There are many ways to quantify reversibility
  - $\Sigma_{i,j} |\pi_i Q^{UN}_{i,j} - \pi_j Q^{UN}_{j,I}|$

Acknowledgements

- Goldman Group
- TAC committee – Nick Goldman, Jan Korbel, John Marioni, Simon Tavaré
- Sergei Kovakovsky-Pond
- Greg Jordan