A Mathematical Model of Correlated Evolution

Patrick Fried
Marie-Hélène Descary
Chris Lester

Our Aims (1)

- DNA sequences evolve, but do so whilst preserving structural stability and functionality.
- Numerous studies estimate the interaction between amino acids, but it almost always use alignment that consider evolution at different sites independently.
- However, biologists have shown that evolution at a site is very much affected by its surroundings.

Our Aims (2)

- We aim to devise a model which incorporates pair-wise dependencies. This model would eventually allow for prediction of protein structure from sequence data alone.

- Three major building blocks:

1. Describe the way amino acids interact.

2. Consider all the possible dependency structures, possibly stochastically.

3. Efficiently assign a likelihood to each design. Choose most suitable one.

Brief Timeline (1)

- First week saw research conducted, start on coding project in R.
- Initial coding was promising, although mostly creating and structuring data structures and checking methods.
- Using binary alphabet, we adapted Felsenstein's Pruning Algorithm to calculate a structure's likelihood. Simulated data used.
- Then we expanded the alphabet, and made substitution rates realistic. Relied heavily on Miyazawa & Jernigan's work.

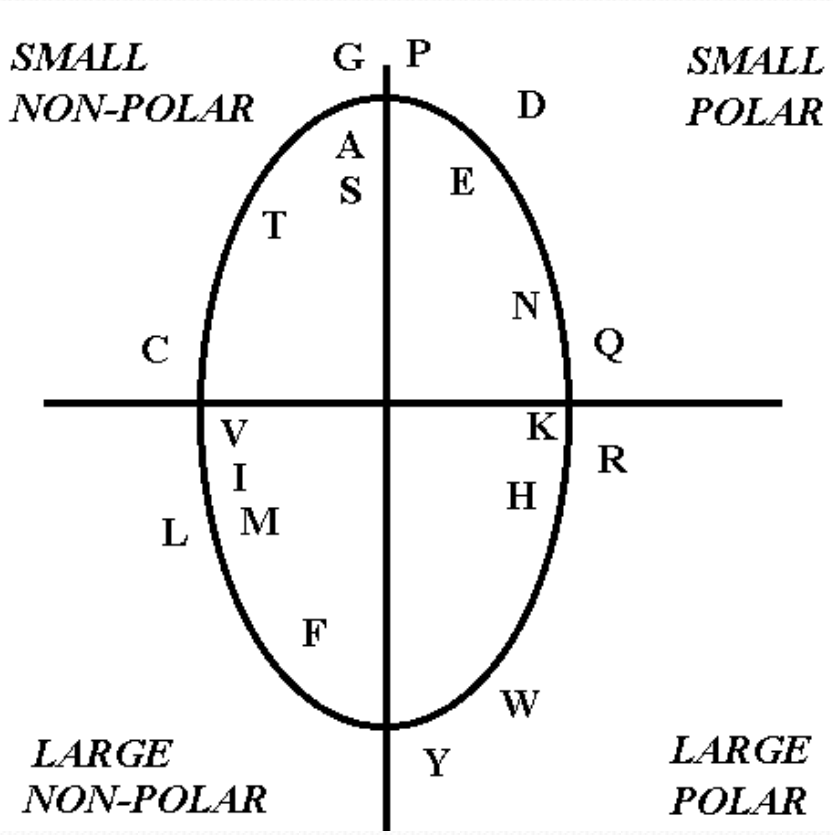
Brief Timeline (2)

- In fifth week, Markov chain Monte Carlo allowed for searching over possible tree structures. More sequences could be inputted. Code also moved over to C++.
- In sixth week, managed to get results. Added some rigour. Compiled all our research.
- Approximation with 6 groups of amino acids tested.

The Method

Reduced Amino Acid

- Reducing the alphabet size from 20 to 6 is computationally significant. Q(2) has 1296 entries, not 160 000.
- Behaviour of amino acids is correlated, as shown by Dayhoff.
- Thus group acids.

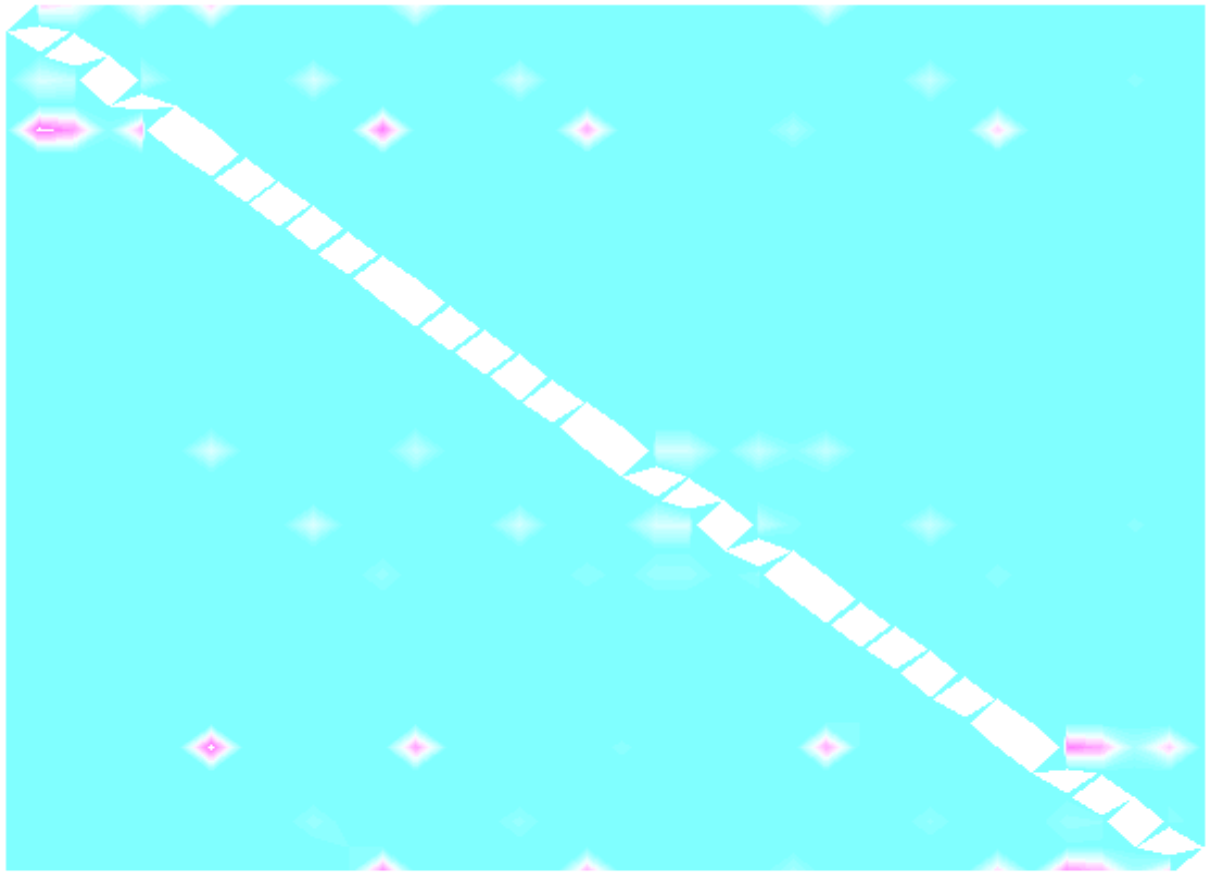


http://prowl.rockefeller.edu/aainfo/daysub.htm

Reduced Alphabet (2)

- Taken Miyazawa Jernigan matrix, grouped together amino acids and calculated average potential. This isn't ideal.
- Alternative was to simply sum up energies. This proved even worse.
- Ideally, although not feasible at last minute, would be to work back from Q(2) calculated for all twenty amino acids.

Reduced Alphabet (3)



Monte Carlo Markov Chain

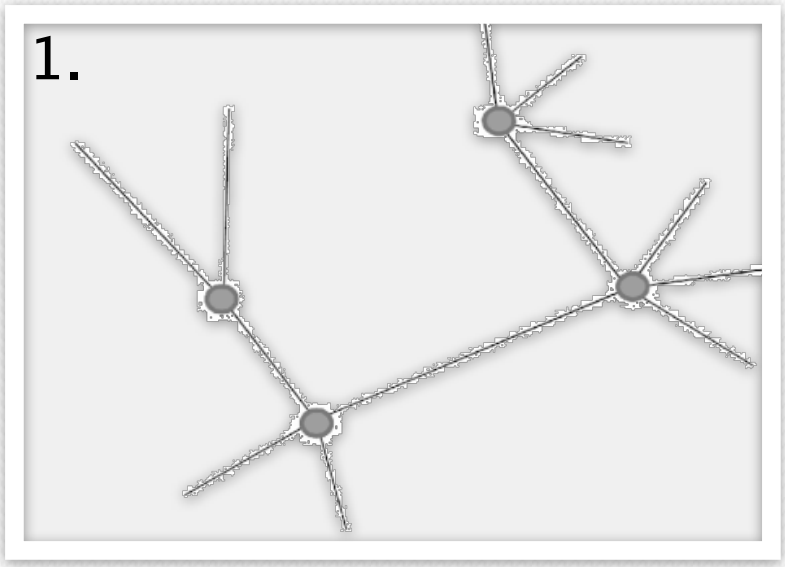
- Since the state space of spanning trees is huge, it's infeasible to calculate the likelihood of every possible dependency structure. To get around this, we implement "MCMC".
- What is the process ?
  - «Trial and improvement» game. Start off with candidate dependency structure.
  - Perturb it by choosing a connected node pair, and then considering a different way of connecting nodes.
  - Calculate likelihood of structure efficiently, and calculate Metropolis-Hastings ratio.
  - Accept or reject change, according to this ratio.

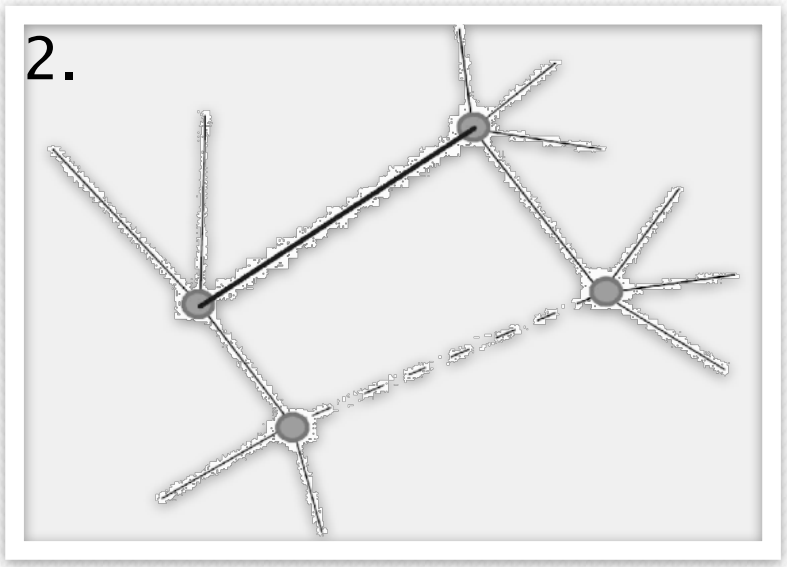
Monte Carlo Markov Chain

- In particular, we use the Metropolis-Hasting method. We calculate:

$$\alpha(\theta, \phi) = \min\left\{1, \frac{p(\phi)p(X|\phi)q(\theta, \phi)}{p(\theta)p(X|\theta)q(\phi, \theta)}\right\}$$

- Accept changes with probability alpha – else reject.



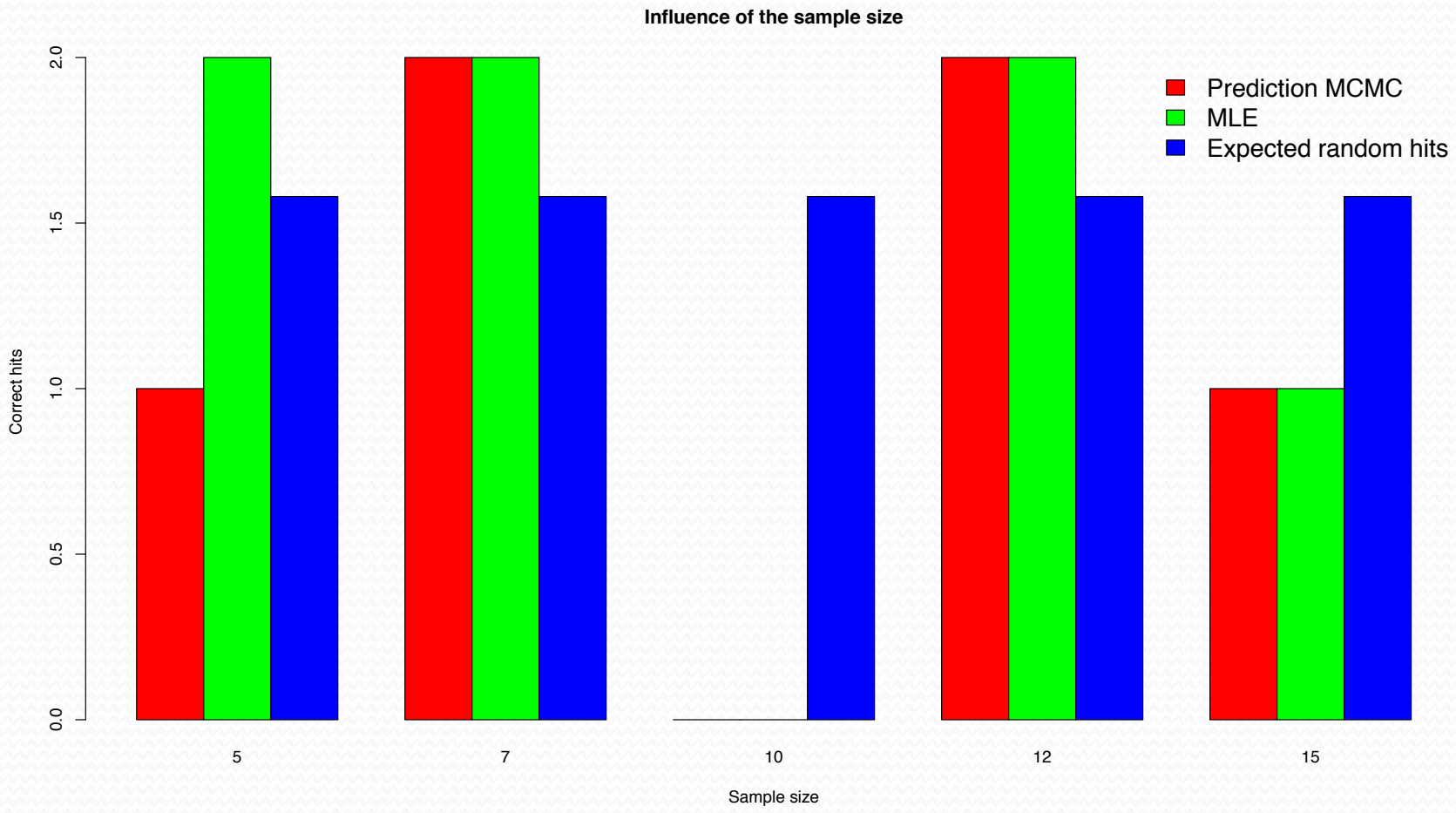


The Results

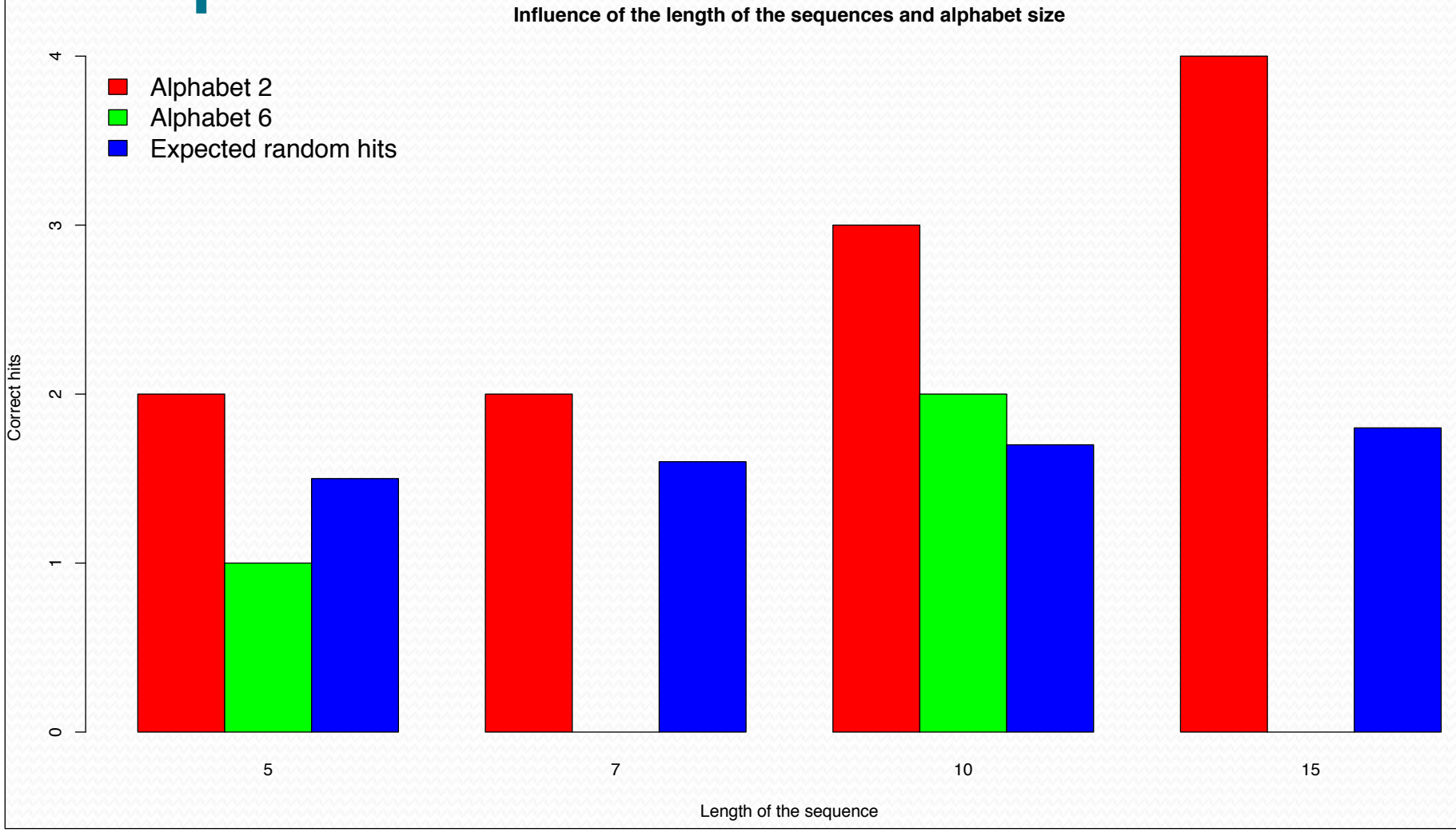
Analysis of the results

- We test our method with differents parameters:
  - Size of the sample
  - Length of the sequence
  - Size of the alphabet
  - «Heating» the MCMC
  - Different inital starting point

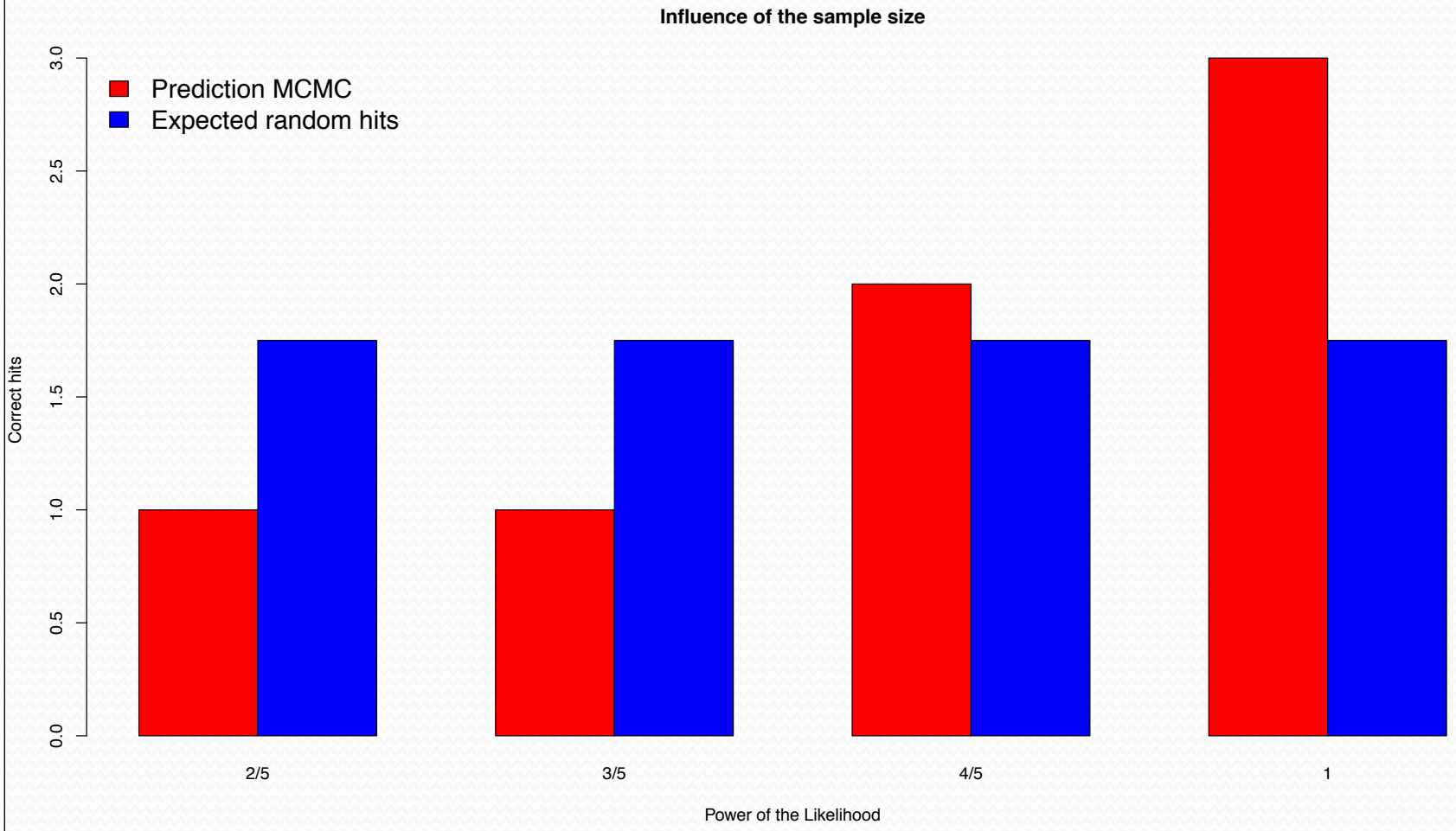
Sample size (L=7)



Sequence Length and Alphabet size



Heating of the MCMC (L=10,N=12)



Different initial starting point for the MCMC

- Real dependency structure:
  - (1,2) (2,3) (2,5) (3,4) (3,7) (5,6) (5,12) (5,13) (7,8) (7,9) (10,12) (11,12) (13,14) (13,15)
- When we start the MCMC with a dependency structure that shares 5 edges with the real one we obtain the following results:
- Predicted dependency structure :
  - (11,12) (8,15) (1,2) (1,6) (2,3) (3,4) (5,9) (6,7) (7,13) (9,10) (10,14) (13,14) (5,12)
- When we start the MCMC with a dependency structure that shares 10 edges with the real one we obtain the following results:
- Predicted dependency structure :
  - (4,6) (11,12) (1,2) (1,15) (2,5) (3,7) (3,13) (5,10) (7,8) (8,9) (13,14) (14,15) (6,9) (10,12)

The Future

Improvements

- Verification with actual data needed.
- Only substitutions considered: could incorporate deletions and insertions.
  - TKF91 model could be adapted.
- Alignment of input sequences needs further consideration.
  - Incorporation of project into alignment programmes helpful.
- Need to establish extent of improvement over current methods.

Towards a D.Phil.

- Real need for more rigour.
  - Substantial verification of substitution rate matrix needed. Current approach could be too coarse.
  - Establish significance of likelihood estimates, and sensitivity to parameters.
- Substantially increase the scope.
  - “Indel” needs consideration.
  - Integration with alignment packages.
- Advance MCMC method : Tempered MCMC
- An amazing result.

Questions?

1. Aims and Ambitions
2. Methods
3. Results

Thanks!