Detecting an evolutionary signal between pairs of circular genomes

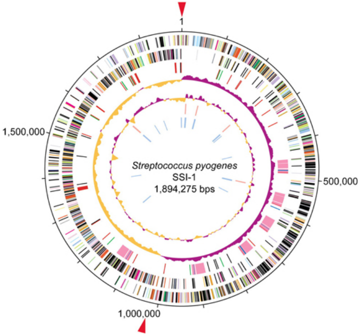
Venta Terauds and Jeremy Sumner
with David Bryant, Andrew Francis and Peter Jarvis

Discipline of Mathematics
University of Tasmania

MAM10
February 2019

The motivation

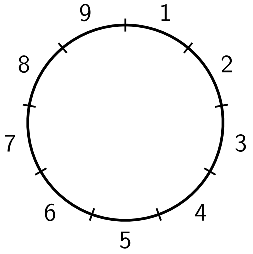
Bacterial genomes are circular and evolve via a combination of processes.

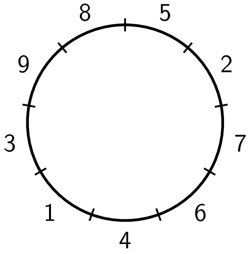


To model bacterial evolution, we focus on differences in genomic **structure**, rather than **content**.

The theory

Given two circular genomes that share N regions of interest . . .



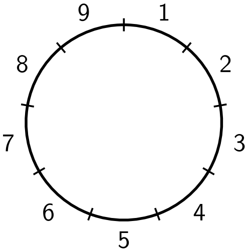


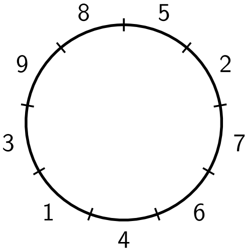
. . . we

- use a **rearrangement model** to find possible '**evolutionary paths**' from one genome to the other;

The theory

Given two circular genomes that share *N* regions of interest . . .





. . . we

- use a **rearrangement model** to find possible ‘**evolutionary paths**’ from one genome to the other;
- then apply a **distance method** to estimate the **evolutionary distance** between them.

Rearrangement models

We represent a **genome with N regions** by a permutation $\sigma \in S_N$, where $\sigma(i) = j \iff$ region i is in position j.

Rearrangement models

We represent a **genome with N regions** by a permutation $\sigma \in S_N$, where $\sigma(i) = j \iff$ region i is in position j.

A **rearrangement of the genome** σ occurs when a permutation, $a \in \mathcal{S}_N$, acts on σ (on the left):

$$\sigma \mapsto a\,\sigma\,.$$

Rearrangement models

We represent a **genome with N regions** by a permutation $\sigma \in S_N$, where $\sigma(i) = j \iff$ region i is in position j.

A **rearrangement of the genome** σ occurs when a permutation, $a \in \mathcal{S}_N$, acts on σ (on the left):

$$\sigma \mapsto a\,\sigma\,.$$

To specify a **rearrangement model**, we need

- a set of allowed rearrangements $\mathcal{M} = \{a_1, a_2, a_3, \ldots, a_R\} \subseteq \mathcal{S}_N \setminus D_N$;

Rearrangement models

We represent a **genome with N regions** by a permutation $\sigma \in S_N$, where $\sigma(i) = j \iff$ region i is in position j.

A **rearrangement of the genome** σ occurs when a permutation, $a \in \mathcal{S}_N$, acts on σ (on the left):

$$\sigma \mapsto a\,\sigma\,.$$

To specify a **rearrangement model**, we need

- a set of allowed rearrangements $\mathcal{M} = \{a_1, a_2, a_3, \ldots, a_R\} \subseteq \mathcal{S}_N \setminus D_N$;
- a set of rearrangement probabilities $\{w(a_i) : a_i \in \mathcal{M}\}$;

Rearrangement models

We represent a **genome with N regions** by a permutation $\sigma \in S_N$, where $\sigma(i) = j \iff$ region i is in position j.

A **rearrangement of the genome** σ occurs when a permutation, $a \in \mathcal{S}_N$, acts on σ (on the left):

$$\sigma \mapsto a\,\sigma\,.$$

To specify a **rearrangement model**, we need

- a set of allowed rearrangements $\mathcal{M} = \{a_1, a_2, a_3, \ldots, a_R\} \subseteq \mathcal{S}_N \backslash D_N$;
- a set of rearrangement probabilities $\{w(a_i) : a_i \in \mathcal{M}\}$;
- a distribution of events in time, *dist*.

The evolutionary distance measure

Our evolutionary distance measure is the **maximum likelihood estimate of time elapsed** (MLE).

This is the **most probable amount of time** taken for the reference genome to evolve into a target genome under the given model.

The evolutionary distance measure

Our evolutionary distance measure is the **maximum likelihood estimate of time elapsed** (MLE).

This is the **most probable amount of time** taken for the reference genome to evolve into a target genome under the given model.

Precisely, for a genome represented by $\sigma \in \mathcal{S}_N$, it's the time, T, at which the likelihood function $L(\sigma|T)$ attains its maximum*, where

$$
\begin{aligned}
&L(\sigma|T) := \mathrm{P}(\mathrm{id} \to [\sigma] \text{ in time } T) \\
&\quad = \sum_{k=0}^{\infty} \mathrm{P}(\mathrm{id} \to [\sigma] \text{ via } k \text{ rearrangements })\mathrm{P}(k \text{ rearrangements in time } T)
\end{aligned}
$$

Calculating the MLE – rewriting the likelihood function

Define $\mathbf{s} := \sum_{a \in \mathcal{M}} w(a)a$ in the group algebra $\mathbb{C}[\mathcal{S}_N]$ and observe that

$$\mathbf{s}^k = \sum_{\sigma \in \mathcal{S}_N} \beta_k(\sigma)\sigma\,.$$

Calculating the MLE – rewriting the likelihood function

Define $\mathbf{s} := \sum_{a \in \mathcal{M}} w(a)a$ in the group algebra $\mathbb{C}[\mathcal{S}_N]$ and observe that

$$\mathbf{s}^k = \sum_{\sigma \in \mathcal{S}_N} \beta_k(\sigma)\sigma \,.$$

where for each permutation σ, the coefficient $\beta_k(\sigma)$ in the expansion is the probability of obtaining σ from the identity in k rearrangements.

Calculating the MLE – rewriting the likelihood function

Define $\mathbf{s} := \sum_{a\in\mathcal{M}} w(a)a$ in the group algebra $\mathbb{C}[\mathcal{S}_N]$ and observe that

$$\mathbf{s}^k = \sum_{\sigma\in\mathcal{S}_N} \beta_k(\sigma)\sigma\,.$$

where for each permutation σ, the coefficient $\beta_k(\sigma)$ in the expansion is the probability of obtaining σ from the identity in k rearrangements.

For any σ, we can rewrite the above so that $\beta_k(\sigma)$ is the coefficient of the identity in the expansion of $\sigma^{-1}\mathbf{s}^k$.

Calculating the MLE – rewriting the likelihood function

Define $\mathbf{s} := \sum_{a \in \mathcal{M}} w(a)a$ in the group algebra $\mathbb{C}[\mathcal{S}_N]$ and observe that

$$\mathbf{s}^k = \sum_{\sigma \in \mathcal{S}_N} \beta_k(\sigma)\sigma\,.$$

where for each permutation σ, the coefficient $\beta_k(\sigma)$ in the expansion is the probability of obtaining σ from the identity in k rearrangements.

For any σ, we can rewrite the above so that $\beta_k(\sigma)$ is the coefficient of the identity in the expansion of $\sigma^{-1}\mathbf{s}^k$.

Now using the regular representation of $\mathcal{S}_N$ extended to $\mathbb{C}[\mathcal{S}_N]$, we have for each $\sigma \in \mathcal{S}_N$

$$\beta_k(\sigma) = \tfrac{1}{N!}\chi_{\mathrm{reg}}(\sigma^{-1}\mathbf{s}^k)\,,$$

Calculating the MLE – rewriting the likelihood function

Define $\mathbf{s} := \sum_{a \in \mathcal{M}} w(a)a$ in the group algebra $\mathbb{C}[\mathcal{S}_N]$ and observe that

$$\mathbf{s}^k = \sum_{\sigma \in \mathcal{S}_N} \beta_k(\sigma)\sigma \, .$$

where for each permutation σ, the coefficient $\beta_k(\sigma)$ in the expansion is the probability of obtaining σ from the identity in k rearrangements.

For any σ, we can rewrite the above so that $\beta_k(\sigma)$ is the coefficient of the identity in the expansion of $\sigma^{-1}\mathbf{s}^k$.

Now using the regular representation of $\mathcal{S}_N$ extended to $\mathbb{C}[\mathcal{S}_N]$, we have for each $\sigma \in \mathcal{S}_N$

$$\beta_k(\sigma) = \tfrac{1}{N!}\chi_{\mathrm{reg}}(\sigma^{-1}\mathbf{s}^k) \, ,$$

so that

$$\mathrm{P}(e \to [\sigma] \text{ via } k \text{ rearrangements }) = \tfrac{1}{N!}\chi_{\mathrm{reg}}(\sigma^{-1}\mathbf{d}\mathbf{s}^k) \, ,$$

where we have incorporated the symmetries of the genome using $\mathbf{d} := \sum_{d \in D_N} d \ \in \mathbb{C}[\mathcal{S}_N]$.

Rearrangement model → Markov model

Now, setting the distribution of events in time to be Poisson(1), we have

$$\begin{aligned} L(\sigma|T) &= \tfrac{1}{N!}\sum_{k=0}^{\infty}\chi_{\mathrm{reg}}(\sigma^{-1}\mathbf{ds}^{k})\tfrac{\mathrm{e}^{-T}T^{k}}{k!} \\ &= \tfrac{1}{N!}\chi_{\mathrm{reg}}(\sigma^{-1}\mathbf{d}\mathrm{e}^{(\mathbf{s}-\mathrm{id})T}) \\ &= \tfrac{1}{N!}\chi_{\mathrm{reg}}(\sigma^{-1}\mathbf{d}\mathrm{e}^{QT})\,, \end{aligned}$$

where $Q = \rho_{\mathrm{reg}}(\mathbf{s} - \mathrm{id})$.

Observe that $\rho_{\mathrm{reg}}(\mathbf{s})$ is in fact the transition matrix for a discrete Markov chain with state space $\mathcal{S}_N$.

Thus Q is the generator matrix for a continuous time Markov chain and we see that the rearrangement model gives rise to a 'group-based' Markov model.

Rearrangement model $\rightarrow$ Markov model

Now, setting the distribution of events in time to be Poisson(1), we have

$$L(\sigma|T) = \tfrac{1}{N!}\sum_{k=0}^{\infty}\chi_{\mathrm{reg}}(\sigma^{-1}\mathbf{ds}^k)\tfrac{\mathrm{e}^{-T}T^k}{k!}$$

$$= \tfrac{1}{N!}\chi_{\mathrm{reg}}(\sigma^{-1}\mathbf{d}\mathrm{e}^{QT}),$$

where $Q = \rho_{\mathrm{reg}}(\mathbf{s} - \mathrm{id})$.

Observe that $\rho_{\mathrm{reg}}(\mathbf{s})$ is in fact the transition matrix for a discrete Markov chain with state space $\mathcal{S}_N$.

Thus Q is the generator matrix for a continuous time Markov chain and we see that the rearrangement model gives rise to a 'group-based' Markov model.

Rearrangement model → Markov model

Now, setting the distribution of events in time to be Poisson(1), we have

$$\begin{aligned} L(\sigma | T) &= \frac{1}{N!} \sum_{k=0}^{\infty} \chi_{\mathrm{reg}}(\sigma^{-1} \mathbf{d} \mathbf{s}^k) \frac{\mathrm{e}^{-T} T^k}{k!} \\ &= \frac{1}{N!} \chi_{\mathrm{reg}}(\sigma^{-1} \mathbf{d} \mathrm{e}^{(\mathbf{s} - \mathrm{id}) T}) \\ &= \frac{1}{N!} \chi_{\mathrm{reg}}(\sigma^{-1} \mathbf{d} \mathrm{e}^{Q T}) \,, \end{aligned}$$

where $Q = \rho_{\mathrm{reg}}(\mathbf{s} - \mathrm{id})$.

Rearrangement model → Markov model

Now, setting the distribution of events in time to be Poisson(1), we have

$$\begin{aligned} L(\sigma | T) &= \tfrac{1}{N!} \sum_{k=0}^{\infty} \chi_{\mathrm{reg}}(\sigma^{-1}\mathbf{d}\mathbf{s}^k) \tfrac{\mathrm{e}^{-T} T^k}{k!} \\ &= \tfrac{1}{N!} \chi_{\mathrm{reg}}(\sigma^{-1}\mathbf{d}\mathrm{e}^{(\mathbf{s}-\mathrm{id})T}) \\ &= \tfrac{1}{N!} \chi_{\mathrm{reg}}(\sigma^{-1}\mathbf{d}\mathrm{e}^{QT}) \,, \end{aligned}$$

where $Q = \rho_{\mathrm{reg}}(\mathbf{s} - \mathrm{id})$.

Observe that $\rho_{\mathrm{reg}}(\mathbf{s})$ is in fact the transition matrix for a discrete Markov chain with state space $\mathcal{S}_N$.

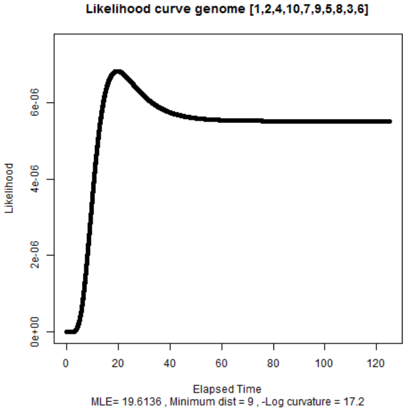
Thus Q is the generator matrix for a continuous time Markov chain and we see that the rearrangement model gives rise to a ‘group-based’ Markov model.

Computing the likelihood

To compute, we **decompose** into irreducible representations of $\mathcal{S}_N$ and, assuming time reversibility of the stochastic model (this is equivalent to $\mathcal{M} = \mathcal{M}^{-1}$ with $w(a^{-1}) = w(a)$ for all $a \in \mathcal{M}$), we **diagonalise**, obtaining

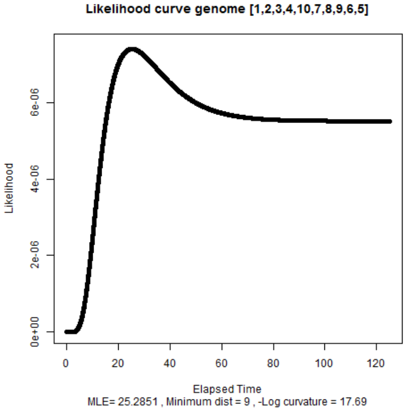
$$L(\sigma | T) = \tfrac{1}{N!} \sum_{p \dashv N} D_p \sum_{i=1}^{r_p} \mathrm{tr}(\rho_p(\sigma^{-1}\mathbf{d}) E_{p,i}) e^{\lambda_{p,i} T} .$$

Some likelihood plots - “Model 1”



Likelihood curve genome [1,2,4,10,7,9,5,8,3,6]

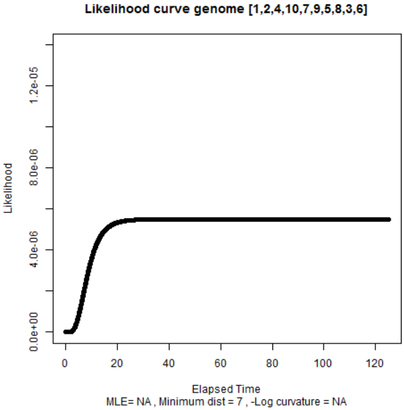
MLE= 19.6136 , Minimum dist = 9 , -Log curvature = 17.2

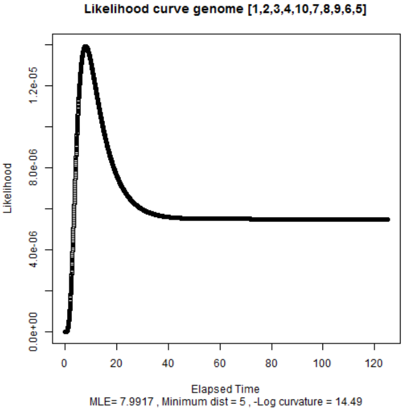


Likelihood curve genome [1,2,3,4,10,7,8,9,6,5]

MLE= 25.2851 , Minimum dist = 9 , -Log curvature = 17.69

Some more likelihood curves - “Model 2”





What can we actually compute?

In the above form, calculating the MLE has complexity approx $\sqrt{N!}$.

What can we actually compute?

In the above form, calculating the MLE has complexity approx $\sqrt{N!}$.

However.... even without knowing the value of the MLE, knowing **whether or not it exists**, ie, whether or not two genomes are related under a particular model, can still be useful.

What can we actually compute?

In the above form, calculating the MLE has complexity approx $\sqrt{N!}$.

However.... even without knowing the value of the MLE, knowing **whether or not it exists**, ie, whether or not two genomes are related under a particular model, can still be useful.

Observe that each likelihood function is just a finite (weighted) sum of exponentials,

$$L(T|\sigma) = b_0 e^{\lambda_0 T} + b_1 e^{\lambda_1 T} + b_2 e^{\lambda_2 T} + b_3 e^{\lambda_3 T} + \ldots + b_m e^{\lambda_m T}\,,$$

where each $b_i \neq 0$ and the eigenvalues λ_i are decreasing, ie

$$0 = \lambda_0 > \lambda_1 > \lambda_2 > \ldots > \lambda_m \geq -2\,;$$

What can we actually compute?

In the above form, calculating the MLE has complexity approx $\sqrt{N!}$.

However.... even without knowing the value of the MLE, knowing **whether or not it exists**, ie, whether or not two genomes are related under a particular model, can still be useful.

Observe that each likelihood function is just a finite (weighted) sum of exponentials,

$$L(T|\sigma) = b_0 e^{\lambda_0 T} + b_1 e^{\lambda_1 T} + b_2 e^{\lambda_2 T} + b_3 e^{\lambda_3 T} + \ldots + b_m e^{\lambda_m T} \,,$$

where each $b_i \neq 0$ and the eigenvalues λ_i are decreasing, ie

$$0 = \lambda_0 > \lambda_1 > \lambda_2 > \ldots > \lambda_m \geq -2 \,;$$

Taking the derivative, we see that as $T \to \infty$,

$$L'(T|\sigma) \approx b_1 \lambda_1 e^{\lambda_1 T} \,.$$

Does an evolutionary signal exist?

Does an evolutionary signal exist?

Theorem

If $b_1 > 0$, then the likelihood function has a maximum, i.e., an MLE exists.

Does an evolutionary signal exist?

Theorem

If $b_1 > 0$, then the likelihood function has a maximum, i.e., an MLE exists.

This is a simple consequence of our observations above. The exponential function is always positive, and $\lambda_1 < 0$, so we see that if $b_1 > 0$, then the slope of the likelihood curve, as $T \to \infty$, is negative.

Does an evolutionary signal exist?

Theorem

If $b_1 > 0$, then the likelihood function has a maximum, i.e., an MLE exists.

This is a simple consequence of our observations above. The exponential function is always positive, and $\lambda_1 < 0$, so we see that if $b_1 > 0$, then the slope of the likelihood curve, as $T \to \infty$, is negative.

What about $b_1 < 0$?

If it is the case that the likelihood function has either no maximum or one maximum, then $b_1 < 0$ means that we have no MLE.

Does an evolutionary signal exist?

Theorem

If $b_1 > 0$, then the likelihood function has a maximum, i.e., an MLE exists.

This is a simple consequence of our observations above. The exponential function is always positive, and $\lambda_1 < 0$, so we see that if $b_1 > 0$, then the slope of the likelihood curve, as $T \to \infty$, is negative.

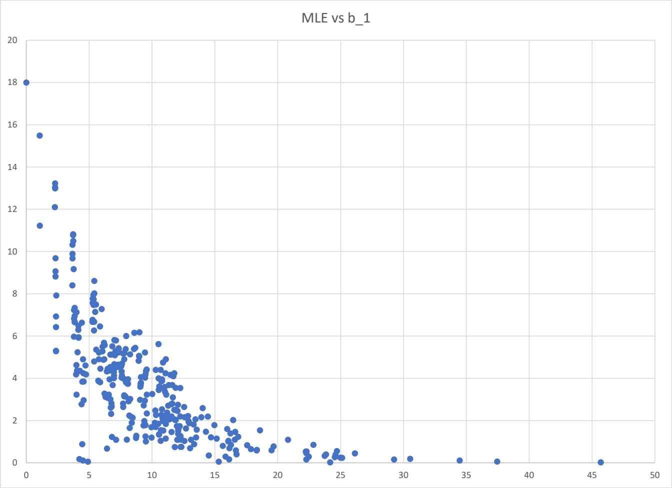
What about $b_1 < 0$?

If it is the case that the likelihood function has either no maximum or one maximum, then $b_1 < 0$ means that we have no MLE.

One can easily create sums of exponentials that have multiple optima.

However, using actual models (for genomes with up to 12 regions), we have only ever been able to create likelihood functions with zero or one maximum.

‘Model 2’; $\mathcal{S}_9$: MLE vs b_1 for genomes with an MLE



How much easier is this question?

We know exactly where to look for the “second biggest eigenvalue”, λ_1.

It’s in the ‘third’ irreducible representation, that is, it’s an eigenvalue of the matrix

$$\rho_{[N-2,2]}(\mathbf{s}) \, .$$

How much easier is this question?

We know exactly where to look for the “second biggest eigenvalue”, λ_1.

It’s in the ‘third’ irreducible representation, that is, it’s an eigenvalue of the matrix

$$\rho_{[N-2,2]}(\mathbf{s})\,.$$

How much easier is this question?

We know exactly where to look for the “second biggest eigenvalue”, λ_1.

It’s in the ‘third’ irreducible representation, that is, it’s an eigenvalue of the matrix

$$\rho_{[N-2,2]}(\mathbf{s})\,.$$

(nb. Observationally, this is always true, but the algebraic proof is still pending.)

How much easier is this question?

We know exactly where to look for the “second biggest eigenvalue”, λ_1.

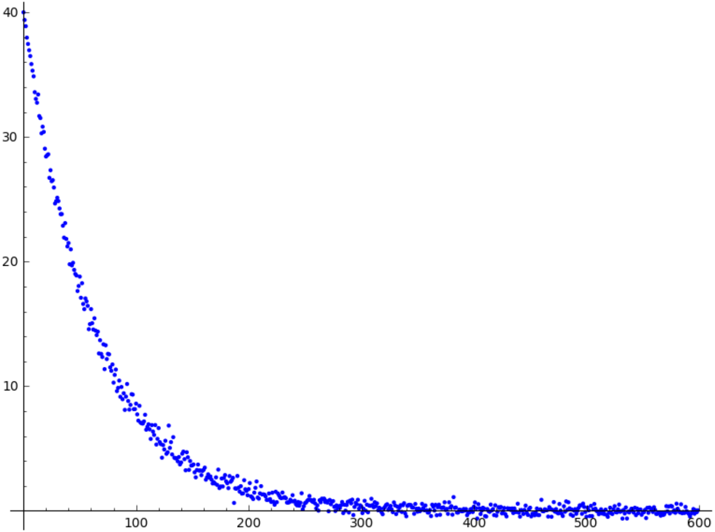
It’s in the ‘third’ irreducible representation, that is, it’s an eigenvalue of the matrix

$$\rho_{[N-2,2]}(\mathbf{s})\,.$$

(nb. Observationally, this is always true, but the algebraic proof is still pending.)

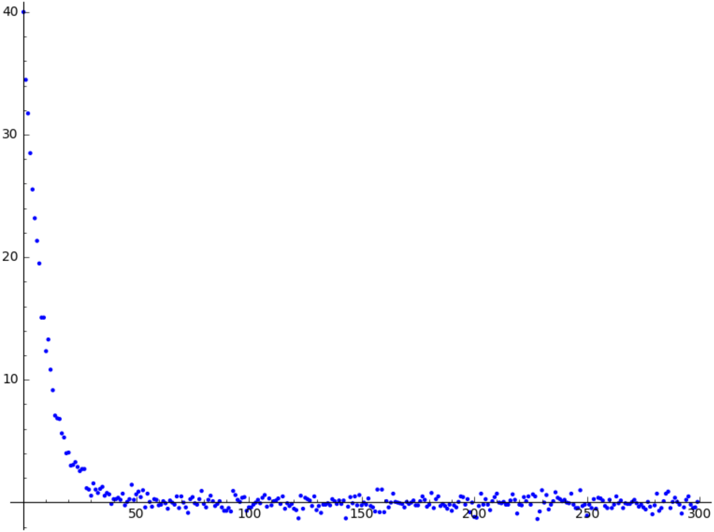
In any case, for N regions, this matrix has dimension $\frac{N^2-3N}{2}$.... which makes computations simple.

simulating: mean b_1 vs T



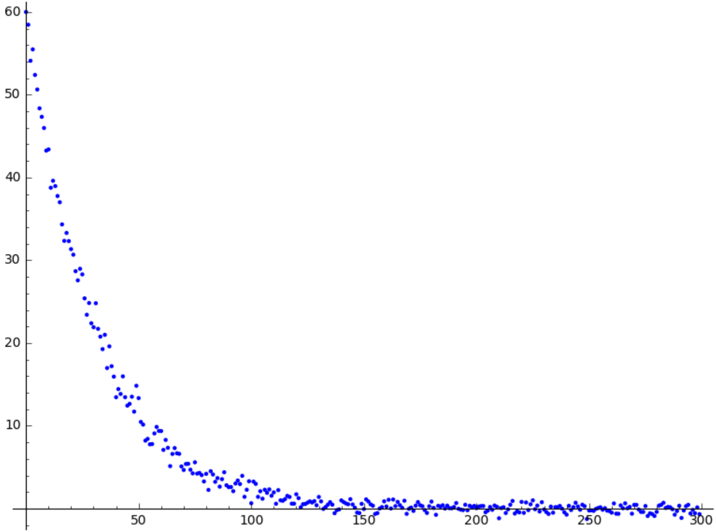
$\mathcal{S}_{20}$, model T_1, 100 repetitions, 600 time steps

simulating



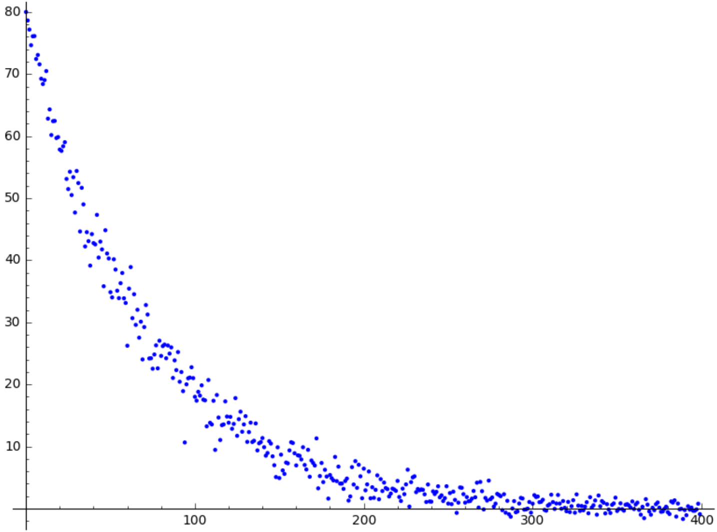
$\mathcal{S}_{20}$, all inversions model, 40 repetitions, 300 time steps

simulating



$\mathcal{S}_{30}$, inv7 model, 40 repetitions, 300 time steps

simulating



$\mathcal{S}_{40}$, inv7 model, 10 repetitions, 400 time steps

What next?

As far as this predictor goes, we have a couple of gaps to fill in (eg prove that $b_1 < 0 \implies$ no MLE under our model/symmetry assumptions).

More generally, a priority is to increase the number of regions for which we can calculate MLEs. In particular/in parallel...

- Most eigenvalues that we calculate do not contribute to the final likelihood function (as their coefficient b_i is zero). We now understand why this is and are working on a way to apply this (which will massively reduce our computational load!).
- We may still have to start to use some real numerical approximations (as opposed to the ones the computer does in order to actually calculate anything).
- Investigate further applying the technique to compare models – eg what is the 'most likely model' for some given data?
- Apply/adapt this technique to slightly different genome models. eg include an origin and terminus of replication, include gene orientation ... etc

References

S. Bhatia, P. Feijao, A. R. Francis, Position and content paradigms in genome rearrangements: the wild and crazy world of permutations in genomics. *Bull Math Biol* (2018) 80: 3227.

S. Serdoz, A. Egri-Nagy, J. G. Sumner, B. Holland, P. D. Jarvis, M. M. Tanaka, and A. R. Francis. Maximum likelihood estimates of pairwise rearrangement distances, *J. Theoret. Biol.* **423** (2017), 31–40.

J. G. Sumner, P. D. Jarvis, A. R. Francis, A representation-theoretic approach to the calculation of evolutionary distance in bacteria. *J. Phys A: Math. Theor.*, **50** (2017) 335601 (14pp).

V. Terauds, J. G. Sumner, Maximum likelihood estimates of rearrangement distance: implementing a representation-theoretic approach, *Bull Math Biol* **81** (2019), 535–567.

Thanks

This work was supported by

- ARC Discovery Project grant DP180102215 (CIs Jeremy Sumner, David Bryant, Andrew Francis and Peter Jarvis);
- the Nectar Research Cloud, a collaborative Australian research platform supported by the National Collaborative Research Infrastructure Strategy.

We used open-source software, SageMath and R, for all computations.

sage

nectar

Australian Government
Australian Research Council

R