**Intro –**

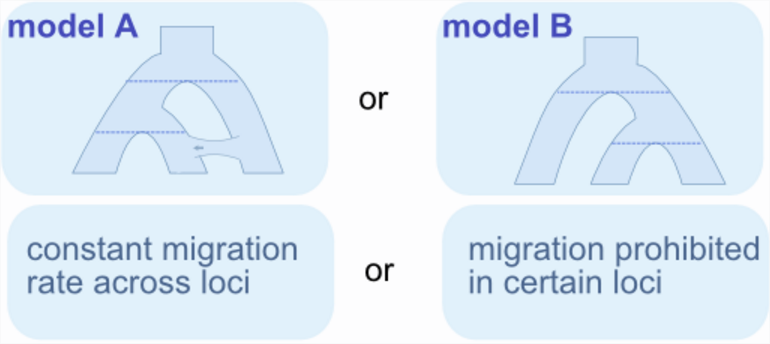
* explanation about Gphocs purpose and implementation
* motivation –
  + Talk about the model compare problem (example birds).
* Previous approaches and disadvantages
  + harmonic mean
    - Issues with harmonic mean
  + Other standard methods

**Theory –**

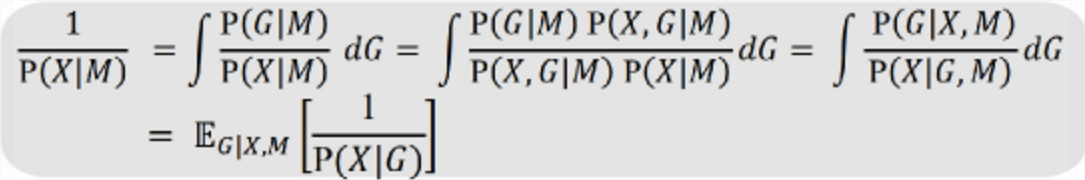
* **The model comparison problem**
  + The problem we are trying to solve is:

Given two demographic population models of a species, which better fits the Data?

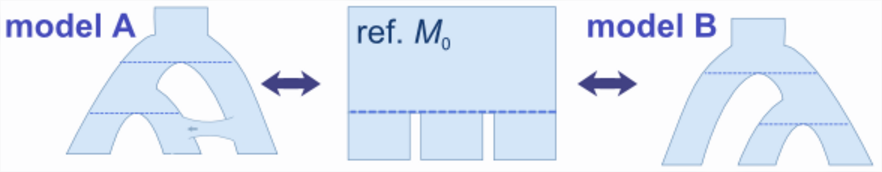
For example:

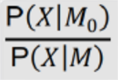


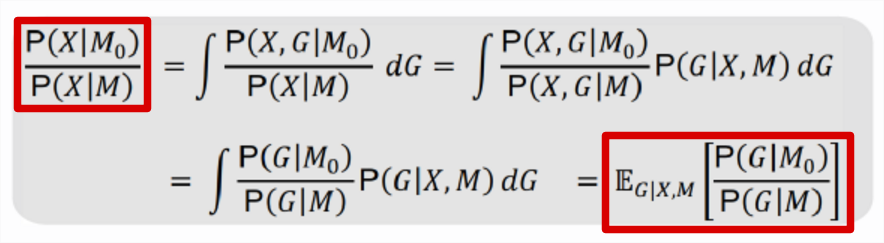
* **An attempt – standard harmonic mean**
  + One straight forward approach to the model-comparison problem would be to estimate the posterior likelihood of the model, P(X|M). This is hard to assess as X and G are only remotely related (via G). what we **can** do, using gphocs, is sample genealogies given the model and dna data. We use gphocs to estimate the last expression of the following -



* + This method suffers from one major disadvantage – since P(X|G) is very small, 1/P(X|G) is vary unstable, thus calculating its expectance would require a very large number of samples.
* **An improvement - Relative Bayes factors**
  + In an attempt to decrease instability, we will use a reference model M0 -



* + The expression we are trying to estimate -  - has a few advantages; It is much closer to 0 and much more stable. We can even choose a specific reference model M0 to minimize its variance.
  + To estimate the desired likelihood we will again rely on Gphocs’ existing framework of sampling genealogies -



**current work –**

For the past couple of months we’ve implemented and experimented with a small proof of concept.

* Simulation M1, M0. Running GPHOCS. Running modelcompare and validating results
* preliminary results

**open issues == future work**

* **Theoretical limitations of M0** –
  + **explanation:**
  + When calculating E[P(ZM0)/ P(ZM)] we are actually estimating – calculating a discrete average from sample in an estimated space. For some M0s, this estimation breaks, for example if some Genealogy z of M1 simply do not exist in M0. This would lead to P(zM0)/ P(zM) = 0.
  + A first goal of the following research would be to formalize the reference model –
    - What are the hidden assumptions on the model of E[P(ZM0)/ P(ZM)]?
      * For example for P(Z1,Theta|M0) should be well-defined
    - What Criteria must M0 meet to uphold these assumtions?
      * For example – any model M0 which gives non-zero probabilities to any genealogy in M1
  + Later for practical reasons we might choose lo limit ourselves only to M0s with a well-defined projection from M1
* **Difference in parametrization between models**
  + **explanation:**
  + M1 & M0 (& M2) will often have differing parametrizations. Usually M0 would have fewer params. These params manifest in the calculation E[P/P].
    - לקשר יותר טוב עם הנוסחא E[p/p] ועם הבעיה התאורטית. לאתר את ההנחות החביות
  + Currently we are ignoring priors completely.
    - What effect does this have?
    - Under what assumptions is this legal?
    - Can we still do this when also comparing with M2? Under what assumptions
  + A mile-stone in our work would be to develop a theory and technique to map\project parameters from M1 to M0
    - Option – make priors smarter by adding flags (binomial random vars) – isSplitPopulation, isMigrationStart, …
* **Supporting more than M0** –
  + **explanation:**
  + The core model-compare algorithm outputs a comparison between the main thesis model “M1” and a reference model “M0”. In an ideal world the algorithm would support **any** model as a reference model (instead of the current M0-only).

In practice this is not possible since we rely on existing GPHOCS calculations and these obviously do not apply to general population trees. This narrows the scope to population trees

Some ideas:

* + gphocs outputs in each iteration all genealogies. Freely choose M0 and use genealogies to exactly calculate gen-likelihood.
  + for each major event in the model (population split, mig start, mig end) defining a band in the model, output sufficient statistics for specific band. Create M0 through some reassembly of the bands and calculate gen-likelihood using sufficient stats on the bands.

פה לעבות, להדגיש את החשיבות, את זה שזה מוסיף לנו גמישות, את האתגר המדמ"חניקי, ושזה יהיה האתגר הראשון ואולי העיקרי.

* **Choosing a reference model**
  + Different M0s would give different variances to E[P/P]. *Benefit of good M0*
  + The choosing grows more complex when also considering M2. What is the best model to compare M1 & M2?
    - We need develop the theory and practice of this question.
  + A necessary ability to make model\_compare useful would be to formalize and implement an approach to finding the best M0
    - For example, exhaustive approach over legal M0s which measures the variance of E[P/P]

**Experiment road-map**

**Near Future**

1. Fill in some theoretical gaps.

Focus on one of the following

* 1. **Theoretical Limitations on M0**
  2. Supporting multiple M0s in single gphocs run
  3. Comparing models with different parameters

1. Expand implementation of gphocs & mcmcmc
2. Experiment – build tests, execute mcmcmc, predict expected results and compare with actual results
3. Repeat

את ההסבר ללמה תמיכה בריבוי M0 היא הכי חשובה למרכז התוכן:

We will start with supporting multiple M0s. this is a major technical achievement which will later allow give us flexibility in post processing. … . . . ……

TODO –

להוסיף פרק שמגדיר בצורה ברורה מה 'חוקי המשחק' של האלגוריתם שלנו (לדוגמא למה זה חוקי להתעלם מהpriors?)

להגדיר אנוטציות

לוגו לmcmcmc

להוסיף בהקדמה כמה נוסחאות יפות ושרטוטים של עצים

העיקר שההצעה תהיה ברורה – לא צריך בהחלט לומר מה תהיה הגישה לפתרון