Kai Pilger / Wikimedia

The Dilthey Lab

Computational immunogenomics

Graph-based genome inference

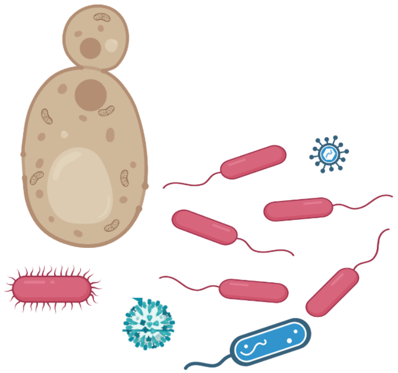
Long reads methods development

Sequencing-based diagnostics

→ Translation: Computer science and statistical modeling to generate biological insight!

Taxonomic classification and the EM algorithm

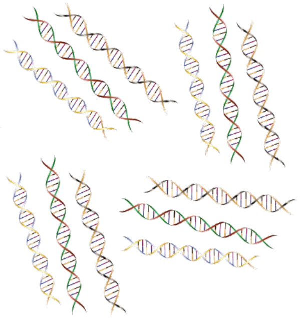
Taxonomic assignment



Sample
(e.g. from patient)

DNA extraction
(involves fragmentation)

Unordered pool
of (shorter) DNA fragments
from many organisms



Library
preparation
OR
16S amplification +
library preparation

Sequencing

Sequencing reads
(Sequences of sequenced DNA fragments)

Which organisms do these reads emanate from?

Why taxonomic assignment?

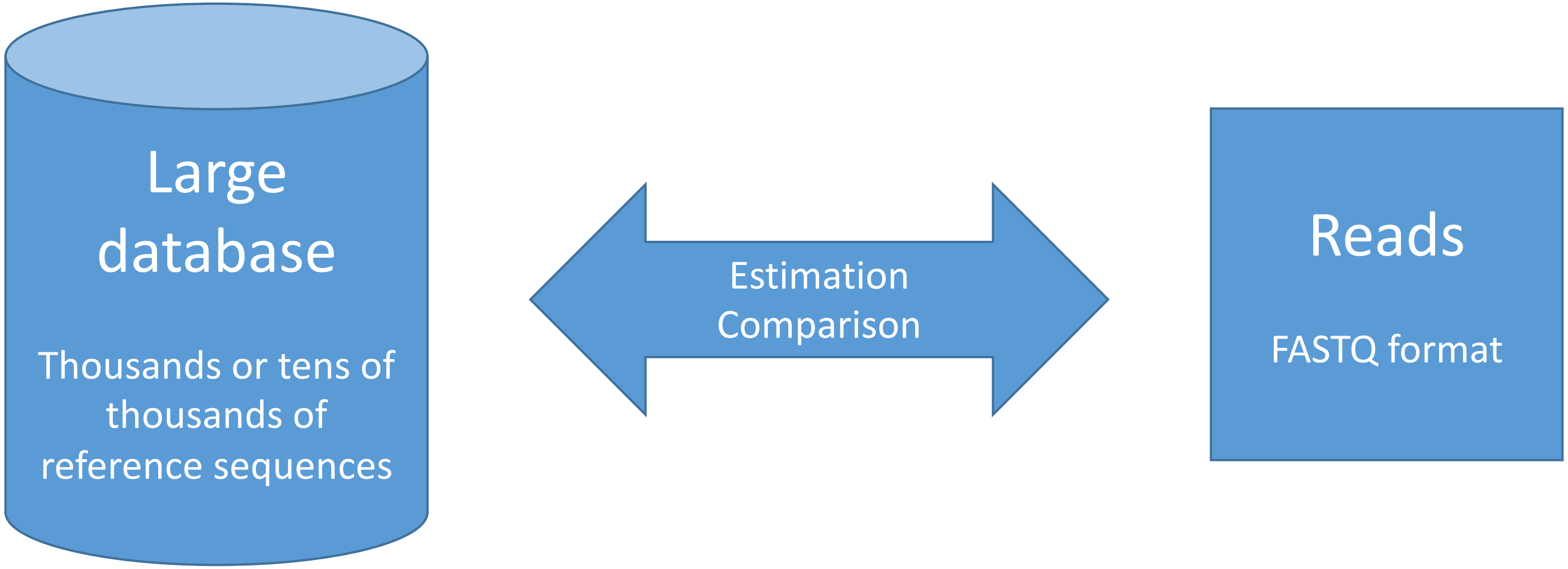
- Overall community composition

  „Which organisms are present in my sample, and in which relative abundances?“

- Investigating individual reads

  „Which organism does this individual read, which encodes an antiobiotic resistance gene, come from?“

Taxonomic assignment from 10,000 feet



Results:

- A „most likely“ taxonomic assignment for each read (i.e. an entry from „large database“), often with an assigned confidence score
- Alternatively: a probability distribution over taxonomic assignments for each read (general case)
- Community composition generally follows from summing over the values for individual reads.

Just BLAST!

Mapping the input reads against your input database is a reasonable first approach to taxonomic assignment.

Let's try it out...

Task [5 minutes]

You are given the following reference database:

```
>E. coli O157:H7
CGTGTTTAACGTTTAACTCTGCTTATTATTAAAAACAGGGCGAAACTTGCCCTGTTATCGCAACCCGCGC

>Klebsiella strain ABC
GACGCTGTGAAAGCAGACGCCAGGCCTCCTGCCAGCGGGCGTTAAACCGTTCGGGCTGACCTGCGCAATC

>Candida albicans SC5314
CCAACACCACTGCAACAACAACAACAACCACAATAGCTCCTCCAACTAGGGCGAAACTGAGTCTATACAA
```

... and the following reads:

```
@r1
ACTCTGCTTATTATTAA
@r2
CAGACGCCAGGCC
@r3
TTGCCCTGTTA
```

➔ Find out which references the reads belong to, and where they „map“ (exact matches)

Task [2 minutes]

We are given the same references and, in addition to the same reads as in the previous task, the additional read:

```
@r4
AGGGCGA
```

➔ Find out which reference the read @r4 belong to, and where it „maps“ (exact matches)

Issues

- What happens if we find multiple possible mapping locations per read?

- Relatedly, if we want to support non-exact matches (typically increasing the number of mapping locations per read), how do we deal with the fact that some matches are better than others (e.g., fewer mismatches)?

Occam's Razor

> *"Occam's razor is the problem-solving principle that recommends searching for explanations constructed with the smallest possible set of elements. It is also known as the principle of parsimony or the law of parsimony."*

(Wikipedia)

- .... so perhaps we just want to assume that @r4 comes from an *E. coli* genome?
- .... but what happened if we had 20 other *E. coli* reads, and 2 *Candida* reads?

Probabilistic read assignment

- In such instances, it is not possible to be 100% confident about where a read come from
- But we can set up a probabilistic model...

- Let R denote the set of observed reads and S the set of species for which we have reference genomes.
- Let O_r be a random variable that denotes the taxonomic origin of read $r \in R$ and let $o_r \in S$ be a specific value of that random variable
    - Interpretation: "$O_r = o_r$" $\Leftrightarrow$ "Read r emanates from species o_r"

- We are looking for a probability distribution $\mathrm{P}(O_r = o_r | r)$
    - Note: „Probability distribution“$\Leftrightarrow$ $\sum_{s \in S} \mathrm{P}(O_r = o_r | r) = 1$ and $\mathrm{P}(O_r = o_r | r) \geq 0 \; \forall \, s \in S$

- Bayes' Theorem: $\mathrm{P}(O_r = o_r | r) = \frac{\mathrm{P}(r | O_r = o_r) \times \mathrm{P}(O_r = o_r)}{\sum_{s \in S} \mathrm{P}(r | O_r = s) \times \mathrm{P}(O_r = s)}$, where:
    - $\mathrm{P}(r | O_r = s)$ is the probability of observing read r, conditional on r emanating from species s
    - $\mathrm{P}(O_r = s)$ is the so-called „prior“ probability of a read emanting from species s

$$\boxed{\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A) \times \mathrm{P}(A)}{\mathrm{P}(B)}}$$

Probabilistic read assignment

$\mathrm{P}(r|O_r = s)$:

- In a world without sequencing errors, we could generate reads from the genome of species s by:
  - Selecting a read length l
  - Uniformly selecting a possible start position j within the reference sequence of s, conditional on l
  - The sequence of then would then be equal to the subsequence of the reference of s from positions $s\,..(s + l\ - 1)$
  - I.e. $\mathrm{P}(r|O_r = s) = P\big(\text{length}(r)\big)\times\frac{1}{\text{length}\big(\text{sequence}(s)\big)-\text{lengt}h\,(r)+1}$ if there is an exact match between r and the reference sequence of s and 0 otherwise.

- In practice, we often ignore length differences in reference genomes and do not explicitly model read lengths. Hence, we often set $\mathrm{P}(r|O_r = s) = 1$ henever there is an exact match between r and the reference sequence of s, and 0 otherwise. [The result of this is an improper probability distribution, but this often does not matter]

- However, in a world of inexact matches, we often do care about the quality of the match between r and s, and hence modify $\mathrm{P}(r|O_r = s)$ to take this quality into account (smaller values correspond to lower alignment qualities).

Probabilistic read assignment

$\mathrm{P}(O_r = s)$:

- „Prior“ = Probability of sampling a read r from s, prior to observing the actual sequence of r

- In a generative model for read generation, this is, when sampling a read, the probability of selecting species for generating the actual read sequence. That is, $\mathrm{P}(O_r = s)$ is the abundance of species s in our sample!

- $\mathrm{P}(O_r = s)$ needs to sum to 1 and we let F denote the „composition vector“ of prior probabilities.
  - I.e.: for all $s \in S$, $F_s \coloneqq \mathrm{P}(O_r = s)$, $\sum_{s \in S} F_s = 1$ and $F_s \geq 0 \;\forall\, s \in S$

- When we analyze a sample, F is not generally known (the composition vector is often what we want to infer)

Probabilistic read assignment

- In the following, we use the notation $\mathrm{P}(O_r|r) = \mathrm{P}_r(O_r|F)$ for notational clarity.

- I.e. $\mathrm{P}_r(O_r = o_r|F)$ is the probability that read r emanates from species o_r, conditional on sample composition F.

- $\mathrm{P}_r(O_r|F) = \frac{\mathrm{P}(r|O_r = o_r) \times F_{o_r}}{\sum_{s \in S} \mathrm{P}(r|O_r = s) \times F_s}$ (this is just re-stating our earlier definitions)

Maximum likelihood

- How do we go about the fact that F is typically unknown and our object of interest?

  ➔ We can try to learn or estimate F from the data!

- Conditional on F, the probability of observing a specific read r is $\mathrm{p}(r|F) = \sum_{s \in S} \mathrm{P}(r|O_r = s) \times \mathrm{P}(O_r = s)$

- As reads are independent, the probability of the read set R is $\mathrm{p}(R|F) = \prod_{r \in R} \mathrm{p}(r|F)$

- When we treat $\mathrm{p}(R|F)$ as a function of F, we use the notation $\mathrm{L}(F) \coloneqq \mathrm{p}(R|F)$.

  $\mathrm{L}(F)$ is called the „likelihood function".

- „Maximum likelihood": Find the value $\hat{F} = \underset{\left\{F \in \mathbb{R}^{|S|}:\ \sum_{s \in S} F_s = 1 \text{ and } F_s \geq 0 \forall s \in S\right\}}{\operatorname{argmax}} \mathrm{L}(F)$

  $\hat{F}$ is called the „maximum-likelihood estimate"of F.

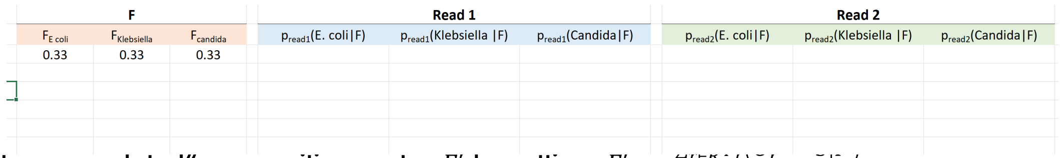
- How easy or difficult is it to find $\hat{F}$? For large values of $|S|$, it can become computationally difficult.

Task [10 minutes]

1. Start with an initial guess for F: F_S = 1/3 for all species $s \in S$.

2. For each of the 4 reads, compute $P_r(O_r = s|F)$ for all species $s \in S$ (resulting in 12 values in total: 4 reads x 3 species)
    1. We use $\mathrm{P}(r|O_r = s) = 1$ whenever there is an exact match between r and the reference genome of s, and 0 otherwise

3. Set up a simple spreadsheet (Excel or Google Sheets) with 3 + (4 x 3) columns. To fill the first row of that spreedsheet,
    1. Fill the first 3 columns with the current values of F
    2. Fill the next 3 columns with the values $P_r(O_r = s|F)$ for the first read, in the same order of species you also used for the first 3 column
    3. Fill the next 3 columns with the values $P_r(O_r = s|F)$ for the second read...



F			Read 1			Read 2		
$F_{E\ coli}$	$F_{Klebsiella}$	$F_{candida}$	p_{read1}(E. coli|F)	p_{read1}(Klebsiella |F)	p_{read1}(Candida|F)	p_{read2}(E. coli|F)	p_{read2}(Klebsiella |F)	p_{read2}(Candida|F)
0.33	0.33	0.33						

4. Compute an „updated“ composition vector F' by setting $F'_s = \frac{\sum_{r \in R} P_r(O_r = s|F)}{4}$

5. Set F = F', go back to Step 2, and fill the next row of the spreadsheet.

- ➔ Do 3 rounds of this (i.e. until you have filled four rows of the spreadsheet)
- ➔ Ideally use formulas instead of hard-coding the values in the spreadsheet
- ➔ Observe what happens with @r4
- ➔ What would happen if @r2 also mapped to *E. coli* (instead of *Klebsiella*)?
- ➔ What would happen if we used a different initial guess for F?

The EM algorithm

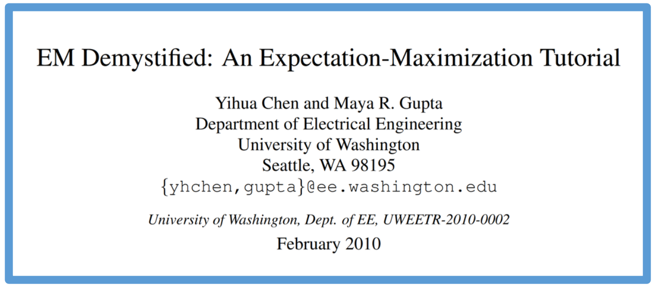
- The EM algorithm is an approach that can be used for the optimization of $\mathrm{L}(F)$.

- Key idea: Assume that, for the inference problem at hand, there exists a set X of „complete" data that one wishes one had to tackle the inference problem; Y, the observed data, need to be related via a deterministic function T, i.e. Y = T(X); if likelihood inference becomes easier to tackle if one assumes X is known, then EM may be a good approach.
    - In our case, the „complete" data include o_r, i.e. the taxonomic origin of each read, i.e. $\mathrm{X} = \{(r, o_r)\}\ \forall\, r\ \in R$
    - o_r is not actually observed (this is the trick!) – but we can make a probabilistic guess of o_r that we iteratively improve

<u>EM algorithm:</u>

- Let $\theta \in \Theta$ be the parameters we want to optimize.
- In order to apply EM, we need a density $\mathrm{P}(x|\theta)$ and a density $\mathrm{P}(X|\theta, y)$
- The trick is to start with an initial estimate $\theta^{(1)}$, fix this estimate, and find a new value $\theta^{(2)}$ by maximizing the function $\mathrm{E}_{x\sim P(x|\theta^{(1)},y)} \log P\left(x\middle|\theta^{(2)}, y\right)$
- ... and then iterate.

The EM algorithm



EM Demystified: An Expectation-Maximization Tutorial

Yihua Chen and Maya R. Gupta
Department of Electrical Engineering
University of Washington
Seattle, WA 98195
{yhchen,gupta}@ee.washington.edu

University of Washington, Dept. of EE, UWEETR-2010-0002
February 2010

Step 1: Pick an initial guess $\theta^{(m=0)}$ for θ.

Step 2: Given the observed data y and pretending for the moment that your current guess $\theta^{(m)}$ is correct, calculate how likely it is that the complete data is exactly x, that is, calculate the conditional distribution $p(x \mid y, \theta^{(m)})$.

Step 3: Throw away your guess $\theta^{(m)}$, but keep Step 2's guess of the probability of the complete data $p(x \mid y, \theta^{(m)})$.

Step 4: In Step 5 we will make a new guess of θ that maximizes (the expected) $\log p(x \mid \theta)$. We'll have to maximize the *expected* $\log p(x \mid \theta)$ because we don't really know x, but luckily in Step 2 we made a guess of the probability distribution of x. So, we will integrate over all possible values of x, and for each possible value of x, we weight $\log p(x \mid \theta)$ by the *probability of seeing that* x. However, we don't really know the probability of seeing each x, all we have is the guess that we made in Step 2, which was $p(x \mid y, \theta^{(m)})$. The expected $\log p(x \mid \theta)$ is called the Q-function:[3]

$$Q(\theta \mid \theta^{(m)}) = \text{expected } \log p(x|\theta) = E_{X|y,\theta^{(m)}}[\log p(X \mid \theta)] = \int_{\mathcal{X}(y)} \log p(x \mid \theta) p(x \mid y, \theta^{(m)}) dx, \qquad (2.3)$$

where you integrate over the support of X given y, $\mathcal{X}(y)$, which is the closure of the set $\{x \mid p(x \mid y) > 0\}$. Note that θ is a free variable in (2.3), so the Q-function is a function of θ, and also depends on your old guess $\theta^{(m)}$.

Step 5: Make a new guess $\theta^{(m+1)}$ for θ by choosing the θ that maximizes the expected log-likelihood given in (2.3).

Step 6: Let $m = m + 1$ and go back to Step 2.

Applying EM to taxonomic classification

- $\mathrm{y} \Leftrightarrow R$
- $\mathrm{X} \Leftrightarrow (R, O) = \{(r, O_r)\}\ \forall\, r\ \in R$ (i.e. O is the set of taxonomic origins of all reads)
- $\log P(X|\theta) \Leftrightarrow \sum_{r \in R} \log \mathrm{p}\,(r, O_r|F)$
- $\mathrm{E}_{X|y,\theta^{(m)}} \Leftrightarrow \mathrm{E}_{(R,O)|R,F^{(m)}} = \mathrm{E}_{O|R,F^{(m)}}$ *{i.e. we take the expectation w.r.t. $P_r(O_r|F)$ for each read}*

What about $\log \mathrm{p}\,(r, O_r|F)$?

This is simple, assuming a specific value o_r:

$$\log \mathrm{p}\,(r, O_r = o_r|F) = \log\big(\mathrm{P}(r|O_r = o_r) \times F_{o_r}\big) = \log \mathrm{P}(r|O_r = o_r) + \log F_{o_r}$$

Applying EM to taxonomic classification

Our goal is now to maximize $\mathrm{E}_{O|R,F^{(m)}} \sum_{r\in R} \log \mathrm{p}\left(r, O_r \middle| F^{(m+1)}\right)$ as a function of $F^{(m+1)}$:

$$\mathrm{E}_{O|R,F^{(m)}} \sum_{r\in R} \log \mathrm{p}\left(r, O_r \middle| F^{(m+1)}\right) = \sum_{r\in R} \mathrm{E}_{O_r|r,F^{(m)}} \log \mathrm{p}\left(r, O_r \middle| F^{(m+1)}\right) = \sum_{r\in R} \sum_{o_r\in S} \log \mathrm{p}\left(r, O_r \middle| F^{(m+1)}\right) \times P_r\left(o_r \middle| F^{(m)}\right)$$

$$= \sum_{r\in R} \sum_{o_r\in S} \left[\log \mathrm{P}(r|O_r = o_r) + \log F_{o_r}^{(m+1)}\right] \times P_r\left(o_r \middle| F^{(m)}\right)$$

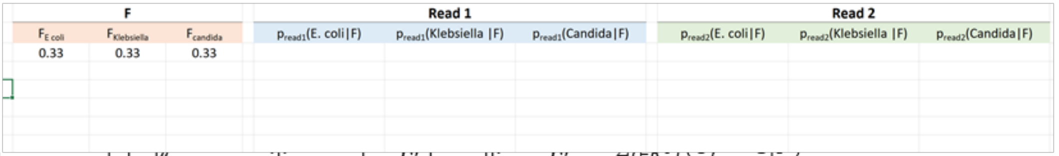
$$= \sum_{s\in S} \left(\log F_s^{(m+1)} \times \sum_{r\in R} P_r\left(s \middle| F^{(m)}\right)\right) + \sum_{r\in R} \sum_{o_r\in S} \log \mathrm{P}(r|O_r = o_r) \times P_r\left(o_r \middle| F^{(m)}\right)$$

… which is maximized by $F_s^{(m+1)} = \frac{\sum_{r\in R} P_r\left(s \middle| F^{(m)}\right)}{|R|}$

1. Start with an initial guess for F: $F_s = 1/3$ for all species $s \in S$.

2. For each of the 4 reads, compute $P_r(O_r = s|F)$ for all species $s \in S$ (resulting in 12 values in total: 4 reads x 3 species)
    1. We use $\mathrm{P}(r|O_r = s) = 1$ whenever there is an exact match between r and the reference genome of s, and 0 otherwise

3. Set up a simple spreadsheet (Excel or Google Sheets) with 3 + (4 x 3) columns. To fill the first row of that spreadsheet,
    1. Fill the first 3 columns with the current values of F
    2. Fill the next 3 columns with the values $P_r(O_r = s|F)$ for the first read, in the same order of species you also used for the first 3 column
    3. Fill the next 3 columns with the values $P_r(O_r = s|F)$ for the second read…



F			Read 1			Read 2		
$F_{E\ coli}$	$F_{Klebsiella}$	$F_{candida}$	p_{read1}(E. coli|F)	p_{read1}(Klebsiella |F)	p_{read1}(Candida|F)	p_{read2}(E. coli|F)	p_{read2}(Klebsiella |F)	p_{read2}(Candida|F)
0.33	0.33	0.33						

4. Compute an „updated" composition vector F' by setting $F'_s = \frac{\Sigma_{r\in R} P_r(O_r = s|F)}{4}$

5. Set $F = F'$, go back to Step 2, and fill the next row of the spreadsheet.

E $P_r\left(o_r \middle| F^{(m)}\right)$

M $F_s^{(m+1)} = \frac{\sum_{r\in R} P_r\left(s \middle| F^{(m)}\right)}{|R|}$

- ➔ Do 3 rounds of this (i.e. until you have filled four rows of the spreadsheet)
- ➔ Ideally use formulas instead of hard-coding the values in the spreadsheet
- ➔ Observe what happens with @r4
- ➔ What would happen if @r2 also mapped to *E. coli* (instead of *Klebsiella*)?
- ➔ What would happen if we used a different initial guess for F?

Using an alignment-quality aware read likelihood

- What if we want to use a more sophisticated approach to computing $\mathrm{P}(r|O_r = s)$, e.g. one that allows for mismatches between r and the reference genome of s?

- This will have an effect on $\mathrm{P}_r(O_r|F) = \frac{\mathrm{P}(r|O_r = o_r)\times F_{o_r}}{\sum_{s\in S}\mathrm{P}(r|O_r = s)\times F_s}$; everything else remains unchanged.