Bacterial Genomics: Molecular Evolution at the Level of Ecosystems

Lecture by Eric Alm on December 3rd 2010

Scribe Notes

Overarching ideas: The microbiome as an ecosystem inside the human body.

Data source: the Human Microbiome Project (HMP) generated much of the data referred to in this lecture

Relevant numbers:

-10 fold more bacterial cells in the human body than human cells $O(1e14)$
-100x more distinct genes than in the human genome

Goal: model the microbial ecosystem. This problem can be divided into several parts as follows:

1) Survey: survey by high throughput sequencing of 16S ribosome, barcodes for species: who is present and where?
2) Rules: Analysis of rules/modules/motifs/clusters in ecosystem occupancy patterns as a function of space and time
3) Rules and motifs $\rightarrow$ ODE model (species as a function of time)
4) Incorporate spatial structure $\rightarrow$ PDE model (time and space)

Further details on ODE modeling: As an example of a model in step 3, consider a simple coupled ODE model for the population of an e.coli species within a niche in the microbiome:

$$N_i = e_{coli}$$

The differential is proportional to abundance of species times some other function of the abundance of other species and a lumped parameter $E$ for environmental factors.

$$a_{ij} = \text{species-species interaction term}$$

$$r_i = \text{intrinsic growth rate}$$

$$K_j(E(t)) = \text{carrying capacity for species } j$$

Further details on data generation process:

- extract DNA from body site
- amplify 16S rDNA to get prokaryotic DNA marker
- sequence this to get large # of “molecular barcodes” (don’t talk about “species”)
-instead of species classify (cluster) into OTU (operational taxonomic unit)
-get out vector of abundances of the different “species” (normalized counts of OTUs)

**Species-environment associations:** Applications in disease diagnosis (look at bacteria to diagnose human physiology) and natural “biosensors”.

Current unsupervised learning approach:

-PCA (principal components analysis) on the microbial abundance vector space
-cluster in the reduced dimensional space

Limitations of this approach:

-the resulting clusters correspond reasonably well to a few different environments, but have trouble distinguishing MANY factors – not getting data out that is as rich as the data itself

-PCA always gives the same plot – clustering always gives the same clusters! No way to incorporate external information. This is a major limitation.

SLIME (supervised learning in microbial ecology) is a supervised learning approach which overcomes the above limitations. Here the axes also become biologically meaningful, unlike in PCA.

Challenges:

-large inter-patient variation
-low abundance discriminative taxa

Techniques:

- Random forests
- Support vector machines

General rule = avoid unsupervised learning when know categories already

HMP body part study using SLIME:

18 body sites on 250 individuals – 16s rDNA sequences

--high discriminative accuracy between closely spaced sites
--consider supperingival plaque vs. subgingival plaque (above or below gumline):

73 percent cross-validated accuracy at predicting differences

key bacteria that distinguish are: strict anaerobic vs. facultative anaerobic species (facultative = can tolerate oxygen)
makes sense – more oxygen above the gumline vs. below

Vaginal sites: dominated by lactobacillus, except that some patients have much fewer lactobacillus and higher abundance of other species – interestingly, these were perfectly healthy patients – but these bacteria were the characteristic ones of bacterial vaginosis... therefore, the bacteria population could be a normal state of healthy people who, however, have a greater propensity to get bacterial vaginosis for OTHER reasons → the bacteria may not be the causative factor

**Pediatric Inflammatory Bowel Disease: collaboration w/ Childrens hospital**

(Athos Bousvaros)

goal = diagnose without colonoscopy

Random forest classifier on the species with known categories sick vs. non-sick patients with large additional sources of inter-patient variability

Note: “Shannon diversity” of an ecosystem (assuming this is the same as the Shannon entropy measure on the species occupancies). As patients become more diseased, ecosystem diversity decreases. Could be the ecosystem as a whole that is controlling disease status, not one particular species!

-effect of sampling depth

What about unsupervised learning approaches?

**Topic models: basic concepts**

-plaques divide into contributions from sub-gingival, supra-gingival and rest of mouth = basic idea is assume underlying sub-populations or sources of a sample

-infer probability of a sample from a given topic model of that sample

-run maximum likelihood methods to find best topic model for a sample set

Strongest prediction: one individual female has a different topic model – turns out to come from chloroplasts (some sort of face product?)

Bioremediation study at Oak Ridge NL: need to clean up chromium

-identify taxa responsive to nitrate

-nitrate inhibits Cr bioremediation

-infer environmental conditions from DNA

-want to identify shift in microbial community structure as fn of nitrate concentration
-available data gives rise to tricky prediction problem

----Dessulfovibrio are predictive of nitrate level. These are the only ones.

----Strongly suggests dessolfovibrio species are playing a key role in the bioremediation process.

----Can use DNA sequences give a very good regressor to nitrate concentration

Species-species interactions:

co-occurrence networks:

find species-species correlations across different environment

currently: lots of strong correlations but the involved species appear different between different techniques

need to normalize data because of experimental sample-sample variability: this totally changes the nature of the analysis problem (compositional effects)

also see anti-correlations: can draw in a colored graph

graduate student wanted to prove existence of compositional effects:

so scrambled the data (lost sample-identity information) and got the SAME correlations

after all... Pearson said don't use his correlation coefficient for rational functions of the variables you're interested in, such as frequency:

\[ f_i = \frac{\text{abundance}_i}{\sum_i (\text{abundance}_i)} \]

the basic idea is that you're seeing spurious positive and negative correlations when you deal with percentages – they tell about overall abundance of species but not about correlated “fluctuations” in bacterial abundances

solution: use log frequency ratio matrix \( y_{ij} = \log(f_i/f_j) \) and compute the variance of each element in this matrix. Overall abundances cancel out in the ratio. Can relate this variance matrix to the species-species covariance matrix which is what we're really interested in! (this is assuming many components with sparse correlations)

what we understand now is that with the pearson correlation there is a low **effective number of species**. Only with high #s of species do we get something meaningful.

But with sparse correlation inference method, can infer the right correlations even at low effective number of species.
interesting observation: something as simple as normalizing your data and taking a correlation is an open problem in this field

**Time series models: moving to ODEs**

Experiment: inject new bacteria and see how population dynamics is affected

HuGE network: effects of environmental factors (sleep, happiness...) on gut ecosystem

Horizontal gene transfer (HGT) in the human microbiome: maybe the ODEs totally break down due to gene exchange

--looking only at the most recent gene transfers

--want to identify 100% identical DNAs in unrelated species

--there is a significant amount of identical DNA at all levels of phylogenetic distance

--inside human genomes there is MORE hgt: almost all hgt in genbank is due to bacteria isolated from human bodies, and much is happening in the same site of the body (in different people but same SITES) – outside the body human food also has a lot of DNA swapping (human food to gut transfer)

--highest levels of swapping are from the same micro-environment

--also see exchange from commensal bacteria to pathogens and pathogen to pathogen → innovations in pathogens are rapidly shared