Canadian Bioinformatics Workshops

www.bioinformatics.ca
Module 6
Microbiome biomarker discovery

John Parkinson on behalf of Thea Van Rossum, Anamaria Crisan, Mike Peabody & Fiona Brinkman
Analysis of Metagenomic Data
June 24-26, 2015

What are biomarkers and how do we find them?

bi·o·mark·er
ˈbīōˌmärkər/

Measureable biological property that can be indicative of some phenomena, such as an infection, disease, or environmental disturbance

Functional biomarkers

Biological functions that may be specific to a single organism or shared among multiple organisms

Taxonomic biomarkers

Can be a specific species, but most often is a category of organisms – also called Operational Taxonomic Unit (OTU).

Anamaria Crisan
What are biomarkers and how do we find them?

**DISCOVERY**
- Bioinformatics Software - Takes the raw digitized genomic data and performs QC and biomarker quantification
- Use math – applied statistical methods can help us find useful biomarkers

**VALIDATION**
- Design Primers – biological “hooks” that pick out our biomarker of interest from a sample
- qPCR – measures how many times primers (hooks) manage to snag our biomarker of interest

Anamaria Crisan
### What will put the “bio” in biomarker?

**A) Bacteria**
- Shotgun or 16S amplicon
- Best studied, most methods developed

**B) Viruses**
- Shotgun or amplicon (RdRp, g23)
- Can be challenging to get enough DNA
- Host-specificity and population “bursts” hold promise

**C) Eukaryotes**
- Amplicon (18S, ITS) (large genomes make shotgun difficult)
- Best studied, most methods developed

---

### What kind of biomarker do we want?

**A) TAXONOMIC**
- Can use amplicon or shotgun data
- Strain-level diversity can lead to false positives/negatives
- More variable across environments (for better or worse)

**B) GENE-BASED**
- Requires shotgun data (DNA or RNA)
- Need good sequencing depth to reach specialised genes
- Domain-based gene architecture can be tricky

**C) OTHER?**
- Diversity metrics, using microbiome analysis to suggest other metabolic markers, etc

---

*Thea Van Rossum, Fiona Brinkman*
What is biomarker selection?

The process for removing non-informative or redundant OTUs from an analysis.

Anamaria Crisan, Fiona Brinkman

What makes a good biomarker?

Good biomarkers are those with:
- Class means (i.e. average OTU abundance in each group) that are far apart
- Tight variance (i.e. consistent OTU abundance in each group)
- Abundance across samples follow a normal distribution

OTU1

OTU2

Abundance

Tight Variance (little overlap)

Anamaria Crisan

Lots of overlap
What makes a good biomarker?

In this example we want to find biomarkers that separate between the red and blue class labels.

<table>
<thead>
<tr>
<th>OTU</th>
<th>Blue Measures</th>
<th>Red Measures</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✔️</td>
<td>✔️</td>
<td>Clear difference, consistent size</td>
</tr>
<tr>
<td>2</td>
<td>❓</td>
<td>☑️</td>
<td>Inconsistent, not sure.....</td>
</tr>
<tr>
<td>3</td>
<td>❌</td>
<td>☑️</td>
<td>No difference at all</td>
</tr>
</tbody>
</table>

Anamaria Crisan

What makes a good biomarker?

1. How can biomarkers add value to current testing procedures?

2. How can we make biomarkers easily accessible to those that routinely test water quality?

Anamaria Crisan
Biomarker selection relies on statistical techniques

- These techniques can range from the very simple (like a t-test) to more complex
- You can write your own statistical analysis (using programs like R, Matlab, Python, STATA, SAS etc.)
- OR you can use more complex methods developed by others:
  - Popular metagenomics / microbiome methods include:
    - LEfSE
    - metagenomeSeq

- Whatever you choose to use it’s important to understand the statistical methods especially:
  - Your assumptions about the data
  - The statistical method’s assumptions about the data
  - The statistical method’s limitations
  - How to interpret your results from the output of the statistical method

Biological Data is difficult

- Most biological data is correlated and, especially with microbial community data, quite sparse
- Statistical methods vary with respect to how correlation and data sparseness are taken into account.

- So how do others deal with the problem?
  - They ignore correlation and use only parametric methods (most common approach); use a hard filter to look at only abundant OTUs
  - Use more complex approaches that take into account correlation and sparseness
    - BUT it can be more computationally intensive and time consuming to use these methods
    - AND it can be difficult to understand and interpret what the method is doing
So you might be wondering … what are these mysterious “statistical methods”?

Generally, statistical techniques either try to predict labels or continuous values.

**Classification**
- Sample 1
- Sample 2
- Sample 4
- Sample 3
- Sample 5

Classification attempts to predict the label, or class, of some sample. A common example would be classifying sick and healthy patients.

**Regression**
- Over-estimate
- Ideal
- Under-estimate

Regression attempts to predict the future value for some variable. Common example would be attempting to predict tomorrow’s stock prices.
And...they also generally belong to one of two categories

**Supervised**

- Sample 1
- Sample 2
- Sample 3
- Sample 4
- Sample 5

Samples come from **known classes**

Use knowledge of the classes in a training set to optimize the model's ability to separate classes. Evaluate the generalizability by using a test set

**Unsupervised**

- Sample 1
- Sample 2
- Sample 3
- Sample 4
- Sample 5

**Don't know classes** that samples belong to or how many there are

"Letting the data drive" - using biomarker data to find patterns or structure in the data to define classes by clustering similar patterns

(I lied a little : Semi-supervised methods also exist) Anamaria Crisan

Module 6

bioinformatics.ca

**Supervised**

<table>
<thead>
<tr>
<th>OTU 1</th>
<th>OTU 2</th>
<th>OTU 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample 5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Advantages**

- Far easier and faster to do
- Allows one to create a simpler study design, so biomarkers may be more robust and relevant
- Biomarkers are easy to validate because success depends on ability to separate classes

**Disadvantages**

- Classes may not be well defined, so it's difficult to find biomarkers that support the pre-defined classification

Anamaria Crisan

Module 6

bioinformatics.ca
Note, the same techniques are called different things in different fields

<table>
<thead>
<tr>
<th>Glossary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine learning</td>
</tr>
<tr>
<td>network, graphs</td>
</tr>
<tr>
<td>weights</td>
</tr>
<tr>
<td>learning</td>
</tr>
<tr>
<td>generalization</td>
</tr>
<tr>
<td>supervised learning</td>
</tr>
<tr>
<td>unsupervised learning</td>
</tr>
<tr>
<td>large grant = $1,000,000</td>
</tr>
<tr>
<td>nice place to have a meeting:</td>
</tr>
<tr>
<td>Snowbird, Utah, French Alps</td>
</tr>
</tbody>
</table>

Anamaria Crisan
The best approach to use depends on your research

Now that we have biomarkers, we should validate them.
How do we validate our biomarkers?

• Once you ID the gene or taxonomic group you want to use as a biomarker, you need to design a test for it

(q)PCR is a good option:

1. Identify biomarker specific sequence
   – If used marker-based tool to identify biomarker, you’re all set (e.g. MetaPhlAn2)
   – Otherwise, can cluster reads or align them to find conserved sequences
     – need to verify that “representative” sequence is still a good biomarker
2. Design primers (& probe)
   – PrimerProspector: designs primers from a sequence alignment
   – PrimerBLAST: designs primers specific to a clade

Case Study – Categorical Biomarker from Bacterial Shotgun Data

Example of fast track to marker identification and PCR test development – IDing markers of water quality
(see www.watersheddiscovery.ca for project)

• Bacterial shotgun data using Illumina HiSeq platform
• Comparing riverwater microbiomes at an agricultural unimpacted site (AUP) versus two impacted downstream sites (“at pollution” (APL) & “downstream” (ADS))

• Using MetaPhlAn
  – Low sensitivity, high precision
  – Based on clade-specific gene sequences
  – 3000 reference genomes
  – Fast: 3 million reads (100-150 bp) in 10 minutes
1. Processed and validated data

Quality trimmed and normalised data across samples

**MOCK COMMUNITY VALIDATION**
- Validated with mock community: DNA-free water spiked with DNA from lab-cultured bacteria
- 7% of reads were assigned to a species (low sensitivity)
- Of those, 84% were correctly assigned

Thea Van Rossum

2. Identified differential taxa

**Taxonomic Classifications – Rural Site**
- Prioritised high abundance taxa
- Use White’s non-parametric t-test with false discovery rate multiple test correction to find differentially abundant taxa (alternative: RandomForests)

Thea Van Rossum
3. Identified sequences characteristic of taxa

- 57,016 reads assigned to Taxon 1
- 2,176 reads assigned to Taxon 2
- Prioritised taxon 1 due to our research indicating high abundance taxa are most accurately predicted

- Extracted taxon-specific sequences from Metaphlan database
  - 607 sequences for Taxon 1

- Aligned reads against these sequences
- Chose regions of Metaphlan sequences with most hits

Thea Van Rossum, Fiona Brinkman
4. Designed qPCR primers and probes from marker sequence

- Used Primer3 for primer and probe design
- Checked relative occurrence rates of candidate primers and probes
  - Considered matches that are exact or have 1-2 mismatches
  - Chose sequences that minimize non-specific matches
- Confirmed we can amplify a product of the right size

Next: iteratively test & improve qPCR

Thea Van Rossum

“Fast track” Case Study Summary

- Initial analysis identified a PCR marker based on differential abundance of a bacterial species that is being used to pilot our iterative validation process
- Benefits of this approach
  - Fast: sequence data to PCR primers in a couple days
  - Simple: does not require large amounts of processing power
- Limitations of this approach
  - Depends on differential abundance of known bacteria (if the differential bacteria aren’t highly similar to those in the Metaphlan database, this approach will not work)
  - Depends on taxa, which have been shown to be more variable across environments than (gene) functions
- This is a “low hanging fruit” approach; a good first step

Thea Van Rossum, Fiona Brinkman

Module 6
bioinformatics.ca
Bacterial Shotgun – alternative methods

• More complete taxonomic analysis
  – Kraken, Discriminate
• Gene function analysis
  – MEGAN4 (SEED, KEGG)
  – RealtimeMetagenomics (FIGFams)
• Cluster based analysis:

Get predicted proteins  Cluster  Find differential clusters  Design PCR

From Biomarker to Lab Test

Identify discriminative taxa or functions

Identify informative region for primer design (CD-HIT)

Design primers (Primer-BLAST or IDT Realtime PCR Tool)

Validate primers in silico (Primer Prospector, Primer-BLAST)

Validate primers in vitro (qPCR)
Remember other markers…

Community diversity could be a useful indicator of ecosystem health

Microbiome analysis may suggest other types of screening tests… (metabolites, etc)

*Markers are only as good as the data they are based on, so design experiments carefully, include +ve and -ve controls*

Anamaria Crisan, Fiona Brinkman

We are on a Coffee Break & Networking Session