SBI2$_{N}otes$

yusuf.roohani

September 13 -15 2017, San Diegooooo

Contents

Educational courses

1 High content phenotpying

1.1 Target based approaches versus phenotypic screening

- Phenotypic based on a biological trait of the system.
- Target based screen in HCS example - translocation of a protein from cytoplasm but even then that needs to be linked back to a phenotype
- On the other hand, very broad phenotypes can make it hard to define targets

1.2 Assay development guidelines

- Biology in a format that can be measured accurately

Figure 1: San Diego Marina

- Have control substance or condition
- What assay window do you have and how variable is your phenotype
- Is there a need for more replicates to increase statistical power
- Is the library appropriate - gene based/small molecule
- Understand assay endpoint but don't ignore what the data is telling you
- Don't use single measures of shape
- Be careful using ratios
- Be careful using complex methods you don't understand or can't explain

1.3 Data distribution

Is endpoint continuous or discrete

1.4 Data Normalization

- Percentage of control
- Scale phenotpye measure zero to one

- Z-score - is the change in phenotype significant compared with variability in control
- KNIME offers many other solutions

If there's a lot of variability in the biology it's good to have a lot of negative controls

Choices of normalizing between a plate or a run, but must be careful

1.5 How big of a difference do I need

- What is the assay window - Z', SSMD, S/N
- Do I know how large a chance is biologically relevant
- rule of thumb for z-score +- 3

1.6 Hit calling

Examples

- Cell death survivial
- Protein translocation
- Shape descriptor
- Count of cellular compartments

The one thing in common is that there's a single readout that determines hit calling. But we want to look at multivariate hit calling

1.7 Lymphangiogensis example

presented by someone from Novartis, Basel

Quantify the formation of lymph vessels. Used a univeariate threshold - sprout length (Why - biological prior knowledge)

Usupervised model (SOM) - for a qualitative assessment of the experiment. Very similar to tSNE. Data points with similar descriptors are grouped in adjacent areas of the map if the map doesnt clearly separate out the controls then the assay is probably not well defined.

Supervised component (RF) - classification

1.8 Another example from literature

RNAi screens for Rho GTPase regulators of cell shape and YAP/TAZ localisation in triple

Used Columbus to measure 126 features to describe phenotype

- Cell regions intensity and morphology
- YapTaz localization ring regions compared to nucleus and some others

Used a linear classifier to group cells with interesting phenotypes

Had 5 groups observed within the dataset (arbitrarily defined)

- spindly, large, triangular, fan, small/round
- classes were determined using prioir knowledge of biology

Took the z-score across the PCA of the features for picking hits

What features are important to classify phenotype? Used hierachical clustering to pick out important clusters for different 'clusters'

Nilgen Tasedmir, Cancer Discovery 2016 [4]

2 Systems biology

Search for the incredibly important novel phenotype

Data can be used to identify hidden relationships - important starting points for hypothesis based studies

Skill sets for hich content screening: - Biology - FLuoresence microscopy - Informatics - Instrumentation - Data Analysis

Images and data - What can be measured/quantified?

Eg: for nuclei - size (area) - shape (elongation) - variation (punctuate regions) - Intensity

Doe ficwn qwll, features can be summaries using mean, median, mode, max, min

/textbfDistributions

We are trained to assume gaussian distributions.

But that's not always true, the skewness itself can change with treatment conentration

heterodascicity [2]

2.1 How much data do we have - what does the aggregated data look like

A lot, especially if looking at the cell level instead of the well level

96 wells 1400 cells per row 12 rows x 1400 rows - 17k cells

Data per cell: 30-700 per cell

2.2 Pre-processing

Metadata - well location, image location within a well, cell location within an image,

Measurements - Feature data

Data elimination is expresly fit for purpose

Data reduction

Features can be highly correlated eg: actin fiber alignemnt against nuclear size

A common pathway could be come dysregulated, so this loss of correlation could itself indicate a phenotypic effect

Sometimes scaling issues can make it hard to spot important correlations and other relationships - so large correlation analysis can be very helpful

2.3 Data reduction - Normalization of data

Can't necessarily be done by eye

May correlate under one set of conditions but not under others

2.4 Data assessment and refinement

First **unsupervised learning**, for all the things that we tested what hypothesis can we develop

A literal hypothesis is not defined a priori - eploratory data analysis - correlations no causations

main question - how do we define general trends within the data

Requires a separate validation set for connecting correlation to causation

One of the most popular approaches - PCA Lot of dicussion on how PCA works that I didn't make notes on

Some discussion on k-means, self-organizing maps and hierchical clustering

hierarhical clustering is about separating points as opposed to measuring similarity between points as in k-means.

Different types: single linkage, complete linkage, average linkage

2.5 How can we tell

elbow plots for k-means clustering help determine the number of clusters, similar to Scree plots

2.6 Supervised learning

3 Intro to HCS/HCA - Image Analysis

Data/Image analysis are a part of assay development - should not be an afterthought

Stages -

- Assay development
- Hardware
- Image analysis

3.1 Software solutions

Application modules - turn key solutions

Development environments - good for new assay development, more flexibility

3.2 Illumination correction

Approximate background and perform illumination correction per plate, per well or per channel

Need to be sure about being clear of illumination artifacts versus systematic biology

3.3 Top-hate ("rolling ball") filtering

3.4 Thresholding

Sometimes it's straightforward, when you don't have a lot of noise (eg: Otsu etc.)

You can also use machine learning - eg: Ilastik

3.5 Identifying cell objects

Nuclei more easily separated than cells - DNA markers are specific - Yield is good (foreground/background contrast) - Uniform shape

Identifying cells is more difficult - Markers gnerally lower contrast - boundaries aren't as clear

Identifying secondary objects

Growing the primary objects to identify cell boundaries - eg: using nuclei as seeds by using a cell stain channel

what's a good marker - a good gradient works better - cell mask, beta-tubulin

3.6 Identifying sub cellular structures

same approaches apply but now you can use enclosing object for better preprocessing, thresholding

> Everything before this was the hardest part, it's all downhill fomr there

3.7 Extracting features

Most common readouts - Num of cells per image - num of organelles per image - num of organelles per cell

Maasuring object morphology - Area - Perimieter - Eccentricity - Major/minor axis length (elongation) - Form factor - measure of compactness - Zernike features

> If it touches the edge of the image, it's probably best to throw it out

Measuring object intensity - Integrated intensity - statistical measure of intensity

Measuring object texture determine whether the staining pattern is smooth or coarse

> Dealing with **uninterpretable features** - often ahrd to report but can be linked to more easily describable features or clusters (eg: linking one of the Harelik features to nuclear intensity) and then it's easier to link them to a specific biological phenotype. But yea, don't normally just toss them away because you don't know what they are

At the moment, cell profiler doesn't simply let you identify good cell phenotypes just by clicking them

Often hard to interpret HArelick texture features

3.8 questions

confocal (helps to zoom in on something particular) v/s wide field

4 Quality control and validation

4.1 Importance of giving a confidence interval for Z'

Can't measure effect size without confidence interval on - Z' - d (separation of standard deviations of two groups) - SSMD (Strictly standardized mean difference) - β

4.2 Interpretatin of metrics

lot more detail in the slides

β or the probability of superiority - much more apporpriate for HCS as opposed to univariate metrics

d, SSMD and Z' are all effect size measure

Cohen's U values d can also be represented as the odds ratio

4.3 Designing a robust assay

p values are useless by themselves, unless they are used to get back to the effect size

d can be retrieved from the p value (when based on a t distribution)

Sometimes, HCS perspective metrics (Z', d etc) may show much greater impact than HTS metrics (eg: p value)

4.4 Statistical power

Misinterpreting the p-value. Ensure adequate statistical power if it's not a big cost issue

Try power analysis available in most statistical packages

4.5 Dealing wtih multiple dimensions

Two inaccurate screens of lower accuracy could be combined to create secreens of a higher accuracy

Mahalanobis distance

When you work in multiple dimensions make sure your classifiers are also non-linear

Normalization (mean,sd shift) Feature selection

- provide simple models that simplify interpretation
- discuss the constant factor of complex algorithms
- mitigate the 'curse of dimensionality'
- reduce variance (reduce overfitting). curse of dimensionality - how many rows versus how many columns

5 Cytometry

5.1 How to choose the right platform choice

Good introductory paper - [6]

- suspension of cells or adherent cells
- number of markers
- endogenous vs exogenous probes
- live cells or fixed
- dynamics
- number of samples
- Application - assay development, screening, systems biology

insert table on key features of cytometry platforms

can common reagents be used across platforms - fluorescent proteins, labelled antibodies can assays be ported between platforms - FP expressions/dynamics, immunofluoresence, ion flux can measurements be standardized across platforms marker abundance, fluoresence intensity,

6 Brenda Andrews

Systematic analysis of genetic determinants of sub-cellular morphology

6.1 Integration of autmiated genetics and HTP microscopy

Double mutant map [1]

Combining SGA with HCS

6.2 Understanding the genotype to phenotype system

6.3 Quantitative cell biological analysis of the proteome

Y: Deep learning seems to be the approach of choice mainly for simplicity and time saving when working with complex networks, as opposed to increased accuracy alone

DeepLoc - Transfer learning protocol saved a lot of time and effort on training

Also reduced overfitting

cytoplasm, nucleus, septin marker to identfy bud neck

7 Organ-on-a-chip session

7.1 MIMETAS

All of these use image based readouts. They use a lot of automated image analysis. Done several projects with GSK.

[5]

- Using dyes for barrier integrity assays: - Kidney-on a chip - VAsculature-on-a chip (can see angiogenic sprouting!)
- Co-culture models Renal clearance model Neruovascular model of the blood brain barrier

7.2 3D pancreatic model for high throughput cancer drug screening

Clear increase in assay sensitivity in 3D. Drugs not active in 2D screening, show activity in 3D screens

7.3 High Content Imaging of Microtissues

Visikol - 3D BioImaging - Able to generate a spatial map of drug distribution within a microtissue model

8 Mitocondrial toxicity

Final result

Image based cellular analyses to monitor CLL patients to identify which need treatment and which are benefiting form treatment

9 Informatics session

9.1 Godinez

Dataset used - BBBCX

hold one compound out validation for mechanism of action training model

Tried an **unsupervised approach -** trained a new network on the images, did NOT use features from a pre-trained network and then cluster.

Appears to not have a clear understanding of why the algorithm seems to work

Drawbacks: don't provide single cell information

9.2 Genedata

Talking about workflow improvements that I spoke aobut during my talks at GSK

Have two workflows,

Assay development workflow:

- Generate training data, explore and dissect phenotypic space
- Create similarity maps on the first pass and then use that to decide what you want to trian
- Very visual and GUI-ey
- After training show saliency maps, confusion matrices etc.
- Object detection (single cells) -¿ Similarity maps -¿ Training -¿ Analyze
- used a Human Fibrosis Assay from AZ to test their workflow and saw 'very good' results with deep learning

Production workflow Similar to the deployment channel

Representation challenge Looking for the right embedding - Cell profiler - Autoencoder - VGG16

Understanding the decisions made by the network

> Good question by Ann Are there metrics that can capture the things that we don't already know. It's great that we can speed up analyses that we can already do, but are there ways in which we can target the full content of the image. Learn things that we don't already know.

Y:Really fascinating researchq question

9.3 Recursion Pharmaceuticals

Identification of therapeutics for rare genetics

> HCS Phenotyping + AI -¿ Disrupting the pharma industry in terms of how drug discovery is done

Utilize a single approach to answer all these different questions, mainly using a cell painting approach

Strucutral markers that light up a lot of compartments in the cell. It's target agnostic or biomarker agnositc. Aren't looking for a specific phenotype - very 'unbiased'. One single assay an give us a lot of the information that we want.

Still using a mix of segmentation and deep learning to get what they need to.

Detection of subtle phenotypic states and how to move between these. What might be useful for one therapetic area may not be for another. So want to look broadly and understand how to move from one section of the phenotypic map to another. Use hi dimensional phenotypic signature to describe a disease state.

Recursion is trying to repurpose compounds that have been abandoned by other drug companies. Trying to repurpose these compounds for rare diseases.

Plot drug response curves in their high dimensional phenotypic space. So you can track undesirable 'off-phenotype' effects instead of a univariate drug response.

> 2D monoculture seems to work quite well. Maybe we're not getting all the ifnormation out of these models that we need to.

> With phenotypic deiversity you automatically get chemical diversity but it's not true the other way around.

Significantly shorten hit to lead process

Use phenotypic hit expansion to find other molecules that achieve the sam effect.

Also have an immuno oncology platform - trying to modulate macrophage polarity (M1 or M2)

It's all about that high dimensional phenotypic signature

10 High Content Analysis - November 2017 - Cambridge, MA

10.1 Pfizer talk

Difference between phenotypic drug discovery and target based drug discovery

- Finding MoA that are dseirable but also lead to healthy state

- 90% of compounds traige away
- coounterscreens based on MoA, toxicology etc.
- Compare $\frac{CytoIC_{50}}{EfficacyIC_{50}}$
- Delay structure based hit teaiging
- Hit expansion

10.2 Novartis - Christophe Antczak - 5 years of Phenotypic Drug Discovery at Novartis

Mixed notes from different slides

- PDD - main readouts - imaging, reporters, gene expression
- Moving towards smaller screens
- talking about the trade off between throughput and physiological relevance when designing assays. Also about the trade off between robustness and sensitivity
- First step is iterative assay development involving quick prototyping. Goal is to get as physiologically close to patient as possible. Eventually get a phenotypic assay that is predictive of applications in vivo.
- In Vitro hit validation informs assay optimization
- In Vivo hit validation informs follow up screens
- Target ID is not necessarily an outcome of PDD
- could potentially bypass PK by using complex organoid models [3]

10.3 Panel Discussion

- **Differences between High content and phnotypic screens. Phenotypic** essentially means target-agnositc. Attempt to approzimate in vivo biology. **High content** - on the other hand - an assay modality - analysis procedures - eg: deep learning
- **What limits throughput.** Most common question: What is most likely to translate into animal model. **Academia**: Whether the disease and assay is relevant is more important than the assay/screen itself. **Industry**: Is there clear phenotype both in vivo and in vitro that can be measured. NV: If there is a link with a gene and that gene produces a phenotype -

then we go after the phenotype. Recursion does it the other way around - they look for phenotypes.

- **Future of Phenoptyic Screening.** Important to understand the biology around a specific process for the team to be successful. R: From morphology alone, trying to ascertain that there are a lot of different phenotypes - and then build phenotypic screens that detect these - thus becoming more independent of our initial assumptions.

- Phenotypic space is determined in large part by the design of assay. How do you make this decision? A: We need a quick profile to determine all phenotypes that are relevant.

- **Will AI make exiting methods obsolete** Reviewers, regulatory agencies - how do they take it? May replace some methods - but require tons of data to be very good - at the moment they're going to be complementary for a long time. N: A less (human) biased way of identifying trends is very valuable - but there's still a bias generated through the data. - Humans always want to know what's going on - and FDA and others would like to know what's going on. R: We rely more on latent representations as opposed to causal ones

- **Mixing up features from conventional and deep approaches** - Depth on the kind of featurization you use depend upon the task and the kind of data that you have.

- **What does it mean to have an understandable algorithm** - "Here are five images that fit this pattern - here are five that don't", "What are the most valuable/informative features?", "Industry wants to know that it works v/s researcher who wants to know why it works", "We may not have an absolute beginning to end explanation - but if it's saving lives do we need to answer all those questions"

References

[1] Michael Costanzo, Benjamin VanderSluis, Elizabeth N Koch, Anastasia Baryshnikova, Carles Pons, Guihong Tan, Wen Wang, Matej Usaj, Julia Hanchard, Susan D Lee, et al. A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 353(6306):aaf1420, 2016.

[2] Steven A Haney. Rapid assessment and visualization of normality in high-content and other cell-level data and its impact on the interpretation of experimental results. *Journal of biomolecular screening*, 19(5):672–684, 2014.

[3] Robert E Hynds and Adam Giangreco. Concise review: the relevance of human stem cell-derived organoid models for epithelial translational medicine. *Stem cells*, 31(3):417–422, 2013.

[4] Nilgun Tasdemir, Ana Banito, Jae-Seok Roe, Direna Alonso-Curbelo, Matthew Camiolo, Darjus F Tschaharganeh, Chun-Hao Huang, Ozlem Aksoy, Jessica E Bolden, Chi-Chao Chen, et al. Brd4 connects enhancer remodeling to senescence immune surveillance. *Cancer discovery*, 6(6):612–629, 2016.

[5] Sebastiaan J Trietsch, Elena Naumovska, Dorota Kurek, Meily C Setyawati, Marianne K Vormann, Karlijn J Wilschut, Henriëtte L Lanz, Arnaud Nicolas, Chee Ping Ng, Jos Joore, et al. Membrane-free culture and real-time barrier integrity assessment of perfused intestinal epithelium tubes. *Nature Communications*, 8, 2017.

[6] Shauna H Yuan, Jody Martin, Jeanne Elia, Jessica Flippin, Rosanto I Paramban, Mike P Hefferan, Jason G Vidal, Yangling Mu, Rhiannon L Killian, Mason A Israel, et al. Cell-surface marker signatures for the isolation of neural stem cells, glia and neurons derived from human pluripotent stem cells. *PloS one*, 6(3):e17540, 2011.