Outline

• Experimental design
  - What to think about when approaching a new problem?
  - How do you know that your method produces useful results?
  - How to quantify a method's performance?
    - What is performance?
      - Accuracy, precision, recall, and F-score
    - Dice score
    - Hausdorff distance
    - Training, validation and testing
    - Quality control
  - How can this be used in relation to the course projects?

• What is performance?
  - Accuracy, precision, recall, and F-score
  - Dice score
  - Hausdorff distance
  - Training, validation and testing
  - Quality control
  - How can this be used in relation to the course projects?

• Examples from High Throughput Screening in biomedicine
  - Inventing the wheel, or using available resources?
    - Short introduction to some different software tools

Theory

- Object detection, segmentation, and classification algorithms, including Deep Learning approaches
- Automated methods for high-throughput phenotyping of cells and model organisms (C. elegans and zebrafish)
- Large-scale image-based screening and compound profiling on cultured cells (including patient-derived cancer stem cells)
- Automated tissue analysis and sequencing of mRNA in situ

Applications

In situ sequencing for analysis of RNA in tumor tissue
Analyze of virus particles and other vesicles in electron microscopy images.
Time-lapse studies of bacteria and fluorescent signals.
Understand fatty metabolism and cancer using primary cell cultures from patients.
Zebrafish assays, screening, development, and coronary heart disease.

Linked references to publications available at http://www.cb.eu/~carolina/carlina_publications.html

Support & Education

A future scenario:

Can you write a program that automatically counts all the dots in these images?

Sure.

Just give me the images. Easy thing, will deliver tomorrow.

A better future scenario:

Can you write a program that automatically counts all the dots in these images?

How many dots are there?

Here is this cell or background noise? How important is it to separate touching dots? Are 'vines' also dots or DNA? What is a dot anyway? And why should they be counted?

Many hours later:

Here is this cell or background noise? How important is it to separate touching dots? Are 'vines' also dots or DNA? What is a dot anyway? And why should they be counted?

Can you write a program that automatically counts all the dots in these images?

You

An expert (and very important collaborator)

You
Method Evaluation

- You only know whether your method produces useful results if you know what results you expect.
- Thus: first try to solve a solved problem!
- “Ground truth” or ‘gold standard’
  - the ‘correct’ solution to the problem
- Compare your method’s results to ground truth
  - Use some kind of metric to measure performance
  - Compare alternative methods using the same metric

A clear hypothesis or goal will be helpful for experimental design

- What is the trade-off when it comes to improvements (e.g., consistency of image acquisition and/or sample preparation?)
- Higher resolution vs number of spots per field?
  - using auto-focusing
  - imaging all samples using the same illumination source
  - avoiding glare and shadows etc.
- Is this collection of images representative of the images you want to analyze?
- What do the worst images look like?

What is ‘ground truth’

- The word comes from observations made on the ground when validating measurements from remote sensing (satellite or airplane)
- Often the output of visual assessment; manually drawn outlines, counts and/or visual classification (e.g., visual assessment of license plate numbers to optimize and validate an automated car toll system)
- Also referred to as ‘benchmarking data’

Performance evaluation

- Predictive capability is usually more interesting than speed.
- In order to evaluate performance we need to know the actual class (of an image: an object or a pixel). This is often referred to as ‘ground truth’.

<table>
<thead>
<tr>
<th>ACTUAL CLASS</th>
<th>PREDICTED CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>TP  FN</td>
</tr>
<tr>
<td>No</td>
<td>FP  TN</td>
</tr>
</tbody>
</table>

Accuracy = (TP + TN) / (TP + FN + FP + TN)

Warning: Accuracy is misleading! E.g., consider a 2-class problem with 9999 examples from class 1, and 1 example from class 2, and a model predicting everything to be class 1... Would give 99.9% accuracy!
Performance evaluation

<table>
<thead>
<tr>
<th>ACTUAL CLASS</th>
<th>PREDICTED CLASS</th>
<th>TP</th>
<th>FN</th>
<th>FP</th>
<th>TN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>TP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td></td>
<td>FN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>FP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>TN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Accuracy = \( \frac{TP + TN}{TP + FN + FP + TN} \)
Recall (=Sensitivity) = \( \frac{TP}{TP + FN} \)
Precision = \( \frac{TP}{TP + FP} \)
F-score = \( \frac{2 \times TP}{FP + FN + 2 \times TP} \)

Measurements that are better than Accuracy:

- Precision tells us what proportion of the positives we can trust, while Recall tells us what proportion of the positives we found.
- The F-score is high when both the precision and the recall are reasonably high.

Recall, precision, accuracy, and the F-score

\( TP = \text{True Positive pixels (correct detections)} \)
\( FP = \text{False Positive pixels (type 1 error)} \)
\( TN = \text{True Negative pixels (often not relevant)} \)
\( FN = \text{False Negative pixels (type 2 error)} \)

Accuracy = \( \frac{TP + TN}{TP + FN + FP + TN} \)
Recall (=Sensitivity) = \( \frac{TP}{TP + FN} \)
Precision = \( \frac{TP}{TP + FP} \)
Specificity = \( \frac{TN}{TN + FP} \)

The F-score (sometimes called F-factor or F-measure) = \( \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \)
referred to as a harmonic mean of recall and precision.

How does the F-score compare to the Dice score?

\( DSC = \frac{2 \times TP}{TP + FN + 2 \times TP} = \text{F-score}!!! \)

What if you have hundreds of objects?

Associate each detected object 1-5 with its most overlapping reference object A-E. If the reference object is shared with another detected object, associate the two objects with the most overlap. Then calculate the F-factor (or DICE) on an object basis.

1—A, F-score = 0.60
2—none, F-score = 0
3—B, F-score = 0.60
4—D, F-score = 0.65
5—none, F-score = 0

False negatives, F-score = 0

Comparing image segmentation results: Dice similarity coefficient

The Dice similarity coefficient (DSC) measures spatial overlap

\( DSC = \frac{2 \times \text{Intersection of Ground Truth and Method's Segmentation}}{\text{Union of Ground Truth and Method's Segmentation}} \)

What F-factor is ‘good enough’?

Hausdorff distance

These two pairs of objects have similar Dice scores, but different Hausdorff distances.

For each point on \( a \) on \( a \), find the shortest distance to any point on \( g \). \( H_{ag} \) is the longest of these distances.

\( H_{ag} = \max(\min(\text{dist}(a,g))) \) and \( H_{ga} = \max(\min(\text{dist}(g,a))) \)

The Hausdorff distance \( H = \max(H_{ag}, H_{ga}) \)

Automated lung segmentation
Ground-truth is needed for optimization/training, validation and testing

- **Training set**
  - this is the data (images) you use to develop your method
- **Validation set**
  - this is the data (images) you use to optimize your method
- **Test set**
  - this is the data (images) you use to evaluate your method

- The training set + validation set will let you:
  - tweak your method until it does what you want (also e.g. machine learning)
  - optimize the parameters of your method

- The test set will let you:
  - see how well your method works on an unseen dataset (accuracy, precision)
  - compare different methods

Training, validation and testing

- **Training set**
- **Validation set**
- **Test set**

Learning and methods ‘tweaking’

- Performance on the training data is NOT a good indicator of performance on future data!

Over-fitting means your classifier not only fits the training observations, but also fits the random noise in the data.

Instead, tweak the parameters until your model fits your validation set!

Finally evaluate on your unseen test set.

k-fold cross-validation

- Used when you don’t have enough data with ground truth
- Repeated training of parameters and method validation with different subsets of the data
  - Divide data into k subsets
  - Pick k-1 subsets for training, last subset validation
  - Repeat so that each subset has been used for validation once
  - Average performance measure over all k repetitions
- Yields unbiased estimate of method performance

Finally, train method parameters with all data and test on an ‘unseen’ test set

- When k = N (the number of images/cells/...)
  - leave-one-out cross-validation

The ROC curve ‘Receiver Operator Characteristics’

- Points map positive and false positive rate as one parameter in the method is changed
  - e.g. changing the classifier boundary (threshold) in a binary classifier
- Often used for cost/benefit analysis

AUC: area under curve

- Simply compute the area under the ROC curve; a larger area indicates better performance.
- Summarizes the ROC curve, but might not provide useful information.
- The F-factor is typically a better summary for a method’s performance
  - or e.g. determine distance to the top-left corner
discriminative power; the Z-factor

A common metric when evaluating the ‘power’ or ability of an assay to discriminate between positive and negative controls. Note: What if distributions aren’t Gaussian?

Image collections with ‘ground truth’

Other resources:
http://people.mcs.anl.gov/~alanm/benchmarks.html
A collection of images with benchmark/ground truth info from a range of different fields (medical, buildings, fingerprints, etc).

Quality control

• Is the input what you expect?
  – i.e. similar enough to the training set
• Things that can go wrong:
  – staining subpar
  – imaged region contains something unexpected
  – camera was out of focus
  – illumination not aligned (uneven illumination)
  – etc.
  – etc.
• Do you build a test for each possible issue?

Data quality control

• If we have 100,000 images, and 10% of them are out of focus or contain debris that skews our measurements, there is a large risk of getting many false hits and also missing hits.
  • How can we find these ‘bad’ images?
  • What to do once we’ve found them?

Three important means of QC

• good experimental design
  – helps to identify systematic errors (especially those linked for example with well position) and determine what normalization should be used to remove/reduce the impact of systematic errors
• the selection of effective positive and negative (chemical/biological) controls
• the development of effective QC metrics to measure the degree of differentiation so that images with inferior data quality can be identified and flagged/excluded.
Controls (a type of ground truth)

- Every experiment (data set) requires a positive and a negative control
  - controls = data with ground truth
  - every experiment can be tested for sensitivity and specificity
    - if these deviate from expectation, something went wrong
- Every experiment should be independently replicated
  - in the ideal case
  - but it is expensive to do everything twice
- Controls:
  - known to be positive / negative
  - treated differently to look like positive / not treated
  - ...

Image based screening

- Screening: to have a large number of samples and identify those that deviate from the norm. The deviation could either be well defined or undefined.
- Many applications other than biomedical (control of product quality, damage detection, sorting etc.)
- Focus today: high throughput screening with biomedical applications.

Image-based screening

- Search for small molecules (or siRNA) that in vivo mimic/induce expected changes in phenotype.
- 10^7 images, 10^7 cells/each image:
  - Total of 10^7 cells/experiment
- Delineate each cell and measure a large number features.

Morphology based predicting a compound’s mechanism of action

- The collection of measurements describing the appearance of a cell
  - Hundreds of features per cell...
  - Thousands of cells per sample...
  - Tens of thousands of samples...

Cytological profile

- Morphological features = cytological profile
Image-based drug screening using model organisms

Current problem:
People infected with bacteria resistant to known antibiotics

Possible solution:
Collections of thousands of different chemical compounds are available. Perhaps one of them works as a drug?

Now what?
Take a thousand sick patients and try a different chemical on each?

Using C elegans to search for novel anti-infectives

1. Infect worms with human pathogens (bacteria)
2. The worms get sick
3. Place sick worms in wells, add a different compound to each well (~4000 potential drugs)
4. Wait
5. Figure out if any treatment cured the worms: ‘live/dead scoring’

Challenges with per-worm measurements:
Worms touch, overlap and cluster
In a typical screen on anti-infectives:
- 15 worms in each well to maximize statistics while minimizing the cost of the screen.
- ~50% of worms are clustered (based on visual examination of 100 wells, 1500 worms)

Methods evaluation
‘Ground truth’ computer result

How to ‘untangle’ worms
Find a combination of control
2. Figure out cluster
3. Have a worm-like shape and length

Robustness to large screens?

1800 random images + 200 hits selected out of 130,000 images from previous screen by Moy et al. and scored visually or by fully automated approach (shape + texture).
Time for visual screen: 20 hours, spread out over 2 weeks
Automated scoring: 15 min on computer cluster
Accuracy = (TP+TN)/total×100%, precision=TP/(TP+FP)=83%
Quantification of variations in reporter protein expression

Once worms are delinated, we can reaample each worm and align it with a common baseline resolution atlas.

Metabolism screen:
Find genetic pathways that regulate fat metabolism

Conservation during evolution will give us hints about humans.

Genome-wide screen: Turn off 20,000 C. elegans genes (one at a time) using RNAi and see which ones affect fat metabolism.

High-throughput phenotyping in zebrafish

untreated treated


Optical Projection Tomography in HTS

1 fish processed every 18s
A new biobank of glioblastoma derived CSCs

Goals:
1. Explain and predict biological effects of siRNA and drugs.

Phenotypic analysis: We are currently testing 250 compounds (Prestwick library + collection from AstraZeneca) at 10 doses on 100 cell lines.

~250 wells (250 compounds, 10 doses per compound)

~200 cells per image at 20x, 4 images per well, three image channels (Hoechst, phalloidin and bright-field)

~10^6 images

Automated analysis using CellProfiler running on a computer cluster.
Project-wide web access to readouts and raw data via CellProfiler Analyst.

Microscopy Analysis of Adipogenesis Models

Avoid inventing the wheel again

Free and open-source imaging software tools
(focused on microscopy, but useful for a number of other applications as well)

Benefits of open-source software

- Educational value: anyone can go in and look at the source and learn.
- You often have the possibility to add your own algorithms.
- A user community for an open-source software will often be a more responsive and efficient source of help than what can be provided through an expensive service package for a commercial software.
- Warning: Always validate the code before trusting it. Everyone makes mistakes. ‘Reproducible research’; with an open source solution, you can provide your analysis pipeline as part of the supplementary material of your published paper (along with data), and use the macro recorder of ImageJ to document exactly what you do with each image.
- You can easily share your analysis approach with colleagues.
Take home message

Knowledge about the image formation, possibilities and limitations, can greatly improve the scientific value of an experiment.

A better understanding of digital image processing, possibilities and limitations, can greatly improve the scientific value of an experiment.

A better communication between experts of from different fields can greatly improve the scientific value of an experiment.

Metrics and data for optimization/training and testing/evaluation should be considered already at the design of a project.

Carolina Wahlby, carola@cb.uu.se