

Metrics and Benchmarking in Bioimage Analysis

Martin Maška

Centre for Biomedical Image Analysis,
Faculty of Informatics, Masaryk University

Defragmentation Training School #2
Porto, Portugal, May 8-10, 2023

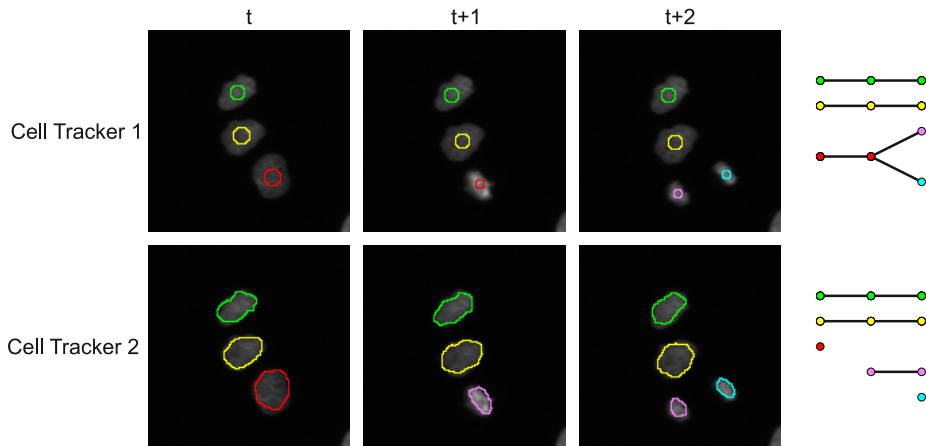


https://www.fi.muni.cz/~xmaska/NEUBIAS23/presentation_neubias23_ts.pdf

- 1 Importance of Benchmarking
- 2 Designing a Benchmark/Challenge
- 3 Performance Evaluation

Motivation

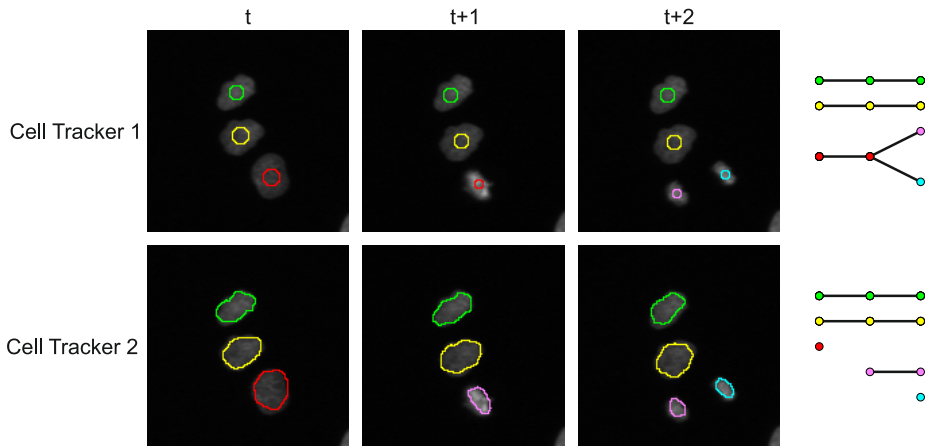
- **Automated image analysis** is a must when quantitatively analyzing bioimage data
- A bunch of **suboptimal tools** often available



Which of these two tools should one prefer?

Motivation

- **Automated image analysis** is a must when quantitatively analyzing bioimage data
- A bunch of **suboptimal tools** often available



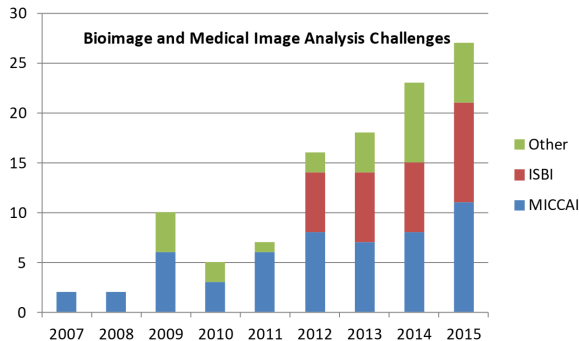
Which of these two tools should one prefer?

The choice is application-dependent!

Bioimage and Medical Image Analysis Challenges

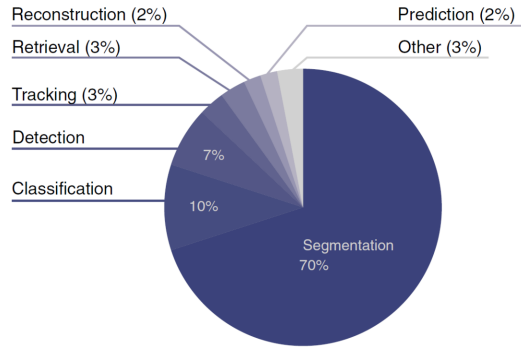
Objectives

- Standardization of reference datasets and algorithm performance measures
- Comparison of the performance of existing and newly developed tools
- Dissemination of the evaluated tools to the community



Kozubek, *Adv Anat Embryol Cell Biol*, 2016

Algorithm categories



Maier-Hein *et al.*, *Nature Communications*, 2018

Lifecycle of a Challenge

Establishing a benchmark dataset

- Select representative image data for a particular task
- Define an annotation protocol and prepare reference annotations

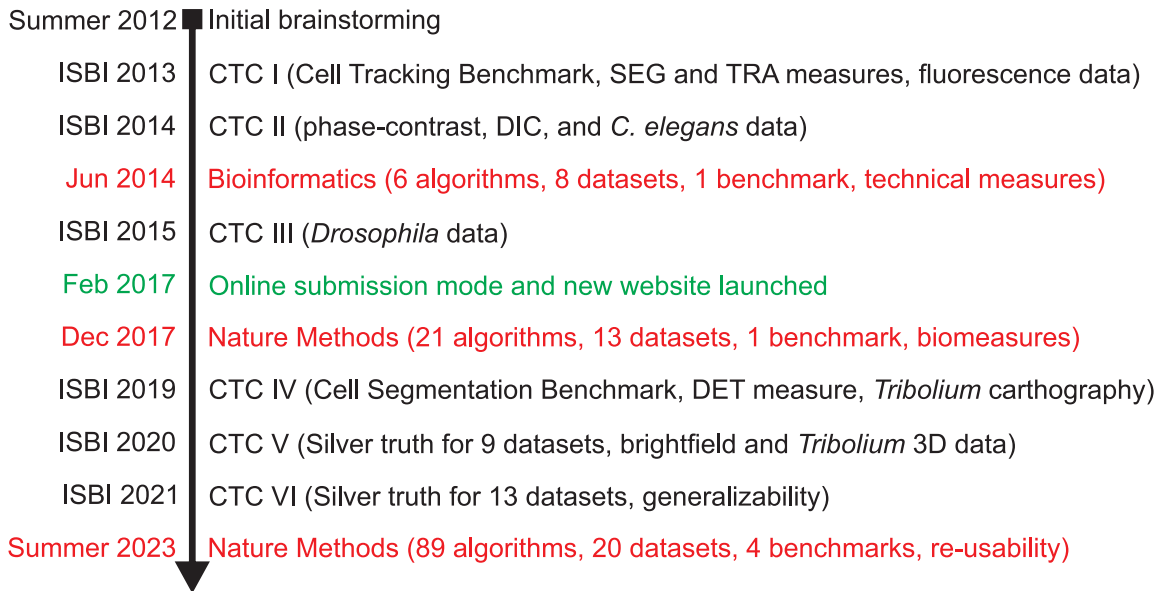
Establishing a challenge

- Split the benchmark dataset into training and test data
- Define evaluation protocol and create evaluation tools


Organizing the challenge

- Attach the challenge to a prestigious event (to attract more participants)
- Release datasets, evaluation tools, and timeline on a web page
- Verify and evaluate the submitted results, and compile rankings
- Publish results (on a web page and possibly in a journal)
- Keep the challenge alive by reflecting new trends in the field

Cell Tracking Challenge Milestones



Cell Tracking Challenge Milestones



Summer 2012	Initial brainstorming
ISBI 2013	CTC I (Cell Tracking Benchmark, SEG and TRA measures, fluorescence data)
ISBI 2014	CTC II (phase-contrast, DIC, and <i>C. elegans</i> data)
Jun 2014	Bioinformatics (6 algorithms, 8 datasets, 1 benchmark, technical measures)
ISBI 2015	CTC III (<i>Drosophila</i> data)
Feb 2017	Online submission mode and new website launched
Dec 2017	Nature Methods (21 algorithms, 13 datasets, 1 benchmark, biomeasures)
ISBI 2019	CTC IV (Cell Segmentation Benchmark, DET measure, <i>Tribolium</i> carthography)
ISBI 2020	CTC V (Silver truth for 9 datasets, brightfield and <i>Tribolium</i> 3D data)
ISBI 2021	CTC VI (Silver truth for 13 datasets, generalizability)
Summer 2023	Nature Methods (89 algorithms, 20 datasets, 4 benchmarks, re-usability)

More details will be revealed tomorrow at 10:30!

Selection of Representative Bioimage Data

Covering natural variability of imaged targets of interest

- Size, shape, texture, density, motility patterns, etc.

Covering natural variability of events/processes

- Cell division, cell death, cell fusion, overlapping cells, etc.

Covering common and rare artifacts

- Fluorescence bleaching, uneven illumination, presence of debris, level of noise, etc.

Widefield (10×), HeLa cells

Phase contrast (20×), U373 cells

DIC (63×), HeLa cells

Real versus Computer-Generated Bioimage Data

Real bioimage data

- + Actual source of biologically relevant information
- No reference output exists (manual annotations needed)
- Expensive, irreproducible, and less variable image acquisition

Computer-generated bioimage data

- A tool that realistically mimics real bioimage data needed
- Lack of natural variability
- + Inherently generated reference annotations (ground truth)
- + Cheap generation of a similar phenomenon under different imaging conditions

Annotation of Real Data

Gold truth

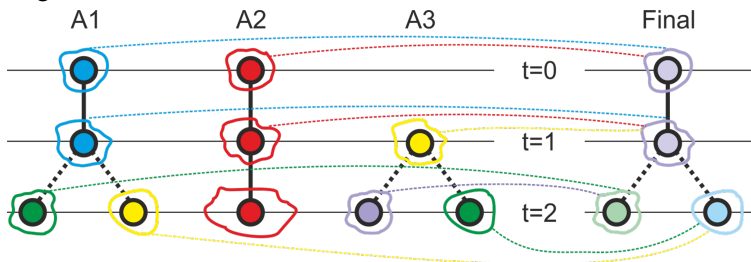
- Fusion of manual annotations created by human experts
- Laborious to obtain, limited quantity, higher reliability (suitable for training and testing)

Silver truth

- Fusion of computer-generated annotations (e.g., created by several algorithms)
- Easy to obtain, higher quantity, lower reliability (suitable for training only)

Annotation fusion

- Reduction of the subjectivity and error-proneness of experts' opinions
- Majority voting often followed



Training versus Test Data

Why to split datasets

- Training phase: developers fine-tune their methods using training datasets with reference annotations available
- Test (competition) phase: developers and/or challenge organizers apply the fine-tuned methods to previously unseen test data with secret reference annotations

How to split datasets

- It is suggested to use majority of the data for training and minority for testing
- A **50:50 rule** is however often taken in practice due to limited gold truth availability
- The split must be conducted in a balanced way (i.e., both training and test data are representative and have similar properties)

Classification of Performance Evaluation Measures

Technical measures

- Precision, Recall, F_1 -score, Accuracy, Average precision, Mean average precision
- Jaccard similarity index, Dice similarity coefficient, Hausdorff distance
- Root-mean-square error

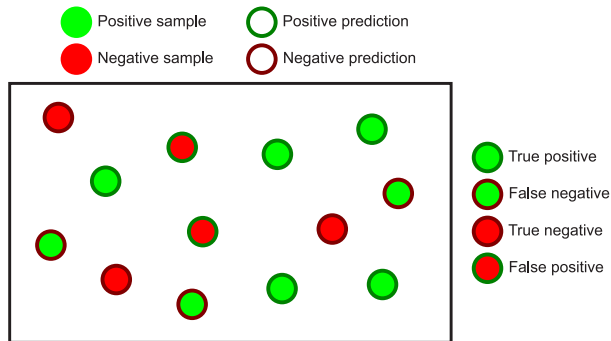
Biologically oriented measures

- Complete tracks (F_1 -score of entirely reconstructed tracks)
- Track fractions (percentage of correctly reconstructed tracklets)
- Branching correctness (F_1 -score of correctly detected divisions)

Usability measures

- Number of tunable parameters required
- Generalizability and availability
- Execution time and peak memory consumption

Binary Classification (valid also for object detection)



Confusion matrix

		True class	
		Positive	Negative
Predicted class	Positive	TP	FP
	Negative	FN	TN

Precision = $\frac{TP}{TP+FP}$ (percentage of correct positive predictions)

Recall = Sensitivity = $\frac{TP}{TP+FN}$ (percentage of correctly predicted positive samples)

Specificity = $\frac{TN}{TN+FP}$ (percentage of correctly predicted negative samples)

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ (percentage of correct predictions)

F₁-score = $\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$ (harmonic mean of Precision and Recall)

Binary Classification: Important Notes

- The scores of all the five measures range from 0 (worst) to 1 (best)
- Accuracy and F_1 -score reflect the overall performance as a single number
- Precision and Recall are **mutually dependent** and of the **same importance** in F_1 -score

How to favor Precision or Recall

$$F_{\beta}\text{-score} = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$$

- Precision is favored for $0 < \beta < 1$
- Recall is favored for $\beta > 1$
- F_2 -score or F_3 -score often used in medicine not to miss cancerous targets

Multiclass Classification

Confusion matrix

		True class		
		Cat	Dog	Fish
Predicted class	Cat	A	B	C
	Dog	D	E	F
	Fish	G	H	I

Class-Level Performance

	Precision	Recall	F1-score
Cat	J	K	L
Dog	M	N	O
Fish	P	Q	R

Macro-averaging

$$\text{Macro-Precision} = (J + M + P) / 3$$

$$\text{Macro-Recall} = (K + N + Q) / 3$$

$$\text{Macro-}F_1 = (L + O + R) / 3$$

Micro-averaging

$$\begin{aligned}\text{Accuracy} = \text{Micro-Precision} = \text{Micro-Recall} = \text{Micro-}F_1 &= \\ &= (A + E + I) / (A + B + C + D + E + F + G + H + I)\end{aligned}$$

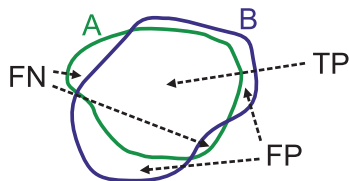
Weighted averaging

$$\text{Weighted-Precision} = [(A + D + G) \cdot J + (B + E + H) \cdot M + (C + F + I) \cdot P] / 3$$

$$\text{Weighted-Recall} = [(A + D + G) \cdot K + (B + E + H) \cdot N + (C + F + I) \cdot Q] / 3$$

$$\text{Weighted-}F_1 = [(A + D + G) \cdot L + (B + E + H) \cdot O + (C + F + I) \cdot R] / 3$$

Binary Segmentation: Overlap-Based Measures



Jaccard Similarity Index = Intersection over Union = $\frac{|A \cap B|}{|A \cup B|}$

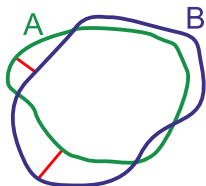
Dice Similarity Coefficient = pixel-level F_1 -score = $\frac{2 \cdot |A \cap B|}{|A| + |B|}$

- The scores of both measures range from 0 (worst) to 1 (best)
- Both measures yield the same ranking and are less sensitive to fine details

How to deal with multiple objects per image or per whole dataset

- For each reference mask A, find a segmented mask B ($|A \cap B| > 0.5 \cdot |A|$ or $|A \cap B| > 0.5 \cdot |A \cup B|$ to include or exclude non-splits, respectively) and compute their overlap-based score
- Average the overlap-based scores over all reference masks

Binary Segmentation: Boundary-Based Measures



$$\text{Hausdorff Distance} = \max \left\{ \max_{a \in A} \min_{b \in B} d(a, b), \max_{b \in B} \min_{a \in A} d(a, b) \right\}$$

where $d(a, b)$ is a Euclidean distance between a and b

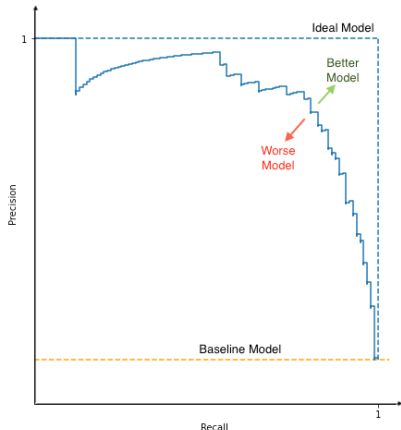
- The scores range from 0 (best) to ∞ (worst) and are sensitive to outliers
- Percentile Hausdorff Distance (the inner maxima replaced by a percentile – often 95th percentile) and Average Distance (the inner maxima replaced by averaging)

How to deal with multiple objects per image or per whole dataset

- For each reference mask A , find a segmented mask B ($|A \cap B| > 0.5 \cdot |A|$ or $|A \cap B| > 0.5 \cdot |A \cup B|$ to include or exclude non-splits, respectively) and compute their boundary-based score
- Average the boundary-based scores over all reference masks

Instance Segmentation

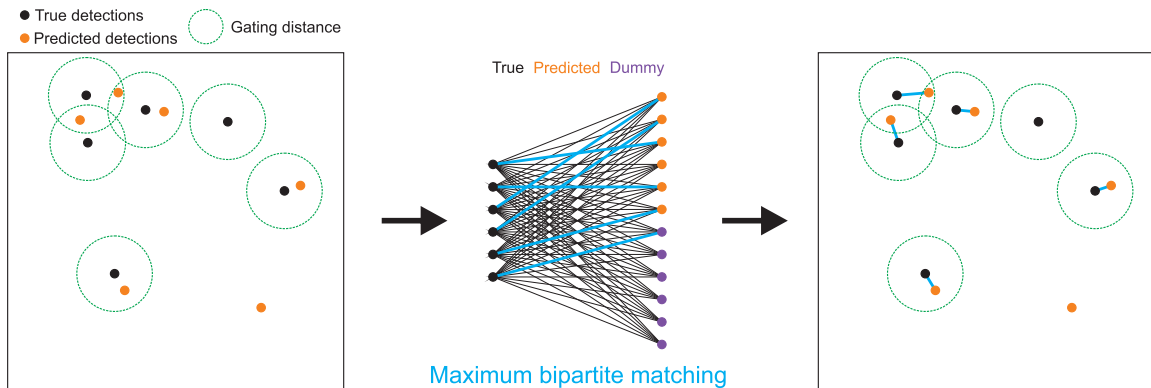
- **Wrong approach 1:** compute an overlap-based measure over the whole image
- **Wrong approach 2:** compute an object detection measure for a fixed IoU threshold
- Correct approach: compute an object detection measure and an overlap-based measure per each instance (+ averaging over all instances)
- Alternative approach:



Average Precision = area under the Precision-Recall curve (from 0 (worst) to 1 (best))

Mean Average Precision = Average Precision averaged over all available classes

Object Localization

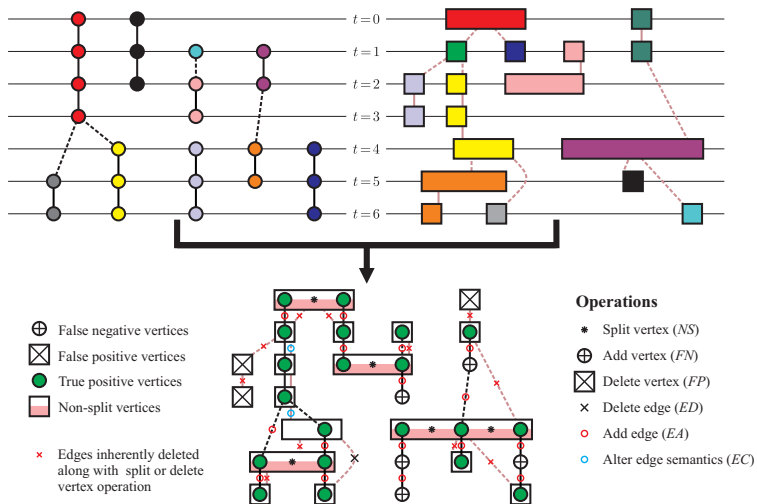


$$\text{Root-Mean-Square Error} = \text{Root-Mean-Square Distance} = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^N d(a_i, b_i)^2} \quad \text{where}$$

$d(a_i, b_i)$ is a Euclidean distance between a matched pair of true and predicted detections

Object Tracking

- Weighted matching of acyclic oriented graphs
- Capable of assessing the detection and linking steps separately



Matula et al., PLoS ONE, 2015

Further Reading and Useful Links

Performance Evaluation Measures

- <https://arxiv.org/abs/2104.05642>
- <https://arxiv.org/abs/2206.01653>

Benchmarks and Challenges

- https://doi.org/10.1007/978-3-319-28549-8_9
- <https://doi.org/10.1038/s41467-018-07619-7>
- <https://data.broadinstitute.org/bbbc/>
- <https://doi.org/10.6075/J0S180PX>
- <http://cbia.fi.muni.cz/datasets/>

Database of Challenges

- <https://grand-challenge.org/challenges/>

Invitation to the AMBIA 2023 Summer School



- AMBIA = **A**dvanced **M**ethods on **B**iomedical **I**mage **A**nalysis
- Dates: September 10-16, 2023
- Deadline for application: May 31, 2023
- Target audience: PhD students and junior researchers
- Location: Masaryk University, Brno, Czech Republic
- Official language: English
- Number of participants: 20-25
- Structure: lectures/PC labs/poster session
- Web pages: <https://ambia.fi.muni.cz/>

