CYTOMINE
Large-scale bioimage analysis with web services and machine learning

Raphaël Marée
PhD, GIGA Bioinformatics Core Facility
Systems and Modeling, Dept. EE&CS
University of Liège, Belgium
Biomedical research heavily relies on bioimages.

Whole-slide scanners

High-content screening

PET scan

Maldi imaging

Electron microscope
e.g. Cytology: many preparation protocols, cell types, ...
Biomedical research heavily relies on bioimages.

Whole-slide scanners

PET scan

High-content screening

Electron microscope

Maldi imaging
e.g. many Zebrafish imaging screens
Automated bioimaging

Sample preparation

Kurabo AS-200 auto slide preparation system
100 slides / hour

Slicing  Mounting  Smoothing  Drying

200 GByte/hour  5TByte/day

Image acquisition

Nanozoomer (Hamamatsu) whole-slide scanner
320 slide loader, 35 s / slide (area: 15 mm x 15 mm),
0.23 μm/pixel. One image ~ (100K x 100K pixels)

+/- 2 GByte / image
Bioimage Informatics challenges

How to transfer bioimaging data?

How to store?

How to annotate?

How to share?

How to find/search?

How to derive new biomedical knowledge?
Bioimage informatics: a new area of engineering biology

Hanchuan Peng
Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, Virginia, USA

Received on June 6, 2008; revised on July 1, 2008; accepted on July 2, 2008
Advance Access publication July 4, 2008
Associate Editor: Jonathan Wren

Bioimage informatics: a new category in Bioinformatics
Hanchuan Peng¹, Alex Bateman², Alfonso Valencia³ and Jonathan D. Wren⁴
¹Janelia Farm Research Campus, Howard Hughes Medical Institute, VA 20147, USA, ²Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK, ³Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre, E-28029 Madrid, Spain and ⁴Oklahoma Medical Research Foundation, Oklahoma City, OK 73104, USA

Why bioimage informatics matters

Gene Myers

Driven by the importance of spatial and physical factors in cellular processes and the size and complexity of modern image data, computational analysis of biological imagery has become a vital emerging sub-discipline of bioinformatics and computer vision.
Bioimage Informatics challenges

How to transfer bioimaging data?

How to store?

How to annotate?

How to share?

How to find/search?

How to derive new biomedical knowledge?
Biomedical research heavily relies on annotation & quantification of images

(Source: Google Images)
Until recently, annotations/quantifications were still often:

- performed manually

- performed only within subregions of subgroups of samples
  (might not be statistically significant or one might miss specific patterns)

- created by isolated experts and stored locally
  (sometimes lost)

- not saved in standardized format so hardly reusable
  (e.g. Photoshop draws for paper figures)
Computer scientist's mission

Until recently, annotations/quantifications were still often:

- performed manually

  Automatic quantifications (as automated as possible)

- performed only within subregions of subgroups of samples
  (might not be statistically significant or one might miss specific patterns)
  
  Analyze many, large, samples

- created by isolated experts and stored locally
  (sometimes lost)

  Enable sharing of data (images, annotations, results)

- not saved in standardized format so hardly reusable
  (e.g. Photoshop draws for paper figures)
Our Cytomine software relies on...

- Modern web development tools & open-source libraries
  - HTML
  - CSS
  - GRAILS
  - jQuery
  - OpenLayers™
  - VARNISH CACHE
  - PostGIS
  - NO SQL
  - BackboneJS
  - Hibernate
  - IIP
  - OpenSlide
  - RabbitMQ
  (+/- 500 person-years)

- Recent algorithms in machine learning and image analysis

- High-performance computing and mass storage equipments
  (1 Petabyte GPFS at GIGA + Montefiore Institute)
RESTful Web API

Client WEB (Javascript)

WEB

HTTP(s)

CYTOMINE CACHE

HTTP(s)

CYTOMINE API

Serveur WEB

Serveur BDD

STATE APIS

HTTP(s)

Connector

APP 1

HTTP(s)

WEB

HTTP(s)

CYTOMINE CACHE

Multi-formats HTTP Image Server

Serveur d'images

Espace Disque

Connector

APP 2

python

java
Software features:
Organize and centralize on the web

Create and manage multiple projects:

- **Upload** images to centralized server or keep data local (distributed image tile servers)
- Support for **various formats** (TIFF, JP2000, Aperio SVS, Hamamatsu NDPI, 3DHistech MRXS, Leica SCN, Roche BIF...)
- Users with **authentication** (e.g. LDAP), **access rights**, and **roles**
- **Specific ontologies** with user-defined, vocabulary terms
Software features:
Visualize

- Explore large (>gigabyte pixel) images at multiple resolutions
- GoogleMaps/OpenStreetmap browsing style (zoom in/out, pyramid tile-based)

1 tissue slice = 35000 x 30000 pixels (0.23µm/pixel)

4 fluo channels 83000 x 100 000 pixels = 4 x 16GB image
Software features: Annotate

- **Annotate** images using various **drawing tools**, with **user-specific layers**
- Describe ROIs with **ontology terms**
- Describe images and ROIs with any **key-value properties** or **text description**
- Build up pathology atlases and gather annotation statistics

![Image of annotation features]
Software features:
Search

- Visual search of regions of interest

Marée et al., Incremental indexing and distributed image search using shared randomized vocabularies, Proc. MIR 2010
Software features:

Share

- **Share images** through simple URLs

http://beta.cytomine.be/#tabs-image-83151073-86503947-

- **Share annotations** through simple URLs & e-mail mechanisms

http://beta.cytomine.be/#share-annotation/92024416
Software features:
Live broadcast
Integrate third-party, distributed, softwares through web services with HTTP requests and import/export data through JSON messages.

- http://beta.cytomine.be/api/project.json
- http://beta.cytomine.be/api/annotation.json?&project=60&term=4777&users=14,16

Software features:
- Extend and reproduce

Software parameters and results are recorded in the centralized database to ease traceability and reproducibility.
Software features: Analyze and proofread

- **Generic machine-learning** based image recognition


- Built-in interfaces for **algorithm evaluation** and **collaborative proofreading**

LUNG tumor tissue quantification
(ongoing collaboration with D. Cataldo, N. Rocks, at LBTD, GIGA)

What is the impact of condition X/Y/... on lung tumor onset and progression?

Condition X

Condition Y

One image = 40 K x 30 K pixels
Tens or hundreds of glass slides to be quantified per study...
Hybrid human-computer workflow
Hybrid human-computer workflow

1. Manual region contouring and labelling to provide training examples
Hybrid human-computer workflow

2. Automatic training of image recognition model based on training examples

VS

Multi-threaded implementation of Random subwindows and Extremely Randomized Trees
Hybrid human-computer workflow

3. Distributed processing of slide images
Hybrid human-computer workflow

3. **Automatic** segmentation of tumors in new images

One image ~ 40 K x 30 K pixels
4. Proofreading automatic segmentations
Hybrid human-computer workflow

4. Export statistics

> 500 whole-slide images analyzed with > 20 000 validated tumoral islets

Roles of polarized neutrophils on lung tumour development in an orthotopic lung tumour mouse model
Rocks et al., European Respiratory Society Annual Congress, 2013
Tumor/necrosis (IHC)
(ongoing work with Ph. Martinive, N. Leroi at LBTD, GIGA)
Ongoing application: cell counting

H&E nucleus counting in whole-slide images
(R. Longuespée, GIGA)
Ongoing application: RNA molecule counting
Bioimaging prospects

2D and 3D MALDI-imaging: Conceptual strategies for visualization and data mining☆

Herbert Thiele a,b, Stefan Heldmann b, Dennis Trede a, Jan Strehlow d, Stefan Wirtz d, Wolfgang Dreher f, Judith Berger e, Janina Oetjen e, Jan Hendrik Kobarg h, Bernd Fischer b, Peter Maass a,c

Image-based transcriptomics in thousands of single human cells at single-molecule resolution

Nico Battich1-3, Thomas Stoeger1-3 & Lucas Pelkmans

Nature Methods | Vol.10 No.11 | November 2013 | 1127
Summary

CYTOMINE: a rich internet application

- Uses generic software design, web services, and distributed processing using machine learning

- Fosters collaboration between pathologists, life scientists, and computer scientists
  - Eases sharing of whole slides and annotations
  - Speeds up large-scale image quantifications
  - Offers mechanisms to integrate novel algorithms / image formats

- Since 2011: ~100 users, 150 projects, >12K images, >100M « objects »
Acknowledgments


- CYTOMINE software beta-testing, etc.: Didier Cataldo, Natacha Rocks, Fabienne Perin, Christine Fink, Sandrine Bekaert, Myriam Remmelink, Caroline Degand, Isabelle Salmon, Sandrine Rorive, Audrey Voncken, Jessica Aceto, Yoann Curé, Benoist Pruvot, Marc Muller, Natacha Leroi, ...

Raphaël Marée is funded by GIGA FEDER grant and the CYTOMINE (2010-2014) research grant n° 1017072 of the Wallonia (DGO6). Benjamin Stévens is funded by SMASH spin-off grant n° 1217606 of the Wallonia.

www.montefiore.ulg.ac.be/~maree/ www.cytomine.be
Related publications

- Marée et al., "A rich internet application for remote visualization and collaborative annotation of digital slide images in histology and cytology". BMC Diagnostic Pathology, 8(Suppl 1):S26, 30th September 2013
