

# Evaluating Quality of and Enhancing Optical Chemical Structure Recognition Tools

**Nikolay Kapralov**

Scientific Advisors:

**Alexey Gurevich**, CAB, SPbSU

**Hosein Mohimani**, CSE, UCSD

December 12<sup>th</sup>, 2015

## Previously on...

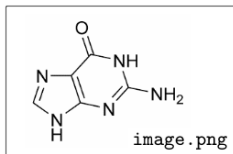
- I. Initial setup
- II. OCSR tools' contest — OSRA was chosen
  - Generate test dataset and write checking script
  - Study existing OCSR tools and choose the best
- III. Test data — first look at errors
  - Run it on test dataset
  - Analyse errors & try to improve identification
- IV. Real data — collecting errors
  - Run it on small real dataset
  - Analyse errors & try to improve identification again
- V. Run it on full real dataset and estimate the error rate

---

### Main goal

find errors → correct them → estimate quality

# Image processing with OSRA



**OSRA**  
atoms & bonds  
recognition

HO2C ▶ COOH  
CH2OH ▶ HOH2C  
spelling.txt

HOH2C ▶ CO  
C3H7 ▶ CCC  
superatom.txt

...c(=O)[nH]c...  
SMILE format

**Issue:** unrecognized symbols may occur:

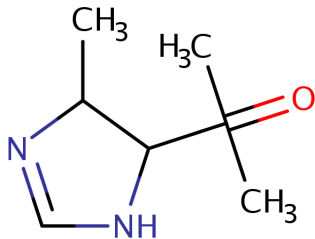
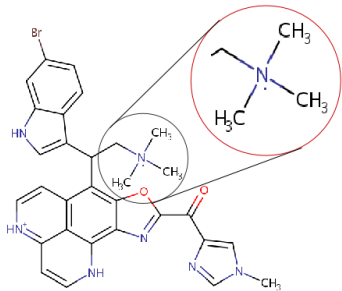
HoH2C --> ??

HOCH2 --> ??

**Solution:** remove hydrogen & convert  
to uppercase.

# Atom charge

**Issue:** OSRA may fail to detect atom charge.



**Solution:** store expected degrees for each atom type, correct or mark incorrect recognitions.

C	N	O	H	Cl	Br
4	3	2	1	1	1

Table 1: Expected degrees of atom degrees

# Quality assessment

Atom quality  $q_i$ :

Value	Description
0	unrecognized atom
0.5	wrong degree
0.75	spelling correction
1	otherwise

Overall quality  $Q$ :

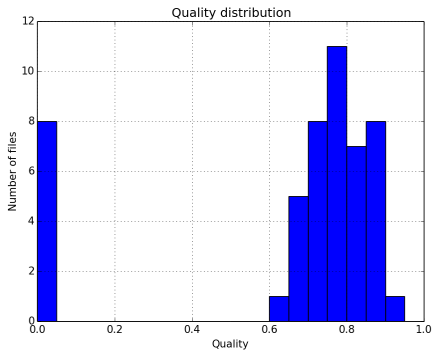
$$Q = \frac{1}{N} \sum_{i=1}^N q_i$$

# Output format

report.txt

filename	unrec	spell	deg	qual
00000102.png	5	10	0	0.6
00000103.png	1	3	1	0.7

- ▼ job
  - ▶ all
  - ▶ good
  - ▶ tmp
  - report.txt
  - result.png



## Future plans

- Quality metric improvement & threshold value adjustment for better classification of recognition results.
- Add some situational tests for specific cases (e.g. limits for cycle length)

## Acknowledgements

### **Alexey Gurevich**

for his guidance & mentoring during this project development.

# Questions?

Thanks for your attention!