Gene mention normalization using GNAT and LINNAEUS

Illés Solt, **Martin Gerner**, Philippe Thomas, Goran Nenadic, Casey M. Bergman, Ulf Leser, Jörg Hakenberg

Budapest University of Technology and Economics

MANCHESTER 1824

HUMBOLDT-UNIVERSITÄT ZU BERLIN

ASU ARIZONA STATE UNIVERSITY

System overview

- Species name recognition
- Gene name recognition
- Disambiguation
  - Context profile matching and filtering
  - Species disambiguation
- False positive filtering
  - Approximate string matching
  - Reliability score adjustments

Species detection

- Primarily performed by LINNAEUS[1]
- LINNAEUS "proxy" dictionary additions: cell lines[2], genera
  - Example: HeLa -> Human
  - Example: Drosophila -> *D. melanogaster*
- Species MeSH annotations loaded from MEDLINE

1: Gerner M, Nenadic G, Bergman CM: **LINNAEUS: a species name identification system for biomedical literature.** *BMC Bioinformatics* 2010, **11:**85.

2: Sarntivijai S, Ade AS, Athey BD, States DJ: **A bioinformatics analysis of the cell line nomenclature.** *Bioinformatics* 2008, **24:**2760-2766.

Gene name dictionaries

- Constructed from Entrez Gene and UniProt
  - CD95R → (CD|Cd|cd)[–]?95[–]?(R|r)(eceptor)?
  - Hfn-3beta → (HFN|Hfn|hfn|HfN)[–]?(3|III|iii)[–]?(B|b|)(eta)?
- Implemented as network services, using deterministic finite-state automatons (DFAs)
- Limited to 32 species (chosen based on MEDLINE mention frequency)
- Dictionaries are chosen based on the species discussed in the document.

Context filters for GN

- FP filtering: Rule-based filtering based on immediate context
  - Example: EPC *culture*

- Disambiguation: mention contexts are matched against gene term profiles (function, mutations, associated diseases and tissues, etc.)

treatment-related effect on the level of mRNA for proteins known to be involved in the control of hepatocyte cell division or apoptosis (e.g. P21, Cyclin D1, PCNA, CDKN1A). Furthermore, there was minimal indication of oxidative stress. Thus, there was no evidence

Comments

- ***FUNCTION***: Removal of H(2)O(2), oxidation of toxic reductants, biosynthesis and degradation of lignin, suberization, auxin catabolism, response to environmental stresses such as wounding, pathogen attack and oxidative stress. These functions might be dependent on each isozyme/isoform in each plant tissue.
- ***FUNCTION***: Might function as heat shock-like defense protein. May be implicated in the systemic acquired resistance response.
- ***CATALYTIC ACTIVITY***: Donor + H_2O_2 = oxidized donor + 2 H_2O.
- ***COFACTOR***: Binds 1 heme B (iron-protoporphyrin IX) group per subunit *(By similarity)*.
- ***COFACTOR***: Binds 2 calcium ions per subunit *(By similarity)*.
- ***TISSUE SPECIFICITY***: Preferentially expressed in roots and leaves, slightly in stems.
- ***DEVELOPMENTAL STAGE***: Up-regulated during leaf development.
- ***INDUCTION***: Late induced after mechanical wounding. Enhanced expression following incompatible bacterial pathogen attack. Expressed under a diurnal rhythm (circadian clock control).
- ***MISCELLANEOUS***: There are 73 peroxidase genes in A.thaliana.
- ***SIMILARITY***: Belongs to the peroxidase family. Classical plant (class III) peroxidase subfamily.

process (3)			
☐		response to oxidative stress	GO:0006979
☐		hydrogen peroxide catabolism	GO:0042744
☐		rhythmic process	GO:0048511
function (7)			
☐		peroxidase activity	GO:0004601
☐		peroxidase activity	GO:0004601
☐		peroxidase activity	GO:0004601
☐		iron ion binding	GO:0005506
☐		calcium ion binding	GO:0005509
☐		oxidoreductase activity	GO:0016491
☐		metal ion binding	GO:0046872

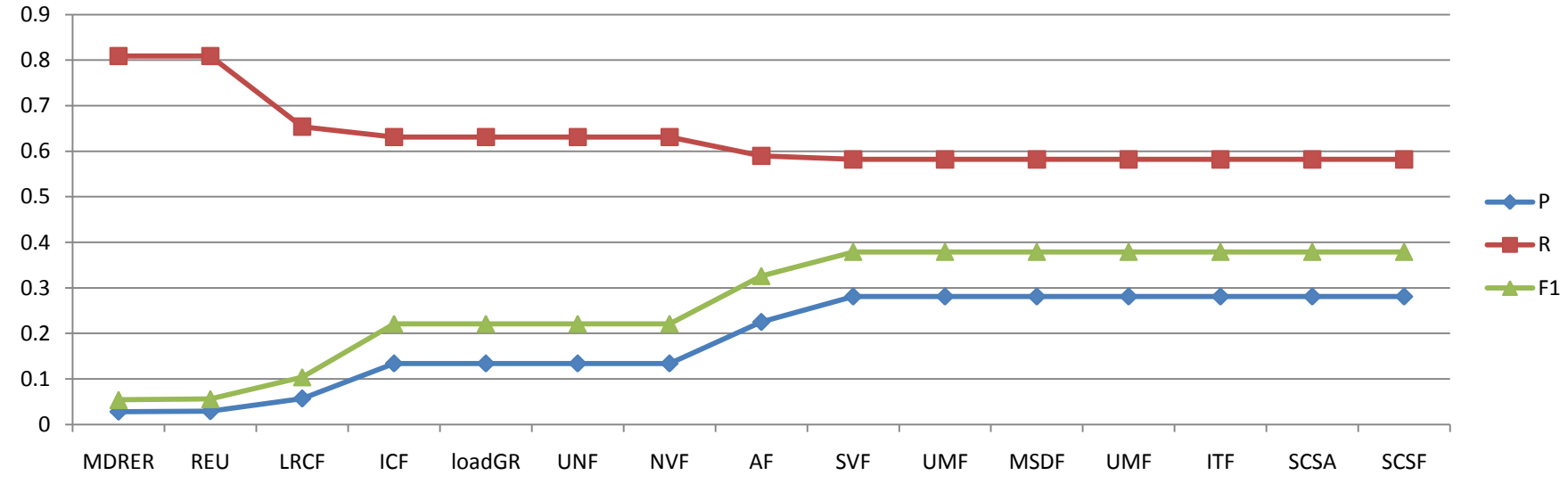
Additional filters

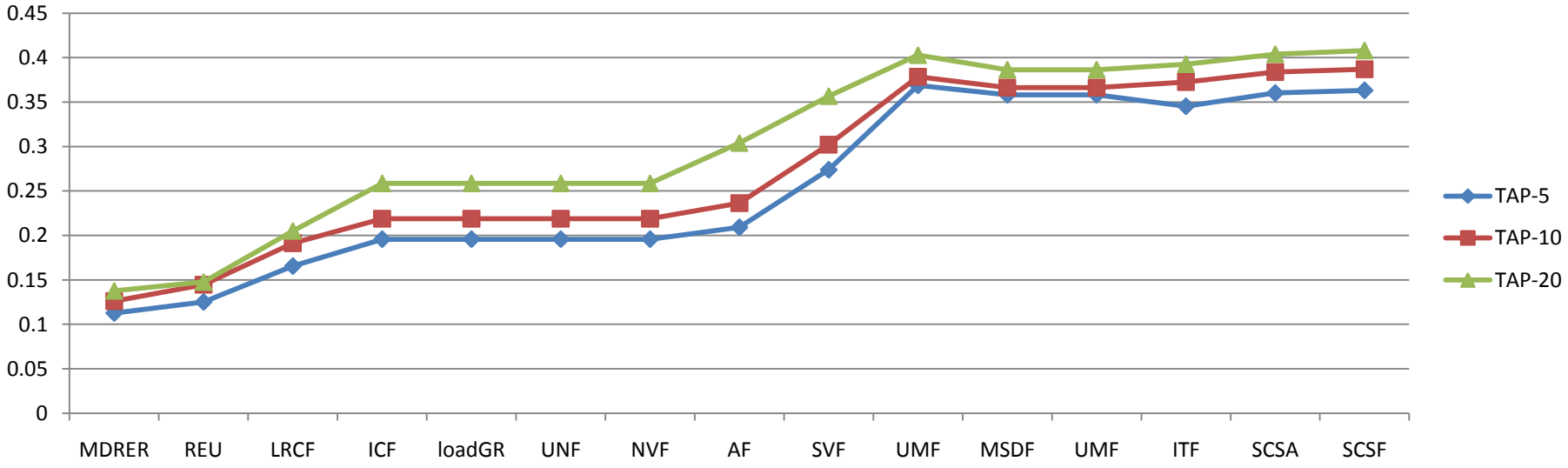
- Species disambiguation
- String similarity (vs unexpanded dictionaries)
- Score mentions based on italicized term occurences in PubMed Central
- Reliability threshold filter

Training results (manual set)

- Species:
  - 95% of gene entries covered by the 32 dictionaries
- TAP
  - TAP-5: 0.363
  - TAP-20: 0.408

Training results (2)

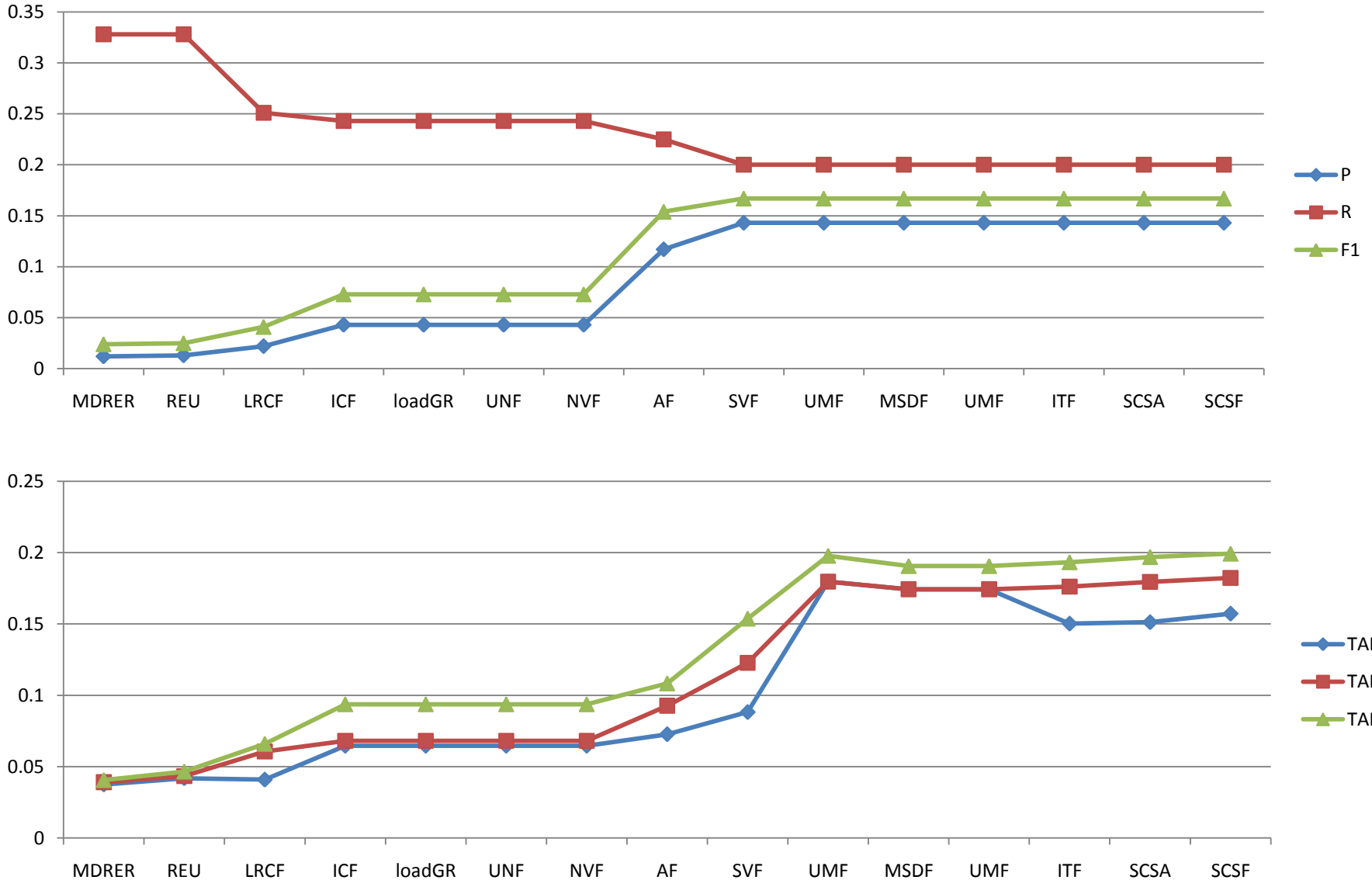




Test results (manual set)

- Species:
  - 44% of gene entries covered by the 32 dictionaries
- TAP
  - TAP-5: 0.157
  - TAP-20: 0.199
- Caused by very different species composition
  - Example: *Enterobacter sp. 638:* 22% of genes

Test results (2)



Conclusions

- TAP test results (0.157-0.199) significantly lower than training results (0.363-0.408)
  - Probable cause: species composition differences
- Most benefitial methods:
  - Context filters
  - Similarity test for recognized terms
  - Species disambiguation
- Going forward: reduce species-specificity

Acknowledgements

- IS: Alexander von Humboldt Foundation
- MG: University of Manchester, BBSRC
- PT: German Federal Ministry of Education and Research (BMBF)
- UL: Humboldt-University of Berlin
- GN, CMB: BBSRC
- JH: Arizona State University

GNAT: http://gnat.sourceforge.net (coming soon)
LINNAEUS: http://linnaeus.sourceforge.net (now with proxy dictionaries)