# Accessing biological data as Prolog facts

Nicos Angelopoulos and Jan Wielemaker

nicos.angelopoulos@sanger.ac.uk

jan@swi-prolog.org

Cancer Genome Project, Sanger Institute, Cambridge

CWI, Amsterdam, Netherlands

# the big picture

new-wave AI (for small size players)

- high level of abstraction
- open source: available and functioning
- ability to reason/program with large scale data

application areas:

- computational biology, bioinformatics
- data science
- social media data analysis
- recommender systems

# SWI-Prolog packs: open source for LP

Infrastructure for user specific libraries

`http://eu.swi-prolog.org/pack/list`

235 "packs"

```
?- pack_install('PACK').
?- pack_rebuild('PACK').
```

includes (versioned) pack dependency resolution

# introduction

*bio_db*

is an SWI-Prolog pack for serving biological data

- high-quality data
- data from primary sources
- convenience to end-user
- encourage use of Prolog
  in bioinformatics and computational biology

# key features

- data as Prolog facts
- served from flat files (and bytecode precompiles), or
- RocksDB (facebook), Berkeley DB, SQLite databases
- on-demand downloading from server
- maps between biological products
- interaction databases

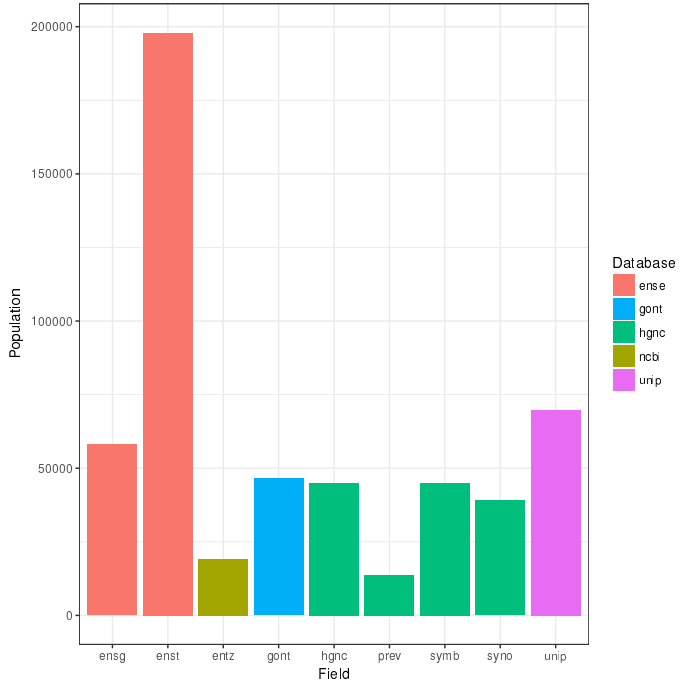
# availability

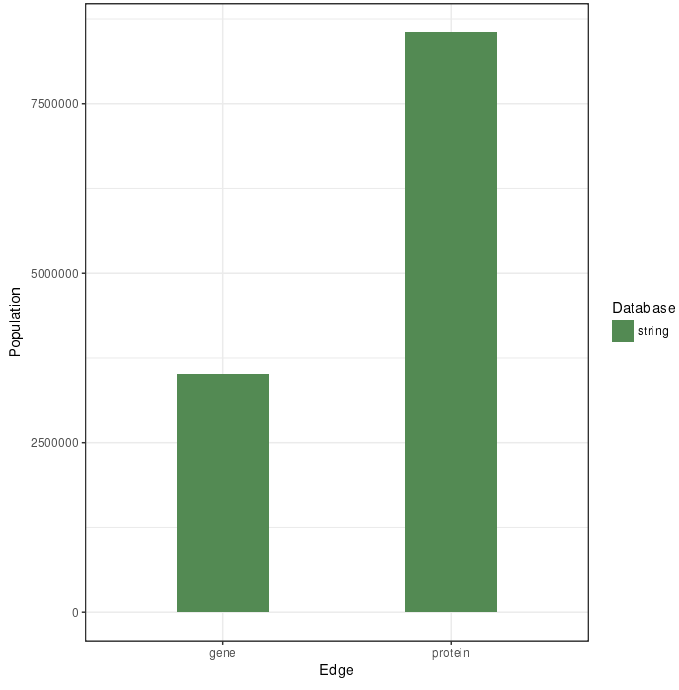
```
?- pack_install(bio_db).

?- debug(bio_db).
?- bio_db_interface(Iface).
Iface = prolog.

?- map_hgnc_prev_symb(Prev,Symb).
...
%Loading prolog db:.../map_hgnc_prev_symb.pl
Prev = 'A1BG-AS',
Symb = 'A1BG-AS1';
Prev = 'A1BGAS',
Symb = 'A1BG-AS1'...
```

# database resources

| Database | Abbv. | Description |
|---|---|---|
| HGNC | hgnc | HUGO Gene Nomenclature Committee |
| NCBI/entrez | entz | Nat. Center for Biot. Inf. |
| Uniprot | unip | Universal Protein Resource |
| GO | gont | Gene Ontology |
| Interactions database | | |
| String | string | protein-protein interactions |

# database populations



# map relations

translate between products

- gene <-> protein
- gene name <-> gene identifier

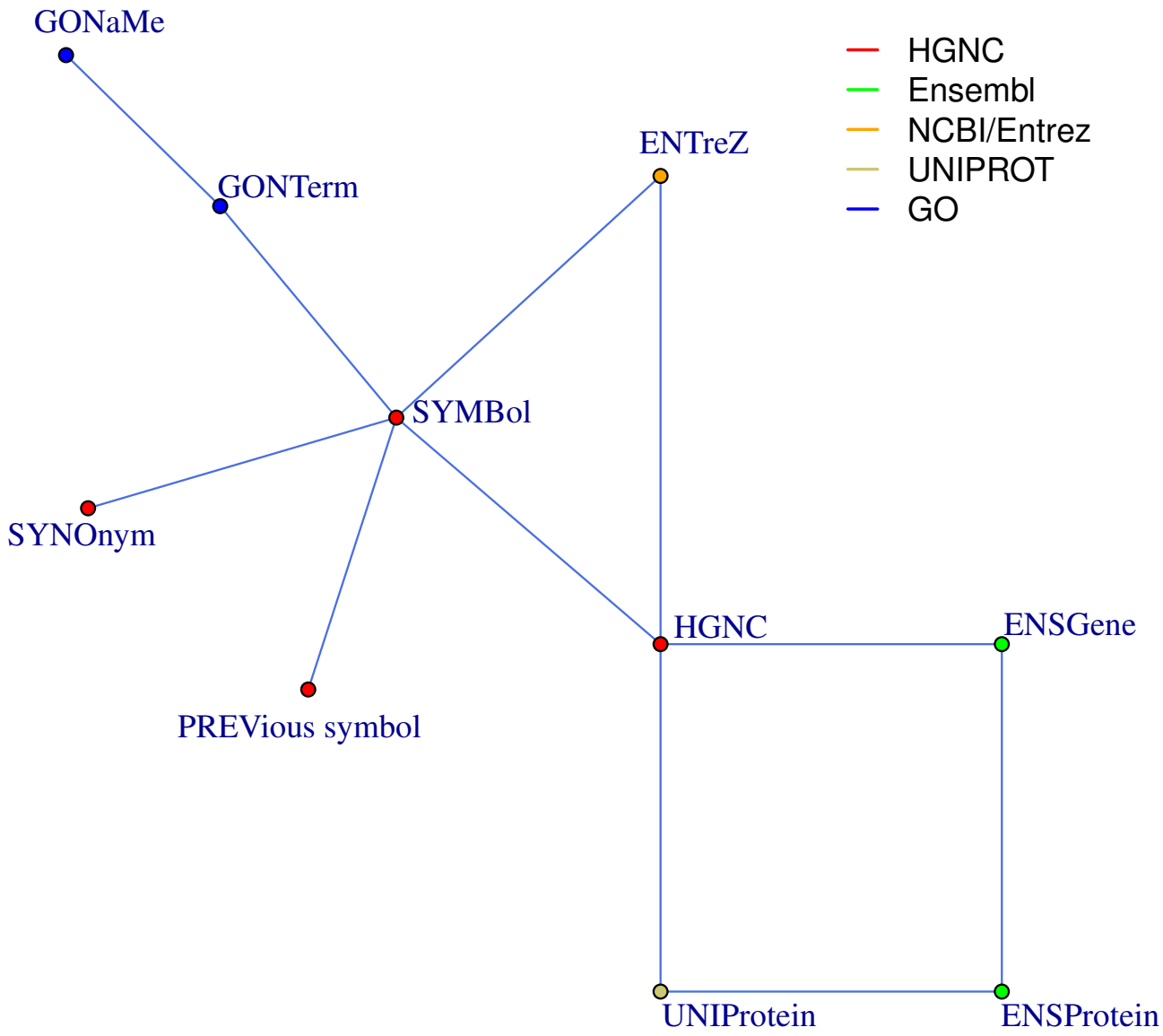
map products to groups

- gene <-> GO term

name convension: map_<DB>_<From>_<To>

- map_hgnc_hgnc_symb(19295, 'LMTK3').
- map_gont_symb_gont('LMTK3', 'GO:0003674').

# key map relations



# gene ontology terms for LMTK3

```
lmtk3_go :-
   map_gont_symb_gont('LMTK3', Gont),
   findall(Symb,
      map_gont_gont_symb(Gont,Symb),
   Symbs),
   map_gont_gont_gonm(Gont, Gonm),
   sort(Symbs,Oymbs), length(Oymbs, Len),
   write(Gont-Gonm-Len), nl, fail.

lmtk3_go.
```

# gene ontology terms for LMTK3

| GO term | GO name | population |
|---|---|---|
| GO:0003674 | molecular_function | 764 |
| GO:0004674 | protein serine/threonine kinase activity | 340 |
| GO:0004713 | protein tyrosine kinase activity | 89 |
| GO:0005524 | ATP binding | 1488 |
| GO:0005575 | cellular_component | 497 |
| GO:0006468 | protein phosphorylation | 557 |
| GO:0010923 | negative regulation of phosphatase activity | 53 |
| GO:0016021 | integral component of membrane | 200 |
| GO:0018108 | peptidyl-tyrosine phosphorylation | 131 |

# weighted graphs

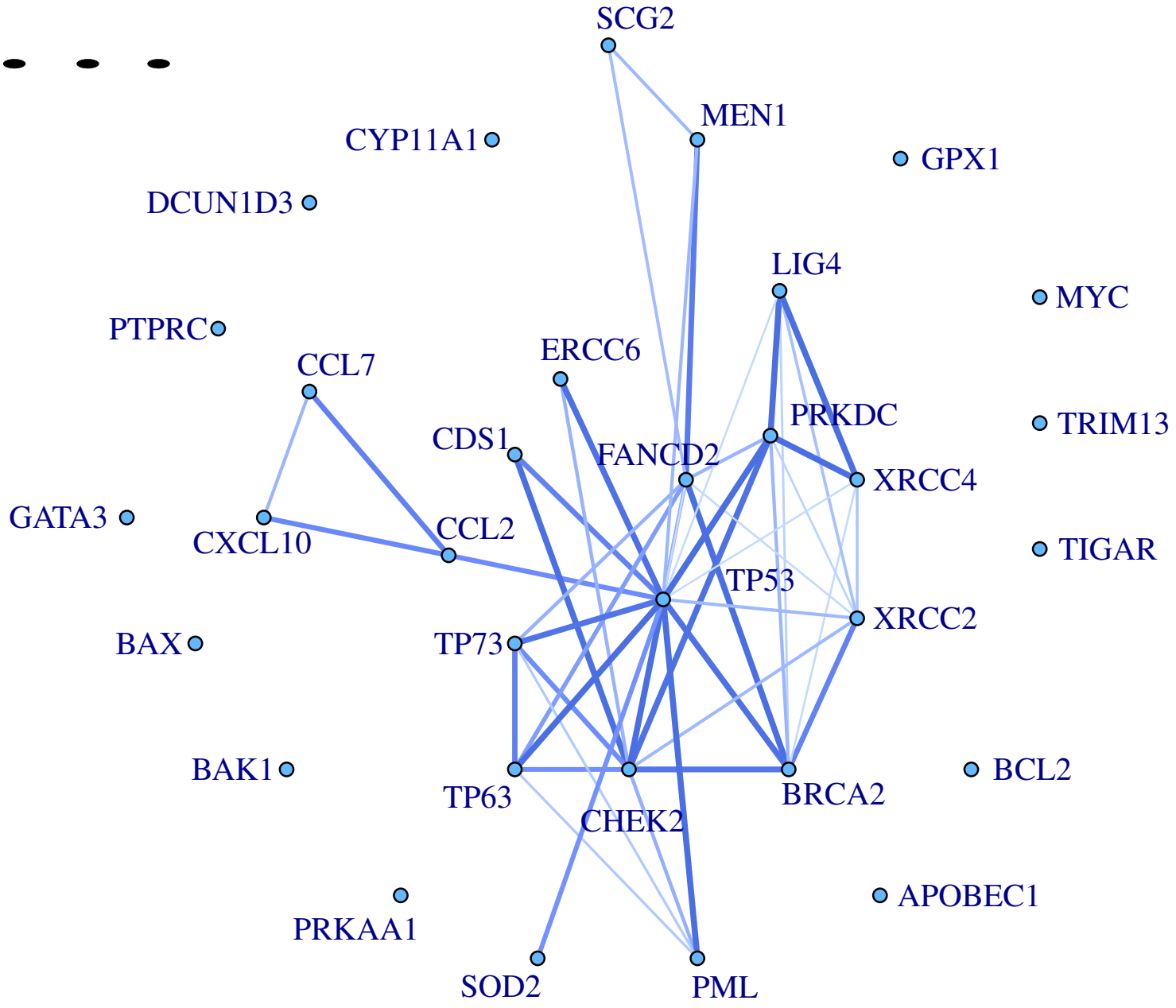
String database of protein-protein interactions.
Weight is strength of belief in physical interaction between 2 genes ($0 \leq i < 1000$).

- edge_string_hs_symb('AATK', 'LMTK3', 203).

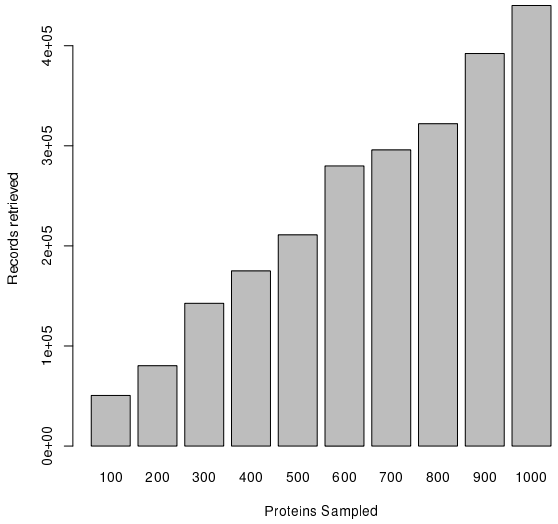
# graph construction

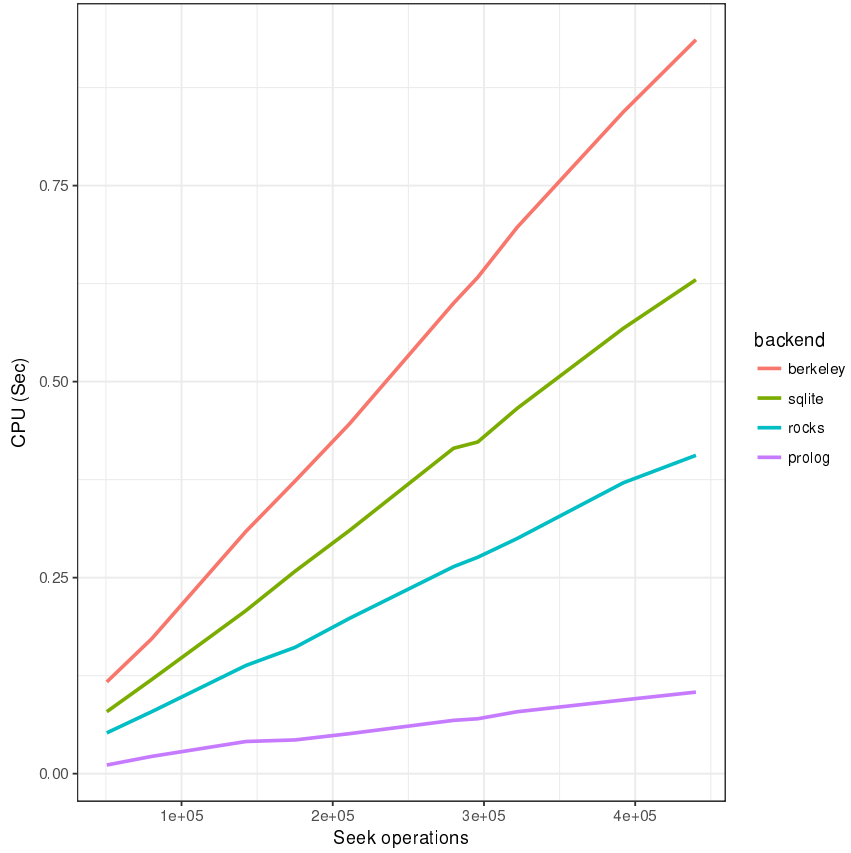
```
go_term_graph(GoTerm,Min,Graph):-
   findall( Symb, map_gont_gont_symb(Gont,Symb), Symbs ),
   findall( Symb1-Symb2:W, ( member(Symb1,Symbs),
                             member(Symb2,Symbs),
                             edge_string_hs_symb(Symb1,Symb2,W),
                             Lim < W ),
                                                  Graph ).
```

# String net for GO:10332



# relative performance



# loading and disk

Loading *edge_string_hs/3*

Prolog 190 sec

convert 207 sec

QLF 4 sec !

Disk space for *edge_string_hs/3*

qlf: 224

rocksdb: 229

bdb: 373

prolog: 481

sqlite: 1100

# web-page

```
?- bio_db_version(Vers, Date).
```

| Vers | Date | |
|---|---|---|
| 0:6:0 | *date*(2016, 10, 13) | 1 |

```
?- once(map_hgnc_hgnc_symb(_,_)),
   bio_db_info(map_hgnc_hgnc_symb/2, Key, Value).
```

| Key | Value | |
|---|---|---|
| interface | prolog | 1 |
| source_url | 'ftp://ftp.ebi.ac.uk/pub/databases/genenames/hgnc_complete_set.txt.gz' | 2 |
| datetime | *datetime*(2016, 9, 10, 0, 2, 14) | 3 |
| data_types | *data_types*(integer, atom) | 4 |
| unique_lengths | *unique_lengths*(44266, 44266, 44266) | 5 |
| relation_type | *relation_type*(1, 1) | 6 |
| header | *row*('HGNC ID', 'Approved Symbol') | 7 |

## Map and edge predicates

```
?- edge_string_hs(EnsP1, EnsP2, W).
```

| EnsP1 | EnsP2 | W | |
|---|---|---|---|
| 'ENSP00000000233' | 'ENSP00000003084' | 150 | 1 |

*3.395 seconds cpu time*

Next | 10 | 100 | 1,000 | Stop

```
?- edge_string_hs_symb(Symb1, Symb2, W).
```

# piece-meal prolog bioinformatics

| | | |
|---|---|---|
| Real | 261 | Swi/Yap <-> R interface |
| bio_db | 27 | this pack |
| pubmed | 19 | access pumed citation records |
| proSQLite | 314 | Swi/Yap <-> SQLite interface |
| db_facts | 106 | Swi/Yap facts <-> SQLite relations interface |
| wgraph | 21 | graph visualisation via R functions |
| silac | | functional analysis of quantative proteomics |

versus the more holistic

blip : `http://www.blipkit.org/`

# bottom-line

key-points

- extending Prolog relations to huge fact bases
- multiple back-ends
- re-usable techniques
- enables powerful analysis of biological datasets

future work

- pathway databases such as Reactome
- other back-ends (ODBC)
- web-analysis workflows
- generalise to non-biological datasets