Working with biological databases

Nicos Angelopoulos and Georgios Giamas

nicos.angelopoulos@imperial.ac.uk

Department of Surgery and Cancer, Imperial College, London

introduction

bio_db

is an SWI-Prolog library/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

- biological data as Prolog relations
- served from fact files, or
- SQLite databases
- on-demand downloading from server
- maps between biological products
- interaction databases

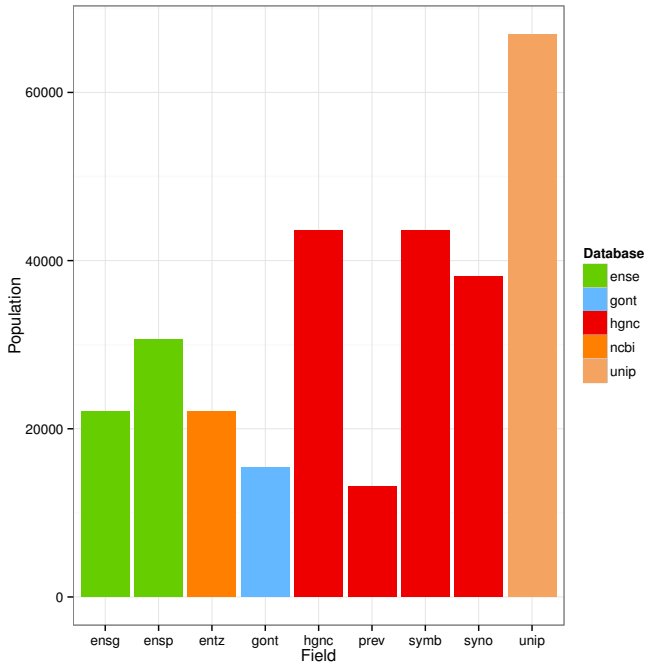
availability

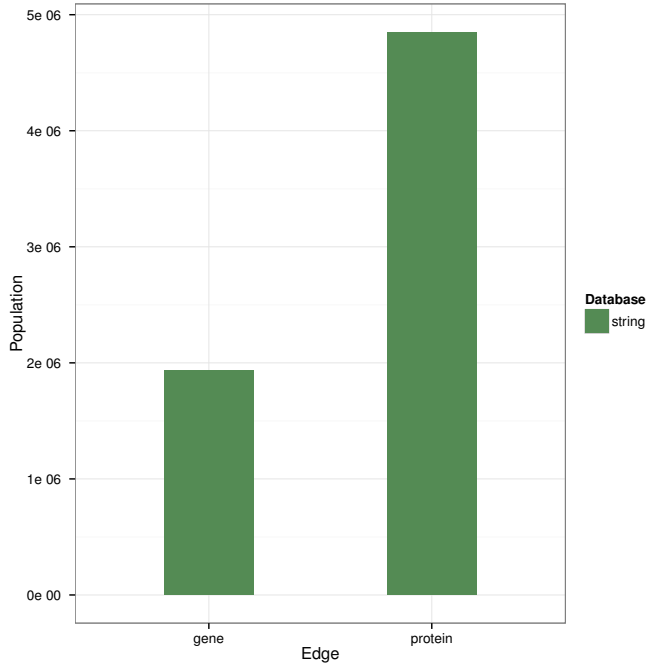
```
?- pack_install(bio_db).

?- library( 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
		genenames.org
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

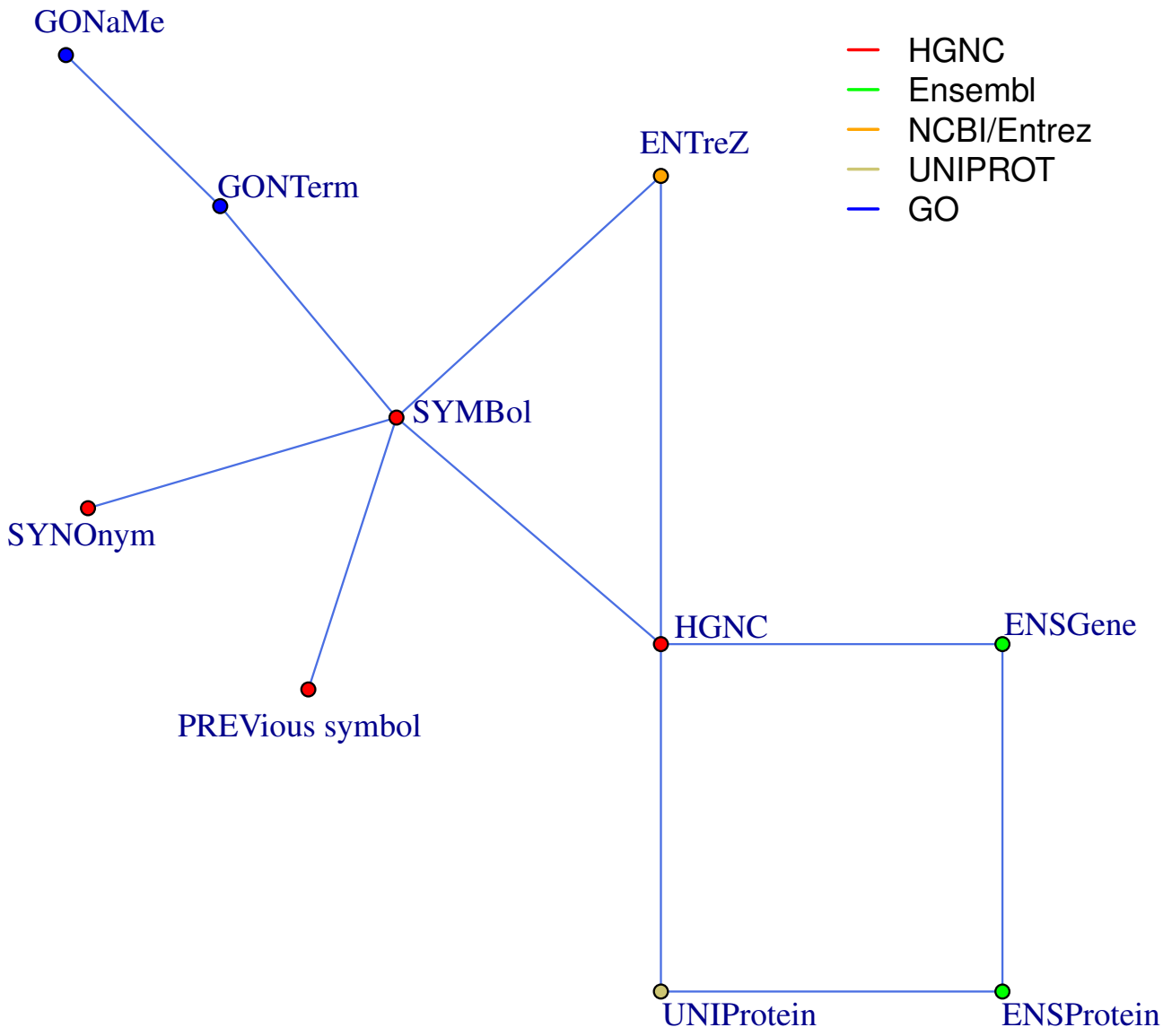
map products to groups

- 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

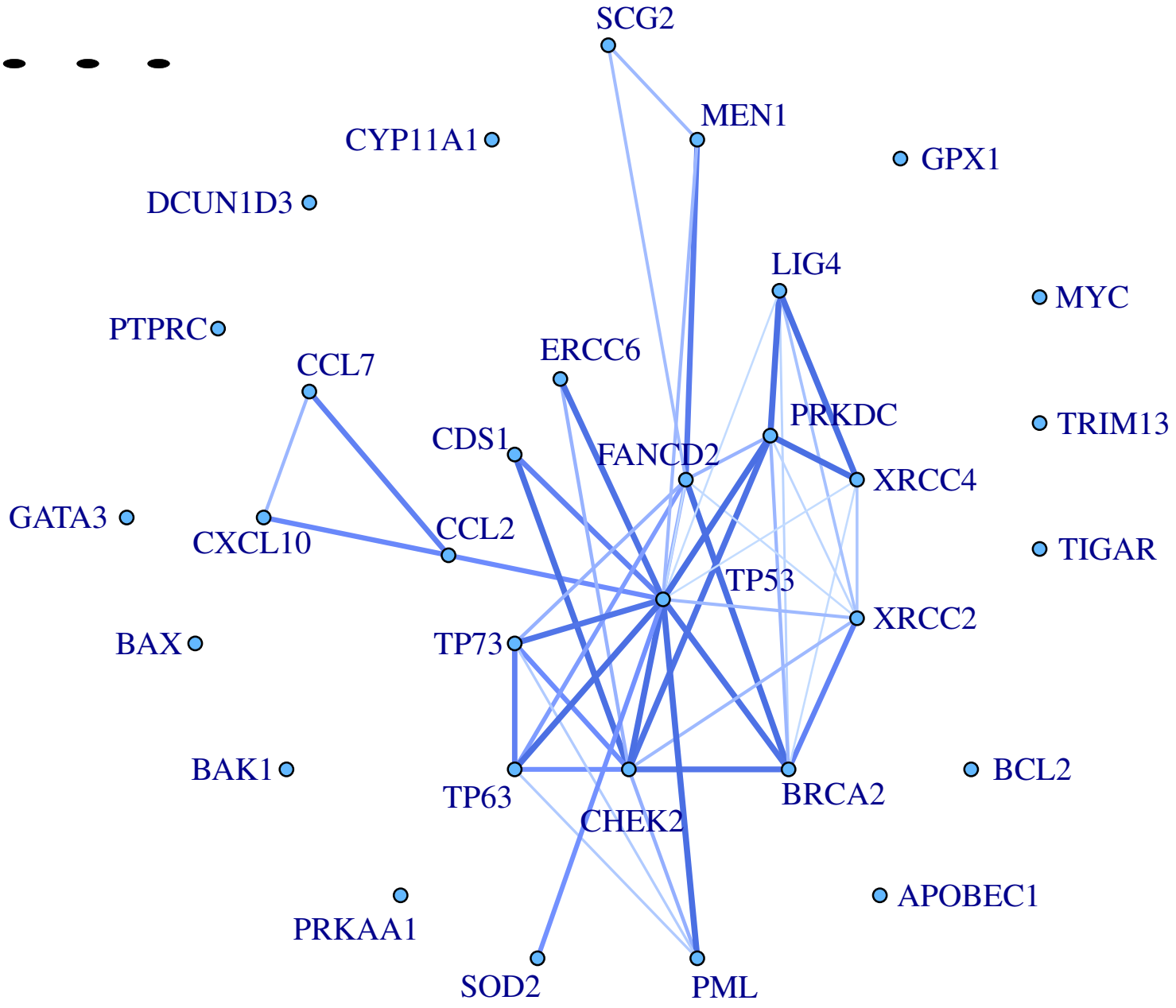
String database of protein-protein interactions.
Weight is strength of belief in physical interaction between 2 genes ($0 \leq i < 1000$).

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

key map relations

```
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



piece-meal prolog bioinformatics

Real	147	Swi/Yap <-> R interface
proSQLite	180	Swi/Yap <-> SQLite interface
db_facts	61	DB tables as Prolog facts
bio_db	5	biological databases
pubmed	16	access pumed citation records
wgraph	5	graph visualisation via R functions
silac		functional analysis of quantative proteomics

versus the more holistic

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

bottom-line

key-points

- re-usable techniques
- high-quality, precise biological data
- infrastructure for logical bioinformatics.

future work

- gene ontology term relations: *is*, *part_of*
- pathway databases (Reactome, KEGG, biopax)