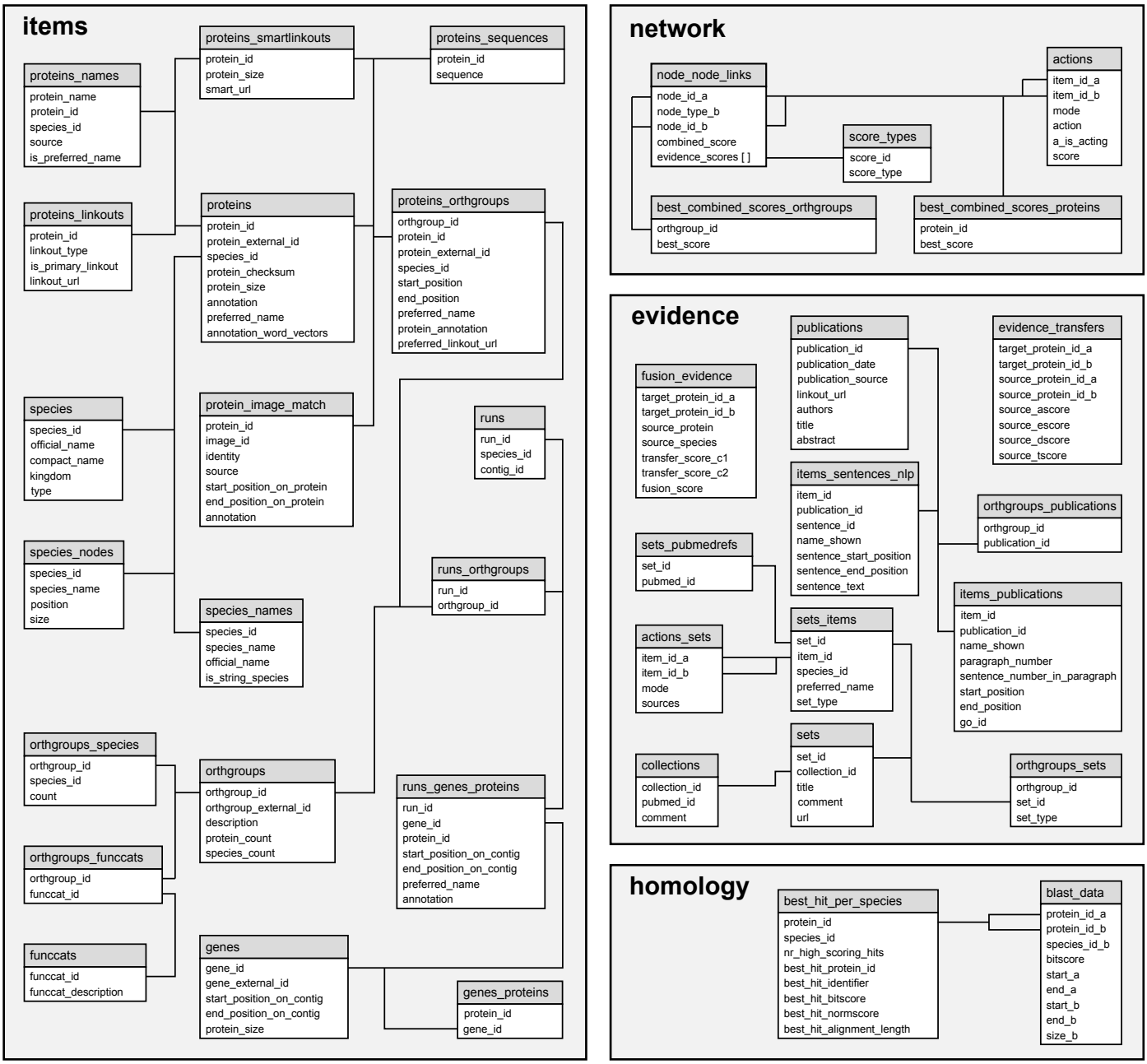
STRING v10.0 Database Schema



comments

- not all connections are shown here (for readability).
- the database is designed for speed: no constraints, or triggers - plus, tables are heavily de-normalized (i.e. redundant).
- the tables are served by PostgreSQL (currently in version 9.2.4).
- STRING is locus-based: only a single translated protein per locus is stored (usually the splice-form with the longest open reading frame).

a) internal identifiers in STRING are simple numericals; they usually remain stable until STRING's major version number changes.
b) species identifiers are numericals as well, but they are external and refer to the taxon-identifiers at NCBI.

c) external protein identifiers:

"83333.b1261"

taxon-id (species) — locus/protein accession (RefSeq, or Ensembl)

d) orthologous groups:

"COG0159"

COG: original groupings (E. Koonin, NCBI)
KOG: "eukaryotic orthologous group" (NCBI)
NOG: "non-supervised orthologous group" (STRING/eggNOG)

e) external gene identifiers:

"83333.b1261.NC_000913.1315246"

species — accession — chromosome — position