STRING v9.1 Database Schema

items

proteins_smartlinkouts
- protein_id
- protein_size
- smart_url

proteins_sequences
- protein_id
- sequence

proteins_names
- protein_name
- protein_id
- species_id
- source
- is_preferred_name
- linkout

meshterms
- mesh_id
- description

proteins_orthgroups
- orthgroup_id
- protein_id
- protein_external_id
- species_id
- start_position
- end_position
- preferred_name
- protein_annotation
- preferred_linkout_url

proteins_meshterms
- mesh_id
- protein_id

proteins
- protein_id
- protein_external_id
- species_id
- protein_checksum
- protein_size
- annotation
- preferred_name
- annotation_word_vectors

runs
- run_id
- species_id
- contig_id

species
- species_id
- official_name
- compact_name
- kingdom
- type

protein_image_match
- protein_id
- image_id
- identity
- source
- start_position_on_protein
- end_position_on_protein
- annotation

species_nodes
- species_id
- species_name
- position
- size

runs_orthgroups
- run_id
- orthgroup_id

species_names
- species_id
- species_name
- official_name
- is_string_species

orthgroups_species
- orthgroup_id
- species_id
- count

orthgroups
- orthgroup_id
- orthgroup_external_id
- description
- protein_count
- species_count

runs_genes_proteins
- run_id
- gene_id
- protein_id
- start_position_on_contig
- end_position_on_contig
- preferred_name
- annotation

orthgroups_funccats
- orthgroup_id
- funccat_id

funccats
- funccat_id
- funccat_description

genes
- gene_id
- gene_external_id
- start_position_on_contig
- end_position_on_contig
- protein_size

genes_proteins
- protein_id
- gene_id

network

node_node_links
- node_id_a
- node_type_b
- node_id_b
- combined_score
- evidence_scores [ ]

score_types
- score_id
- score_type

actions
- item_id_a
- item_id_b
- mode
- action
- a_is_acting
- score

best_combined_scores_orthgroups
- orthgroup_id
- best_score

best_combined_scores_proteins
- protein_id
- best_score

evidence

publications
- publication_id
- publication_date
- publication_source
- linkout_url
- authors
- title
- abstract

evidence_transfers
- target_protein_id_a
- target_protein_id_b
- source_protein_id_a
- source_protein_id_b
- source_ascore
- source_escore
- source_dscore
- source_tscore

fusion_evidence
- target_protein_id_a
- target_protein_id_b
- source_protein
- source_species
- transfer_score_c1
- transfer_score_c2
- fusion_score

items_sentences_nlp
- item_id
- publication_id
- sentence_id
- name_shown
- sentence_start_position
- sentence_end_position
- sentence_text

orthgroups_publications
- orthgroup_id
- publication_id

sets_pubmedrefs
- set_id
- pubmed_id

items_publications
- item_id
- publication_id
- name_shown
- paragraph_number
- sentence_number_in_paragraph
- start_position
- end_position
- go_id

actions_sets
- item_id_a
- item_id_b
- mode
- sources

sets_items
- set_id
- item_id
- species_id
- preferred_name
- is_database_set

collections
- collection_id
- pubmed_id
- comment

sets
- set_id
- collection_id
- title
- comment
- url

orthgroups_sets
- orthgroup_id
- set_id
- is_database_set

homology

best_hit_per_species
- protein_id
- species_id
- nr_high_scoring_hits
- best_hit_protein_id
- best_hit_identifier
- best_hit_bitscore
- best_hit_normscore
- best_hit_alignment_length

blast_data
- protein_id_a
- protein_id_b
- species_id_b
- bitscore
- start_a
- end_a
- start_b
- end_b
- size_b
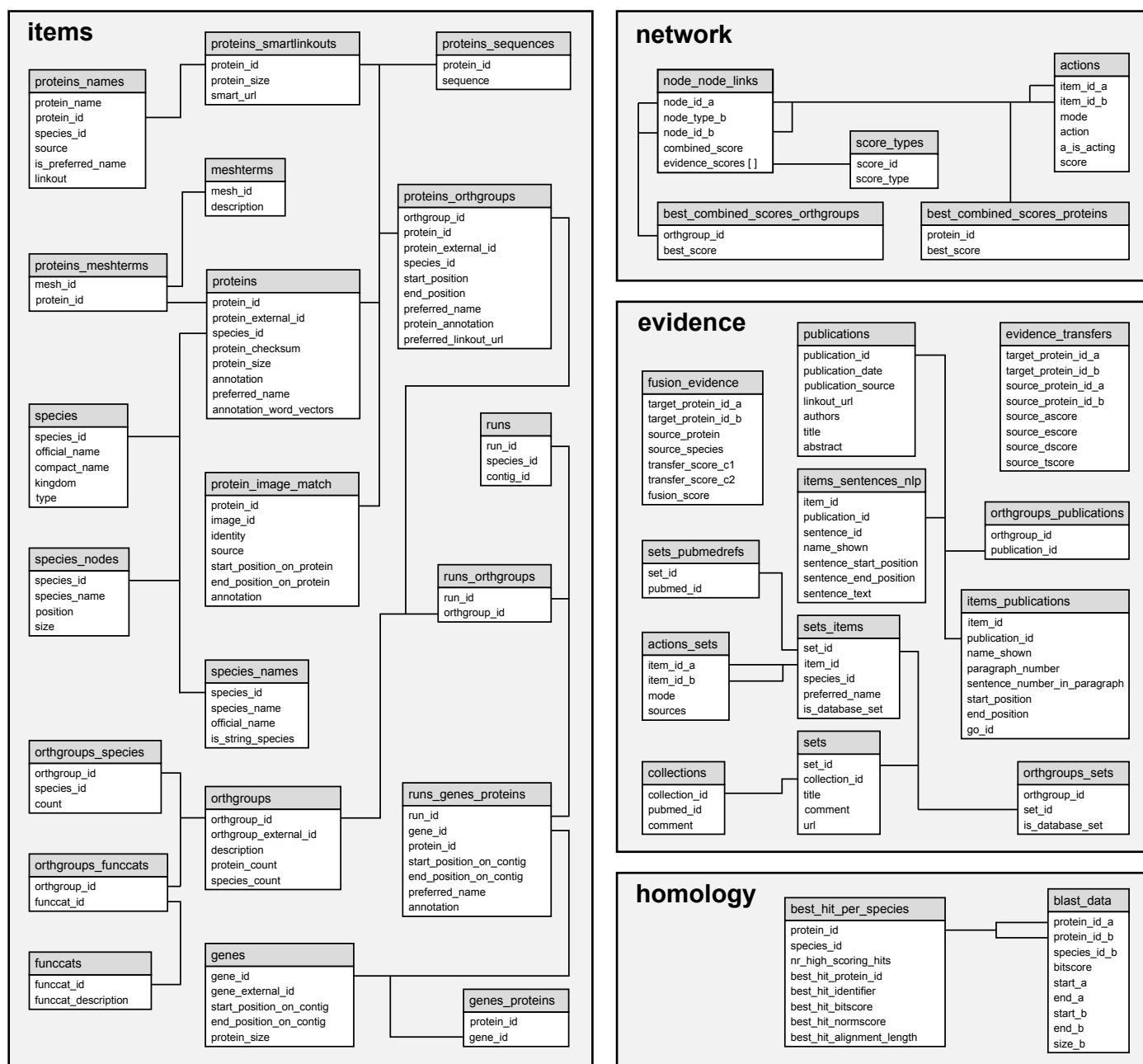
comments

- 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