The LudwigNR database: a comprehensive non-identical protein sequence database suitable for uninterpreted tandem mass spectrometry (MS/MS) database searches

Eugene A. Kapp[1*] and Christian Iseli[2*]

Joint Proteomics Facility, Ludwig Institute for Cancer Research (Melbourne Branch) & The Walter and Eliza Hall Institute of Medical Research, Victoria, Australia[1]
Computational Genomics Group, Ludwig Institute for Cancer Research (Lausanne Branch), Lausanne, Switzerland[2]

***Both authors contribute equally to this work.**

The LudwigNR database is a comprehensive non-identical protein sequence database suitable for the analysis of uninterpreted MS/MS data of a single peptide or a complete LC-MS/MS run. A number of key features distinguish this database from other large comprehensive protein sequence databases: 1) the inclusion of variant splice isoforms and minimal sequence redundancy within species; 2) consistent and standardized FASTA header lines appropriate for MS/MS search algorithms and 3) taxonomy enabled for species-specific searches. The database is updated weekly at the LICR Lausanne branch with quarterly updates made available by the Joint Proteomics Informatics group (LICR Melbourne branch) to the Australian Proteomics Computational Facility (APCF, (http://www.apcf.edu.au) as well as ProteinProspector server (http://jpsl.ludwig.edu.au). The current version (LudwigNR_Q410) comprises 13112897 sequence entries (5.52Gb uncompressed) (see Appendix 3 for revision history).

The procedure to create a suitable FASTA file for database searching involves the following:

The nr_prot.tar.gz file is downloaded from the Swiss Institute of Bioinformatics by public FTP (ftp://ftp.ch.embnet.org/pub/databases/nr_prot/). This file is produced by the Computational Genomics Group of the Ludwig Institute for Cancer Research. The compressed collection includes a readme file (see Appendix 1) as well as STATS file (see Appendix 2).

1) The individual .seq files are reformatted ensuring consistent header lines for each protein sequence entry as well as unique accession numbers for all entries. The NCBI Tax_Id number as well as organism name (for backward compatibility) is included in the header line.
2) The individual *.seq files are then grouped according to database source (e.g. ENSEMBL or plasmodb).
3) The LudwigNR database is then created by concatenating the grouped databases in the following order: Swiss-Prot, Swiss-Prot variant splice isoforms, Trembl, worm, yeast, ensembl, plasmo, gemmata, sludge, graminicola, truncatula and user-specific.

4) Final sanity checks are performed to ensure that all accession numbers are unique across all sequence entries.

Appendix 1: (latest Readme file)

The non-identical protein database is constructed from UniProt (SwissProt, SwissProt splice variants, TrEMBL, TrEMBL splice variants), yeastpep, wormpep, EnsEMBL peptides (http://www.ensembl.org/), SLUDGE, Gemmata, Graminicola, Truncatula and the PlasmoDB (http://plasmodb.org/) database.

Please note that duplicate proteins are removed only within the same taxa, based on NCBI's TaxID number.

The method is to produce non-redundant subsets of each member of the nr set, using a CRC method. The CRC method means we assign a checksum value to each protein sequence, then keep only one protein sequence per unique checksum value and per taxonomic ID (as assigned by the NCBI).
The probability of two different protein sequences having the same checksum is extremely small (but not null). We start with SwissPROT, from which we remove all the duplicates, and this produces swiss_nr.seq. We then take TrEMBL and remove all duplicates plus the sequences already found in swiss_nr.seq; this produces trembl_nr.seq. And the process is repeated for all the following members. Below is an example output of this method, which also gives the order in which the databases are processed, i.e., we keep all (almost) of SwissProt, then take entries from TrEMBL, etc.

```
DB                tag   description
-------------------------------------------------------------
swiss              sp   SwissProt + updates
swiss_varsplic  sp_vs   SwissProt splice variants
trembl             tr   TrEMBL + updates
wormpep            wp   WormPep from the Sanger center
yeastpep           yp   Yeast ORFs from Stanford
Aaegypti          ens   Aedes aegypti from EnsEMBL
Acarolinensis     ens   Anolis carolinensis from EnsEMBL
Agambiae          ens   Anopheles gambiae from EnsEMBL
Amellifera        ens   Apis mellifera from EnsEMBL
Btaurus           ens   Bos taurus from EnsEMBL
Cbriggsae         ens   Caenorhabditis briggsae from EnsEMBL
Celegans          ens   Caenorhabditis elegans from EnsEMBL
Cfamiliaris       ens   Canis familiaris from EnsEMBL
Choffmanni        ens   Choloepus hoffmanni from EnsEMBL
Cintestinalis     ens   Ciona intestinalis from EnsEMBL
Cjacchus          ens   Callithrix jacchus from EnsEMBL
Cporcellus        ens   Cavia porcellus from EnsEMBL
Csavignyi         ens   Ciona savignyi from EnsEMBL
Dmelanogaster     ens   Drosophila melanogaster from EnsEMBL
Dnovemcinctus     ens   Dasypus novemcinctus from EnsEMBL
Dordii            ens   Dipodomys ordii from EnsEMBL
Drerio            ens   Danio rerio from EnsEMBL
Ecaballus         ens   Equus caballus from EnsEMBL
Eeuropaeus        ens   Erinaceus europaeus from EnsEMBL
```

```
Etelfairi          ens   Echinops telfairi from EnsEMBL
Fcatus             ens   Felis catus from EnsEMBL
Gaculeatus         ens   Gasterosteus aculeatus from EnsEMBL
Ggallus            ens   Gallus gallus from EnsEMBL
Ggorilla           ens   Gorilla gorilla from EnsEMBL
Hsapiens           ens   Homo sapiens from EnsEMBL
Lafricana          ens   Loxodonta africana from EnsEMBL
Mdomestica         ens   Monodelphis domestica from EnsEMBL
Mlucifugus         ens   Myotis lucifugus from EnsEMBL
Mmulatta           ens   Macaca mulatta from EnsEMBL
Mmurinus           ens   Microcebus murinus from EnsEMBL
Mmusculus          ens   Mus musculus from EnsEMBL
Oanatinus          ens   Ornithorhynchus anatinus from EnsEMBL
Ocuniculus         ens   Oryctolagus cuniculus from EnsEMBL
Ogarnettii         ens   Otolemur garnettii from EnsEMBL
Olatipes           ens   Oryzias latipes from EnsEMBL
Oprinceps          ens   Ochotona princeps from EnsEMBL
Pcapensis          ens   Procavia capensis from EnsEMBL
Ptroglodytes       ens   Pan troglodytes from EnsEMBL
Ppygmaeus          ens   Pongo pygmaeus from EnsEMBL
Pvampyrus          ens   Pteropus vampyrus from EnsEMBL
Rnorvegicus        ens   Rattus norvegicus from EnsEMBL
Saraneus           ens   Sorex araneus from EnsEMBL
Scerevisiae        ens   Saccharomyces cerevisiae from EnsEMBL
Sscrofa            ens   Sus scrofa from EnsEMBL
Stridecemlineatus ens   Spermophilus tridecemlineatus from EnsEMBL
Tbelangeri         ens   Tupaia belangeri from EnsEMBL
Tguttata           ens   Taeniopygia guttata from EnsEMBL
Tnigroviridis      ens   Tetraodon nigroviridis from EnsEMBL
Trubripes          ens   Takifugu rubripes from EnsEMBL
Tsyrichta          ens   Tarsius syrichta from EnsEMBL
Ttruncatus         ens   Tursiops truncatus from EnsEMBL
Vpacos             ens   Vicugna pacos from EnsEMBL
Xtropicalis        ens   Xenopus tropicalis from EnsEMBL
Pberghei        plasmo   Plasmodium berghei ANKA from PlasmoDB
Pchabaudi       plasmo   Plasmodium chabaudi from PlasmoDB
Pfalciparum     plasmo   Plasmodium falciparum 3D7 from PlasmoDB
Pknowlesi       plasmo   Plasmodium knowlesi H from PlasmoDB
Pvivax          plasmo   Plasmodium vivax SaI-1 from PlasmoDB
Pyoelii         plasmo   Plasmodium yoelii 17XNL from PlasmoDB
Tgondii         plasmo   Toxoplasma gondii from PlasmoDB
sludge_aus      sludge   Australian sludge
sludge_us1      sludge   US sludge, Jazz Assembly
sludge_us2      sludge   US sludge, Phrap Assembly
Gemmata        gemmata   Gemmata obscuriglobus UQM 2246
Mgraminicola       jgi   Mycosphaerella graminicola from jgi
Mtruncatula       IMGA   Medicago truncatula from jcvi

The resulting files are in FASTA format, with a header of the form:
>DBtag|AccessionNb|OtherId Tax_Id=XXXX (GeneName)description[species]

But there are some exceptions:
 - for TrEMBL, OtherId is the EMBL accession number.
 - many TrEMBL entries have an associated gene name, but not all. The
other databases don't have it, and sometimes their description line
starts with something in ()...
 - yeastpep entries do not have an OtherId entry at all.
```

- the splice variants entries are a special case: the OtherId is the accession number of the original entry, and the AccessionNb is made unique for each splice variant.

Appendix 2: (latest Readme.stat file)

```
swiss_nr.seq             :  519348 total,  519127 kept,     221 dups.
swiss_varsplic_nr.seq    :   29373 total,   29369 kept,       4 dups.
trembl_nr.seq            : 11636205 total, 11414317 kept,  221888 dups.
wormpep_nr.seq           :   24705 total,     664 kept,   24041 dups.
yeastpep_nr.seq          :    5885 total,      92 kept,    5793 dups.
Aaegypti_nr.seq          :   16789 total,      27 kept,   16762 dups.
Acarolinensis_nr.seq     :   17672 total,   17632 kept,      40 dups.
Agambiae_nr.seq          :   13133 total,     485 kept,   12648 dups.
Amellifera_nr.seq        :   27755 total,   27626 kept,     129 dups.
Btaurus_nr.seq           :   26977 total,   20490 kept,    6487 dups.
Cbriggsae_nr.seq         :   14713 total,   12288 kept,    2425 dups.
Celegans_nr.seq          :   27975 total,     303 kept,   27672 dups.
Cfamiliaris_nr.seq       :   25559 total,   24947 kept,     612 dups.
Choffmanni_nr.seq        :   12435 total,   12396 kept,      39 dups.
Cintestinalis_nr.seq     :   19858 total,   19379 kept,     479 dups.
Cjacchus_nr.seq          :   45278 total,   43143 kept,    2135 dups.
Cporcellus_nr.seq        :   19774 total,   19464 kept,     310 dups.
Csavignyi_nr.seq         :   20143 total,   19980 kept,     163 dups.
Dmelanogaster_nr.seq     :   21309 total,     806 kept,   20503 dups.
Dnovemcinctus_nr.seq     :   14846 total,   14679 kept,     167 dups.
Dordii_nr.seq            :   15853 total,   15797 kept,      56 dups.
Drerio_nr.seq            :   28630 total,   16715 kept,   11915 dups.
Ecaballus_nr.seq         :   22641 total,   22344 kept,     297 dups.
Eeuropaeus_nr.seq        :   14592 total,   14551 kept,      41 dups.
Etelfairi_nr.seq         :   16562 total,   16533 kept,      29 dups.
Fcatus_nr.seq            :   15048 total,   15007 kept,      41 dups.
Gaculeatus_nr.seq        :   27576 total,   27211 kept,     365 dups.
Ggallus_nr.seq           :   22194 total,   20101 kept,    2093 dups.
Ggorilla_nr.seq          :   27473 total,   27067 kept,     406 dups.
Hsapiens_nr.seq          :   79063 total,   19426 kept,   59637 dups.
Lafricana_nr.seq         :   25622 total,   25576 kept,      46 dups.
Mdomestica_nr.seq        :   32541 total,   32376 kept,     165 dups.
Mlucifugus_nr.seq        :   16232 total,   16186 kept,      46 dups.
Mmulatta_nr.seq          :   36384 total,   35075 kept,    1309 dups.
Mmurinus_nr.seq          :   16319 total,   16271 kept,      48 dups.
Mmusculus_nr.seq         :   50068 total,   11648 kept,   38420 dups.
Oanatinus_nr.seq         :   26836 total,   26726 kept,     110 dups.
Ocuniculus_nr.seq        :   23910 total,   23459 kept,     451 dups.
Ogarnettii_nr.seq        :   15448 total,   15405 kept,      43 dups.
Olatipes_nr.seq          :   24661 total,   24387 kept,     274 dups.
Oprinceps_nr.seq         :   15993 total,   15927 kept,      66 dups.
Pcapensis_nr.seq         :   16101 total,   16068 kept,      33 dups.
Ppygmaeus_nr.seq         :   23533 total,   22958 kept,     575 dups.
Ptroglodytes_nr.seq      :   34142 total,   33040 kept,    1102 dups.
Pvampyrus_nr.seq         :   17053 total,   17017 kept,      36 dups.
Rnorvegicus_nr.seq       :   32971 total,   10824 kept,   22147 dups.
Saraneus_nr.seq          :   13192 total,   13126 kept,      66 dups.
Scerevisiae_nr.seq       :    6698 total,     111 kept,    6587 dups.
Sscrofa_nr.seq           :   19083 total,   17639 kept,    1444 dups.
Stridecemlineatus_nr.seq:   14830 total,   14815 kept,      15 dups.
```

```
Tbelangeri_nr.seq       :   15462 total,   15417 kept,      45 dups.
Tguttata_nr.seq         :   18191 total,   17911 kept,     280 dups.
Tnigroviridis_nr.seq    :   23118 total,   21129 kept,    1989 dups.
Trubripes_nr.seq        :   47841 total,   47490 kept,     351 dups.
Tsyrichta_nr.seq        :   13662 total,   13621 kept,      41 dups.
Ttruncatus_nr.seq       :   16598 total,   16562 kept,      36 dups.
Vpacos_nr.seq           :   11793 total,   11768 kept,      25 dups.
Xtropicalis_nr.seq      :   27710 total,   24570 kept,    3140 dups.
Pberghei_nr.seq         :   12235 total,     754 kept,   11481 dups.
Pchabaudi_nr.seq        :    5098 total,    4502 kept,     596 dups.
Pfalciparum_nr.seq      :    5512 total,     127 kept,    5385 dups.
Pknowlesi_nr.seq        :    5110 total,    5098 kept,      12 dups.
Pvivax_nr.seq           :    5435 total,    5430 kept,       5 dups.
Pyoelii_nr.seq          :    7724 total,      46 kept,    7678 dups.
Tgondii_nr.seq          :    7793 total,    6464 kept,    1329 dups.
sludge_aus_nr.seq       :   30590 total,   30381 kept,     209 dups.
sludge_us1_nr.seq       :   16840 total,   16004 kept,     836 dups.
sludge_us2_nr.seq       :   34254 total,   27467 kept,    6787 dups.
Gemmata_nr.seq          :    7989 total,    7763 kept,     226 dups.
Mgraminicola_nr.seq     :   10952 total,   10923 kept,      29 dups.
Mtruncatula_nr.seq      :   53423 total,   48865 kept,    4558 dups.
Overall                 : 13648311 total, 13112892 kept,  535419 dups.
```

Appendix 3: Revision History

Name	#of entries	Size (Gb)
LudwigNR_Q410	13112897	5.52
LudwigNR_Q310	12595433	5.31
LudwigNR_Q210	12085361	5.11
LudwigNR_Q110	11872793	4.83
LudwigNR_Q409	10752610	4.57
LudwigNR_Q309	9870917	4.19
LudwigNR_Q209	8777915	3.87
LudwigNR_Q109	8195435	3.62
LudwigNR_Q408	7647660	3.30
LudwigNR_Q308	7254532	2.95