Tutorial for AAI-profiler

1. Aims and scope

Whole-genome shotgun sequencing has propelled the re-evaluation of taxonomic classifications and the emergence of single-cell genomics is vastly expanding knowledge about biodiversity. Especially bacterial systematics evolves constantly. Meta-data in sequence databases may be using old synonyms, and some samples may be misclassified. Direct comparison of sequence data is a quicker way to get an overview of the taxonomy and phylogenetic relationships than searching the original literature on taxonomic classification.

AAI-profiler shows the species neighbours of a query species. The approach is intermediate between phylogenomic trees of a selected subset of species and taxonomic profiles in metagenomics. Phylogenomic trees approximate the pairwise distances of species, where distances are measured by ANI (Average Nucleotide Identity) or AAI (Average Amino-acid Identity). AAI-profiler shows the distances from the query species (sample) to all similar species in the sequence database. The sample is assumed to consist of one species (exceptions give rise to characteristic patterns to the AAI-profile). In metagenomics, the species composition of the sample is unknown. Taxonomic profiling maps each read (protein) to the nearest neighbour species in the database. Taxonomic profiling is a border case (k=1) for AAI-profiler, which searches for a large number (k=100) of sequence neighbours of each query protein.

The query species is represented by its proteome (protein sequences in FASTA format). AAI-profiler plots the AAI values between the query proteome and species in the Uniprot database. AAI-profiler compares amino acid sequences rather than nucleotide sequences, which makes it practical also with eukaryotic queries. Eukaryotic genomes are hundreds to thousands times longer than bacterial genomes. However, gene density in eukaryotic genomes is much lower, so that the size of a eukaryotic proteome is typically only ten times larger than a bacterial proteome.

AAI-profiler is powered by SANS and the processing time for a bacterial proteome is a few minutes and less than an hour for a eukaryotic proteome. The homology search detects neighbours down to about 50% sequence identity which shows neighbouring bacterial genera or mammalian families.

The main uses of AAI-profiler are in **explorative analysis** and **quality control** in selecting data sets for comparative genomics and phylogenomic or phylogenetic trees. AAI-profiler reports sequence-based distances from the query proteome to other species. One expects that taxa are monophyletic and therefore distances within a taxon should be smaller than distances between species from different taxa. Exceptions that can be detected using AAI-profiler include:

- misidentified species
- mislabelled multi-isolate samples
- contaminated samples
- corrupted data included in bacterial pan proteomes

Correct meta-data is important because many inference methods test the congruence of sequence trees with the species tree (taxonomy) assuming that species assignments of protein sequences are correct. Such applications are outside the scope of AAI-profiler but include:

- tree reconciliation to identify speciation and gene duplication events
- tree reconciliation to identify lateral gene transfer events
- LCA (last common ancestor) approach for taxonomic profiling in metagenomics
- assignment of the last common ancestor taxon to a cluster of sequences

AAI-profiler is available as a web server at http://ekhidna2.biocenter.helsinki.fi/AAI. The scripts an also be downloaded and run locally using remote databases.

2. Methods

We compute onesided and bidirectional Average Amino-acid Identity (AAI) profiles for a query proteome (protein sequences in FASTA format). We use SANSparallel to retrieve homologous proteins from Uniprot. The OS tag of Uniprot headers is used to retrieve species information. Taxonomic metadata is retrieved using DictServer. For each query protein, we retain the match to all database species with the highest bitscore. Onesided AAI profiles are based on a many-to-one mapping from query proteins to the target proteins of a database species. Multiplicity is the number of query proteins divided by the number of distinct target proteins. Bidirectional AAI profiles are based on a one-to-one mapping, where we exclude the match of a query protein to a database protein, if a higher scoring match exists for either of them. The counts of matches per species are tallied in sequence identity bins one percentage-point wide. Sequence identity is computed per aligned positions in the alignment returned by SANSparallel. The Average Amino-acid Identity (AAI) is the average of sequence identities of all matched pairs between the query proteome and a database species, where each query protein has unit weight. Query proteins with no matches (as reported by SANSparallel) have zero weight. Species are ranked based on the sum of sequence identities of their matched proteins.

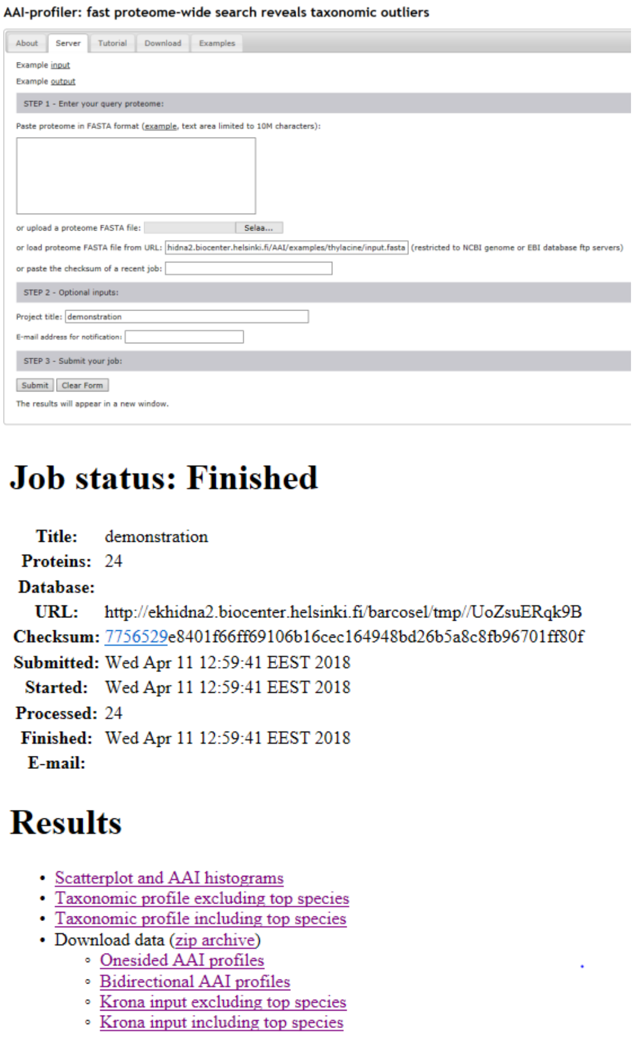
While AAI profiles are computed for all species in the database, for taxonomic profiles we retain only a single best match for each query protein and tally the species. Ties are broken arbitrarily. When the query species is already included in the database, it will dominate the taxonomic profile. Therefore we generate also a second taxonomic profile, which excludes hits to the top ranked species.

3. Web interface

3.1. Submission

The web submission form (Figure 1 top) has one required input field which is the FASTA file of protein sequences. Plain text or gzip is accepted. The file can be uploaded from the user's computer. Alternatively, the file can be uploaded from a hyperlink to the NCBI genomes ftp server, the EBI database ftp server, or the AAI-profiler host. The project title is an optional field which is echoed in the results page. Leave an email address if you wish to be notified by email when the job has finished.

After submission, the results page opens in a new window (Figure 1 bottom). The status fields changes from Queued to Running to Finished. While the job is running, the number of queries processed so far (in increments of 100) is shown in the Processed: field. The checksum is computed from the input file and used to identify duplicated requests, which use cached sequence database search results. When the job has finished, the Results section appears with links to various plots and downloadable taxonomic profiles.



AAI-profiler: fast proteome-wide search reveals taxonomic outliers

Example input

Example output

STEP 1 - Enter your query proteome:

Paste proteome in FASTA format (example, text area limited to 10M characters):

or upload a proteome FASTA file:

or load proteome FASTA file from URL: hidna2.biocenter.helsinki.fi/AAI/examples/thylacine/input.fasta (restricted to NCBI genome or EBI database ftp servers)

or paste the checksum of a recent job:

STEP 2 - Optional inputs:

Project title: demonstration

E-mail address for notification:

STEP 3 - Submit your job:

The results will appear in a new window.

Job status: Finished

Title:	demonstration
Proteins:	24
Database:	
URL:	http://ekhidna2.biocenter.helsinki.fi/barcosel/tmp//UoZsuERqk9B
Checksum:	7756529e8401f66ff69106b16cec164948bd26b5a8c8fb96701ff80f
Submitted:	Wed Apr 11 12:59:41 EEST 2018
Started:	Wed Apr 11 12:59:41 EEST 2018
Processed:	24
Finished:	Wed Apr 11 12:59:41 EEST 2018
E-mail:	

Results

- Scatterplot and AAI histograms
- Taxonomic profile excluding top species
- Taxonomic profile including top species
- Download data (zip archive)
    - Onesided AAI profiles
    - Bidirectional AAI profiles
    - Krona input excluding top species
    - Krona input including top species

Figure 1. Web submission form (top) and result page (bottom).

3.2. Visualization of results

AAI-profiler generates three types of plots: scatterplots of species, AAI histograms per species, and Krona plots of taxonomic profiles (Figure 2).

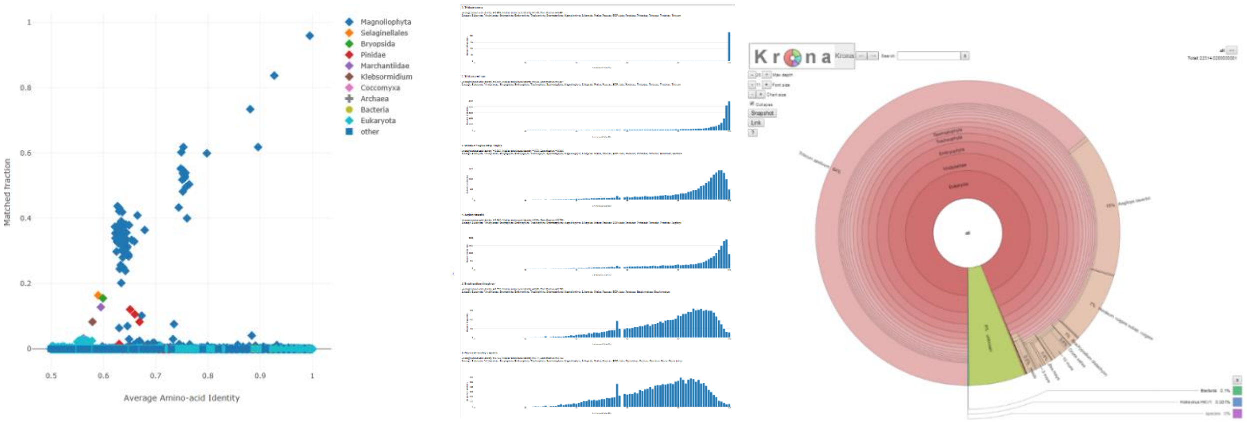


Figure 2. AAI-profiler scatterplot and AAI histograms (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/Triticum_urartu/AAI.html) and taxonomic profiling (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/Triticum_urartu/AAIkrona.html) for wild wheat *Triticum urartu*. The proteome was downloaded from the "Download sequences in FASTA format for **protein**" link of https://www.ncbi.nlm.nih.gov/genome/?term=txid4572[orgn].

Scatterplots

The plots are generated using Plotly. Select "Show closest data on hover" from the Plotly menu (upper right corner of scatterplot). Hover the mouse over data points to see the species labels. Species to the right are more closely related to the query than species further to the left. Data points are coloured according to genus (bacteria) or order (eukaryota). Eukaryotic species are marked by diamonds, bacteria by circles, archaea by crosses and anything else by squares. The horizontal axis shows the AAI (Average Amino-acid Identity) between the query and database species. The average is computed over the best match per database species over those query proteins, for which SANS reports a match. The vertical axis shows the "core fraction" which is the fraction of query proteins that have a match in the species. If the query proteome is present in the database (Uniprot), you see a dot near the top right corner (1.0, 1.0). Related species form a cloud to the left-and-down. More distantly related taxa have low AAI and low "matched fraction" because match counts are based on ~100 nearest hits in the database. At the bottom, there is a band of matches from species from which only individual proteins have been sequenced.

Barcharts

The plots are generated using Plotly. The histograms show the distribution of AAI values of the top ranked species. Species are ranked by the product of AAI and matched fraction, i.e., the sum of sequence identity values over all matched query proteins. The AAI distributions are skewed. Comparing species with fully sequenced genomes, the mode shifts to lower AAI values and the peak gets wider as the species diverge. Sometimes you see a low ranked species with a narrow peak at the right; its position in the ranking is then due to a low total count of matches.

Krona plots

The plots are generated using KronaTools. Krona plots allow the user to select different levels of the taxonomic hierarchy for inspection. The taxonomic profile of the query proteome is similar to what you would get from metagenomics tools. Each query protein is assigned to the species of the closest match (ranking by alignment bit score). Class unknown are query sequences with no hits in the protein database. Two versions are generated. If the query species is in the database, you obviously get an uninteresting profile of self-hits. For this reason, we generate a second taxonomic profile excluding the top ranked species (usually near coordinates (1.0, 1.0) in the scatterplot).

3.3. Data download

The AAI profile data contains the following tab-separated columns:

1. wsum, sum of sequence identities of the best match from each query protein to this species

2. count, number of query proteins that have a match in this species
3. target_count, number of proteins that have a match from this species
4. multiplicity, count divided by target_count
5. kingdom, the first level of the taxonomic lineage used to select the marker in scatterplots
6. genus, a selected level of the taxonomic lineage used to group and colour data points in scatterplots
7. species, as given by the OS= field of Uniprot entries
8. matched_fraction, count divided by the number of query proteins
9. average_identity, wsum divided by the number of query proteins
10. median, median of sequence identities of matched proteins
11. pide_bins, comma-separated list of counts of query proteins that have a best match to this species at sequence identity 0.00, 0.01, …, 1.00
12. lineage, taxonomic lineage from Uniprot

The format of Krona inputs is tab separated columns, where the first column is a weight and the taxonomic hierarchy, starting from the most general class, are given in the other columns. The weight is the sum of sequence identities of the best match in the database to each query protein.

4. Interpretation of results

Here, we discuss examples of AAI-profiler analyses. We use published data and corroborate many findings described in the scientific literature. The examples include good, bad and ugly cases.

4.1. Brush up your biology

The whole genome of the extinct thylacine (Tasmanian tiger) was recently sequenced. To check the relationship of the thylacine to other marsupials, download all thylacine proteins from Uniprot (currently, they number 24) and run AAI-profiler. In our limited sample, the closest species are carnivorous dasyurid marsupials (orange diamonds in Figure 3), in agreement with the genome paper.

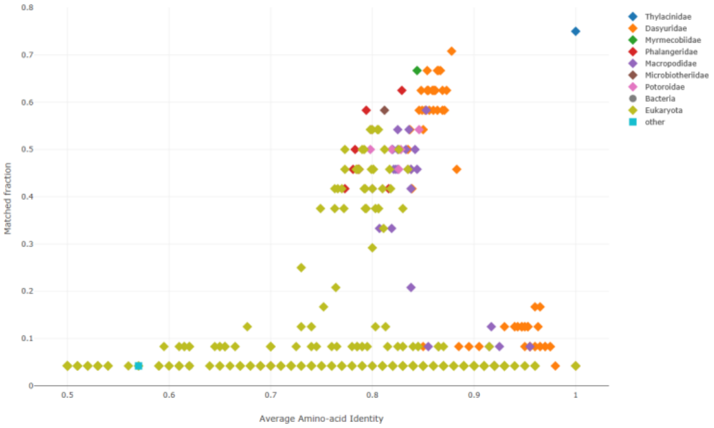


Figure 3. AAI-profiler scatterplot (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/thylacine/) for the thylacine. The query proteome consisted of only 24 proteins, which gives rise to discrete stripes in the plot.

4.2. Species identification

Particular bands of AAI values correspond to taxonomic ranks in the classification of bacteria. For example, MyTaxa (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4005636/) is a taxonomic classifier which was trained on the comparison of 410 fully sequenced genomes and derived boundary values to discriminate between species (80%<AAI<100%), genera (50%<AAI<84%) and phyla (40%<AAI<50%). (MyTaxa is from the Kostas lab, analyses metagenomics samples and only outputs a Krona plot.)

Facey et al. (2015; https://www.ncbi.nlm.nih.gov/pubmed/26185096) took a comprehensive comparative genomics approach to investigate two prominent bacterial symbionts (BFo1 and BFo2) isolated from geographically separated populations of western flower thrips (*Frankliniella occidentalis*), an important pest insect in agriculture. They concluded that BFo1 is a close relative to *Erwinia aphidicola* and that BFo2 represents a highly novel species that maybe related to known *Pantoea*. These conclusions are confirmed by AAI-profiler (and this only takes the push of one button). What is more, we can identify a genus and possible species assignment to BFo2.

BFo1 clearly clusters with *Erwinia* species (Figure 4). *Erwinia aphidicola*, the orange dot at (0.999, 0.002), has few proteins in the Uniprot database but they are almost identical to BFo1. The orange dot at (0.933, 0.288) is *Erwinia dacicola* which is the closest *Erwinia* species with a larger number of proteins in the database. The Krona plot shows overwhelming support for the genus *Erwinia*. Click on the label 'Erwinia', which shows a panel of small piecharts on the right, though the best hits are spread between several *Erwinia* species (ties are broken arbitrarily when best hits are assigned).

Until January 2018, the closest proteomes to BFo2 were found around 75% AAI, which suggests that BFo2 represents a novel genus within *Erwiniaceae* (Figure 5). The taxonomic profile is divided bet ween *Tatumella* and *Pantoea* as the most closely related genera. The closest hit of BFo2 at an AAI value of 0.995 was *Rosenbergiella epipactidis*, although only two proteins are matching. Fortunately, the genome of *Rosenbergiella nectarea* has been sequenced and its predicted proteome is available from NCBI (https://www.ncbi.nlm.nih.gov/genome/?term=Rosenbergiella). Though the proteome is yet to be included in Uniprot, a reverse AAI-profiler analysis with the *R. nectarea* proteome as query shows BFo2 as its closest relative [dot at (0.936, 0.759) in Figure 6]. The AAI value between *R. nectarea* and BFo2 is comparable to the distance between BFo1 and several *Erwinia* species, which supports the assignment of BFo2 within the genus *Rosenbergiella*.

In February 2018, the proteome of tentatively assigned *Tatumella sp.* OPLPL6, a symbiont of *Orius* species, was added to Uniprot and AAI-profiler picks it up as a very close relative of BFo2 at 98% AAI. We propose that OPLPL6 should follow BFo2 and be reclassified as a member of genus *Rosenbergiella*. The proposal is supported by phylogenetic analysis of the atpD, gyrB and rpoB proteins which show a monophyletic clade composed of *Rosenbergiella*, OPLPL6 and BFo2. Interestingly, *Orius* pirate bugs, *Frankliniella* thrips and *Rosenbergiella*, isolated from floral nectar, are linked by a food chain which helps to explain the world-wide spread of the enlarged genus *Rosenbergiella*.

For comparison, a search for sequence neighbours using NCBI's nucleotide BLAST server using the genome of BFo2 as query found hits to *Tatumella* and *Pantoea*. Both the genomes of BFo2 and *Rosenbergiella* are draft quality scaffolds and have not been entered into the nr database searched by the BLAST server. Moreover, the results for each contig have to be browsed separately. The proteome-wide combined report produced by AAI-profiler led to a productive conclusion in fewer steps.

The cyan diamond at (0.968, 0.026) and (0.89, 0.026) in the scatterplots of BFo1 and BFo2, respectively, and the green wedge in the Krona plots, are *Triticum urartu*, the progenitor of cultivated wheat. The taxonomic profile of *Rosenbergiella nectarea* also shows 0.2% of *Triticum urartu*. *Triticum urartu* has many close plant relatives, which do not show up in the scatterplots. This suggests a small amount of bacterial contamination in the *Triticum urartu* genome sample. Indeed, a close look at the taxonomic profile of *Triticum urartu* shows 0.1% Bacteria, of which 16% is BFo1 (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/Triticum_urartu/krona.html).

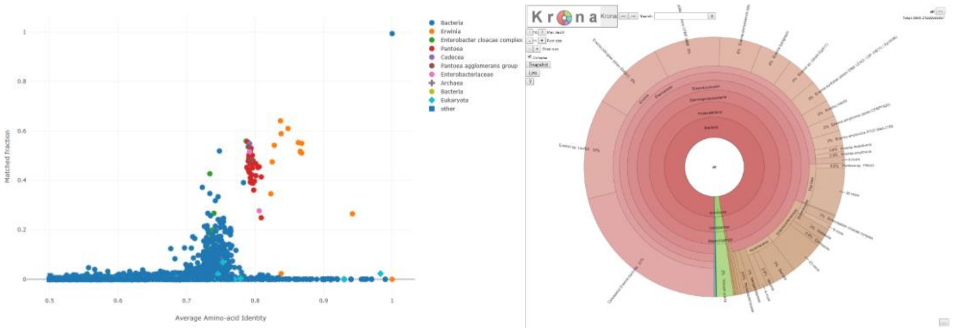


Figure 4: AAI-profiler scatterplot (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/symbiont/AAI.html) and taxonomic profiling (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/symbiont/krona.html) for bacteria symbiont BFo1 of *Frankinella occidentalis*.

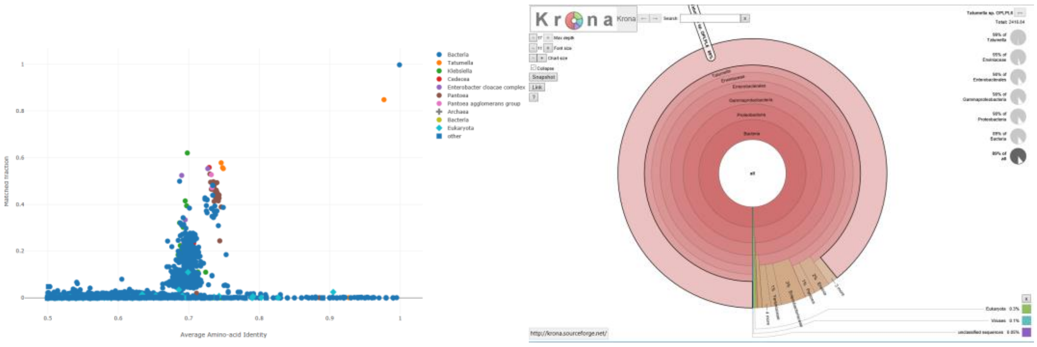


Figure 5: AAI-profiler scatterplot (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/symbiont_BFo2/AAI.html) and taxonomic profiling (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/symbiont_BFo2/krona.html) for bacteria symbiont BFo2 of *Frankiniella occidentalis*.

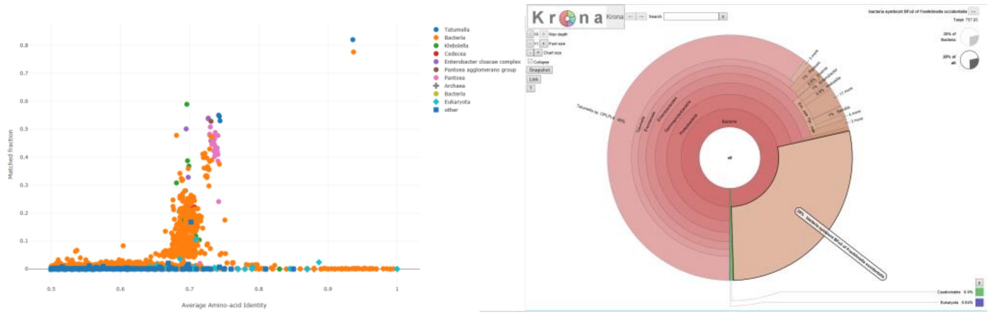


Figure 6: AAI-profiler scatterplot (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/Rosenbergiella/AAI.html) and taxonomic profiling (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/Rosenbergiella/krona_self.html) for *Rosenbergiella nectarea*. The blue dot near the top right corner is bacteria symbiont BFo2 of *Frankiniella occidentalis*.

4.3. Misidentified samples

Proteus sp. HMSC10D02 (http://www.uniprot.org/proteomes/UP000178120) is part of the *Raoultella ornithinolytica / Klebsiella ornithinolytica* pan proteome in the Proteome section of the Uniprot database. However, the name of the organism is given as *Proteus*, which shows up as incongruent colour of the dot at (1,1) in the scatterplot by AAI-profiler (Figure 7). The *Klebsiella* cluster is shown in orange and includes *Raoultella* as well as individual genomes assigned to *Enterobacter* or *Escherichia* (Figure 7). Published phylogenetic analysis (https://www.ncbi.nlm.nih.gov/pubmed/24905728) has suggested abandoning the *Raoultella* genus designation. Similarly, incongruent colours in AAI-profiler scatterplots argue for placing *Klebsiella michiganensis* RC10 and *Enterobacter sp.* Ag1 in the genus *Cedecea* (see *Klebsiella* in http://ekhidna2.biocenter.helsinki.fi/AAI/examples/).

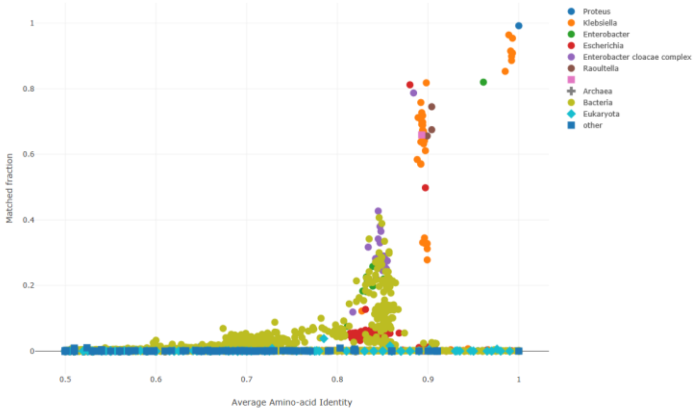


Figure 7: AAI-profiler scatterplot (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/HMSC10D02/AAI.html) for *Proteus sp.* HMSC10D02.

4.4. Multi-isolate samples

Beauveria are fungi that attack insects and therefore have potential in pest control. *Beauveria bassiana* D1-5 has a complex composition (Figure 8 left-middle). The majority of neighbour species are related fungi, but a single arthropod stands out at (0.71, 0.211) and a band of bacteria are found in lower right part of the plot. The green dot at (0.993, 0.151) is *Klebsiella michiganensis*, at high enough AAI to assign species membership (to a part of the genome sample). The taxonomic profile (Krona plot) also shows the contamination. The arthropod is the firefly *Photinus pyralis*. No other arthropoda match the *Beauveria* proteome. The bacterial band is absent from other *Beauveria* strains, e.g. ARSEF 2860, while the firefly match remains (Figure 8, right). While the *Beauveria* genome sample is contaminated by a bacterium, Figure 6 shows that the firefly genome sample contains a component of fungal origin. The fungus found at highest AAI values is *Metarhizium*. The fungus represents 28 % of the genome sample (Figure 9).

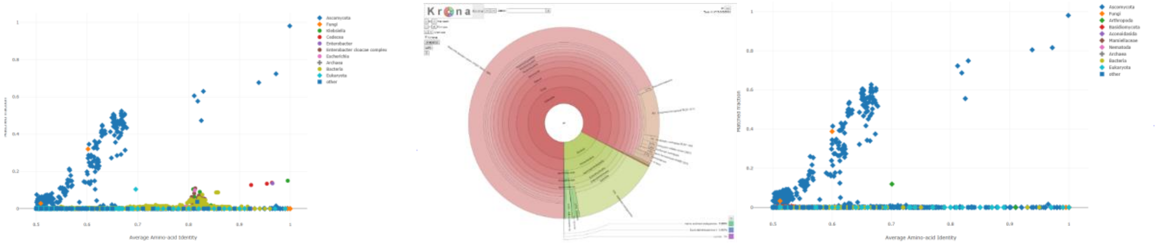


Figure 8. AAI-profiler scatterplot (left, see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/beauveria/AAI.html) and taxonomic profile (middle, see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/beauveria/krona.html) for Beauveria bassiana D1-5. AAI-profiler scatterplot (right, see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/beauveria_bassiana_ARSEF_2860/AAI.html) for *Beauveria bassiana* ARSEF 2860.

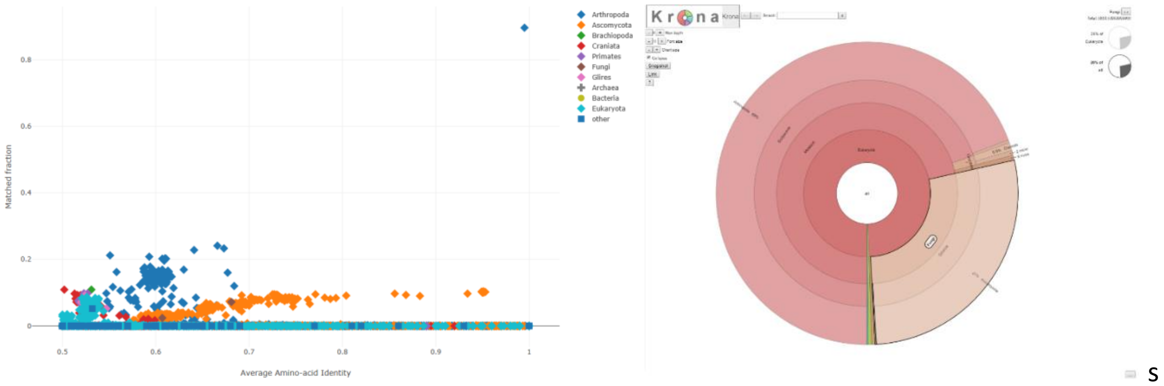


Figure 9. AAI-profiler scatterplot (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/photinus/AAI.html) and taxonomic profile (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/photinus/krona.html) for the firefly *Photinus pyralis*.

4.5. Polyphyletic taxa

The genus *Lactobacillus* currently contains over 180 species and encompasses a wide variety of organisms. The genus is polyphyletic, with the genus *Pediococcus* dividing the *L. casei* group, and the species *L. acidophilus*, *L. salivarius*, and *L. reuteri* being representatives of three distinct subclades (Wikipedia). The scatterplot for a *Lactobacillus brevis* (Figure 10, left) shows green and orange *Pediococci* dots embedded among the blue *Lactobacilli* dots. *Lactobacillus crispatus* (Figure 10, right) belongs to another subclade which shows affinity to *Chlamydia*. The source of *Chlamydia* is discussed in the next section.

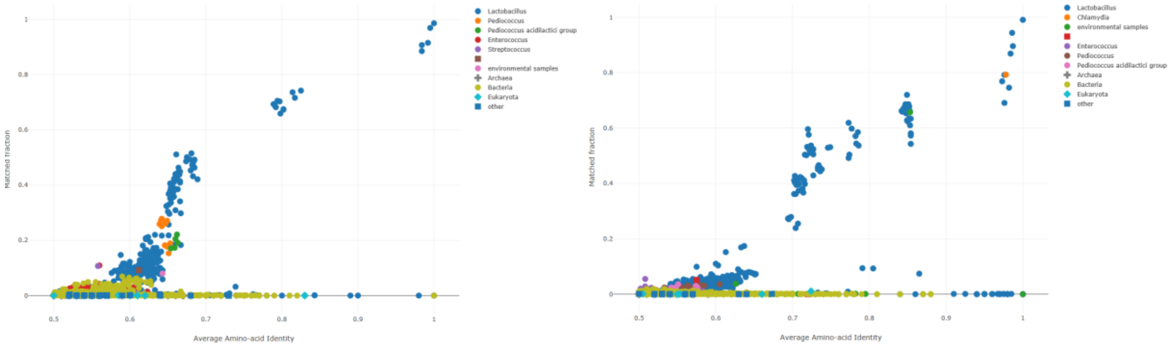


Figure 10. Left: AAI-profiler scatterplot (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/BREVIS/AAI.html) for *Lactobacillus brevis*. Right: AAI-profiler scatterplot (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/lactobacillus_crispatus_125-2-CHN/AAI.html) for *Lactobacillus crispatus* 125-2-CHN.

4.6. Pan proteomes

The taxonomic profile of *Lactobacillus crispatus* 125-2-CHN shows 10 % *Chlamydia trachomatis*. This is the orange dot at (0.971, 0.836) in Figure 10, right. *Chlamydia* are taxonomically classified in a separate order *Chlamydiales*. Many strains of these pathogens have been sequenced, and tracking down the origin of this label required a bit of detective work. The AAI profiles of several *Chlamydia* strains show no trace of *Lactobacilli*. It turns out that Uniprot's organism source (OS=) meta-data can use the species as label for several strains. Thus, the information about which strain was sequenced is hidden from AAI-profiler. The *Chlamydia trachomatis* D/UW-3/Cx pan proteome is composed of 16 strains (http://www.uniprot.org/proteomes/?query=chlamydia+trachomatis+redundant%3Ano&sort=score) and contains 25,858 proteins. (The pan proteome seems to have been removed from Uniprot in February 2018 but multiple strains are still collectively labelled *Chlamydia trachomatis*, CHLTH.) The protein counts of the component strains range from 884 to 7320. Figure 10 show that the pan proteome has an extremely diverse composition, with only 17 % of the proteins having a nearest neighbour in other *Chlamydiales* (self hits to *Chlamydia trachomatis* were excluded). For example, one of the component strains, SwabB4 (http://www.uniprot.org/proteomes/UP000044845) has 3922 predicted proteins and is a mixture of 72% *Lactobacillus* and 25% *Chlamydia* (Figure 12). Contamination of the genomic samples of other strains that were incorporated into the pan proteome are likely to account for the remaining irregularities of the AAI-profiler scatterplot (Figure 11). Clearly, a check of genome quality with AAI-profiler would be beneficial to ensure high quality derived data in databases.

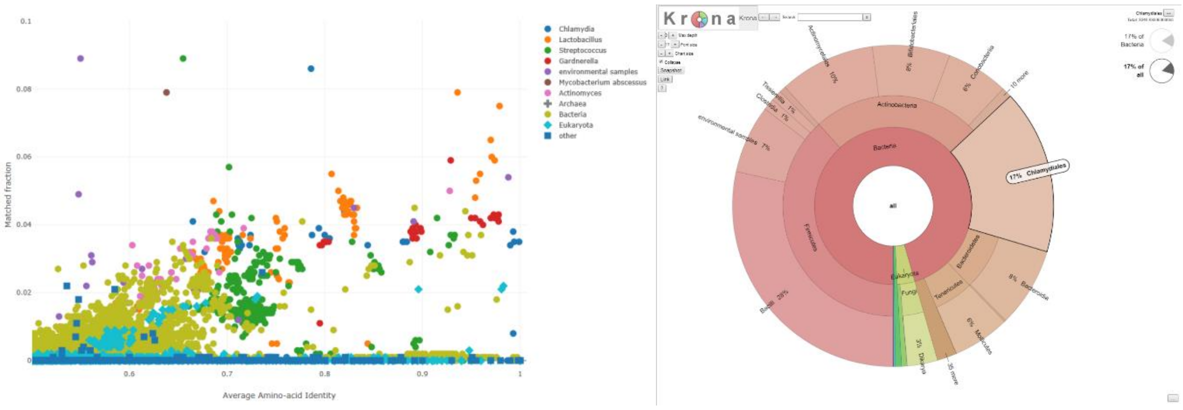


Figure 11. AAI-profiler scatterplot (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/Chlamydia_pan/AAI.html) and taxonomic profile (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/Chlamydia_pan/krona.html) of the *Chlamydia* pan proteome.

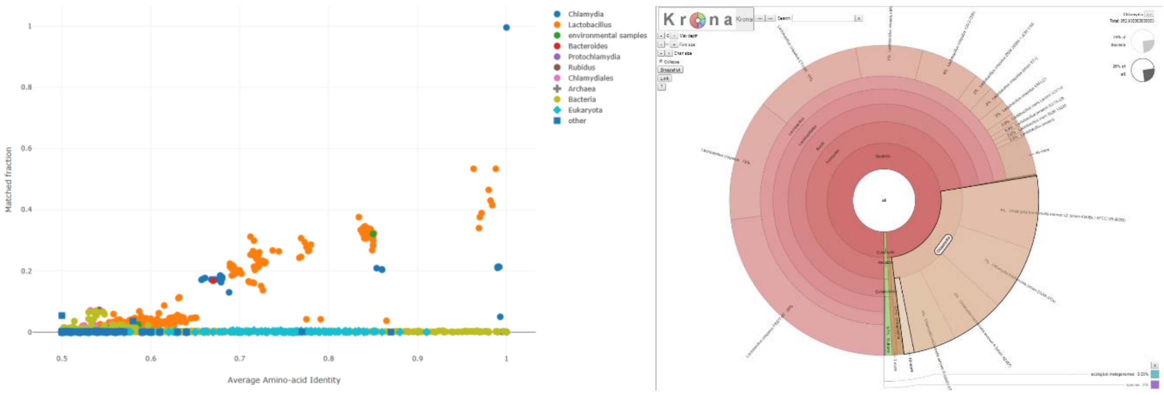


Figure 12. AAI-profiler scatterplot (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/Chlamydia_trachomatis_SwabB4/AAI.html) and taxonomic profile (see http://ekhidna2.biocenter.helsinki.fi/AAI/examples/Chlamydia_trachomatis_SwabB4/krona.html) of *Chlamydia trachomatis* strain SwabB4.

5. Related tools

To our knowledge, the scatterplot visualization of species neighbours is unique to AAI-profiler. A strength of AAI-profiler is that it reveals inconsistencies in database meta-data, unlike taxonomic profiling, and that it finds any neighbours of one query proteome, unlike gene trees which require a distance matrix. Taxonomic profiling assumes that taxonomic annotations in the database are correct when assigning reads (translated proteins) to the nearest species or to the last common ancestor of the nearest hits.

Many inconsistencies revealed by AAI-profiler have been noted before by sequence database curators, but they are buried in notes and remarks in an unsystematic way. AAI-profiler performs the search on the whole proteome and presents a summary report. Many NCBI species with whole-genome sequences of multiple strains have links to a Genome Tree report. For some, but not all, species with multiple strains, NCBI genomes shows a dendrogram based on genomic BLAST and presents clade identifiers and a precomputed Genome neighbour report.

The MiGA web site (http://microbial-genomes.org/) implements a suite of metagenome and genome comparative analyses including AAI distance searches similar to AAI-profiler. Using SANSparallel instead of BLAST, AAI-profiler has a faster response time but the most important difference is that the Uniprot database searched by AAI-profiler has a larger representation of species (809,540 unique labels in March 2018) than the prokaryotic reference genome collections searched by MiGA (1,927 references in NCBI

RefSeq and 11,566 references in NCBI Prok). For example, MiGA returned *Tatumella citrea* at 70% AAI as the closest match to the BFo2 symbiont whereas AAI-profiler indicated OPLPL6 and *Rosenbergiella* at 98% AAI.

There are many servers to derive phylogenomic trees of a selected subset of species (review). For example, AAI (Average Amino Acid Identity) and ANI (Average Nucleotide Identity) matrices can be generated using EDGAR 2.0 or the servers of the Kostas lab. PhyloSearch is a specialized tool that "can help you identify bacterial strains by building a phylogeny that takes into account your data and the data of the French Collection of Plant associated Bacteria (CFBP)". AAI-profiler generates one-sided and bidirectional AAI profiles to show the species neighbourhood of one query proteome, whereas you have to specify a set of species as input to matrix methods.

The Krona plots generated by AAI-profiler are analogous to taxonomic profiling of metagenomics samples. Tools for the taxonomic profiling of metagenomics samples take the raw sequence reads (or translated proteins) as input and classify them taxonomically. The goal of AAI-profiler is to show a larger species neighbourhood (and taxonomic namespace) of an isolate. If your only goal is taxonomic profiling, the large search radius of AAI-profiler is overkill and we recommend dedicated metagenomics analysis tools like Kraken, MEGAN6, or MG-RAST.

OMIC tools has an extensive catalogue of computational tools including phylogenetic analysis and metagenomics sequence analysis.