Tracing functional protein interaction networks using a feature-aware phyletic profiling

SENCKENBERG world of biodiversity

Holger Bergmann,[◈] Ngoc Vinh Tran,[◈] Bastian Greshake,[◈] Julian Dosch,[◈] Bardya Djahanschiri,[◈] Sachli Zafari,[◈] and Ingo Ebersberger[◈*]

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

[◈] Applied Bioinformatics Group, Institute of Cell Biology and Neuroscience, Goethe University, Frankfurt am Main, Germany

[*] Senckenberg Biodiversity and Climate Research Centre (BiK-F), Frankfurt am Main, Germany

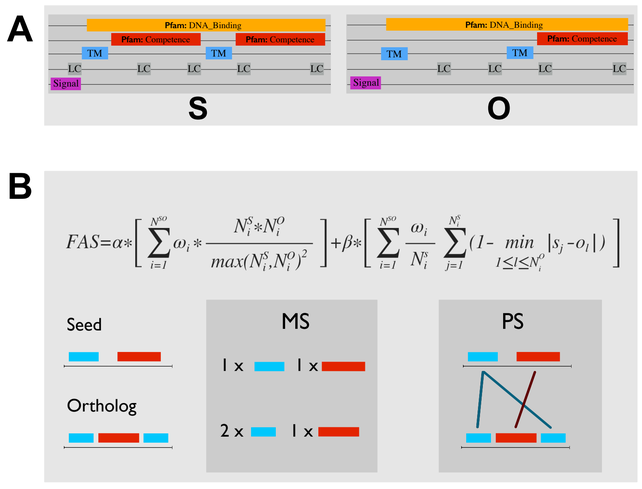
Motivation and Background

Tracing the phyletic distribution of protein interaction networks across large taxon collections calls for reliable and scalable methods for functional annotation transfer. Standard ortholog inferences resulting in so called 'phyletic profiles' do not suffice in many cases, as the functional similarity between orthologs decays with time. Here, we integrate the search for orthologs with an automated scoring of the pair-wise feature architecture similarity (FAS) between a protein of interest, the 'seed' and its orthologs. The resulting score serves as a proxy for the functional equivalence to the two proteins. As a use-case, we apply our 'feature-aware' phyletic profiling framework to investigate the evolution of 204 extracellular proteins from the human nosocomial pathogen *Acinetobacter baumannii* across 1,985 bacterial genomes. On this basis, we identify proteins that changed their feature architecture specifically on the lineage separating *A. baumannii* from its non-pathogenic relatives. These proteins serve as promising candidates for hitherto undetected virulence factors.

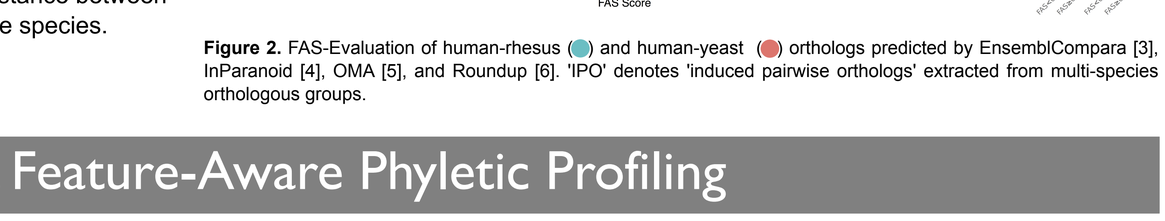
1. Scoring Feature Architecture Similarity (FAS)

- **Features:** We build the feature architecture (Fig. 1A) for a seed protein S and its ortholog O from the following default feature classes:
  - (i) Pfam and SMART domains
  - (ii) Secondary structure elements
  - (iii) Transmembrane domains
  - (iv) Signal sequences
  - (v) Low complexity regions

- **Scoring:** The FAS score (Fig. 1B) comprises **(i)** the *multiplicity score* (MS) capturing the identity and copy number of shared features, and **(ii)** the *positional score* (PS) capturing the similarity in relative position of two shared features. The score is defined on the interval [0,1].



- **Data structure:** The feature architecture is implemented as a directed acyclic graph. When redundant features of the same class overlap in the architectures of ***S*** and/or ***O***, we identify the pair of linearized paths maximizing the FAS score (Fig. 1C).



2. FAS score evaluation of human orthologs

- We determined the FAS score distribution across human orthologs in rhesus and yeast assigned by four ortholog search tools.
- The fraction of orthologs with FAS scores below 0.75 varies with the prediction tool and with evolutionary distance between the species.

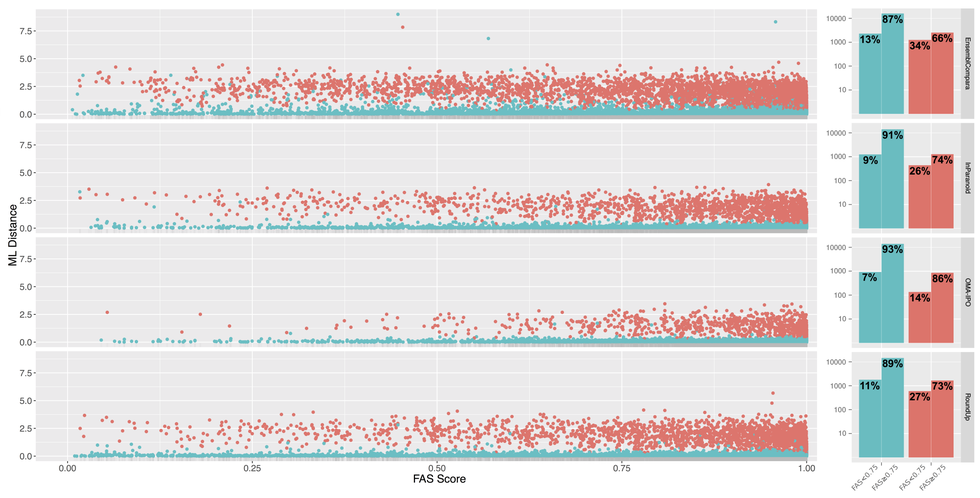


Figure 2. FAS-Evaluation of human-rhesus (●) and human-yeast (●) orthologs predicted by EnsemblCompara [3], InParanoid [4], OMA [5], and Roundup [6]. 'IPO' denotes 'induced pairwise orthologs' extracted from multi-species orthologous groups.

3. Feature-Aware Phyletic Profiling

- For a seed protein, we establish the presence/absence pattern of orthologs across a collection of target species using a targeted ortholog search [1]. Orthologs are weighted by their feature architecture similarity (FAS) to the seed protein. Orthologs in the phyletic profile with a FAS score above an individually chosen cut-off are considered functionally equivalent.

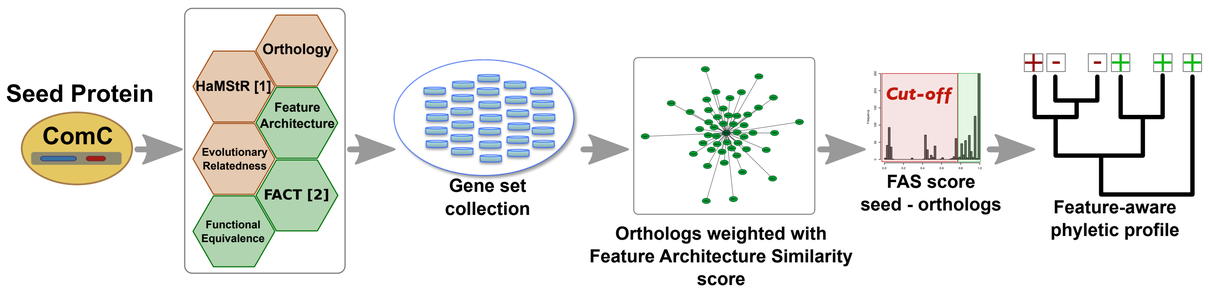


Figure 3. Feature-aware phyletic profiling. '+' and '-' represent present and absent orthologs of the seed protein, respectively. A red '+' indicates an ortholog where the FAS implies a change in functionality.

4. Functional Annotation Transfer

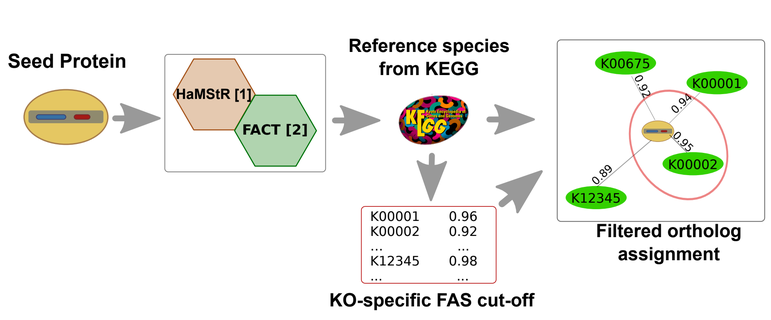


Figure 4. Feature-aware pipeline for annotating proteins with a KEGG orthology (KO) id. For a seed protein we identify orthologs in a collection of annotated KEGG reference species. If the FAS score between seed and KO annotated ortholog exceeds the KO-specific FAS cut-off, the KO id will be transferred to seed protein. FAS cut-off values are determined from the pairwise FAS score between reference proteins annotated with the same KO.

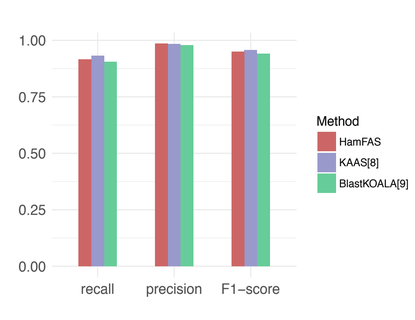


Figure 5. Recall and precision values for the re-assignment of KO ids for 3,457 pre-annotated yeast proteins using three alternative annotation tools. Prior to analysis, we excluded yeast from the KEGG reference species.

5. Search for bacterial pathogenicity factors

- *Acinetobacter baumannii* ranks among the top six most threatening human pathogens worldwide.
- Due to spreading resistances to most if not all antibiotics, protein targets for new therapeutic approaches are required. Species-specific virulence factors are promising targets.
- We screened 204 *A. baumannii* proteins forming the host-interaction interface for candidates with an altered feature architecture in *A. baumannii* indicative of a lineage-specific and virulence-related shift in function.

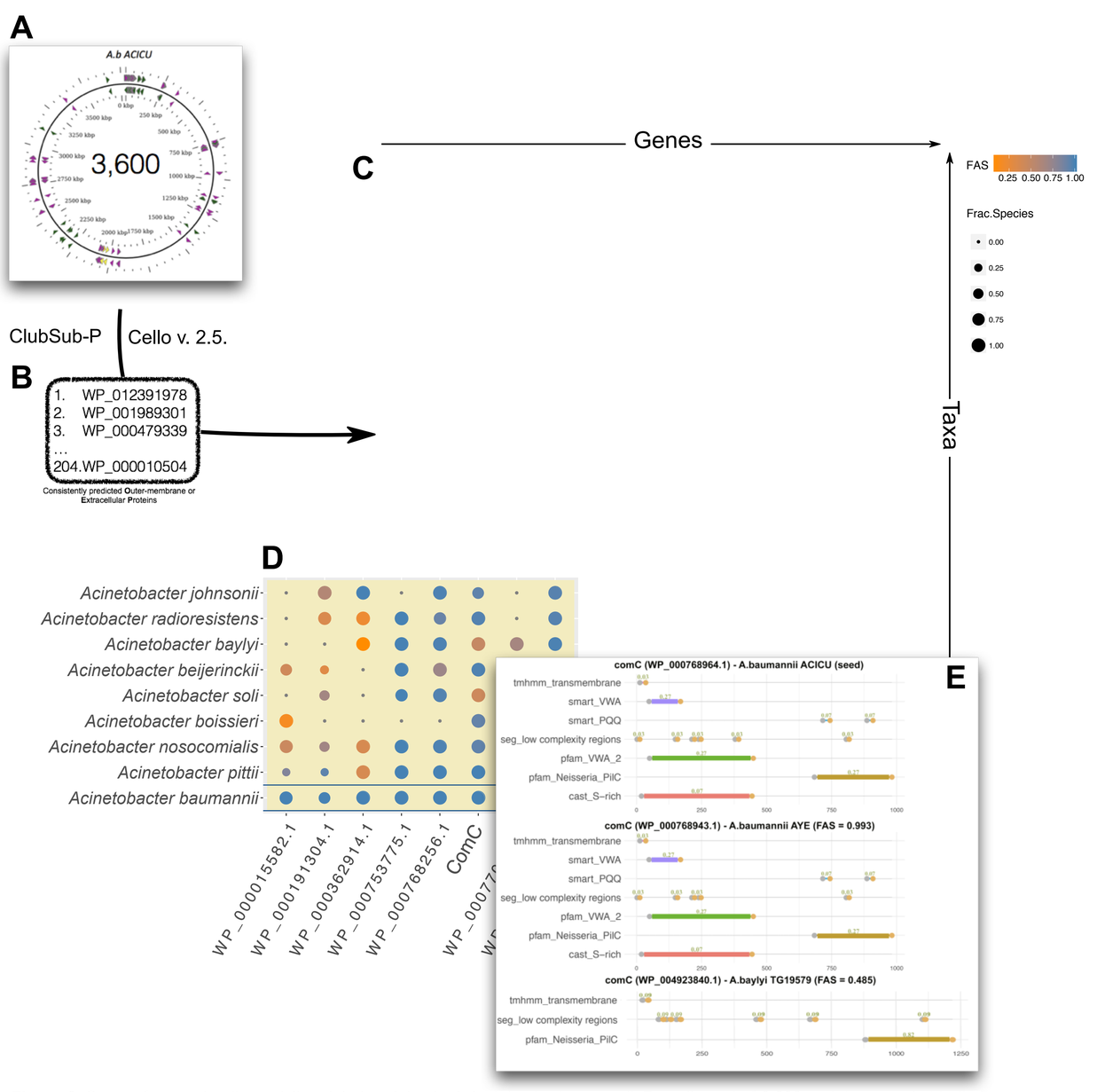


Figure 6. Screen for novel pathogenicity related proteins in *Acinetobacter baumannii*. From the set of 3,600 proteins encoded in the genome of *A.b.ACICU* **(A)**, ClubSubP [10] and Cello [11] agreed in predicting 204 proteins as outer-membrane or extracellular located **(B)**. FAS aware ortholog search obtained a phyletic profile that can be visualized with the *PhyloProfile* app [12] **(C)**. Within PhyloProfile, the information can be adjusted to individual genes and taxa of interest **(D)**, and the feature architecture of candidate proteins can be visualized **(E)**. Here, we identified ComC, a factor involved both in cell adhesion and in direct DNA uptake as a putative candidate driving A. baumannii virulence.

6. Summary

- **The FAS score serves to identify orthologs where changes in the feature architecture implies a change in functionality.**
- **FAS supported phyletic profiles can be routinely applied for functional annotation transfer, and for the identification of lineage-specific shifts in functionality, e.g. in the screen for novel pathogenicity factors. The software is available from https://github.com/BIONF**
- **The PhyloProfile app facilitates an intuitive and dynamic exploration of FAS supported phyletic profiles.**
- **Profiling 204 proteins of the human pathogen A. baumannii across 1,985 bacterial genomes identifies, among other candidates, ComC as a promising virulence associated factor.**

Contact

Ingo Ebersberger
ebersberger@bio.uni-frankfurt.de
Goethe University, Frankfurt am Main, Germany
Max-von-Laue-Straße 13, 60438 Frankfurt am Main

References

[1] Ebersberger, Strauss, Haeseler. BMC Evolutionary Biology (2009), 9:157.
[2] Koestler, Haeseler, Ebersberger. BMC Bioinformatics (2010), 11:417.
[3] Vilella, Severin, Ureta-Vidal, Durbin, Heng, Birney. Genome Research (2009), 19:327-335.
[4] Ostlund, Schmitt, Forslund, Kostler, Messina, Roopra, Frings, Sonnhammer. Nucleic Acids Res (2010). 38:D196-D203.
[5] Roth, Gonnet, Dessimoz. BMC Bioinformatics (2008), 9:518.
[6] Deluca, Cui, Jung, St Gabriel, Wall. Bioinformatics (2012), 28: 715-6.
[7] Ogata, Goto, Sata. Nucleic Acids Res (1999), 27:29-34.
[8] Moriya, Itoh, Okuda. Nucleic Acids Res (2007), 35.
[9] Kanehisa, Sato, Morishima. Journal of Molecular Biology (2016), 48:726-731
[10] Paramasivam, Linke. Front Microbiol (2011), 2:218
[11] Yu, Chen, Lu, Hwang. Proteins: Structure, Function and Bioinformatics (2006), 64:643-651
[12] https://github.com/BIONF