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

SENCKENBERG world of biodiversity

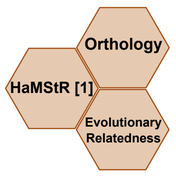
Holger Bergmann[◇], Julian Dosch[◇], and Ingo Ebersberger[◇*]

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

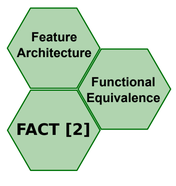
[◇] Department for Applied Bioinformatics, Institute for Cell Biology and Neuroscience, Goethe University, Frankfurt am Main, Germany

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

Motivation and Background



Mining environmental genetic diversity via the direct uptake of free DNA allows naturally competent bacteria a rapid adaptation to changing environments. While natural competence is a highly versatile mechanism for accomplishing genetic innovation, its prevalence in contemporary bacteria is unknown. Phylogenetic profiles for individual building blocks of known DNA uptake machineries - i.e. the presence-absence patterns of orthologs to the corresponding genes across the bacterial domain - provide the means for rapidly identifying novel naturally competent bacteria. Yet, orthology of two sequences is a poor proxy for their functional equivalence, posing the risk of unspecific predictions. Here we present the integration of orthology inference with an automated assessment of feature architecture similarity to facilitate a phylogenetic profiling that is aware of protein features and their associated function.



1. Approaches of Phylogenetic Profiling

A **Conventional:** A targeted ortholog search generates for a seed protein of interest a presence-absence pattern across the set of analysed species. Orthologs remain unweighted and their functional similarity is not further addressed.

B **'Feature-aware':** Orthologs identified as in (A) are weighted by their feature architecture similarity (FAS) to the seed protein. In the phylogenetic profile orthologs with a FAS score above an individually chosen cut-off are considered functionally equivalent.

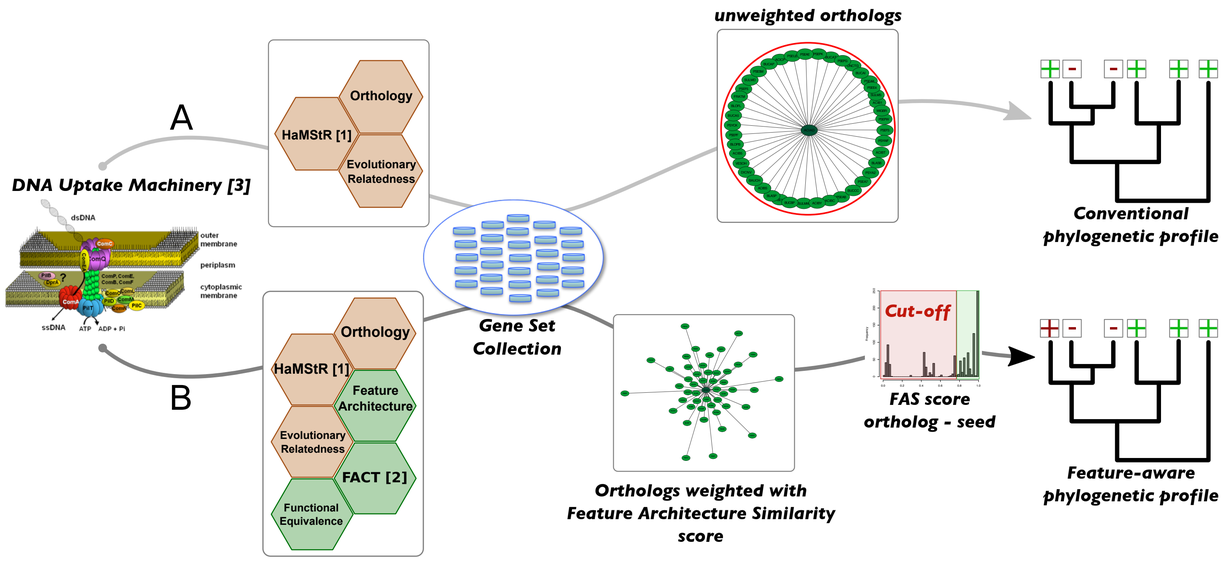


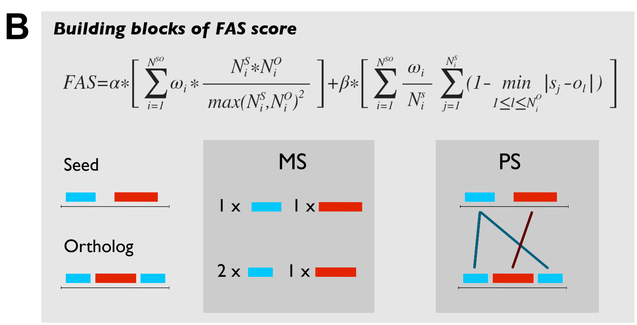
Figure 1. Conventional (**A**) and feature-aware (**B**) phylogenetic profiling. '+' and '-' represent present and absent orthologs to the seed protein, respectively. A red '+' indicates an ortholog which FAS does not suffice for functional equivalence inference.

2. Scoring Feature Architecture Similarity (FAS)

A **Features:** We score the FAS [2] between a seed protein ***S*** and its ortholog ***O***.
We consider the following feaures by default:

(i) Pfam and SMART domains
(ii) Secondary structure elements
(iii) Transmembrane domains
(iv) Low complexity regions

B **Scoring:** The FAS score captures copy number similarity and type of shared features (*MS*), as well as their similarity in relative position (*PS*). The score is defined on the interval [0,1].



C **Data structure:** The feature architecture is implemented as a directed acyclic graph. When redundant features overlap in the architectures of ***S*** and/or ***O*** we identify the linearized paths maximizing the FAS.

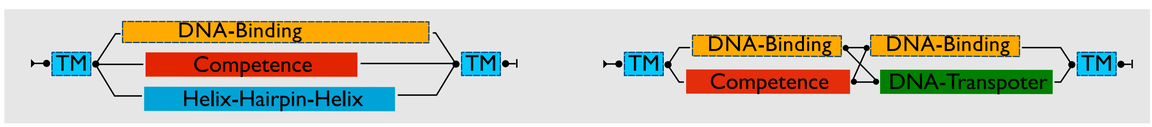


Figure 2. (A) A feature architecture represents annotated features based on the amino acid sequence of the protein. **(B)** The building blocks of *FAS* contribute with different weightings (α = 0.7, β = 0.3) to the score. **(C)** Features are considered as vertices whose order is defined by the edges of the graph.

3. FAS score evaluation of orthologous pairs

- Ortholog pairs assigned by four common orthology predictors display FAS scores ranging from 0 to 1.
- The fraction of FAS scores below 0.7 varies with the prediction tool and with evolutionary distance between the species. It is lowest for OMA groups [6] (5.1%) and highest for EnsemblCompara [4] (20.1%).

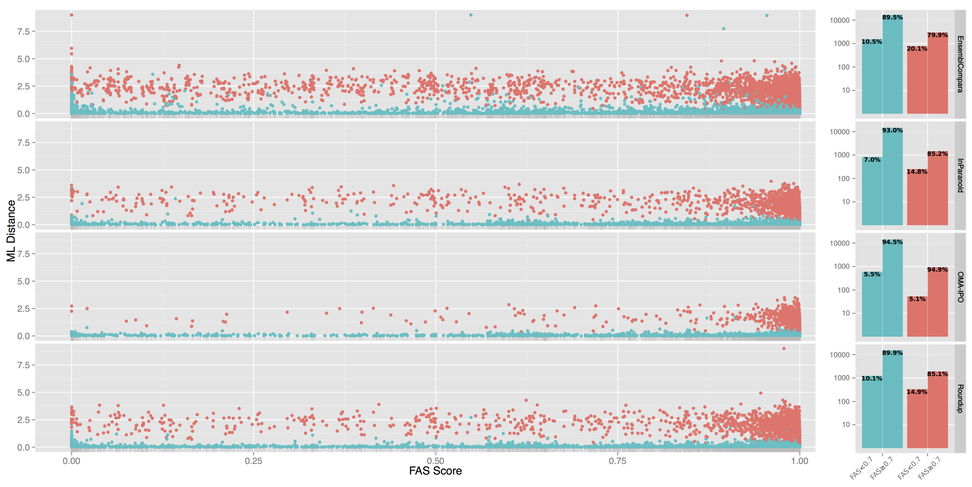
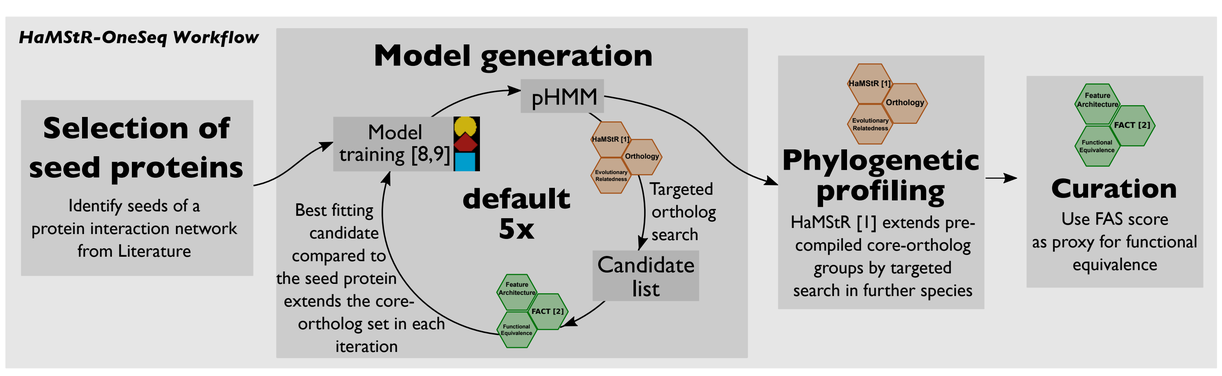


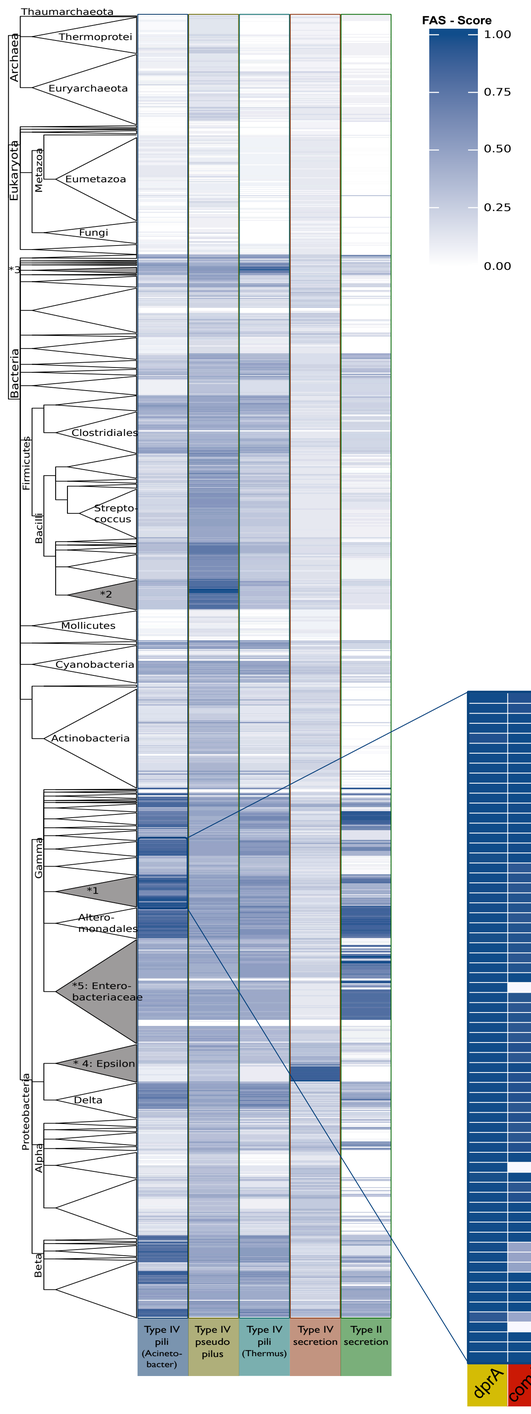
Figure 3. FAS-Evaluation of human-rhesus (●) and human-yeast (●) orthologs predicted by EnsemblCompara [4], InParanoid [5], OMA [6], and Roundup [7].

4. HaMStR-OneSeq: FAS supported targeted ortholog search

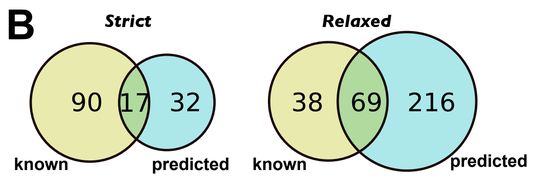


5. Example Application

A The phyletic distribution of 5 DNA uptake machineries



- 5 DNA uptake machineries comprising 81 proteins were traced in 1,316 taxa (1,074 bacteria, 107 archaea and 135 eukaryotes).
- **A** Orthologs were curated by their FAS score. The average score of the respective competence machinery is shown in a single tile. The enlarged subprofile displays per-gene results.
- **B** Based on the profile we classify bacteria as potentially naturally competent. We apply two criteria:
  Strict: All but one gene of the respective machinery must be present. Each ortholog must have a FAS score ≥ 0.7.
  Relaxed: Up to four missing genes are allowed. Genes with a FAS score ≥ 0.7 are regarded as present.



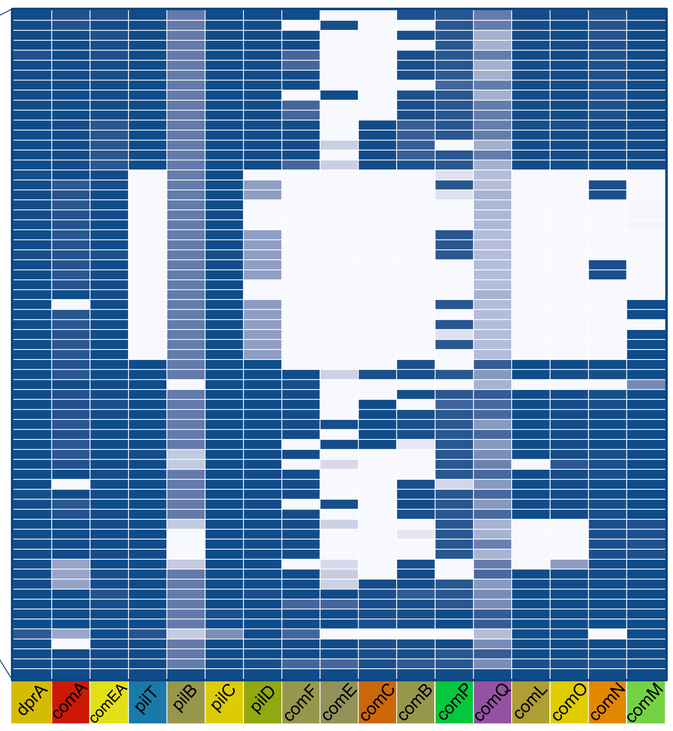


Figure 4. (A) Heat map of a feature-aware phylogenetic profile. Legend: 1: Pseudomonadales (Type IV pili related [3]), 2: Bacillaceae (Type IV pseudopilus [10]), 3: Deinococci (Type IV pili related [11]), 4: Epsilonproteobacteria (Type IV secretion [12]), 5: Enterobacteriaceae (Type II secretion [13,14]). Gene names in the subprofile follow the color code in Figure 1. **(B)** Known nat. competent strains vs. predicted strains (strict/relaxed).

6. Summary

- **We present a scalable approach to establish feature-aware phylogenetic profiles.**
- **Our method integrates a targeted ortholog search with feature architecture similarity scoring.**
- **The phyletic distribution of 5 DNA uptake machineries predicts hitherto uncharacterized bacteria as naturally competent.**

Contact

Holger Bergmann
bergmann@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] Averhoff , Graf. Acinetobacter Mol. Biol., Caister Academic (2008), p. 119-140.
[4] Vilella, Severin, Ureta-Vidal, Durbin, Heng, Birney. Genome Research (2009), 19:327-335.
[5] Ostlund, Schmitt, Forslund, Kostler, Messina, Roopra, Frings, Sonnhammer. Nucleic Acids Res (2010). 38:D196-D203.
[6] Roth, Gonnet, Dessimoz. BMC Bioinformatics (2008), 9:518.
[7] Deluca, Cui, Jung, St Gabriel, Wall. Bioinformatics (2012), 28: 715-6.
[8] Katoh, Toh. Briefings in Bioinformatics (2008), 9: 286-298.
[9] Eddy. PLoS Computational Biology (2011), 7(10): e1002195.
[10] Hamoen, Venema, Kuipers. Microbiology (2003), 149(Pt 1): 9–17.
[11] Averhoff. FEMS Microbiol. Rev. (2009), 33 : 6611-6626.
[12] Stingl, Mueller, Scheidgen-Kleyboldt, Clausen, Maier. PNAS USA (2010), 107:1184–1189.
[13] Chen, Dubnau. Nature reviews. Microbiology (2004), 2:241–249.
[14] Sandkvist. Molecular microbiology, 40:271–283.