Fakultät für Informatik
Technische Universität München

Predicting functional capabilities of microbes using a similarity graph approach

Yannick Mahlich

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:
Prof. Dr. Daniel Cremers

Prüfende der Dissertation:
1. Prof. Dr. Yana Bromberg
2. Prof. Dr. Florian Erhard, JMU Würzburg

Die Dissertation wurde am 16.05.2019 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 01.07.2019 angenommen.

Abstract

Microbes play a major role in our everyday lives. While some pose a deadly threat, many of the "good" bacteria are crucial for our bodies' optimal functionality and disease-free survival. The microbial communities persistent on or in individual body sites are known as microbiomes. For example. the skin microbiome fends off invasive microbial species, while the gut-brain axis links the gut microbiome to neurodevelopment and neurodegeneration.

Advances in high throughput sequencing have deposited a vast amount of microbial genetic data into public databases. This data describes the proteins that microbes are able to synthesize and molecular functions they facilitate. Detailed understanding microbial functionality, in turn, provides for a better understanding of organism interactions with each other and their surroundings. Moreover, as microbial, and particularly bacterial, phylogeny-driven taxonomy is a dynamic and hotly debated field, I postulate in this work that standardized functional assessment of organisms can provide for a more stable and meaningful way for organism classification.

Earlier smaller scale efforts to assess similarity between microbes highlighted the importance of functional evaluations. Here, I set out to evaluate the functional similarities between bacteria using new protein sequence alignment algorithms and network clustering techniques. The presented approaches are flexible and robust in the face of the exponential growth of bacterial genetic data. First, I optimized the process of establishing functional similarities between proteins, increasing the speed of existing methods by over 40-fold without sacrificing precision. This optimization has produced a set of over 250 billion alignments of 31.5 million available microbial proteins, representable as a similarity network. I evaluated many available techniques for clustering a network of this size, which is expected to consistently grow over the coming years. I note that none of the evaluated approaches worked within the limitations of available hardware and I detail a possible new, t-SNE algorithm-based, solution. Going forward, both methods will be incorporated into my existing *fusion* (functional similarity of organisms network) platform. This platform, which is available to the general scientific public, provides a means of functionally annotating new bacterial genomes in the context of others, thereby facilitating further detailed research into emergence of microbial functionality, environmental adaptation of microbes and functional interactions between microbes to name a few.

List of Publications

The cumulative dissertation at hand is based on three peer-reviewed and published publications as well as preliminary results part of a publication in preparation. **Chapters 2**, **4** and **5** describe the methodology and results discussed in the reviewed and published publications (attached to the respective chapters):

- **Yannick Mahlich**, Martin Steinegger, Burkhard Rost, Yana Bromberg, HFSP: High speed homology- driven function annotation of proteins. *Bioinformatics*, vol. 34, no. 13, pp. p. i304 – i312, 2018
- Chengsheng Zhu, **Yannick Mahlich**, Maximilian Miller, and Yana Bromberg. fusionDB: assessing microbial diversity and environmental preferences via functional similarity networks. *Nucleic Acids Res* 46(D1): D535-D541., 2018
- **Yannick Mahlich**, Jonas Reeb, Maximilian Hecht, Maria Schelling, Tjaart Andries Petrus de Beer, Yana Bromberg, and Burkhard Rost. Common sequence variants affect molecular function more than rare variants? *Scientific Reports*, 2017.7:1608, 2017.

Chapter 3 discusses preliminary results that are content of a publication currently in preparation.

In addition to the three publications above, an additional peer-reviewed publication co-authored during my doctoral studies is not discussed in this dissertation:

- Jonas Reeb, Maximilian Hecht, **Yannick Mahlich**, Yana Bromberg, and Burkhard Rost. Predicted Molecular Effects of Sequence Variants Link to System Level of Disease. *PLoS Comput Biol*, 12(8): p. e1005047, 2016.

Acknowledgements

First and most importantly I want to thank my *Doktormutter* Yana Bromberg from the bottom of my heart. Without her unwavering support I would not have been able to achieve what often felt unachievable. I am deeply grateful for her giving me the opportunity to spend the last four years of my live in her lab and the United States. Not only did she enable me to grow professionally but also personally. I can truly say that those last four years had a huge impact on me and my live, and I am very thankful for that.

In that line I also want to thank my whole "academic family". All members current and past of the Bromberg Lab, especially Chengsheng Zhu, who made the transition from living in Germany to living in New Jersey an absolute breeze. Thanks to Maximillian Miller for bringing a bit of Germany to the lab. And thanks as well to Yanran Wang, Zishou Zeng, Kenneth McGuinnes and Adrienne Hoarfrost for making the lab such a great place to work at. All of them are truly excellent people, that one can only wish for having as co-workers. I will certainly miss them once the time has come to move on.

I also want to thank the German side of my academic family, the Rostlab. Thank you to Jonas Reeb, Max Hecht, Tatyana Goldberg, Tim Karl, Inga Weise, and the rest of the whole Rostlab who made me always feel welcome and at home, during the brief period that I spent in Munich. More importantly though I want to extend my deepest gratitude to Burkhard Rost. He single handedly rekindled the flame burning for science inside of me. If Burkhard hadn't welcomed me into his lab during my time as undergrad, I probably would not have pursued my Master let alone the PhD.

Additionally I would like to thank all with whom I had the fortune to collaborate with during my time as PhD. Many thanks as well to all who gave me the opportunity to present my work during international conferences and symposia, especially Dr. Iddo Friedberg and Dr. Tamar Barkay.

I also want to thank the secretaries at the Rutgers Department of Biochemistry and Microbiology, especially Jessie Maguire and Audrey Andrews for helping me through the slew of paperwork ensuring my stay at Rutgers University in New Jersey.

I want to thank Prof. Daniel Kremers and Prof. Florian Erhard for agreeing last minute to be on my thesis committee.

Thanks to my friends back in Munich who welcomed me with open arms whenever I stopped by and but also to everyone else from all across the world whom I met along the way. You made the last four years truly special.

Finally I want to extend my deepest gratitude my family, for their supported during this whole endeavor. To my sister Carine who was there for me whenever I needed her but also virtually kicked my butt all the way from Germany whenever it was necessary. And of course to my parents Helene and Axel, for helping me to turn into the person

that I am. Without their sacrifices I would not stand where I stand today. I can not thank you enough for always believing in me. Last but not least my deepest gratitude and love goes out to my wonderful wife. I wouldn't know what I would do without you Jenny. I can not thank you enough for the countless times I was doubting my self and you were there helping me regain trust in my abilities.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Microbial similarities and functional capabilities

The study of bacteria, their relations to each other, and especially their functional properties (e.g. pathogenicity) has been of great scientific interest for centuries. From the early discoveries of microorganisms in the late 17^{th} century by Antoni van Leeuwenhoek[1, 2], to the complete sequencing of *Haemophilus influenzae* in 1995[3] many important discoveries were made. Other central discoveries specifically towards understanding the functional impact of microbes in our lives were made by Louis Pasteur and Robert Koch in the late 19^{th} century. Pasteur established bacteria to be one of the causative effects of food spoilage[4, 5], more specifically beer and wine. His discoveries ultimately lead to methods improving food safety, a process widely know as pasteurization. Robert Koch on the other hand was one of the first to link bacteria to the development of anthrax, tuberculosis and cholera[6, 7, 8]. Those early works predating modern taxonomy clearly demonstrate the need to understand functionality facilitated by microbes. The first attempts at bacterial classification were done by Ferdinand J. Cohn in 1875[9]. Those ultimately culminated in Bergey's Manual of Determinative Bacteriology by David Hendrick Bergey[10]. Bergey's Manual of Determinative Bacteriology (later renamed to Bergey's Manual of Systematic Bacteriology[11, 12]) first described prokaryotic relations based on their morphological and phenotypical properties. In later volumes taxonomic classification was adapted to be based solely on genetic phylogeny, i.e. evolutionary relationships between individual organisms determined by (dis-)similarity of the 16s rRNA. This process of identifying taxonomic relations remains the gold standard to this date. This concept was pioneered in the late 1970s and improved upon through the 1980s by Carl R. Woese and George E. Fox [13, 14, 15]. However, with the rise of modern genetics and the ability to sequence and assemble complete bacterial genomes, *Haemophilus influenzae* being the first fully sequenced bacterium in 1995[3], the case has been made that taxonomic assessment using the whole genome will result in better classification. An improved taxonomic classification generated on the full genomic information can be established. However, if following the same principles as current taxonomic classification, organisms will still be grouped based on their evolutionary relations. A classification scheme like this will not reflect similarities in the functional capabilities microbial organisms are able to facilitate. Here I present an approach to generate functional profiles for microbes that is to pick up on the intricate functional differences of microbes, and relate them to each other accordingly.

One of the base requirements to establish functional profiles is the availability of completely sequenced microbial genomes. Since the publication of *H. influenzae*'s whole genome the number of completely sequenced bacteria available in public databases has

been increasing steadily. In early 2019 Genbank[16] reported a total of 12,000 completely sequenced and assembled bacterial genomes. One critical observation is that the rate at which new species were sequenced to completion, i.e. their complete genome was fully assembled, stayed relatively stable until 2014 (see Figure 1.1). Interestingly enough this is also the first time that twice as many new bacterial strains ("subspecies") than novel species were completely assembled. This trend continued, and grew even stronger in recent years. By 2018 roughly 2,500 novel strains as opposed to 500 novel species were fully sequenced and assembled. So far this amounts to 3,700 distinct species of the 12,000 sequenced organisms. One reason for this observation is that today's assemblies are often driven by mapping to preexisting references, rather than doing de novo assembly. This of course comes specifically in handy for novel strains, where the assumption is that strains only differ little from the reference species. Another factor for this observation are sequencing projects that investigate differences in strains found in localized areas e.g disease outbreak areas or across different patients exhibiting the same infections. In those cases, species extracted from different patients are labeled as distinct strains. It is not unlikely that one such sequencing project adds more than 50 different strains of one species to NCBI genome. While the majority of strains of one species will share large parts of their genome, unique functional traits that set them apart can still be detected. This is a particular problem in the state-of-the-art identification of bacterial presence in environmental samples.

1.2 Current concept in microbial classification

One of the pillars of prokaryotic classification, respectively detection has been 16s rRNA sequencing. 16s rRNA is a small, highly conserved sub-unit of the prokaryotic ribosomal RNA sequence that can be used to distinguish if two organisms are of different species. In fact, the current taxonomic classification scheme laid out in the most recent edition of Bergey's Manual is based on 16s rRNA similarity in conjunction with phenotypic traits. However, with whole-genome sequencing getting ever cheaper and easier to accomplish, it has become more apparent that high 16s rRNA similarity does not warrant close similarity of the whole genome. Studies have shown, that there are many instances where taxonomically close organisms (e.g. same genus different species) are functionally more diverse, than some other organisms of a higher taxonomic rank[17, 18]. This of course is even more difficult if horizontal gene transfer is taken into account, where microbes share genetic material across species. This is in strong contrast to evolution observed in eukaryotic organisms, where genetic traits are generally only acquired or passed on from one generation to another.

This led me to ask, whether functional classification of microbial samples might be beneficial over traditional taxonomic classification. Ultimately the question is what microbes are capable of, rather than how close they are on the tree of life.

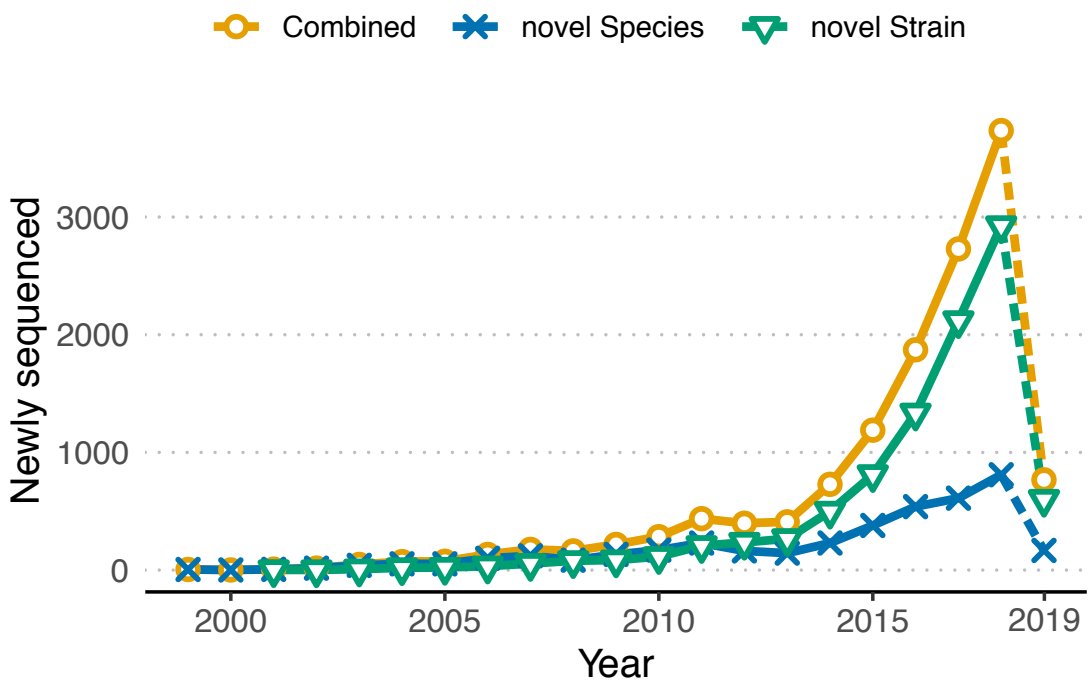


Figure 1.1: Drastic increase in completely sequenced bacteria since 2014. Displayed are the number of sequenced and fully assembled bacterial genomes available in Genbank since 1999. Counts are separated into novel species (blue, cross), novel strains of previously sequenced species (green, triangle) and both combined (yellow, circle). Until 2010 the year by year increase stayed roughly stable. 2011 is the first year where a larger increase was visible. 2012 presented the first year in which more strains of previously sequenced species than novel species were completely assembled. The trend curve from 2018 to 2019 is dashed as the count only reflect the the number until beginning of April 2019.

1.3 Towards identifying functional similarities

To discern which functional capabilities a microbe offers we first have to establish what we actually constitute as a function. I followed the approach of gene ontology and looked at molecular function of proteins, i.e. any activity of a gene product. This is not only restricted to enzymatic activity of proteins but also includes proteins like transporter, signaling, regulation or binning proteins. Identifying functional capabilities of proteins has been a staple in the field of computational biology for many years. A multitude of databases like Brenda[19], Pfam[20] or Gene Ontology[21, 22] are available, describing functional properties and protein similarities in various ways. Since I was primarily interested in functional similarity between two proteins, I specifically looked at the Enzyme Commission[23] (E.C.) classification and molecular function portion of Gene Ontology. Both of these schemata follow a hierarchy in describing the functional properties of proteins. The E.C. classification follows a strict four level, tree like hierarchy, where each consecutive level describes the enzymatic functional in more detail. As an example EC 1.1.1.1 is classified as Oxidoreductase (EC 1.-.-.-) acting on the CH-OH group of donors (EC 1.1.-.-) with NAD+ or NADP+ as acceptor (EC 1.1.1.-). The fourth and final level is the generic name, i.e. alcohol dehydrogenase. Due to the nature of this strict hierarchical structure, defining functional similarity based on EC number is fairly straight forward. Depending on the chosen threshold, functional identity is a simple binary decision, i.e. are the two EC annotations identical down to the defined level in the hierarchy.

Gene ontology on the other hand is structured as three separate directed acyclic graphs. Descriptors (nodes) are interconnected to each other with directed edges describing their relation (e.g. 'is a', 'part of', 'regulates'). The only other restriction imposed on the graph is that any path followed though the graph cannot loop back to the originating node. In other words, the graph has to be acyclic. Out of the three GO subgraphs, molecular function contains the information on functionality of proteins. GO cannot be transferred into a tree structure since multiple paths leading from one node to another with differing lengths can exists. Defining a strict hierarchy for GO is therefore impossible. Additionally we can observe examples, where GO terms reach a detail depth that is not reflected in EC annotations (see Figure 1.3). For example inositol 2-dehydrogenase (GO:0050112) is listed as a child term of alcohol dehydrogenase (GO:0004022) in the annotation scheme of Gene Ontology. One could argue that this represents a more detailed description of the enzymatic activity. In the EC hierarchy however inositol 2-dehydrogenase (EC 1.1.1.18) is described on the same level as alcohol dehydrogenase (1.1.1.1), both being "children" of EC 1.1.1.-. While the potential for increased detail in GO is in general very desirable, it also leads to an increased complexity when trying to establish functional similarity. This is especially significant if proteins are annotated to different levels of completeness.

Many different techniques to assess similarity of graphs (or subgraphs) are available. One that was specifically designed to work with GO is employed by the hosts of the CAFA[24, 25] competition (Critical Assessment of Functional Annotations). While their approach[26] is generalized and can be applied to much more than just binary data (i.e.

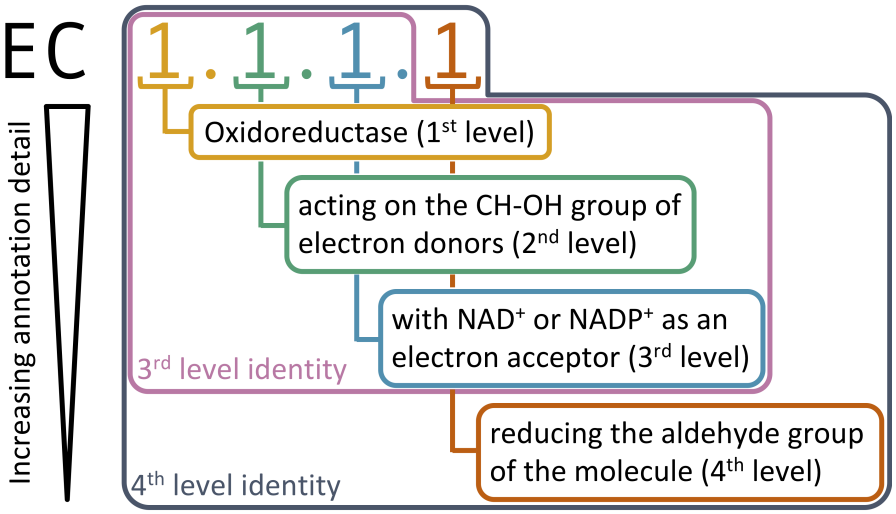


Figure 1.2: Exemplary break down of the hierarchical EC annotation for alcohol dehydrogenase. The Enzyme Commission annotation of is a hierarchical classification scheme consisting of four levels. Each consecutive level describes the enzymatic functionality in increasing detail. Displayed is the EC annotation for alcohol dehydrogenase (EC 1.1.1.1). Different levels of EC annotation are used to describe predictive quality of function prediction algorithms like HFSP.

is a node present in a graph or not). If the method is simplified to suit the binary data, it yields a distance that is the inverse of the Jaccard index. GO annotation has a much higher complexity than the enzyme commission classification scheme. On the flip-side the limited amount of data, respectively varying degrees of annotation completeness for proteins of the same function introduce new challenges. While I ultimately see GO as the superior annotation at its current state it proved too sparse to be useful in creating a measure to compare functional similarity between proteins. Thus I used the Enzyme Commission annotation in this work to establish a measure that can asses functional similarity between proteins. In future applications it might be worth revisiting GO. Apart from being potentially more descriptive and detailed, GO describes not only enzymatic function but rather all molecular function.

Using a set of enzymes I optimized a measure that relies on sequence identity and (ungapped) alignment length of pairwise protein sequence alignments to establish the functional similarity between any two proteins. Homology-derived functional similarity of proteins (or HFSP in short) follows principles similar to HSSP (Homology-derived secondary structure or proteins)[27]. In its essence, HFSP states that sequence identify by itself is not enough to warrant two proteins performing the same molecular function. Short alignments (eg. 50 amino acids) require much higher sequence identity to make reliable assumptions about functional similarity. On the other hand increasing alignment length, allows for much larger variability in overall sequence identity, while still retaining the same or similar function.

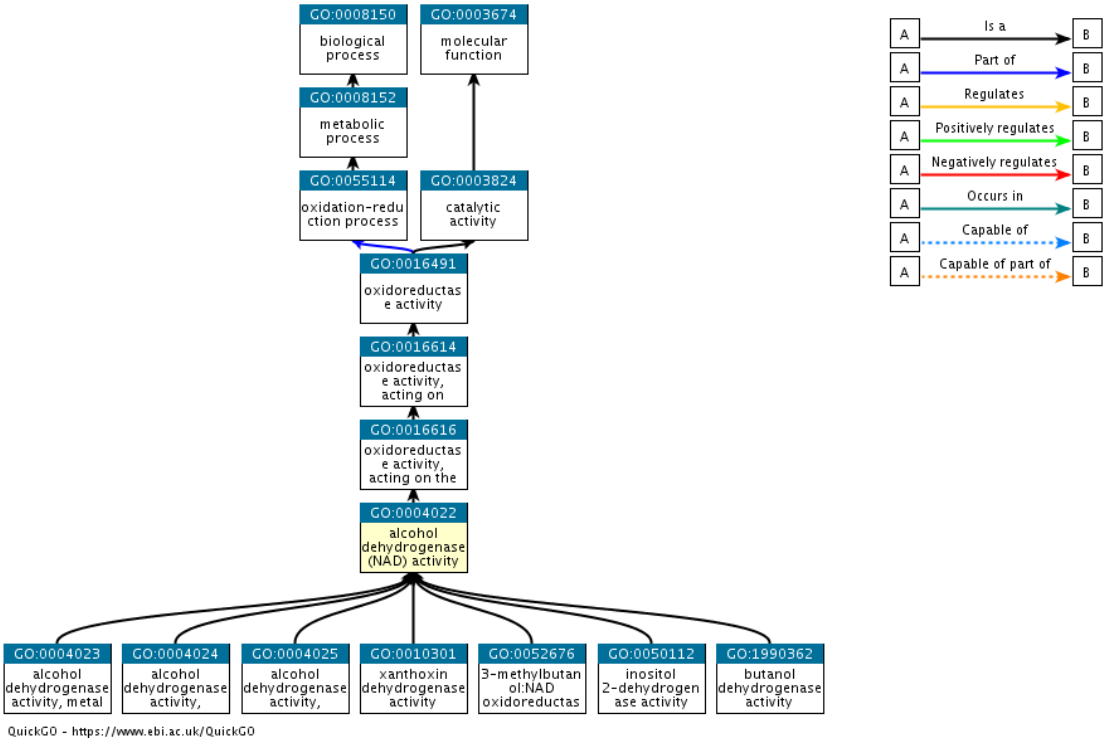


Figure 1.3: Gene Ontology annotation for alcohol dehydrogenase shows important difference to EC annotation. Visualized is the Gene Ontology annotation of alcohol dehydrogenase (EC 1.1.1.1). Individual GO terms are represented as box, while term relations are visualized as arrows. Terms for both the biological process as well as molecular function subtree as shown. Alcohol dehydrogenase is highlighted in yellow. In this specific case the EC annotation of alcohol dehydrogenase is precisely reflected in the molecular function branch of the GO annotation. EC 1.-.-.- maps to GO:0016491; EC 1.1.-.- maps to GO:0016614, EC 1.1.1.- maps to GO:0016616; EC 1.1.1.1 maps to GO:0004022. However here the GO annotation of inositol 2-dehydrogenase activity (GO:0050112) is considered to be a child of alcohol dehydrogenase (GO:0004022). In a sense it is a more detailed description the catalytic activity. In the EC annotation hierarchy however inositol 2-dehydrogenase has its own fourth level EC annotation, namely 1.1.1.18.

1.4 Functional similarity over taxonomic similarity

With the ability to determine if any two proteins share similar molecular function, the idea can be applied to whole bacterial proteomes. To determine if two proteomes carry similar functional capabilities, the entire pool of possible functions across all proteomes has to establish first. Using HFSP, functional similarities between all proteins extracted from a given set of proteomes are generated. The proteins are then related to each other creating a similarity network. Using network analysis methods, proteins of similar function are clustered into units representing specific functional properties. This process generates the aforementioned pool of possible functionality. Now each organism can be represented by a functional profile, a vector that list the presence of absence of functions encoded for by the organisms proteome. It is of note that functional units can be specific to proteomes, i.e. unique to certain organisms. Comparing the functional profiles of individual microbes, functional similarity between organisms can finally be established.

Zhu *et. al*'s work[18] serves as a proof of concept for this process. They determined that classifying microbes by examining their functional potential correlates with traditional taxonomic classification. However, they were also able to pick up clear signals suggesting cases where close taxonomic relation ship between organisms does not guarantee high functional similarity and vice versa.

The major drawback of Zhu *et. al*'s study is their uses of limited data at that time. The lack of information during the creation of the microbial functionality pool, has a direct result in the resolution of final classification. For example, they suggest that roughly two thirds of microbial functionality is organism unique, i.e. not present in any but one organism. Additionally many of those functions are derived from hypothetical proteins, in other words proteins predicted during the gene finding process of genetic assembly. This fraction of the assumed functionality, will most certainly contain emergent functionality that previously has not been described and might even be unique to an organism. On the flip side however it will also introduce a significant error margin potentially leading to a lower than presumed organism similarity. By introducing much larger numbers of proteomes into the generation of the initial collection of microbial functions, I will be able to achieve better resolution at describing the functional similarity of microbes. Clades of functionally similar microbes should separate much clearer from each other than previously observed. At the same time functionality that is unique to single or a small group of organisms will be be more significant due to a reduction of "false uniques". Additionally, with the knowledge of whether certain functionality ascribed to organisms due to plasmids (i.e. horizontal gene transfer), lateral spread of functionality in microbes can be traced as well. This plasticity of microbial capabilities and as a matter of fact evolution, is something that is near impossible to trace by relying on the more static nature of traditional taxonomy.

1.5 Bridging the gap to communities

Given that a microbial community is a collection of microbial organisms, co-existing in a confined environmental space, we can make certain assumptions about them. One assumption is that if we are representing individual microbes as a repertoire of functionality, the same can be applied to microbial communities. In this view, at its core a microbial community is nothing but a collection of functional repertoires, a meta functional repertoire. This assumption holds true, independent of whether they exist in varying degrees of symbiosis or competition, i.e. whether microbes are capable of existing on their own, or whether they need some kind of functional interaction with other members of the community to survive. It can easily be the case that in one microbiome a specific subset of functions originates from one organism alone, whereas in a different microbiome the same subset of functionality might be distributed over two or more microbes. However, making the move to a grander scheme of understanding i.e. away from the functional properties of a single organism towards regarding the microbial community as an entity on its own it matters little which organism provides which functionality.

By today's standards a first step in investigating microbial communities is establishing taxonomic composition i.e. clade abundances in the sample. Popular metagenome analysis suites to perform analysis like this are Qiime2[28], MOTHUR[29], bioBakery[30] or MG-RAST[31]. The taxonomic composition is in most cases determined by one of two common ways. Algorithms included for example in Qiime2 use abundances of 16s rRNA information in the sample and compare those against databases like SILVA[32, 33] or Greengenes[34, 35] to establish taxonomic clade compositions. The other increasingly more common option is a hybrid approach that utilizes both 16s rRNA and other marker genes. Examples for those hybrid approaches are MetaPhlan[36] (part of bioBakery) and the algorithm used in MG-RAST. Taxonomic compositions established this way are then often used in conjunction with prior knowledge of phenotypic behavior of organisms belonging to detected clades to make inferences on the functional properties of said metagenomic communities. Additionally, 16s rRNA fragments (and fragments of other marker genes) can also be used as input into functional predictors, like PICRUSt[37]. However, PICRUSt still predicts function based on presence of marker genes, not the actual functional properties of the microbiome's gene content.

A more robust way of establishing functional properties of a microbial community would of course be to assess the functional properties not simply based of the community's taxonomic composition, or based of a few marker genes, but to directly infer function of all present gene content. Zhu et al. developed mi-faser[38], an algorithm to directly and reliably map reads from unassembled (meta-) genomes to a small but very strict and well described set of enzymes and their respective KEGG[39, 40, 41] pathways. The beauty of mi-faser is that this concept works with any reference database, i.e. the database could be replaced without the need to retrain or reoptimize the method. The aforementioned concept of assigning functional profiles to individual microbes is based on sorting proteins into functional units. By choosing a representative for each described functional unit, a new reference database can be created. Combining such a

reference database with a method like mi-faser could be used to infer molecular function to all reads present in a metagenomic sample. A functional profile for the whole microbial community can be generated this way instead of relying on a few marker genes, and inferring functional capabilities from taxonomic distribution. An approach like this is becoming increasingly vital, as more and more "whole meta-genomes", i.e. more than just ribosomal RNA has been amplified and sequenced, are publicly available. Databases like the Sequence Read Archive (SRA)[42, 43] contain large collections of reads, on both metagenomic as well as single organism scale, and are widely used endpoints for researchers to retrieve datasets as well as deposit their sequencing data. Aggregation services like MetaSeek[44] on the other hand increasingly simplifies narrowing down available dataset to the desired data using a set of metadata properties. Being able to compare functional compositions between metagenomes collected through aggregation services will be more and more important in the future.

1.6 Genetic variation of host influences microbial composition of microbiome

In the last last few years increasing evidence demonstrates a direct link between genetic variation of host organisms and the microbial composition of their microbiomes [45, 46, 47]. The functional impact of the hosts genetic variations could be seen as an environmental factor to the proliferation of certain microbial organisms. Understanding the interaction between those two variables can be a crucial part of developing new predictive models, as well as aid in optimizing clinical applications.

1.7 Contents of this work

In this work I postulate that a standardized functional assessment of organisms can provide a far more stable and meaningful way for organism classification. I present the concept of *fusion* and describe ideas and concepts to cope with the massive amount of sequencing data. I detail the progress towards establish functional similarities for today's collection of fully sequenced microbial organisms ($\sim$ 9000 bacteria & $\sim$ 45 million proteins). I discuss the challenges in clustering resulting from the massive dataset and offer possible solutions. Finally I describe how I provide already available data to the scientific community.

Chapter 2 introduces HFSP[48] (Homology-derived Functional Similarity of Proteins) the measure that is used to establish the functional similarities between proteins. I use MMSeqs2[49] over PSI-BLAST [50] to greatly improve the speed at which protein sequence alignments can be generated, i.e. a more than 40-fold reduction of compute time.

Chapter 3 details the challenges I encountered in the attempt to cluster the protein similarity network into functional units, and offer suggestions how to overcome those challenges. Additionally, I lay out ideas and projects that could be enabled once the clustering is solved.

Chapter 4 describes *fusionDB* my effort to create a database of functional profiles for microbes. *fusionDB* currently contains the functional profiles of 1,374 taxonomically distinct bacterial organisms resulting from of a prove of concept analysis of *fusion*. Additionally, *fusionDB* includes available environmental metadata associated with these 1,374 organisms.

Finally **Chapter 5** highlights methodological approaches to variation analysis. Ideas presented there could be used to establish links between genetic variation of host organisms and their functional impact to the microbial composition of bacterial communities inhabiting the host.

1.8 References

[1] A. V. Leeuwenhoek. More observations from mr. leewenhook, in a letter of sept. 7. 1674. sent to the publisher. *Philosophical Transactions of the Royal Society of London*, 9(108):178–182, 1674. doi:doi:10.1098/rstl.1674.0057.

[2] A. V. Leeuwenhoek. Observations, communicated to the publisher by mr. antony van leewenhoeck, in a dutch letter of the 9th octob. 1676. here english'd: concerning little animals by him observed in rain-well-sea- and snow water; as also in water wherein pepper had lain infused. *Philosophical Transactions of the Royal Society of London*, 12(133):821–831, 1677. doi:doi:10.1098/rstl.1677.0003.

[3] R. Fleischmann, M. Adams, O. White, R. Clayton, E. Kirkness, A. Kerlavage, C. Bult, J. Tomb, B. Dougherty, J. Merrick, and e. al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269:496–512, 1995.

[4] L. Pasteur. *Études sur le vin: ses maladies, causes qui les provoquent, procédés nouveaux pour le conserver et pour le vieillir*. Simon Raçou et Comp., 1873.

[5] L. Pasteur. *Études sur la bière: ses maladies, causes qui les provoquent, procédé pour la rendre inaltérable; avec une théorie nouvelle de la fermentation*. Gauthier-Villars, 1876.

[6] R. Koch. Die Ätiologie der milzbrand-krankheit, begründet auf die entwicklungsgeschichte des bacillus anthracis. *Beiträge zur Biologie der Pflanzen*, 2(2):277, 1876.

[7] R. Koch. Die aetiologie der tuberkulose. *Berliner Klinischen Wochenschrift*, 19:221–230, 1882.

[8] R. Koch. Ueber den augenblicklichen stand der bakteriologischen choleradiagnose. *Zeitschrift für Hygiene und Infektionskrankheiten*, 14(1):319–338, 1893. doi:10.1007/bf02284324.

[9] F. Cohn. *Untersuchungen über Bacterien: I*, volume 1, pages 127–222. JU Kern, 1875.

[10] J. H. Brown. *Bergey's manual of determinative bacteriology.* American Public Health Association, 1939.

[11] J. G. Holt and N. Krieg. Bergey's manual of systematic bacteriology, vol. 1. *The Williams and Wilkins Co., Baltimore*, pages 1–1388, 1984.

[12] P. H. Sneath, N. S. Mair, M. E. Sharpe, and J. G. Holt. *Bergey's manual of systematic bacteriology. Volume 2.* Williams & Wilkins, 1986.

[13] C. R. Woese and G. E. Fox. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, 74:5088–5090, 1977.

[14] C. R. Woese, E. Stackebrandt, T. J. Macke, and G. E. Fox. A phylogenetic definition of the major eubacterial taxa. *Syst Appl Microbiol*, 6:143–51, 1985.

[15] C. R. Woese. Bacterial evolution. *Microbiological reviews*, 51(2):221–271, 1987.

[16] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic Acids Res*, 41(Database issue):D36–42, 2013. doi:10.1093/nar/gks1195.

[17] G. E. Fox, J. D. Wisotzkey, and P. Jurtshuk. How close is close: 16s rrna sequence identity may not be sufficient to guarantee species identity. *International Journal of Systematic and Evolutionary Microbiology*, 42(1):166–170, 1992. doi:doi:10.1099/00207713-42-1-166.

[18] C. Zhu, T. O. Delmont, T. M. Vogel, and Y. Bromberg. Functional basis of microorganism classification. *PLoS Comput Biol*, 11(8):e1004472, 2015. doi:10.1371/journal.pcbi.1004472.

[19] A. Chang, I. Schomburg, L. Jeske, S. Placzek, and D. Schomburg. Brenda in 2019: a european elixir core data resource. *Nucleic Acids Research*, 47(D1):D542–D549, 2018. doi:10.1093/nar/gky1048.

[20] A. Bateman, A. Smart, A. Luciani, G. A. Salazar, J. Mistry, L. J. Richardson, M. Qureshi, S. El-Gebali, S. C. Potter, R. D. Finn, S. R. Eddy, E. L. Sonnhammer, D. Piovesan, L. Paladin, S. C. Tosatto, and L. Hirsh. The pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1):D427–D432, 2018. doi:10.1093/nar/gky995.

[21] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9, 2000. doi:10.1038/75556.

[22] The Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic Acids Research*, 47(D1):D330–D338, 2018. doi:10.1093/nar/gky1055.

[23] A. Bairoch. The enzyme database in 2000. *Nucleic Acids Res*, 28(1):304–5, 2000.

[24] P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, G. Pandey, J. M. Yunes, A. S. Talwalkar, S. Repo, M. L. Souza, D. Piovesan, R. Casadio, Z. Wang, J. Cheng, H. Fang, J. Gough, P. Koskinen, P. Törönen, J. Nokso-Koivisto, L. Holm, D. Cozzetto, D. W. A. Buchan, K. Bryson, D. T. Jones, B. Limaye, H. Inamdar, A. Datta, S. K. Manjari, R. Joshi, M. Chitale, D. Kihara, A. M. Lisewski, S. Erdin, E. Venner, O. Lichtarge, R. Rentzsch, H. Yang, A. E. Romero, P. Bhat, A. Paccanaro, T. Hamp, R. Kaßner, S. Seemayer, E. Vicedo, C. Schaefer, D. Achten, F. Auer, A. Boehm, T. Braun, M. Hecht, M. Heron, P. Hönigschmid, T. A. Hopf, S. Kaufmann, M. Kiening, D. Krompass, C. Landerer, Y. Mahlich, M. Roos, J. Björne, T. Salakoski, A. Wong, H. Shatkay, F. Gatzmann, I. Sommer, M. N. Wass, M. J. E. Sternberg, N. Škunca, F. Supek, M. Bošnjak, P. Panov, S. Džeroski, T. Šmuc, Y. A. I. Kourmpetis, A. D. J. van Dijk, C. J. F. t. Braak, Y. Zhou, Q. Gong, X. Dong, W. Tian, M. Falda, P. Fontana, E. Lavezzo, B. Di Camillo, S. Toppo, L. Lan, N. Djuric, Y. Guo, S. Vucetic, A. Bairoch, M. Linial, P. C. Babbitt, S. E. Brenner, C. Orengo, B. Rost, et al. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10:221, 2013. doi:10.1038/nmeth.2340.

[25] Y. Jiang, T. R. Oron, W. T. Clark, A. R. Bankapur, D. D'Andrea, R. Lepore, C. S. Funk, I. Kahanda, K. M. Verspoor, A. Ben-Hur, C. E. Koo da, D. Penfold-Brown, D. Shasha, N. Youngs, R. Bonneau, A. Lin, S. M. Sahraeian, P. L. Martelli, G. Profiti, R. Casadio, R. Cao, Z. Zhong, J. Cheng, A. Altenhoff, N. Skunca, C. Dessimoz, T. Dogan, K. Hakala, S. Kaewphan, F. Mehryary, T. Salakoski, F. Ginter, H. Fang, B. Smithers, M. Oates, J. Gough, P. Toronen, P. Koskinen, L. Holm, C. T. Chen, W. L. Hsu, K. Bryson, D. Cozzetto, F. Minneci, D. T. Jones, S. Chapman, D. Bkc, I. K. Khan, D. Kihara, D. Ofer, N. Rappoport, A. Stern, E. Cibrian-Uhalte, P. Denny, R. E. Foulger, R. Hieta, D. Legge, R. C. Lovering, M. Magrane, A. N. Melidoni, P. Mutowo-Meullenet, K. Pichler, A. Shypitsyna, B. Li, P. Zakeri, S. ElShal, L. C. Tranchevent, S. Das, N. L. Dawson, D. Lee, J. G. Lees, I. Sillitoe, P. Bhat, T. Nepusz, A. E. Romero, R. Sasidharan, H. Yang, A. Paccanaro, J. Gillis, A. E. Sedeno-Cortes, P. Pavlidis, S. Feng, J. M. Cejuela, T. Goldberg, T. Hamp, L. Richter, A. Salamov, T. Gabaldon, M. Marcet-Houben, F. Supek, Q. Gong, W. Ning, Y. Zhou, W. Tian, M. Falda, P. Fontana, E. Lavezzo, S. Toppo, C. Ferrari, M. Giollo, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol*, 17(1):184, 2016. doi:10.1186/s13059-016-1037-6.

[26] R. Yang, Y. Jiang, S. Mathews, E. A. Housworth, M. W. Hahn, and P. Radivojac. A new class of metrics for learning on real-valued and structured data. *arXiv preprint arXiv:1603.06846*, 2016.

[27] B. Rost. Enzyme function less conserved than anticipated. *J Mol Biol*, 318(2):595–608, 2002. doi:10.1016/s0022-2836(02)00016-5.

[28] E. Bolyen, J. R. Rideout, M. R. Dillon, N. A. Bokulich, C. Abnet, G. A. Al-Ghalith, H. Alexander, E. J. Alm, M. Arumugam, F. Asnicar, Y. Bai, J. E. Bisanz, K. Bittinger, A. Brejnrod, C. J. Brislawn, C. T. Brown, B. J. Callahan, A. M. Caraballo-Rodríguez, J. Chase, E. Cope, R. Da Silva, P. C. Dorrestein, G. M. Douglas, D. M. Durall, C. Duvallet, C. F. Edwardson, M. Ernst, M. Estaki, J. Fouquier, J. M. Gauglitz, D. L. Gibson, A. Gonzalez, K. Gorlick, J. Guo, B. Hillmann, S. Holmes, H. Holste, C. Huttenhower, G. Huttley, S. Janssen, A. K. Jarmusch, L. Jiang, B. Kaehler, K. B. Kang, C. R. Keefe, P. Keim, S. T. Kelley, D. Knights, I. Koester, T. Kosciolek, J. Kreps, M. G. I. Langille, J. Lee, R. Ley, Y.-X. Liu, E. Loftfield, C. Lozupone, M. Maher, C. Marotz, B. D. Martin, D. McDonald, L. J. McIver, A. V. Melnik, J. L. Metcalf, S. C. Morgan, J. Morton, A. T. Naimey, J. A. Navas-Molina, L. F. Nothias, S. B. Orchanian, T. Pearson, S. L. Peoples, D. Petras, M. L. Preuss, E. Pruesse, L. B. Rasmussen, A. Rivers, I. I. M. S. Robeson, P. Rosenthal, N. Segata, M. Shaffer, A. Shiffer, R. Sinha, S. J. Song, J. R. Spear, A. D. Swafford, L. R. Thompson, P. J. Torres, P. Trinh, A. Tripathi, P. J. Turnbaugh, S. Ul-Hasan, J. J. J. van der Hooft, F. Vargas, Y. Vázquez-Baeza, E. Vogtmann, M. von Hippel, W. Walters, Y. Wan, M. Wang, et al. Qiime 2: Reproducible, interactive, scalable, and extensible microbiome data science. *PeerJ Preprints*, 6:e27295v2, 2018. doi:10.7287/peerj.preprints.27295v2.

[29] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, and C. J. Robinson. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, 75(23):7537–7541, 2009.

[30] L. J. McIver, G. Abu-Ali, E. A. Franzosa, R. Schwager, X. C. Morgan, L. Waldron, N. Segata, and C. Huttenhower. biobakery: a meta'omic analysis environment. *Bioinformatics*, 34(7):1235–1237, 2018. doi:10.1093/bioinformatics/btx754.

[31] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards. The metagenomics rast server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9(1):386, 2008. doi:10.1186/1471-2105-9-386.

[32] C. Quast, E. Pruesse, J. Gerken, J. Peplies, P. Yarza, P. Yilmaz, T. Schweer, and F. O. Glöckner. The silva ribosomal rna gene database project: improved data

processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596, 2012. doi:10.1093/nar/gks1219.

[33] C. Quast, E. Pruesse, J. Gerken, J. Peplies, L. W. Parfrey, P. Yarza, T. Schweer, W. Ludwig, F. O. Glöckner, and P. Yilmaz. The silva and "all-species living tree project (ltp)" taxonomic frameworks. *Nucleic Acids Research*, 42(D1):D643–D648, 2013. doi:10.1093/nar/gkt1209.

[34] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and Environmental Microbiology*, 72(7):5069–5072, 2006. doi:10.1128/aem.03006-05.

[35] D. McDonald, M. N. Price, J. Goodrich, E. P. Nawrocki, T. Z. DeSantis, A. Probst, G. L. Andersen, R. Knight, and P. Hugenholtz. An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The Isme Journal*, 6:610, 2011. doi:10.1038/ismej.2011.139.

[36] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9:811, 2012. doi:10.1038/nmeth.2066.

[37] M. G. I. Langille, J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. Vega Thurber, R. Knight, R. G. Beiko, and C. Huttenhower. Predictive functional profiling of microbial communities using 16s rrna marker gene sequences. *Nature Biotechnology*, 31:814, 2013. doi:10.1038/nbt.2676.

[38] C. Zhu, M. Miller, S. Marpaka, P. Vaysberg, M. C. Ruhlemann, G. Wu, F. A. Heinsen, M. Tempel, L. Zhao, W. Lieb, A. Franke, and Y. Bromberg. Functional sequencing read annotation for high precision microbiome analysis. *Nucleic Acids Res*, 46(4):e23, 2018. doi:10.1093/nar/gkx1209.

[39] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, 2000.

[40] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*, 45(D1):D353–d361, 2017. doi:10.1093/nar/gkw1092.

[41] M. Kanehisa, Y. Sato, M. Furumichi, K. Morishima, and M. Tanabe. New approach for understanding genome variations in kegg. *Nucleic Acids Res*, 47(D1):D590–d595, 2019. doi:10.1093/nar/gky962.

[42] R. Leinonen, H. Sugawara, M. Shumway, and C. International Nucleotide Sequence Database. The sequence read archive. *Nucleic Acids Res*, 39(Database issue):D19–21, 2011. doi:10.1093/nar/gkq1019.

[43] Y. Kodama, M. Shumway, R. Leinonen, and C. International Nucleotide Sequence Database. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res*, 40(Database issue):D54–6, 2012. doi:10.1093/nar/gkr854.

[44] A. Hoarfrost, N. Brown, C. Brown, and C. Arnosti. Sequencing data discovery with metaseek. *Bioinformatics*, 2019, in review.

[45] R. Blekhman, J. K. Goodrich, K. Huang, Q. Sun, R. Bukowski, J. T. Bell, T. D. Spector, A. Keinan, R. E. Ley, D. Gevers, and A. G. Clark. Host genetic variation impacts microbiome composition across human body sites. *Genome biology*, 16(1):191–191, 2015. doi:10.1186/s13059-015-0759-1.

[46] J. K. Goodrich, E. R. Davenport, A. G. Clark, and R. E. Ley. The relationship between the human genome and microbiome comes into view. *Annual review of genetics*, 51:413–433, 2017. doi:10.1146/annurev-genet-110711-155532.

[47] C. C. R. Smith, L. K. Snowberg, J. Gregory Caporaso, R. Knight, and D. I. Bolnick. Dietary input of microbes and host genetic variation shape among-population differences in stickleback gut microbiota. *The ISME journal*, 9(11):2515–2526, 2015. doi:10.1038/ismej.2015.64.

[48] Y. Mahlich, M. Steinegger, B. Rost, and Y. Bromberg. Hfsp: high speed homology-driven function annotation of proteins. *Bioinformatics*, 34(13):i304–i312, 2018. doi:10.1093/bioinformatics/bty262.

[49] M. Steinegger and J. Soding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*, 35(11):1026–1028, 2017. doi:10.1038/nbt.3988.

[50] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, 1997.

2 HFSP: high speed homology-driven function annotation of proteins

2.1 Preface

The first step in the *fusion* protocol to establish functional similarities between microbes is to establish the functional similarity between all microbial proteins. In order to achieve this, I developed HFSP (Homology-derived Functional Similarity of Proteins). HFSP follows the principles of HSSP (Homology-derived Secondary Structure of Proteins). Hence, HFSP's basis is a homology search that establishes alignments between the microbial proteins. HSSP uses PSI-BLAST for this step. Despite being the defacto default homology search algorithm in the Bioinformatics community, PSI-BLAST is a well known bottleneck in large scale alignment analysis. For my purposes where I have to generate alignments between millions of proteins, this is especially noticeable. To this end I opted to use the state-of-the-art homology search algorithm MMSeq2 in PSI-BLAST's stead.

Using MMSeqs2 I generated alignments between enzymes of an up to date set of experimentally validated enzymes. The protein pairs' resulting ungapped alignment length and sequence identity places all enzyme pairs into a two dimensional space. I also know for each pair if they share the same functional annotation determined by the Enzyme Commission. Combining the knowledge of functional similarity and the spacial placement of the enzyme pair I optimized a gradient decay curve that optimally separates functionally identical from functionally different enzymes based on their shared alignment length and sequence identity.

Evaluation of HFSP revealed more than 40-fold gain in speed over traditional methods, while maintaining high precision of the predictions. Additionally, I used HFSP to evaluate functional annotation of proteins deposited in Swiss-Prot and Uniprot, and discovered large parts with potential miss-annotation.

Implementation, evaluation and execution of the work was done by me. Martin Steinegger help with the utilization of MMSeqs2, and provided adaptions to MMSeqs2 in order to be fully functional for the purpose of HFSP. The manuscript was drafted by all authors.

2.2 Journal Article: Mahlich et al. Bioinformatics 2018

Bioinformatics, 34, 2018, i304–i312
doi: 10.1093/bioinformatics/bty262
ISMB 2018

OXFORD

HFSP: high speed homology-driven function annotation of proteins

Yannick Mahlich[1,2,3,*], Martin Steinegger[2,4,5], Burkhard Rost[2,3,6,7,8] and Yana Bromberg[1,3,9,*]

[1]Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ 08873, USA, [2]Computational Biology & Bioinformatics - i12 Informatics and [3]Institute for Advanced Study, Technical University of Munich (TUM), Munich 85748, Germany, [4]Quantitative and Computational Biology Group, Max-Planck Institute for Biophysical Chemistry, Göttingen 37077, Germany, [5]Department of Chemistry, Seoul National University, Seoul 08826, Korea, [6]TUM School of Life Sciences Weihenstephan (WZW), Technical University Munich (TUM), Freising 85354, Germany, [7]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA, [8]New York Consortium on Membrane Protein Structure (NYCOMPS), New York, NY 10032, USA and [9]Department of Genetics, Human Genetics Institute, Rutgers University, Piscataway, NJ 08854, USA

*To whom correspondence should be addressed.

Abstract

Motivation: The rapid drop in sequencing costs has produced many more (predicted) protein sequences than can feasibly be functionally annotated with wet-lab experiments. Thus, many computational methods have been developed for this purpose. Most of these methods employ homology-based inference, approximated via sequence alignments, to transfer functional annotations between proteins. The increase in the number of available sequences, however, has drastically increased the search space, thus significantly slowing down alignment methods.

Results: Here we describe homology-derived functional similarity of proteins (HFSP), a novel computational method that uses results of a high-speed alignment algorithm, MMseqs2, to infer functional similarity of proteins on the basis of their alignment length and sequence identity. We show that our method is accurate (85% precision) and fast (more than 40-fold speed increase over state-of-the-art). HFSP can help correct at least a 16% error in legacy curations, even for a resource of as high quality as Swiss-Prot. These findings suggest HFSP as an ideal resource for large-scale functional annotation efforts.

Contact: ymahlich@bromberglab.org or yanab@rci.rutgers.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The recent rapid drop in the cost of DNA-sequencing has produced a large number of fully sequenced genomes. For prokaryotes, for example, this represents a more than 6-fold growth (1400–9000 in GenBank (Benson *et al.*, 2013)) in the last 5 years alone. While this increase in data enables many types of research, experimental annotation lags far behind. In particular, the speed (or lack thereof) of experimental evaluation and validation of protein molecular functionality clearly necessitates computational approaches. In fact, many methods (Jiang *et al.*, 2016; Radivojac *et al.*, 2013) have already been developed for this purpose, the vast majority of which rely on transfer of functional annotation by homology (Loewenstein *et al.*, 2009). Mistakes in available annotations (Schnoes *et al.*, 2009), inconsistencies in experiments as well as simply missing or yet unknown functions make these sequence similarity-based methods error-prone (Clark and Radivojac, 2011). Furthermore, organism-focused research interests result in more detailed annotations for a non-random subset of proteins, where homologous proteins of identical functionality in another species are often annotated significantly less thoroughly. Evaluating the performance of computational annotation methods is complicated by the absence of large, well curated and 'evenly' functionally annotated protein sets, representing the entire breadth of available biomolecular functionality.

Protein sets that are used as benchmarks of prediction employ annotation ontologies, i.e. standardized terms and their relationships. One such benchmark set is enzymes with Enzyme Commission (Bairoch, 2000) (EC) numbers. EC numbers reflect a four level hierarchy, where each consecutive level is a more precise specification of

the annotation on the previous level. For example, enzymes classified as EC: 1.1.1.- are oxidoreductases (first level), acting on the CH-OH group of electron donors (second level), with NAD^+ or $NADP^+$ as an electron acceptor (third level). The fourth and most specific level might then annotate an enzyme as alcohol dehydrogenase (EC: 1.1.1.1), i.e. reducing the aldehyde group of the molecule. Note that dashes ('-') in EC numbers indicate lack of specificity of functional annotation at that level. While EC numbers facilitate comparison of functions across enzymes, the annotation specificity at the same EC level varies; e.g. the class of serine/threonine protein kinases (EC: 2.7.11.-) contains a category EC: 2.7.11.1 (fourth level annotation = 1) that collects all kinases that are non-specific or whose specificity has not been analyzed to date. On the other hand, serine/threonine protein kinases with the fourth level annotations between 2 and 32 are very specifically annotated, with each category limited to proteins that act on a particular substrate. Using EC annotations as a benchmark, thus, comes at the expense of variability in annotations even at the same level of the hierarchy. This, in turn, complicates establishing functional similarity of two proteins in a precise and balanced manner across the entire enzymatic activity spectrum.

By definition, using EC annotations also means missing out on non-enzymatic functionality. Other ontologies, like the molecular function branch of Gene ontology (Ashburner *et al.*, 2000) (GO) do not have this limitation. GO, however, employs a different, even more detailed, strategy in defining function than EC. The number of GO annotation levels varies by ontology sub-branch. Moreover, one protein can (and likely does) have multiple functional GO terms assigned to it (e.g. both copper ion binding and DNA binding terms describe the function of P53; AmiGo 2.4.6; PMID: 15358771, PMID: 7824276). Thus, comparing GO annotations may lead to much stronger distortions of similarity than skewed or even incomplete EC numbers. Note that moonlighting (Khan *et al.*, 2014) proteins, i.e. proteins that can be assigned multiple specific functions, further confuse functional similarity metrics.

As a consequence of the drastic increase in genomic and protein sequences in need of annotation, the search space for all computational function assignment methods has also increased. A centerpiece of much of sequence analysis efforts is the Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990; Altschul *et al.*, 1997) family. We note that with the quasi exponential growth in search space, while PSI-BLAST (Altschul *et al.*, 1997) may still remain viable for the analysis of a single protein, large scale evaluations are not time-feasible. Many methods that reduce runtime while retaining or increasing alignment accuracy have been developed over the last years, including caBLASTp (Daniels *et al.*, 2013), HHblits (Remmert *et al.*, 2012) and MMseqs2 (Steinegger and Soding, 2017). However, replacing (PSI-) BLAST in any bioinformatics pipeline with another alignment method requires parameter re-optimization or even a complete method overhaul.

Existing function prediction methods are very sophisticated, using a variety of inputs (e.g. structure and literature mining) and computational techniques (e.g. machine learning). However, here we focused on Homology-derived Secondary Structure of Proteins (HSSP) (Rost, 1999; Rost, 2002; Sander and Schneider, 1991)—a simple distance metric that infers protein function and structure similarity from sequence identity and alignment length. We optimized HSSP parameters to classify protein pairs as functionally identical or different using the results of MMseqs2, a lightning-fast alignment method. We found that our newly developed Homology-derived Functional Similarity of Proteins (HFSP) method is 40-fold faster than HSSP, while retaining HSSP precision in annotating enzymatic functionality of proteins (85% precision; Fig. 1).

Analyzing existing protein databases with our method, we showed that currently available computationally determined annotations in even the manually curated Swiss-Prot (The UniProt, 2017) database are incorrect for at least a sixth of the cases. We suggest that these errors are likely due to loosely defined rules of homology-based propagation of functional annotations. With the number of protein sequences in public databases bordering on 100 million and growing, HFSP is well suited to help improve the quality of existing and newly assigned functional annotations.

2 Materials and methods

2.1 Extraction of datasets

We extracted a set of reviewed proteins from Swiss-Prot with only one, EC (Bairoch, 2000) annotation per protein (complete at all four levels; 214 000 proteins; *Swiss-Prot set*). The 2002 (latest) formula for computing the HSSP (Rost, 1999; Rost, 2002) distances was developed on a combined set of Swiss-Prot (The UniProt, 2017) and Protein Data Bank (Berman *et al.*, 2002) proteins. To validate the performance of HSSP reported in Rost (1999) and Rost (2002), we extracted proteins from the Swiss-Prot set that had experimental evidence of protein existence (e.g. crystal structure, protein detection by antibodies, etc.) and an EC annotation in BRENDA (Placzek *et al.*, 2017). The resulting proteins (*Swiss-Prot 2017* set; 7022 proteins) were further filtered to retain entries appearing in the database before January 2002 (*Swiss-Prot 2002*, 3, 908 proteins). Both *Swiss-Prot 2017* and *2002* datasets were extracted in October 2017 (Uniprot release 2017_09) and redundancy reduced to 98% sequence similarity and 98% target sequence coverage with CD-HIT (Fu *et al.*, 2012; Li and Godzik, 2006). *Swiss-Prot 2002* contained 3801 proteins with 1481 unique EC annotations and *Swiss-Prot 2017* containing 6835 proteins with 2552 unique EC annotations (Supplementary Material).

Swiss-Prot 2017 was further split into sets containing only prokaryotic (*Swiss-Prot$_{pro}$ 2017*, 2572 proteins) or eukaryotic (*Swiss-Prot$_{euk}$ 2017*, 4263 proteins) proteins. Finally, we extracted two more Swiss-Prot subsets from: (i) proteins that did not have an EC annotation (293 058 proteins) and (ii) proteins with incomplete or multiple EC annotations (48 536 proteins).

2.2 Aligning proteins

To augment the homology profiles used in alignments [by both PSI-BLAST (Altschul *et al.*, 1997) and MMseqs2], we computed alignments of all proteins in our datasets (Swiss-Prot 2002, Swiss-Prot 2017, Swiss-Prot$_{pro}$ 2017 and Swiss-Prot$_{euk}$ 2017) against proteins in the full (non-reduced) Swiss-Prot (Uniprot release 2017_09). For each specific dataset, we then extracted only those alignments, where both proteins were present in that set (e.g. *both query and target protein in Swiss-Prot 2002*).

PSI-BLAST alignments where created with NCBI-BLAST version 2.2.29+. We ran three iterations of PSI-BLAST (-num_iterations 3). In each iteration, the top 500 hits (*E*-value 10^{-10}, -inclusion_ethresh 1e-10) were included into the profile. After the third round all alignments that satisfied the *E*-value $\leq 10^{-3}$ threshold (-evalue 1e-3) were considered for evaluation of performance.

MMseqs2 (Steinegger and Soding, 2017) parameters were chosen to mirror the PSI-BLAST runs. The alignment-mode (—alignment-mode 3) was set to calculate sequence identity between query and target over the full alignment length, i.e. analogous to BLAST. We ran three iterations (—num-iterations 3) of alignments including hits with an *E*-value $\leq 10^{-10}$ into the generated profile (—e-profile 1e-10).

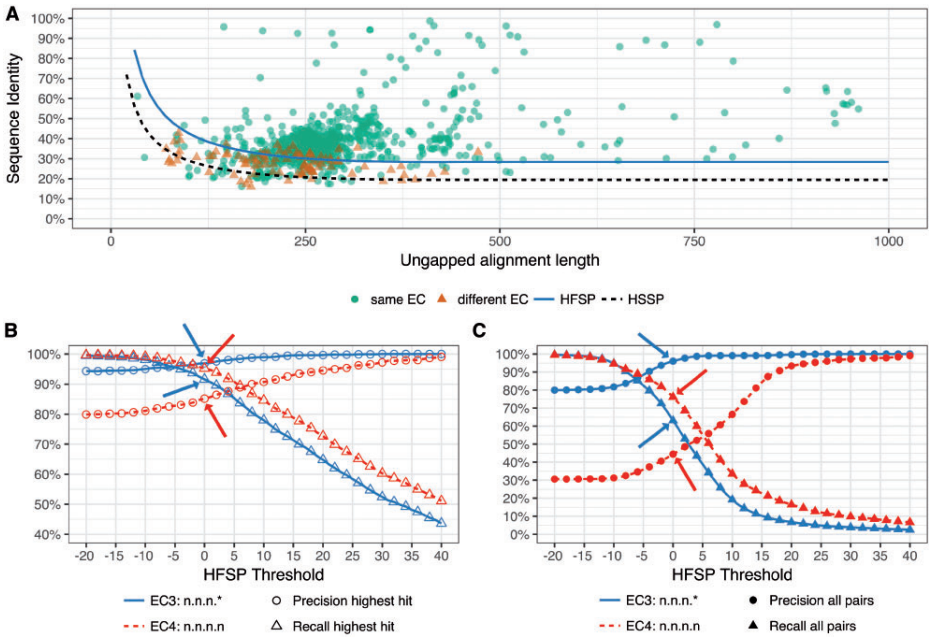


Fig. 1. *HFSP precisely predicts functional identity.* All Swiss-Prot 2002 protein pairwise alignments were mapped into the sequence identity versus ungapped alignment length space. In (**A**) protein pairs were differentiated according to identity of their EC level 3 (same EC annotation are green circles; different annotations are red triangles). The HFSP curve (HFSP = 0, light blue solid line) is shown relative to the HSSP curve (black dashed line). Protein pairs above the curve are predicted to be of same function, pairs below the curve of different function. In (**B**, **C**) precision (circles) and recall (triangles) in predicting functional identity, at third (blue, solid curve) and fourth (red, dashed curve) EC level for Swiss-Prot 2002. Arrows indicate performance at default cutoff of HFSP = 0. In (B) prediction was done using the highest HFSP scoring alignment per protein. In (C) all alignments were used, resulting in significantly worse performance

Only alignments of protein pairs with and *E*-value $\leq 10^{-3}$ were reported in the final result (-e 1e-3). The sensitivity (-s) cutoff for MMseqs2 prefiltering step was set to 5.6 (default value).

It had taken MMseqs2 1228 CPU hours to complete the alignment of our Swiss-Prot enzyme set (214 000 proteins) to the full (non-reduced) Swiss-Prot (555 594 proteins). Although MMSeqs2 was exceedingly fast for this set, note that it has been optimized to deal with much larger databases and, thus, it did not reach its full potential in speed. In earlier testing (Zhu *et al.*, 2015; Zhu *et al.*, 2018) with a dataset of ∼4.2 million proteins, the all-to-all protein alignment time for the MMseqs2 was ∼30 000 CPU hours (4.2e6 × 4.2e6 = ∼1.8e13 comparisons in roughly 4 days on 12 compute nodes with 24 CPUs each). In comparison, creating the same PSI-BLAST alignments took ∼1.3 million CPU hours (∼3 months on 78 compute nodes with 8 CPUs each). From these numbers, the HFSP speed-up (using MMseqs2) over HSSP (using PSI-BLAST) was estimated at over 40-fold and expected to grow significantly with database size.

2.3 Defining functional identity

Proteins sharing the same EC annotation at chosen (third or fourth level) were assigned functional identity. For example, L-lactate dehydrogenase and D-lactate dehydrogenase have EC assignments 1.1.1.27 and 1.1.1.28, respectively. Thus, at EC level 4, the proteins are different, but at EC level 3 they are the same, 1.1.1.

2.4 Retraining HSSP curve with MMseqs2

We used the *Swiss-Prot 2002* proteins and their third EC level annotations to develop the HFSP measure. Investigating the protein distribution of EC categories at the third EC level, we realized a strong distortion toward a few EC categories with exceptionally many associated proteins (Fig. 2C). This is in addition to other differences between EC categories (Fig. 2A and B). To compensate for this category bias, we limited the size of EC categories to no more than 50 proteins (randomly chosen for the 19 larger categories, Supplementary Table S1). We then extracted all MMseqs2 alignments for all *Swiss-Prot 2002* protein pairs in our set.

It has been previously shown that using class-balanced training sets is beneficial in the development of data driven classification models (Rost and Sander, 1993; Wei and Dunbrack, 2013). We therefore balanced the results in training to contain equal numbers of protein pairs with the same versus different third level EC annotations.

We first used cross-validation for training/testing our method; i.e. we split the data into 10 sets such that no sequence in one set shared more than 40% identity with a sequence in another set (CD-HIT clusters). In each of in 10 rounds of training, 1 set was retained for testing and the other 9 were used for training. Note that in each round of

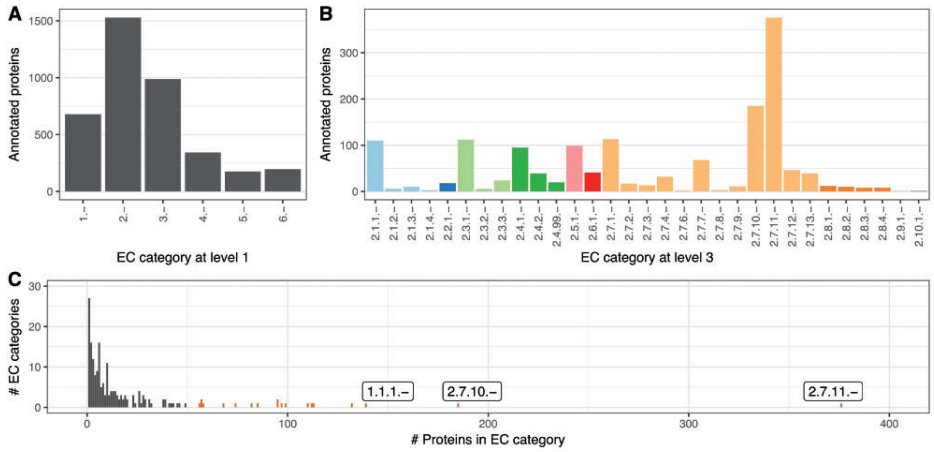


Fig. 2. *Strong bias in EC distribution.* Different EC categories contain different numbers of proteins with both general (**A**) EC level 1 and (**B**) more specific EC annotations. (**C**) This bias is particularly obvious for third level EC categories, with 2.7.11.-, 2.7.10.- and 1.1.1.- being the most prominent (first three bars from right; all ECs with more than 50 proteins are red)

cross-validation, we reintroduced into the testing set those proteins, which were originally removed for class balancing purposes. We optimized the parameters [originally factor = 480 and exponent = −0.32; Equation. (1), Supplementary Table S2] of the 2002 HSSP formula (Rost, 2002) to fit a new curve separating protein pairs of identical function from those of different functions in the two-dimensional space of sequence identity (y-axis) and ungapped alignment length (alignment length−number of gaps; x-axis). Pairs of same function proteins (identical annotation for EC) and a given threshold distance away from the curve along the y-axis were true positives (TP). Pairs that did not have the same function but were also above the threshold were false positives (FP). False negatives (FN) were pairs of same function but scoring below the threshold. We optimized for F_1 score [Equation (3)] using R's implementation of the Nelder–Mead method (Nelder and Mead, 1965), searching for a local optimal F_1 score, using combinations of exponents from −0.3 to −0.9 in steps of 0.05, and factor from 300 to 1500 in steps of 50.

$$HSSP = PIDE - \begin{cases} 100, & for\ L \leq 11 \\ 480 \cdot L^{-0.32 \times \left(1+e^{-\frac{L}{1000}}\right)}, & for\ 11 < L \leq 450 \\ 19.5, & for\ L > 450 \end{cases} \quad (1)$$

$$PIDE = Percent\ sequence\ identity\ of\ the\ alignment$$

$$L = ungapped\ alignment\ length$$

$$precision = \frac{TP}{TP+FP},\ recall = \frac{TP}{TP+FN} \quad (2)$$

$$F_1\ score = 2 \times \frac{precision \times recall}{precision + recall} \quad (3)$$

HFSP values for protein pairs were calculated using MMseqs2 results; Pearson correlation coefficient of HFSP to the HSSP values computed using PSI-BLAST results for same pairs. For each dataset, we calculated precision (i.e. how often a prediction of identical function is correct), recall (i.e. how many identical function pairs were correctly identified) and the F_1 score [Equations (2) and (3)] using HSSP and HFSP distance thresholds to determine true/false positives/negatives.

After evaluation was completed, we retrained as described above, but without testing, one HFSP curve on the complete balanced set of *Swiss-Prot 2002* protein pairs for all further use.

2.5 Using HFSP to make function predictions

We used the 6835 experimentally annotated proteins with 2552 unique EC annotations of *Swiss-Prot 2017* as the reference database for all further function predictions. For every protein, only the highest HFSP-scoring protein match (≥0; excluding self-matches) was used to annotate function. We thus predicted functions of proteins in the complete Swiss-Prot set of enzymes. Curiously, some EC numbers used in Swiss-Prot protein annotation did not have any members in the experimentally annotated *Swiss-Prot* 2017 reference set. The proteins annotated with these EC numbers (32 201 proteins at fourth and 381 proteins at third EC level, respectively) were considered false positives by default. Note that we are still unclear about the origins and experimental support of these annotations. Additionally, some proteins did not produce any alignments, and for others the highest hits did not reach our HFSP cutoff = 0. For these, no functional assignment could be made.

3 Results

3.1 HFSP scores correlate with HSSP, but are produced more than 40-fold faster

We trained, evaluated, and defined the HFSP [Homology-derived Functional Similarity of Proteins; Equation (4)] as described in Materials and Methods.

$$HFSP = PIDE - \begin{cases} 100, & for\ L \leq 11 \\ 770 \cdot L^{-0.33 \times \left(1 + e^{-\frac{L}{1000}}\right)}, & for\ 11 < L \leq 450 \\ 28.4, & for\ L > 450 \end{cases} \quad (4)$$

HFSP uses MMseqs2 iterative profiles as they have three major advantages over PSI-BLAST: (i) compositional bias correction to suppress high scoring non-homologous alignments, (ii) profile computation by only considering the 1000 most diverse sequences (PSI-BLAST uses the *n* BEST scoring hits) and (iii) realignment to reduce over-extension (Frith *et al.*, 2008); over-extension includes sequences into the profile at the edges of the alignment threshold in consecutive iterations. Thus, MMseqs2 alignments of smaller and more distant proteins tend to be more compact, favoring higher sequence identity, and thus leading to slightly higher HSSP scores calculated using the original equation [Equation (1)]. These differences in alignment methods, however, do not significantly affect the HSSP scores across the entire spectrum, especially for high sequence identity alignments (Pearson correlation coefficient between BLAST-based and MMseqs2-based HSSP scores = 0.95; Fig. 3).

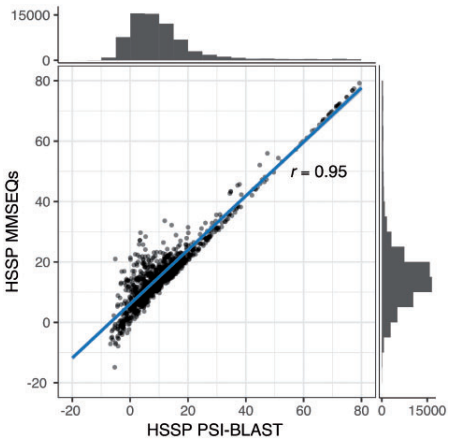


Fig. 3. *HSSP scores derived from MMSeqs2 and PSI-BLAST alignments strongly correlate.* HSSP scores derived from PSI-BLAST alignments (x-axis) and MMSeqs2 (y-axis), respectively. The histograms display the number of protein pairs in the respective ranges of HSSP scores. HSSP scores for both methods highly correlate (Pearson correlation coefficient=0.95)

3.2 HFSP precisely identifies the third, but not fourth, level of EC annotations

In identifying pairs of proteins sharing the same function at the fourth level of EC (Materials and Methods), HFSP attained precision of 44.1% ± 3.6 at HFSP 0 and recall of 71.5% ± 1.6 (in cross-validation). This disappointing performance suggests that the increasing resolution/fine-tuning of experimental molecular function annotation is prohibitive for large-scale computational analyses of proteins; i.e. for any given alignment scoring HFSP ≥ 0, it is more likely that the proteins in the alignment are not functionally identical.

In exploring this problem, we found that many highly sequence similar protein pairs of different EC annotations contained homologous proteins that were assigned slightly different functionality in different organisms. For example, proteins from the squalene cyclase family (Interpro: IPR018333, Pfam: PF13243 and PF13249) were annotated with different ECs; e.g. GERS_RHISY, a germanicol synthase in the red mangrove, is assigned EC: 5.4.99.34 and has 93% sequence identity (alignment length = 758) to BAS_BRUGY, a Beta-amyrin synthase of the Burma mangrove, which is annotated as EC: 5.4.99.39. This combination of sequence identity and alignment length produces an HFSP score of 64.6. At this HFSP level protein pairs are predicted to share the same EC annotation at fourth EC level with a precision of >99%. Note that GERS_RHISY is the only EC 5.4.99.34 protein to date. The publication describing its catalytic activity (Basyuni *et al.*, 2007), suggests that GERS_RHISY activity warrants a brand new EC number (germanicol synthase), because it primarily catalyzes germanicol synthesis. From our perspective, GERS_RHISY should additionally carry the beta-amyrin synthase annotation, since beta-amyrin (and lupeol) are synthesized in addition to germanicol albeit at a lower rate. Note that this example also recalls the problem of moonlighting proteins.

The above example reflects the general problem of unbalanced annotation detail of different EC categories at the same level of annotation. For example, EC: 5.4.99.- is by choice of the EC meant to temporarily 'house' a collection of enzyme reactions that have yet to be more thoroughly categorized. Many members of EC: 5.4.99.- fall into the same PFAM families, while catalyzing the conversion of the same reactant into similar chemical compounds; i.e. the fourth level EC annotations of these proteins convey only a small amount of functional difference. However, 5.4.99.- also contains significantly different proteins catalyzing different reactions, where fourth level annotations convey very large differences. Note that in this scheme, automated protein function annotation is significantly limited by lack of awareness of what individual EC numbers represent; i.e. it is incorrect to assume that the fourth, most precise, level EC annotations, across the entire EC system, are similarly defined in terms of depth of functional understanding and/or functional distances between proteins of the same third level EC. Note, however, that increasing the HFSP threshold for calling protein functions identical leads to significantly improved precision (if at significant cost to recall). For example, at HFSP cutoff = 20, 93% of the protein pairs are correctly annotated to share functionality. In other words, protein pairs with higher HFSP score represent more reliable predictions. This improvement is unsurprising as it is due in large part to increasing sequence identity and is very likely reflective of closer evolutionary relationships between proteins.

In identifying pairs of proteins sharing the same function at the third level of EC, we found that performance improved drastically at the default HFSP cutoff = 0. Here, our method attained precision of 96% ± 1.2 at HFSP 0 and recall of 64% ± 1.6 (in cross-validation, Fig. 1). These results suggest that in the absence of additional knowledge about an aligned protein pair, it is prudent to only accept higher scoring HFSP alignments (for fourth digit annotations) or to move up in the required resolution of functional annotation (i.e. to third EC level).

Finally, we tested HFSP precision and recall on proteins in *Swiss-Prot 2017* that were NOT in *Swiss-Prot 2002* (which was used for training of the HFSP curve), i.e. proteins that were added to Swiss-Prot after January 2002. We found that performance for this subset was similar to the expected performance at both the third and fourth EC levels (Fig. 4), suggesting that our measure remains applicable for newly added proteins AND enzyme classes (EC numbers).

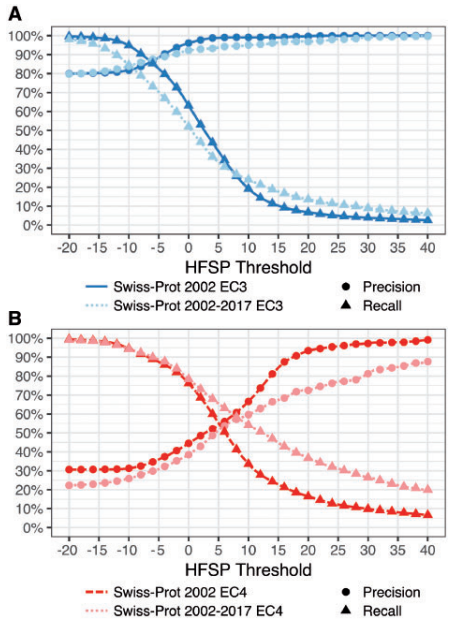


Fig. 4. *HFSP performs as expected on newly added proteins.* Precision and recall of function prediction at (**A**) third (dark blue) EC level of proteins in Swiss-Prot 2002 and of those added since 2002 (Swiss-Prot 2002–2017; light blue) are similar. However, for the fourth EC level, the (**B**) performance on newly added proteins (dark red) is worse than for older ones (light red)

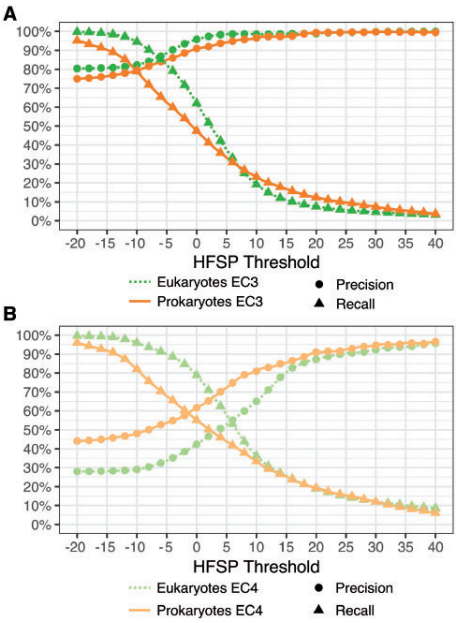


Fig. 5. *Differing annotation performance for prokaryotic and eukaryotic proteins at third and fourth EC level.* (**A**) For the third EC level at default cutoff of HFSP=0, eukaryotic protein pairs are assigned functional similarity correctly more often than prokaryotic ones. However, for high thresholds, i.e. higher precision at the expense of recovered protein pairs, performance is similar. (**B**) Performance is better for prokaryotes than eukaryotes at the fourth EC level

3.3 HFSP performance differs in annotating prokaryotic versus eukaryotic proteins

We additionally evaluated the HFSP performance in annotating the eukaryotic versus prokaryotic proteins of the entire *Swiss-Prot 2017 set* (Methods, Fig. 5A) at the third EC level. At our default cutoff of HFSP = 0, eukaryotic protein pairs were assigned functional similarity correctly more often than prokaryotic ones (precision/recall 96 ± 1.5/62% versus 91 ± 1.5/47%, respectively). Note that there were more eukaryotic proteins in our data than prokaryotic ones, which may have contributed to this disparity during HFSP curve optimization. This larger number of proteins can be explained by the eukaryotes (i) trending toward bigger proteomes and, perhaps more importantly, (ii) making up a bigger fraction of model organisms, which are better studied. Curiously, at the fourth EC level this trend was reversed, i.e. precision was better for prokaryotes than for eukaryotes (precision/recall 62/55% versus 42/79%, respectively, Fig. 5B). This observation may potentially be due to a smaller number of homology-confusing multi-domain proteins in prokaryotes. It may also reflect a lower enzymatic diversity of prokaryotic proteins in our set: 1522 distinct EC annotations in eukaryotes versus 1403 in prokaryotes. Whether this difference is due to actual diversity or a result of experimental bias remains unclear.

3.4 HFSP accurately predicts unknown protein function at all EC levels

There is a conceptual difference between annotating functionality of an unknown protein and measuring functional similarity of two proteins. That is, in assigning ONE specific protein function to a newly obtained amino acid sequence is not the same as relying on homology to identify proteins sharing the similar functionality in a particular database. To use HFSP as a method of function prediction we proposed simply relying on the 'highest hit'; we have previously shown that this approach is best for transferring functional annotations with HSSP (Zhu *et al.*, 2018) and suggest that similar logic should apply here.

By mapping the highest HFSP match (at cutoff = 0 and excluding self-hits) for the experimentally annotated proteins of the *Swiss-Prot 2017* set, we were able to correctly identify the fourth level EC function of 4668 (~83% of 5647) proteins. As expected, the numbers were higher for the third level EC (5425 of 5647 proteins, 96%). Note that this performance is the upper limit of actual HFSP performance, as *Swiss-Prot 2002*, on which our method was developed, is a subset of *Swiss-Prot 2017*. Also note that (i) 625 proteins in our *Swiss-Prot 2017* set did not reach our HFSP cutoff =0 and (ii) 563 proteins did not align to any others in our set. Of these, 645 proteins (291 and 354, respectively) proteins were unique in our set; i.e. there was no other protein with the same EC number at fourth EC level. Thus, 1188 proteins in our set (~17% of 6835 in the set) could not be assigned function at all—~8% due to HFSP limitations and ~9% due to the absence of homologs.

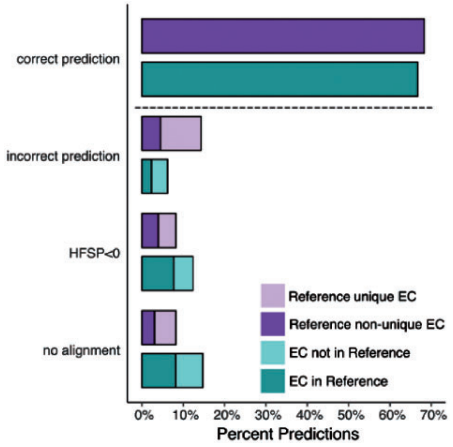


Fig. 6. *More proteins in Swiss-Prot enzyme could not be assigned to function than expected.* Function predictions for proteins in Swiss-Prot 2017 with unique (light purple) and non-unique (dark purple) fourth level EC annotation and all proteins in Swiss-Prot with EC annotation complete on all four levels that either share an EC with proteins in Reference (light teal) or not (dark teal)

3.5 Functional annotations even in manually curated databases are often incorrect

We applied the highest HFSP hit measure to evaluate EC annotations in the entire Swiss-Prot set (Materials and Methods) on the basis of their alignment to our experimentally annotated *Swiss-prot 2017* set. We estimate that 142 831 of the 214 000 Swiss-Prot enzymes (67%) are correctly annotated at the fourth level of EC (Fig. 6). Curiously, 32 201 (15%) of the enzymes in Swiss-Prot had no corresponding fourth level ECs (381 third level ECs) in *Swiss-Prot 2017*, raising questions as to the accuracy of these annotations. Another 4937 are deemed wrongly annotated (highest hit at HFSP $\geq$ 0 has a different EC number). While these proteins may indeed be assigned wrong functionality, this may also be due to error in HFSP assignments at this level (17% false positives at this cutoff, as described above for the *Swiss-Prot 2017* experimentally-annotated set). A more interesting finding, however, is that 34 031 (19%) of the proteins in this set could not be annotated at all by HFSP, whether due to lack of alignments (17 519 proteins) or HFSP highest hits unable to reach the cutoff (16 512 proteins). These 19% of proteins that could not be annotated represent a more than 2-fold higher number than expected (~8% as described above for the *Swiss-Prot 2017* set). We, thus, suggest that the Swiss-Prot EC annotations of many of these 34 031 proteins, a sixth of the total number of annotations, are incorrect.

3.6 Identifying proteins of new functionality is simplified with HFSP

One problem of function transfer by homology methods is their inability to identify proteins of completely novel, i.e. not found in the reference database, functionality. Note that sequence similar proteins are also likely functionally *similar*, but are clearly not necessarily functionally *identical*. To evaluate how HFSP deals with proteins of novel functionality, we extracted a set of proteins from *Swiss-Prot 2017*, where no other protein in our set had the same fourth

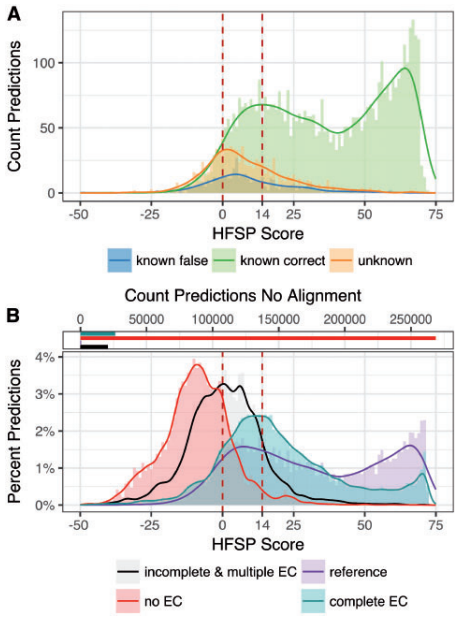


Fig. 7. *HFSP is robust to previously unseen enzymatic functionality.* (**A**) Proteins with no known homologs—approximated by investigating experimentally annotated proteins which fall into a EC category unique to the protein (orange)—show on average smaller highest scoring HFSP hits than proteins with existing homologs (green—correct predictions, blue—incorrect predictions). Of all predictions at HFSP score $\geq$14, <10% of proteins with 'unknown' and 'known' but falsely predicted function where observed (**B**, bottom panel): highest HFSP score predictions for different protein subsets of the non-reduced Swiss-Prot: (i) experimentally verified enzymes (reference—purple), (ii) enzymes with EC annotation complete on all four levels that are not experimentally verified (complete EC—teal), (iii) enzymes with incomplete EC annotation or multiple EC annotations (incomplete & multiple EC—black) and (iv) proteins that are not annotated as enzymes (no EC—red); note that for most proteins with no EC annotation there were no matched to the reference database (268 857 proteins, 91%; **B**, top panel)

EC level annotation ('unknown' functionality). These 'unknown' proteins, i.e. assigned to a fourth EC level category, which appear just once in our set, are a minority (19%; 1317 of 6835 proteins), albeit a significant one. We asked if we could in advance identify these 'unknown' proteins, for which prediction of function could not be made, rather than making incorrect predictions.

We separated function predictions for the 6835 proteins in Swiss-Prot 2017 into three subsets (i) 'unknown', as described above, and (ii) correctly and (iii) incorrectly predicted 'known', i.e. proteins with fourth EC level annotations containing more than one protein. We then compared the highest hit HFSP score distributions for all three sets (Fig. 7). HFSP scores for correctly annotated proteins with known functionality appear to come from a mixture of two distributions. These are likely to be evolutionarily distant (peak of the distribution at HFSP = ~20) versus close (peak at HFSP = ~65) homologs. The peak of the distribution of 'unknown' protein scores is obviously different from either (HFSP = ~2). However, the distribution of incorrect predictions for 'known'

proteins closely follows the 'unknowns' (Fig. 7A and Supplementary Fig. S2A and B). Combined, 'known incorrect' and 'unknown', make up less than 10% of all predictions at HFSP $\geq$ 14 (false discover rate, FDR = 9.6%), whereas between the default cutoff and HFSP = 14 ($0 \leq$ HFSP $<$ 14) this fraction is nearly 40%. Despite the fact, that at this threshold only ~6% of all predictions are of 'unknown' origin, these are still 30% of all 'unknown' proteins; similarly ~3% of all predictions, but 29% of all 'known incorrect' proteins are at HFSP $\geq$ 14. These observations suggest that while we cannot differentiate incorrect predictions from missing-reference ones, HFSP handles new protein function, as well as that which it has already seen, with higher scores indicating more reliable/correct annotations.

Given the vast number of proteins that yet have to be functionally annotated (e.g. TrEMBL is currently approaching 109 million proteins), the number of potential EC functionalities missing from our reference set, as well as the understanding that the total number of enzymes among the unannotated proteins may not mirror the Swiss-Prot distribution (where ~47% of all proteins are annotated enzymes including those with incomplete and multiple EC annotations), we suspect that accurately estimating the HFSP cutoff at which the FDR would fall below some threshold, e.g. 5% (currently at HFSP $\geq$ 28), is not possible. For example, given the current distribution of scores, 29% of 1384 'unknowns' and incorrect 'knowns' present at HFSP $\geq$ 14 make up only 407 proteins. If we were annotating tens of millions of proteins, however, this error rate can be expected to produce hundreds of thousands of annotations. On the other hand, given the limited size of our reference database, we cannot necessarily expect that the true positive findings would grow accordingly.

We further predicted EC annotation for all Swiss-Prot (555 594 proteins in October 2017, Fig. 7B). Importantly, the majority (91%) of the non-enzymes (no EC annotations; 293 058 proteins) did not generate any matches to our reference database. Of the remaining non-enzymes, 21% (4987 proteins) scored at HFSP $\geq$ 0, making up 3% of all predictions (false positives, 1% for all predictions at HFSP $\geq$ 14). Predictions could be made for 57% of the enzymes with multiple or incomplete EC annotations (27 717 of 48 536 proteins); 53% (14 668 proteins) of these scored at HFSP $\geq$ 0 and 13% above HFSP $\geq$ 14 (3653 proteins). If these proteins were like our 'unknowns', we would expect at least twice as many with a match at HFSP $\geq$ 14. Thus, we suspect, that the enzymes in this set are not especially novel and can likely be annotated using HFSP and our reference dataset. This further suggests that at least 73% (43% no hits and 30% below HFSP = 0) of proteins with incomplete or multiple EC annotations could be proteins with no homologous sequence in our reference database.

In light of our findings, we note that without further experimental work to elaborate on the functions of the yet-unannotated proteins, even the best function prediction methods will soon reach their limits. We suggest that using HFSP cutoffs can help in both more accurately annotating protein function and, arguably even more importantly, in identifying new frontiers of molecular function exploration.

4 Conclusion

While experimental function annotation of proteins is more accurate, computational methods are more readily available for the vast amount of sequences currently in our databases. Here we demonstrated that our newly developed HFSP is a fast an accurate method applicable to this task. Applying HFSP to evaluate existing annotations we also highlighted inconsistencies in existing annotations of enzymatic activity reported in Swiss-Prot. We thus suggest that HFSP provides both a way to (i) enrich functional annotation analysis on a large scale, as well as to (ii) narrow down the space of proteins of interest for further experimental analysis.

Acknowledgements

We thank Drs Chengsheng Zhu and Max Miller (both Rutgers), for all help with interpreting our data and providing feedback on the manuscript. We also want to extend our deepest gratitude to Dr Peter Kahn (Rutgers) for his valuable feedback and help in finalizing the manuscript. Big thanks to Dr Predrag Redivojac, Jiang Yuxiang (both Indiana U) and Yanran Wang (Rutgers) for valuable discussions. Last but not least, we thank all those who deposit their experimental data in public databases and those who maintain these databases.

Funding

Y.B. was supported by the NSF CAREER Award 1553289, NIH U01 GM115486, and USDA-NIFA 1015: 0228906; Y.B. and Y.M. by the TU Munich (TUM)—Institute for Advanced Study Hans Fischer Fellowship (TUM-IAS), funded by the German Excellence Initiative and the EU Seventh Framework Programme, grant agreement 291763.

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLASsT: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.

Basyuni,M. *et al.* (2007) Triterpene synthases from the Okinawan mangrove tribe, Rhizophoraceae. *Febs. J.*, **274**, 5028–5042.

Benson,D.A. *et al.* (2013) GenBank. *Nucleic Acids Res.*, **41**, D36–D42.

Berman,H.M. *et al.* (2002) The protein data bank. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 899–907.

Clark,W.T. and Radivojac,P. (2011) Analysis of protein function and its prediction from amino acid sequence. *Proteins*, **79**, 2086–2096.

Daniels,N.M. *et al.* (2013) Compressive genomics for protein databases. *Bioinformatics*, **29**, i283–i290.

Frith,M.C. *et al.* (2008) The whole alignment and nothing but the alignment: the problem of spurious alignment flanks. *Nucleic Acids Res.*, **36**, 5863–5871.

Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Jiang,Y. *et al.* (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol.*, **17**, 184.

Khan,I. *et al.* (2014) Genome-scale identification and characterization of moonlighting proteins. *Biol. Direct*, **9**, 30.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Loewenstein,Y. *et al.* (2009) Protein function annotation by homology-based inference. *Genome Biol.*, **10**, 207.

Nelder,J.A. and Mead,R. (1965) A simplex method for function minimization. *Comput. J.*, **7**, 308–313.

Placzek,S. *et al.* (2017) BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res.*, **45**, D380–D388.

Radivojac,P. *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nat Methods*, **10**, 221–227.

Remmert,M. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.

Rost,B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.*, **318**, 595–608.

Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.

Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Schnoes,A.M. *et al.* (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.*, **5**, e1000605.

Steinegger,M. and Soding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.

The UniProt, C. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.

Wei,Q. and Dunbrack,R.L., Jr. (2013) The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLoS One*, **8**, e67863.

Zhu,C. *et al.* (2015) Functional basis of microorganism classification. *PLoS Comput. Biol.*, **11**, e1004472.

Zhu,C. *et al.* (2018) fusionDB: assessing microbial diversity and environmental preferences via functional similarity networks. *Nucleic Acids Res.*, **46**, D535–D541.

Supporting online material
for:

HFSP: High speed homology-driven function annotation of proteins

Mahlich, Y.[1,2,3*], Steinegger, M.[2,4,5], Rost, B.[2,3,6,7,8], Bromberg, Y.[1,3,9*]

1 Department of Biochemistry and Microbiology, Rutgers University, 76 Lipman Dr, New Brunswick, NJ 08873, USA
2 Computational Biology & Bioinformatics - i12 Informatics, Technical University of Munich (TUM) Boltzmannstrasse 3 85748 Garching/Munich Germany
3 Institute for Advanced Study, Technische Universität München, Lichtenbergstrasse 2 a, D-85748 Garching, Germany
4 Quantitative and Computational Biology group, Max-Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany
5 Department of Chemistry, Seoul National University, Seoul, Korea
6 TUM School of Life Sciences Weihenstephan (WZW), Freising, Germany
7 Columbia University, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, USA
8 New York Consortium on Membrane Protein Structure (NYCOMPS), New York, USA
9 Department of Genetics, Rutgers University, Human Genetics Institute, Life Sciences Building, 145 Bevier Road, Piscataway, NJ 08854, USA

Table of Contents for Supporting Online Material

Supplementary Data 1

Excel file with two sheets, each containing the UniProt ID, EC number and eukaryote/prokaryote mapping of proteins in Swiss-Prot 2002 and Swiss-Prot 2017 respectively

Supplementary Data 2

Excel file includes multiple sheets, each containing the results of individual predictions for proteins in datasets in the manuscript. Each sheet contains 5 columns, protein_reference, ec_reference, protein_prediction, ec_prediction & hfsp_score.

protein_reference & protein_prediction contain the Uniport IDs of the reference protein and the aligned protein, respectively. ec_reference and ec_prediction are the EC numbers of the reference protein and aligned protein, respectively. hfsp_score is the HFSP score for the alignment.

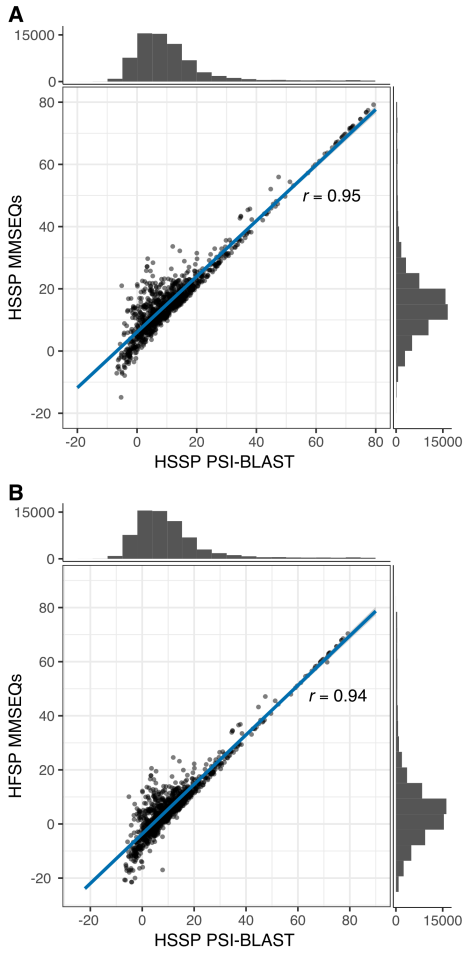
Table 1: 3rd level EC categories with over 50 proteins, sorted according to the number of proteins.

EC3	# Proteins
2.7.11	332
2.7.10	172
1.1.1	136
3.2.1	130
2.7.1	113
2.3.1	111
2.1.1	110
4.1.1	97
2.5.1	97
3.4.21	93
2.4.1	91
3.1.3	85
4.2.1	81
6.1.1	74
2.7.7	66
3.5.1	56
3.1.4	56
3.1.1	55
3.4.22	55

Table 2: F1-scores for each optimization run of HFSP-training.

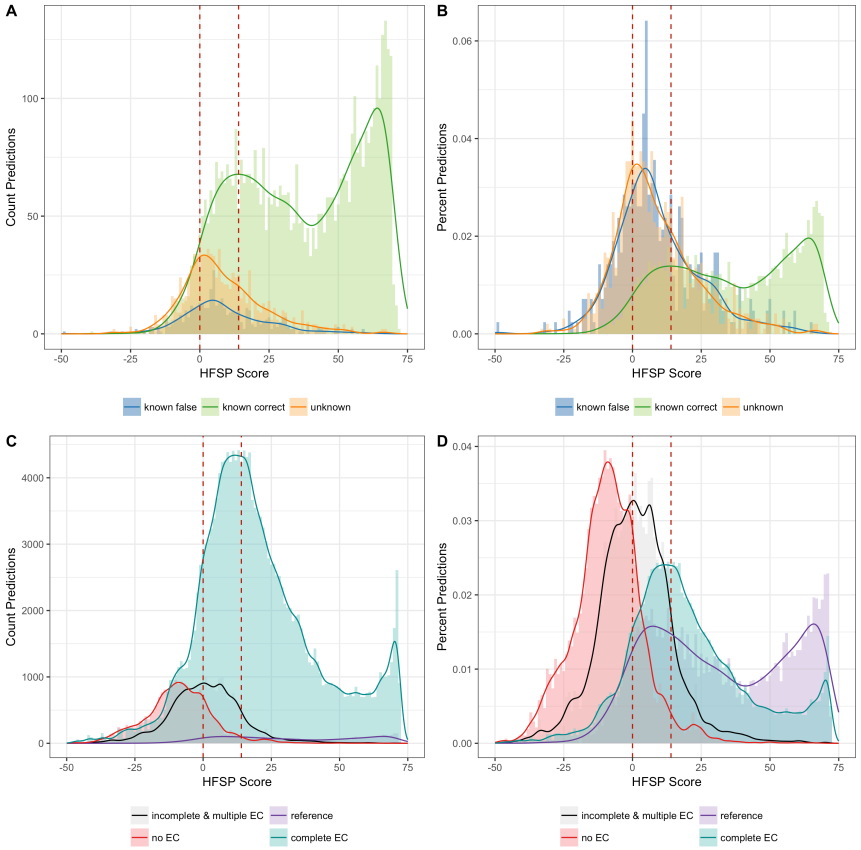
split	F1-score	Exponent	Factor
1	0.75	0.33	770
2	0.74	0.32	660
3	0.74	0.32	660
4	0.74	0.32	658
5	0.74	0.34	823
6	0.74	0.33	770
7	0.74	0.33	770
8	0.73	0.32	660
9	0.74	0.33	770
10	0.73	0.41	1646

Supplementary Figure 1



Supplementary Fig. 1: HSSP scores derived from MMSeqs2 and PSI-BLAST alignments strongly correlate. HSSP scores derived from PSI-BLAST alignments (x-axis) vs. **(A)** HSSP scores and **(B)** HFSP scores derived from MMSeqs2 (y-axis). The histograms display the number of protein pairs in the respective ranges of HSSP scores. In both scenarios HSSP/HFSP scores derived from MMSeqs2 highly correlate with HSSP scores from PSI-BLAST (Pearson-correlation coefficient = 0.95 / 0.94).

Supplementary Figure 2



Supplementary Fig. 2: Newly emerging enzyme functionality difficult to differentiate from incorrect function predictions. Proteins with no known homologs – approximated by experimentally annotated proteins, which have a unique EC number (orange) – show on average smaller highest scoring HFSP hits than proteins with homologs (green – correct predictions, blue – incorrect predictions). **(A/B)** Comparison of HFSP score distributions for highest scoring protein pair for Swiss-Prot 2017, **(A)** showing the distribution of raw counts and **(B)** the corresponding percentages of the respective datasets. **(C/D)**: Panels of counts and percentages as in (A/B), data is the Comparison of HFSP distributions for different subsets of the non-reduced Swiss-Prot: (i) experimentally verified enzymes (reference - purple), (ii) not experimentally verified enzymes with EC annotation complete on all 4 levels (complete EC - teal), (iii) enzymes with incomplete or multiple EC annotations (incomplete & multiple EC – black) and (iv) proteins that are not annotated as enzymes (no EC).

3 Clustering massive protein functional similarity networks: a scalable approach

3.1 Preface

In the previous chapter (HFSP: high speed homology-driven function annotation of proteins) I established how functional similarity between any to proteins can be determined. Ultimately the goal of *fusion* is to establish functional similarity between microbes. The simplest comparison that can be made between two microbes would be which molecular functionality they share between each other. If each microbe is represented by a set that lists which function is present (absent functions are omitted), the solution to this problem becomes trivial. Similarity simply can be represented by the intersection of both sets S_1 and S_2 divided by the larger of the two functional profiles (see Equation 3.1).

$$sim(S_1, S_2) = \frac{|S_1 \cap S_2|}{\max(|S_1|, |S_2|)} \tag{3.1}$$

The next step in generating the functional profiles for each microbe is to assign each of the microbes proteins to a functional unit. This can be done by following a simple three step process. (1) I establish functional similarities between all microbial proteins with the aid of HFSP [1]. (2) I generate a functional similarity network using the previously established protein similarities. (3) I cluster the network into functional units, where each protein is assigned to only one of those units.

This chapter describes all of these three steps in more detail, mostly focussing on the problems and challenges presented by step (3).

3.2 Introduction

The last five years has seen a rapid increase of completely sequenced and fully assembled bacterial genomes, deposited in publicly available database. A *complete assembly* is specified as a gapless assembly of all chromosomes. For bacteria this normally means one circular "primary" genome, with the addition of plasmids if existing in the bacterium. Technological leaps in sequencing technology have lead to a large uptick in completely assembled genomes, especially in the last 5 years. An increase in sequencing speed, as well as reliability has lead to a continually decreasing costs for whole genome sequencing. A dataset of microbes extracted approximately five years ago, used in a prove of concept of *fusion*[2] consisted of 1,374 completely assembled bacterial genomes. In contrast,

at the start of 2018 (Feb. 12^{th} 2018), the date at which the dataset for this analysis was extracted, close to 9,000 fully sequenced bacterial genomes were available in NCBI GenBank[3]. Even more so, within one year from January 2018 to January 2019, another ∼3,000 new complete assemblies of bacterial genomes where deposited in public databases, totaling ∼12,000 completely assembled bacterial genomes.

Curiously enough the majority of novel completely assembled bacterial genomes are not newly discovered or sequenced species but rather strains of species with known reference genome. One explanation for is that de-novo assemblies from scratch are still difficult to complete [4, 5]. Assembling reads into a genome by mapping them against a reference sequence, is vastly easier than de-novo assembly. Another observation reflecting the quick growth of available genomes, is that some recent sequencing project extend far beyond sequencing only one species or strain. Examples are studies on strains of *Bordetella pertussis* by Weigand MR et al.[6, 7, 8] or the NCTC 3000 project[9], a joint venture between the Public Health England's National Collection of Type Cultures, the Welcome Trust Sanger Institute and Pacific Biosciences. The studies by Weigand et al. are examples where strains of a specific species are targeted to get a better understanding of pathogenesis. To this date Weigand et al. have deposited 345 completely assembled strains of Bordetella pertussis. On the other hand NCTC 3000 is a ongoing project that aims to fully assemble 3,000 novel bacterial reference genomes counting close to 650 to this date. This includes both strains of previously sequenced as well as novel species.

Given the almost 10-fold increase of the completely assembled genomes within the last 5 years, (∼70% of that in the last two years alone), the problems I had to tackle are a direct result of this huge size of the dataset. This is mainly due to the computational requirements of the two key concepts of *fusion*: (a) Establishing functional similarities between all proteins present in the set of microbial proteomes. (b) Clustering the proteins into functional groups based on their shared functionality (see Figure 3.1)

In the proof of concept of *fusion* the functional similarities between proteins were established by calculating HSSP (Homology-derived Secondary Structure of Proteins)[10] scores for the protein pairs. In order to calculate the HSSP score a PSI-BLAST[11] alignment between the two proteins has to be generate first however. While the calculation of the HSSP score by itself does not significantly contribute to the necessary compute time, creating the alignments between all proteins with PSI-BLAST is a huge bottle neck. For the $\sim 1,400$ organisms used during the proof of concept this step required approximately three months. Even without an increase in dataset size, this is a step in the protocol that needed urgent improvement. I achieved this by introducing HFSP (Homology-derived Functional Similarity of Proteins)[1] (see **Chapter 2**) that utilizes MMSeqs2[12] instead of PSI-BLAST for the alignment generation.

The second key concept (clustering the proteins into functional units based on their shared functional similarity) unfortunately is far less straightforward to improve upon. The problem is that the clustering algorithm that was used initially (Markov Clustering[13]) despite performing very well on the limited size of the initial dataset, rapidly increases in its computational requirements with the number of proteins to cluster. Unfortunately this is a problem that is not unique to MCL, but observable with a wide range of other clustering algorithms. In the following sections I will briefly describe the dataset in more

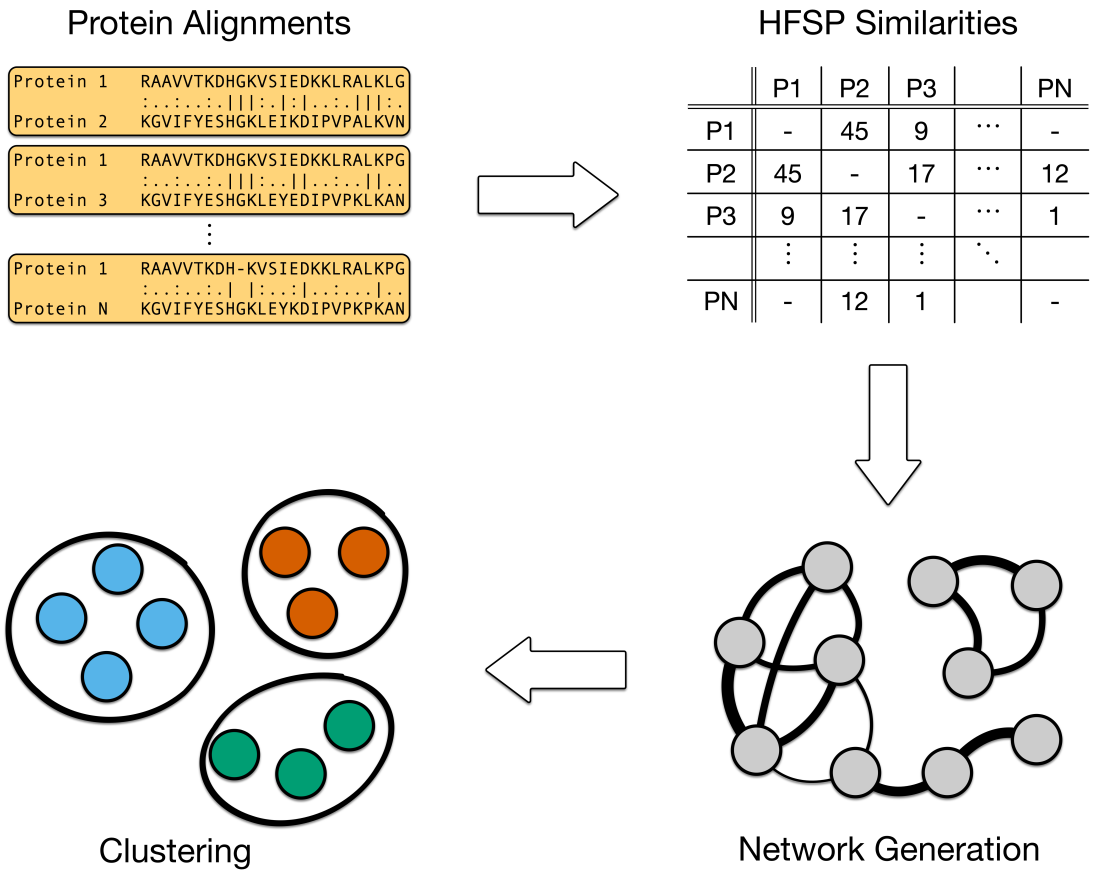


Figure 3.1: Schematic workflow of *fusion*. *fusion* consist of two major parts: (1) generation functional similarities between all bacterial proteins (top row) and (2) clustering the nodes in the resulting similarity network into functional units (bottom row). Once the functional units are established, functional similarities between organisms can be established by comparing their functional profiles, which are a collection of functional units associated with the individual microbes.

detail. The I explore the problems during clustering using MCL as a example. Finally I discuss how I plan to solve the clustering problem.

3.3 Data and Methods

3.3.1 Bacterial proteomes

The set was extracted from the GenBank[3] subdirectory (`/genomes/genbank/bacteria`[14]) of the NCBI ftp server on February 12th 2018. Two criteria have to be fulfilled for the bacterium to be included into the dataset: (1) The assembly of the bacterial genome has to be complete. As previously described a *complete assembly* is defined as a gapless assembly, of all chromosomes, i.e. the primary cyclic bacterial DNA plus all plasmids if present. (2) A proteome has to be predicted from the genomic sequence. Out of a total of 130,209 bacterial assemblies present at the date of extraction 8,906 bacteria fulfilled both criteria (9,329 only fulfilled criterion 1). Those 8,906 bacteria, account for 31,566,498 coding sequences.

Many bacteria contained in this set are strains of a common species. Strains of a species tend to share a lot of genetic similarity, resulting in a large redundancy in the dataset. Two redundancy reductions of this protein set were performed, one to 100% sequence identity the other one to 71.6%. Both reductions were done with the use of CD-HIT[15, 16]. By default, CD-HIT calculates the sequence identity over the length of the shorter of the two aligned sequence. Due to the greedy nature of CD-HIT's algorithm it is therefore possible that very short sequence pieces get aligned to long sequences with a reported high sequence identity. To make sure that I get to "true" 100% sequence identity redundancy reduction, i.e. a set of unique protein sequences, I forced CD-Hit to only add proteins to a cluster if the longer of the two sequences (the cluster representative) is covered completely by the alignment. This 100% sequence identity clustering cut the number of proteins roughly in half (49.5%) from 31,566,498 protein to 15,629,432. This set of proteins is used to establish functional similarities between all proteins (see subsection 3.3.2).

An additional second redundancy reduction to 71.6% sequence identity was performed in order to reduce the number of potential nodes in a consecutive protein similarity network clustering even further. 71.6% was specifically chosen to create an artificial upper boundary mirroring the lower boundary of 29.4% sequence identity that is at minimum required to create HFSP score of ≤ 0 if the alignment is longer than 450 residues long. Analogous to the 100% sequence identity reduction with CD-Hit, parameters were chosen in such a fashion to eliminate the chance that very short sequences are assigned to clusters with long sequences. The clustering ultimately resulted in a reduction from 31,566,498 proteins to 5,792,278 representative proteins, i.e. a reduction to about 1/5th (18.3%) of the original protein set.

converted reliability scores											
-20	0.44	0.44	0.44	0.45	0.45	0.45	0.46	0.46	0.47	0.47	-11
-10	0.48	0.49	0.50	0.51	0.53	0.54	0.55	0.56	0.58	0.60	-1
0	0.62	0.63	0.65	0.68	0.70	0.72	0.75	0.77	0.79	0.80	9
10	0.81	0.83	0.83	0.84	0.85	0.85	0.87	0.87	0.89	0.90	19
20	0.91	0.91	0.91	0.91	0.92	0.92	0.93	0.94	0.94	0.94	29
30	0.95	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.95	0.96	39
40	0.97	0.96	0.96	0.97	0.96	0.97	0.97	0.97	0.97	0.97	49
50	0.99	0.99	0.99	0.98	0.98	0.98	1.0	1.0	1.0	1.0	59
60	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	69

Table 3.1: HFSP to reliability score conversion table. Score conversion is based on predicted HFSP precision transferring EC annotations for prokaryotic enzymes. The first and last column lists the HFSP score range for which each row lists the converted reliability score.

3.3.2 Generating the functional similarities

I obtained functional similarities between proteins using HFSP (see **Chapter 2**). HFSP[1] requires a pairwise alignment of two proteins. Using the ungapped alignment length (number of positions in the alignment with aligned amino acids in both sequences), and the sequence identity (percentage of aligned position with the same amino acid in both sequences) an HFSP score can be calculated. I generated the alignments (as per HFSP protocol) with MMSeqs2's[12] `search` module. Parameters chosen for the searches were `-e 1e-3` (reporting e-value cut-off set to (1×10^{-3}), `--e-profile 1e-10`, (e-value cut-off for sequence inclusion into PSSM generation set to (1×10^{-10}), `--num-iterations 3` (three iterations of homology search), `--max-seqs 100000` (a maximum of 100k results per query sequence are reported), `-s 5.7` (default sensitivity cut-off for pre-filtering step), `--diff 1000` (at least 1000 sequences go into the generation of the PSSM while redundancy is reduced to 90% sequence identity). Based on the resulting alignments HFSP scores were generated for each available protein pair.

3.3.3 Generating the similarity network

I generated protein similarity networks, using the previously established functional similarities. Nodes represent proteins, and directed edges between proteins are added, if the two protein's shared HFSP score exceeds a defined threshold. Each edge was assigned a weight determined by the HFSP score. Using the predictive precision for prokaryotic enzymes of HFSP, I established an HFSP to reliability score conversion (see Table 3.1). The predictive precisions were obtained during the performance evaluation of HFSP (see also **Chapter 2**).

The main network I created contains all $\sim$ 15.6 million representatives of the redundancy reduced protein set as nodes. Edges were introduced if the HFSP score between proteins is ≤ 0. This resulted in $\sim$ 50 billion edges in the network. Since nodes can

be representatives of many proteins the converted edge weight was multiplied by the number of proteins represented by the node.

Additionally to the main network I also generated an expanded network. The expanded network reintroduced all proteins previously removed during the alignment stage. This network results in $\sim$ 31 nodes connected by $\sim$ 150 billion edges. Edge weights here are simple conversions from HFSP score to reliability score, without an further scaling.

The third and final network I generated is based on the set of representatives of the 71% sequence identity reduction. Each of the 5,792,278 representatives is added as node. Edges weights were calculated by summing up all edge weights between cluster members of the any two representatives. This network contains approximately 4.5 billion edges.

3.4 Results & Discussion

3.4.1 HFSP significantly speeds up generation of protein functional similarities.

As described earlier improving *fusion*'s key concepts became a necessity due to the much larger size of the data set in comparison to the proof of concept. To put the growth into perspective: Using the initial dataset of $\sim$ 4.2 million proteins, results in approximately 18 trillion ($\sim 17.6 \times 10^{12}$) protein pairs that in theory have to be compared. PSI-BLAST[11] required approximately 1.3 million CPU-hours to compute all alignments ($\sim$ 3 months on a compute cluster with 78 compute nodes with 8 CPUs each, 624 CPUs in total). With the updated complete dataset of $\sim$ 31.6 million proteins this would amount for $\sim 9.98 \times 10^{14}$ pairs. Those close to 1 quadrillion pairs, represent a more than 50-fold increase over the initial dataset. Even using the protein set reduced to 100% sequence identity we would still end up with $\sim$ 243 trillion possible protein pairs. Assuming the same number of calculated alignments per CPU per hour (approximately 14 million), this would put us at a requirement of $\sim$ 17.4 million CPU-hours. This translates to a runtime of slightly more than 3 years on the same system. Given that the required run times for PSI-BLAST searches scale linearly with database growth we have to expect an even longer compute time per protein alignment and therefore an overall longer run time. Additionally, only within the year of 2018, the set of coding sequences predicted from complete assemblies has increased by an additional 50% from $\sim$ 31.6 million to $\sim$ 45.1 million. While it can be expected that a large fraction of those coding sequences is redundant at 100% sequence identity, the underlying problem remains the same: Even with larger compute clusters being readily available, generating the alignments with PSI-BLAST is clearly not sustainable.

Switching the alignment algorithm from PSI-BLAST to MMSeqs2[12] in HFSP significantly reduced the runtime required. The total runtime of MMSeqs2 to calculate alignments between all possible protein pairs required roughly four weeks. Note that I switched to a more modern and powerful compute cluster (12 available nodes, equipped with 28 cores and 128GB RAM each; 336 cores in total). On this infrastructure 4 weeks of runtime equals roughly 226,000 CPU hours. During the development of HFSP I estimated based on a much smaller dataset, that MMSeqs2 should give us at least a 40-fold

speed increase in calculating the alignments over PSI-BLAST. As stated in **Chapter 2** I was expecting the gain over PSI-BLAST to be even bigger for larger datasets, which seems to observable here. Summarizing I can say that replacing PSI-BLAST with MMSeqs2 in the generation of functional similarities between proteins vastly improves the computational needs of *fusion*'s first key concept.

Out of the roughly 243 trillion possible protein pairs (2.43×10^{14}), approximately a 1000^{th} (249,624,795,920) generated alignments that passed the defined e-value cut-off. For those protein pairs, functional similarities were calculated as outlined in the section "Generating the functional similarities". Networks were created as described in section "Generating the functional similarities". This resulted in a network with $\sim$ 50 billion directed edges between the 15.6 million nodes. Due to the way HFSP is designed it is not guaranteed that the alignment from Protein P_A to Protein P_B results in the same score as P_B being aligned to P_A. In fringe cases this can result in one alignment falling above the decided HFSP threshold while the other does not. However, I did observe that the distribution for in and out degree of nodes is nearly identical in this network. This suggests if an edge from P_A to P_B exists the reverse is true as well for in the majority of cases. In other words it suggests that differences between the HFSP scores are only minimal, and those aforementioned fringe cases only occur very rarely.

3.4.2 Clustering very large networks still a challenge

During the conceptual phase of *fusion* MCL[13] (Markov Clustering) was used to group proteins into functional units based on their previously established functional similarity. It had previously shown that MCL generates biologically meaningful results when clustering protein similarity networks [17].

Markov Clustering revolves around two basic steps: (1) inflation of a transition probability matrix A by squaring and multiplying all values with a constant, and (2) pruning the columns of the resulting matrix. The inflation reinforces or lessens existing transition probabilities. The degree of the inflation can be determined by adjusting the constant the squared matrix is multiplied with. Imagine tightly connected nodes being pulled closer together by increasing the weights of the edges connecting the nodes, while loose connections are weakened. Iteratively repeating only step 1 should already lead to a network that stabilizes, i.e. transition probabilities only marginally change after inflation. To speed up this process (and aid in the decomposition of the network into clusters) the matrix is pruned after each inflation step, by only retaining the strongest connections, i.e. transition probabilities for each node.

While this approach of MCL's original implementation works very well for networks of up to a certain size the complexity quickly leads to exceeding computational resource requirements. In my specific case I evaluated a high performance compute cluster implementation of MCL called HipMCL[18] that already circumvents some of the short coming of the original implementation. Yet I ran into problems detailed in the following paragraphs.

The initial similarity network contains roughly 50 billion edges, i.e. transition probabilities in the stochastic matrix. By default the transition probability matrix only gets

pruned for the first time after the first iteration of inflation. We can assume a required 24 bytes per entry in the matrix: 8 bytes for the numerical value of the transition probability and 16 bytes (8 bytes each) for the i, j index describing the location of the probability in the matrix. Simply loading the matrix into system memory would therefore already require a minimum of roughly $50 \times 10^9 \times 24$bytes which equals to approximately 1.2TB.

Since (Hip)MCL needs to inflate and prune the matrix in each iteration, more than the minimal memory required to store the matrix is needed. HipMCL's minimal memory requirement is estimated to be $|V| \times R \times 4 \times 24$ bytes, where $|V|$ is the number of vertices in the network and R is the maximal number of recoverable edges. R is one of the parameters, that determines how much the transition matrix is pruned after each step of inflation. (Hip)MCL's pruning is defined by three-parameters: (1) a transition probability cut-off P, (2) a parameter defining the number of edges to be selected S, (3) a parameter defining the number of edges to be recoverable R. By default, MCL / Hip-MCL removes all transition probabilities below the cut-off P. If more edges than defined by S remain after removal, the column is pruned further to contain less or equal to S transition probabilities. If too many edges are discarded in this process, up to R transition probabilities are recovered. This serves two purposes: (a) It enables the considerably quick clustering heuristic of MCL and (b) it significantly reduces memory requirements, since RAM requirements no longer scale with the number of non-zero edges, but rather the number of nodes. By default the parameters are set to $P = 1/10000; S = 1100; R = 1400$. As previously described only $|V|$ and R directly influence the memory requirement. In me case this would lead an estimated requirement of $15 \times 10^6 \times 1400 \times 96$ bytes or approximately 2TB. Remember that this network already has been reduced to contain only $\sim$ 15.6 million nodes from the initial $\sim$ 31 million nodes. The full network would require about twice as much system memory.

While 4TB of shared system memory can easily be attained in modern compute clusters using a message passing interface, the default parameters of HipMCL introduce another problem. A R of 1400 restricts the pruning in a way that at most 1400 edges per node can be retained. In very sparse networks with a wide distribution of edge weights per node this might not immediately pose a problem. However under certain network architectures this can lead to faulty clusterings. As soon as nodes exists in the network with more than $1/P$ edges and more than R edges fulfill the pruning cut-off with similar transition probabilities every edge pruned will lead to a loss of information. An example for this would be a hub or a fully connected subgraph within the network where each protein shares high similarity with each other. Given that my dataset is based on $\sim$ 9000 organisms a presence of house keeping genes existing in almost all organisms is a guarantee. Those house keeping genes are coding for highly similar proteins, that will result in the behavior described above. For hubs like this a finer granularity of the clustering than desired will likely be the outcome. Ultimately it is likely that this will lead to functional clusters showing differing degrees of granularity, depending on how many proteins are represented by said cluster. Especially in my case where the aim is to cluster proteins into functional units this is a highly unfavorable effect. Increasing the number of edges retained after pruning might alleviate the problem. However, doing so directly correlates with the memory requirements of MCL and leads to an increase.

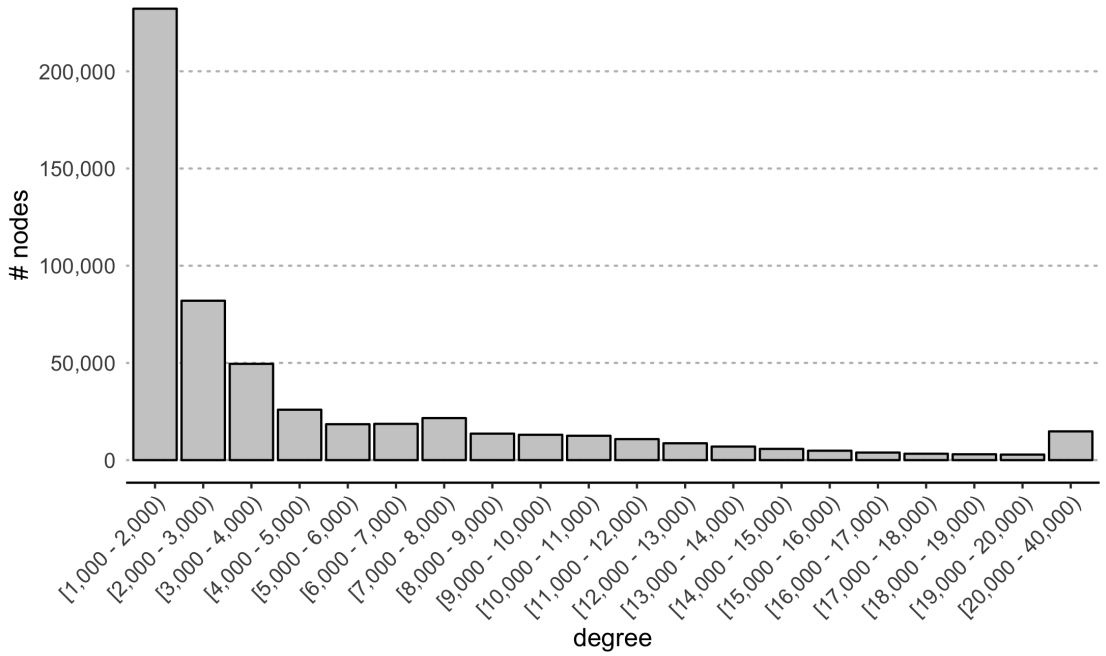


Figure 3.2: Substantial number of nodes with $\gg 1000$ outgoing edges in largest connected component. Displayed is the out-degree distribution for nodes in the larges strongly connected component, of the protein similarity network at 71.6% sequence redundancy reduction. By far the larges fraction (~ 3.1 million) of nodes have < 1000 out going edges (not displayed in the distribution for better visualization). Yet more > 700.000 than nodes which amounts for $\sim 20\% > 1000$ outgoing edges.

The high memory requirements, in conjunction with potential miss clustering resulting from the topology of the input network, present huge challenges to MCL. In an effort to circumvent those issues, I decided further reduce the network. I used the representative proteins of the 71% sequence identity reduction to generate the network. The resulting network contained ~ 6 million nodes and ~ 4.5 billion edges. Despite this reduction of size, I was still able to observe network topology as previously described. An investigation of the largest connected component in the network revealed that while the network is very sparse overall, a substantial number of nodes with $\gg 1000$ outgoing edges still exists (see Figure 3.2).

3.4.3 t-SNE able to overcome MCL's shortcomings

t-distributed stochastic neighbor embedding[19] or t-SNE technically is a visualization approach for high dimensional data. Part of t-SNE, akin to a principal component analysis PCA or multidimensional scaling MDS is to perform a dimensionality reduction of the dataset, most often into two dimensions. The assumption is that similarity information is retained in the dimensionality reduction, and individual nodes cluster together

more visibly. The advantage of t-SNE is that once an initial embedding is generated, this embedding can be used to re-embed new nodes, or nodes removed from the initial embedding set. In theory I should be able to chose a representative subset of nodes from the original network, to generate a initial "training" embedding, and then re-embed the remaining nodes into the previously generated embedding. While the initial embedding step might still be computationally expensive, re-embedding of further nodes, can be parallelized and is considerably less expensive. After embedding all nodes, the final clustering can be extracted. This methodology would potentially also open up a way for easy small scale updates to the clustering. Functional similarities for coding sequences from newly sequenced organisms to existing proteins can be generated and used to directly incorporate them into the existing cluster. One caveat is that new nodes would increase dimensionality if added into the original data set. It can be assumed that a new t-SNE embedding would have to regenerated from scratch after a certain amount of data that has been added in retroactively. The point at which this is the case still has to be evaluated.

3.5 Conclusion

As described during this chapter, the significant increase in raw data resulting from new microbial organisms presents an increasingly difficult problem to solve. Especially with more wide spread availability and increasing accuracy of single cell sequencing this amount in data can only be expected to grow fast than the already exponential growth observed today. Additionally, availability of computational power is outpaced by the amount of data made available today already. Generally speaking, we are at a stage where we can observe a higher influx of novel strains of previously sequenced species, as opposed to never seen before species. Assuming the strains them self are mostly redundant in terms of their genetic makeup but the few coding sequences that are differentiating them are what make them unique, the added data on a yearly or monthly basis leading to potential *fusion*DB (see the next chapter: "*fusion*DB: assessing microbial diversity and environmental preferences via functional similarity networks") updates, should not change the functional clusters in a significant way.

In that light I think that t-SNE might be not only able to solve the problem of dealing with the massive amounts of data in the first place, but might also enable me to create smaller incremental updates, that incorporate a small number of novel genomes. I can also envision a way how t-SNE could be incorporated into respectively replace the algorithm that currently maps previously unseen proteomes to existing functional units in *fusion*DB. Currently a simple homology search against all available proteins in *fusion*DB is executed and the novel proteins are assigned the functional unit that contains the best matching HSFP hit. One major drawback of this approach is however that functional similarity for proteins within the novel proteome will not be considered in case they don't match anything that has previously been seen, i.e. functionality encoded by one or more proteins that is unique to this specific proteome. t-SNE could eventually aid in overcoming this drawback. Looking forward I also hope that t-SNE

will be able to discover near optimal representatives of the functional units, that can then be used to create a reference database of proteins, for applications like mi-faser[20]. Utilizing this capability would bring *fusion*DB and mi-faser closer together, enabling researchers to annotate metagenomic reads, with *fusion*DB's functional units. I see this as a first step towards discovering functional pathways within a selection of organisms, respectively metagenomic microbial community. With all those updates in mind, it still will be difficult and in fact get increasingly difficult to deal with the amount of data that is generated. There is an immense need for clustering algorithms that are capable to deal with huge networks, preferably without segmenting and breaking them down into smaller problems, as some information might potentially be lost in doing so.

3.6 References

[1] Y. Mahlich, M. Steinegger, B. Rost, and Y. Bromberg. Hfsp: high speed homology-driven function annotation of proteins. *Bioinformatics*, 34(13):i304–i312, 2018. doi:10.1093/bioinformatics/bty262.

[2] C. Zhu, T. O. Delmont, T. M. Vogel, and Y. Bromberg. Functional basis of microorganism classification. *PLoS Comput Biol*, 11(8):e1004472, 2015. doi:10.1371/journal.pcbi.1004472.

[3] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic Acids Res*, 41(Database issue):D36–42, 2013. doi:10.1093/nar/gks1195.

[4] S. L. Salzberg, A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, G. Marcais, M. Pop, and J. A. Yorke. Gage: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res*, 22(3):557–67, 2012. doi:10.1101/gr.131383.111.

[5] J.-i. Sohn and J.-W. Nam. The present and future of de novo whole-genome assembly. *Briefings in bioinformatics*, 19(1):23–40, 2016.

[6] M. R. Weigand, Y. Peng, V. Loparev, T. Johnson, P. Juieng, S. Gairola, R. Kumar, U. Shaligram, R. Gowrishankar, H. Moura, J. Rees, D. M. Schieltz, Y. Williamson, A. Woolfitt, J. Barr, M. L. Tondella, and M. M. Williams. Complete genome sequences of four bordetella pertussis vaccine reference strains from serum institute of india. *Genome Announc*, 4(6), 2016. doi:10.1128/genomeA.01404-16.

[7] M. R. Weigand, Y. Peng, V. Loparev, D. Batra, K. E. Bowden, M. Burroughs, P. K. Cassiday, J. K. Davis, T. Johnson, P. Juieng, K. Knipe, M. H. Mathis, A. M. Pruitt, L. Rowe, M. Sheth, M. L. Tondella, and M. M. Williams. The history of bordetella pertussis genome evolution includes structural rearrangement. *J Bacteriol*, 199(8), 2017. doi:10.1128/jb.00806-16.

[8] M. R. Weigand, L. C. Pawloski, Y. Peng, H. Ju, M. Burroughs, P. K. Cassiday, J. K. Davis, M. DuVall, T. Johnson, P. Juieng, K. Knipe, V. N. Loparev, M. H. Mathis, L. A. Rowe, M. Sheth, M. M. Williams, and M. L. Tondella. Screening and genomic characterization of filamentous hemagglutinin-deficient bordetella pertussis. *Infect Immun*, 86(4), 2018. doi:10.1128/iai.00869-17.

[9] P. H. England. Nctc 3000 project, 2019. URL: https://www.phe-culturecollections.org.uk/collections/nctc-3000-project.

[10] B. Rost. Enzyme function less conserved than anticipated. *J Mol Biol*, 318(2):595–608, 2002. doi:10.1016/s0022-2836(02)00016-5.

[11] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, 1997.

[12] M. Steinegger and J. Soding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*, 35(11):1026–1028, 2017. doi:10.1038/nbt.3988.

[13] S. M. Van Dongen. *Graph clustering by flow simulation*. PhD thesis, 2000.

[14] N. GenBank. Genbank ftp directory, 2019. URL: ftp://ftp.ncbi.nlm.nih.gov//genomes/genbank/bacteria.

[15] W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–9, 2006. doi:10.1093/bioinformatics/btl158.

[16] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–2, 2012. doi:10.1093/bioinformatics/bts565.

[17] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 30(7):1575–84, 2002.

[18] A. Azad, G. A. Pavlopoulos, C. A. Ouzounis, N. C. Kyrpides, and A. Buluc. Hipmcl: a high-performance parallel implementation of the markov clustering algorithm for large-scale networks. *Nucleic Acids Res*, 46(6):e33, 2018. doi:10.1093/nar/gkx1313.

[19] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[20] C. Zhu, M. Miller, S. Marpaka, P. Vaysberg, M. C. Ruhlemann, G. Wu, F. A. Heinsen, M. Tempel, L. Zhao, W. Lieb, A. Franke, and Y. Bromberg. Functional sequencing read annotation for high precision microbiome analysis. *Nucleic Acids Res*, 46(4):e23, 2018. doi:10.1093/nar/gkx1209.

4 fusionDB: assessing microbial diversity and environmental preferences via functional similarity networks

4.1 Preface

*fusion*DB is a novel database that enables the investigation for functional similarity between microbes and relating their environmental preferences. The database consists of the functional classification of 1,374 taxonomically distinct microbes derived generated using the *fusion* protocol, with added annotation of available metadata. This metadata augmentation enables highlighting and potentially uncovering functional relations of microbes sharing similar environmental niches. The investigation of functional niches can be further enhanced by using the build in visualization tool of *fusion*DB. The visualization displays a clustering of organisms according to their functional profiles, in conjunction with the functional niches that has be selected for visualization. I showed that organisms sharing the same environmental preferences tend to be functionally more similar to each other than microbes from different environmental niches. For example anaerobic bacteria share $\sim 31\%$ of their functional repertoire between each other, whereas they only share $\sim 24\%$ of their functional repertoire with aerobic microbes. I also demonstrated that *fusion*DB can be used to predict the functional profile of organisms not yet available in the database, and reliably place them into a neighborhood of functionally highly similar microbial organisms.

The database, the backend, and the functional profile prediction algorithm was implemented and evaluated by me. The web interface was implemented by me and Maximilian Miller. The analyses were conducted by me and Chengsheng Zhu. The manuscript was drafted by all authors.

4.2 Journal Article: Zhu, Mahlich, Miller, et al., Nuclear Acids Research 2017

Published online 3 November 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D535–D541***
doi: 10.1093/nar/gkx1060

*fusion*DB: assessing microbial diversity and environmental preferences via functional similarity networks

Chengsheng Zhu[1,†], Yannick Mahlich[1,2,3,4,*,†], Maximilian Miller[1,2,3,†] and Yana Bromberg[1,4,*]

[1]Department of Biochemistry and Microbiology, Rutgers University, 76 Lipman Dr, New Brunswick, NJ 08873, USA, [2]Computational Biology & Bioinformatics - i12 Informatics, Technical University of Munich (TUM), Boltzmannstrasse 3, 85748 Garching/Munich, Germany, [3]TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), Technical University of Munich (TUM), 85748 Garching/Munich, Germany and [4]Institute for Advanced Study, Technical University of Munich (TUM), Lichtenbergstrasse 2 a, 85748 Garching/Munich, Germany

Received August 11, 2017; Revised September 24, 2017; Editorial Decision October 12, 2017; Accepted October 22, 2017

ABSTRACT

Microbial functional diversification is driven by environmental factors, i.e. microorganisms inhabiting the same environmental niche tend to be more functionally similar than those from different environments. In some cases, even closely phylogenetically related microbes differ more across environments than across taxa. While microbial similarities are often reported in terms of taxonomic relationships, no existing databases directly link microbial functions to the environment. We previously developed a method for comparing microbial functional similarities on the basis of proteins translated from their sequenced genomes. Here, we describe *fusion*DB, a novel database that uses our functional data to represent 1374 taxonomically distinct bacteria annotated with available metadata: habitat/niche, preferred temperature, and oxygen use. Each microbe is encoded as a set of functions represented by its proteome and individual microbes are connected via common functions. Users can search *fusion*DB via combinations of organism names and metadata. Moreover, the web interface allows mapping new microbial genomes to the functional spectrum of reference bacteria, rendering interactive similarity networks that highlight shared functionality. *fusion*DB provides a fast means of comparing microbes, identifying potential horizontal gene transfer events, and highlighting key environment-specific functionality.

INTRODUCTION

Microorganisms are capable of carrying out much of molecular functionality relevant to a range of human interests, including health, industrial production, and bioremediation. Experimental study of these microbes to optimize their uses is expensive and time-consuming; e.g. as many as three hundred biochemical/physiological tests only reflect 5–20% of the bacterial functional potential (1). The recent drastic increase in the number of sequenced microbial genomes has facilitated access to microbial molecular functionality from the gene/protein sequence side, via databases like Pfam (2), COG (3), TIGRfam (4), RAST (5) and others. Note that the relatively low number of available experimental functional annotations limits the power of these databases in recognizing microbial proteins that provide novel functionality. Additional information about microbial environmental preferences can be found, e.g. in GOLD (6). While it is well known that environmental factors play an important role in microbial functionality (7), none of the existing resources directly link environmental data to microbial function.

We mapped bacterial proteins to molecular functions and studied the functional relationships between bacteria in the light of their chosen habitats. We previously developed *fusion* (8), an organism functional similarity network, which can be used to broadly summarize the environmental factors driving microbial functional diversification. Here, we describe *fusion*DB – a database relating bacterial *fusion* functional repertoires to the corresponding environmental niches. *fusion*DB is explorable via a web-interface by querying for combinations of organism names and environments. Users can also map new organism proteomes to the functional repertoires of the reference organisms in *fusion*DB; including, notably, matching proteins of yet unannotated function across organisms. The submitted organisms are vi-

*To whom correspondence should be addressed. Tel: +1 848 932 5638; Fax +1 848 932 8965; Email: ymahlich@bromberglab.org
Correspondence may also be addressed to Bromberg Yana. Tel: +1 646 220 3290; Email: yanab@rci.rutgers.edu
†These authors contributed equally to this work as first authors.

sualized, and can be further explored, interactively as *fusion* networks in the context of selected reference genomes. Additionally, the web interface generates *fusion*+ networks, *i.e.* views that explicitly indicate shared microbial functions.

Our overall analyses of the *fusion*DB data for the first time give quantitative support to the fact that environmental factors drive microbial functional diversification. To demonstrate *fusionDB* functionality for individual organisms, we mapped a recently sequenced genome of a freshwater *Synechococcus* bacterium to *fusion*DB. In line with our previous findings (8), we demonstrate that this microorganism is more functionally related to other fresh water Cyanobacteria than to the marine *Synechococcus*. In a case study on *Bacillus* microbes, we use *fusionDB* to track organism-unique functions and illustrate the detection of core-function repertoires that capture traces of environmentally driven horizontal gene transfer (HGT). *fusionDB* is a unique tool that provides an easy way of analysing the, often unannotated, molecular function spectrum of a given microbe. It further places this microbe into a context of other reference organisms and relates the identified microbial function to the preferred environmental conditions. Our approach allows for detection of microbial functional similarities, often mediated via horizontal gene transfer, that are difficult to recover via phylogenetic analysis. We note that, in the near future, *fusionDB* may also be useful for the analysis of functional potentials encoded in microbiome metagenomes. We expect that *fusionDB* will facilitate the study of environment-specific microbial molecular functionalities, leading to improved understanding of microbial lifestyles and to an increased number of applied bacterial uses.

METHODS

Database setup

*fusion*DB is based on alignments of 4 284 540 proteins from 1374 bacterial genomes (December 2011 NCBI GenBank (9). For each bacterium, we store its (a) NCBI taxonomic information (10) and, where available, (b) environmental metadata (temperature, oxygen requirements, and habitat; GOLD (6). The environments are generalized, e.g. *thermophiles* include hyper-thermophiles. '*No data*' is used to indicate missing annotations (Supplementary Online Material, SOM Table S1, SOM Figure S1). The general *fusion* (functional repertoire similarity-based organism network) protocol is described in our previous work (8). Briefly, all proteins in our database are aligned against each other using three iterations of PSI-BLAST (11) and the alignment length and sequence identity are used to compute Homology-derived Secondary Structure of Proteins (HSSP) distances (12). A network of protein similarities is then clustered using the Markov Clustering Algorithm (MCL) (13). For *fusion*DB the original *fusion* algorithm was modified to use less stringent protein functional similarity criteria (with HSSP distance cutoff = 10), which resulted in 457 576 functions (protein clusters; Table 1). Each bacterium was thus mapped to a set of functions, its functional repertoire (∼2400 functions on average, ranging from 118 to 6134 functions). Note that our functional repertoires include all the bacterial functions, regardless of annotation. We are thus able to make function predictions for proteins in new bacteria, even if these functions have not been annotated before.

Mapping new organisms to fusion

User submitted microbial proteomes and the associated functions are stored in a separate database (SOM Figure S2). For each query protein of the new organism, the mapping pipeline (SOM Figure S3, SOM Methods) (a) runs PSI-BLAST (reporting e-value 1e–10, inclusion e-value 1e–3, three iterations) against reference proteins in *fusion*DB and (b) maps the query to a *fusion* functional cluster, which contains the reference with the highest hit HSSP score. Note that novel proteins that cannot be assigned to existing functional groups (do not match any reference at HSSP distance $\geq$10) are reported as functional singletons even if they are similar among themselves. Additionally, protein alignments that exceed 12 CPU hours of run-time are currently eliminated from further consideration. In testing, we found that no >0.1% of the proteins fall into this category. Although long run-times usually indicate that query proteins likely align to many others in our database, they contribute only a small fraction to the overall bacterial similarity and are eliminated for the sake of a faster result turnaround. Note that we also evaluated a number of other algorithms for mapping organism functional repertoires, of which the above-described algorithm performed best (SOM Methods).

All functional cluster assignments of proteins in the query proteome are then combined into a functional repertoire where each functional cluster is unique; i.e. if two query proteins are assigned to the same functional cluster, this cluster is listed only once in the final repertoire.

Evaluating *fusion*DB performance

We evaluated the accuracy of functional mapping of new proteomes by iteratively mapping each of the *fusion*DB organisms back to the remaining 1373. We aligned each protein of the query organism to all proteins in other organisms and selected the alignment with highest HSSP score. We then assigned the query protein to the functional cluster of its match as described above for mapping new organisms.

The performance of this approach was evaluated on a per-function basis, i.e. for each function of each 'newly added' organism we retrieved counts of true positives (TP, proteins correctly assigned to this *fusion*DB function), false positives (FP, proteins falsely assigned to this *fusion*DB function), and false negatives (FN, proteins that are part of this *fusion*DB function in the reference database, but not correctly assigned). Note that reference singleton proteins that were not assigned to any *fusion*DB function were considered true positives. Averaged across all functions, the mean per-function precision and recall of correctly assigning proteins were 97.2% and 96.6%, respectively (3.1×10^{-8} mean per function false positive rate, FPR), while the overall precision of assigning any protein to a function was 98.2% (Eq. 1).

Individual organisms were assigned to their functional repertoires with 99.5% precision and 98.9% recall (Eq. 1,

Table 1. Annotation status of (HSSP-based) function groups

	Function groups (>1 sequence)	Function groups (1 sequence)	Total
Known (Kn)	54 522	15 738	70 260
Hypothetical (Hy)	85 252	89 282	174 534
Unknown (Un)	22 802	189 980	212 782
Total	162 576	295 000	457 576

SOM Figures S4 and S5). For this estimate we evaluated to overlap between reference and assigned repertoire; i.e. functional clusters that appear in both the reference and mapped functional repertoire are true positives. False positives are functional clusters in the mapped functional repertoire but not the reference repertoire, false negatives vice versa. The reported precision and recall are the mean precision and recall values averaged over all organism submissions.

$$precision = \frac{TP}{TP+FP},\ recall = \frac{TP}{TP+FN},\ FPR = \frac{FP}{FP+TN} \quad (1)$$

Web interface

*fusion*DB web interface has two functions: *explore* and *map new organisms*. The *explore* section contains access to all the 1374 bacteria and their metadata. Users can search these with (combinations of) organism names and environmental preferences by using text box input or built in filters. A user-selected organism set can be used to create a *fusion* network, in which organism nodes are connected by functional similarity edges. The *fusion* network can be viewed in an interactive display, as well as downloaded as network data files or static images. The user-defined color labels of the organism nodes reflect microbial taxonomy or environment. In the interactive display clicking an organism node reveals its taxonomic information and environmental preferences, while clicking an edge between two organisms yields a list of their shared functions. A *fusion*+ network can further be generated from the same list of organisms. There are two types of vertices (nodes) in *fusion*+: organism nodes and function nodes. Organism nodes are connected to each other only through the function nodes they share. The number of edges (degree) of an organism node represents the total number of functions of the organism; the relative position of each organism node is determined by the pull *toward* other organisms via common functions and *away* from others via unique functions (8). Like *fusion*, *fusion*+ can be interactively displayed, downloaded, and colored by the users' choices. For both network types, users can further retrieve the functions shared by the selected organisms—the core-functional repertoire of the set. Note that the primary function annotation of each functional cluster is the myRAST (5) description most commonly assigned to the cluster members. For each cluster we also include the corresponding Pfam (2) families. This feature is an efficient tool for investigating functions underlying organism diversification, particularly within different environment conditions.

In the *map* section, users can submit their own new organism proteomes (in fasta format) to our server (SOM Figure S3). The server sends out emails to users when mapping is finished. The *map* result page contains two tables containing (a) functional annotations, including the associated *fusion*DB reference sequences and proteins of the query organism that mapped to each functional cluster, as well as (b) similarity (Eq. 2) to the reference organisms in *fusion*DB, including functional repertoire size, functional overlap with the query, and metadata. Tables can be easily sorted, searched and exported as comma-separated files. The submitted proteome is further mapped to user-selected reference organisms with *fusion* and/or *fusion*+ as described above (Figure 1).

$$\text{similarity} = \frac{\text{shared functions}}{\text{the larger functional repertoire size}} \quad (2)$$

Analysis of environment-driven organism similarity

For each environmental condition in *fusionDB*, we sampled organism pairs where organisms were from (a) the same condition (SC, e.g. both mesophiles) and (b) different conditions (DC, e.g. thermophile versus mesophile). To alleviate the effects of data bias, the organisms in one pair were always selected from different taxonomic groups (different families). The smallest available set of pairs, SC-psychrophile contained 33 organisms from 17 families (SOM Table S1; 136 pairs—48 same phylum, 88 different phyla; due to high functional diversity of *Proteobacteria,* its classes were considered independent phyla). For all other environmental factors we sampled (bootstrap with 100 resamples) 136 organism pairs for both SC and DC sets, covering this same minimum taxonomic diversity. We calculated the pairwise functional similarity (Eq. 2) distributions and discarded organism pairs with <5% similarity.

RESULTS AND DISCUSSION

Mapping a new Synechococcus genome to fusionDB

We downloaded the full genome of *Synechococcus* sp. PCC 7502 (GCA_000317085.1) as translated protein sequence fasta (.faa file) from the NCBI Genbank (9) and submitted it to our web interface. This 3,318 protein fresh water Cyanobacteria is isolated from a Sphagnum (peat moss) bog (6). 86% (2,853) of the bacterial proteins mapped to 2208 *fusion*DB functions, while 462 (14%) were functional singletons; three proteins exceeded runtime and were excluded (Methods). The whole process from submission to results notification e-mail took under three and a half hours. The mapping indicates that *Synechococcus* sp. PCC 7502 is most functionally similar (56%) to *Synechocystis* PCC 6803, a fresh water organism evolutionarily closely related to *Synechococcus*. It also shares a high functional similarity with a mud *Synechococcus* (*S.sp.* PCC 7002; 53%) and with other fresh water *Synechococcus* (*S. elongatus* PCC 7942 and *S. elongatus* PCC 6301; 52%). Notably, but not surprisingly, *Synechococcus* sp. PCC 7502 shares much less functional

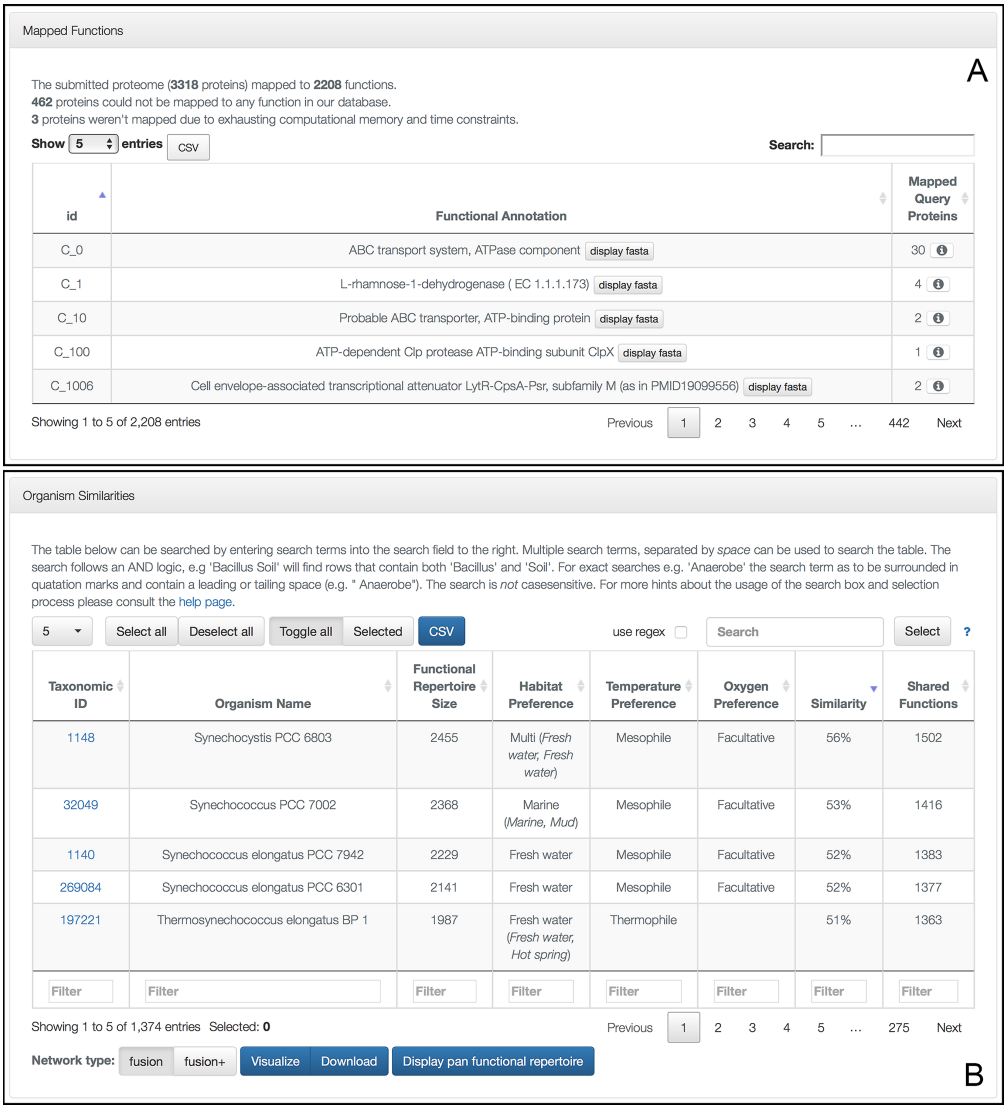


Figure 1. Screenshot of the organism mapping result page. (**A**) The 'Mapped Functions' table lists the functions that the submitted organism is mapped to. For each function, associated proteins from *fusion*DB and mapped query proteins can be displayed. (**B**) The 'Organism Similarities' table displays, all 1374 *fusion*DB organisms and their functional similarities to the query organism, including additional information such as environmental metadata; the view can be toggled between all and user-selected organisms. *fusion*(+) networks of the query and user-selected organisms can be created for on-site visualization (see Figure 2) or download.

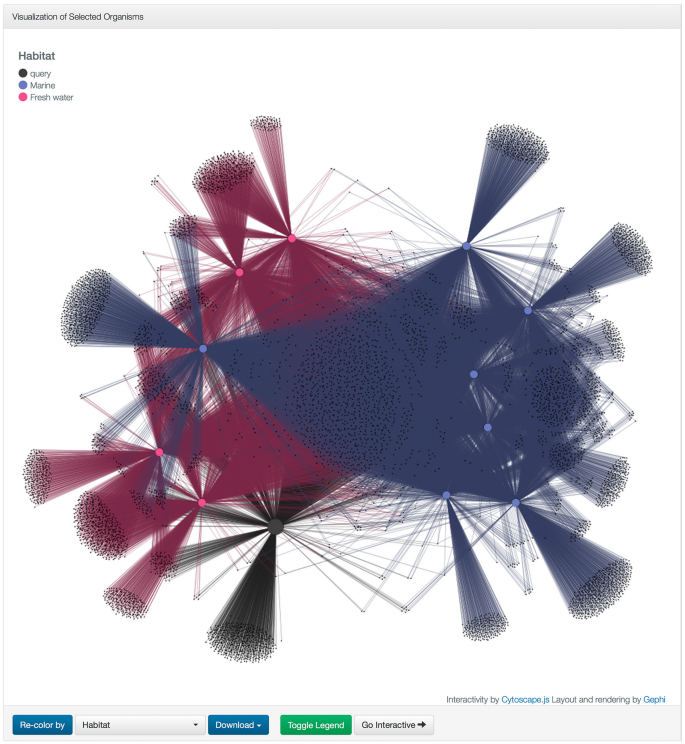


Figure 2. Screenshot of the fusion+ visualization of all Synechoccocus genomes. The submitted *Synechococcus* sp. PCC 7502 (query, black) clusters with the fresh water *Synechococcus* organisms (magenta). Note that *Synechococcus* sp. PCC 7002 – clustered among fresh water organisms; colored dark blue (marine) – is isolated from marine mud. It is salt tolerant but does not require salt for growth).

similarity (40–42%) with the marine *Synechococcus* bacteria. This relationship is clearly demonstrated by the *fusion*+ networks (Figure 2). There are 874 functions shared by all the twelve *Synechococcus* (SOM Data 1), the core-function repertoire for this genus, and 1128 functions shared among only the fresh water *Synechococcus* (SOM Data 2). These differential 254 functions (SOM Data 3) are likely important for living in fresh water, as opposed to marine, environment, e.g. low salinity and low osmotic pressure.

Environment significantly affects microbial function

In our evaluation of the effects of environmental pressures on microbial functionality we found that, in general, same environmental condition (SC) organisms across all environmental factors are more functionally similar than DC organisms (from different environments; Figure 3; with some exceptions mentioned below, Kolmogorov-Smirnov test (14) *P*-value < 2.5e–6). This finding is intuitive and many studies have demonstrated the presence of horizontal gene transfer (HGT) within environment-specific microbiomes (15–17). Our results, however, for the first time, quantify on a broad scale the environmental impact on microorganism function diversification.

SC-thermophile and SC-psychrophile pairs demonstrate significantly higher similarities when compared to DC pairs (Figure 3A). Notably, the higher functional similarity between thermophiles than between psychrophiles suggests that protein functional adaptation to low temperature may be less taxing than to high temperature – an interesting finding in itself. When contrasted with the extremophiles, mesophiles seem to have much larger functional diversity; in fact, SC-mesophile similarities are comparable to those of DC pairs (Figure 3A).

Different molecular pathways of aerobic-respiration and anaerobic-respiration/fermentation may explain the high level of dissimilarity between the aerobes and anaerobes (DC-anaerobe-aerobe; Figure 3B). Interestingly, the SC-anaerobe similarities are higher than the SC-aerobe similarities, likely because the more ancient anaerobic-respiration/fermentation machinery tends to be simpler (fewer reactions) (18) and more conserved.

D540 *Nucleic Acids Research, 2018, Vol. 46, Database issue*

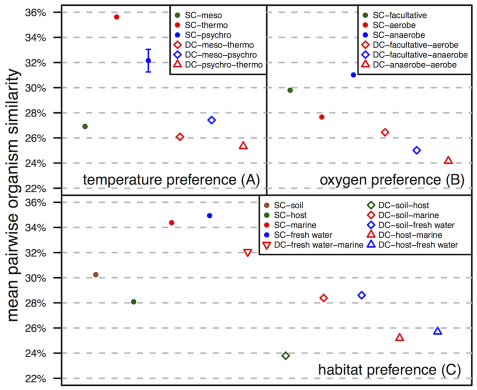


Figure 3. Organism pairwise similarity is higher among organisms living in the same environmental conditions. The mean pairwise similarity for same (SC) and different (DC) condition organisms according to (**A**) temperature, (**B**) oxygen and (**C**) habitat preferences. For all points without error bars, the standard errors are vanishingly small.

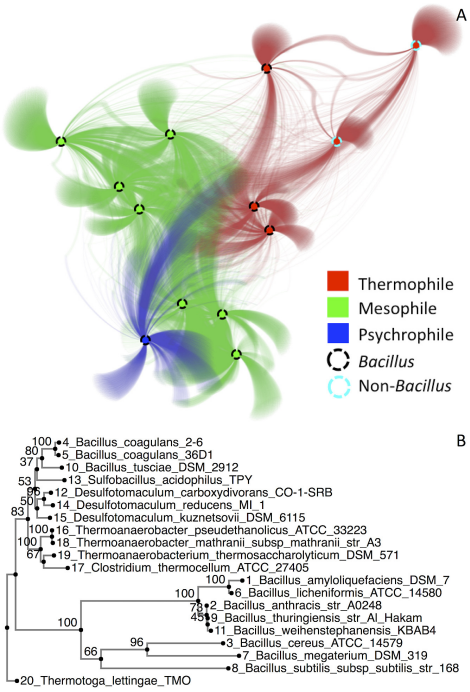


Figure 4. *fusion*DB reveals an HGT event between thermophilic *Bacilli* and thermophilic *Clostridia*. (**A**) *fusion*+ visualization of Bacillus and thermophilc Clostridia**.** Large organism nodes are connected via small function nodes. The two thermophilic *Clostridia* are connected to the thermophilic *Bacilli* via functions that are possibly horizontally transferred; (**B**) phylogenetic analysis of pyruvate, phosphate dikinase (PPDK) gene suggests HGT between thermophilic *Bacilli* and thermophilic *Clostridia.* The PPDK genes in thermophilic *Bacilli* are evolutionarily more related to those in thermophilic *Clostridia* than those in other *Bacilli.*

Different habitat (DC) samples show lower pairwise organism similarity than SC samples as well (Figure 3C). Interestingly fresh water and marine organism similarity (DC-fresh water-marine) is fairly high, likely due to overlaps in requirements of the aquatic conditions. Note however, that the dissimilarity across fresh water and marine conditions is still high enough to differentiate organisms of the same taxa (e.g. strains of *Synechococcus* in Figure 2). SC-host has the lowest mean organism similarity of the habitat SC samples; we speculate this to be a result of differential adaptations necessary to deal with diverse host defense mechanisms (19). The soil organisms also share low functional similarity, which is likely due to soil heterogeneity at physical, chemical, and biological levels, from nano- to landscape scale (20).

Case study of a temperature driven HGT event

Using the *fusion*DB *explore* functionality, we extracted thermophilic, mesophilic, and psychrophilic species representatives (one per species) of the *Bacillus* genus. We also added two other thermophilic *Clostridia*, *Desulfotomaculum carboxydivorans* CO-1-SRB and *Sulfobacillus acidophilus* TPY, to generate a *fusion*+ network (SOM Table S2; Figure S4A). As expected, note here that overall thermophilic bacteria are further removed from psychrophiles than from mesophiles. Moreover, the thermophilic *Bacilli* were more closely related to the non-*Bacillus* thermophiles than to other *Bacilli*. The three *Bacilli* thermophiles share 29 functions (SOM Data 4) that are not found in other *Bacilli* in this organism set, three of which also exist in the two thermophilc *Clostridia*. One is a likely pyruvate phosphate dikinase (PPDK) that, in extremophiles, works as a primary glycolysis enzyme (21). The thermophilic *Bacilli*'s PPDK proteins are more similar to those in thermophilic *Clostridia* (sequence identity $= 0.65 \pm 0.03$), than to those in mesophilic/psychrophilic *Bacilli* (sequence identity $= 0.17 \pm 0.05$). Phylogenetic analysis of the genes with additional thermophilic organisms (SOM Methods) suggests a likely HGT event between the thermophilic organisms (Figure 4B). The other two shared functions are carried out by proteins translated from mobile genetic elements (MGEs) that mediate the movement of DNA within genomes or between bacteria (22). Shared closely-related MGEs in distant organisms imply HGT (23). We thus suggest that *fusion*DB offers a fast and easy way to trace likely functionally necessary HGT events within niche-specific microbial communities.

In this work, we have highlighted the importance of environmental factors for microbial function, and demonstrated the capability of *fusion*DB to not only annotate functions, but also directly link function to environment. Although it was developed for mapping new microbial genomes, *fusion*DB also has the potential for microbiome

annotations. By mapping metagenome assemblies to *fusion*DB, both the functional and taxonomical annotations can be obtained. Moreover, our recent work (Zhu *et al.* 2017, Functional sequencing read annotation for high precision microbiome analysis, *submitted*) suggests that accurate functional annotations can also be obtained without assembly. We thus also expect to make *fusion*DB useful in this type of analyses in the near future.

CONCLUSIONS

*fusion*DB links microbial functional similarities and environmental preferences. Our analysis reveals environmental factors driving microbial functional diversification. By mapping new organisms to the reference functional space, our database offers a novel, fast, and simple way to detect core-function repertoires, unique functions, as well as traces of HGT. With more microbial genome sequencing and further manual curation of environmental metadata, we expect that *fusion*DB will become an integral part of microbial functional analysis protocols in the near future.

AVAILABILITY

*fusion*DB is publicly available at http://services.bromberglab.org/fusiondb/

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

We thank Drs Burkhard Rost (TU Munich), Max Haggblom, Tamar Barkay (both Rutgers), and Tom O. Delmont (U Chicago) for all help with interpreting our data and understanding the community needs. Big thanks to Yanran Wang and Dr Anton Molyboha (both Rutgers) for all discussions. We want to thank the anonymous reviewers for their thorough review and suggestions to improve this manuscript. We are also grateful to all those who deposit their data in public databases – fusionDB wouldn't be possible without them.

FUNDING

NSF CAREER Award [1553289 to Y.B. and C.Z.]; USDA-NIFA [1015:0228906]; TU Mu̇nchen – Institute for Advanced Study Hans Fischer Fellowship, funded by the German Excellence Initiative and the EU Seventh Framework Programme [291763 to Y.B. and Y.M.]; German Research Foundation (DFG) and the Technical University of Munich (TUM) in the framework of the Open Access Publishing Program.
Conflict of interest statement. None declared.

REFERENCES

1. Garrity,G.M., Boone,D.R. and Castenholz,R.W. (eds). (2001) *Bergey's Manual of Systematic Bacteriology*. 2nd edn. Springer, NY, Vol. **1**.
2. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
3. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
4. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
5. Aziz,R.K., Bartels,D., Best,A.A., DeJongh,M., Disz,T., Edwards,R.A., Formsma,K., Gerdes,S., Glass,E.M., Kubal,M. *et al.* (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, **9**, 75.
6. Pagani,I., Liolios,K., Jansson,J., Chen,I.M., Smirnova,T., Nosrat,B., Markowitz,V.M. and Kyrpides,N.C. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **40**, D571–D579.
7. Sun,W., Yu,G., Louie,T., Liu,T., Zhu,C., Xue,G. and Gao,P. (2015) From mesophilic to thermophilic digestion: the transitions of anaerobic bacterial, archaeal, and fungal community structures in sludge and manure samples. *Appl. Microbiol. Biotechnol.*, **99**, 10271–10282.
8. Zhu,C., Delmont,T.O., Vogel,T.M. and Bromberg,Y. (2015) Functional basis of microorganism classification. *PLoS Comput. Biol.*, **11**, e1004472.
9. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.
10. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
11. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
12. Rost,B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.*, **318**, 595–608.
13. Dongen,S.V. (2000) Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht.
14. Massey,F.J. (1951) The Kolmogorov-Smirnov test for goodness of fit. *J. Am. Statist. Assoc.*, **46**, 68–78.
15. Kim,S.E., Moon,J.S., Choi,W.S., Lee,S.H. and Kim,S.U. (2012) Monitoring of horizontal gene transfer from agricultural microorganisms to soil bacteria and analysis of microbial community in soils. *J. Microbiol. Biotechnol.*, **22**, 563–566.
16. Liu,L., Chen,X., Skogerbo,G., Zhang,P., Chen,R., He,S. and Huang,D.W. (2012) The human microbiome: a hot spot of microbial horizontal gene transfer. *Genomics*, **100**, 265–270.
17. Saye,D.J., Ogunseitan,O., Sayler,G.S. and Miller,R.V. (1987) Potential for transduction of plasmids in a natural freshwater environment: effect of plasmid donor concentration and a natural microbial community on transduction in Pseudomonas aeruginosa. *Appl. Environ. Microbiol.*, **53**, 987–995.
18. Raymond,J. and Segre,D. (2006) The effect of oxygen on biochemical networks and the evolution of complex life. *Science (New York, N.Y.)*, **311**, 1764–1767.
19. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
20. Lehmann,J., Solomon,D., Kinyangi,J., Dathe,L., Wirick,S. and Jacobsen,C. (2008) Spatial complexity of soil organic matter forms at nanometre scales. *Nat. Geosci.*, **1**, 238–242.
21. Chastain,C.J., Failing,C.J., Manandhar,L., Zimmerman,M.A., Lakner,M.M. and Nguyen,T.H.T. (2011) Functional evolution of C4 pyruvate,orthophosphate dikinase. *J. Exp. Bot.*, **62**, 3083–3091.
22. Frost,L.S., Leplae,R., Summers,A.O. and Toussaint,A. (2005) Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Micro*, **3**, 722–732.
23. Krupovic,M., Gonnet,M., Hania,W.B., Forterre,P. and Erauso,G. (2013) Insights into dynamics of mobile genetic elements in hyperthermophilic environments from five new Thermococcus plasmids. *PLoS ONE*, **8**, e49044.

*fusion*DB: assessing microbial diversity and environmental preferences via functional similarity networks

Zhu C[1+], Mahlich Y[1,2,3,4*+], Miller M[2,3+], Bromberg Y[1,4*]

[1]Department of Biochemistry and Microbiology, Rutgers University, 76 Lipman Dr, New Brunswick, NJ 08873, USA
[2]Computational Biology & Bioinformatics - i12 Informatics, Technical University of Munich (TUM) Boltzmannstrasse 3 85748 Garching/Munich Germany
[3]TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), Technische Universität München, 85748 Garching/Munich, Germany
[4]Institute for Advanced Study, Technische Universität München, Lichtenbergstrasse 2 a, D-85748 Garching, Germany.

* Corresponding authors: ymahlich@bromberglab.org, yanab@rci.rutgers.edu
Tel: +1.848.932.5638; Fax +1.848.932.8965
+ These authors contributed equally to this manuscript.

Supplementary Online Material

5. SOM Reference

Supplementary Online Methods

Pipeline performance analysis. Historically, we used PSI-BLAST for pairwise sequence alignments as per requirements of HSSP computations (Rost, 2002). Here we evaluated using BLASTP instead of PSI-BLAST to speed-up the mapping process.

The performance of both algorithms, in mapping any one of the 1,374 organisms to *fusion* back to the remaining 1,373, was evaluated using the precomputed BLAST and PSI-BLAST results from the original *fusion* (Zhu, et al., 2015). For each protein of the query organism, we extracted the highest HSSP scoring hit that was not part of the query organism's proteome. We then created the functional repertoires per organism as described above, including singletons as well.

Using this approach, the PSI-BLAST-based pipeline attained a mean precision of 99.5% and recall of 98.9% in finding fusion-identified organism repertoires (SOM Figure 4), while BLASTP was lower (precision: 98.7%, recall: 98.2%). Additionally, the upper and lower quantiles for both precision and recall were tighter around the mean for PSI-BLAST than for BLASTP. Thus, we use PSI-BLAST pipeline for all further mapping of individual organisms to *fusion*.

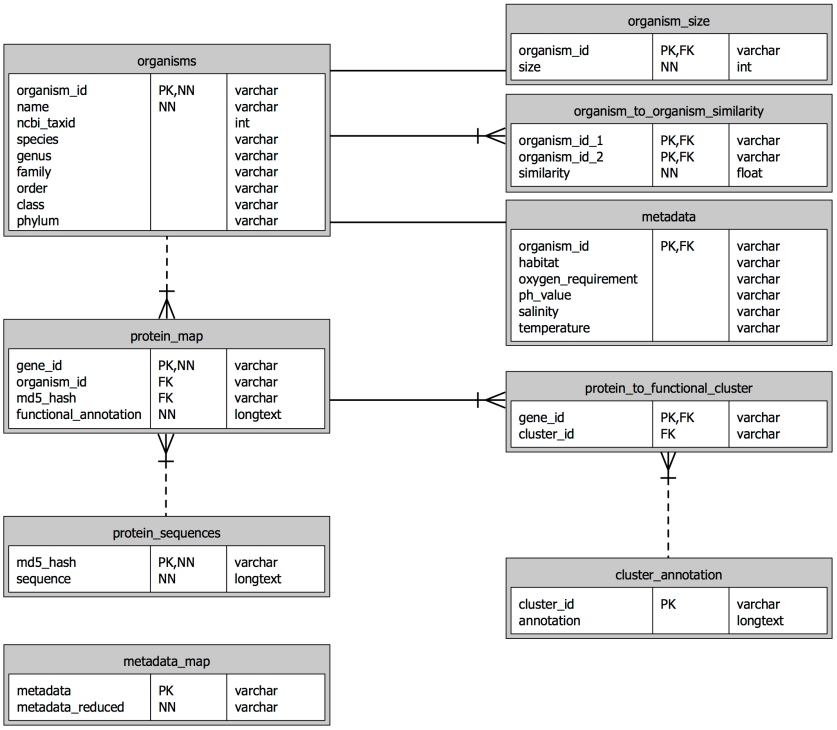
In order to avoid recomputing functional clusters with every new organism, we also tested three methods of assigning proteins to function clusters. (1) *majority vote:* we choose the functional cluster containing most (HSSP >= 10) matches to the query; (2) *weighted HSSP*: we computed the mean of all HSSP scores (≥ 10) of all query matches per functional cluster. The query protein is then assigned to the functional cluster with the highest mean; (3) *highest hit*: the query protein is assigned to the functional cluster containing the highest HSSP-score match.

When comparing precision and recall of recognizing functional repertoires for these three methods (SOM Figure 5), both majority vote and highest HSSP perform better than weighted HSSP. Majority vote and highest HSSP perform almost equally well, with the highest HSSP method being slightly better. The

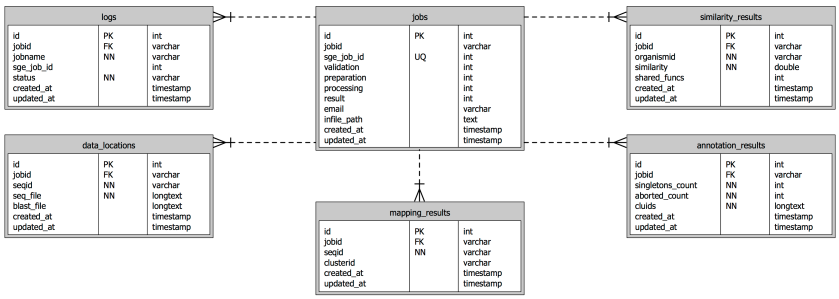
latter is also least computationally and complex, and was thus chosen for the final pipeline implementation.

Phylogenetic analysis. Homologs of pyruvate, phosphate dikinase (PPDK) were extracted from the selected organism genomes via BLASTP (best hit at E value cut-off of 1e-3). We computed the multiple sequence alignment and reconstructed the neighbor-joining tree with the online version of Mafft (Punta, et al., 2012). The phylogenetic tree was visualized with FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

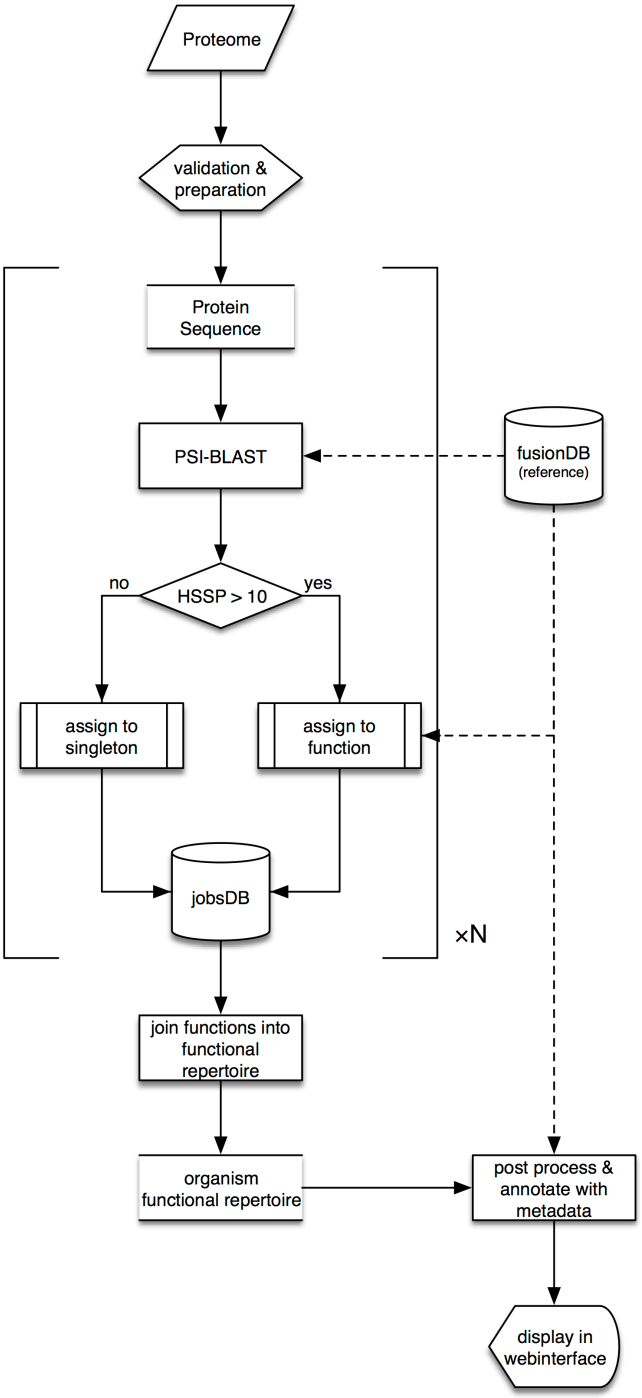
SOM Figures



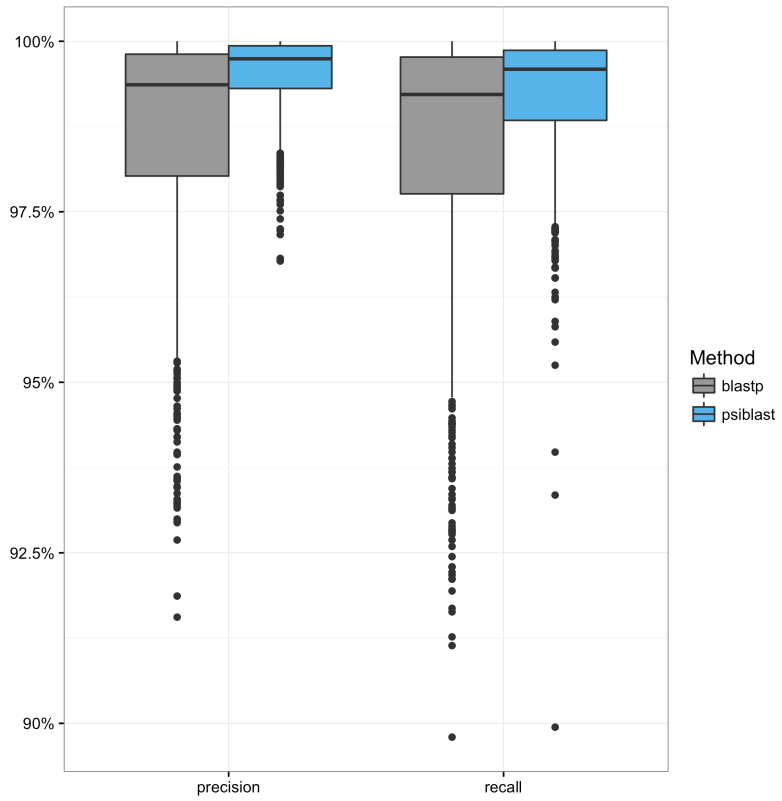
SOM Figure 1. Entity-relationship (ER)-model of *fusion*DB. For each organism, we store taxonomy information and additional organism metadata in *fusion*DB. Note that some metadata is generalized (*e.g.* hyper-thermophiles are considered thermophiles). Other information in *fusion*DB includes the proteins' associations to functional clusters, their functional annotations, and the pairwise organism functional similarities



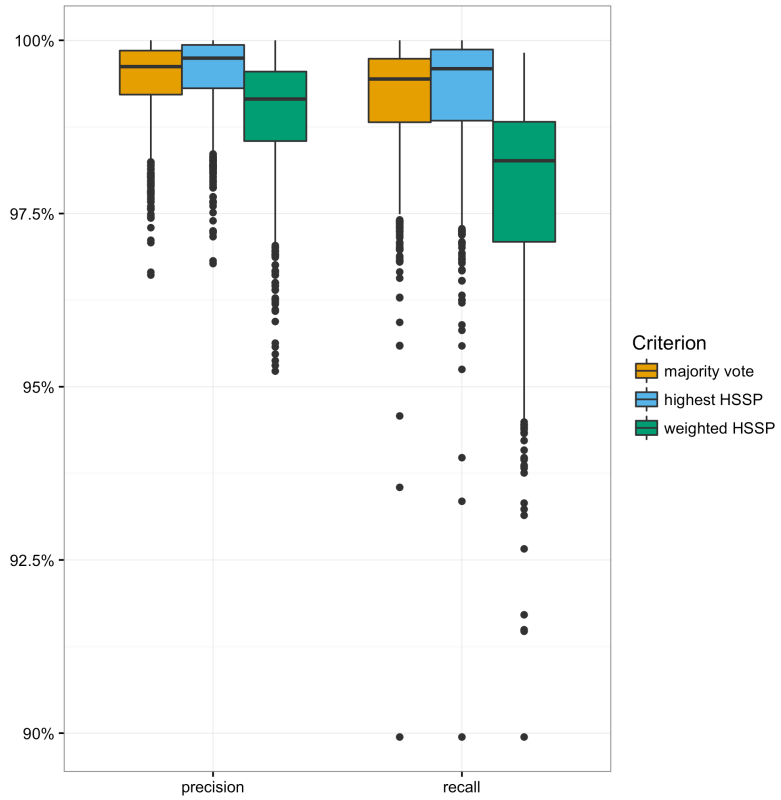
SOM Figure 2. ER-model of job database. User submitted microbes are stored in a separate database. Each organism proteome creates a job with a unique identifier. Related results and logs are also associated with this identifier.


Proteome
validation & preparation
Protein Sequence
PSI-BLAST
fusionDB (reference)
no
HSSP > 10
yes
assign to singleton
assign to function
jobsDB
×N
join functions into functional repertoire
organism functional repertoire
post process & annotate with metadata
display in webinterface

SOM Figure 3. Pipeline of mapping new proteomes. A submitted proteome is validated (*i.e.* checked for correct formatting and content) and split into single sequences. PSI-BLAST runs against *fusion*DB are distributed across computing resources per sequence. Each individual sequence is either assigned to a functional cluster in *fusion*DB or counted as a singleton. In a final step, the functional repertoire (the union of assigned function clusters and novel singletons) is compared to other organisms in *fusion*DB before displaying in the web interface.



SOM Figure 4. PSI-BLAST outperforms BLASTP in recovering functional repertoires. PSI-BLAST-based pipeline attains a mean precision of 99.5% and recall of 98.9%, while BLASTP is slightly lower (precision: 98.7%, recall: 98.2%). Additionally, upper and lower quantiles are tighter around the mean precision and recall for PSI-BLAST. Precision and recall are calculated from the overlap between the mapped functional repertoire and the reference repertoire.



SOM Figure 5. Highest HSSP offers best performance in assigning functional repertoires to query organisms. "Highest HSSP" hit-based method of mapping proteins to functional clusters is best at recalling organism functional repertoires.

SOM Tables

SOM Table 1. Taxonomic composition of environmentally distinct groups.

		# phylum	# family	# organisms
Temperature	Mesophile	23	166	1083
	Thermophile	16	42	115
	Psychrophile	3	16	33
	*No Data	14	68	143
Oxygen requirement	Facultative	10	51	207
	Aerobe	14	115	481
	Anaerobe	20	71	245
	*No Data	15	88	232
Habitat*	Soil	8 (11)	43 (75)	78 (279)
	Host	11 (15)	59 (94)	329 (706)
	Marine	10 (15)	24 (49)	61 (116)
	Fresh water	10 (18)	37 (84)	69 (271)
	*No Data	15	88	206

*No Data indicates missing annotations.
** One organism can be annotated with multiple habitats (*e.g.* both soil and host). The first number includes only organisms with one annotation, whereas the number in parenthesis includes organisms with multiple habitats.

SOM Table 2. Temperature preferences of organisms used in the case study.

Id	Temperature
Bacillus amyloliquefaciens DSM 7	Mesophile
Bacillus anthracis A0248	Mesophile
Bacillus cereus ATCC 14579	Mesophile
Bacillus coagulans 2 6	Thermophile
Bacillus coagulans 36D1	Thermophile
Bacillus licheniformis ATCC 14580	Mesophile
Bacillus megaterium DSM319	Mesophile
Bacillus subtilis 168	Mesophile
Bacillus thuringiensis Al Hakam	Mesophile
**Bacillus tusciae* DSM 2912	Thermophile
Bacillus weihenstephanensis KBAB4	Psychrotolerant
Desulfotomaculum carboxydivorans CO 1 SRB	Thermophile
Sulfobacillus acidophilus TPY	Thermophile

*Note that *Bacillus tusciae* is reclassified as *Kyrpidia tusciae* (Klenk, et al., 2011), another genus that is still evolutionarily more related to the *Bacilli* organisms than to the *Clostridia* organisms.

Supplementary Online References

Klenk, H.-P.*, et al.* (2011) Complete genome sequence of the thermophilic, hydrogen-oxidizing Bacillus tusciae type strain (T2(T)) and reclassification in the new genus, Kyrpidia gen. nov. as Kyrpidia tusciae comb. nov. and emendation of the family Alicyclobacillaceae da Costa and Rainey, 2010, *Standards in Genomic Sciences*, **5**, 121-134.
Punta, M.*, et al.* (2012) The Pfam protein families database, *Nucleic Acids Research*, **40**, D290-D301.
Rost, B. (2002) Enzyme function less conserved than anticipated, *J Mol Biol*, **318**, 595-608.
Zhu, C.*, et al.* (2015) Functional Basis of Microorganism Classification, *PLoS Comput Biol*, **11**, e1004472.

5 Common sequence variants affect molecular function more than rare variants?

5.1 Preface

Genetic variants can have wide spread impact on organism. This can range from simple functional changes of a protein, to less immediate changes like the microbial composition of microbiomes at various host body sites.

There is plentitude of variation analysis tools available including SNAP2 the tool I used here to exemplify approaches taken in variant analysis. Here I explore the functional effect of amino acid variants within the human population as well as amino acid variation between different species. The study serves as example to demonstrate the methodology that can be applied to investigate the interaction of genetic variation of hosts with their microbiome composition.

In this study I demonstrate that amino acid variants common within the human population ($\geq 5\%$) tend to be functionally more impactful on average, than functional impacts caused by rare variations. I hypothesize this is due two factors: (1) most common variants carry small functional effects. If they have a positive outcome for the organism, they are fixed over time. If they are damaging to the organism they are selected against, i.e. become rare variants over time. (2) rare variants on the other hand, have either such a strong functional impact that disappear again very rapidly (the minority of rare variants) or show no functional change all together, and are therefore not fixed over time, i.e. stay rare.

Furthermore I observe that inter-species variants (variants fixed at speciation) are also more neutral than variants in an evolving population.

Implementation, evaluation and execution of the work was done by me. Maximilian Hecht is the author of SNAP2 and provided assistance during the interpretation of the results. Jonas Reeb, provided the SNAP2 performance reevaluation. Maria Schelling provided the localization prediction of proteins. The manuscript was drafted by all authors.

5.2 Journal Article: Mahlich et al., Scientific reports 2017

www.nature.com/scientificreports

OPEN

Common sequence variants affect molecular function more than rare variants?

Received: 14 June 2016
Accepted: 28 February 2017
Published online: 09 May 2017

Yannick Mahlich[1,2,3,4], Jonas Reeb[1,3], Maximilian Hecht[1], Maria Schelling[1], Tjaart Andries Petrus De Beer[5], Yana Bromberg[2,4] & Burkhard Rost[1,4,6]

Any two unrelated individuals differ by about 10,000 single amino acid variants (SAVs). Do these impact molecular function? Experimental answers cannot answer comprehensively, while state-of-the-art prediction methods can. We predicted the functional impacts of SAVs within human and for variants between human and other species. Several surprising results stood out. Firstly, four methods (CADD, PolyPhen-2, SIFT, and SNAP2) agreed within 10 percentage points on the percentage of rare SAVs predicted with effect. However, they differed substantially for the common SAVs: SNAP2 predicted, on average, more effect for common than for rare SAVs. Given the large ExAC data sets sampling 60,706 individuals, the differences were extremely significant (p-value $<$ 2.2e-16). We provided evidence that SNAP2 might be closer to reality for common SAVs than the other methods, due to its different focus in development. Secondly, we predicted significantly higher fractions of SAVs with effect between healthy individuals than between species; the difference increased for more distantly related species. The same trends were maintained for subsets of only housekeeping proteins and when moving from exomes of 1,000 to 60,000 individuals. SAVs frozen at speciation might maintain protein function, while many variants within a species might bring about crucial changes, for better or worse.

Single nucleotide variants (SNVs) constitute the most frequent form of human genetic variation[1]. Here, we focus on non-synonymous SNVs, *i.e.* genomic variants that result in single amino acid variants (SAVs) in protein sequences. Children differ by about two SAVs from their parents (*de novo* variation), while any two unrelated individuals can differ by as many as 10–20 K[2]. The vast majority (99%) of the known unique SAVs are rare, *i.e.* observed in less than 1% of the population[1,3]. Only about 0.5% of the unique SAVs are common, i.e. observed in over 5% of the population[1,3]. SAVs can impact protein function in many ways.

We might be inclined to classify SAVs according to what they affect or do not affect. Effects are commonly distinguished upon protein function and structure. This distinction has limited value because what changes structure often tends to affect function. Similarly, we might distinguish between the effect upon molecular function (*e.g.* binding stronger or not binding), upon the role of a protein in a process (native process hampered, blocked, or non-native role acquired), or upon the localization of a protein (*e.g.* protein makes it to the membrane or not). Again the problem of this distinction is that these aspects are coupled: for instance, effects upon molecular function and localization might affect the process or not. All of the above, we might classify as effects upon the protein. Unfortunately, from all experiments monitoring SAV effects in many model organisms, just a few tens of thousands effects are available in public databases. For a tiny subset of these, enough detail is available to consider all effect types (structure vs. function, molecular vs. process vs. localization). We might consider the effect upon protein as *molecular* as opposed to the effect upon the organism, such as diseases. Toward this end, the distinction is often made between SAVs that cause severe monogenic diseases[4] (referred to as *OMIM-type SAVs*) or contribute to complex diseases[5] and low-effect SAVs, which are only cumulatively linked to our phenotypic

[1]Computational Biology & Bioinformatics - i12, Informatics, Technical University of Munich (TUM), Boltzmannstrasse 3, 85748, Garching/Munich, Germany. [2]Department of Biochemistry and Microbiology, Rutgers University, 76 Lipman Dr, New Brunswick, NJ, 08901, USA. [3]TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), Technische Universität München, 85748, Garching/Munich, Germany. [4]Institute of Advanced Study (TUM-IAS), Lichtenbergstr. 2a, 85748, Garching/Munich, Germany. [5]European Molecular Biology Laboratories, European Bioinformatics Institute (EMBL-EBI), Welcome Trust Genomes Campus, Cambridge, Cambridgeshire, UK. [6]Institute for Food and Plant Sciences WZW–Weihenstephan, Alte Akademie 8, Freising, Germany. Correspondence and requests for materials should be addressed to Y.M. (email: ymahlich@bromberglab.org)

individuality[6]. This latter distinction is the only one for which ample data is available. Methods predicting the effect of SAVs differ in many ways, including in what experimental data they use and how they use it. Typically, the methods inherit the bias of the data with respect to the type of effect considered.

Almost any SAV will have some effect under some condition. However, some SAVs clearly have stronger effects. A set of SAVs with stronger effect is likely to affect more aspects of function (molecular, process, localization) simultaneously. For example, OMIM-like SAVs are assumed to largely affect the biological process through strong effects upon structure and/or molecular function (or localization). Many of these SAVs also affect the organism as a whole manifesting as disease. Prediction methods add another level of complexity: some methods provide values reflecting the strength of a prediction that correlates with the reliability of the prediction (i.e. its accuracy) and the strength of the experimental effect. Assume we analyzed two subsets of effect predictions from a method: those with predictions stronger than threshold T1 and those stronger than T2, where T2 > T1. For methods for which the score is well-balanced, two statements are true: (1) setT2 is predicted at higher accuracy than setT1, and (2) setT2 has, on average, stronger effect than setT1. No matter what conditions or populations we analyze, results valid at T1 are also valid at T2, as long as the thresholds are chosen in a regime in which the prediction method works.

Analyzing the effect of all known variants experimentally remains unfeasible. Computational methods may incorrectly predict the effect for some SAVs, but they successfully capture trends for large sets of SAVs[7–11]. Furthermore, computational predictions are available for all variants and inherit only some of the bias from today's experimental techniques. Both experimental and computational assays often fail to infer the impact of variation on the organism as a whole from individual SAV molecular effects.

The 1000 Genomes Project (*1KG*)[12,13] sequenced 1,092 individuals from 14 populations recording about 268,115 SAVs. In August 2016, the MacArthur lab at the Broad expanded on this collection by reporting 7,599,572 SAVs between 60,706 people[3]. How much of this variation impacts protein function? Only *in silico* tools can fully address this question. Here, we present a comprehensive and detailed analysis of the known human SAVs (1KG) and of SAVs that differentiate human proteins from their homologs in other species (hominids, primates, rodents, and fly).

Results

Many 60KE SAVs predicted with effect. SNAP2 spreads its predictions into a wide interval of scores, in addition to predicting a reduced binary outcome (effect/neutral). The scale ranges from −100 (SAV strongly predicted as neutral) to +100 (SAV strongly predicted as effect–either deleterious or beneficial). For a binary prediction, SNAP2 is optimized for the threshold = 0 (*i.e.* neutral: −100 ≤ SNAP2-score ≤0 and effect: 0 < SNAP2-score ≤100). At this default threshold, 78% of the known neutral and 79% of the known effect SAVs are predicted correctly (Fig. 1 lower panel). Higher absolute values of scores imply more reliable predictions, as illustrated by the three example scores. For instance, zooming into more reliable neutral predictions at scores ≤−42 or into stronger effect predictions at scores ≥+50 raises the accuracy to 85%; going further to effect scores ≥+75 reaches 88% accuracy (Fig. 1 lower panel). We use these three examples throughout the manuscript to highlight trends. Note that to avoid two possible misunderstandings we point out that: Firstly, there is no single threshold for "strong" predictions, the higher the value the stronger the effect (below) and the higher the probability for the prediction to be correct. Secondly, zooming into some higher threshold such as SNAP2-score ≥75 does not imply that all with SNAP2-scores <75 are predicted neutral. Instead, we simply focus on a subset of strongly predicted SAVs.

We used the raw SNAP2-score to summarize our central results for all data sets cumulatively, *i.e.* showing at each SNAP2-score which percentage of the data set was predicted at or above that score. For instance, of the 7,599,572 60KE SAVs in healthy humans, for which predictions were available, 1,642,225 (about 21%) were predicted to have functional effects at a SNAP2-score ≥+50 (Fig. 1A, upper panel *60KE* purple line with triangle markers intersecting the blue vertical dashed line of score = +50). This threshold corresponds to an estimated accuracy of 85% (red line in Fig. 1B). On the other hand, SNAP2 predicted 100 − 77 = 23% of the 60KE SAVs as clearly neutral (Fig. 1A leftmost blue vertical dashed line of score −42) at an expected accuracy of 85% (Fig. 1B).

Another extreme point was SNAP2-score >+75 (Fig. 1, rightmost blue vertical dashed line): if we considered only SAVs predicted above this effect level, we would capture half of the OMIM SAVs (Fig. 1A, orange line with circular markers; Fig. 1B, 88% accuracy). For the same threshold about 1/15th of all 60KE SAVs (496,854) were predicted to have an effect. Loosely put, one of every 15 SAVs in healthy individuals is predicted to have as strong an effect as the top 50% of known disease variants. Note that the OMIM SAVs were predicted by a version of SNAP2 that was not trained on any OMIM or HumVar SAVs (Methods). Another SNAP2 version, that did use such disease-related SAVs, predicted a much stronger effect (Supplementary Fig. S1).

SNAP2 has been evaluated in comprehensive cross-validation tests. However, the performance estimates provided here (Fig. 1B) depend crucially on what data is included in the assessment (they are higher when using OMIM-type SAVs and lower when using the small subset of experimental-only neutrals). The error estimates in performance curves for the 60KE data (Fig. 1A upward pointing triangles) were obtained by boot-strapping[14], i.e. by testing how the results depend upon changes in the data set. Note that for all 60KE data these error estimates were visually indistinguishable from the curves shown (at a 99.7% confidence interval, i.e. SNAP2-score ±3 standard error of mean: *denisovan*: ±8.46, chimp: ±0.45, 60k all: ±0.06, 60KE common: ±0.80, 60KE rare: ±0.06).

For our previous method SNAP1, we have shown[6,15] that the SNAP-score correlates with effect strength, *e.g.* SAVs predicted closer to +100 tend to have stronger impact on molecular function than SAVs predicted closer to +50. Here, we confirmed the same for SNAP2 (Fig. 2). SNAP2 was trained on PMD variants, but never with fine-grained classification by degree of effect; instead SNAP2 was trained to classify binary labels (effect/neutral).

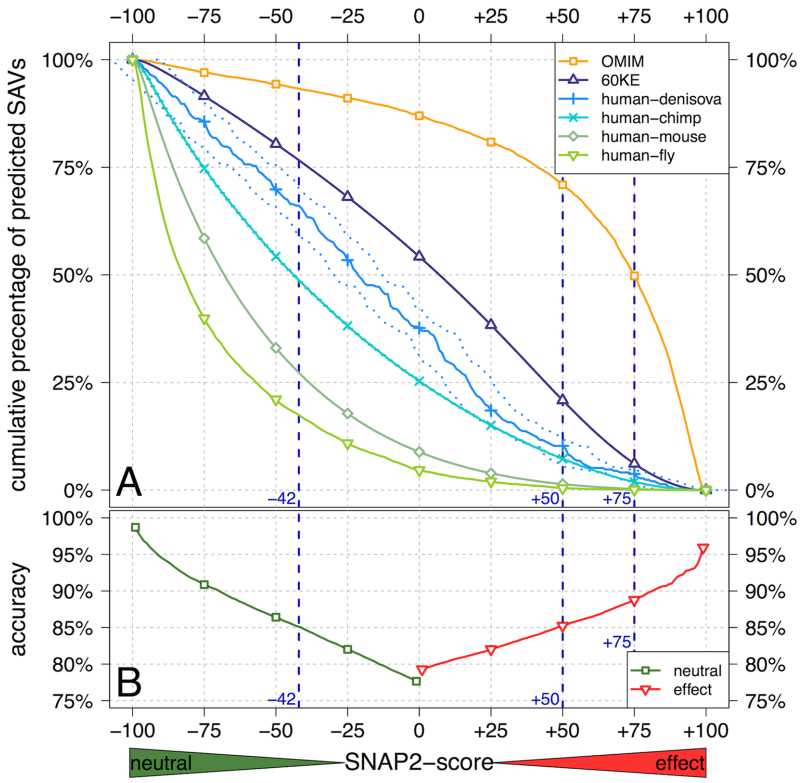


Figure 1. 60KE SAVs predicted to have more effect than cross-species variants. SNAP2 predicts the effect of single amino acid sequence variants (SAVs) upon protein function: the higher the score, the more reliable the prediction (horizontal x-axis, toward right); the more negative, the stronger the prediction that the variant is neutral (horizontal x-axis, toward left). The top panel (A) gives cumulative percentages, i.e. the percentage of SAVs in a data set predicted above a certain value, *e.g.* for SNAP2-score ≥+75, about 6% of all 60KE SAVs are predicted to have an effect; at the same threshold about half of all disease-causing SAVs are predicted to affect function. For *60KE*, *denisova* and *chimp*, 99.7% confidence intervals (SNAP2-score ±3 standard error of mean) are indicated by dotted lines (indistinguishable for *60KE*, barely distinguishable for *chimp*, clearly visible for *denisova*). Lower panel (B) gives cumulative accuracy (red: effect-SAVs correctly predicted to have effects, green: neutral-SAVs correctly predicted); here the values accumulate from the extremes to 0, *i.e.* left-to-right for neutral (green −100 to 0) and right-to-left for effect (red +100 to 0); estimates from cross-validation using only molecular function[21]. For instance, at SNAP2-scores ≥+75 about 88% of all effect-SAVs are correctly predicted. On the other hand, variations between homologs in human and other species (human-denisova, human-chimp, human-mouse, and human-fly) were predicted to be much more neutral (all curves shifted toward lower left corner of neutral variants).

Hence, the observed difference between mild, moderate, and severe could not have originated from SNAP2 training other than through the simple fact that stronger effects yield more consistent data and therefore effect strength is captured by the method. This correlation was exactly what we wanted to show.

Cross-species variants predicted with less effect than variation within human. We analyzed all amino acid differences between human proteins and their *homologs* in other species ("cross-species variation"). For simplicity, we referred to those variants as to SAVs. There is an important caveat for the comparison between 60KE and cross-species SAVs. With the 60KE set we compared a population using essentially the same gene pool (pairs of people differ by few SAVs spread across their ~20 K proteins). In contrast, the cross-species comparison had to be limited to subsets of corresponding proteins. The size of this subset is inversely proportional to the

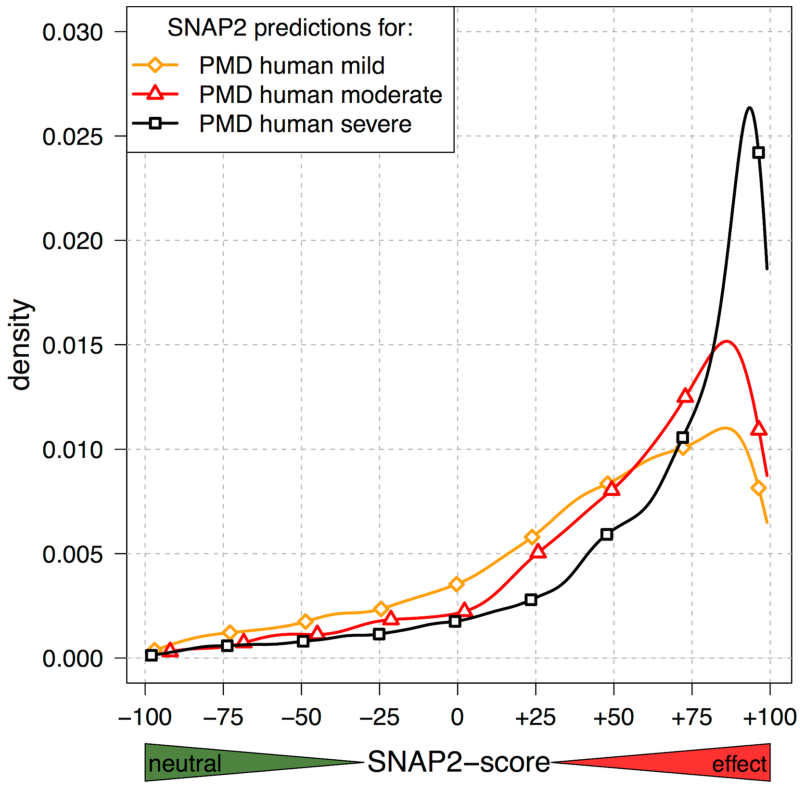


Figure 2. Higher SNAP2-scores imply stronger effect upon molecular function. We classified SAVs from the Protein Mutant Database (PMD) according to their impact upon molecular protein function into three classes (mild, moderate, and severe). Here, we repeat this analysis applying SNAP2 to the subset of human SAVs in PMD. We show density distributions, instead of cumulative. Although the three curves overlap, the shift is significant and consistent (black curve with most effect highest shift to right, orange curve with weakest shift most to the left). Thus, the SNAP2-score correlated with the strength of the effect upon molecular function.

evolutionary distance; *i.e.* fewer proteins for more distantly related organisms. To simplify: 60KE compared few SAVs in all proteins, while cross-species compared many SAVs in a few proteins.

We began with the extinct hominid *Denisova hominin*. In all orthologs, we considered the effect of substituting a *Denisova* amino acid that was introduced into the corresponding human protein back to the human reference amino acid. When using all available orthologs, the SNAP2 predictions for the human-denisovan cross-species SAVs were moved toward "less effect", *i.e.* a lower fraction of SAVs was predicted with effect for human-denisovan than for the 60KE-set (Fig. 1A, human-denisovan below 60KE). This implied that sequence variants that became *fixed* in the modern human population were more neutral, on average, than the SAVs within the living human population (60KE). The strength of this shift depended on the SNAP2-score: the numeric difference was highest thresholds of 0 and the relative difference was highest around thresholds of +50 (Fig. 1A). The shift between 60KE and cross-species SAVs was increased with species divergence (Fig. 1A: curves shifted toward the lower left implying less effect for chimp, mouse, and fly). Note that any overlap with SNAP2 training data was excluded from this study (Supplementary Note).

Next we addressed the problem of different "gene pools" (sets of proteins) for the comparison within human and between human and other species. Toward this end, we restricted our analysis to subsets of identical proteins, *i.e.* by restricting the SAVs to the subset of orthologs that were common to human, chimp, and mouse. The subset was restricted further by the constraint that 1KG SAVs be also observed in the same protein. The resulting curves were shifted toward "less effect" (curves higher in Fig. 1A than in Fig. 3); the standard errors of the mean increased only slightly (human: from 0.10 to 0.14; chimp: from 0.16 to 0.22; mouse: no change). However, the

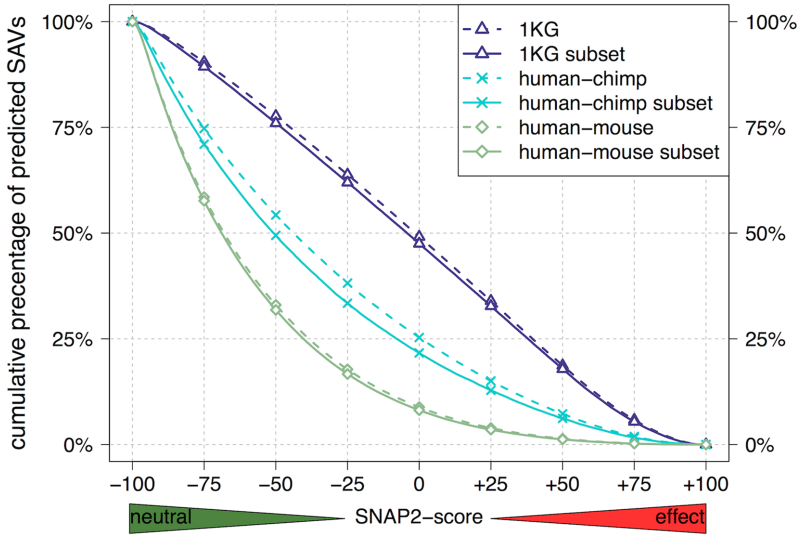


Figure 3. Subsets of "house-keeping" proteins confirmed findings for entire proteomes. We reduced the analysis to SAVs from subsets of orthologs between three organisms (human, chimp, mouse), and with SAVs observed in the 1KG data. For brevity, we referred to those as to "house-keeping" proteins. With respect to the observation for the entire data set (Fig. 1), the curves shifted less strongly, but the main trend remained: a higher fraction of the SAVs in cross-species comparison (human-chimp and human-mouse) was predicted as neutral than for the SAVs between healthy individuals (1KG). Furthermore, the shift between cross-species and 1KG was higher for larger evolutionary distances (more neutral for larger distance).

SAV set		SNAP2*	CADD	PolyPhen-2	SIFT
effect/all	common	61%	19%	18%	19%
	rare	50%	50%	42%	40%
	all	51%	46%	39%	47%
neutral/all	common	39%	81%	82%	81%
	rare	50%	50%	58%	60%
	all	49%	54%	61%	63%

Table 1. Similar ratios of neutral/effect predicted by four methods for 1KG SAVs*. *Data sets: Rare SAVs (LDAF < 0.01): all methods agreed within ten percentage points on the ratio; common SAVs (LDAF ≥ 0.05); the values for "all" also included uncommon SAVs (0.01 ≤ LDAF < 0.05). Methods: SNAP2* implies error-corrected estimates for SNAP2 (below). The four methods compared here have different aims and use different thresholds. Here, we applied defaults to simplify the comparison and interpreted predictions as binary (effect/neutral) according to those thresholds. In particular, we chose the following thresholds. Effect: SNAP2-score > 0, CADDv1.3 raw score > 3, PolyPhen-2 = *probably or possibly damaging*, SIFT = *deleterious*; neutral: SNAP2-score ≤ 0, CADDv1.3 raw score ≤ 3, PolyPhen-2 = *benign*, SIFT = *tolerated*. SNAP2* error correction: Values for SNAP2 were corrected for false positives and false negatives (e.g. Neff* = Neff(raw) − FPR(Neff(raw)) + FNR(Nneu(raw)). Error correction lowered the estimates for common effect and increased that for rare effect.

major characteristics of the curve shifts were not altered by the constraint to the subsets (Fig. 3: 1KG curve highest, higher evolutionary distance corresponds to lower curves). For less restrictive sets of orthologs, we observed similar trends (Supplementary Fig. S3).

Common 60KE SAVs predicted with more effect than rare SAVs. Rare SAVs (allele frequency [LDAF] < 1%) dominate the set of unique SAVs in the 60KE data (7,530,337, *i.e.* 99.1% of all SAVs) and therefore dominate the overall analysis. If our objective were to assess the per-person effect instead of the per-SAV effect, we could have removed this bias by counting each SAV exactly once, *i.e.* by letting SAVs observed in ten individuals count ten-times more than those observed only once.

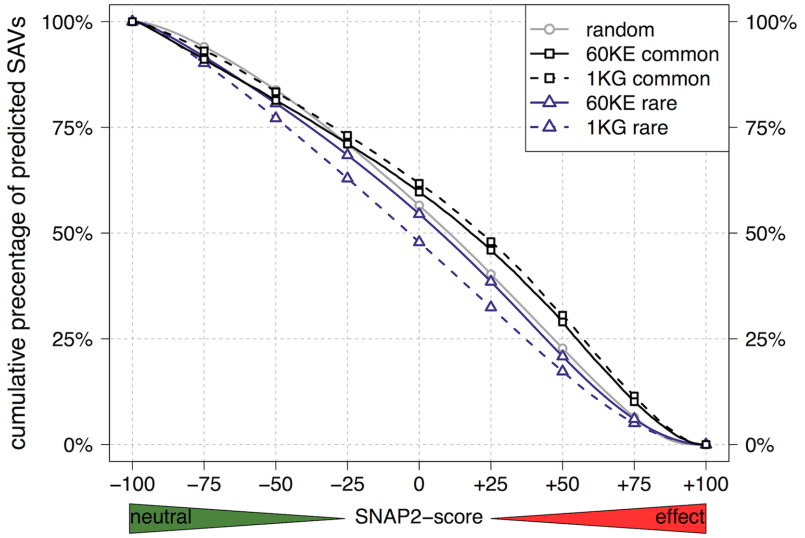


Figure 4. Common SAVs predicted with more effect than rare SAVs. We grouped SAVs by their observed frequency in 1KG and 60KE exome data: rare (LDAF < 1%: dark blue triangles), uncommon (1% ≤ LDAF < 5%: not displayed), and common (LDAF ≥ 5%: black squares). The potential mutational background for human was estimated by randomly selecting a set of SNV-possible SAVs (gray circles). The curves for rare SAVs were similar to the results for all SAVs (Fig. 1, purple triangles for 60KE) since counting only unique SAVs the results were dominated by rare SAVs. Rare SAVs were predicted below randomly chosen SNV-possible SAVs, although the recent 60KE set came close to random. In contrast, the set of common SAVs remained substantially above the random curve for both common-1KG and common-60KE (Kolmogorv-Smirnov, estimated p-value < 2.2e-16 in both cases).

	SNAP2	CADD	PolyPhen-2	SIFT
SNAP2		$0.36 \pm 6.8*10^{-3}$	$0.31 \pm 6.9*10^{-3}$	$0.41 \pm 6.6*10^{-3}$
CADD	$0.57 \pm 1.8*10^{-3}$		$0.70 \pm 5.2*10^{-3}$	$0.63 \pm 5.7*10^{-3}$
PolyPhen-2	$0.53 \pm 1.9*10^{-3}$	$0.77 \pm 1.6*10^{-3}$		$0.47 \pm 6.4*10^{-3}$
SIFT	$0.57 \pm 1.8*10^{-3}$	$0.65 \pm 1.7*10^{-3}$	$0.49 \pm 1.9*10^{-3}$	

Table 2. Predictions more correlated for rare than for common 1KG SAVs*. *Pearson correlation coefficients above the diagonal show the agreement for common 1KG SAVs, those below the diagonal for rare 1KG SAVs. Correlation values were calculated using the predicted raw scores of SAVs for which predictions were available for each method. The correlation in predictions between the four methods was higher for all pairs below the diagonal, *i.e.* for rare SAVs. Standard error of r: $SE_r = \text{sqrt}((1 - r^2)/(n - 2))$; $n_{common} = 18{,}876$; $n_{rare} = 209{,}928$.

We might expect rare SAVs to sample the space of all possible SAVs. We investigated this through two sets of random SAVs introduced *in silico* into human proteins. One random set contained SAVs replacing the native amino acid by one of the 19 non-native amino acids. The other set was restricted to SNV-possible SAVs, *i.e.* amino acid substitutions that can be reached by a single nucleotide change (SNV; the set *SNV-possible* constitutes a subset of *19-non-native*). SNAP2 predicted the *19-non-native* SAVs to have slightly but significantly (two-sample Kolmogorov-Smirnov, KS, test: D = 0.068; n,n' = 268,115; p-value < 2.2e-16) higher effects than SNV-possible SAVs (Supplementary Fig. S2).

One important aspect in the shift from the analysis of rare SAVs for 1000 people (1KG) to that for 60,000 people (60KE) was that the effect predicted for these variants by SNAP2 was very similar to the effect predicted for the random subset of SNV-possible SAVs (Fig. 4: solid blue line "60KE rare" much closer to gray random than dashed blue line "1KG rare"). The differences between the curves were small in absolute terms but statistically very significant (two-sample KS test, D = 0.033; $n_{rare} = 7{,}530{,}277$, $n_{random} = 268{,}115$; p-value < 2.2e-16).

SNAP2 predicted a similar fraction of rare SAVs (LDAF < 1%: 209,928 variants, *i.e.* 81% of all *1KG*) to have an effect as did PolyPhen-2[16], CADD[17], and SIFT[18] (Table 1; range from 40–50%). This statement comes with the caveat that the binary classification requires the introduction of a threshold (effect/neutral) that may not be appropriate for a particular tool. For instance, for CADD, and to a large extent for SNAP2, the particular

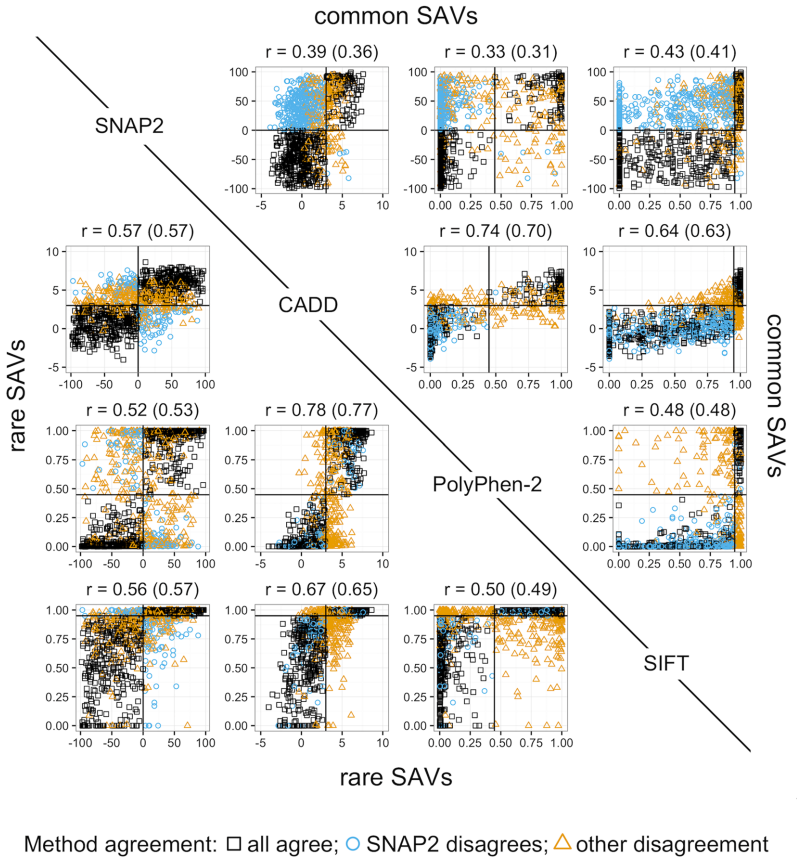


Figure 5. Methods correlated more for rare than for common 1KG SAVs. Each plot shows the correlation of functional effect scores between one pair of prediction methods for two samples of 1,000 rare and 1,000 common SAVs from 1KG. Results for common SAVs are shown above diagonal, those for rare SAVs are given below the diagonal. With the order of the plots being 1 = SNAP2, 2 = CADD, 3 = PolyPhen-2, and 4 = SIFT, this implied that the plot corresponding to matrix element Pmn compared common SAVs between methods m and n (above diagonal), and the element Pnm rare SAVs between those two (below diagonal). For instance, row = 1/column = 2 gave the correlation between SNAP2 and CADD for common SAVs, while the transposed element row = 2/column = 1 correlated rare SAVs for SNAP2 and CADD. Each point represents a pair of scores for a single SAV, e.g. from SNAP2 and CADD. The predicted score for SIFT has been inverted (1-SIFT) to ease the comparisons. The shape and color reflect the overall method agreement. We use the following code: black squares mark SAVs for which all four methods agree on the binary classification. Blue circles mark SAVs for which all methods but SNAP2 agree; orange triangles mark all other points. The Pearson Correlation Coefficient for all 1,000 SAVs was added above each plot, along with the corresponding value for the full set of all SAVs (in brackets, as in Table 2).

threshold used depends on the data set and the question being asked. Overall, the predictions for rare SAVs also correlated pairwise between SNAP2, PolyPhen-2, CADD, and SIFT (Table 2, Fig. 5: below diagonal: correlations from 0.49–0.77). However, the set of rare SAVs for which all four methods predicted an effect was around 28% (data not shown), *i.e.* the methods correlated on average, but differed in detail in the trends captured.

PolyPhen-2, CADD and SIFT also largely agreed in their classifications for the overall amount of effect/neutral for common SAVs (LDAF ≥ 5%: 18,876 variants, *i.e.* 7% of all *1KG*, Fig. 5: above diagonal, agreeing predictions in lower left and upper right quadrants). In contrast SNAP2 predicted larger fractions of the common than of the rare SAVs to have an effect, *i.e.* SNAP2, on average, predicted stronger effects for common than for rare SAVs (Fig. 4: *common* moved toward upper right, *i.e.* more effect). The shift could also be expressed by

calculating the "effect AUC" of the density distribution for SNAP2 scores, *i.e.* the Area-Under-the-Curve (AUC) from $0 \leq$ SNAP2-score ≤ 100: AUC = 0.61 for common and AUC = 0.48 for rare SAVs (Supplementary Fig. S4). Another example highlighting the difference between common and rare SAVs: over 12% of the common, as opposed to fewer than 6% of the rare, SAVs were predicted to affect function as strongly (same SNAP2-score) as 50% of the strongest OMIM SAVs (Fig. 1 *OMIM*).

SAVs with effects not distributed randomly. We investigated whether or not SAVs predicted with strong effect (SNAP2-score >50) were randomly distributed through two analyses: simple statistical enrichment and the enrichment of sub-cellular localization.

Firstly, we investigated the distribution of common SAVs per protein. For about 18k proteins the 60KE data set annotated rare and common SAVs in the same protein. For 1,254 of those there was more than one common SAV and SNAP2 strongly predicted over 50% of the common SAVs to have effect (SNAP2-score >50). An analogous filter on the rare SAVs ($\geq$2 rare SAVs in same protein + $\geq$50% of rare SAVs at SNAP2-score >50) led to only 304 out of 18k proteins. Furthermore, nine proteins (ENSP00000363436, ENSP00000369568, ENSP00000408146, ENSP00000337397, ENSP00000363433, ENSP00000353078, ENSP00000415517, ENSP00000310338, ENSP00000413079) display ten or more common variants, and more than 50% have a SNAP2-score >50. All those nine proteins contain more rare variants with SNAP2-score $\leq$50 than with SNAP2-score >50.

Secondly, we predicted the sub-cellular localization for all proteins with LocTree3[19] (given the many annotations available for human proteins, most predictions were based on homolog-inference). Of all the proteins with SAVs, ~28% were predicted to be nuclear and about ~23% as secreted or cell membrane. We compared these numbers to different ways of scanning for proteins enriched in effect predictions. For the subset of all proteins with at least 50% of their residue positions observed as a SAV, and 50% of those predicted with SNAP2-score >50, 17% were nuclear (reduction to 60% of expected) and 40% secreted or cell membrane (1.7 fold over-representation over expected. We found a similar over-representation of "secreted + cell membrane" in proteins with many effect SAVs when looking at the subset of all proteins with at least 4 common SAVs for which at least 30% were predicted at SNAP2-score >50 (value of 30% guided by the average expected at that SNAP2-score, cf. Fig. 4-common). About 44% of those proteins were predicted as secreted or cell membrane, i.e. 1.9 times more than expected. Finally, only 29 proteins were so enriched in common SAVs that 10% of all residues had common SAV. LocTree3 predicted 21 to be secreted.

Discussion

Prediction distributions are different for sets of SAVs. We analyzed sets of SAVs that were predicted to have effect or not. To estimate the number of expected mistakes, we have to provide extensive performance evaluations of our method. Instead, of these approximations we compared distributions of different sets of SAVs and their predicted effects. Even with relatively high rates of errors, methods can set such distributions reliably apart as long as their mistakes are not systematic with respect to the results. Below we argue that such systematic bias explains the differential results for prediction methods with respect to human SAVs that are common and rare. For instance, if we measured the height of women in Greece and Germany, we might find out that they differ by barely 2 cm, while the standard deviation for the measures are 5-times this difference. Despite the considerable standard deviation, we can easily distinguish between two countries of millions of individuals. The same is true for the differences between the distributions for the prediction of SAV effects.

Cross-species vs. 1KG and 60KE: changes from a point frozen in time. The cross-species "SAVs" appear to be describing something very different from the SAVs in the contemporary human population (1KG/60KE sets). Our analysis, nevertheless, assessed the effects of sequence variation in the same way. Our first results (Fig. 1A) compared orthologous proteins between human and other species; i.e. each curve was based upon a different set of proteins–the subset of proteins orthologous between human and the other species. We chose these comparisons because using the subset of proteins common to all species is so small that is clearly not representative of all proteins, a set that we might refer to as the "house keeping" proteins. Any results obtained exclusively for these specially selected proteins might be biased. Surprisingly, the conclusions did not change between taking "all orthologs" and restricting the analysis to "house keeping" genes (Fig. 3 dashed vs. solid lines). In fact, by filtering the data differently, we could even confirm the major trends to a level as far diverged as human-fly (Supplementary Fig. S3).

How can SAVs between people affect function more than those between species? We compared SAVs within a dynamic, evolving population (human) with SAVs that describe a speciation event, *i.e.* were frozen in time. At speciation, SAV effects between descendants likely randomly sampled the 1KG data. Thereafter, each protein carrying a "speciation SAV" had two possible fates. Either, the protein has drifted away in sequence and function to an extent that it has not been considered in our comparison of evolutionarily related proteins (orthologs). Or, it has maintained function by constraining variation to neutral SAVs. Thus, inter-species SAVs might appear to be more neutral with increasing evolutionary distance due to the process of removing proteins with too many effect SAVs from the comparison. Indeed, the subset of "house-keeping" proteins differed substantially in their effect from all proteins (difference between dashed and solid lines in Fig. 3 and Supplementary Fig. S3). However, the reported effect that 1KG SAVs were predicted, on average, with more effect than cross-species SAVs was valid both for only "house keeping" and for "all proteins". This signal could therefore neither be explained by data bias nor by data inconsistencies.

Will all possible SAVs be observed? Not all possible SAVs have been observed in healthy people[3]: some may be lethal and prevent the carrier from being sequenced; others might prevent development in much earlier

stages, even long before birth. Will we observe all remaining ones? Our *in silico* mutagenesis experiment randomly sampled impact predictions for every possible SAV (19-non-native Supplementary Fig. S2) and every SNV-possible SAV (SNV-possible in Supplementary Fig. S2). This enabled us to gauge the expected effect of random variation in comparison to those observed in 1KG. We noted small but significant differences between what has been observed between healthy people and a random subset (Supplementary Fig. S2). This implied that the SAVs observed in 1KG were NOT random subsets of all possible SAVs. This in turn enabled the estimation of what to expect from observing SAVs for a larger population.

When we analyzed the data for the 60KE set, we confirmed the above observation; namely that the score distributions for rare SAVs approached random (SNV-possible, Fig. 4). Although the difference between random and rare became smaller, it remained significant (p-value $<$ 2.2e-16). There will always be a difference between what can be observed and what can be simulated: some SAVs will simply be "too fatal" to observe[20]. With the 60KE curve for rare SAVs so close to random, did this imply that such an effect was relatively minor, *i.e.* that very few SAVs are that deadly? Answers remain speculative. One problem lies in the scale of the numbers: even if thousands of SAVs were deadly, their effect could easily be overshadowed by $>$7 million SAV set size.

Common SAVs affect function more than rare SAVs?. Many colleagues who we have confronted with our data had expected the predictions for rare SAVs (observed in $<$1% of population) to be moved toward "more effect" (moved toward right in Fig. 4) than those for common SAVs (observed in $>$5% of the population). Predictions from CADD, PolyPhen-2, and SIFT confirmed this expectation (Supplementary Fig. S5), while SNAP2 predicted the opposite, *i.e.* predictions for common SAVs were shifted toward "more effect" than for rare SAVs (Fig. 5). We also established that SNAP2 on the one hand and CADD, PolyPhen-2, and SIFT on the other hand largely agreed on the fraction of effect/neutral predictions for rare SAVs (Tables 1 and 2). Although all four methods agreed for only about half of effect predictions for the rare SAVs, two-method correlations were fairly high for rare SAVs (Table 2, lower diagonal).

Could SNAP2 predictions be wrong more often when they disagree with three other methods that agree with each other? First off: CADD, Polyphen-2 and SIFT agreed with each other more for rare than for common SAVs (Table 2, values higher on lower than on upper diagonal). Thus, the view of "three over one" is not fully supported by the details.

The extreme hypothesis is that for the majority of common SAVs for which SNAP2 predicts more effect than the other methods, SNAP2 is wrong. We have published two findings that clearly refuted this hypothesis for two of the methods (PolyPhen-2 and SIFT). Firstly, all three methods performed significantly worse for difficult-to-predict SAVs. Secondly, for a subset of human SAVs that had been used to train PolyPhen-2 and was biased by a predominance of effect, SNAP2 still predicted much better than SIFT (Q2-SNAP2 = 58% vs. Q2-SIFT = 44%) and *on par* with PolyPhen-2 although the latter had the advantage over SNAP2 of having been optimized on these data[21]. Although these results suggest that SNAP2 was right for most of the SAVs for which it's prediction differed from the others, SNAP2 might still systematically mis-predict common SAVs.

Could details in methods development result in systematic mistakes for common SAVs? SNAP2 was developed predominantly on rare SAVs. More explicitly, of all the SAVs trained to have effect, about one quarter were *OMIM-like*, *i.e.* resembled rare SAVs. Thus, SNAP2 is likely biased towards predicting rare SAVs as having more effect than common. Correcting for this bias, we expect the "true" common curve to be moved even more to the right (towards more effect). In contrast, PolyPhen-2, CADD, and SIFT have been built using principles that might explain the bias toward "SAVs common to a population neutral". PolyPhen-2 was optimized to differentiate between monogenic disease-causing SAVs (rare by definition) and variation between orthologs. This implies that the machine learning may have enforced a much more substantial and explicit bias toward "rare have effect" by teaching "100% of the effect is rare", as opposed to "25%" for SNAP2. CADD implicitly shares some of this bias as it optimizes the separation between simulated variants and variants that differentiate orthologs. The simulated variants, by definition, are not observed, most likely because they cause disease or are even lethal, *i.e.* impact the whole organism. SIFT is built upon a similar idea, namely that SAVs conserved throughout evolution are likely to be neutral, suggesting that common effect are systematically unlikely to be captured by this method. While SNAP2 also uses conservation, many of the SAVs used for training SNAP2 *effect* were not conserved; conversely, many SAVs used to train *neutral* were conserved. Thus, the neural networks underlying SNAP2 might have put SNAP2 into the unique position of correctly spotting SAVs that are conserved, yet might affect function[21].

The differences in bias in the training data between SNAP2/SNAP1 on the one side and between PolyPhen-2, CADD, and SIFT on the other side might explain why common SAVs were predicted so differently. A related explanation pertained to different objectives. SNAP aims at predicting the effect of SAVs upon protein function; most training data measures molecular function rather than biological process. In contrast, PolyPhen-2, CADD, and SIFT, have been optimized with a more focused view upon pathogenicity for an organism. SNAP2 might correctly identify trends in the changes of molecular function because it avoids labeling variants in the context of the pressure exerted upon the organism.

SNAP2 cannot distinguish between SAVs that help and those that hurt the organism[15]. Although common variants might affect molecular function significantly, as suggested by SNAP2, they are unlikely to cause severe diseases, as illustrated by the predictions of PolyPhen-2 and CADD, since such extreme functional disruptions are unlikely to spread in a population. Nevertheless, SNAP2, CADD, PolyPhen-2, and SIFT predicted similar ratios for neutral/effect for rare SAVs and are fairly correlated for those (Tables 1 and 2). The big disparity that flips the prediction from "common less effect" to "rare less effect" between SNAP2 and the others is based only on the common SAVs.

An alternative way to explain the difference between SNAP and the other three (CADD, PolyPhen-2 and SIFT) is the following hypothesis. (1) All methods might capture the effects of monogenic-disease causing SAVs of the OMIM-type equally well, and therefore agree in their predictions for rare SAVs. (2) Effects for common

SAVs are less likely to impact the organism although they might strongly affect molecular function, and SNAP2 might be the best of the four methods in capturing such effects upon molecular function.

To prove this hypothesis, we have to show that many common SAVs have effects and have never been used by any of the methods for training. Unfortunately, there is no experimental data available to support or refute this point. However, recent deep-scanning experiments might at least show trends as to which method is best at capturing effects upon molecular function. In one case study, we compared the correlation between CADD and SNAP for one particular protein[22] (BRCA1 in Supplementary Fig. S6). Hopf *et al.* have recently analyzed the performance of prediction methods for several deep-scanning experimental data sets[23,24]. This analysis confirmed the trends that we observed: SNAP2 captured effects upon molecular function better than the other three methods. Incidentally, those analyses also suggested that the low-throughput experiments previously used to develop and assess prediction methods might over-estimate performance, at least if taking high-throughput deep-scanning experiments as a better proxy for reality. Thus, while many SNAP2 predictions might be wrong, all analyses converge upon the explanation that SNAP2 captures the effect of common SAVs upon molecular function better than CADD, PolyPhen-2, or SIFT. If so, the SNAP2 predictions discovered an unexpected reality, namely that common SAVs have, on average, more effect upon molecular function than rare SAVs. Common SAVs, in fact, might be relatively enriched through evolutionary selection as those that affect molecular function in a way that might help to drive the evolution of the species.

Conclusion

Our results are compatible with the distinction of two types of Single Amino acid Variants (SAVs). (1) *Point punches*, *i.e.* genetic alterations leading to large molecular changes that significantly diversify proteins and molecular pathways of individuals. (2) *Additive small effects*, *i.e.* near-neutral variants in many genes whose cumulative interplay is responsible for the impact upon pathway-wide molecular functionality. In this view, isolated strong-effect SAVs usually do not drive speciation. Within the individuals of one species, however, many of the common variants strongly impact protein function for better or worse (gain- vs. loss-of-function). Together, these findings might imply that common SAVs are unlikely to drive speciation. How much variation is beneficial to the individual and how much is necessary for the survival of the species? To answer this conundrum, we need better experimental and computational tools that distinguish the directionality of change and bridge from the micro-molecular view of single sequence variants to the macro-systems view of phenotypic impact for the organism.

We also observed that although *in silico* methods often agree in the effects they predict for SAVs, their differences are substantial enough to completely invert predicted trends as extreme as from "common SAVs have more effect" to "rare SAVs have more effect". We argued that such crucial differences originated from the way the methods were trained, and that SNAP2 picked up a crucial aspect of molecular function that were missed by others. While it remains unclear how the impact of SAVs upon molecular function translates to the impact upon the organism, the inference from the micro- to the macro-level will remain obfuscated. Specialized *in silico* methods combined with experimental deep scanning might bring about more clarity in the future. Until then, we remain with very surprising findings for the impact of sequence variation upon molecular function.

Methods

Data variants (SAVs). Our work focused entirely on sequence variants that alter a single amino acid in the protein. We referred to those as **SAV** (single amino acid variant; abbreviations found in the literature for the same include: nsSNV, nsSNP, and SAAV). We analyzed the following subsets separately.

OMIM. Set of disease-causing variants reported in OMIM[4] as extracted from SNPdbe[25]. SNAP2 scores were calculated for 5,661 SAVs in 1,547 unique protein sequences. The scores reported for OMIM in all figures were calculated with a special version of SNAP2 trained without using OMIM and HumVar[16] SAVs. Additionally, we generated a second set of SNAP2-scores for the above mentioned 5,611 OMIM SAVs through cross-validation. For this purpose, we retrained SNAP2 on the full training set (including OMIM and HumVar SVAs) holding out a small subset of OMIM SAVs as test-set in each iteration of the cross-validation (Supplementary Fig. S1).

60KE. SAVs reported by the Exome Aggregation Consortium (ExAC) at the Broad Institute reporting SAVs for 60,706 exomes[3]. We extracted all SAVs from ExAC release 0.3.1 labeled as 'missense_variant' and 'SNV' in the 'CSQ' information field. The resulting total was 10,474,468 SAVs; for 7,599,572 of these SNAP2 could predict the impact on molecular function. 40,446 were classified as common (LDAF $\geq$ 0.05), 28,789 as uncommon (0.01 $\leq$ LDAF $<$ 0.05), and 7,530,337 as rare (LDAF $<$ 0.01).

1KG. SAVs in human reported by the 1000 Genomes Project[12,13]. In particular, we included SAVs labeled as NON_SYNONYMOUS retrieved from the CADDv1.3 dataset[17]. 1KG variation in CADDv1.3 is based on 1000 Genomes Project Phase1 release v3.20101123. This set contained 292,848 variants, of which SNAP2 scores could be obtained for 268,115 SAVs (91.6%) in 19,696 sequence-unique proteins. Almost all missing SNAP2 predictions originated from problems with the underlying SIFT runs. Less than 1% of the SAVs were not predicted due to the SNAP2 limitation to exclude proteins with over 6,000 residues. Out of the 268,115 variants 20,352 were classified as common (LDAF $\geq$ 0.05), 30,543 SAVs as uncommon (0.01 $\leq$ LDAF $<$ 0.05), and 217,220 variants as rare (LDAF $<$ 0.01).

Inter-species orthologs. Orthologs were extracted from ENSEMBL Genes 73 entries (release Sep. 2013) using the Biomart[26] interface. To determine homology, ENSEMBL uses EnsemblCompara GeneTrees[27]. Pairs of human

proteins (hg19) and other species' orthologous proteins were aligned using the Needleman-Wunsch EMBOSS implementation[28] with default parameters (BLOSUM 62, gap open = 10, gap extend = 0.5). All alignments with PIDE < 70% (excluding gaps) were discarded. For all remaining pairs of aligned proteins, every amino acid variant was considered a SAV. Each SAV was evaluated in the context of the non-human sequence such that the residue position with a difference was mutated to the human amino acid. For multiple orthologs of the same protein (*e.g.* FoxP in Fly to FoxP1/2/3 and 4 in human) the highest ungapped sequence identity alignment was chosen to extract the SAVs. We avoided bias by excluding all variants that were used for SNAP2 training. The numbers of SAVs in resulting sets were as follows: chimp – 95,624 SAVs in 14,361 proteins; mouse – 379,795 SAVs in 12,616 proteins; fly – 16,403 SAVs in 364 proteins.

"House-keeping". This was a subset of all proteins that were considered as orthologs in cross-species comparisons (human-X). For the main figure (Fig. 3), we compared predictions for human-chimp variants, human-mouse variants and 1KG SAVs to the subset of variants from orthologs common to all species. Additionally, we further restricted the comparison to proteins for which SAVs were available in all three orthologs of the protein in three species (e.g. excluding cases for which human-chimp had no SAVs between protein X_{human} and X_{chimp}, while human-mouse had SAVs between proteins X_{human} and X_{mouse}). Number of SAVs in resulting sets: human–133,500 SAVs in 8,535 proteins; chimp 46,288 in 8,147 proteins; mouse 309,516 SAVs in 8,235 proteins.

Denisovan. Amino acid differences between *Homo sapiens* and *Denisova hominin* were extracted from the data published by the groups of Svante Pääbo and Janet Kelso[29]. In our implementation, the *denisovan* amino acid (ancestor) was introduced into the corresponding position in the human protein and then mutated back to the human reference. For example, if the human sequence Xp contained amino acid L at position 42 and the corresponding *denisovan* residue was V, we created a protein sequence Xp', equivalent to Xp except for V at position 42. For simplicity, we referred to these variants as SAVs because technically they originated from the same "edit procedure", i.e. the change of a single amino acid. We then predicted the effect of the Xp' SAV V42L. This set contained 236 SAVs in 292 proteins.

Random. We created two 'random' sets of human SAVs. Both sets are a random sample (n = 268,511) of two 'supersets': (1) all 19-non-native SAVs; (2) all SNV-possible SAVs, i.e. SAVs that can be reached by mutating one single nucleotide, which is in turn a subset of (1). The size of the random sample was chosen to be 268,511 to be in line with the size of the 1KG SAV set.

Methods

Effect scores for SAVs in all sets were computed using SNAP2[21], an improved version of SNAP1[15]. SNAP2 uses a protein sequence and a list of SAVs as input to predict the effect of each substitution on the protein molecular function. The prediction scores range from −100 for fully neutral to +100 for strong effect. In its original form, SNAP scores served only as reliability index, where the confidence that the assigned class (neutral/effect) for a specific mutation is higher for SNAP scores closer to the −100 and +100 maxima. However, the score also correlates clearly with the severity of the effect[15], i.e. scores slightly above 0 are not as severe as those closer to +100. Interestingly, for mis-predicted neutral SAVs (i.e. SAVs with known effect, incorrectly predicted as neutral), the scores closer to 0 indicate higher than for those scoring closer to −100. For all SAVs in all sets, SNAP2 scores were computed.

For a binary projection (effect/neutral), SNAP learned to optimize the experimental annotations such that SNAP2-score ≤0 implied neutral and SNAP2-score >0 implied effect. The experimental evidence for effect is much more reliable than evidence for neutral. Therefore, the point that optimally fits the known data might not describe reality best[21]. This discrepancy calls for introducing additional thresholds for the binary distinction effect/neutral. All of those are arbitrary, i.e. are meaningful only to highlight trends. The raw data is the full spectrum of the prediction (−100 to +100). Therefore, we showed this full spectrum in all figures.

To simplify the communication of trends, we defined three example thresholds for SNAP2 in addition to the default of 0: (1) SNAP2-score ≤−42 (below this point we expect 85% of all predictions to be correctly predicted as neutral, Fig. 1– lower panel), (2) SNAP2-score ≥+50 (above this point we expect 85% of all predictions to be correctly predicted as effect, Fig. 1– lower panel), (3) SNAP2-score ≥+75 (chosen because above this point 50% of all OMIM SAVs are correctly predicted; effect prediction accuracy of 88%). Accuracy values are based on the number of variants that are reported above (for effect) and below (for neutral) the respective thresholds, i.e. for the effect prediction accuracy at the threshold ≥+50 SAVs with predicted SNAP2-scores between +50 to +100 were considered and are predicted correctly with an accuracy of 85%.

Other methods. PolyPhen-2[16] measures the likelihood whether a SAV is pathogenic or benign based on family conservation and structural information. The method has been trained on disease-related SAVs from HumDiv and HumVar, and on SAVs between orthologs in human and mammal considered as neutral. SIFT[18] uses family conservation to measure the probability for a SAV to be deleterious or tolerated (as expected from alignment). CADD[17] aims at distinguishing between "variants that survived natural selection" and simulated mutations. CADD has been trained on mutations between an inferred human-chimp common ancestor and the human reference genome (excluding common SAVs in 1KG; however, including those where the reference allele carries the ancestral variant and the derived 1KG allele occurs in more than 95% of the population). Simulated variants are created through a process of mutating nucleotides based on parameters extracted from the inferred human-chimp ancestor and sequence alignments between multiple primate species. Scores and classification for

PolyPhen-2, CADD and SIFT were extracted from the CADDv1.3 data set. PolyPhen-2 effect scores were available for 260,039 SAVs (common: 19,272 - uncommon: 29,281 - rare: 211,486). CADD scores were available for all SAVs. SIFT scores were available for 264,565 SAVs (common: 19,704 - uncommon: 29,898 - rare: 214,963).

Estimated method performance. SNAP2[15,21] has been estimated to perform at a sustained positive accuracy (TP/(TP + FP)) 78% and a negative accuracy (TN/(TN + FN)) of 77% (at the default SNAP-score of 0, Fig. 1 lower panel). Prediction accuracy rises with increasing thresholds (effect: toward SNAP-score +100 on the right in Fig. 1; neutral: toward SNAP-score −100 on the left). For the example thresholds SNAP2-score >+50 and SNAP2-score >+75 the accuracy of predicted functional effect increased to 85% and 88% respectively, whereas at SNAP2-score <−42 the expected accuracy for predicting neutral increased to 85%.

From molecular function to disease. SNAP1 correctly predicted over 80% of the SAVs for which a monogenic disease (OMIM[4]) was known four years ago (1,105) at its default threshold[25]. For this work, we repeated this analysis with a fivefold larger set of monogenic disease-causing variants in OMIM (5,611) using a version of SNAP2 that had not used OMIM-type of variants for training at all (Fig. 1 *disease*). In fact, for all versions of SNAP the OMIM SAVs that had not been used for method development were predicted at much higher average scores than variants in our training set[21]. The SNAP2-score primarily serves as a reliability index. However, we previously demonstrated[6,15] that effect strength also correlates well with the SNAP-score, e.g. more reliably predicted effect SAVs tend to have stronger effect.

Standard error estimates. The standard errors of the mean were estimated by bootstrapping the SNAP2-scores of the respective datasets. The datasets were resampled with replacement a hundred times, calculating 100 means of the SNAP2-score distribution. Standard error was estimated as the standard deviation of the means.

References

1. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74, doi:10.1038/nature15393 (2015).
2. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682, doi:10.1016/S0140-6736(12)61480-9 (2012).
3. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291, doi:10.1038/nature19057 (2016).
4. Hamosh, A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2004).
5. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics* **9**, 356–369, doi:10.1038/nrg2344 (2008).
6. Bromberg, Y., Kahn, P. C. & Rost, B. Neutral and weakly nonneutral sequence variants may define individuality. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 14255–14260, doi:10.1073/pnas.1216613110 (2013).
7. Stitziel, N. O., Kiezun, A. & Sunyaev, S. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome biology* **12**, 227, doi:10.1186/gb-2011-12-9-227 (2011).
8. Cline, M. S. & Karchin, R. Using bioinformatics to predict the functional impact of SNVs. *Bioinformatics* **27**, 441–448, doi:10.1093/bioinformatics/btq695 (2011).
9. Mah, J. T., Low, E. S. & Lee, E. In silico SNP analysis and bioinformatics tools: a review of the state of the art to aid drug discovery. *Drug discovery today* **16**, 800–809, doi:10.1016/j.drudis.2011.07.005 (2011).
10. Thusberg, J., Olatubosun, A. & Vihinen, M. Performance of mutation pathogenicity prediction methods on missense variants. *Human mutation* **32**, 358–368, doi:10.1002/humu.21445 (2011).
11. Reeb, J., Hecht, M., Mahlich, Y., Bromberg, Y. & Rost, B. Predicted molecular effects of sequence variants link to system level of disease. *PLoS computational biology* **12**, e1005047, doi:10.1371/journal.pcbi.1005047 (2016).
12. Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073, doi:10.1038/nature09534 (2010).
13. Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65, doi:10.1038/nature11632 (2012).
14. Efron, B. & Tibshirani, R. J. *An introduction to the bootstrap*. Vol. 57 (Chapman & Hall; CRC Monographs on Statistics & Applied Probability (Book 57), 1993).
15. Bromberg, Y. & Rost, B. In SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res* **35**, 3823–3835, doi:10.1093/nar/gkm238 (2007).
16. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248–249, doi:10.1038/nmeth0410-248 (2010).
17. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**, 310–315, doi:10.1038/ng.2892 (2014).
18. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073–1081, doi:10.1038/nprot.2009.86 (2009).
19. Goldberg, T. *et al.* LocTree3 prediction of localization. *Nucleic Acids Res* **42**, W350–355, doi:10.1093/nar/gku396 (2014).
20. Miller, M., Bromberg, Y. & Swint-Kruse, L. Computational predictors fail to identify amino acid substitution effects at rheostat positions. *Scientific Reports* **7**, 41329, doi:10.1038/srep41329 (2017).
21. Hecht, M., Bromberg, Y. & Rost, B. Better prediction of functional effects for sequence variants. *BMC Genomics* **16** Suppl 8, S1, doi:10.1186/1471-2164-16-S8-S1 (2015).
22. Starita, L. M. *et al.* Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* **200**, 413–422, doi:10.1534/genetics.115.175802 (2015).
23. Hopf, T. A. *et al.* Quantification of the effect of mutations using a global probability model of natural sequence variation. *arXiv:1510.04612 e-prints* (2015).
24. Hopf, T. A. Phenotype prediction from evolutionary sequence covariation Dr. rer. nat. (PhD) thesis, TUM (2015).
25. Schaefer, C., Bromberg, Y., Achten, D. & Rost, B. Disease-related mutations predicted to impact protein function. Disease-related mutations predicted to impact protein function. *BMC Genomics* **13** Suppl 4, S11, doi:10.1186/1471-2164-13-S4-S11 (2012).
26. Kasprzyk, A. In BioMart: driving a paradigm change in biological data management. *Database* (*Oxford*) Vol. 2011, bar049–bar049, doi:10.1093/database/bar049 bar049 (2011).

27. Vilella, A. J. *et al.* In EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research* Vol. 19, 327–335, doi:10.1101/gr.073585.107 (2008).
28. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276–277, doi:10.1016/S0168-9525(00)02024-2 (2000).
29. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226, doi:10.1126/science.1224344 (2012).

Acknowledgements
Thanks at the TUM to Tim Karl and Laszlo Kajan for invaluable help with hardware and software; to Marlena Drabik and Inga Weise for support on many levels; to Shaila Roessle, Christian Schaefer, Dominik Achten, Rebecca Kassner (all TUM) and Chengsheng Zhu (Rutgers) for providing valuable input during the beginning of our quest; thanks to Boyang Yin and Dietrich Wins for their help in bringing the topic into TEDxTUM. We are particularly grateful to Janet Thornton (EBI Hinxton) for a vivid dialogue that started this work, and to Janet Kelso (MPI Leipzig) and Martin Kircher (Wash U Seattle) for the discussions and help that shaped the results. We are also grateful for the critical patience of the anonymous reviewers and the Associate Editor who all helped crucially improve work and paper. The work was partially supported by the Alexander von Humboldt foundation through the German Ministry for Research and Education (BMBF: Bundesministerium fuer Bildung und Forschung) and by the German Research Foundation (DFG) and the Technische Universität München within the funding programme Open Access Publishing. YM and YB were supported by a grant from the National Institute of General Medical Sciences (1U01GM115486-01). Last but not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these resources.

Author Contributions
Y.M. collected data sets and conducted the analysis; M.H. developed SNAP2, contributed data sets, and helped in the analysis. J.R. provided the analysis of the OMIM set, the performance estimates for SNAP2, and maintains the SNAP2 software. M.S. provided the subcellular localization predictions. B.R. conceived the concept and supervised the work. Y.B. developed SNAP, provided crucial advice that helped to avoid traps, and supervised last stages of the work. TAPDB helped crucially in the original data collection, and thereby affected our initial focus on the 1KG data.

Additional Information
Supplementary information accompanies this paper at doi:10.1038/s41598-017-01054-2

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supporting online material (SOM) for:

Common sequence variants affect molecular function more than rare variants?

Yannick Mahlich [1, 2, 3*], Jonas Reeb [1], Maximilian Hecht [1], Tjaart Andries Petrus De Beer [4], Yana Bromberg [2, 3] & Burkhard Rost [1, 3, 5]

1 Computational Biology & Bioinformatics - i12, Informatics, Technical University of Munich (TUM), Boltzmannstrasse 3, 85748 Garching/Munich, Germany
2 Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ 08901, USA
3 Institute of Advanced Study (TUM-IAS), Lichtenbergstr. 2a, 85748 Garching/Munich
4 European Molecular Biology Laboratories, European Bioinformatics Institute (EMBL-EBI), Welcome Trust Genomes Campus, Cambridge, Cambridgeshire, United Kingdom
5 Institute for Food and Plant Sciences WZW – Weihenstephan, Alte Akademie 8, Freising, Germany
* Corresponding author: Yannick Mahlich (ymahlich@bromberglab.org)

Table of Contents for SOM

Material

Fig. S1:

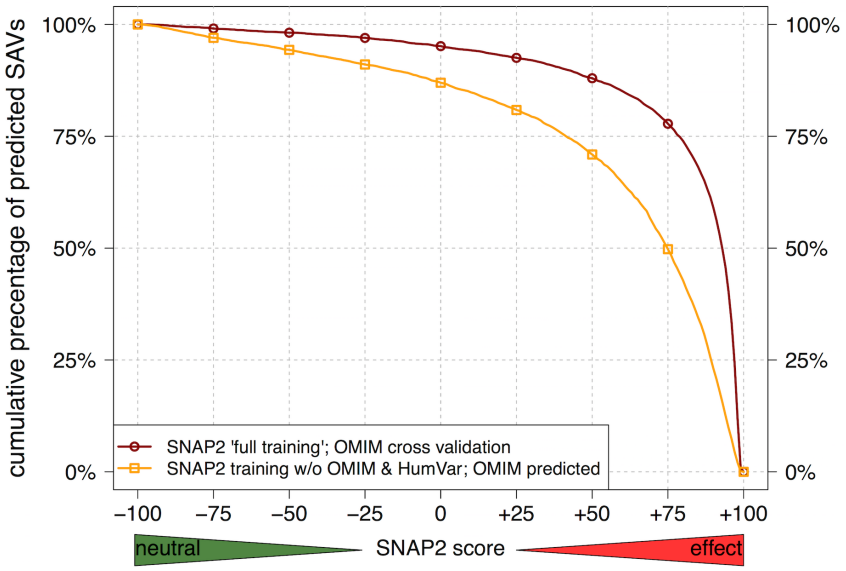


Fig. S1: Training on disease-causing SAVs improves prediction for those. The curves refer to monogenic SAVs from OMIM directly implied in disease (c.f. Method section in main manuscript). Here, we compared two different SNAP2 versions: the first (dark red, circles) was trained using SAVs from OMIM and HumVar. Results were obtained through cross-validation, i.e. the SAVs shown here were not in proteins sequence-similar (HVAL>0 AND PSI-BLAST EVAL<10-3) to proteins used for training. In contrast, the SNAP2 version labeled “SNAP2 training w/o OMIM & HumVar” (orange, squares) never used any disease-impact SAV for training. The difference between the two versions of SNAP2 demonstrated how much training on disease-causing SAVs helps to predict (dark red much higher than orange). In turn this suggested that methods trained on features relevant to disease-causing SAVs capture different aspects than methods not using such SAVs.

Fig. S2:

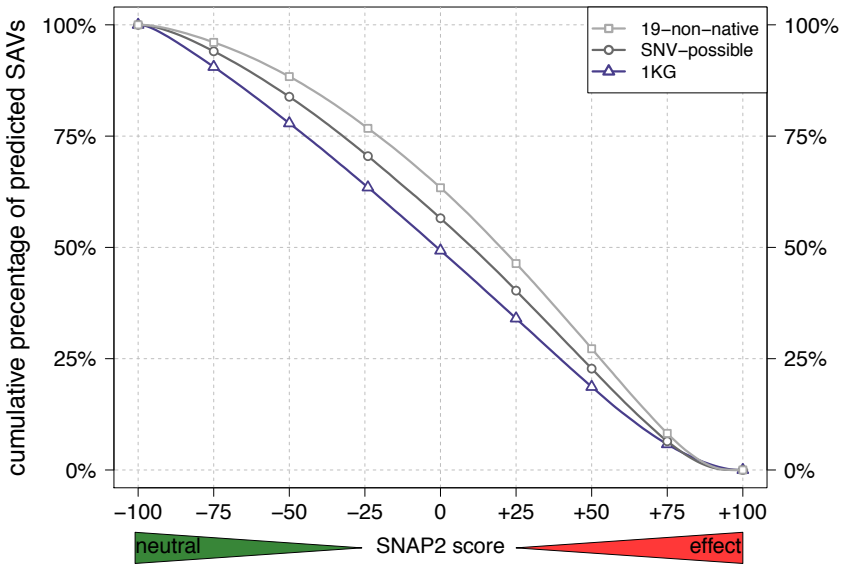


Fig. S2: 1KG SAVs differ from random SAVs. All curves show predictions from the standard SNAP2 version. We compared the predictions for all SAVs in the 1KG set of healthy people (blue curve, triangles) to random subsets of all possible SAVs that we could generate *in silico*. We have two options to realize "all possible": replace all native amino acids at all residue positions by (1) all 19-non-native amino acids ("19-non-native"), and (2) all amino acid substitutions that can be reached by a single nucleotide variant ("SNV-possible"). We then selected a subset of these large data sets with the same number of SAVs as in the 1KG set. This gave the sets random-19-non-native (light gray, squares) and random-SNV-possible (dark gray, circles). Although the difference between all curves appeared to be minor, they were statistically significant. We tested the significance by the two-sample Kolmogorov-Smirnov (KS) test (1KG vs. SNV-possible, D = 0.072; 1kg vs. 19-non-native, D = 0.141; 19-non-native vs. SNV-possible, D = 0.069; n,n' = 268115, estimated p-value < 2.2e-16 for all three KS-tests). Since the standard error of mean (SEM) for all three SNAP2 score distributions was < 0.5, error bars and confidence intervals were omitted.

Fig. S3:

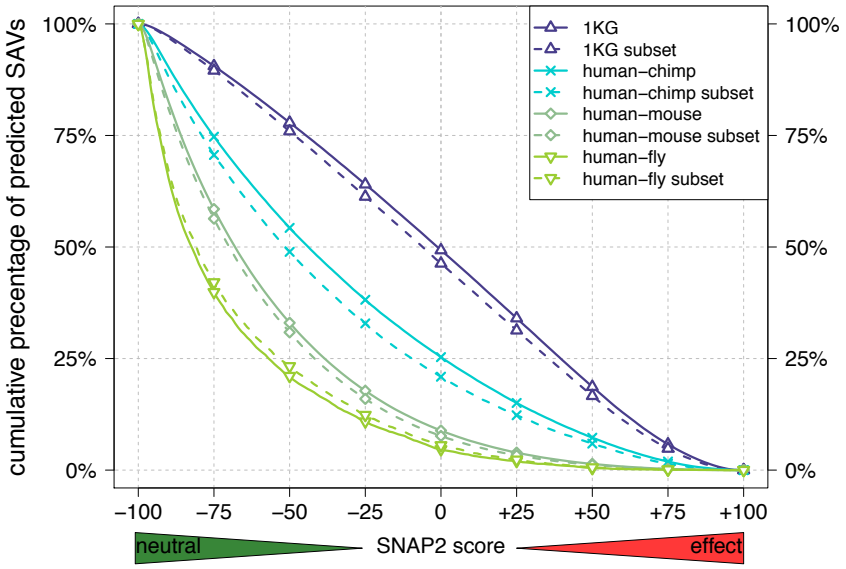


Fig. S3: Orthologs across four species same trend as entire proteomes. While Fig. 1 in the main manuscript compares cross-species and 1KG SAVs for all proteins (and thereby compares different sets of proteins), and Fig. 2 compares a much smaller data set of orthologs shared between three organisms and 1KG, here we compiled results for SAVs that is less restrictive. The main restriction, that a protein used to extract SAVs has to have an ortholog in the other species still applies, however, we do not exclude proteins for which in one or more of the human-X inter-species comparisons do not yield SAVs. To give an example: The orthologous proteins X_{human}, X_{chimp}, X_{mouse} (and X_{fly}) are only considered for SAV extraction in Fig. 3 if all proteins contain SAVs. Here we include SAVs from those proteins into the analysis even if this does not hold up, *e.g.* X_{chimp} does not contain SAVs, when compared to X_{human}; on the other hand, X_{mouse} and X_{fly} do. This figure confirms the main trend: inter-species variants are shifted to the left (less effect) with respect to the 1KG SAVs between healthy people and the shift is higher the more divergent an organism from human (e.g. human-chimp shifted less to the right than human-fly).

Fig. S4:

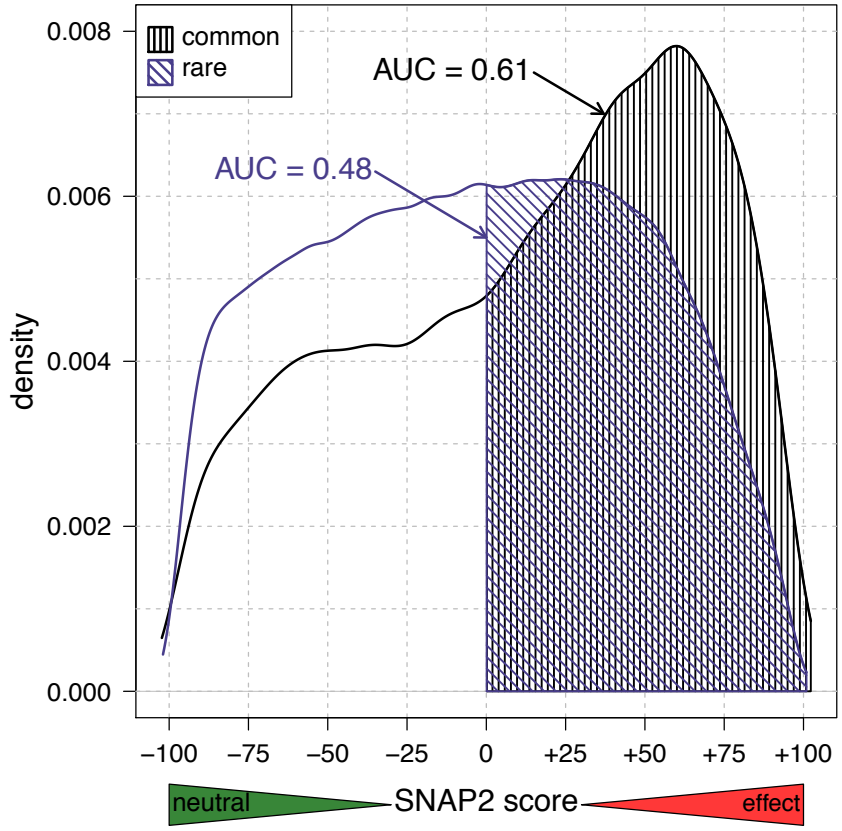


Fig. S4: SNAP2 predicts more common than rare SAVs to be effective. Displayed are the density curves of predicted SNAP2 scores for common (black) and rare (dark blue) 1KG SAVs. Examining the area under the curve for effect scores (SNAP score >= 0, common shaded black vertical, rare shaded dark blue diagonally) of both curves it is clearly visible that SNAP predicts a larger fraction of common SAVs to be effective than rare SAVs (AUC 0.61 vs. 0.48).

Fig. S5:

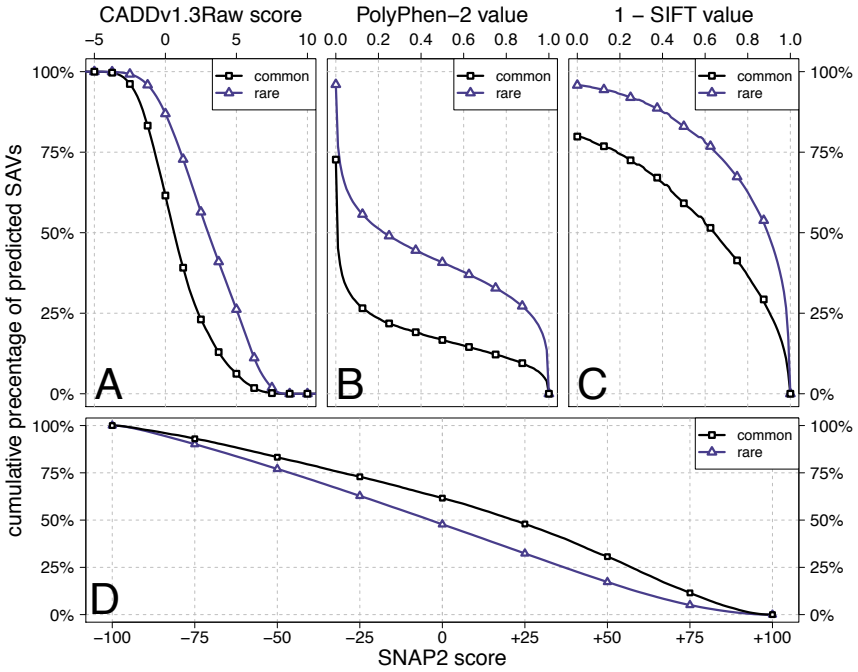


Fig. S5: CADD, PolyPhen-2 and SIFT predict a higher fraction of rare variants to be functionally effective than common variants. The cumulative percentages (read as Y% of SAVs are predicted to have higher score than X) for predicted SAVs across the range of scores for each method displayed a clear trend. CADD (A), PolyPhen-2 (B) and SIFT (C) predict rare variants (dark blue, triangles) to impact function of the protein more often than common variants (black, squares). This is in stark contrast to our predictions for common and rare SAVs by SNAP2 (D).

Fig. S6:

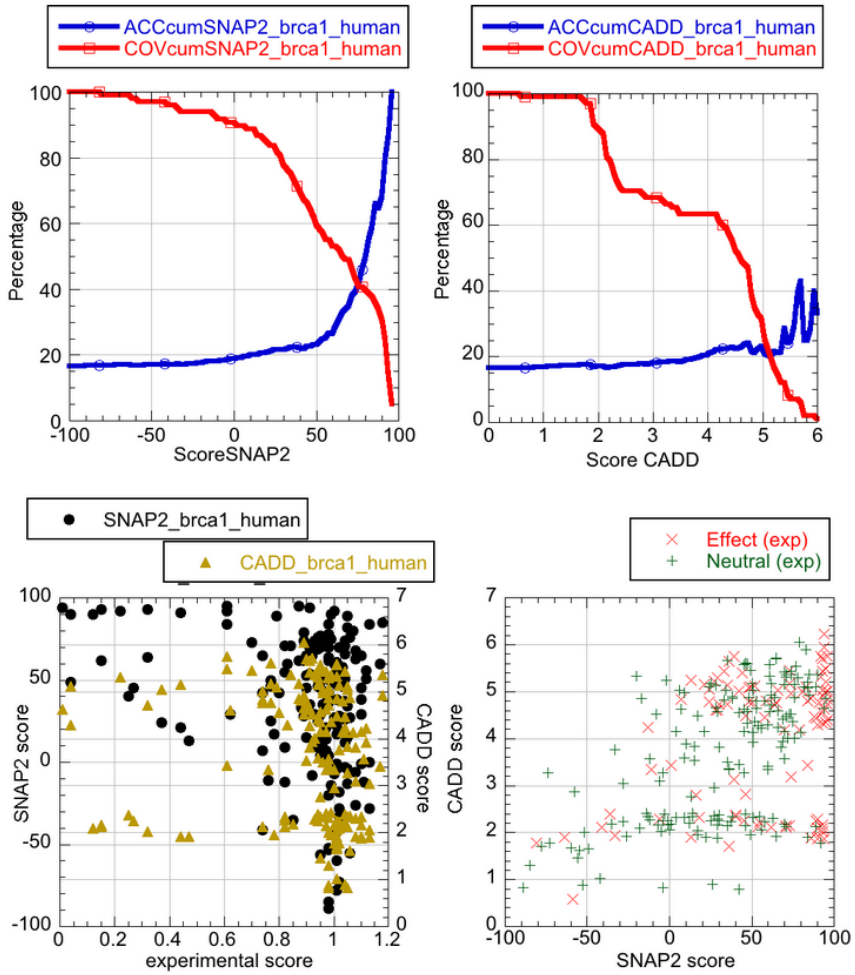


Fig. S6: SNAP2 captured molecular function better than CADD for deep scanning BRCA1 dataset. The lower left panel showed the experimental score from an experimental deep scanning testing the impact of SAVs upon molecular function (xxbr: put in quote of experiment here). SNAP2 results marked by black disks; the SNAP2-scores on the left y-axis ranged from strongly predicted as neutral (-100, bottom of the y-axis) to strongly predicted as effect (+100, top of the y-axis). The yellow-brown triangles gave results for CADD; CADD-scores on the right y-axis of the left panel. Visually, it seems that

SNAP2 correlated much better with the experimental score than CADD. The lower right panel tried to simplify by projecting the raw experimental score upon a binary classification, i.e. effect (red x) vs. neutral (green +). Using any threshold for SNAP2 and CADD scores (defaults: effect: SNAP2-score>0 and CADD>3), SNAP2 captured the simplified experimental impact of sequence variation upon molecular function better than CADD. The upper panels explicitly compile the performance for SNAP2 (upper left panel; x-axis SNAP2-score) and CADD (upper right panel; x-axis CADD-score). The y-axis of the upper panels showed cumulative percentages (blue line: accuracy=correctly predicted as effect/all predicted as effect; red line: coverage=correctly predicted as effect/all observed as effect). For instance, while for about 20% of the most strongly predicted SAVs SNAP2 reached over 80% accuracy (SNAP2-score>90: red curve ~20%, green curve ~80%), while CADD saturated at 20% accuracy for the same coverage (CADD-score>5: red curve ~20%, green curve ~20%).

Note that CADD never attempted to predict the impact of sequence variation upon molecular function. This figure gives one particular example proving that SNAP2 achieves the goal it was optimized for better than CADD the goal it was NOT optimized for. Although this was not much of a surprise, in light of the reverse of the predicted effect for rare and common SAVs, this confirmation constituted important evidence. The BRCA1 data was chosen for the only reason that it was the first large experimental deep scanning experiment that was made available to us. In a separate analysis, we have recently compiled additional data sets that confirm these findings for larger data sets (Theresa Wirth, TUM, in preparation); similar findings have been submitted by others (Thomas Hopf, Chris Sander & Debbie Marks, Harvard University)

Supplementary Note – SNAP2 training data

SNAP2 was trained on a set of ~100k mutations. The majority of these (~52%) were obtained from experimental effect annotations recorded in the Protein Mutant Database (PMD [2]). OMIM [3] and HumVar [4] effect variants accounted for another ~22% of the data and a set ~26% putative neutral variants was derived from alignments of enzymes with identical EC (Enzyme Commission) numbers [5]. Note that some of the latter might not be actually neutral due to compensating mutations (*i.e.* other sequence differences in same alignment), as well as due to differences in levels of ortholog activity between species. To avoid introducing a bias in this study towards predicting variants in orthologous sequences as neutral, we excluded all variants that were used for SNAP2 training from comparison.

References for Supporting Online Material

1. Bromberg, Y., Kahn, P.C. & Rost, B. Neutral and weakly nonneutral sequence variants may define individuality. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 14255-14260 (2013).
2. Kawabata, T., Ota, M. & Nishikawa, K. The Protein Mutant Database. *Nucleic acids research* **27**, 355-357 (1999).
3. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33**, D514-517 (2005).
4. Capriotti, E., Calabrese, R. & Casadio, R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* **22**, 2729-2734 (2006).
5. Webb, E.C. Enzyme Nomenclature 1992. Recommendations of the Nomenclature committee of the International Union of Biochemistry and Molecular Biology., Edn. 1992. (Academic Press, New York; 1992).

6 Conclusion & Outlook

As laid out in this work, gaining a better understanding of "what makes microbes tick" is a difficult, time and resource consuming task. Nevertheless it is important to answer this question to further future applications in biomedical research. Modern assessment methods just start to scratch the surface of what is possible. The *fusion* protocol and *fusion*DB (described in **Chapters 3** and **4**) is a step towards answering exactly this question. We can use this knowledge to improve our understanding of how species functionally relate to each other. Additionally, one of *fusion*DB's future applications can be to serve as starting point to investigate microbial communities. By selecting representative proteins we would be able to generate a reference database for tools like mi-faser. This could in turn be used to evaluate the functional properties of microbial communities. Furthermore we would be able to combine the functional property knowledge of a community, with the individual functional profile information of *fusion*DB. This combination of information could lead to discover symbiotic relations between microbial species. An example for a specific pattern to look for would be complete metabolic pathways present in the microbial community, but not in individual organisms. In other words, different species "coming together" to complete a otherwise incomplete pathway.

At the date of writing (April 2019) NCBI GenBank contains more than 12,000 complete bacterial genomes. Incorporating those into *fusion*DB will lead to a much better resolution of any analysis. However as also described in **Chapter 3** updating *fusion*DB to contain all those organisms is a difficult task.

The first hurdle to overcome was to find a way to cut down the necessary compute time to establish functional similarities between any two proteins. To this end I developed HFSP (see **Chapter 2**). HFSP uses MMSeqs2 to generate many versus many sequence alignments in a much faster fashion, while fundamentally following the same principles established by PSI-BLAST. HFSP is over 40-fold faster in comparison to traditional methods, establishing functional similarities by examining protein homology. I evaluated the functional annotations of multiple subsets of Uniprot during the development of HFSP, and came to the conclusion that HFSP will be useful not only for the purposes of *fusion*, but will also be able to uncover a large portion of potential function miss annotation in public databases.

The second key concept of fusion was described in **Chapter 3**. Using the established functional similarities between proteins to generate a protein similarity network results in very large directed graphs. The network size very quickly reaches tens of millions of nodes and tens to hundreds of billions of edges. Clustering algorithms for networks of this size are not widely available. Algorithms that claim to work for networks of this size are often highly specialized on individual types of networks. MCL (Markov Clustering) which was used in the proof of concept for *fusion* scales poorly with increasing numbers of nodes

in the network. HipMCL which is a high performance compute cluster implementation of MCL is able to cope with much larger networks than MCL. However, HipMCL is still prone to the same scaling issues present in MCL. (Hip)MCL therefore was no longer a viable option as clustering algorithm.

In recent years multi dimensional scaling algorithms have become increasingly popular as clustering algorithms. Originally devised as graph visualization method, t-SNE could potentially solve my problem. The graph can be interpreted as multidimensional data. Each outgoing edge of a node (similarity to another protein) serves as a distance in one dimension. The result is as many dimensions as nodes present in the similarity graph. The initial embedding of the graph can then be used to re-embed nodes initially omitted. Looking forward, the ultimate goal would be to generate an easy-to-update scheme. t-SNE could provide exactly this with the possibility of re-embedding. Novel proteins not present in the clustering could be added, by establishing functional similarity to all existing nodes, and using the re-embedding technique.

The massive influx of newly and fully sequenced bacterial organisms in the past 5 years, clearly demonstrates the need for incremental updates. This could also be especially interesting going forward for comparative studies of bacterial strains. Comparing many strains of the same bacterium can for example give insight into variations of phenotypes. One field where this will be very important, is understanding the development of resistances to antimicrobials or vaccines.

Additionally, as more microbial organisms are fully sequenced, we can also increase our understanding of the contribution of individual organisms to a larger community. Using the t-SNE embedding I will be able to identify representative proteins for the functional units in *fusion*DB. As briefly touch upon earlier, I can use those representatives as a reference database to identify functional properties of whole microbial communities. If regarding the community of organisms as a "meta-organism" I will therefore also be able to detect pathways of molecular function that only emerge if the community is investigated as a whole.

As laid out during this work, the incredible speed at which new species and strains are sequenced offers a treasure throve of available data. It should also be clear however, that the race between data and tools to analyze the data will not be one that is over soon. I offered solutions to process the current amount of data in a time and resource sensitive manner. Yet I see this as being only a temporary solution. There clearly is a strong need to continue the development of efficient algorithms, especially in the realm of network analysis and clustering.