Characterizing metagenomic novelty with unexplained protein domain hits

Thomas Lingner, Peter Meinicke

Department of Bioinformatics
University of Göttingen
Goldschmidtstr. 1
37077 Göttingen
thomas@gobics.de
peter@gobics.de

Abstract: In metagenomics, the discovery of functional novelty has always been pursued in a gene-centered manner. In that way, sequence-based analysis has been restricted to particular features and to a sufficient length of the sequences. We propose a statistical approach that is independent from the identification of single sequences but rather yields an overall characterization of a metagenome. Our method is based on the analysis of significant differences between the functional profile of a metagenome and its reconstruction from a combination of genomic profiles using the Taxy-Pro mixture model. Here, protein families with a large proportion of domain hits that cannot be explained by the model are interesting candidates for the exploration of metagenomic novelty. The results of three case studies indicate that our method is able to characterize metagenomic novelty in terms of the protein families that significantly contribute to unexplained domain counts. We found a good correspondence between our predictions and the discoveries in the original studies as well as specific indicators of functional novelty that have not yet been described.

1 Background

Metagenomics rigorously extends the exploration of microbial life beyond the borders of culturable organisms. Therefore the vast amount of metagenomic sequence data potentially provides a gold mine for the discovery of novel genes. Several approaches have been proposed for gene mining on metagenomic data. The corresponding methods aim at the identification of candidate sequences using either gene neighbourhood context [HSD+07], protein domain architecture [MCR10], clustering of open reading frames [YSR+07] or phylogenetic trees of known protein families [KTZ+07]. The different methods have specific strengths and limitations, and they can be combined to identify interesting candidate sequences for further analysis [SDL+09]. However, none of these methods is capable to characterize the novelty of metagenomic data beyond the identification of a number of sequences that can be detected according to the above-mentioned criteria.

We here present an approach for the characterization of metagenomic novelty based on

the Taxy-Pro mixture model [KALM13]. Taxy-Pro performs a reconstruction of the functional profile of a metagenome using a linear combination of genomic reference profiles of known organisms. The reconstruction error in terms of the fraction of protein domain hits unexplained (FDU) can be analyzed with respect to the most contributing domain families. We propose that all domain families that are overrepresented in the metagenome when compared to its genomic reconstruction are candidates for the characterization of functional novelty. In contrast to all previous methods, our approach does not detect single candidate sequences but rather predicts the functional categories that contribute to metagenomic novelty. Therefore our method can be applied prior to the identification of candidate sequences to narrow down the search space for gene mining. Furthermore, for the first time we are able to identify and select metagenomes that are most promising in terms of the predicted novelty.

We conducted a number of case studies on real-world metagenomic datasets to evaluate the potential of our approach. The results indicate that our method identifies protein domain-specific functional novelty that is known or suspected with respect to particular environments and also show how hypotheses on novel genes can be obtained from large metagenomic dataset collections.

2 Materials and methods

Our approach to characterization of novelty in metagenomes is based on the statistical analysis of overrepresented Pfam domain families in a metagenome's functional profile as compared to a reconstruction from genome profiles by a mixture model. We analyze our method in terms of case studies for functional description of novelty in real-world metagenomic datasets from different environments. In the following, we will describe the datasets and methods that we used for our case studies.

2.1 Metagenomic datasets

From the 'Cow rumen' metagenome collection [HSE+11] we selected the largest sample (EBI accession SRR094415, downloaded from www.ebi.ac.uk/metagenomics/sample/ as of April 2014), which contains 11,334,156 Illumina GAIIx paired-end reads of 2 x 125 bp average length (2.8 Gbp total).

The 'Sediment' dataset ([BWTH11], EBI acc. SRS004796) has been collected from marine sediments at the Brazos-Trinity Basin in the western Gulf of Mexico and contained after quality control 402,793 reads with ≈264 bp length (106.4 Mbp total). The sediment metagenome was sequenced using the Roche 454 GS FLX technology.

The Human Microbiome Project (HMP, [PGG+09]) provides an extensive collection of samples from human body sites of healthy individuals for large-scale comparative studies. For our evaluation, we used 750 data samples of the HMP as described in detail in [KALM13]. Briefly, the samples have been taken from five major human body sites ('Uro-

genital tract', 'Oral', 'Gastrointestinal tract', 'Skin', 'Airways') that encompass up to nine subsites. Sequencing was performed using the Illumina HiSeq 2000 platform resulting in paired-end reads of about 2 x 100 bp average length.

2.2 Taxonomic profiling

The Taxy-Pro approach to taxonomic profiling of metagenomes is based on mixture models of functional profiles in terms of protein domain frequencies and is described in detail in [KALM13]. Briefly, a metagenome's profile $\mathbf{y}$ is estimated from Pfam domain counts and then reconstructed by a linear combination of genomic profiles $\mathbf{x_1}, ..., \mathbf{x_N}$ from a reference database containing several thousand domain signatures of microbial genomes according to $\hat{\mathbf{y}} = \sum_i^N w_i \mathbf{x_i}$. The coefficients w_i of the linear combination then represent estimated contributions of particular organisms to the explanation of the metagenomes' signature. The estimation of the coefficients can in principle be implemented as a linear least-squares regression problem. However, due to the unit-sum and positivity contraints on the weights one has to resort to quadratic programming for a solution. Furthermore, because many of the protein domain frequencies are estimated from small counts the errors can be far from normal deviates making a squared-error criterion less adequate. For these reasons we use the classical EM-optimization in case of protein domain features [KALM13]. As a unique feature of the Taxy-Pro method, the approximation error $\frac{1}{2}\sum_j^D |y_j - \hat{y}_j|$ of the mixture model provides a quality measure in terms of the fraction of domains unexplained (FDU), i.e. the entirety of the D domain-specific deviations of the metagenome's actual protein domain profile from its reconstruction. In this way, the FDU can reflect metagenomic novelty in terms of a lack of explanation according to the discrepancy between the observed domain profile and the mixture model.

To obtain the domain signatures we used the ultrafast protein classification (UProC) tool (uproc.gobics.de) in combination with version 27 of the Pfam protein family database [FBC+14]. The domain detection significance threshold was left at the default value (0.1% FPR). Depending on the average read length of the metagenomic dataset we applied the default ORF mode (length >200 bp) or the short read mode of UProC.

2.3 Unexplained domain counts from overrepresentation analysis

Our model is based on the assumption that the protein domain frequencies in a metagenome that cannot be well explained by genomic reference profiles are likely to be indicators of metagenomic novelty. This novelty, in general, results from unknown organisms in the microbial community that are missing in the database. In particular, genomes with unusual protein domain signatures that have no evolutionary close neighbors among the references will contribute to the deviations between the observed domain frequencies and their model-based approximation.

The domain-specific deviations of the mixture model $d_j = y_j - \hat{y}_j$ can be used to identify

over- ($d_j > 0$) or underrepresented ($d_j < 0$) domain families in the actual metagenome's functional profile as compared to its reconstruction. We here focus on the overrepresented families that correspond to domain counts that are significantly higher than the reconstructed counts of the model. To identify significantly overrepresented domain families, we analyzed the difference of domain frequencies for each family using a binomial test. Here, the relative domain frequencies $\hat{y}_j$ of the reconstructed profile are used as estimators of the event probabilities and the total number of hits n to Pfam domains in the metagenome are used as the number of draws in terms of Bernoulli trials. In order to avoid the singular case of zero event probabilities that would always result in significant overrepresentation, we used a unit pseudocount value for the estimation of probabilities. The value $P(c_j|n, \hat{y}_j)$ of the cumulative binomial distribution function calculated for the actual domain count value c_j then corresponds to the probability of observing up to c_j counts in n hits. The value $1 - P$ corresponds to the probability (p-value) of y_j being significantly larger as compared to $\hat{y}_j$. To account for multiple testing with $D >$14,000 Pfam families, we calculated p-values using the family-wise error rate (FWER) correction method [Hol79]. For selection of significantly overrepresented domains we used a p-value threshold of 0.05.

For our evaluation we use two domain-specific novelty indices: firstly, the number of expected unexplained domain counts (EUC), which we define as the rounded product of the domain-specific deviation d_j and the total number of domain counts n in the dataset. Furthermore, we use the original count (OC) value c_j associated with a domain to calculate the "novelty ratio" according to the fraction EUC/OC. We integrated the statistical overrepresentation analysis in the Taxy-Pro toolbox available at gobics.de/TaxyPro/.

2.4 GO enrichment analysis

The Pfam family descriptions usually characterize enzymatic, regulatory or structural properties of the associated domains on a very specific and non-standardized level. Conveniently, the Pfam database provides a mapping of a large number of domain families to the standardized vocabulary of the Gene Ontology (GO, [BLH+10]) database. In order to simplify the inspection of long lists of significantly overrepresented domain families w.r.t. to common properties, we perform an aggregation ("enrichment analysis") of Pfam domain hits into GO term categories. For this purpose, we calculate for each GO term $g_1, \ldots, g_M$ the sum of family-specific deviations for the domains for which a mapping to this GO term exist: $g_j = \sum_i^D d_i \times I_{i,j}$. Here, the $I_{i,j}$ are the elements of an indicator matrix which contains non-zero elements for those entries that correspond to mappings from a particular Pfam domain family to GO terms. Note that some Pfam families lack a mapping to GO terms and thus systematically do not contribute to the aggregation, which may introduce a bias towards more positive or negative deviations. Therefore, we excluded these families from the analysis and renormalized the vector of domain-specific deviations to zero-mean before computing the GO enrichment.

3 Results and discussion

Our approach to characterization of functional novelty in metagenomes is based on the analysis of unexplained protein domain hits as obtained from a mixture model for taxonomic profiling. In order to evaluate the utility of our method, we conducted several case studies using different real-world metagenomic datasets. While the first two studies focus on the characterization of novelty in single datasets, the third case represents a comparative analysis of the proposed novelty indicators for large metagenome data collections.

3.1 Characterization of metabolic novelty in complex communities

Metagenomes of complex communities such as found in biomass-degrading environments have been intensively investigated with respect to novel genes because of their relevance regarding biotechnological applications [HSE$^+$11, SBD$^+$08]. Therefore, as a first case study we analyzed a cow rumen metagenomic dataset (see also section 2.1). The Pfam domain detection resulted in a high fraction of sequences without valid domain assignments (FSU) of $\approx$73% but still provided a sufficient amount of domain hits ($\approx 6.2 * 10^6$) for our overrepresentation analysis (see section 2.3). A brief inspection of the taxonomic profile as estimated by the Taxy-Pro method revealed neglectable proportions ($<$1%) of eukaryotes or viruses in the dataset, with Proteobacteria ($\approx$42%) and Firmicutes ($\approx$13%) accounting for the largest fractions of bacteria. Furthermore, the approximation error of the mixture model in terms of the fraction of domains unexplained (FDU: $\approx$19%) indicates a good model fit and a modest level of the overall novelty.

The analysis of unexplained domain counts in the cow rumen dataset resulted in a high number (1039) of potentially interesting families (p-value $\leq$0.05), which can partly be explained by the large number of reads associated with this dataset. However, by concentrating on the novelty indicators of our approach in terms of the number of expected unexplained domain counts (EUC, see section 2.3) and their fraction as compared to the original counts (OC), we restrict the discussion of results to the most promising candidates. Figure 1 shows a scatter plot of the significantly overrepresented Pfam families in terms of their EUC and EUC/OC ratio, whereby the nine data points associated with the highest EUC peak out of the distribution. Here, the "Leucine rich repeats" (LRR) family consisting of 6 copies (PF13306) shows the highest EUC (27,392) and a relatively high "novelty ratio" (EUC/OC) of $\approx$62%. According to the family-specific Pfam summary, these leucine-rich repeat motifs are found in many (functionally unrelated) protein families and represent a structural rather than a specific functional property. Further inspection of figure 1 reveals functionally more specific domains such as two glycosyl hydrolases, a carboxylesterase and an aldo/keto reductase family. Glycosyl hydrolases are known to play an important role in fermentation and have been associated with novel genes during the analyses of the Cow rumen metagenome [HSE$^+$11, FGC$^+$05, FGB$^+$12]. Interestingly, multi-domain proteins consisting of glycosyl hydrolase domains and leucine-rich repeats have been identified from newly sequenced species (e.g. UniProt accession I0TBE2, www.uniprot.org/uniprot/I0TBE2). However, the combined appearance of such domains

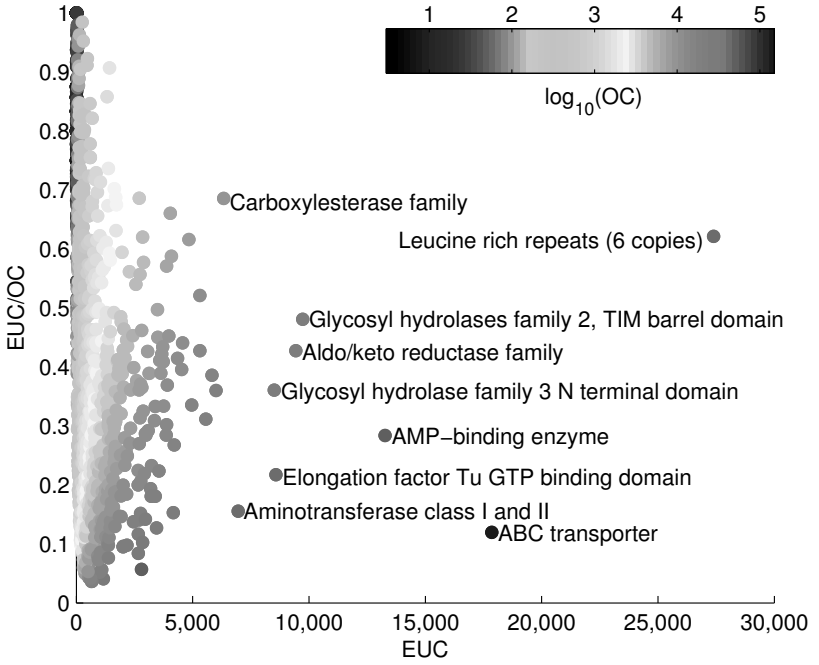

Figure 1: Scatter plot of the number of expected unexplained counts (EUC) and the fraction of EUC and original counts (OC) for 1039 significantly overrepresented domain families of the Cow rumen dataset. The marker color corresponds to the logarithmically scaled OC and Pfam domain descriptions are provided for the nine families associated with the highest EUC.

can usually not be identified in short-read data prior to an assembly of the reads. Our findings suggest that a large amount of novelty in the Cow rumen dataset may be represented by yet unknown combinations of LRR with enzymatic domains.

In principle, metagenomes with a high degree of overall novelty (and thus a high FDU) are particularly interesting for identification of novel genes, however, an FDU of $>$80% indicates an insufficient reconstruction of the metagenomic profile by reference organisms [KALM13]. As a second case study and as an example of an environment with extreme conditions, we analyzed a deep-sea sediment metagenome from the western Gulf of Mexico (see also section 2.1). The 'Sediment' dataset showed a high FSU ($\approx$91%) with only 37,318 domain hits and a fairly high FDU ($\approx$58%) that already indicates the limits of the mixture model. Analysis of unexplained domain counts resulted in 229 significantly overrepresented families many of which can be related to iron supply to the cell under anaerobic conditions. The highest degree of novelty was observed for a "Ferrous iron transport protein B" (PF02421; EUC: 1079, EUC/OC: 99%) which has already been identified during the initial analysis of this sample [BWTH11]. The large amount of novelty related to iron processing can easily be identified in the list of Gene Ontology terms as obtained by our enrichment analysis (see section 2.4). Table 1 shows the top ten GO terms associated

GO ID	GO name	SOD
GO:0006812	cation transport	0.050
GO:0006811	ion transport	0.045
GO:0003674	molecular function	0.045
GO:0006810	transport	0.042
GO:0005488	binding	0.040
GO:0044765	single-organism transport	0.040
GO:0000041	transition metal ion transport	0.037
GO:0015684	ferrous iron transport	0.036
GO:0015093	ferrous iron transmembrane transporter activity	0.036
GO:0072511	divalent inorganic cation transport	0.036

Table 1: Top ten overrepresented GO categories for the Sediment dataset according to the sum of deviations (SOD, column three) of associated domain families.

with the overrepresented families. Here, the GO categories are sorted according to the sum of deviations of the associated domain families. The list allows an intuitive interpretation of the distribution of novelty over different domain families and a quick characterization of the functional novelty in terms of a standardized vocabulary of metabolic processes.

3.2 Large-scale comparative analysis on human microbiome data

Our third case study describes the analysis of novelty in large collections of short-read data and the comparison of novelty profiles. Here, we used 750 metagenome datasets from the human microbiome project (HMP) associated with five major body sites (see section 2.1). The average number of significantly overrepresented domain families over a body site varied between $\sim$200 (Airways) and $\sim$330 (Oral cavities). The average FDU ranged from $\approx$8% (Oral cavities and Gastrointestinal tract) to 17% (Urogenital tract), indicating a good fit of the mixture model in general.

For a more detailed analysis of the HMP dataset we focused on Oral samples, which showed the highest degree of novelty in terms of the number of significant domain families in our overrepresentation analysis. Here, we compared samples from three out of nine subsites with more than 100 associated datasets: tongue dorsum (135 samples), supragingival plaque (128) and buccal mucosa (122). In order to aggregate the results, we calculated the average EUC for each domain family and subsite over all samples. Table 2 shows the top ten domain families associated with a high overrepresentation in terms of maximum average EUC regarding the oral subsites. Here, the "IgA1-specific Metalloendopeptidase" (PF07580) on rank one can be identified as representing a high amount of novelty in the buccal mucosa samples. IgA1 peptidases cleave specific peptide bonds in mammalian immunoglobulin A1 (IgA1) and can be found in particular in pathogenic bacteria. At mucosal sites of infection they can destroy the structure and function of human IgA1 and eliminate an important aspect of host defense [MS06]. On rank five in the list, a "G5 domain" (PF07501) also shows a high degree of novelty associated with buccal

Pfam ID	name	tongue	plaque	mucosa
PF07580	M26 IgA1-specific Metallo-endopeptidase C-terminal region	11	100	**1568**
PF00496	Bacterial extracellular solute-binding proteins, family 5 Middle	432	285	**1452**
PF07690	Major Facilitator Superfamily	59	**779**	242
PF12698	ABC-2 family transporter protein	12	61	**771**
PF07501	G5 domain	18	94	**735**
PF01610	Transposase	165	12	**718**
PF08428	Rib/alpha-like repeat	9	25	**650**
PF00005	ABC transporter	193	92	**626**
PF07564	Domain of Unknown Function (DUF1542)	5	58	**585**
PF00593	TonB dependent receptor	324	**499**	65

Table 2: Average expected unexplained counts (EUC, rounded) for Pfam domains associated with significantly overrepresented domain families for three oral subsites from HMP datasets. Domain families are sorted from high to low according to the maximum average EUC regarding the subsites (in boldface type).

mucosa (see also summary of Pfam database entry). The G5 domain is found in the N-terminus of peptidases belonging to the M26 family and is suspected to have an adhesive function. Furthermore, on rank nine we can see a mucosa-specific domain of unknown function ("DUF1542", PF07564). This domain is found in several cell surface proteins some of which are involved in antibiotic resistance and/or cellular adhesion (see also Pfam summary). Our findings suggest that a high degree of novelty that we identified is related to yet unknown mucosa-specific bacteria providing many proteins related to pathogenicity. This first glimpse of metagenomic novelty in the human microbiome already indicates the potential utility of our approach for medical research.

4 Conclusion

We presented an approach for the characterization of metagenomic novelty based on the analysis of the functional profile of a metagenome. Using the unexplained protein domain hits as obtained from the UProC tool for domain detection and the Taxy-Pro mixture model estimation, the identification of significant domain families usually takes a few minutes for a real-world metagenomic dataset. The results in terms of lists of domain families highlight the functional and structural properties that are not well-explained by the mixture model and therefore provide interesting candidates for further analysis. We are aware that using binomial distributions under the assumption of statistical independence is a coarse approximation which is likely to overestimate the number of significant differences. This shortcoming can be overcome if biological (or technical) replicates are available which would allow the application of more sophisticated models, such as the negative binomial distribution that is often used in RNA-seq analysis.

Classical gene mining approaches are based on a bottom up strategy that requires the identification of complete genes, which typically extend to lengths around 1000 bp for microbial organisms. Therefore, the methods require datasets with long sequencing reads or have to rely on an assembly of reads prior to gene identification. In contrast, our mixture model is based on functional profiles that are estimated by detecting and counting Pfam protein domains. Therefore the method can even be applied to short-read data as shown in our studies on rumen and HMP samples. In that way, our approach can be used for novelty mining on large metagenomic data collections to identify interesting habitats and samples that may be further explored by functional screening methods. On the other side, datasets with a high degree of novelty usually yield a large proportion of sequence reads without domain assignments, which, as a consequence, do not contribute to the characterization of novelty. However, this problem affects all approaches for gene mining and can only be solved by a higher coverage of functionally annotated sequences in reference databases.

Using sequences without similarities to known families, the interpretation of results from a classical bottom up analysis can be rather difficult. In the extreme, a cluster analysis may end up in a large number of putatively new protein families with no or only little evidence for functional properties [YSR+07]. In contrast, our top down approach always yields a prediction of novelty in terms of biologically defined categories. However, the descriptions of Pfam domain families are often not easy to interpret in terms of metabolical relevance. In this study we showed how a mapping of domains to Gene Ontology (GO) categories can be used to facilitate the interpretation of functional annotations. On the downside, the lack of mappings for many Pfam domains and the hierarchical structure of GO also result in a systematic overrepresentation of functionally unspecific terms. Future work will focus on the extension of our approach to protein families that can directly be associated with metabolic pathways [KGS+14]. Furthermore, we plan to integrate the characterization of functional novelty into the CoMet web server for comparative functional profiling of metagenomes [LASM11].

5 Acknowledgments

We thank Heiner Klingenberg for technical support. This work was supported by the Deutsche Forschungsgemeinschaft (grant Me3138 to P.M.).

References

[BLH+10] T. Z. Berardini, D. Li, E. Huala, et al. The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, 38(Database issue):D331–335, Jan 2010.

[BWTH11] J. F. Biddle, J. R. White, A. P. Teske, and C. H. House. Metagenomics of the subsurface Brazos-Trinity Basin (IODP site 1320): comparison with other sediment and pyrosequenced metagenomes. *ISME J*, 5(6):1038–1047, Jun 2011.

[FBC+14] R. D. Finn, A. Bateman, J. Clements, et al. Pfam: the protein families database. *Nucleic Acids Res.*, 42(Database issue):D222–230, Jan 2014.

[FGB+12] M. Ferrer, A. Ghazi, A. Beloqui, et al. Functional metagenomics unveils a multifunctional glycosyl hydrolase from the family 43 catalysing the breakdown of plant polymers in the calf rumen. *PLoS ONE*, 7(6):e38134, 2012.

[FGC+05] M. Ferrer, O. V. Golyshina, T. N. Chernikova, et al. Novel hydrolase diversity retrieved from a metagenome library of bovine rumen microflora. *Environ. Microbiol.*, 7(12):1996–2010, Dec 2005.

[Hol79] Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):pp. 65–70, 1979.

[HSD+07] E. D. Harrington, A. H. Singh, T. Doerks, et al. Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *Proc. Natl. Acad. Sci. U.S.A.*, 104(35):13913–13918, Aug 2007.

[HSE+11] M. Hess, A. Sczyrba, R. Egan, et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, 331(6016):463–467, Jan 2011.

[KALM13] H. Klingenberg, K. P. Aßhauer, T. Lingner, and P. Meinicke. Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. *Bioinformatics*, 2013.

[KGS+14] M. Kanehisa, S. Goto, Y. Sato, et al. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, 42(Database issue):199–205, Jan 2014.

[KTZ+07] N. Kannan, S. S. Taylor, Y. Zhai, J. C. Venter, and G. Manning. Structural and functional diversity of the microbial kinome. *PLoS Biol.*, 5(3):e17, Mar 2007.

[LASM11] T. Lingner, K. P. Asshauer, F. Schreiber, and P. Meinicke. CoMet–a web server for comparative functional profiling of metagenomes. *Nucleic Acids Res.*, 39(Web Server issue):W518–523, Jul 2011.

[MCR10] L. V. Mello, X. Chen, and D. J. Rigden. Mining metagenomic data for novel domains: BACON, a new carbohydrate-binding module. *FEBS Lett.*, 584(11):2421–2426, Jun 2010.

[MS06] D. Mistry and R. A. Stockley. IgA1 protease. *Int. J. Biochem. Cell Biol.*, 38(8):1244–1248, 2006.

[PGG+09] J. Peterson, S. Garges, M. Giovanni, et al. The NIH Human Microbiome Project. *Genome Res.*, 19(12):2317–2323, Dec 2009.

[SBD+08] A. Schlüter, T. Bekel, N. N. Diaz, et al. The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. *J. Biotechnol.*, 136(1-2):77–90, Aug 2008.

[SDL+09] A. H. Singh, T. Doerks, I. Letunic, J. Raes, and P. Bork. Discovering functional novelty in metagenomes: examples from light-mediated processes. *J. Bacteriol.*, 191(1):32–41, Jan 2009.

[YSR+07] S. Yooseph, G. Sutton, D. B. Rusch, et al. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.*, 5(3):e16, Mar 2007.