On the estimation of metabolic profiles in metagenomics*

Kathrin Petra Aßhauer and Peter Meinicke†

Department of Bioinformatics, Institute for Microbiology and Genetics
University of Göttingen
37077 Göttingen, Germany
peter@gobics.de

Abstract

Metagenomics enables the characterization of the specific metabolic potential of a microbial community. The common approach towards a quantitative representation of this potential is to count the number of metagenomic sequence fragments that can be assigned to metabolic pathways by means of predicted gene functions. The resulting pathway abundances make up the metabolic profile of the metagenome and several different schemes for computing these profiles have been used. So far, none of the existing approaches actually estimates the proportion of sequences that can be assigned to a particular pathway. In most publications of metagenomic studies, the utilized abundance scores lack a clear statistical meaning and usually cannot be compared across different studies. Here, we introduce a mixture model-based approach to the estimation of pathway abundances that provides a basis for statistical interpretation and fast computation of metabolic profiles. Using the KEGG database our results on a large-scale analysis of data from the Human Microbiome Project show a good representation of metabolic differences between different body sites. Further, the results indicate that our mixture model even provides a better representation than the dedicated HUMAnN tool which has been developed for metabolic analysis of human microbiome data.

1998 ACM Subject Classification J.3 Life and Medical Sciences – Biology and genetics

Keywords and phrases metagenomics, metabolic profiling, taxonomic profiling, abundance estimation, mixture modeling

Digital Object Identifier 10.4230/OASIcs.GCB.2013.1

1 Introduction

In metagenomics a central task is to characterize the metabolic potential of a microbial community. The metabolic profile of a metagenome quantifies the amount of genetic material that can be attributed to metabolic pathways. The abundance of a pathway is usually estimated by the number of sequences that can be mapped to gene families with functional roles within that pathway. Several heuristics exist to compute a corresponding estimate. Using for instance the KEGG database, an abundance may be estimated by counting all BLAST best hit matches to KEGG Orthologs which are annotated for the particular pathway (see e.g. [4]). There are two major difficulties with this classical approach of metabolic profiling: First, the computational effort for the identification of homologs can become burdensome. Usually the BLASTX tool is required, which takes a considerable amount of

* This work was partially supported by the Deutsche Forschungsgemeinschaft (ME3138 "Compositional descriptors for large scale comparative metagenome analysis").

† corresponding author

German Conference on Bioinformatics 2013 (GCB'13).
Editors: T. Beißbarth, M. Kollmar, A. Leha, B. Morgenstern, A.-K. Schultz, S. Waack, E. Wingender; pp. 1–13
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

CPU-time even for a moderately sized data set. Second, the usual counting scheme lacks a probabilistic model that would provide a clear statistical interpretation of the resulting quantities. To our knowledge, none of the existing heuristics actually yields an estimate of the fraction of sequence material that can be mapped to a particular pathway. Depending on the particular method the existing tools merely provide different kinds of abundance scores [14, 12, 1, 5, 4]. Although these scores may be used for comparative analysis as well, they do not provide a strictly probabilistic description of metagenomic sequence data. Therefore, the comparison and combination with other methods or models is at least problematic. We address both problems, the algorithmic and statistical efficiency within a metabolic mixture model in terms of a mixture of pathways (MoP). This model is capable to provide both, a sound statistical basis and a fast estimation of pathway abundances. Our results on a large-scale analysis of data from the Human Microbiome Project (HMP) show the utility of our method for fast model-based estimation of pathway abundances. Further, the results for the mixture-based metabolic profiles indicate a better separation between body sites than for the profiles of the HUMAnN tool which has particularly been developed for analyis of HMP data.

2 Material

2.1 Human Microbiome Project (HMP)

Within the scope of the Human Microbiome Project (HMP) [3] an extensive collection of samples from healthy individuals from diverse human body sites was established allowing an insight into the functions of the healthy human microbiome. More than thousand HMP data sets are recorded in HMP's Data Acquisition and Coordination Center (DACC) Project Catalog (http://www.hmpdacc.org/resources/data_browser.php) providing a comprehensive data basis for large-scale comparative studies investigating the associations of the human microbiome in healthy and diseased states.

From the HMP-DACC website we assessed the available metadata for the metagenomic samples (http://www.hmpdacc-resources.org/hmp_catalog/main.cgi) and the metabolic reconstruction data (http://www.hmpdacc.org/HMMRC/). The metabolic reconstruction data is obtained through the HMP Unified Metabolic Analysis Network (HUMAnN) pipeline [1]. HUMAnN performs functional and metabolic profiling directly from high-throughput metagenomic short sequence reads. The pipeline starts with a similarity search against a functional sequence database including the KEGG Orthologs (Release 54) using an accelerated translated BLAST implementation. Subsequently, the output is used for a series of gene- and pathway-level quantification, noise reduction, and smoothing steps resulting in the identification of present/absent pathways and modules together with their relative abundances. From the available metabolic reconstruction data, we used the "KEGG pathway abundance values – Summary file" (as of February 2013).

For our mixture modeling approach we used the reduced data samples of the HMP as describes in [10]. For comparability, the available samples and pathway abundances of HUMAnN and our mixture modeling approach were reduced to a subset of samples and pathways available in both methods. The final dataset includes 680 data samples from 14 specific body subsites, which can be grouped into five major body sites.

2.2 KEGG database

For the metabolic mixture modeling approach introduced here, we use the Kyoto Encyclopedia of Genes and Genomes (KEGG) database as reference knowledge base for estimating the pathway abundances of metagenomic samples [9, 8]. KEGG integrates a variety of information and provides links from gene catalogs to higher-level systematic functions of the organisms enabling biological interpretation of genomes and high-throughput datasets.

An essential part of the database with respect to metabolic profiling are the KEGG Orthologs (KO) that consist of gene groups with specific functions directly linked to known pathways in the KEGG Pathway database. Further, the KEGG Orthology is structured as a hierarchy of four flat levels: top, second, third level, and leaf nodes. While the leaf nodes represent the KEGG Orthologous groups, the third level represents the KEGG Pathways, which can be further summarized in higher level pathway classes (top and second level).

For the mixture modeling the required data reference was extracted from the KEGG database (Release 64.0).

2.3 MarVis

The MarVis-Suite (**Mar**ker **Vis**ualization) [7, 6], a toolbox originally developed for the analysis of metabolomic data, was used for filtering, clustering, and visualization of the pathway abundances. For exploration of complex pattern variation within the samples of the different body sites/subsites we used the MarVis-Cluster interface which permits high-level visualization and cluster analysis based on a one-dimensional self-organizing map (1D-SOM). The MarVis-Filter software was used for the identification of pathways overrepresented in the gastrointestinal tract samples compared to the other body subsites.

3 Methods

3.1 Taxonomic mixture modeling

The mixture model based Taxy approach provides a fast and direct estimation of taxonomic abundances in metagenomes. Taxy-Oligo [13] and Taxy-Pro [10] do not perform a taxonomic classification of sequencing reads but instead apply a mixture model to approximate the overall metagenome distribution of oligonucleotides and protein domain hits, respectively. The discrete distribution of oligonucleotides/protein domains is modeled by a mixture of organism-specific profiles as obtained from sequenced reference genomes. Because of the computational efficiency of the taxonomic mixture model approach, both methods were able to perform a large-scale analysis of sequence data from the HMP without using a computer cluster or special hardware. All reference profiles were obtained from the bacterial and archaeal genomes available in the KEGG database (Release 64.0). These genomes were also used for pre-computing the organism-specific pathway abundances for the metabolic profiling of metagenomes. For Taxy-Pro, all protein domain profiles according to the Pfam database [2] were obtained from the CoMet web server [11].

3.2 Metabolic mixture modeling

For metabolic profiling, we assume that the genomic sequence material to some degree can be explained by a mixture of pathways. The mixture approach accounts for the fact that in most cases a putative gene function as observed in a sequence fragment provides evidence for more than one metabolic pathway. The statistical representation of this ambiguity of the function-to-pathway mapping was the main motivation for the development of the following model. With M pathways P_i the probability to observe a function F encoded in sequences under this model is:

$$\tilde{p}(F) = \sum_{i=1}^{M} p(P_i) p(F|P_i) \tag{1}$$

The tilde indicates that $\tilde{p}(F)$ only is an approximation of the functional profile $p(F)$ because not every function can be explained in terms of metabolic pathways. The prior pathway probabilities $p(P_i)$ denote the overall sequence-based abundance of functions associated with pathway P_i and correspond to the mixture weights of the model. These weights are the central model parameters, which can directly be used and interpreted in terms of the relative abundances of a metabolic profile. The conditional probability $p(F|P_i)$ denotes the i-th pathway-specific distribution over N possible gene functions F_j. The annotation in current databases, such as KEGG, can be represented by some $M \times N$ assignment matrix $\mathbf{A}$ with binary entries $A_{ij} = 1$ denoting that function j is associated with pathway i. From that assignment it follows that all functions not associated with pathway i must attain a zero conditional probability. Just from the annotation, we cannot draw any conclusions about the other probabilities. Without further knowledge the only reasonable assumption is that the $p(F|P_i)$ are proportional to the corresponding overall function probabilities, i.e.

$$\forall i, j: \; p(F_j|P_i) \propto A_{ij} p(F_j). \tag{2}$$

This constraint implies that the ratio between any two non-zero function probabilities in a pathway is equal for all pathways these two functions are associated with and must equal the global ratio of the corresponding probabilities of the functional profile $p(F)$. With the N estimates $\hat{p}(F_j)$ of the specific function probabilities of the profile as derived from the observed frequencies, e.g. from BLAST hit counts, we have the following estimator of the conditional probabilities:

$$\hat{p}(F_j|P_i) = \frac{A_{ij}\hat{p}(F_j)}{\sum_{k=1}^{N} A_{ik}\hat{p}(F_k)}. \tag{3}$$

Now let us consider the assignment probability

$$p(P|F_j) = \frac{p(P)p(F_j|P)}{\sum_{i=1}^{M} p(P_i)p(F_j|P_i)} \tag{4}$$

which denotes the responsibility of a pathway for a given function F_j, i.e. the contribution of a pathway to the explanation of that function. We assume that this probability is equal for all pathways the function is associated with. Without further knowledge, just with the underlying pathway annotation, there is no reason to prefer a particular pathway for the explanation of an observed function. This implies the following additional constraint:

$$\forall i, j, k: \; A_{kj} p(P_i|F_j) = A_{ij} p(P_k|F_j). \tag{5}$$

For a function F_j that is annotated in two pathways P_i and P_k we can obtain the ratio of the corresponding pathway abundance estimators using the former three equations (3), (4) and (5):

$$\frac{\hat{p}(P_i)}{\hat{p}(P_k)} = \frac{\hat{p}(F_j|P_k)}{\hat{p}(F_j|P_i)} = \frac{\sum_{s=1}^{N} A_{is}\hat{p}(F_s)}{\sum_{t=1}^{N} A_{kt}\hat{p}(F_t)}. \tag{6}$$

From the above proportionality, we finally obtain the estimator of the pathway probabilities:

$$\hat{p}(P_i) = \frac{\sum_{j=1}^{N} A_{ij}\hat{p}(F_j)}{\sum_{k=1}^{M}\sum_{l=1}^{N} A_{kl}\hat{p}(F_l)}. \tag{7}$$

Using matrix vector algebra we can compute the whole metabolic profile vector $\mathbf{p}$ with entries $\hat{p}(P_i)$ from the functional profile vector $\mathbf{f}$ with entries $\hat{p}(F_j)$ by

$$\mathbf{p} = \frac{\mathbf{Af}}{\mathbf{1}^T\mathbf{Af}} \tag{8}$$

where $\mathbf{1}$ is an M-vector of ones. In an application of the above mixture model most time will be spent for the computation of the functional profile which usually requires a costly BLASTX matching of metagenomic reads against a database of functionally annotated protein sequences, such as the KEGG Orthologues. However, with our formulation in terms of a statistical model we are able to provide a shortcut that utilizes the combination with another model to obtain a hierarchical mixture of pathways. Assume that we have the functional profiles of K reference organisms as columns in an $N \times K$ matrix $\mathbf{F}$ and we have estimated the relative abundances of the reference organisms in a taxonomic profile vector $\mathbf{t}$. Then we can approximate the functional profile of the metagenome by a linear combination of reference profiles $\mathbf{Ft}$. In Taxy-Pro [10] we use this mixture model in combination with Pfam functional profiles to estimate the taxonomic abundances in a metagenome. Here, we propose a combination with K pre-computed KEGG reference profiles to predict the functional profile of a metagenome from its taxonomic profile which may be obtained by some fast method such as the oligonucleotide-based Taxy tool [13]. The estimator of the metabolic profile is then

$$\mathbf{p} = \frac{\mathbf{AFt}}{\mathbf{1}^T\mathbf{AFt}}. \tag{9}$$

Note that also the matrix product $\mathbf{AF}$ can be pre-computed to obtain K organism-specific metabolic profiles which are then just combined by the taxonomic weights $\mathbf{t}$ of a metagenome to obtain its metabolic profile. In principle, this gives rise to a nested model where a mixture of pathways is first used for each reference organism to estimate its metabolic profile. This step has only to be performed once for each organism and therefore even a costly BLASTX analysis may be used for the “offline” training of the organism-specific models. When applied to metagenomic data a mixture of the utilized reference organisms has to be estimated by some taxonomic profiling method. In order to combine the two models the second step requires a profiling method that actually estimates the abundances in terms of the amount of sequence material that can be attributed to a particular organism. For example, this requirement is automatically fulfilled when using Taxy-Oligo [13] or Taxy-Pro [10], which we both included in the evaluation of our approach, as described above. For an application of the MoP model, it is important to check whether the metagenome composition can actually be approximated by a mixture of known reference organisms. If the reference is completely insufficient for a description of the metagenome composition, the mixture approach in general would become inadequate. Therefore, it is desirable, that the taxonomic profiling method gives us an indication of the fidelity of the abundance estimates. Both, Taxy-Oligo and Taxy-Pro provide a specific error measure to assess the adequacy of the underlying model. In this case, the fraction of oligonucleotides unexplained (FOU) and the fraction of domain-hits unexplained (FDU) should be inspected when using Taxy-Oligo and Taxy-Pro, respectively.

3.2.1 Workflows

For the evaluation of our model, we implemented the direct application of the metabolic mixture model as well as the nested model.

The direct application of the mixture model starts with a BLASTX analysis where the metagenomic reads are mapped against a reference database consisting of KO amino acid sequences of bacterial or archaeal origin. By default BLAST hits with E-value $\leq 10^{-2}$ were considered to be significant. The functional profile vector $\mathbf{f}$ is obtained by counting the KO-specific BLAST hits using a fractional increment of $1/K$ if K different KOs simulataneously show significant hits for a particular sequence. Note that due to the computational expense of BLASTX on metagenomes, we restricted the correlation analyis (see section 4.1) to a subset of six HMP data samples from different body subsites (SRS013825, SRS016752, SRS022621, SRS024265, SRS024428, and SRS055401). For the computation of the assignment matrix $\mathbf{A}$ the association of KOs with KEGG Pathways was extracted from the database and transformed into a binary matrix. Finally, the mixture model was applied as described above using the functional KO profile vector $\mathbf{f}$ and matrix $\mathbf{A}$ as input.

For the nested model, we first pre-computed the organism-specific metabolic profiles from reference genomes using all bacterial and archaeal KEGG Genomes. The KEGG Genomes were downloaded and subsequently fragmented in overlapping reads of length 400 bp with 200 bp overlap simulating a two-fold coverage of the genomes as previously described in [10]. For each reference organism, first the functional profile vector is calculated and then the metabolic profile is estimated applying the steps as described for the direct mixture approach. By combining the weights $\mathbf{t}$ of a metagenome with the pre-computed organism-specific metabolic profiles the metabolic profile of a metagenome can be obtained in an efficient manner. Note that a BLASTX/KO analysis of the metagenome is not required in this case. For the estimation of the taxonomic profile $\mathbf{t}$, we were using both, Taxy-Oligo and Taxy-Pro. According to the utilized taxonomic profiling method we denote our metabolic mixture model MoP-Oligo and MoP-Pro, respectively.

4 Results

To validate the metabolic mixture model on a well-studied dataset, we analyzed metagenomic sequences from the Human Microbiome Project (HMP) [3]. Originally, the metabolic profiles of the HMP data have been investigated by means of the HMP Unified Metabolic Analysis Network (HUMAnN) pipeline [1]. In the following, we use the metabolic profiles of HUMAnN for comparison with the abundance estimates that we obtained from our mixture of pathways model.

4.1 Correlation analysis

To study the similarity of metabolic profiles across different methods we computed the Pearson and Spearman (rank) correlation coefficients of the pathway abundance estimates. First, we evaluated the fast approximation scheme using pre-computed reference profiles based on Taxy-Pro taxonomic profiles (MoP-Pro). The resulting metabolic profiles were compared with the direct application of the mixture model to KO frequencies, which were obtained from a more time consuming BLASTX analysis. For each data sample, the correlation of the pathway abundances on two different pathway hierarchy levels (second and third level) was calculated.

The means and standard deviations of all data examples of the Pearson and Spearman correlation coefficients are shown in Table 1. The results show a very high correlation of the approximation-based and the directly obtained abundances. By reducing the number of pathways from 340 to 38 according to the third and second pathway hierarchy levels an increase of the correlation from 0.9558 to 0.9804 and 0.9491 to 0.9842 could be observed for the Taxy-Pro-based approximation. These results indicate that the approximative approach is very close to the direct approach and therefore provides a computationally attractive alternative to the BLAST-based estimation.

Table 1 Correlation analysis based on the metabolic abundances obtained by applying the Taxy-Pro-based approximation and the direct mixture approach. The correlation is calculated according to Spearman and Pearson and at the third and second pathway hierarchy level.

	Pearson	Spearman
Third level	0.9557 (± 0.0409)	0.9491 (± 0.0124)
Second level	0.9803 (± 0.0150)	0.9842 (± 0.0110)

The correlations are similarly high for the even faster Taxy-Oligo variant (MoP-Oligo) with results shown in Table 2.

Table 2 Correlation analysis based on the metabolic abundances obtained by applying the Taxy-Oligo-based approximation and the direct mixture approach. The correlation is calculated according to Spearman and Pearson and at the third and second pathway hierarchy level.

	Pearson	Spearman
Third level	0.9575 (± 0.0409)	0.9466 (± 0.0105)
Second level	0.9796 (± 0.0138)	0.9813 (± 0.0087)

In contrast to the high similarity of results between different variants of the mixture approach the correlation between the mixture-based pathway abundances and the HUMAnN-based profiles is comparatively low with a Pearson correlation of 0.5290 as shown in Table 3. However, the correlation is increasing when considering the second pathway level or when using the Spearman rank correlation. A maximum rank correlation of 0.9080 indicates that the coarse shape of metabolic profiles is still rather similar between different approaches. Note that the correlation with HUMAnN profiles was averaged over all 680 HMP samples.

Table 3 Correlation analysis based on the metabolic abundances obtained by applying HUMAnN and the TaxyPro-based mixture model. The correlation is calculated according to Spearman and Pearson and at the third and second pathway hierarchy level.

	Pearson	Spearman
Third level	0.5290 (± 0.0206)	0.7588 (± 0.0242)
Second level	0.7884 (± 0.0308)	0.9080 (± 0.0135)

4.2 Nearest neighbor classification

To assess the quality of the estimated metabolic profiles we first investigated whether the body site (subsite) classification of HMP samples can be reproduced by the corresponding pathway abundances. For that purpose, we evaluated the predictive power of metabolic profiles by some nearest neighbor classification scheme using different profile distance measures. We utilized a leave-one-out cross validation, measuring the classification rate for Euclidean distance, City block metric and Shannon-Jensen divergence on profiles.

The results for body sites and subsites as shown in Table 4 reveal that the nearest neighbor classification rate is rather high and varies between 0.9735 and 0.9897 for the five body sites and between 0.8853 and 0.9235 for the 14 body subsites. For both classification problems, HUMAnN shows the highest prediction accuracy irrespective of the distance measure used. However, the two mixture variants are always very close with a maximum difference of 2.94% for the Euclidean distance on body subsite level between HUMAnN and MoP-Oligo.

Table 4 Nearest neighbor classification performing a leave-one-out cross validation with the Euclidean distance, City block metric and Shannon-Jensen divergence as distance measure for the approaches HUMAnN, MoP-Pro, and MoP-Oligo.

	Body site			Body subsite		
	Euclidean	City block	Jensen-Shannon	Euclidean	City block	Jensen-Shannon
HUMAnN	0.9838	0.9897	0.9868	0.9147	0.9235	0.9132
MoP-Pro	0.9794	0.9809	0.9779	0.9103	0.9103	0.9059
MoP-Oligo	0.9735	0.9750	0.9779	0.8853	0.8956	0.9132

4.3 Clustering performance

For a more comprehensive analysis of profile distances, we compared the body site (subsite) classification of samples with a profile-based clustering of the data. For clustering, we used a standard hierarchical approach with average linkage, also known as UPGMA. In this context, we evaluated the same three distance measures as for the nearest neighbor classification experiment. The quality of the cluster partitioning was assessed by the Jaccard coefficient, measuring the overlap of the resulting clusters with the HMP body site (subsite) groups.

The results obtained through the application of HUMAnN, MoP-Pro, and MoP-Oligo are presented in Table 5 which shows a large variation of the clustering performance.

Table 5 Cluster partitioning quality in terms of the Jaccard coefficient based on Euclidean distance, City block metric and Shannon-Jensen divergence for metabolic profiles of HUMAnN, MoP-Pro, and MoP-Oligo

	Body site			Body subsite		
	Euclidean	City block	Jensen-Shannon	Euclidean	City block	Jensen-Shannon
HUMAnN	0.4335	0.4342	0.4325	0.2361	0.3715	0.2344
MoP-Pro	0.6958	0.8817	0.6971	0.4791	0.4603	0.4801
MoP-Oligo	0.6577	0.7251	0.6382	0.3671	0.3008	0.3939

The Jaccard coefficient varied between 0.4325 and 0.8817 at body site level and between 0.2344 and 0.4801 at body subsite level. The partitioning of the MoP-Pro approach always showed the highest values on body site and subsite level. For both levels, the clustering performance of the MoP-Oligo approach is superior to HUMAnN except for the City block metric at body subsite level.

4.4 1D-SOM clustering and visualization

In order to study the overall variation of pathway abundance patterns over the whole range of HMP samples, we analyzed the estimated metabolic profiles with the MarVis tool. A one-dimensional self-organizing map (1D-SOM) was created using MarVis-Cluster (see Figure 1) to obtain a set of ordered prototypes well-suitable for visualization of profile variations.

Here, we utilized the second pathway level where we reduced the profiles to include just the top 10 pathways with the highest variance over all samples. Taking the union of the top 10 MoP and HUMAnN pathways we achieved a total of 13 profile dimensions that we used for 1D-SOM clustering with 14 prototypes and a unit 2-norm scaling of profile vectors. The resulting visualization indicates that most of the body sites are separated into distinct clusters (Figure 1). For the MoP profiles three major groups of clusters can be identified: gastrointestinal tract (GI tract, left side), urogenital tract (UG tract, right side), and an intermediate set of clusters from airways, oral and skin sites. Furthermore there are some interesting gradients (left to right) that show a decreasing relative abundance for *Amino Acid Metabolism*, *Carbohydrate Metabolism*, and *Signal Transduction* pathways and an increasing abundance for *Membrane Transport*, *Nucleotide Metabolism*, *Replication and Repair*, and *Translation* pathways. In contrast, the 1D-SOM based on the HUMAnN pathway profiles shows a distinct picture of the overall variation. The different body sites are not as clearly separated as for the MoP-based SOM and the overall abundance gradients of selected pathways are not as prominent as for the MoP results. The visible gradients (left to right) that show a decreasing abundance include the *Amino Acid Metabolism* and *Metabolism of Cofactors and Vitamins* pathways while an increasing abundance can be observed for *Metabolism of Other Amino Acids*, *Replication and Repair*, and *Translation* pathways.

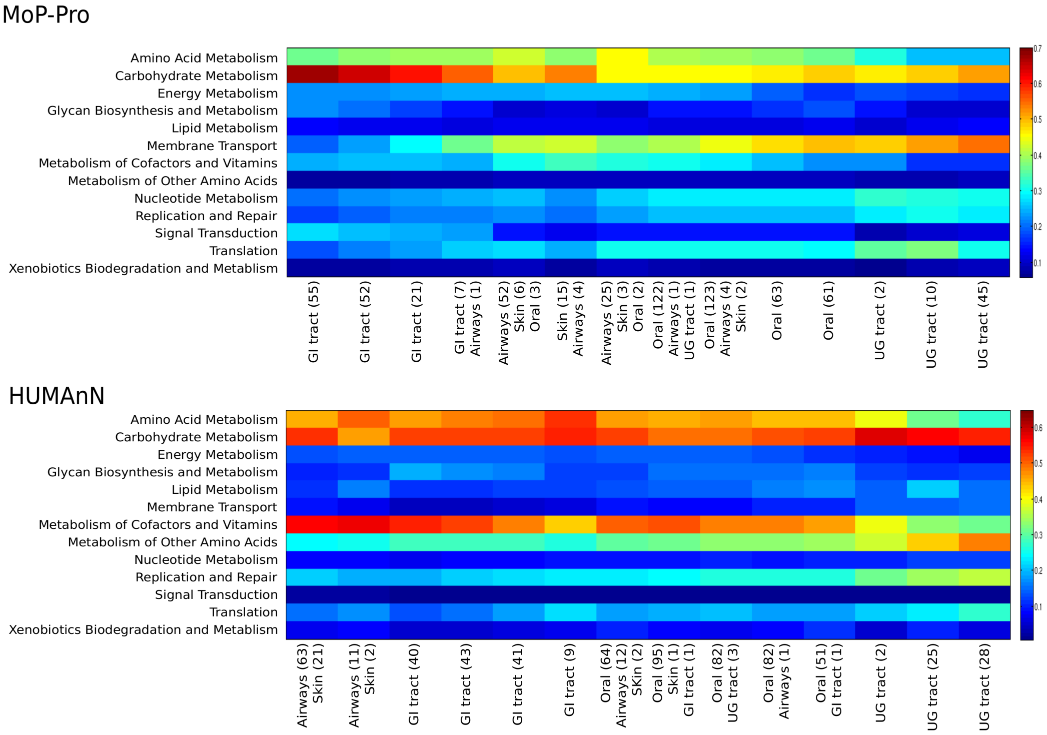

Figure 1 1D-SOM created with MarVis-Cluster at second pathway hierarchy level for MoP-Pro (upper) and HUMAnN (lower) profiles (GI tract – gastrointestinal tract; UG tract – urogenital tract). The numbers (in brackets) indicate the number of profiles (samples) assigned to the corresponding prototype (cluster) above.

4.5 Significant pathways

For a specific analysis of the metabolic profiles in terms of statistically significant differences in pathway abundances between different body sites we compared the gastrointestinal (GI) tract samples with all other HMP samples. To identify overrepresented pathways for the GI body site we applied an ANOVA with Holm-Bonferroni (FWER) correction on pathway abundances of the second level pathway hierarchy, filtered for pathways with an FWER below 0.05, and ranked the remaining pathways according to their fold-change in terms of the corresponding overrepresentation factor on mean abundances. In Table 6 the significant pathways of MoP-Pro and HUMAnN with a calculated fold-change larger than 1 are listed.

Table 6 MarVis-Filter analysis for the identification of overrepresented pathways in the gastrointestinal tract samples in comparison to all other body subsites. All second level pathways obtained through the application of the MoP-Pro and HUMAnN approach with a fold-change larger than 1 are listed.

MoP-Pro

Pathway (second level)	Fold-Change
Transport and Catabolism	2.00
Signal Transduction	1.81
Digestive System	1.79
Biosynthesis of Other Secondary Metabolites	1.53
Nervous System	1.48
Carbohydrate Metabolism	1.23
Glycan Biosynthesis and Metabolism	1.15
Endocrine System	1.13
Immune System	1.12

HUMAnN

Pathway (second level)	Fold-Change
Digestive System	2.04
Endocrine System	1.12
Glycan Biosynthesis and Metabolism	1.07
Amino Acid Metabolism	1.06
Biosynthesis of Other Secondary Metabolites	1.06
Energy Metabolism	1.01

HUMAnN and MoP-Pro identified pathways associated with the *Digestive System*, *Endocrine System*, *Biosynthesis of Other Secondary Metabolites*, *Glycan Biosynthesis and Metabolism* to be overrepresented in GI tract samples. For all these pathways, except for the *Digestive System*, the MoP-Pro fold-change was higher than the corresponding factor of HUMAnN. Exclusively for the HUMAnN approach, pathways associated with *Amino Acid Metabolism* and *Energy Metabolism* are found to be slightly overrepesented. Furthermore, through the application of the MoP-Pro we detected five additional pathways to be overrepresented: *Transport and Catabolism*, *Signal Transduction*, *Nervous System*, *Carbohydrate Metabolism*, and *Immune System*. These additional pathways are possibly related to a mutually beneficial relationship between the gut microbiota and the host, maintaining a normal mucosal immune function and nutrient absorption. Furthermore, the overrepresentation of pathways associated with the nervous system may provide an indication for the bidirectional brain-gut interactions which have an important role in the modulation of gastrointestinal functions and possibly support the hypothesis of a communication pathway between the microbiota and the host's central nervous system [15].

4.6 Runtime

To get an overview of the computational cost of the different variants of the mixture modeling, we measured the approximate runtime averaged over the six selected HMP data samples (average size ~200 MB) used for the correlation analysis. For the selected data sets the mean runtime ranges from minutes to months. The longest CPU times were required by the direct application of the mixture model due to the costly similarity searches against the KO database. On a computer with four CPUs (2.4 GHz) BLASTX searches and calculation of the metabolic profile took approximately 58 days. The fastest method was MoP-Oligo with about half a minute, followed by the MoP-Pro method with about one minute runtime in total. Once the taxonomic profile is estimated, using either MoP-Oligo or MoP-Pro, the resulting matrix vector multiplication for obtaining the metabolic profile of a metagenome can be done within a second.

5 Discussion

We presented a novel metabolic profiling approach for metagenomics, which is based on a mixture of pathways (MoP) model for estimation of pathway abundances. To overcome computationally intense homology searches, we implemented a shortcut to estimate the metabolic profile of a metagenome. Here, we link the taxonomic profile of the metagenome to a set of pre-computed metabolic reference profiles. The combination of the taxonomic abundance estimates, obtained through the fast methods Taxy-Oligo and Taxy-Pro, and the metabolic reference profiles, based on the KEGG database, achieves an unrivaled speed of the metabolic profiling approach.

We are aware of the difficulties in the evaluation that arise when trying to assess the quality of the resulting metabolic profiles. Therefore we restricted our evaluation to the large-scale data from the Human Microbiome Project (HMP) and to the comparison with the observations and findings for this data obtained through the HUMAnN approach. In this setup we tried to provide several views on metabolic profiles considering different aspects of quality: Our correlation analysis has shown that the pathway abundances obtained through our statistical model are slightly different when compared to the HUMAnN abundance predictions. However, we demonstrated through the nearest neighbor classification that our model based approach is at least comparable to the HUMAnN approach when considering the prediction of body sites and subsites. Considering the cluster performance analysis, our approach even outperforms the HUMAnN pipeline in most cases. Furthermore, our case study on statistically overrepresented pathways in the gastrointestinal tract provides additional insight in comparison with the results of the dedicated HUMAnN approach.

To our knowledge, the MoP approach for the first time provides a potentially unbiased estimator of the fraction of sequences that can be attributed to a particular pathway. In addition, our model-based combination with taxonomic abundance estimators also provides the fastest way to estimate the metabolic profile of a metagenome. We intend to make the method accessible via an easy-to-use interface by integration into the CoMet web server [11] (http://comet.gobics.de).

Acknowledgements We would like to thank Heiner Klingenberg, Thomas Lingner, Alexander Kaever, and Manuel Landesfeind for fruitful discussions.

References

1 Sahar Abubucker, Nicola Segata, Johannes Goll, Alyxandria M. Schubert, Jacques Izard, Brandi L. Cantarel, Beltran Rodriguez-Mueller, Jeremy Zucker, Mathangi Thiagarajan, Bernard Henrissat, Owen White, Scott T. Kelley, Barbara Methé, Patrick D. Schloss, Dirk Gevers, Makedonka Mitreva, and Curtis Huttenhower. Metabolic Reconstruction

for Metagenomic Data and Its Application to the Human Microbiome. *PLoS Comput Biol*, 8(6):e1002358, 06 2012.

2 Robert D. Finn, Jaina Mistry, John Tate, Penny Coggill, Andreas Heger, Joanne E. Pollington, O. Luke Gavin, Prasad Gunasekaran, Goran Ceric, Kristoffer Forslund, Liisa Holm, Erik L. L. Sonnhammer, Sean R. Eddy, and Alex Bateman. The Pfam protein families database. *Nucleic Acids Research*, 38(suppl 1):D211–D222, 2010.

3 The NIH HMP Working Group, Jane Peterson, Susan Garges, Maria Giovanni, Pamela McInnes, Lu Wang, Jeffery A. Schloss, Vivien Bonazzi, Jean E. McEwen, Kris A. Wetterstrand, Carolyn Deal, Carl C. Baker, Valentina Di Francesco, T. Kevin Howcroft, Robert W. Karp, R. Dwayne Lunsford, Christopher R. Wellington, Tsegahiwot Belachew, Michael Wright, Christina Giblin, Hagit David, Melody Mills, Rachelle Salomon, Christopher Mullins, Beena Akolkar, Lisa Begg, Cindy Davis, Lindsey Grandison, Michael Humble, Jag Khalsa, A. Roger Little, Hannah Peavy, Carol Pontzer, Matthew Portnoy, Michael H. Sayre, Pamela Starke-Reed, Samir Zakhari, Jennifer Read, Bracie Watson, and Mark Guyer. The NIH Human Microbiome Project. *Genome Research*, 19(12):2317–2323, 2009.

4 Daniel H. Huson, Suparna Mitra, Hans-Joachim Ruscheweyh, Nico Weber, and Stephan C. Schuster. Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, 21(9):1552–1560, 2011.

5 Dazhi Jiao, Yuzhen Ye, and Haixu Tang. Probabilistic Inference of Biochemical Reactions in Microbial Communities from Metagenomic Sequences. *PLoS Comput Biol*, 9(3):e1002981, 03 2013.

6 Alexander Kaever, Manuel Landesfeind, Mareike Possienke, Kirstin Feussner, Ivo Feussner, and Peter Meinicke. MarVis-Filter: ranking, filtering, adduct and isotope correction of mass spectrometry data. *BioMed Research International*, 2012, 2012.

7 Alexander Kaever, Thomas Lingner, Kirstin Feussner, Cornelia Gobel, Ivo Feussner, and Peter Meinicke. MarVis: a tool for clustering and visualization of metabolic biomarkers. *BMC Bioinformatics*, 10(1):92, 2009.

8 Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

9 Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114, 2012.

10 Heiner Klingenberg, Kathrin Petra Aßhauer, Thomas Lingner, and Peter Meinicke. Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. *Bioinformatics*, 2013.

11 Thomas Lingner, Kathrin Petra Aßhauer, Fabian Schreiber, and Peter Meinicke. CoMet - a web server for comparative functional profiling of metagenomes. *Nucleic Acids Research*, 39(suppl 2):W518–W523, 2011.

12 Victor M. Markowitz, I-Min A. Chen, Ken Chu, Ernest Szeto, Krishna Palaniappan, Yuri Grechkin, Anna Ratner, Biju Jacob, Amrita Pati, Marcel Huntemann, Konstantinos Liolios, Ioanna Pagani, Iain Anderson, Konstantinos Mavromatis, Natalia N. Ivanova, and Nikos C. Kyrpides. IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Research*, 40(D1):D123–D129, 2012.

13 Peter Meinicke, Kathrin Petra Aßhauer, and Thomas Lingner. Mixture models for analysis of the taxonomic composition of metagenomes. *Bioinformatics*, 27(12):1618–1624, 2011.

14 Folker Meyer, Daniel Paarmann, Mark D'Souza, Robert Olson, Elizabeth M Glass, Michael Kubal, Tobias Paczian, A Rodriguez, Rick Stevens, Andreas Wilke, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9(1):386, 2008.

15 Sang H Rhee, Charalabos Pothoulakis, and Emeran A Mayer. Principles and clinical implications of the brain–gut–enteric microbiota axis. *Nature Reviews Gastroenterology and Hepatology*, 6(5):306–314, 2009.