BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 2 2013, pages 215–222
doi:10.1093/bioinformatics/bts653

Systems biology

Advance Access publication November 9, 2012

Density parameter estimation for finding clusters of homologous proteins—tracing actinobacterial pathogenicity lifestyles

Richard Röttger[1,2,*], Prabhav Kalaghatgi[1,3], Peng Sun[1,3], Siomar de Castro Soares[4], Vasco Azevedo[4], Tobias Wittkop[5] and Jan Baumbach[1,2,3,6]

[1]Max Planck Institute for Informatics, [2]Center for Bioinformatics, [3]Cluster of Excellence for Multimodal Computing and Interaction, Saarland University, 66123 Saarbrücken, Germany, [4]Department of General Biology, Federal University of Minas Gerais, 31270-901 Belo Horizonte, Minas Gerais, Brazil, [5]Buck Institute for Age Research, Navato, CA 94945, USA and [6]Department of Mathematics and Computer Science, University of Southern Denmark, DK-5230 Odense M, Denmark

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Homology detection is a long-standing challenge in computational biology. To tackle this problem, typically all-versus-all BLAST results are coupled with data partitioning approaches resulting in clusters of putative homologous proteins. One of the main problems, however, has been widely neglected: all clustering tools need a density parameter that adjusts the number and size of the clusters. This parameter is crucial but hard to estimate without gold standard data at hand. Developing a gold standard, however, is a difficult and time consuming task. Having a reliable method for detecting clusters of homologous proteins between a huge set of species would open opportunities for better understanding the genetic repertoire of bacteria with different lifestyles.

Results: Our main contribution is a method for identifying a suitable and robust density parameter for protein homology detection without a given gold standard. Therefore, we study the core genome of 89 actinobacteria. This allows us to incorporate background knowledge, i.e. the assumption that a set of evolutionarily closely related species should share a comparably high number of evolutionarily conserved proteins (emerging from phylum-specific housekeeping genes). We apply our strategy to find genes/proteins that are specific for certain actinobacterial lifestyles, i.e. different types of pathogenicity. The whole study was performed with transitivity clustering, as it only requires a single intuitive density parameter and has been shown to be well applicable for the task of protein sequence clustering. Note, however, that the presented strategy generally does not depend on our clustering method but can easily be adapted to other clustering approaches.

Availability: All results are publicly available at http://transclust.mmci.uni-saarland.de/actino_core/ or as Supplementary Material of this article.

Contact: roettger@mpi-inf.mpg.de

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

Received on July 20, 2012; revised on October 16, 2012; accepted on October 29, 2012

*To whom correspondence should be addressed.

1 INTRODUCTION

Finding clusters of homologous proteins, i.e. clusters containing only paralogous and orthologous proteins, is a long-standing bioinformatics challenge in the post-genome era. Searching the exact phrase 'homology detection' with PubMed leads to 174 hits. The group of Peer Bork published one of the first review articles on 'Predicting functions from protein sequences' as early as 1998 (Bork and Koonin, 1998). The availability of next-generation sequencing technology provided us with almost 2000 whole-genome sequences, scattered over all domains of life (Sayers *et al.*, 2011). The annotation of the emerging sequences is difficult, error prone and impossible to perform in the wet laboratory for each gene/protein of each organism individually without appropriate bioinformatics software (Blanco and Abril, 2009; Tcherepanov *et al.*, 2006). To date, we have more than 5 million bacterial sequenced genes available for download from the National Center for Biotechnology Information (NCBI) database (Sayers *et al.*, 2011).

The usual starting point is a pairwise similarity matrix given by local alignment tools, such as Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1997), that assigns each pair of proteins a similarity value. Afterwards, we pipe this data into clustering tools, i.e. computational methods for partitioning data objects into groups such that the objects share common traits, which have been measured with the similarity function (Hartigan, 1975). Over the past years, many tools have been developed for this purpose. Andreopoulos *et al.* (2009) outline further biological application areas. For protein homology detection, the following tools have proven useful, and their accuracy is well studied: k-means, affinity propagation, Markov clustering and FORCE, as well as transitivity clustering (TC) (Enright *et al.*, 2002; Enright and Ouzounis, 2000; Frey and Dueck, 2007; Paccanaro *et al.*, 2006; Wittkop *et al.*, 2010).

Although most research concentrated on developing new more sophisticated data partitioning methods, one of the major problems has been widely neglected: all clustering tools need a (set of) density parameter(s) that adjust the number and the size of the clusters. A clustering tool cannot 'know' a priori if we seek to find protein families (restrictive parameters) or protein superfamilies (weak parameters), for instance. Although these parameters are crucial they are hard to estimate without gold standard data at

hand. Furthermore, it is difficult and time consuming to define a good gold standard, which consequently limits us to a small number of proteins from a limited number of organisms.

Clusters of homologous proteins across a number of organisms allow for studying lifestyle-specific genetic repertoires, i.e. the genes that have homologous counterparts in all organisms or in a specific set of organisms. Such studies can lead, for instance, to the discovery of mutual proteins shared only among pathogenic strains of a certain phyla, thus suggesting new drug targets and wet laboratory candidates for vaccine design. The quality of such studies is highly dependent on the quality of the clustering process and consequently dependent on the choice of the clustering method and a good estimate of the density parameter(s).

In this study, we present a robust method for selecting a suitable density parameter for TC for the task of protein homology detection. TC is a clustering method that has been shown to perform well when trying to identify protein families and protein superfamilies based on sequence similarity. Using all protein sequences from 89 actinobacteria, we build our method upon two assumptions: (i) clusters of size equal to the number of input organisms (here 89) are likely to contain housekeeping genes and thus should be over-represented, and (2) clusters greater than the number of input organisms are more likely to contain many false positives (non-homologous genes). Maximizing (i) while minimizing (ii) allows us to estimate a meaningful threshold for discovering clusters of homologous proteins without manually curated gold standard associations for any of the >300 000 proteins. Given this threshold, we compute and analyze the core genome of the 89 actinobacteria. We further divide them into four different groups of pathogenicity: non-pathogens (NPs), human pathogens (HPs), animal pathogens (APs) and opportunistic pathogens (OPs) (Supplementary Table S1). We then study the class-specific genetic repertoire of the 89 actinobacteria.

The phylum actinobacteria is one of the biggest clades of bacteria. Their members show a high diversity throughout different lifestyles and can cope with a variety of different habitats (Miao and Davies, 2010). Many of these bacteria are important for biotechnological production processes, as well as human and animal medicine (Ventura *et al*., 2007). Here, we focus on selected species of the following so-called CMNR group: corynebacteria, mycobacteria, nocardia and rhodococcus. Our main motivation for this study and our main focus of attention are the *Corynebacterium pseudotuberculosis*. It causes caseous lymphadenitis in animals (Williamson, 2001), with dramatic effects on livestock all over the world. All CMNR organisms selected for this study share common properties with impact on the design of effective vaccinations; they all share a common cell wall organization (Dorella *et al*., 2006), for instance. For vaccine design, accurate homology information about the protein space in this group is important, e.g. for reducing drug target side effects and negative effects on the other microorganisms that are part of the host's microbiome.

There have been several studies about the actinobacterial evolution [refer to Gao and Gupta (2012a)]. Most of them concentrated on phylogenetic tree reconstruction solely based on the DNA sequence information of the 16S RNA. Despite the many advantages of this method, it cannot provide insights into the evolutionary relationship on a species level (Stackebrandt, 2009). Gao and Gupta (2012b) used only a limited dataset of only a few genes that were expected to be conserved along the phylum for phylogenetic tree reconstruction. In several recent studies, best bidirectional hits from genome-wide all-versus-all BLAST results of all genes were used for homology detection [Karberg *et al*. (2011) or Gao *et al*. (2006), for instance]. This strategy, however, neglects the impact of careful BLAST cutoff evaluation, as well as the effect of transitive dependencies in the similarity function. Gene A may be similar to gene B, which is similar to gene C, but gene C is not similar to gene A. These problem instances can be 'repaired' with clustering tools, such as TC (Wittkop *et al*., 2011a). However, the problem of finding a reasonable density parameter remains with TC, as well as with any other clustering method.

In the following section, we briefly describe the actinobacterial dataset used. Afterwards, we give a short introduction to TC followed by our main contribution: a robust method for estimating a meaningful similarity threshold (TC's density parameter). We will describe how we study the robustness of our approach. We further support our strategy by computing a revised phylogenetic tree based on the whole genetic repertoire of the 89 actinobacteria. We will discuss our results and present core genomes specific to the four aforementioned pathogenicity classes.

2 METHODS

2.1 Data sources

We obtained the protein sequences in FASTA format from NCBI (Sayers *et al*., 2011) for the 89 sequenced and annotated actinobacteria of the CMNR group. See Supplementary Table S1 for a list of all species and a classification into the four pathogencity classes. We also give the associated disease where available. Our dataset comprises 344 421 proteins of 89 species: 27 corynebacteria, 55 mycobacteria, 6 rhodococcus and 1 nocardia.

2.2 Transitivity clustering

We decided to use TC (Wittkop *et al*., 2010) for this study for the following reasons: (i) It has proven to be a well-performing clustering tool for biological data in general and for protein sequence clustering in particular. (ii) TC requires only a single intuitive parameter to control cluster sizes and numbers. (iii) TC is comparably robust against noise in the data [Wittkop *et al*. (2011b) and Wittkop *et al*. (2010)]. (iv) The runtime and memory efficiency of TC allows for an evaluation of hundreds of thousands of sequences for varying parameter settings easily. Here, point (ii) is striking. In contrast to other more complicated clustering tools, with TC, we are not required to optimize two or more such parameters but only a single one. Note that our parameter estimation method would work with other clustering tools, even though the traversal over the parameter setting space can be more runtime intense.

First, we need to obtain a pairwise similarity measure for all protein sequences. The canonical way to archive a similarity between two proteins is using the $-\log_{10}$ of the BLAST *E*-value. Therefore, we performed a BLAST all-versus-all on all amino acid sequences using an *E*-value cutoff of 0.01. TC considers this input as a graph, with proteins being the nodes and the similarities being weighted edges. All edges below the given threshold

(TC's density parameter) are removed from the graph; all others we keep. By solving the weighted transitive graph projection problem, TC converts that graph into a transitive graph with the least edit costs. As edit costs, we use the accumulated differences between the similarities of the modified edges and the similarity threshold. The resulting fully connected cliques represent the final clusters and are reported as result set. For a more detailed description, please refer to Rahmann *et al.* (2007). We study the clustering result sets of TC runs for thresholds ranging from 8 to 100, corresponding to BLAST *E*-values of $1 \cdot 10^{-8}$ and $1 \cdot 10^{-100}$, respectively. We are confident that this range covers all meaningful thresholds.

2.3 Threshold estimation

To reasonably investigate the clustering results, the density parameter has to be set correctly such that most of the clusters actually contain groups of homologous proteins. In our study with >300 000 proteins from 89 different bacteria, we do not have a given gold standard that would allow us to find a reliable threshold. We will present an approach that only uses intrinsic indirect information of the dataset to determine such a threshold.

In what follows, n denotes the number of species. This number, i.e. $n = 89$ in our study, is constant and independent on the chosen threshold. Our first assumption is based on the expectation of observing significantly more clusters of size n than clusters of other sizes, as housekeeping genes and essential genes are expected to be conserved across all bacteria. Thus, they are more likely to cluster together in a group of exactly (or almost exactly) 89 proteins. In our analysis, we observed a peak in the cluster size distribution (Fig. 1) at $n = 89$, with most of these clusters containing exactly one protein from each of the 89 organisms. This gives evidence in favor of our first assumption. Setting the clustering threshold such that we maximize the size (height) of this peak would increase the number of allowed housekeeping gene. However, on the other hand, we cannot assess the number of false positives in these clusters directly, as we do not have a given reliable gold standard. What we require is a second measure for allowing us to minimize these false positives. Here, our second assumption is used: clusters with larger sizes (far bigger than n proteins) are likely to contain non-homologous proteins (i.e. false positives). The more a cluster size exceeds n, the more unlikely it is that this increase can be explained by true-positive paralogous proteins. We will use this assumption to receive a measure for handling the number of false positives.

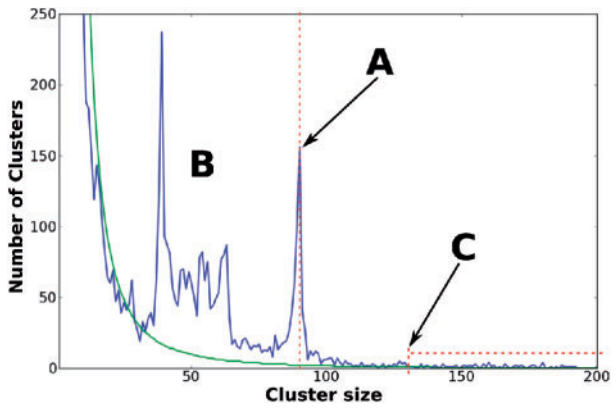

Fig. 1. Cluster size distribution of the 89 actinobacteria for similarity threshold 48. Arrow (A) highlights the core gnome peak at cluster size 89. These peaks in area (B) represent more specific core genomes, for example, all mutual proteins of the different mycobacteria/corynebacteria strains. The beginning of the unspecific clusters is marked by arrow (C)

Put in other words, our strategy is to vary the similarity threshold such that our TC-based clustering results yields the following two optimizations:

(1) Maximize the number of clusters of size n (most likely containing common housekeeping genes).
(2) Minimize number of large clusters (most likely containing many false positives).

To account for the first problem, we have to separate the desired peak from the background distribution to get the relative peak height compared with the surrounding area. The cluster size distribution seems to follow a power law. For that reason, we learned the best fitting discrete power-law:

$$P_{\alpha, x_{min}}(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{min})},$$

with $\zeta(\alpha, x_{min})$ being Rieman's Zeta function for the background distribution using the python tools provided in Clauset *et al.* (2007). Figure 1 depicts the cluster size distribution and the best fit power law for threshold 48 (we will explain below why we picked 48). Let $\hat{\alpha}_t$ and $\hat{x}_{min,t}$ be the approximated parameters for the best fitting power law for the cluster size distribution $D_t(x)$ for threshold t. The function $D_t(x)$ gives the absolute number of clusters of size x. Furthermore, m_t denotes the number of observations, i.e. the total number of clusters, again for threshold t. We now define the relative peak height $h_t(x)$ as:

$$h_t(x) = D_t(x) - P_{\hat{\alpha}_t, \hat{x}_{min,t}}(x) \cdot m_t,$$

where $P_{\hat{\alpha}_t, \hat{x}_{min,t}}(x) \cdot m_t$ denotes the expected number of observation of a perfect power law of the given sample size m_t, as $P_{\hat{\alpha}_t, \hat{x}_{min,t}}(x)$ is a probability function. In the following, we refer to the relative core genome height $h_t(n)$ as h_t.

Selecting the best threshold by optimizing only for h_t would lead to a weak threshold, as it would favor thresholds 'filling up' many of small clusters such that they contain n proteins in the end. To address this issue, we need to penalize the occurrence of unrealistic large clusters (false positives). In this work, we define such a cluster as a cluster containing $>1.5 \cdot n$ proteins. It is unlikely, that there are clusters of that size containing only real homologous and functional identical proteins because our actinobacterial dataset is diverse. A 'real' cluster of size $1.5 \cdot n$ would imply that at least half of the species must have undergone the same duplication event. That means this duplication event most likely happened at an evolutionary early time point in their common ancestor. On the other hand, the genetic variation was small enough such that these paralogous proteins still belong to the same cluster of homologous proteins. If that would happen to be a common case, one would also expect core genome peaks for paralogous proteins, e.g. at $2 \cdot n$ or $3 \cdot n$. We were not able to identify such a peak for any of the similarity thresholds. In conclusion, a cutoff for unrealistic big

cluster at $1.5 \cdot n$ is reasonable, and the accidental punishment of real paralogous clusters is negligible. Thus, we define the number of unrealistic clusters u_t for threshold t as:

$$u_t = \sum_{x > \frac{3}{2} \cdot n} D_t(x)$$

Optimizing only for that measure in turn would decrease the number of false positives but would increase the number of false negatives, i.e. homologous proteins put into two different clusters. Thus, in a final step, we combine both quality measures to a single overall quality value that we can assign to the TC results for the varying thresholds. As h_t and u_t are two completely different measures, we scale them to the range [0, 1], with T being the set of all used thresholds:

$$h'_t = \frac{h_t - \min(h_i, \forall i \in T)}{\max(h_i, \forall i \in T) - \min(h_i, \forall i \in T)}$$

$$u'_t = \frac{u_t - \min(u_i, \forall i \in T)}{\max(u_i, \forall i \in T) - \min(u_i, \forall i \in T)}$$

We calculate our final quality measure $Q(t)$ as the harmonic mean of both of them:

$$Q(t) = 2 \cdot \frac{h'_t \cdot (1 - u'_t)}{h'_t + (1 - u'_t)}$$

We are using $(1 - u'_t)$ such that lower numbers of unrealistic clusters result in the better quality measures. We may now use this approach to find that similarity threshold t of TC, which gives the best quality measure $Q(t)$. Figure 2 plots $Q(t)$ for several TC results for threshold ranging from 8 to 100.

2.4 Robustness analysis

So far, we have derived a quality measure $Q(t)$ using only intrinsic information of the provided dataset. On the other hand, actinobacteria are known to be diverse, suggesting that our dataset is biased. For example, we have 35 different strains of *Mycobacterium tuberculosis*, which are all likely to be more similar to each other than to the other actinobacteria. To respect for this potential bias, we split our datasets to investigate the stability of our approach. The following datasets were created:

- Myco-Only: all organisms of the genus mycobacteria (here: 55 species).
- Coryne-Only: same as Myco-Only but with all corynebacteria (here: 27 species).
- Rand-20: here, we randomly selected 20 of the 89 species without replacement. We created 20 such datasets, to get an impression of the variability of our approach.

As expected, our approach is limited by the level of biological diversity among the studied organisms. Although the actinobacterial phylum already is diverse, we also selected a dataset consisting of 40 different proteobacteria. Proteobacteria resemble one of the largest bacterial phyla with a huge genetic diversity (Stackebrandt *et al.*, 1988). In the remainder of this manuscript, we will call this the Proteo dataset. With Proteo, we aim to assess the stability and the limits of our approach for more diverse genomes. We used the protein sequences of 10 bacteria of each of the following four proteobacterial subgroups: alphaproteobacteria, betaproteobacteria, gammaproteobacteria and the

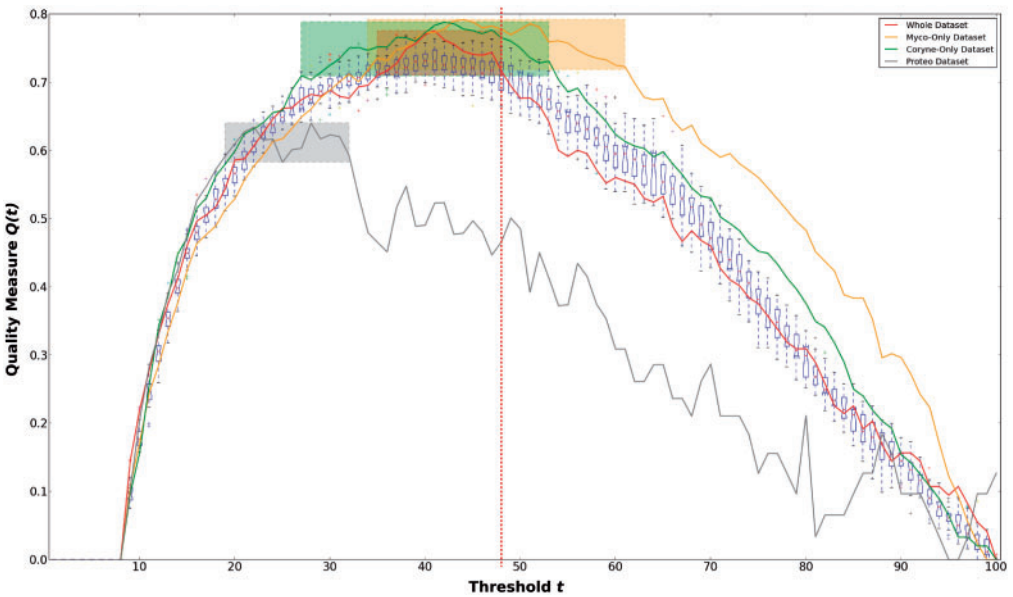

Fig. 2. In this figure, we plot our quality measure against the similarity threshold of TC. The red plot represents the quality measure $Q(t)$ for the entire dataset and the blue box-and-whisker plot in the background represents the variance and mean of all Rand-20 datasets (see text). The green and orange lines plot the quality measure for the Coryne-Only and the Myco-Only dataset respectively (see text). The three boxes in the plot mark the pick range, i.e. that range of thresholds where we see 10% of the best quality hits $Q(t)$. For the two rather phylum-biased datasets, i.e. Myco-Only and Coryne-Only, the pick range is larger than the pick range of the entire dataset. Notably, the pick range for the entire dataset is completely contained in the pick range of both, the Myco-Only and the Coryne-Only datasets. The gray line indicates the quality measure for the Proteo dataset. This dataset is too diverse for the presented quality measure, which is indicated by the generally lower quality values and the shifted box toward a weak threshold. The dotted red line indicates the threshold 48, which was chosen for the core genome analysis (see text)

delta/epsilon subdivisions (40 genomes in total). Please refer to Supplementary Table S2 for a detailed description. As we only use intrinsic information 'hidden' in the dataset, we rely on a certain level of homogeneity among the genomes to receive a reasonably large 'core-genome peak'. Hence, we may expect a slightly lower quality measures for the more diverse Proteo dataset highlighting the limits of our approach.

2.5 The actinobacterial phylogenetic tree

Given a meaningful clustering of homologous proteins, we may now calculate an interspecies similarity. Let $O = \{o_1, \ldots, o_n\}$ be the set of n organisms with $o_i = \{p_{i1}, \ldots, p_{in_i}\}$ as a set of n_i different proteins. Furthermore, let $C = \{c_1, \ldots, c_m\}$ be the set of m clusters. Furthermore, we define $\delta_{o_i}(c_k)$ to be the number of proteins that organism o_i has in cluster c_k. The function

$$\delta_{o_i,o_j}(c_k) = \begin{cases} 0 & \text{if } \delta_{o_i}(c_k) = 0 \vee \delta_{o_j}(c_k) = 0 \\ \delta_{o_i}(c_k) + \delta_{o_j}(c_k) & \text{otherwise} \end{cases}$$

denotes the number of mutual proteins in cluster c_k of organisms o_i and o_j if both organisms are represented by at least one protein. The similarity function $s(o_i, o_j)$ between two organisms o_i and o_j is now defined as:

$$s(o_i, o_j) = \frac{\sum_{c_i \in C} \delta_{o_i,o_j}(c_i)}{n_{o_i} + n_{o_j}}$$

This is basically the number of all mutual proteins of o_i and o_j scaled with the total number of proteins of both species. This scaling is done to prevent a bias of the similarity toward species with larger genomes (more genes).

These interspecies similarities fulfill all properties for a similarity function required for TC. To create a phylogenetic tree, we ran TransClust in hierarchical mode with option 'top-down'. Set in hierarchical mode, TransClust starts with a low threshold that is increased over several iterations. As result, in the first iteration, we obtain one big cluster containing all species. With more restrictive thresholds, the cluster(s) are divided into smaller clusters until each species ends in its own singelton cluster. The clustering result of all iterations is used to generate a phylogenetic tree. This tree is now based on the whole-genome repertoire of all actinobacteria. Note that we construct this (simple) tree for supporting our threshold estimation procedure, rather than introducing a new phylogenetic tree reconstruction methodology. One could also use other phylogenetic tree reconstruction approaches that are based on pairwise similarity functions.

3 RESULTS

3.1 Threshold estimation

First, we will discuss the evaluation of the threshold estimation method. Figure 2 illustrates the stability of our approach. In particular, the results of the 20 randomly sampled Rand-20 datasets are a good indicator for the reliability of our approach (refer to Table 1).

We define a threshold pick range $R_D = \{t_i, \ldots, t_k\}$ as the set of all thresholds, where the quality measure $Q(t)$ exceeds 90% of the best threshold of dataset D, i.e. that similarity threshold area where we find 10% of the best results. The pick ranges for the different datasets are marked with a box in Figure 2. For the complete dataset, we observe a pick range of $R_{\text{All}} = \{35, \ldots, 48\}$. In this range, the SD of the 20 Rand-20 datasets is only ~3% from the mean.

Table 1. This table shows the exact values for the evaluation of the threshold estimation

	20 × Rand-20			All data	
t	$\mu(Q(t))$	$\sigma(Q(t))$	Ratio (%)	$Q(t)$	Δ (%)
35	0.716	0.0152	2.13	0.711	−0.72
36	0.719	0.0196	2.73	0.717	−0.28
37	0.724	0.0195	2.69	0.728	0.56
38	0.729	0.0157	2.15	0.749	2.87
39	0.732	0.0186	2.54	0.757	3.46
40	0.731	0.0200	2.74	0.771	5.36
41	0.726	0.0192	2.64	0.776	6.84
42	0.723	0.0181	2.50	0.763	5.44
43	0.721	0.0215	2.98	0.756	4.90
44	0.719	0.0232	3.22	0.746	3.67
45	0.717	0.0232	3.23	0.741	3.34
46	0.715	0.0259	3.63	0.740	3.49
47	0.711	0.0257	3.61	0.744	4.64
48	0.706	0.0263	3.73	0.717	1.62

The left part represents the results of the 20 Rand-20 datasets, showing the mean [column '$\mu(Q(t))$'], the SD [column '$\sigma(Q(t))$'] and the percentage of the SD with respect to the mean (column 'Ratio'). For comparison, the right part displays values for the entire dataset, subdivided into a column showing the quality measure ['$Q(t)$'] and column 'Δ' displays the percentage deviation of '$Q(t)$' from '$\mu(Q(t))$'.

As we expected, the Proteo dataset (gray line) shows a lower quality $Q(t)$ than the actinobacterial dataset(s). We also observe a left-shifted pick range, i.e. toward a lower threshold, resulting in a less rigorous homology detection. The main reason for that is the smaller size of the proteobacterial core genome. This indicates that a single threshold for all species, ignoring the level or diversity, cannot sufficiently be detected, and, for instance, the alphaproteobacteria should be investigated separately from the betaprotebacteria.

We now discuss the Myco-Only dataset. The quality measure is better than for the other datasets, and the pick range is larger (the range of suitable similarity thresholds is bigger). This is mainly contributed to 35 strains of *M.tuberculosis* in a dataset with a total of only 55 species. As the different strains of *M.tuberculosis* are closely related, there is less variance in the clustering result with respect to the threshold. In other words, the proteins of the core genome cluster together early (for weaker thresholds), and variance only occurs for the less similar proteins of the non-tuberculosis species. Therefore, the relative core genome peak height stays pretty stable for more thresholds.

We suggest that all thresholds from the pick range are good candidates. We decided to choose the most restrictive one, i.e. 48, to further reduce the possibility of false positives in the homology detection and thus enhance the confidence in the presented actinobacterial core genome. We marked this threshold with a dashed line in Figure 2.

3.2 Pathogenicity as a genetic model

In this first application of our previously obtained clusters of homologous proteins, we study the relationship between the genetic repertoire and bacterial lifestyles, pathogenicity classes in our case. In particular, we are looking for genes that we find exclusively in a certain class of species, pathogens, for instance. Most likely, those genes would be conserved across several different pathogenic phyla and thus build a cluster that contains no proteins from NP organism. In the following, we work with the TC clusters that we obtained by using the conservative threshold of 48, estimated as described earlier. In the following, we will distinguish between four different types of pathogenicity:

- HPs (44 bacteria),
- APs (10 bacteria),
- OPs (23 bacteria),
- NPs (12 bacteria).

Note that OPs are generally not infectious but normally act commensal and do not harm the host. However, they can cause diseases in the case of a weak host's (immune) resistance (Rogers, 1963). Figure 3 depicts distributions of the cluster size overlaps between different combinations of the pathogenicity classes. Furthermore, we provide datasets containing all specific core genomes, the general core genome and all possible combinations, for example, clusters containing only proteins from HP and AP but not from OP and NP. These datasets are disjoint, e.g. the combined core genome of HP and AP does not contain the only-HP and only-AP clusters.

Some clusters were bigger than the number of species. Hence, some species must contribute with two or more proteins. That can happen by means of gene duplication events or because of clustering mistakes, i.e. false positives in the homology detection. Therefore, we provide the core genome datasets in two different 'flavors':

- Optimistic: all clusters with three or more proteins (includes paralogs).
- Conservative: only those clusters from the optimistic, where the number of proteins equals the number of involved species (no paralogs).

The general core genome includes only those clusters where all 89 species are involved. In the conservative case, these clusters are additionally limited to a size of exactly 89 proteins.

Figure 4 depicts a Venn diagram containing the number of clusters in the different categories. Of particular interest are 'distinctive clusters', which lack participating organisms of at least one type of pathogenicity. Thus, the core genome and all other clusters containing proteins of species of NP, HP, AP and OP are not on that list because they do not provide information on how to separate the different types of pathogenicity. One can clearly observe a connection between HPs and APs. A total of 1685 of 2888 HP and 3010 AP distinctive clusters are shared, which account for more of half of the respective distinctive clusters in HP and AP. In contrast to this, only 646 distinctive clusters contain proteins from HP and NP, and even less, 587, are conserved across AP and NP. The OPs seem to be less specific, as they overlap well distributed with all other categories (HP:1890, AP:

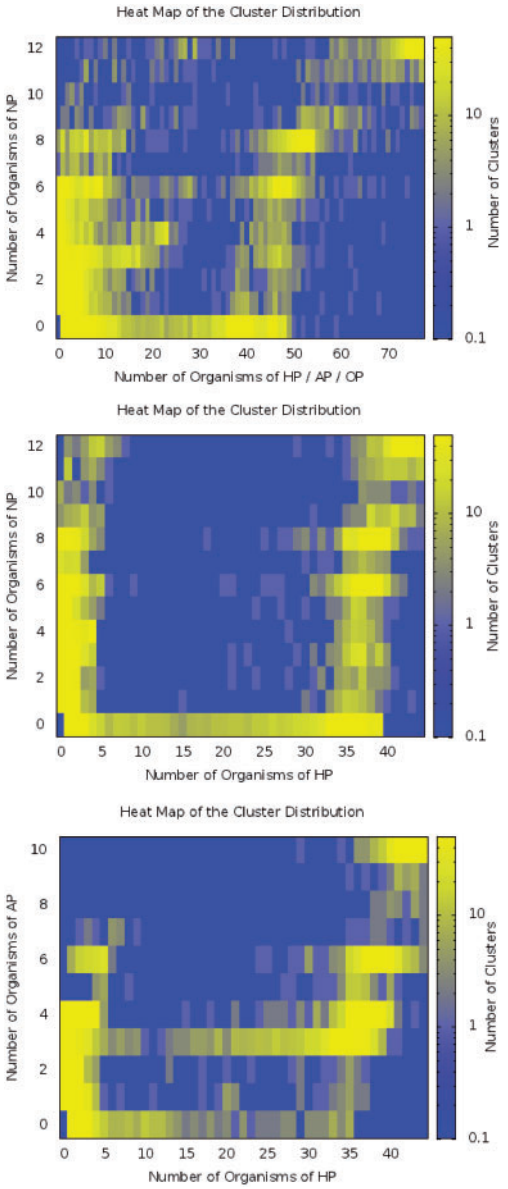

Fig. 3. Pathogenicity-specific cluster size distribution. The top picture, for example, represents on the *x*-axis all pathogens and on the *y*-axis only NPs. The colors encode the number of clusters that contain *x* '*x*-axis-type-pathogenic' and *y* '*y*-axis-type-pathogenic' species. We count each species only once (no paralogs). The core genome can be found in the top-right corner, whereas the top-left and bottom-right corners represent the exclusive core genomes. There are no peaks in the latter two areas, which means that there are no proteins that uniquely distinguish between the two pathogenicity classes. The heat maps for the remaining combinations can be downloaded form the web site of this article. Please note the log-scale of the color range

1820 and NP: 2250). All these results are based on the conservative core genome with TC threshold 48, although a similar tendency can be observed in the optimistic case (data not shown). Although our results do not totally fulfill our hope of seeing 100% pathogenicity class-specific proteins, our findings clearly indicate a certain genetic divergence between the pathogenicity lifestyles.

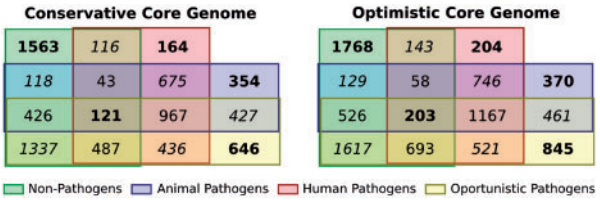

Fig. 4. These Venn diagrams depict the number of shared clusters in each possible intersection of the four different kinds of pathogenicity with at least three proteins (for conservative and optimistic; see text). These intersections are disjoint; for example, the intersection of the non-pathogenic core genome and the human pathogenic core genome does not contain the human-pathogenic-only clusters. The core genome contains all clusters, which contain proteins of all species. We marked the NP/HP/AP/OP-only clusters and the core genome itself with bold font; for an intersection of two areas, we used italic font

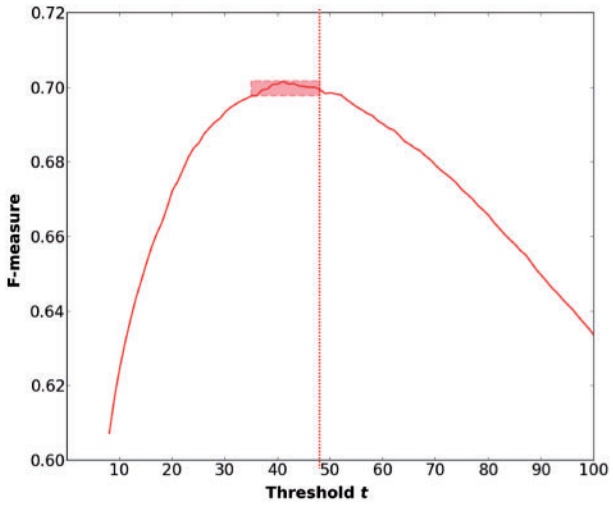

Fig. 5. Agreement with the OMA homology detection tool. Depicted is the development of the F-measure as a function of the clustering threshold. The red box marks the pick range derived by using our model $Q(t)$. Remarkably, the best F-measures (agreements with OMA) are achieved for clustering results with thresholds in pick range that we suggested using our method. The red dotted line indicates the threshold 48

All results are publicly available at http://transclust.mmci.uni-saarland.de/data/actino_core/.

3.3 Quality of the homology detection

As already mentioned, there is no gold standard for our bacteria. This poses problems for assessing the appropriateness of clustering methods for homology detection. Although slightly beyond the scope of this study, we like to discuss the agreement of our results with existing prediction-based homology repositories, EggNOG (Powell *et al*., 2012) and the Ortholog Matrix Project (OMA) (Dessimoz *et al*., 2005), for instance. OMA has the largest number of common species with our study and was shown to perform well (Dessimoz *et al*., 2006) on this task. We mapped 118 000 proteins of 30 species (9 corynebacteria, 17 mycobacteria, 1 nocardia and 3 rhodococcus) against our actinobacterial dataset by their IDs and their sequences. Please refer to Supplementary Table S2 for a list of mapped proteins and species. Finally, we removed all unmapped proteins from both actinobacterial datasets.

To assess the agreement of both datasets, i.e. our with that of OMA, we used the F-measure [a harmonic mean between precision and recall; see e.g. Wittkop *et al*. (2011a)]. The F-measure ranges between 0 and 1, where 1 means perfect agreement between both datasets.

We now varied the TC threshold and compared the results against OMA by using the F-measure to assess the agreement between both results sets (see Fig. 5). For the best threshold(s), the F-measure of 0.7 is good. Most notably, however, is the observation that the F-measure is best for thresholds almost exactly within the pick range that was suggested by our method (between 35 and 48). As with this article, we particularly focus on detecting a meaningful threshold, i.e. density parameter, for clustering algorithms (rather than studying the performance of clustering algorithms for homology detection in general); this observation further strengthens our main conclusion. Furthermore, it would be hard to make a qualified statement about the quality of OMA compared with ours, as both methods are based on computer predictions.

3.4 The actinobacterial phylogenetic tree

We used our aforementioned interspecies similarity to perform a hierarchical clustering. With this, we were able to construct a phylogentic tree based on the whole-genome repertoire of all the 89 actinobacteria. Supplementary Figure S1 depicts the resulting tree. Whenever a cluster is split into subclusters, with increasing threshold, we branch in the tree accordingly. If a cluster sticks together for x decreasing thresholds, we set the length of the branch to $\log(x+1)$. This is necessary mainly for optical reasons because some closely related organisms stick together for many threshold, which would result in long branches. One can see that most of the mycobacteria cluster together, whereas the other CMNR groups are slightly more separated. This observation is reasonable, given the different lifestyles, and was previously reported in other studies, see the review from Ventura *et al*. (2007), for instance. We emphasize that this tree is supposed to support our threshold estimation procedure, rather than introducing a new method for phylogenetic tree reconstruction.

4 CONCLUSION

To sum up, we studied the actinobacterial genetic repertoire with respect to four pathogenicity lifestyles. We used BLAST and TC for this purpose. Here, our main novel contribution was the estimation of a robust similarity threshold for TC. Therefore, we set the density such that we balance the size of the core genome (number of clusters with exactly 89 genes/proteins; putative true positives) and the number of unreasonably larger clusters (putative false positives) based on the cluster size distribution. We studied the robustness of our method by using random sampling and achieve stable and reasonable core genomes for similarity thresholds between 35 and 48. We receive similar results for the exclusive repertoire of the corynebacteria and mycobacteria, respectively. In conclusion, our results suggest that we may use the intrinsic information contained in the cluster size distribution, at least in the phylum actinobacteria, to deduce a reasonable density parameter for robust and accurate protein

homology clustering. For future work with bacterial genomes, we suggest using BLAST *E*-values between 10^{-35} (optimistic) and 10^{-48} (conservative) when using bidirectional BLAST hits only for homology detection. Remarkably, the same range is also suggested by comparing the agreement of our clustering result with the results from the OMA project.

Our method, however, is limited by the level of biological diversity among the set of species to be studied. As we only use the intrinsic information that is 'hidden' in the dataset, we rely on a certain level of homogeneity. Hence, we can expect a reduced accuracy for more diverse sets of genomes.

Here, we applied the methods to prokaryotes only. To use eukaryotic genomes, some adaptions would be necessary. Mainly, the factor that defines the number of false positives (unrealistically large clusters) must account for the fact that eukaryotes underwent more duplication events. We would suggest the training of this parameter against a small gold standard.

ACKNOWLEDGEMENT

J.B., R.R. and P.K. thank the Center for Bioinformatics Saar (ZBI).

Funding: The Cluster of Excellence for Multimodal Computing and Interaction of the German Research Foundation (DFG) (to J.B. and R.R.) and the International Max Planck Research School in Computer Science (to R.R.).

Conflict of Interest: none declared.

REFERENCES

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

Andreopoulos,B. *et al.* (2009) A roadmap of clustering algorithms: finding a match for a biomedical application. *Brief. Bioinform.*, **10**, 297–314.

Blanco,E. and Abril,J.F. (2009) Computational gene annotation in new genome assemblies using GeneID. *Methods Mol. Biol.*, **537**, 243–261.

Bork,P. and Koonin,E.V. (1998) Predicting functions from protein sequences–where are the bottlenecks? *Nat. Genet.*, **18**, 313–318.

Clauset,A. *et al.* (2007) Power-law distributions in empirical data. *SIAM Rev.*, **51**, 661–703.

Dessimoz,C. *et al.* (2005) OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: introduction and first achievements. *Comp. Genomics*, **3678**, 61–72.

Dessimoz,C. *et al.* (2006) Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Res.*, **34**, 3309–3316.

Dorella,F.A. *et al.* (2006) Corynebacterium pseudotuberculosis: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. *Vet. Res.*, **37**, 201–218.

Enright,A.J. and Ouzounis,C.A. (2000) GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics*, **16**, 451–457.

Enright,A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.

Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.

Gao,B. and Gupta,R.S. (2012a) Microbial systematics in the post-genomics era. *Antonie Van Leeuwenhoek*, **101**, 45–54.

Gao,B. and Gupta,R.S. (2012b) Phylogenetic framework and molecular signatures for the main clades of the phylum actinobacteria. *Microbiol. Mol. Biol. Rev.*, **76**, 66–112.

Gao,B. *et al.* (2006) Signature proteins that are distinctive characteristics of Actinobacteria and their subgroups. *Antonie Van Leeuwenhoek*, **90**, 69–91.

Hartigan,J.A. (1975) *Clustering Algorithms*. John Wiley & Sons, New York, NY.

Karberg,K.A. *et al.* (2011) Similarity of genes horizontally acquired by *Escherichia coli* and *Salmonella enterica* is evidence of a supraspecies pangenome. *Proc. Natl Acad. Sci. USA*, **108**, 20154–20159.

Miao,V. and Davies,J. (2010) Actinobacteria: the good, the bad, and the ugly. *Antonie Van Leeuwenhoek*, **98**, 143–150.

Paccanaro,A. *et al.* (2006) Spectral clustering of protein sequences. *Nucleic Acids Res.*, **34**, 1571–1580.

Powell,S. *et al.* (2012) eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.*, **40**, D284–D289.

Rahmann,S. *et al.* (2007) Exact and heuristic algorithms for weighted cluster editing. *Comput. Syst. Bioinformatics Conf.*, **6**, 391–401.

Rogers,F.B. (1963) Medical subject headings. *Bull. Med. Libr. Assoc.*, **51**, 114–116.

Sayers,E.W. *et al.* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, D38–D51.

Stackebrandt,E. (2009) *Phylogeny based on 16SrRNA/DNA. eLS* [Epub ahead of print, doi: 10.1002/9780470015902.a0000462.pub2, September 15, 2009].

Stackebrandt,E. *et al.* (1988) Proteobacteria classis nov., a name for the phylogenetic taxon that includes the purple bacteria and their relatives. *Int. J. Syst. Bacteriol.*, **38**, 321–325.

Tcherepanov,V. *et al.* (2006) Genome Annotation Transfer Utility (GATU): rapid annotation of viral genomes using a closely related reference genome. *BMC Genomics*, **7**, 150.

Ventura,M. *et al.* (2007) Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. *Microbiol. Mol. Biol. Rev.*, **71**, 495–548.

Williamson,L.H. (2001) Caseous lymphadenitis in small ruminants. *Vet. Clin. North Am. Food Anim. Pract.*, **17**, 359–371, vii.

Wittkop,T. *et al.* (2010) Partitioning biological data with transitivity clustering. *Nat. Methods*, **7**, 419–420.

Wittkop,T. *et al.* (2011a) Comprehensive cluster analysis with transitivity clustering. *Nat. Protoc.*, **6**, 285–295.

Wittkop,T. *et al.* (2011b) Extension and robustness of transitivity clustering for protein–protein interaction network analysis. *Internet Math.*, **7**, 255–273.