UNIVERSITÄT AUGSBURG
Universitätsbibliothek

OPEN ACCESS
OPUS AUGSBURG

GBDP on the grid: a genome-based approach for species delimitation adjusted for an automated and highly parallel processing of large data sets

Jan P. Meier-Kolthoff, Alexander F. Auch, Hans-Peter Klenk, Markus Göker

Angaben zur Veröffentlichung / Publication details:

Meier-Kolthoff, Jan P., Alexander F. Auch, Hans-Peter Klenk, and Markus Göker. 2014. "GBDP on the grid: a genome-based approach for species delimitation adjusted for an automated and highly parallel processing of large data sets." In *Hochleistungsrechnen in Baden-Württemberg-Ausgewählte Aktivitäten im bwGRiD 2012: Beiträge zu Anwenderprojekten und Infrastruktur im bwGRiD im Jahr 2012*, edited by Janne Chr. Schulz and Sven Hermann, 83–102. Karlsruhe: KIT Scientific Publishing. https://doi.org/10.5445/KSP/1000039516.

Nutzungsbedingungen / Terms of use:

licsonst

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
Sonstige Open-Access-Lizenz
Weitere Informationen finden Sie unter: / For more information see:
https://www.bibliothek.uni-augsburg.de/opus/lic_sonst.html

Hochleistungsrechnen in Baden-Württemberg *Ausgewählte Aktivitäten im bwGRiD*

Beiträge zu Anwenderprojekten
und Infrastruktur im bwGRiD 2012

Janne Chr. Schulz und Sven Hermann (Hrsg.)

KIT Scientific Publishing

Impressum

Karlsruher Institut für Technologie (KIT)
KIT Scientific Publishing
Straße am Forum 2
D-76131 Karlsruhe

KIT Scientific Publishing is a registered trademark of Karlsruhe Institute of Technology. Reprint using the book cover is not allowed.

www.ksp.kit.edu

This document – excluding the cover – is licensed under the Creative Commons Attribution-Share Alike 3.0 DE License (CC BY-SA 3.0 DE): http://creativecommons.org/licenses/by-sa/3.0/de/

The cover page is licensed under the Creative Commons Attribution-No Derivatives 3.0 DE License (CC BY-ND 3.0 DE): http://creativecommons.org/licenses/by-nd/3.0/de/

Print on Demand 2014

ISBN 978-3-7315-0196-1
DOI 10.5445/KSP/1000039516

GBDP on the grid: a genome-based approach for species delimitation adjusted for an automated and highly parallel processing of large data sets

Jan P. Meier-Kolthoff[1]*†, Alexander F. Auch[2]†,
Hans-Peter Klenk[1], Markus Göker[1]

[1]Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures, Braunschweig
[2]Eberhard-Karls-Universität, Tübingen, Germany

Abstract: The GBDP approach (Genome Blast Distance Phylogeny) is a digital, genome-based method for the calculation of distances between organisms that can be further utilized for the inference of phylogenetic trees. Moreover, it is a technological advancement over the tedious and error-prone wet-lab technology DNA-DNA hybridization (DDH), on which the prokaryotic species concept is ultimately based. GBDP provides for an exact calculation of distances between pairs of entirely or partially sequenced genomes. These are compared using local-alignment tools such as BLAST and the resulting intergenomic matches subsequently transformed into a genome-to-genome distance (GGD). The Genome-to-Genome Distance Calculator (GGDC) web service is implementing the GBDP approach and is publicly available under http://ggdc.gbdp.org/. We advanced the GBDP approach by developing a high-performance cluster (HPC) version that is capable of executing large amounts of genome comparisons by using the parallel nature of compute grids, such as the bwGRiD. Pairwise distances for a novel exemplary data set of 15 eukaryotic Basidiomycota genomes – about order of magnitude larger than common prokaryotic genomes – were calculated, a phylogenetic tree reconstructed and subsequently analysed. The new implementation is boosting the conduction of large genome-based experiments and can thus provide new and even more detailed phylogenetic insights into groups of organisms for which genomic data are available. Benchmarks revealed that the total computation time of the Ba-

* E-mail of corresponding author: jan.meier-kolthoff@dsmz.de
† These authors contributed equally to this work

sidiomycota data set is almost negligible (within 1h) due to the linear speed-up provided by the cluster.

1 Introduction

This article describes an HPC (high-performance cluster) implementation of Genome-Blast Distance Phylogeny [Henz2005, Auch2006, Auch2009, Auch2010, Auch2010a], a bioinformatics approach for the calculation of distances between completely or partially sequenced genomes. GBDP is a basis for the inference of phylogenetic trees or networks and can also be used as a technically improved genome-sequence-based alternative for tedious and error-prone wet-lab techniques such as DNA-DNA hybridization (DDH).

The 2011 EHEC outbreak in Germany and other European countries called to mind that a quick identification and classification of pathogenic microorganisms is of utmost importance for proper reactions to such crises [Beutin2012]. Knowledge about key properties such as infectious spreading, antibiotic resistances, optimal growth conditions and morphology help to develop a proper cure and to eliminate or at least to reduce the risk of new infections. For this and other reasons, one important sub-discipline in biology is taxonomy; the identification and classification of species according to a given scheme. In recent decades, microbial taxonomy was richly informed by phylogenetics, the study of evolutionary relatedness among groups of organisms on the basis of molecular sequencing data. It reached a preliminary climax in 1977 when Carl Woese used the DNA sequence of the 16S ribosomal subunits of bacterial strains to introduce a revolutionary classification scheme that contained the *Archaea* as a third domain along with *Bacteria* and *Eucaryota* [Woese1977]. These ribosomal sequences are ancient and distributed over all lineages of life with little or no horizontal gene transfer. However, the more 16S sequences were obtained from both *Bacteria* and *Archaea* the more they turned out not to be suitable as sole universal phylogenetic marker [Klenk2010], i.e., sequences among some microbial groups are almost identical (high conservation), although the underlying organisms are only distantly related. Hence, new approaches were required and finally became apparent with the advent and rapid advances in

whole-genome sequencing [Mardis2011]. These offered new perspectives for genome-based identification and classification of microorganisms: by using whole genomes – or at least a large number of gene-families – the phylogenetic resolution can be substantially increased [Henz2005].

This principle led to the aforementioned development of the Genome-Blast Distance Phylogeny approach (GBDP) that was originally devised in 2005 [Henz2005] and subsequently improved. The latest installation of the software is especially designed for the use on high-performance clusters – such as those provided by bwGRiD [bwgrid2012] – and thus capable of handling large data sets. The underlying principle of the GBDP software itself is as follows: in the first step two genomes A and B are locally aligned using tools such as BLAST [Altschul1990], which search for local similarities and thus produce a set of high-scoring segment pairs (HSPs; these are intergenomic stretches matching up to a certain extent). In the second step, information contained in these HSPs (e.g., the total number of identical base pairs) is transformed into a single genome-to-genome distance value (GGD) by the use of a specific distance formula. In principle, GBDP could as well process proteomic data instead of genomic ones.

We describe the basic principle of the GBDP approach and the extensions necessary for running it on compute clusters. Finally, we demonstrate the implementation on the basis of an exemplary data set of fungal genomes. These eukaryotic genomes are by an order of magnitude larger than those of *Bacteria* and *Archaea* thus making the computation of intergenomic distances a more challenging task. We observed that the GBDP implementation is able to handle input of this size without requiring any extra adjustments to the algorithms. We further assessed the general suitability of our grid-based implementation for this kind of analyses and could confirm that the highly parallel nature of the bwGRiD is boosting studies of the aforementioned kind: overall computation times are significantly reduced, allowing our experiment to be completed within a single hour; computations that would have taken days – if not weeks – on a standard desktop PC.

2 Materials and Methods

2.1 The GBDP principle

The GBDP approach has been discussed in several publications [Henz2005, Auch2006, Auch2009, Auch2010, Auch2010a], thus, we will only describe the basic mechanisms and principles of the algorithm. The pipeline is primarily subdivided into two phases: in the first phase, a genome X is BLASTed against a genome Y and vice versa. Here, the term "BLASTed" denotes the application of one out of six supported local-alignment programs: BLAST+ [Camacho2009], NCBI-BLAST [Altschul1990], MUMmer [Kurtz2004], BLAT [Kent2002], WU-BLAST [Altschul1990] and BLASTZ [Schwartz2003]. The resulting matches between both genomes are called high-scoring segment pairs (HSPs). In a second phase, GBDP is filtering these HSPs according to one out of three available algorithms: "greedy", "greedy-with-trimming" or "coverage". Each one of these is accounting differently for special cases such as overlapping HSPs (i.e., two HSPs that share a specific part within the query or subject genome). Briefly, the algorithms define (i) whether the smaller overlapping HSP is removed (called "greedy"; can lead to information loss but computationally fast), (ii) the overlapping parts are merged ("greedy-with-trimming"; also prevents overlapping genome parts to be considered twice but is more compute-intense) or (iii) only the amount of the genome is accounted that is actually covered by HSPs. At the end of the second phase, these matches are transformed to a single distance value $d(X,Y)$ by applying one out of ten available distance formulas (d_0 to d_9). These formulas are basically different flavours of how distances between genomes can be computed on the basis of their respective HSP sets. For example d_4 and d_5 are preferable when dealing with partially incomplete genomes, whereas d_1 and d_7 (and their logarithmized variants d_3 and d_9) are especially suitable if two complete genomes are significantly differing in size (see [Auch2010] for detailed descriptions). Optional is the generation of bootstrap or jackknife replicate distances, which is based on a random sampling of the above mentioned HSPs (prior to the distance calculation). The GGDC can be requested to compute these replicates by adding either a bootstrap or jackknife generator object to the YAML request (see

Fig. 3). These sampling implementations should not be confused with the type of bootstrapping/jackknifing that is usually applied to multiple sequence alignments.

Regarding the implementation, the first phase is encapsulated in a so-called "match request" which triggers the aforementioned comparison of two genomes and finally provides the results in a standardized output format (see below for details). In turn, the latter output is read during the second phase and used by a separate "distance request" to finally compare distance values that are also stored in a proper output format. The division into these two types of requests is due to the fact that distance requests can be conducted under different settings without requiring the matches requests to be repeated. This procedure is saving both computation and storage resources.

2.2 Adjusting GBDP for the bwGRiD

The GBDP software was originally devised to be run on local machines, thus requiring a couple of extensions to the initial concept [Auch2010a]. The final workflow of the new pipeline is shown in Fig. 1. The following list summarizes the requirements of the implementation and how they were finally embedded into the bwGRiD environment.

Languages The GBDP software is written in Java (1.5+) whereas the grid-related part (e.g., for the scheduling) is implemented in the Ruby language (1.8.7+).

Interfaces All request and output files are uniformly provided in the YAML cross-language format – a human-readable data serialization format (http://www.yaml.org/). This allowed us to upload and/or download data to and from the bwGRiD. Fig. 2 and 3 show sample requests and output files for both match and distance calculations.

Bulk generation of the request Most of the phylogenomic analyses conducted on the bwGRiD are based on comparably large sets of genomes thus requiring an automated generation of match and distance requests as well as a simultaneous validation of the genome files' for-

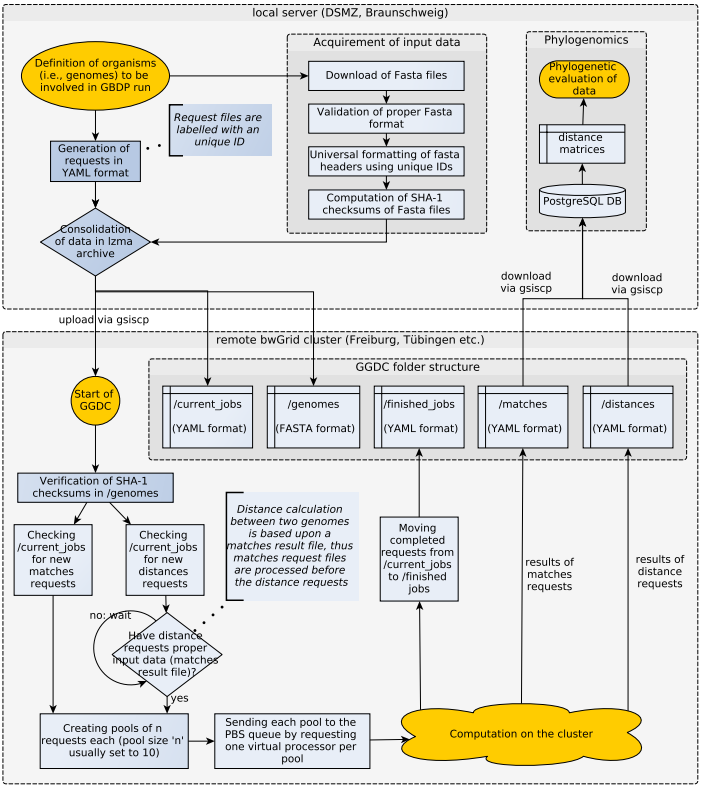

Figure 1: The GBDP pipeline as embedded on the grid. The pipeline is subdivided into three fully automated steps: (i) preparation of input data (both request and FASTA files) and upload to the grid, (ii) start of the GGDC software and submission of jobs to the PBS queue and, (iii) download of the results to the local server and final phylogenetic analysis. Even though the use of the grid infrastructure is preferable, the GBDP software can be run on any type of server or local PC. In that case, the requests are processed consecutively.

mat before the data is transferred to the grid. For each FASTA file an SHA-1 checksum is computed that can be verified once the file has been transmitted to the remote site (i.e., bwGRiD). Moreover, genome files and genome parts are internally represented by unique IDs to avoid the problem of mislabelled or duplicate genome parts.

Job dispatcher Match and distance requests as well as FASTA files are the initial input for the Genome-to-Genome Distance Calculator software (GGDC) – the software which is implementing the aforementioned GBDP approach. The Job dispatcher facility checks the available requests and validates the SHA-1 checksums of the uploaded FASTA files. Afterwards the requests are pooled (default pool size is 10), i.e. they are assigned to one virtual processor on the grid and consecutively executed. On the background of large sets of requests, this strategy substantially reduces the number of virtual processors that have to be requested. The dispatcher task can be either triggered on the grid by a cron job or from an arbitrary external logic.

2.3 GBDP-based phylogenetic analysis of 15 species from the Basidiomycota phylum

Together with the Ascomycota, Basidiomycota is one of two large phyla that comprise the subkingdom Dikarya within the kingdom Fungi [Hibbett2007]. We queried the Genomes Online Database [Pagani2012] for all species from the Basidiomycota phylum and restricted the outcome to those for which genomic data was available. Seven strains were marked as "complete and published" whereas eight strains were in state "incomplete". Genomic data was downloaded from NCBI for 15 strains including incomplete ones (see Tab. 1); GBDP is capable of calculating distances even between these types of data [Auch2010].

As next step, all 105 distinct pairwise distances between these genomes had to be calculated. This means that 105 match requests had to be completed, followed by the same number of distance requests. By default, the GBDP software calculates intergenomic distances under all distance formulas d_0 to d_9, thus resulting in a separate distance matrix for each of them.

```
1  --- !BlastJob
   jobId: 1
3  outputDirectory: matches
   EValueThreshold: 10.0
5  localAlignmentTool: !!BlastPlusWrapper
     lowComplexityFilter: 1
7    softMasking: 1
     wordLength: 11
9    maskedBlastDbAlgorithm: dustmasker
     dbSoftMaskingAlgorithmId: 11
11
   genomes:
13 - !Genome
     name: Coprinopsis_cinerea_okayama7_130
15   genomeId: 7
   - !Genome
17   name: Cryptococcus_neoformans_B_3501A
     genomeId: 6
```

```
---
2  jobId: 1
   genomeId1: 6
4  genomeId2: 7
   matchCollections:
6  - !MatchCollection
     querySequenceId: 6::20546
8    hitSequenceId: 7::20614
     matches:
10   - !Match
       alignmentLength: 1874
12     bitScore: 2480.9185
       eValue: 0.0
14     hitEnd: 2038
       hitStart: 3902
16     identity: 1688
       queryEnd: 2005279
18     queryStart: 2003429
```

Figure 2: A sample match request (**left**) with the resulting output (**right**) in YAML format. Line 12 in the match request denotes an array of the two genomes to be compared, whereas line 5 defines a hash table containing settings under those the genomes are locally aligned. Line 9 in the output file defines an array of `BLAST` hits. For convenience the BLAST statistics are also contained in the matches output file (not shown). The files can be found at `http://www.bw-grid.de/projekte/`.

These matrices were used for the phylogenetic reconstruction, using an improved minimum-evolution approach (FastME [Desper2002]). The resulting trees were compared with the NCBI classification of the included organisms to assess their accuracy in representing evolutionary relationships. NCBI is not an authoritative source for taxonomy but already includes recent improvements of the higher-order classification of fungi [Hibbett2007]. Comparison was done using the c-score [Henz2005, Auch2006] which corrects for the insufficient resolution of the classification-based reference tree.

2.4 Benchmark setup

On the one hand we measured (i) the total execution time (walltime) for each of the aforementioned 105 matches and 105 distances requests, and (ii)

```
--- !DistanceJob
jobId: 106
matchesJobId: 1
distanceAlgorithms:
- Trimming
eValueFilterThreshold: 0.01
outputFile: 106_distances.yaml
genomes:
- !Genome
  name: Coprinopsis_cinerea.fna
  genomeId: 7
- !Genome
  name: Cryptococcus_neoformans.fna
  genomeId: 6

replicateGenerator: !BootstrapGenerator
  numberOfReplicates: 100
  randomSeed: 10027302
```

```
--- !DistanceData
distanceAlgorithm: Trimming
jobId: 106
matchesJobId: 1
replicateId: 0
taxonId1: 6
taxonId2: 7

distanceEntries:
- distance: 0.99
  distanceType: D0
  status: OK
  variance: 3.784E-12
  ...
- distance: 8.23
  distanceType: D9
  status: OK
  variance: 1.914E-4
```

Figure 3: A sample distance request (**left**) with the resulting output (**right**) in YAML format. Line 8 of the distance request denotes an array of two genomes. Line 4 defines a list of algorithms that should be applied during the distance calculation. The additional computation of bootstrap or jackknife replicates can be activated by adding either a "!BootstrapGenerator" or "!JackknifeGenerator" object (here: bootstrap replicates are requested). In the distance output file, the distance values for all ten distance formulas (d_0 - d_9) are listed (in this example only d_0 and d_9). Bootstrap or jackknife replicate distances are numbered according to a replicate ID (line 5). The files can be found at `http://www.bw-grid.de/projekte/`.

the file sizes of the YAML files containing the match results. On the other hand, we approximated the intergenomic search space of each genome pair by calculating the product of both genomes' lengths. We investigated if and how the execution time was affected in dependence of search space size and, secondly, assessed whether the latter also affected the size of the match results. For means of a better interpretation of the results we added information on the relatedness of the denoted genome pairs by calculating patristic distances from the NCBI taxonomy tree of the 15 Basidiomycota genomes.

Strain	Size (Mb)	GOLD ID	NCBI accession
Phanerochaete chrysosporium RP-78	29	Gc00187	AADS00000000
Cryptococcus neoformans JEC 21	19	Gc00247	AE017341
Ustilago maydis 521	20	Gc00507	AACP00000000
Malassezia globosa CBS 7966	8.7	Gc00704	NZ_AAYY00000000
Laccaria bicolor S238N-H82	74	Gc00714	NZ_ABFE00000000
Postia placenta MAD 698-R	59	Gc00946	NZ_ABWF00000000
Schizophyllum commune H4-8	52	Gc01524	NZ_ADMJ00000000
Cryptococcus neoformans gattii R265	17	Gi00179	AAFP00000000
Malassezia restricta CBS 7877	4.7	Gi01942	AAXK01000000
Moniliophthora perniciosa FA553	12	Gi00175	ABRE00000000
Coprinopsis cinerea okayama 7#130	36	Gi01113	AACS00000000
Mixia osmundae IAM 14324	13	Gi07938	BABT00000000
Cryptococcus neoformans var. *grubii* serotypeA H99	19	Gi00180	AACO02000000
Cryptococcus neoformans B-3501A	20	Gi00177	NZ_AAEY00000000
Puccinia graminis tritici CRL 75-36-700-3	87	Gi01690	AAWC01000000

Table 1: Strains from the Basidiomycota phylum used in the GBDP analysis. Information as retrieved from the GOLD database [Pagani2012]. The genome status is either “complete and published" (if the GOLD ID starts with “Gc") or “incomplete" (if the ID starts with “Gi"). The genome sizes are provided in mega base pairs (Mb). The genomic data can be downloaded from `http://www.ncbi.nlm.nih.gov/sites/genome/`.

All calculations were made on the `bwGRiD` cluster in Freiburg which is made up of 140 nodes, each one equipped with an IBM-Bladeserver HS21XM containing two Intel Xeon E5440 CPUs (Harpertown) with a clock frequency of 2.83 GHz. Moreover, each node comes with 16 GByte of main memory.

3 Results

3.1 Phylogenetic analysis of the Basidiomycota data set

The resulting branch support was uniformly high (see Fig. 4 for an example), but the c-scores varied, depending on the distance formula used and to a much lesser degree on whether "greedy", "greedy-with-trimming" or "coverage" was used. The lowest c-scores of 0.333 (i.e., the least correspondence with the reference classification) were obtained with formulas d_4 and d_5, the highest c-scores of 0.917 with formulas d_3 and d_9. Whether "greedy", "greedy-with-trimming" or "coverage" was used did not affect the c-scores of formulas d_3 and d_9.

One of the best trees is shown in Fig. 4. The subphyla (Pucciniamycotina: *Mixia* and *Puccinia*; Ustilaginomycotina: *Malassezia* and *Ustilago*; Agaricomycotina: all other included organisms) are all well recovered. The sole discrepancy with the NCBI classification is that *Phanerochaeta* and *Postia* ("Agaricomycetes incertae sedis") are placed within Agaricales, closer to the other Agaricales than *Schizophyllum*. Their names are boxed in Fig. 4. Because this discrepancy is caused by organisms of uncertain taxonomic placement ("incertae sedis"), the GBDP phylogeny might well be regarded as in full agreement with the classification.

3.2 Benchmark results

The total execution time of all match requests added up to 19.75 hours and that of the distance requests was 59.9 hours. The average execution time of all match requests was about 11 minutes and that of the distance requests about 34 minutes. The average size of all match results was about 40 MByte (total size: 4 GByte). Fig. 5 shows the benchmark results.

4 Discussion

4.1 Phylogenetic analysis

Distance formulas d_4 and d_5 (which is the logarithmized version of d_4) ignore the genome lengths and only relate the total HSP length and the total number of identical base pairs within HSPs to each other; in contrast, d_3 and

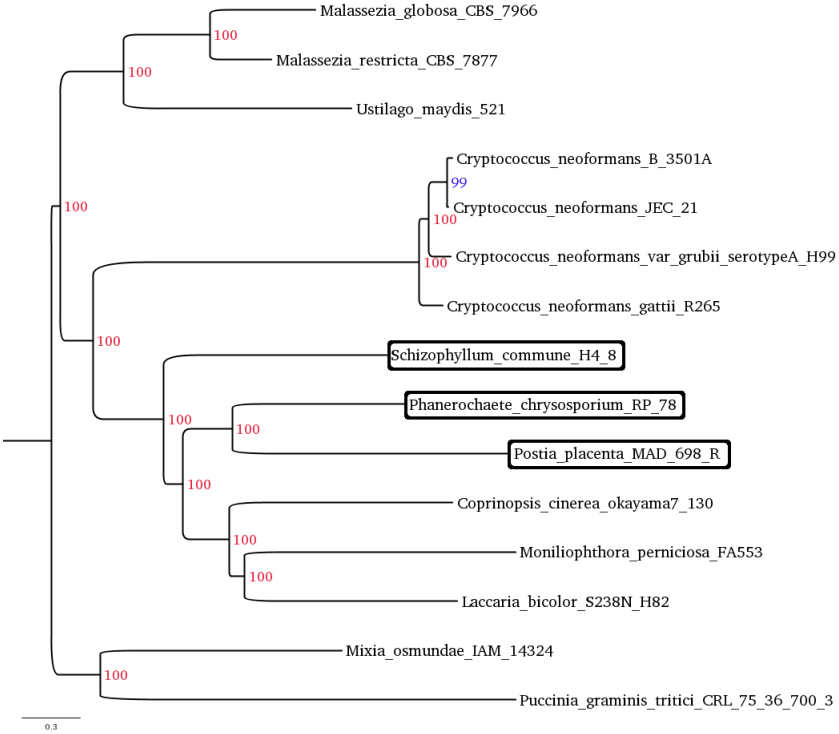

Figure 4: GBDP-based phylogeny of 15 Basidiomycota genomes reconstructed with FastME. The following GBDP settings were used: BLAST+ with default settings, trimming algorithm and distance formula d_9. The root was set via mid-point rooting [Farris1972]. The boxed species names are conflicting with the NCBI classification, but only with respect to taxa of uncertain position.

d_9 are both logarithmized distances and either relate the total HSP length or the total number of identical base pairs within HSPs to the smaller of the two respective genome lengths. When comparing distantly related organ-

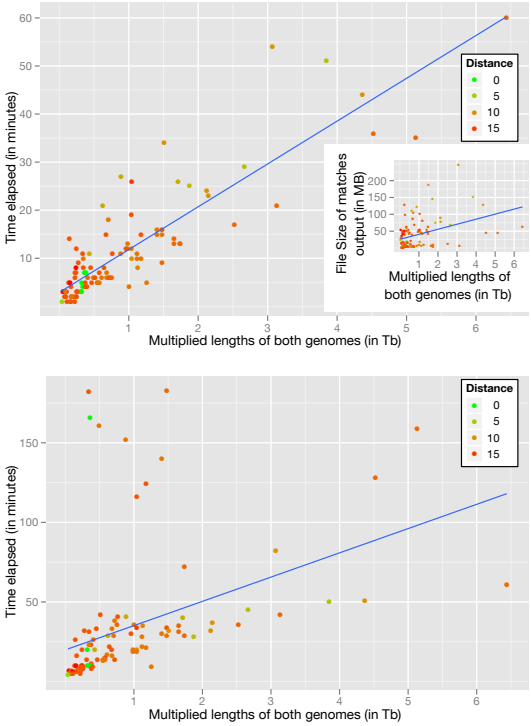

Figure 5: Benchmarks regarding the performance of both matches (top) and distance requests (bottom). In both cases, the size of the intergenomic search space affects the execution time as well as the size of the matches output files (fan-like shape). The right plot contains eight striking outliers with respective total times above 100 minutes opposed to a comparably small search space ($<$ 2Tb). In order not to base the interpretation of the data solely on a search space criteria, we also determined the taxonomic distance (as provided by the NCBI taxonomy) as an additional one (green=low, red=high). Both figures were created using the `R` package `ggplot` [Wickham2009].

isms, d_4 and d_5 may suffer from saturation effects because the HSPs are reduced to matches between strongly conserved genes. For this reason, d_4 and d_5 distances may even decrease with decreasing evolutionary relatedness. The main benefits of d_4 and d_5 are elsewhere [Auch2010, Auch2010a]. Conversely, logarithmizing the distances helps against saturation, and using the smaller of the two respective genome lengths as the denominator in the distance formula corrects against huge differences between genome sizes [Henz2005]. Hence, the relative performance of the distance formulas in the phylogenomic problem studied here is not surprising, given previous results [Henz2005, Auch2006]. Eukaryotic genomes, however, were tested here for the first time, whereas earlier work with GBDP was restricted to prokaryotes or organelles of eukaryotes.

4.2 Benchmarks

The use of 15 eukaryotic (Basidiomycota) genomes was a special test case for the presented grid-based implementation of GBDP as these comparably big genomes had never been processed by that implementation before, thus we entered new territory when assessing the scalability of the algorithms and the hardware (e.g., memory usage). GBDP succeeded in processing this data and provided interesting insights into time and space complexity: in general, we can assume a linear relationship between search space size and the elapsed time for both types of requests. If the work load of a cluster is low, the cluster can process all requests in less than an hour, virtually providing for a linear speed-up.

However, some outliers were detected among the computed distance requests which cannot be explained by the mere size of the respective input data they used (matches). Here, other effects must have occurred that might be due to the partial incompleteness of some genomes. A closer look at the internal structure of these files revealed a high number of sequences (i.e., headers) – ranging from hundreds to thousands of sequences – and thus might have influenced the way GBDP is internally processing hits and transforming them to distance values. Moreover, the time measurement for each request ranges from the process' starting time till the time when the result file was completely written to the file system; measurement errors at any

of these stages might also be the case. Thus, a detailed benchmarking of all the steps performed in a single request could presumably provide more explanations to the aforementioned outliers. On the background of this data set, the taxonomic relatedness only had a minor effect on the computation time.

The benchmark results provide for an estimated a priori calculation of both the expected computation time as well as the order of magnitude of the resulting matches files' sizes. Hence, this information can be used to plan the requirements of future GBDP-based experiments.

4.3 Outlook

GBDP is using local-alignment tools such as BLAST+ for processing match requests. As the latter is providing multi-core support these could also be utilized for speeding-up the overall execution time. Even more speed-up would be achieved with highly-optimized GPU-BLAST software suites [Vouzis2011]. Even though the local alignment phase isn't a computational bottle-neck right now, optimized BLAST could be beneficial in the near future if more complete genomes are sequenced, especially from eukaryotes. In cooperation with Marek Dynowski (Rechenzentrum, University of Freiburg) and Kevin Körner (Zentrum für Datenverarbeitung, University of Tübingen) we are currently working on a GGDC portlet for the bwGRiD. The portlet is targeting the following aspects and features:

- The aforementioned grid-based GGDC variant should be provided to the scientific community as a high-throughput tool. By means of a web-based graphical user interface (similar to the web service already available on http://ggdc.gbdp.org/), users should be able to set-up and launch their HPC-based experiments.

- The conduction of large GBDP-based data sets/experiments in a relatively small amount of time should be possible, thus accessing new types of scientific questions that had previously only been possible in theory.

- Providing large amount of disk space even beyond common user quotas. Data management will be totally left to the portal, thus relieving the user from this tedious task. Data will be always available via a job monitoring and download portlet.
- Generation of match and distance requests: the portlet should translate the user-defined genome comparisons to the YAML request format as required by the GGDC.
- Recycling of results: in order to avoid the repetition of popular genome comparisons which have already been triggered by other users before, the results of match and distance requests could be stored in a central repository such as the bwGRiD storage located at the Karlsruher Institut für Technologie (KIT). However, a concept for this kind of central data management has to be developed beforehand and should preferably consider a broader spectrum of features. The latter would positively affect the development of new applications as these could recourse to existing infrastructure via an universal application programming interface (API).
- GGDC's job submission system will be adopted for grid computing using the GATLET library (http://gatlet.scc.kit.edu/). The presented implementation is currently devised for sending requests to a single cluster and is thus only using a specific part of the grid instead of dynamically sending jobs to those cluster(s) having the smallest work load at a particular time. This would provide for an additional speed-up and, in principle, even allow for the processing of even larger data sets. Notifications would be brought to the user in a similar manner as already implemented on http://ggdc.gbdp.org/.
- A permanent software infrastructure should be established that would directly benefit from future hardware extensions such as those provided by follow-up projects of the successful bwGRiD service.

The work presented here allows for a clearly optimistic view regarding GBDP's suitability for the processing of large (eukaryotic) genomes. Since

each genome comparison is independent of the others, these calculations perfectly fit to the distributed, node-based architecture of compute clusters. The highly parallel processing of genomic data sets thus leads to a dramatically reduced overall computation time, paving the way for experiments that have practically been impossible before.

Bibliography

[Altschul1990] S Altschul, W Gish, W Miller, E Myers, and D Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, 1990.

[Auch2006] A. F. Auch, S. Henz, B. Holland, and M. Göker. Genome BLAST distance phylogenies inferred from whole plastid and whole mitochondrion genome sequences. *BMC bioinformatics*, 7(1):350, 2006.

[Auch2009] A. F. Auch. *A phylogenetic potpourri – Computational methods for analysing genome-scale data*. PhD thesis, Universität Tübingen, Wilhelmstr. 32, 72074 Tübingen, 2009.

[Auch2010] A. F. Auch, M. von Jan, H.-P. Klenk, and M. Göker. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Standards in Genomic Sciences*, 2(1):117–134, 2010.

[Auch2010a] A. F. Auch, H.-P. Klenk, and M. Göker. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. *Standards in Genomic Sciences*, 2(1):142–148, 2010.

[Beutin2012] L. Beutin and A. Martin. Outbreak of Shiga Toxin-Producing Escherichia coli (STEC) o104:h4 Infection in Germany Causes a Paradigm Shift with Regard to Human Pathogenicity of STEC Strains. *Journal of Food Protection*, 75(2):408–418, 2012.

[Camacho2009] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421, 2009.

[Desper2002] R. Desper and O. Gascuel. Fast and Accurate Phylogeny Minimum-Evolution Principle. *Journal of Computational Biology*, 9(5):687–705, 2002.

[Farris1972] J. S. Farris. Estimating phylogenetic trees from distance matrices. *American Naturalist*, 106(951):645–667, 1972.

[Henz2005] S. R. Henz, D. H. Huson, A. F. Auch, K. Nieselt-Struwe, and S. C. Schuster. Whole-genome prokaryotic phylogeny. *Bioinformatics*, 21(10):2329–2335, 2005.

[Hibbett2007] D. S. Hibbett, M. Binder, J. F. Bischoff, M. Blackwell, P. F. Cannon, O. E. Eriksson, S. Huhndorf, T. James, P. M. Kirk, R. Lücking, H. T. Lumbsch, F. Lutzoni, P. B. Matheny, D. J. McLaughlin, M. J. Powell, S. Redhead, C. L. Schoch, J. W. Spatafora, J. A. Stalpers, R. Vilgalys, M. C. Aime, A. Aptroot, R. Bauer, D. Begerow, G. L. Benny, L. A. Castlebury, P. W. Crous, Y.-C. Dai, W. Gams, D. M. Geiser, G. W. Griffith, C. Gueidan, D. L. Hawksworth, G. Hestmark, K. Hosaka, R. A. Humber, K. D. Hyde, J. E. Ironside, U. Kõljalg, C. P. Kurtzman, K.-H. Larsson, R. Lichtwardt, J. Longcore, J. Miadlikowska, A. Miller, J.-M. Moncalvo, S. Mozley-Standridge, F. Oberwinkler, E. Parmasto, V. Reeb, J. D. Rogers, C. Roux, L. Ryvarden, J. Paulo Sampaio, A. Schüßler, J. Sugiyama, R. G. Thorn, L. Tibell, W. A. Untereiner, C. Walker, Z. Wang, A. Weir, M. Weiss, M. M. White, K. Winka, Y.-J. Yao, and N. Zhang. A higher-level phylogenetic classification of the Fungi. *Mycological Research*, 111(5):509–547, 2007.

[Kent2002] W. Kent. BLAT – the BLAST-like alignment tool. *Genome Res*, 12(4):656–664, 2002.

[Klenk2010] H.-P. Klenk and M. Göker. En route to a genome-based classification of Archaea and Bacteria? *Systematic and Applied Microbiology*, 33(4):175–182, 2010.

[Kurtz2004] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg. Versatile and open software for comparing large genomes. *Genome biology*, 5(2):R12, January 2004.

[Mardis2011] E. R. Mardis. A decade's perspective on DNA sequencing technology. *Nature*, 470(7333):198–203, 2011.

[Pagani2012] I. Pagani, K. Liolios, J. Jansson, I-M. Chen, T. Smirnova, B. Nosrat, V. M. Markowitz, and N. C. Kyrpides. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic acids research*, 40(Database issue):D571–D579, January 2012.

[Schwartz2003] S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller. Human - Mouse Alignments with BLASTZ. *Genome Research*, 13(1):103–107, 2003.

[Vouzis2011] P. D. Vouzis and N. V. Sahinidis. GPU-BLAST: using graphics processors to accelerate protein sequence alignment. *Bioinformatics*, 27(2):182–188, 2011.

[Wickham2009] H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[Woese1977] C. R. Woese and G. E. Fox. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977.

[bwgrid2012] BwGRiD. Member of the German D-Grid initiative, funded by the Ministry of Education and Research and the Ministry

for Science, Research and Arts Baden-Wuerttemberg (2007-2012). Technical report, Universities of Baden-Württemberg, 2012.

Acknowledgment

Many thanks are addressed to Marek Dynowski, Rechenzentrum, University of Freiburg, and Werner Dilling, Zentrum für Datenverarbeitung, University of Tübingen, for granting access and for their technical support related to the compute clusters of the `bwGRiD` [bwgrid2012].

Funding

This work was supported by the German Research Foundation (DFG) SFB/TRR 51, which is gratefully acknowledged. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.