The smallest cells pose the biggest problems: High performance computing and the analysis of metagenome sequence data.

Robert A Edwards

Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439

Department of Computer Science, San Diego State University, San Diego, CA 92182

E-mail: redwards@mcs.anl.gov

Abstract.

New high-throughput DNA sequencing technologies have revolutionized how scientists study the organisms around us. In particular microbiology, the study of the smallest, unseen organisms that pervade our lives has embraced these new techniques to characterize and analyze the cellular constituents, and use this information to develop novel tools, techniques, and therapeutics.

Increased technology such as pyrosequencing and other so-called next generation DNA sequencing platforms have resulted in huge increases in the amount of raw data that is generated. Argonne National Laboratory developed the premier platform for the analysis of this new data (mg-rast), and offers a freely available service that is used by microbiologists world-wide. This paper uses the accounting from the computational analysis of more than 10,000,000,000 bp of DNA sequence data and describes an analysis of the advanced computational requirements, the needs of this broad set of scientists, and suggests the level of analysis that will be essential as microbiologists move to understand how these tiny organisms affect our every day lives.

1. Introduction

1.1. DNA sequencing and metagenomics

All around us small organisms are influencing everything we do and how we do it. These organisms, called microbes because they are only visible under the microscope, are more abundant and much more diverse than the plants, trees, and other life we see all around us. Microbes affect us in both positive ways, providing cheese, bread, wine, and antibiotics for example; and they also affect us in negative ways, causing disease, decay, rotting timber, and so forth.

Biologists have struggled for many years to truly understand the role of microbes in different environments; what particular microbes are in an environment, and what are they doing there? With this knowledge in hand, they want to begin manipulating them. Like almost all known life, microbes use DNA (deoxyribonucleic acid) as their genetic material. This complex chemical encodes proteins that perform all cellular functions, from creating new cell walls, to propelling the cells towards food and away from noxious chemicals, to both making and using amino acids. A typical microbe is only about 10^{-6} meters long (1 micron), but has about 2,000,000 base pairs of DNA. Each base pair is comprised of one combination of the four chemical compounds adenine,

guanine, cytosine, and thymidine (abbreviated A, G, C, or T). The particular combination of letters encodes for the genes that encode the proteins. By sequencing the complete complement of DNA in a microbe (its *genome*), and by comparing that to sequences that have been experimentally validated, microbiologists are able to characterize all of the genes in a microbe, and all of the functions that microbe is performing [1]. Rather than studying a single microbe in isolation, more recently a new approach has been taken to understand the role of microbes in complex communities in the environment. This new technique is called *metagenomics* to imply that it is more than just a single genome that is being sequenced. The first of these studies were most widely performed in the oceans [2], and have subsequently been translated to most major microbial habitats, including the human body [3].

These transformative studies of microbes in their habitats have been assisted by new technologies that have revolutionized the speed and accuracy with which the DNA sequence can be determined. Until very recently most sequencing was performed using a technique first described in 1977 [4], and only minor modifications and improvements to the basic approach had been made. High throughput was largely achieved by robotization and miniaturization of the existing technology. A typical sequencing machine would generate the DNA sequence of 96 fragments simultaneously. The first of the so-called next-generation sequencing technologies to market, massively parallel pyrosequencing developed by 454 Life Sciences, Inc, (now part of Roche, Inc) enabled the generation of hundreds of thousands of DNA sequences at a time [5]. The increased volume of sequences came at a cost of decreased length of each single sequence. In the traditional approach a single DNA sequence could exceed 1,000 letters, but the new pyrosequencing approach could only generate about 100 letters from a single piece of DNA. Over time, however, the technology has matured and pyrosequencing currently generates 400 letters from a single DNA fragment, and about 500,000 fragments in a single reaction. All of this is accomplished in less than eight hours. In addition to generating more data, pyrosequencing is about ten-fold to one hundred-fold cheaper than Sanger sequencing (depending on economies of scale). This revolution in technology has enabled microbial researchers to take a sample from the environment, extract the DNA, and generate the sequence of a sample of that DNA in less than a week. The DNA sequence generated by this approach is the raw input for the advanced computing that determines the functions of those DNA fragments and is the focus of this paper.

2. An automated analysis pipeline for metagenomes

As researchers generated more and more sequence data, the need for an accurate, automated, high-throughput annotation and analysis platform became pressing. The SEED database was first developed to annotate complete microbial genomes, using a new approach that exploited the clustering in microbial genomes of genes that perform related functions. (Although it is unknown why this happens, there are several hypotheses but none that are conclusively supported by all of the data.) The SEED was used as the basis of the metagenomics annotation and analysis pipeline, called the *m*eta*g*enomics *r*apid *a*nnotation using *s*ubsystem *t*echnology (mg-rast). This pipeline is depicted in figure 1.

The pipeline consists of a series of discrete steps that are carried out in series on the whole dataset:

(i) Preprocessing
Incoming DNA sequence is checked to ensure that it is really DNA sequence (sometimes our users submit corrupted files, binary files from word-processing software, and so forth). There is also a known artifact of the high-throughput sequencing technology, that occasionally the exact same sequence is generated more than once. These sequences are removed, and each sequence is provided a unique identifier so that it can be tracked through subsequent steps.

(ii) Similarity computation

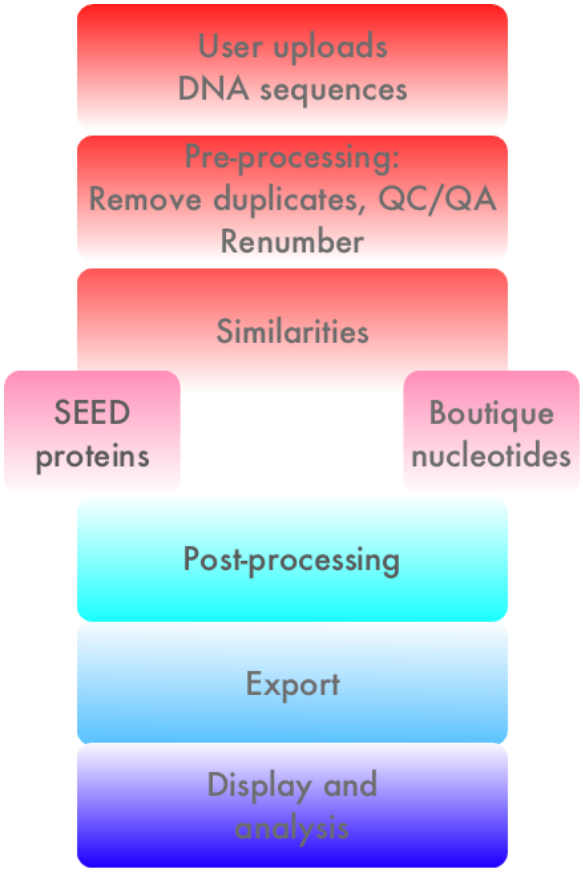

Figure 1. The metagenomics rapid annotation pipeline (mg-rast) consists of four steps, *preprocessing*, *similarity computation*, *postprocessing*, and *export*

The sequences are compared to a number of databases of previously characterized and well understood sequences. These include a database that has all of the sequences from complete genomes from all three domains of life, and four different databases that contain sequences of certain genes of specific interest to microbiologists. (They are typically interested in these genes because they cannot change very quickly, and thus can be used to describe the time-series relationships between organisms.)

(iii) Postprocessing
After the DNA sequences have been compared, the results are parsed to generate a description of each DNA sequence in the sample, and the entire sample itself.

(iv) Export
In the final step, reports are generated for the end-user to either view via an online interface, or to download and peruse at will.

The mg-rast service was released in beta-version in April, 2007; and opened for unfettered public access in July of the same year. The service is freely available to all researchers, although registration with a valid email address is required so that we may contact the users in the case of trouble in the system. Since the initial release, over 800 sequence analysis jobs have been submitted by 99 different users; comprising in excess of 10,000,000,000 base pairs (10 Gbp) of DNA sequence, and the mean sample submitted by our users is currently 15,000,000 base pairs (15 Mbp).

Table 1. Computational resources used in the mg-rast pipeline.

Name In Figures	Processor	Memory	Number of units
bio-big	16 Intel Xeon CPU X7350 @ 2.93GHz	123822988 kB	1
CGAT	8 Intel Xeon CPU X5365 @ 3.00GHz	16436064 kB	1
DGRI	8 Intel Xeon CPU E5335 @ 2.00GHz	16440024 kB	1
Intel	1 Intel Pentium 4 CPU 3.00GHz	3636148 kB	2
MacPro	8 Intel Xeon CPU X5365 @ 3.00GHz	16387776 kB	2
mg-rast	8 Intel Xeon CPU E5450 @ 3.00GHz	16436064 kB	1
RAST	8 Intel Xeon CPU X5365 @ 3.00GHz	16436060 kB	1
PPC	2 PPC970FX, altivec supported	4042892 kB	45

3. Computation

The entire mg-rast computation is currently being performed on a hybrid cluster. Initially developed as an Apple PowerPC cluster running Debian Sarge, over time these machines have been supplemented with Intel Xeon processors also running linux (table 1). The cluster is a shared resource, and in addition to the mg-rast is also used to annotate complete microbial genomes through a related service, the RAST (*r*apid *a*nnotation using *s*ubsystems *t*echnology) [6]. Jobs on the cluster are scheduled using the Sun Grid Engine®, and the accounting system was used as the basis of all timing studies presented in this work. All accounting times were extracted using the PERL Schedule::SGE module written by the author and available from the Comprehensive Perl Archive Network (CPAN) or by contacting the author.

Throughout this paper the input data is measured in terms of base-pairs; individual letters of DNA uploaded by the end users. Of course, this equates to bytes of data with standard unicode encoding. Wall time is generally measured in seconds throughout the analysis.

4. The four stages of the computation

4.1. Preprocessing

The four separate stages of the computation occur in series: each needs to be completed before the next is started. The only exception is the multiple different computations calculated during the similarity processing stage. Each different database can be compared in parallel, since each is an independent entity. At this stage, the input data is also routinely fragmented into smaller units to be processed. The goal is to achieve a single unit of computation that takes between one and two hours to complete, an empirically determined balance between fairness with other users of the clusters; limitations of IO capability and scaling; and sufficient throughput to complete the analysis in a timely fashion.

As shown in figure 2, the preprocessing step is defined by two distinct populations. The first population displays linear complexity with input size, and represents those tasks that complete without problem. The preprocessing step is not memory limited, does not require complex calculations, and there does not appear to be a significant difference between the Intel or PPC architectures used in this approach. The total time taken to process the data is approximately 2 seconds per million base pairs of input sequence, or less than a minute for the average metagenome. The second population consists of those problematic jobs alluded to earlier. It is at this stage of the computation that user-derived errors are usually identified. Malformed files, bad sequence data, and so on result in larger processing times for smaller jobs.

Those jobs likely represent the situations where endless cycles are performed before a submission is terminated.

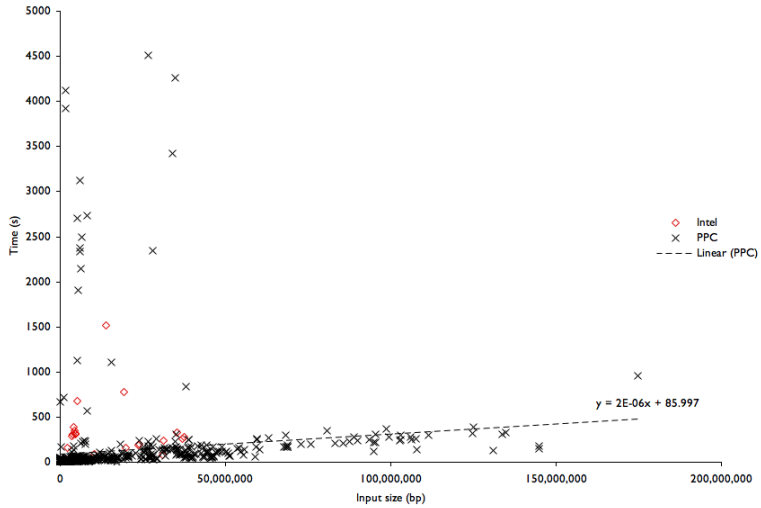

Figure 2. Preprocessing is characterized by two distinct populations. The first computes with linear complexity compared to input size, and represents those jobs that complete without problem. The second population is more scattered, takes longer to compute, and represents those tasks that are problematic.

4.2. Similarity

The similarity computation, as noted above, really consists of several different computations run in parallel. The largest, and most involved comparison is between the DNA sequence submitted by the user and the known sequences from all of the genomes that have ever been sequenced. (Note that for efficiency reasons, we don't currently automatically compute a comparison between all the metagenomes that have been submitted.) In this case, additional processing is added to the computation because the DNA sequence from the user is compared to the known sequences in protein space. Recall that DNA encodes protein (albeit indirectly through an intermediate molecule called mRNA). Therefore, given a particular DNA sequence, the protein sequence can be generated computationally. However, because there are only four DNA letters (the base pairs G, A, T, and C) but twenty amino acids, nature has evolved to use a 3 letter register for the translation of DNA sequence to protein sequence. Therefore, for any given DNA sequence there are *six* possible protein sequences (three each in both the forward and reverse directions). Although this doesn't increase the complexity of the calculation, it does significantly increase the computation time required to complete the analyses.

As shown in figure 3, the similarity against the SEED protein database is significantly demanding, taking almost seven seconds per hundred base pairs of sequence on the PPC architecture and seven seconds per thousand base pairs of sequence on the 16-core Intel Xeon architecture.

In contrast to the large amount of data in the computation of the similarity to all known sequences, the comparison to more boutique databases is decidedly more rapid (figure 4), with between 3,000 and 12,500 sequences being processed per second depending on architecture.

4.3. Postprocessing and Export

The final two steps of the procedure are to parse the results from the similarity computations and assert functions based on those results, and then to export the data into standard formats commonly used by members of the bioinformatics community. These two processes, as shown in figure 5, have linear complexity and are completed rapidly for small sequence datasets, and in less than two hours for the largest datasets.

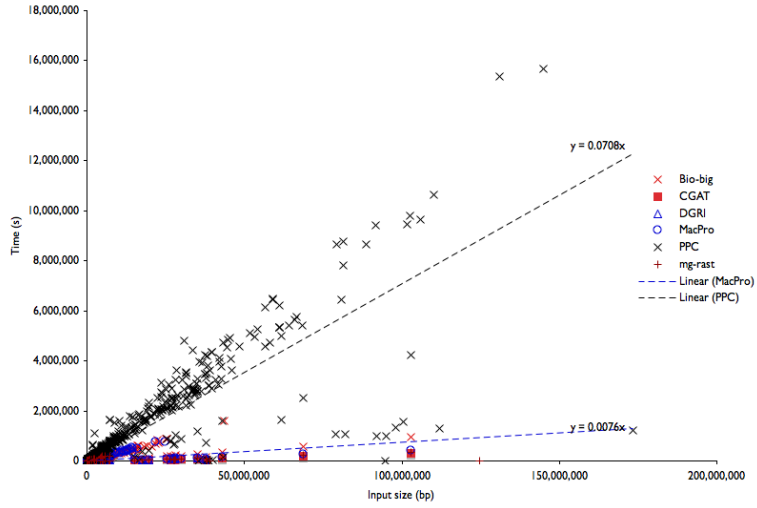

Figure 3. Calculating the similarity between the user-supplied sequences and the known sequences in the publicly available databases is the most time-consuming tasking. It is therefore the task that is most amenable to improvement in the whole pipeline. Using sixteen-core architecture it takes approximately seven seconds to process 1,000 base pairs of DNA sequence.

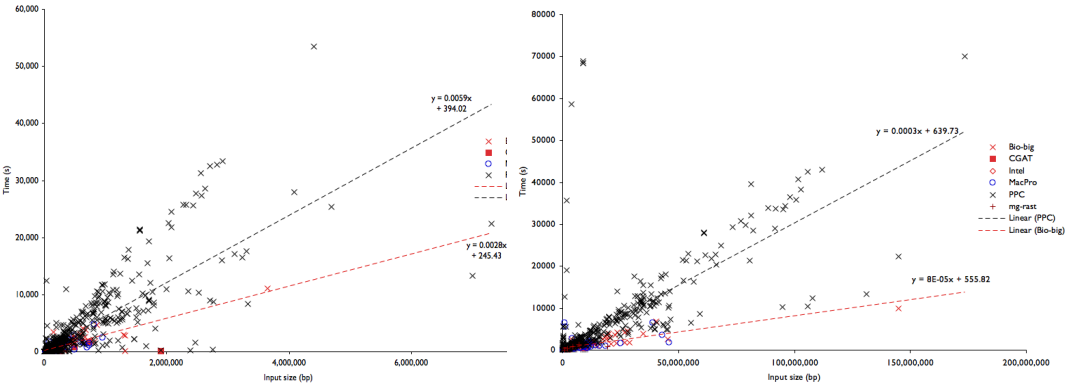

Figure 4. Comparison of all the sequences against the RDP [7] (a; left) and the GreenGenes database [8] (b; right) is much quicker than comparison to all known proteins.

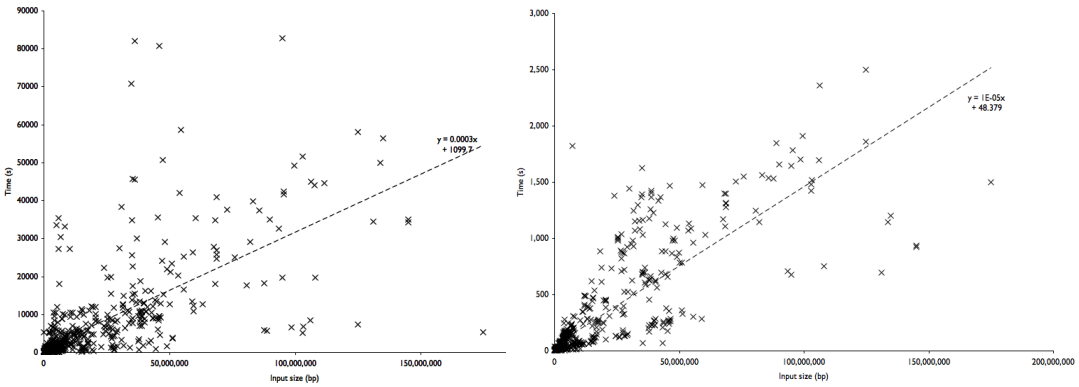

Figure 5. Postprocessing and data export are not computationally intensive steps, taking less than two hours for most large datasets.

5. Total Elapsed Time

Given the very linear nature of the processes involved – all calculations have a computational time complexity of $O(n)$ – it was very surprising to see that the total elapsed time for the

calculation of all similarities between sequences was not linear (figure 6). As shown in figure 6 (b), the amount of data being processed by the metagenomics server is rapidly increasing, and about 1 Gbp of sequence data (1 Gbyte) is being processed per month. As noted above, this queue is not solely used for the metagenomics analyses, and other data being processed on this platform shows a similar trend (data not shown). The limiting factor, therefore, in processing the data in the metagenomics platform is currently access to appropriate high performance computing, and specifically the amount of time a job spends in the queue.

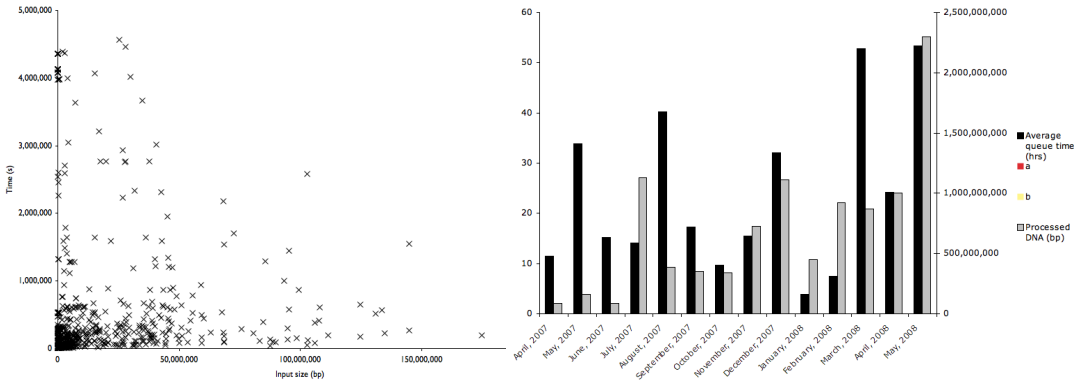

Figure 6. Total elapsed time is not linear overall (a). There is a correlation between the amount of data processed per month, and the amount of time spent in the queue. (Note data for June, 2008 is incomplete and not shown.)

6. Conclusions

This analysis has shown that it is possible to compute the complete comparisons of metagenomes against pertinent databases in linear time, and the computation should be able to keep pace with the generation of new sequences. However, we are currently in a period of rapid deployment of sequencing machines around the country. It is rumored that the number of FLX units is fast approaching 300 in North America, with more coming online every day. These machines could generate an estimated 0.1 Tbp (100,000,000,000 bp) of sequence per day. We need to ensure that the increases in computational efficiency and computer speed maintain our ability to process data for the end users. Second, it is clear that the bulk of the processing happens very quickly. Although these computations are $O(n)$ they are currently processed in with little delay. Of course, as n increases these computations will be challenging, but as long as increased processes speed tracks increased DNA sequencing capacity these computations can be completed in a timely manner. However, the single largest computational step, the comparison of the DNA sequences to the known protein sequences is not only increasing with $O(n)$ complexity, it is taking a significantly longer time to compute than any other step in the process. Better algorithms are required for comparing the sequences; algorithms to triage sequences that will not generate significant similarity may reduce overall computation; and more computational horse-power is needed to complete the analyses in an appropriate timeframe. Currently a single run of a pyrosequencing machine produces about 100 Mbp of data; the time taken for the individual steps in the pipeline are shown in table 2.

Acknowledgments

I thank Robert Olson for assistance with the SGE analysis and insightful discussions; all of the developers of the metagenomics RAST server, RAST server, the SEED, and the National Microbial Pathogen Resource (NMPDR) for assistance, dedication, and tutelage. Special thanks

Table 2. Time taken to compute each of the steps in the pipeline for a typical 100 Mbp DNA sample with no queuing.

Preprocessing	Similarity (SEED)	Postprocessing	Export
200 seconds	700,000 seconds (194 hours)	300 seconds	10 seconds

to Rick Stevens and Ross Overbeek for support and guidance. Part of this work was funded by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract HHSN266200400042C. This work was also funded by the U.S. Dept. of Energy under Contract DE-AC02-06CH11357.

References

[1] Overbeek R, Begley T, Butler R M, Choudhuri J V, Chuang H Y, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank E D, Gerdes S, Glass E M, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy A C, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch G D, Rodionov D A, Ruckert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O and Vonstein V 2005 *Nucleic Acids Res* **33** 5691–5702 ISSN 1362-4962 (Electronic)

[2] Dinsdale E A and Edwards R A 2007 *Oceanography* **20** 56

[3] Turnbaugh P J, Ley R E, Hamady M, Fraser-Liggett C M, Knight R and Gordon J I 2007 *Nature* **449** 804–810 ISSN 1476-4687 (Electronic)

[4] Sanger F, Nicklen S and Coulson A R 1977 *Proc Natl Acad Sci U S A* **74** 5463–5467 ISSN 0027-8424 (Print)

[5] Margulies M, Egholm M, Altman W E, Attiya S, Bader J S, Bemben L A, Berka J, Braverman M S, Chen Y J, Chen Z, Dewell S B, Du L, Fierro J M, Gomes X V, Godwin B C, He W, Helgesen S, Ho C H, Irzyk G P, Jando S C, Alenquer M L I, Jarvie T P, Jirage K B, Kim J B, Knight J R, Lanza J R, Leamon J H, Lefkowitz S M, Lei M, Li J, Lohman K L, Lu H, Makhijani V B, McDade K E, McKenna M P, Myers E W, Nickerson E, Nobile J R, Plant R, Puc B P, Ronan M T, Roth G T, Sarkis G J, Simons J F, Simpson J W, Srinivasan M, Tartaro K R, Tomasz A, Vogt K A, Volkmer G A, Wang S H, Wang Y, Weiner M P, Yu P, Begley R F and Rothberg J M 2005 *Nature* **437** 376–380 ISSN 1476-4687 (Electronic)

[6] Aziz R K, Bartels D, Best A A, DeJongh M, Disz T, Edwards R A, Formsma K, Gerdes S, Glass E M, Kubal M, Meyer F, Olsen G J, Olson R, Osterman A L, Overbeek R A, McNeil L K, Paarmann D, Paczian T, Parrello B, Pusch G D, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A and Zagnitko O 2008 *BMC Genomics* **9** 75 ISSN 1471-2164 (Electronic)

[7] Cole J R, Chai B, Farris R J, Wang Q, Kulam S A, McGarrell D M, Garrity G M and Tiedje J M 2005 *Nucleic Acids Res* **33** D294–6 ISSN 1362-4962 (Electronic)

[8] DeSantis T Z, Hugenholtz P, Larsen N, Rojas M, Brodie E L, Keller K, Huber T, Dalevi D, Hu P and Andersen G L 2006 *Appl Environ Microbiol* **72** 5069–5072 ISSN 0099-2240 (Print)