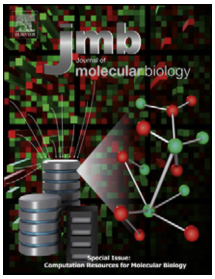
Databases/
Web Servers



BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences

Minoru Kanehisa [1], Yoko Sato [2] and Kanae Morishima [1]

*1 - **Institute for Chemical Research,** Kyoto University, Uji, Kyoto 611-0011, Japan*
*2 - **Healthcare Solutions Department,** Fujitsu Kyushu Systems Ltd., Hakata-ku, Fukuoka 812-0007, Japan*

***Correspondence to Minoru Kanehisa:** kanehisa@kuicr.kyoto-u.ac.jp*
http://dx.doi.org/10.1016/j.jmb.2015.11.006
Edited by M. Sternberg

Abstract

BlastKOALA and GhostKOALA are automatic annotation servers for genome and metagenome sequences, which perform KO (KEGG Orthology) assignments to characterize individual gene functions and reconstruct KEGG pathways, BRITE hierarchies and KEGG modules to infer high-level functions of the organism or the ecosystem. Both servers are made freely available at the KEGG Web site (http://www.kegg.jp/blastkoala/). In BlastKOALA, the KO assignment is performed by a modified version of the internally used KOALA algorithm after the BLAST search against a non-redundant dataset of pangenome sequences at the species, genus or family level, which is generated from the KEGG GENES database by retaining the KO content of each taxonomic category. In GhostKOALA, which utilizes more rapid GHOSTX for database search and is suitable for metagenome annotation, the pangenome dataset is supplemented with Cd-hit clusters including those for viral genes. The result files may be downloaded and manipulated for further KEGG Mapper analysis, such as comparative pathway analysis using multiple BlastKOALA results.

Introduction

Genome sequencing has become a routine task in many areas of biological sciences. It also has the potential to revolutionize healthcare, such as in personalized medicine using personal genome information and in combating infectious diseases and drug resistance using pathogen genome information. There is a growing need for better bioinformatics methods to fully make use of such genome sequencing data. We have been developing the KEGG (Kyoto Encyclopedia of Genes and Genomes) bioinformatics resource as a reference knowledge base for biological interpretation of genome sequences and other high-throughput data, especially for uncovering high-level functions and utilities of the cell, the organism and the ecosystem [1].

One unique aspect of the KEGG resource is that functional information is associated with ortholog groups and stored in the KO (KEGG Orthology) database, separately from the KEGG GENES database that accumulates all fully sequenced genomes. Genome annotation in KEGG is simply to assign KO identifiers (K numbers) rather than to rewrite text definitions of genes in the GENES database. In general, KO grouping of functional orthologs is defined in the context of KEGG pathways and other molecular networks, which are in fact represented as networks of nodes identified by K numbers. Therefore, once genes in the genome are given K numbers, KEGG pathways and other molecular networks can be automatically reconstructed.

Internally, the GENES database is accompanied by the SSDB database, which contains amino acid sequence similarity scores and best-hit relations for all gene pairs in pairwise genome comparisons computed by the SSEARCH program [2]. SSDB is a huge weighted, directed graph and has been used to computationally define Ortholog Clusters as quasi-cliques of bidirectional best hits with significant similarity scores [3]. Genome annotation in KEGG involves both manual and computational procedures.

Basically, members of a KO group are first manually selected defining a core subgraph in the SSDB graph, which is then computationally extended by the KOALA (KEGG Orthology and Links Annotation) program [1]. Here we report automatic annotation servers for outside users, BlastKOALA and GhostKOALA, which are based on the methods similar to those used in internal annotations and the newly developed non-redundant GENES database. Brief comparisons are made to other servers including KAAS [4], RAST [5] and MG-RAST [6].

Results and Discussion

Improved KO and GENES databases

The KO database is a collection of orthologs, most of which are members of the KEGG molecular networks, including KEGG pathways, BRITE hierarchies and KEGG modules. There are also unclassified KO entries whose relationships to molecular networks are unknown. In order to facilitate sequence information-based assignment, each KO entry is defined as a sequence similarity group including some complicated cases. For example, one KO entry may consist of multiple sequence similarity groups, or one KO entry may be the remaining portion of a large sequence similarity group after extracting a closely related portion defined as another KO entry. Efforts are being made to associate each KO entry with experimental evidence of functionally characterized sequence data. As of June 2015, about 19,000 KO entries are made available, and PubMed links and sequence links are included in 75% and 42%, respectively.

By the genome annotation procedure in KEGG, about 46% of 17 million genes are annotated with K numbers. The GENES database can be seen as structured by the assigned K number groups, where each group corresponds to a sequence information-based extension of experimental evidence and functional information. Thus, the sequence similarity search against the structured GENES database is a search for most appropriate K numbers, which can easily be computerized as implemented in the KOALA and KAAS programs. This is a clear advantage over most other sequence databases, where functional information is associated with individual proteins or genes and the sequence similarity search will require processing of search results, which may contain a large amount of data.

The GENES database is a collection of complete genomes taken from RefSeq and GenBank databases, each identified by the three- or four-letter organism code, such as “hsa” for *Homo sapiens*. We have recently added three new categories with two-letter codes: “pg” for plasmid genes taken from the RefSeq plasmid section excluding those already represented in the KEGG complete genomes, “vg” for viral genes taken from the RefSeq viral section and “ag” for the internally developed and publication-based collection of KEGG addendum genes. The “ag” category is a gene-based, not genome-based, collection and now fills the gap of missing sequence data in some pathway maps, where KO entries do not correspond to any GENES entries and simply linked to UniProt database because complete genome sequences have not been determined. With these new categories, the repertoire of KO entries has been significantly expanded toward capturing all knowledge on gene/protein functions.

Non-redundant dataset of pangenome sequences

As many closely related bacterial strains are sequenced, it has become customary to define selected sets of genomes for annotation and other purposes, such as reference and representative genomes in RefSeq [7] and reference proteomes in UniProt [8]. As of June 12, 2015, there are 3904 KEGG organisms (complete genomes): 304 eukaryotes in the ranks of 299 species, 227 genera and 172 families, as well as 3600 prokaryotes in the ranks of 1858 species and 809 genera. See the following Web pages for the current version: KEGG organisms[†], KEGG species list[‡] and KEGG genus list[§]. Each species or genus name in these lists is linked to another Web page for analyzing an organism group, such as how pathway maps and BRITE hierarchies can be reconstructed from the combined set of assigned K numbers in the group.

When a new genome is added in KEGG, the SSEARCH comparison is performed against all genomes in the GENES database for subsequent K number assignment by KOALA. Because of the size of the GENES database, this procedure is too time consuming to be implemented as a Web service. Thus, for use in the BlastKOALA and GhostKOALA servers, we have devised a way to reduce the size of the GENES database while retaining the functional content of the gene set in each taxonomic rank of species, genus or family. When multiple members are present in each species/genus/family group, the first genome in the KEGG organisms list is taken as a representative genome. When the other members in the group contain different K numbers that are not present in the representative genome, those genes are added as if they are present in additional chromosomes or plasmids. Therefore, the resulting dataset is not simply a collection of representative genomes. It is rather a non-redundant dataset of pangenome sequences in terms of the functional content of genes designated by assigned K numbers. The degree of reduction was as follows: 4,660,929 sequences in 304 eukaryotes were reduced to 4,598,413 (98.7%), 3,719,997 (79.8%) and 2,803,038 (60.1%) sequences at the species, genus and family levels, respectively, while 11,315,630 sequences in 3600 prokaryotes were reduced to 6,148,174 (54.3%) and 2,698,820 (23.9%) sequences at the species and genus levels, respectively.

The annotation of metagenome sequences usually involves assignment of both functional and taxonomic information. The non-redundant dataset of pangenome sequences is supplemented by sequence similarity groups of genes without assigned K numbers to better characterize taxonomic compositions of metagenomes. This is accomplished by creating Cd-hit clusters [9] with 50% identity cutoff for each species/genus/family group, and sequences are selected from each cluster when no members have assigned K numbers and no members from the representative genome are included. As a result, the number of family-level eukaryotic sequences increased from 2,803,038 to 3,423,406 (122%) and the number of genus-level prokaryotic sequences increased from 2,698,820 to 3,913,518 (145%).

The new addendum (ag) and plasmid (pg) categories of the GENES database have also been included in the non-redundant dataset of pangenome sequences. The ag and pg sequences are grouped into species with the NCBI taxonomy identifier, such as 562 for *Escherichia coli*, and each species/genus/family group is processed in the same way as the KEGG organisms. For the viral (vg) category, since the K number assignment rate is very low, currently 7% for viruses *versus* 46% for KEGG organisms, a separate dataset is created simply from the Cd-hit clusters with 90% identity cutoff.

K number assignment

In the internal annotation of the GENES database, the GFIT (Gene Function Identification Tool) table is generated from SSDB computation results, where a list of top-scoring genes (target genes) in all other genomes is shown for each query gene in each genome. The assignment of a K number to the query gene is performed by the KOALA algorithm, which computes the weighted sum of SW (Smith–Waterman) scores for each K number group, as well as the no K number group, of target genes. The weighting factors include the bidirectional best-hit information, the overlap length of the alignment, the ratio of the query and target sequence lengths, the degree of matches of taxonomic categories if known and the degree of matches of Pfam [10] domains.

There are also some additional rules for how many target genes are to be examined before deciding a K number to be given or not to be given. A variant of this is a recently introduced rule for identical and almost identical sequences. For a predefined set of K numbers, which are called tight KOs, a single hit of an identical sequence or a few hits of almost identical sequences are considered sufficient to assign the K number of a tight KO. Tight KOs have been used, for example, to characterize antimicrobial resistance and distinguish substrate specificity in biodegradation pathways.

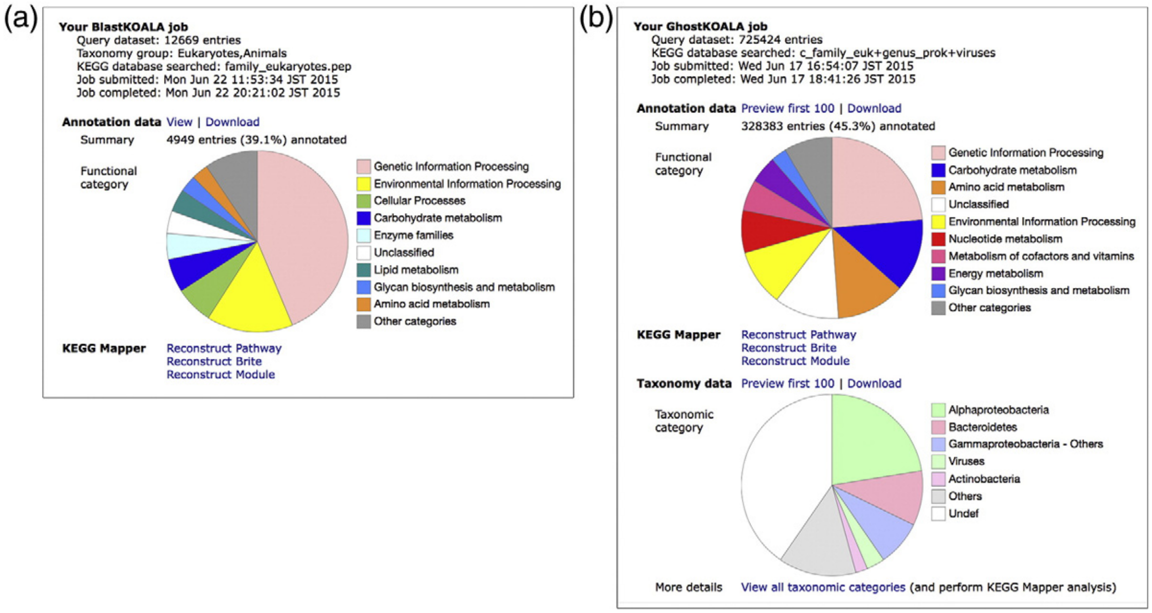
The BlastKOALA and GhostKOALA servers utilize the BLASTP [11] and GHOSTX [12] programs, respectively, for searching the non-redundant GENES database. The search result by BLASTP or GHOSTX is converted into a GFIT-like table where multiple hits (paralogs) in each target genome are combined and only the top-scoring hit is shown. In BlastKOALA, the K number assignment is performed using the weighted sum of BLAST bit scores, where the weighting scheme is the same as the internally used KOALA algorithm excluding the bidirectional best-hit information. In GhostKOALA, the K number assignment is simply based on the sum of GHOSTX normalized scores without considering any weighting factors. The new rule to assign tight KOs is included only in BlastKOALA.

Implementation and comparison

The BlastKOALA and GhostKOALA servers are made freely available at the kegg.jp Web site (see Table 1 for URLs). Both servers are e-mail based. A job request from the Web interface can either be confirmed or be canceled by clicking on the link in the automatically sent e-mail, and the annotation result such as shown in Fig. 1 can be viewed from

Table 1. KEGG automatic annotation servers and associated KEGG Mapper tools

Program	URL	Content
BlastKOALA	www.kegg.jp/blastkoala/	Automatic annotation and KEGG mapping server suitable for functional characterization of complete genomes
GhostKOALA	www.kegg.jp/ghostkoala/	Automatic annotation and KEGG mapping server suitable for functional characterization of metagenomes
Annotate Sequence	www.kegg.jp/kegg/tool/annotate_sequence.html	An interactive version of BlastKOALA suitable for use when related genomes are present in KEGG
Reconstruct Pathway	www.kegg.jp/kegg/tool/map_pathway.html	KEGG Mapper tool for reconstructing KEGG pathway maps
Reconstruct Brite	www.kegg.jp/kegg/tool/map_brite.html	KEGG Mapper tool for reconstructing BRITE functional hierarchies
Reconstruct Module	www.kegg.jp/kegg/tool/map_module.html	KEGG Mapper tool for reconstructing KEGG modules
Map Taxonomy	www.kegg.jp/kegg/tool/map_taxonomy.html	KEGG Mapper tool for reconstructing KEGG taxonomy or NCBI taxonomy



Your BlastKOALA job
Query dataset: 12669 entries
Taxonomy group: Eukaryotes,Animals
KEGG database searched: family_eukaryotes.pep
Job submitted: Mon Jun 22 11:53:34 JST 2015
Job completed: Mon Jun 22 20:21:02 JST 2015

Annotation data View | Download
Summary 4949 entries (39.1%) annotated
Functional category

KEGG Mapper Reconstruct Pathway
Reconstruct Brite
Reconstruct Module

Your GhostKOALA job
Query dataset: 725424 entries
KEGG database searched: c_family_euk+genus_prok+viruses
Job submitted: Wed Jun 17 16:54:07 JST 2015
Job completed: Wed Jun 17 18:41:26 JST 2015

Annotation data Preview first 100 | Download
Summary 328383 entries (45.3%) annotated
Functional category

KEGG Mapper Reconstruct Pathway
Reconstruct Brite
Reconstruct Module

Taxonomy data Preview first 100 | Download
Taxonomic category

More details View all taxonomic categories (and perform KEGG Mapper analysis)

Fig. 1. Examples of (a) the BlastKOALA result page and (b) the GhostKOALA result page.

the link in the job completion notification e-mail. There is an option to run BlastKOALA interactively without giving an e-mail address. It is implemented as the Annotate Sequence tool in KEGG Mapper (Table 1), but this option is limited to searching a single pangenome at the genus or family level. Generally speaking, BlastKOALA is suitable for annotating fully sequenced genomes, while GhostKOALA is recommended for annotating a large amount of metagenome sequences. This is because GHOSTX uses suffix arrays for finding matching sequences and runs 100 times faster than BLAST, although the sensitivity is not as good as BLAST.

Table 2 shows a summary of sample runs with different search programs and different datasets to be searched. The query genome was *Kangiella geojedonensis* [13], before it was included in KEGG with the organism code of kge (T03889). The first row of Table 2, SSDB/KOALA, is the subsequent annotation of this genome using the standard internal procedure of SSDB computation and KOALA assignment. Out of 2208 genes in the genome, 1368 genes were automatically annotated. By taking this set as a reference, we compared the performance of BlastKOALA with two different datasets, GhostKOALA, KAAS and the Annotate Sequence tool. As expected, the best accuracy was achieved by BlastKOALA with the species-level dataset, followed by BlastKOALA with the genus-level dataset and GhostKOALA, but the performance must be evaluated together with the execution time. The Annotate Sequence tool very much depends on the available dataset and may not have been very effective in this case because there was only a single species of *Kangiella* in KEGG. The KAAS server reported a number of false positives, presumably

Table 2. A comparison of K number assignments for the genome of *K. geojedonensis*.

Program	Dataset	Size	Elapsed time[a]	Assigned[b]	Match	Mismatch	False positive	False negative
SSDB/KOALA				1368	1368			
BlastKOALA	species_prokaryotes	6,098,670	1:20:54	1359	1350		9	18
BlastKOALA	genus_prokaryotes	2,687,538	41:11	1342	1331	1	10	36
GhostKOALA	c_genus_prokaryotes	3,889,798	6:57	1289	1278	2	9	88
KAAS	representative set	134,684	25:29	1393	1239	50	104	79
Annotate Sequence	*Kangiella*	2,632	1:55	1309	1280	12	17	76

[a] Elapsed time was measured on Sun Fire X4600 for KAAS and on SGI Altix UV1000 for all other programs. It is dependent on how many other jobs are running at the same time.
[b] The number of genes in the genome was 2208.

because the search is limited to a selected set of genomes and the algorithm to assign K numbers [4] is different from KOALA.

We also made a comparison with the RAST server [5] by submitting the same genome as a query. The server assigned 1530 FIGfams (protein families), which can be compared to 1359 KOs assigned by BlastKOALA. We compared EC numbers given to FIGfams and KOs. The number of genes with full EC numbers was 581 in RAST and 643 in BlastKOALA and 460 were perfect matches. There is a conceptual difference in the design and implementation of the servers. RAST and its metagenome version, MG-RAST, are an environment for the entire processing and subsequent analysis of genome and metagenome sequences, while BlastKOALA and GhostKOALA are meant to be used as a modular unit that may be integrated in the user's environment. BlastKOALA and GhostKOALA can also handle eukaryotic genomes and metagenomes containing eukaryotes, which is a clear advantage.

KEGG Mapper analysis

Figure 1a shows an example of the BlastKOALA result page, where the user can view or download the annotation data (the list of query genes and assigned K numbers) and perform KEGG Mapper analysis. The pie chart shows the functional category of annotated genes according to the KO system[‖]. In BlastKOALA, the BLAST search result in the form of a GFIT table can be examined for each query gene from the “View” page of annotation data. In GhostKOALA, the result page (Fig. 1b) contains an additional pie chart for the taxonomic composition of the query data together with a link for more detailed analysis, which is especially useful for metagenome data. Each query gene is assigned a taxonomic category according to the best-hit gene in the Cd-hit cluster supplemented version of the non-redundant pangenome dataset.

The KEGG Mapper links generate lists of reconstructed KEGG pathways, BRITE hierarchies and KEGG modules according to the assigned K numbers, which will allow interpretation of high-level functions encoded in the genome or the metagenome. The same analysis can be performed by downloading the annotation data and using the Reconstruct Pathway, Reconstruct Brite and Reconstruct Module tools in the KEGG Mapper Web site (Table 1). Since the BlastKOALA/GhostKOALA result page is deleted after 1 week of job completion, it is recommended to locally store the annotation result (list of K number assignments). If the query genome size exceeds the limit of the server, it would be necessary to split the query data, download the result files, concatenate them and perform KEGG Mapper analysis. Furthermore, multiple results from different query genomes may be combined and, for example, displayed with different coloring on KEGG pathway maps by the Reconstruct Pathway tool. In GhostKOALA, the taxonomy data may also be downloaded and used to analyze, for example, metabolic capacities of selected taxonomic groups by using KEGG Mapper tools.

Other types of analyses and usage

In addition to the predefined set of analysis tools in KEGG Mapper, the downloaded data may be used for other types of analyses. One is the chemical reaction network analysis of small molecules, which can be performed by converting K numbers into R numbers of the KEGG REACTION database. Alternatively, K numbers may be converted into EC numbers in the KEGG ENZYME database, which contains most up-to-date enzyme information taken from the Enzyme Nomenclature database [14]. It is recommended, however, to use the K number-to-R number links because one EC number may represent multiple reactions and multiple enzyme sequence families. Another is the phylogenetic analysis. Each KO entry is associated with the taxonomic distribution of assigned genes. This may be used to identify phylogenetically interesting genes, such as eukaryote-like genes in an archaeal genome. The EC numbers and the NCBI taxonomy IDs may also be useful for integration of KEGG annotation results with other data and analysis tools. Furthermore, KEGG contains disease and drug information linked to disease genes and drug targets in human, together with special pathway maps with the five-letter code of “hsadd”. Therefore, if genes in the query genome can be linked to human orthologs through the assigned KOs, the analysis of diseases and drugs may be performed.

The BlastKOALA and GhostKOALA servers reported in this paper have been used internally for automatic annotation of draft genomes in the KEGG DGENES database and metagenomes in the KEGG MGENES database, respectively. In fact, Fig. 1 was the results of annotating a butterfly genome of *Acropora digitifera* (T10031) [15] in DGENES and a sample of Tara Oceans data (T30798) [16] in MGENES. The KEGG Mapper data linked from the result pages of Fig. 1 are equivalent to those made available under the categories of Pathway map, Brite hierarchy and Module in the T10031 and T30798 genome pages at the KEGG Web site.

Acknowledgements

We thank Dr. Hiroyuki Ogata for helpful comments on metagenome analysis. The KEGG project is partially supported by the National Bioscience Database Center of the Japan Science and Technology Agency. The computational resource for the BlastKOALA and

GhostKOALA servers is provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

Received 26 June 2015;
Received in revised form 23 September 2015;
Accepted 5 November 2015
Available online 14 November 2015

Keywords:
genome annotation;
metagenome analysis;
taxonomic composition;
KEGG Orthology;
KEGG pathway mapping

† http://www.kegg.jp/kegg/catalog/org_list.html.

‡ http://www.kegg.jp/kegg/catalog/org_list4.html.

§ http://www.kegg.jp/kegg/catalog/org_list5.html.

|| http://www.kegg.jp/brite/ko00001.

References

[1] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, Data, information, knowledge and principle: Back to metabolism in KEGG, Nucleic Acids Res. 42 (2014) D199–D205.

[2] W.R. Pearson, Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms, Genomics 11 (1991) 635–650.

[3] A. Nakaya, T. Katayama, M. Itoh, K. Hiranuka, S. Kawashima, Y. Moriya, S. Okuda, M. Tanaka, T. Tokimatsu, Y. Yamanishi, A.C. Yoshizawa, M. Kanehisa, S. Goto, KEGG OC: A large-scale automatic construction of taxonomy-based ortholog clusters, Nucleic Acids Res. 41 (2013) D353–D357.

[4] Y. Moriya, M. Itoh, S. Okuda, A. Yoshizawa, M. Kanehisa, KAAS: An automatic genome annotation and pathway reconstruction server, Nucleic Acids Res. 35 (2007) W182–W185.

[5] R.K. Aziz, D. Bartels, A.A. Best, M. DeJongh, T. Disz, R.A. Edwards, K. Formsma, S. Gerdes, E.M. Glass, M. Kubal, F. Meyer, G.J. Olsen, R. Olson, A.L. Osterman, R.A. Overbeek, L.K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G.D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, O. Zagnitko, The RAST Server: Rapid annotations using subsystems technology, BMC Genomics 9 (2008) 75.

[6] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E.M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, R.A. Edwards, The metagenomics RAST server—A public resource for the automatic phylogenetic and functional analysis of metagenomes, BMC Bioinf. 9 (2008) 386.

[7] T. Tatusova, S. Ciufo, B. Fedorov, K. O'Neill, I. Tolstoy, RefSeq microbial genomes database: New representation and annotation strategy, Nucleic Acids Res. 42 (2014) D553–D559.

[8] UniProt Consortium, UniProt: A hub for protein information, Nucleic Acids Res. 43 (2015) D204–D212.

[9] W. Li, A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.

[10] R.D. Finn, A. Bateman, J. Clements, P. Coggill, R.Y. Eberhardt, S.R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E.L. Sonnhammer, J. Tate, M. Punta, Pfam: The protein families database, Nucleic Acids Res. 42 (2014) D222–D230.

[11] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.

[12] S. Suzuki, M. Kakuta, T. Ishida, Y. Akiyama, GHOSTX: An improved sequence homology search algorithm using a query suffix array and a database suffix array, PLoS One 9 (2014) e103833.

[13] H. Choe, S. Kim, J. Oh, A. Nasir, B.K. Kim, K.M. Kim, Complete genome of *Kangiella geojedonensis* KCTC 23420T, putative evidence for recent genome reduction in marine environments, Mar. Genomics (2015) http://dx.doi.org/10.1016/j.margen.2015.05.015.

[14] A.G. McDonald, S. Boyce, K.F. Tipton, ExplorEnz: The primary source of the IUBMB enzyme list, Nucleic Acids Res. 37 (2009) D593–D597.

[15] Heliconius Genome Consortium, Butterfly genome reveals promiscuous exchange of mimicry adaptations among species, Nature 487 (2012) 94–98.

[16] P. Bork, C. Bowle, C. de Vargas, G. Gorsky, E. Karsenti, P. Wincker, Tara Oceans. Tara Oceans studies plankton at planetary scale. Introduction, Science 348 (2015) 873.