OPEN ACCESS Freely available online

PLOS ONE

INDIGO – INtegrated Data Warehouse of MIcrobial GenOmes with Examples from the Red Sea Extremophiles

Intikhab Alam[1*], André Antunes[2], Allan Anthony Kamau[1], Wail Ba alawi[1], Manal Kalkatawi[1], Ulrich Stingl[3], Vladimir B. Bajic[1*]

1 Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia, **2** IBB-Institute for Biotechnology and Bioengineering, Centre of Biological Engineering, Micoteca da Universidade do Minho, University of Minho, Braga, Portugal, **3** Red Sea Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia

Abstract

Background: The next generation sequencing technologies substantially increased the throughput of microbial genome sequencing. To functionally annotate newly sequenced microbial genomes, a variety of experimental and computational methods are used. Integration of information from different sources is a powerful approach to enhance such annotation. Functional analysis of microbial genomes, necessary for downstream experiments, crucially depends on this annotation but it is hampered by the current lack of suitable information integration and exploration systems for microbial genomes.

Results: We developed a data warehouse system (INDIGO) that enables the integration of annotations for exploration and analysis of newly sequenced microbial genomes. INDIGO offers an opportunity to construct complex queries and combine annotations from multiple sources starting from genomic sequence to protein domain, gene ontology and pathway levels. This data warehouse is aimed at being populated with information from genomes of pure cultures and uncultured single cells of Red Sea bacteria and Archaea. Currently, INDIGO contains information from *Salinisphaera shabanensis*, *Haloplasma contractile*, and *Halorhabdus tiamatea* - extremophiles isolated from deep-sea anoxic brine lakes of the Red Sea. We provide examples of utilizing the system to gain new insights into specific aspects on the unique lifestyle and adaptations of these organisms to extreme environments.

Conclusions: We developed a data warehouse system, INDIGO, which enables comprehensive integration of information from various resources to be used for annotation, exploration and analysis of microbial genomes. It will be regularly updated and extended with new genomes. It is aimed to serve as a resource dedicated to the Red Sea microbes. In addition, through INDIGO, we provide our Automatic Annotation of Microbial Genomes (AAMG) pipeline. The INDIGO web server is freely available at http://www.cbrc.kaust.edu.sa/indigo.

Citation: Alam I, Antunes A, Kamau AA, Ba alawi W, Kalkatawi M, et al. (2013) INDIGO – INtegrated Data Warehouse of MIcrobial GenOmes with Examples from the Red Sea Extremophiles. PLoS ONE 8(12): e82210. doi:10.1371/journal.pone.0082210

Editor: Enrique Hernandez-Lemus, National Institute of Genomic Medicine, Mexico

Received July 2, 2013; **Accepted** October 22, 2013; **Published** December 6, 2013

Funding: IA and AAK were supported from the KAUST CBRC Base Fund of VBB. WBa and VBB were supported from the KAUST Base Funds of VBB. US was supported by the KAUST Base Fund of US. This study was partly supported by the Saudi Economic and Development Company (SEDCO) Research Excellence award to US and VBB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: Co-author Vladimir B. Bajic is PLOS ONE Editorial Board members. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: intikhab.alam@kaust.edu.sa (IA); vladimir.bajic@kaust.edu.sa (VBB)

Introduction

The Next Generation Sequencing (NGS) technologies substantially increased the throughput of genome sequencing [1-3]. Annotation of newly sequenced genomes requires a variety of experimental and computational methods [4,5], as well as integration of diverse biological information from multiple sources. Annotations stemming from information integration can be potentially used as a powerful approach in functional genomics that facilitates downstream experiments [6,7]. Data warehouses based on integrated information [8,9] are particularly useful as they open the possibility to explore content based on queries from diverse annotation attributes (e.g. genes, proteins, families, protein domains, ontologies, pathways). InterMine [10] is one of the frameworks that allows construction of such data warehouses. It has previously been

applied to developing data warehouses of model genomes resulting in resources such as FlyMine, modMine, RatMine, YeastMine, etc. For more details on InterMine features and comparison to similar systems, see reference [10] and its supplementary materials.

Here, we introduce INDIGO (Integrated Data Warehouse of Microbial Genomes), a data warehouse for microbial genomes we developed, which allows integration of annotations for exploration and analysis of microbial genomes. Currently, INDIGO contains information from three species: two bacterial species, *Salinisphaera shabanensis* [11] and *Haloplasma contractile* [12], and one archaeal species, *Halorhabdus tiamatea* [13], all isolated from deep-sea anoxic brine lakes of the Red Sea. INDIGO will be regularly updated and expanded by addition of new microbial genomes from Red Sea species.

Our contributions in this study can be summarized as follows:

- Introduction of our Automatic Annotation of Microbial Genomes (AAMG).
- Automation of data warehouse development in a high throughput manner that minimizes the intermediate steps for processing of annotation results.
- Provision to public annotations of microbial genomes being sequenced at KAUST from studies of the Red Sea environment. The number of genomes will gradually increase.

INDIGO data warehouse

Generally, newly sequenced microbial genomes are submitted to archival databases such as GenBank [14] or EMBL [15] and later they become part of curated resources such as NCBI's RefSeq database [16,17]. In order to help research on microbial genomes, a number of microbial data warehouses have been developed. A few examples are Integrated Microbial Genomes (IMG) [18], MicrobesOnline [19] Ensemble Genomes (www.ensemblgenomes.org) and MicroScope [20]. These publicly available data warehouses that contain microbial genomes information allow data browsing and comparison of genomes based on different sequence and functional features. On the other hand, these data warehouses are quite limited in capacity of query building and customized feature/attribute/entity list generation for more specific interrogation of information they contain.

We developed INDIGO, a data warehouse for microbial genomes using the InterMine framework Smith et al. [10] that allows extensive query building, customized feature/attribute/entity list creation and enrichment analysis for Gene Ontology (GO) concepts, protein domains and various pathways. In order to populate INDIGO with information from a newly sequenced genome, one needs a draft or complete genome assembly and functionally annotated the assembled genome. The INDIGO deployment requires the following five functions, namely, 1/ definition of a genomic data model of entities to be stored, 2/ data validation and population of the Postgres database, 3/ data integration, 4/ data post-processing, and 5/ web-application development. These five functions are synchronized through a project xml file that stores the location of different datasets, type of data sources and standard InterMine post-processing steps.

Results and Discussion

Genome assembly

In our case, we reassembled previously reported [11-13], three genomes based on the NGS-generated data available from Roche and Illumina sequencers and using Roche 454 Newbler assembler (www.454.com) with scaffolding option turned on in addition to using SOAPdenovo [21] and Velvet [22]. Furthermore, we use CISA [23] to obtain consensus assemblies. We improved the resulting scaffolds using SSPACE [24], GapFiller [25] and GapCloser [21]. Applying this procedure significantly improved the assemblies by reducing the number of contigs, improved N50 parameter of all three genomes. Consequently, the redundancy in the contigs observed previously using minimus [26] is now resolved. These re-assembled contigs and associated annotations are deposited to NCBI with accession numbers AFNU00000000, AFNV00000000 and AFNT00000000 for HLPCO, SSPSH and HLRTI strains, respectively.

Genome annotation

In our study, we performed genome annotation through a series of steps described in a workflow depicted in Figure 1. First, genomic sequences are passed through fastaclean (Exonerate package) [27]. Before the prediction of coding regions, the genome is masked for RNA using RNAmmer [28]and tRNAscanSE [29]. Predicted 16S rRNA genes are searched for in the NCBI prokaryotic 16S rRNA gene database to retrieve related taxonomic information that is later used in selecting the best BLAST hits. Open Reading Frame (ORF) prediction is performed using Prodigal [30], GeneMark [31] and MetageneAnnotator [32]. A series of BLAST [33,34] searches are then performed against the GenBank non-redundant (nr) [14], UniProt [35] and Kyoto Encyclopedia of Genes and Genomes (KEGG, [36]) databases including Reverse Position Specific (RPS) [37] searches against Conserved Domain Databases (especially COG and Prokaryotic Protein Clusters (PRK)). KEGG ortholog IDs are used to map relevant pathways and to display their presence on KEGG pathway maps. Interproscan analysis is carried out for GO terms and protein signature domains [38,39]. A check for annotation results is carried out using NCBI's tbl2asn and errors are manually corrected. To verify origin of each contig/genomic sequence, a global scan of BLAST results of all genes is carried out and Globally Best Taxonomies (GBT) are assigned based on species from high to low ranked top hits. Ties are broken based on the higher to lower total length of alignment reported in BLAST results by each of the top scoring species.

Benchmarking

Recently, Triplet et al. [40] thoroughly compared and benchmarked four data warehousing systems namely BioMart [41], BioXRT (mentioned in [40]), InterMine [10] and Pathway Tools [42] in a number of aspects covering accuracy, their

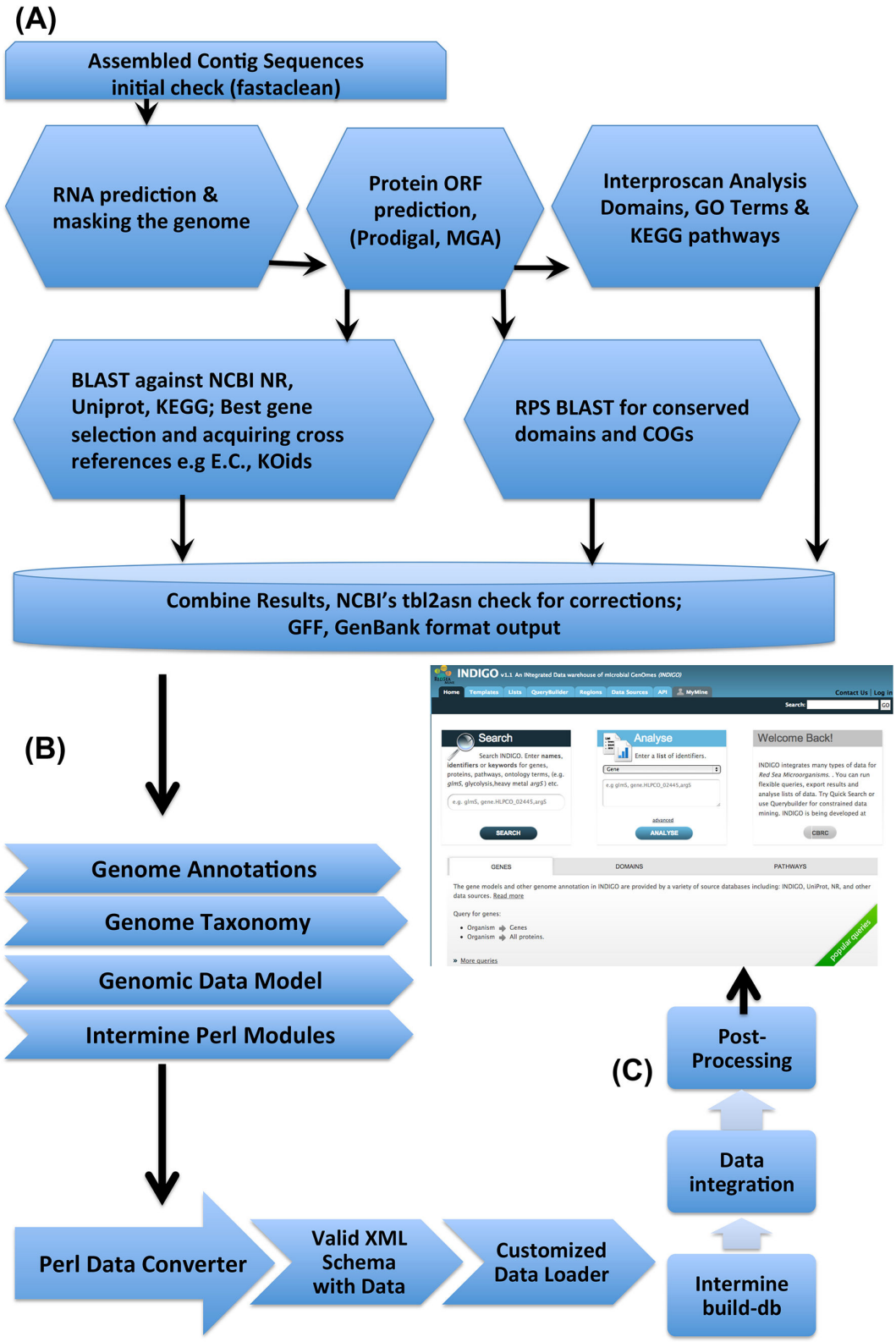

Figure 1. Workflow of annotation process and data warehousing. Here, the section marked (A) shows steps in the annotation process. Section (B) shows a PERL based conversion of annotations into an XML schema - validated using the class attributes and data types defined in the genomic model, and finally, section (C) shows the process of data warehouse development steps.

doi: 10.1371/journal.pone.0082210.g001

computational requirements and development efforts. In that study, InterMine and Pathway Tools superseded other systems. InterMine obtained the highest score, where five different aspects of data retrieval for genomics research were considered, such as aggregation, algebra, graph, data integration and sequence handling. We developed INDIGO system using the InterMine framework, but we extended it by the following features not available in InterMine.

1 Development of an automatic high throughput data warehousing pipeline to process customized annotation and their validation from newly sequenced microbial genomes. As an example, we annotated and processed annotations from three extremophile genomes from Red Sea and added to INDIGO for public data mining.
2 Addition of Genome Browser functionality.
3 Addition of BLAST interface to allow comparison of external user specified sequence data to INDIGO dataset and integration of BLAST results to either explore hit genes annotations in the INDIGO data warehouse or the auxiliary genome browser.
4 We made available special hyperlinks for KEGG assigned INDIGO pathway gene sets to be shown on publication quality pathway diagrams at KEGG website.
5 and more importantly, we made available Automatic Annotation of Microbial Genomes (AAMG) pipeline for public use through the INDIGO server.

We compared INDIGO system to InterMine and few other microbial genome data warehouses such as Integrated Microbial Genomes (IMG) [18], MicrobesOnline [19] and MicroScope [20]. Table 1 shows the list of features compared as being present or not in a data warehouse. InterMine is also included in the comparison to show what are the differences between its basic framework and our INDIGO system. This comparison clearly shows the advantages of the INDIGO system complementing InterMine and providing more control to the user in integrating annotation information that is lacking in other microbial data warehouses. MicroScope microbial genome annotations data warehouse differs from INDIGO by providing a scope for manual annotation for each and every gene individually. However, it thus requires a lot of expert manpower to deal with increasing amount newly sequenced microbial genome data. MicroScope also has a number of similar features to INDIGO, but InterMine-based INDIGO system takes lead in providing several automated and powerful routes for user-defined data integration, particularly keyword, query builder or BLAST based user-controlled gene lists making, which lead to statistically robust GO, pathway or protein domain enrichment analyses.

Benchmarking genomic annotations

To assess the quality and volume of annotations produced using our AAMG pipeline, we compare AAMG annotation results based on three publicly available datasets. Two of these datasets, namely *Escherichia coli* (*E. coli*) K12 strain and *E. coli* TY2482 strain, were recently considered in benchmarking two different annotation pipelines [43]. The third dataset is a very small genome, *Candidatus* Carsonella *ruddii* DC[44].

Recent outbreak of *E. coli* in Germany triggered the sequencing of *E. coli* O104 [45], the cause of enterohemorrhagic diarrhea. Sequencing was carried out in BGI (the strain TY2482) and multiple groups annotated this sequence. The annotation produced by AAMG pipeline for *E. coli* TY2482 is compared with annotations available from BG7 [43]. BG7 pipeline compared the annotations considering an annotations set available from Broad Institute website, http://www.broadinstitute.org/annotation/genome/Ecoli_O104_H4/assets/Ecoli_TY_2482_BGI.gbk. Results depicted in Figure 2, show that we achieve comparable performance in gene calls. Furthermore, considering the annotation in assigning gene product names, our annotation shows a significant increase in non-hypothetical products as compared to Broad Institute annotation and BG7.

BG7 compared its *E. coli* O104 (TY2482) annotation results to RAST-based annotations considering Broad Institute annotation as a gold standard. It was reported for *E. coli* TY2482 assembly version 4 [43] that BG7 predicted 5210 CDS genes, 163 false negatives and 271 false positives, while the number of genes obtained with RAST was 5446 with 116 false negatives and 321 false positives. We report AAMG-based annotation of *E. coli* TY2482 (see Supplementary materials) showing about the same number of genes predicted as BG7, but with higher numbers of functional (non-hypothetical) products and smaller number of orphan (hypothetical) genes when compared to the Broad Institute reference annotations.

In addition to *E. coli* O104 (TY2482), we also compared our results in comparison to existing annotations in NCBI for E. coli K12 and another much smaller genome, *Candidatus* Carsonella *ruddii* DC. Results show that our annotation pipeline is able to minimize hypothetical genes names through scanning multiple full protein and domain databases. Our gene calls are also very close to the existing annotations. Table 2 shows this annotation comparison.

Our annotations for these three genomes are available as a material at http://www.cbrc.kaust.edu.sa/indigo_data/. Data files and results are visualized using interactive graphs based on modified Krona package [46].

Methods and Analysis

Genomic data model

InterMine provides a core genomic data model defined with several genome entities, their attributes, syntax and relationships. We extend this core genomic model to fit our needs so as to cater to all types of annotations we receive from our annotation process. This includes data types and relationships between entities to be stored, such as attributes for organism, contigs, genes, CDS, protein domains, pathways and cross references. An example of genomic data model is provided in the materials at the website.

Data validation and population of the Postgres database

InterMine provides a built-in setup for complex data integration, post-processing and web-application development. Data integration is heavily dependent on genomic model

Table 1. A comparison of features from different microbial data warehouses.

	INDIGO	InterMine	Integrated Microbial Genomes	Microbes Online	MicroScope
Basic Data					
Chromosome/Contigs	Yes	Yes	Yes	Yes	Yes
Genes	Yes	Yes	Yes	Yes	Yes
Proteins	Yes	Yes	Yes	Yes	Yes
Expression data	No	Yes	Yes	Yes	Yes
Functional genomics					
Gene Ontology	Yes	Yes	Yes	Yes	Yes
KEGG Pathways	Yes	Yes	Yes	Yes	Yes
Interpro Domains	Yes	Yes	Yes	Yes	Yes
Cross references	Yes	Yes	Yes	Yes	Yes
Data Integration and Functional analysis					
Showing assigned KEGG pathway diagrams	Yes	No	No	No	No
Individual Feature (Gene/Protein/Pathway) list generation	Yes	Yes	Yes	Yes	Yes
Multiple Feature (Gene/Protein/Pathway) list generation	Yes	Yes	No	Yes, limited	Yes
Keyword search	Yes	Yes	Yes	Yes	Yes
Keyword search against all attributes	Yes	Yes	No	Yes	No
Filter keyword search results based on categories	Yes	Yes	No	Yes	Yes
Keyword search for feature list generation	Yes	Yes	No	No	Yes
BLAST search to feature list generation	Yes	No	Yes	Yes	Yes
Query builder to user selected all/multiple feature list generation	Yes	Yes	No	No	Yes
Save / share queries	Yes	Yes	No	Yes	Yes
Feature list analysis; GO enrichment	Yes	Yes	No	No	No
Feature list analysis; Pathway enrichment	Yes	Yes	No	No	No
Feature list analysis; Protein enrichment	Yes	Yes	No	No	No
Adding additional attribute to generated lists	Yes	Yes	No	No	No
List summary functions	Yes	Yes	No	No	No
List filtering functions	Yes	Yes	Yes	Yes	Limited
List export	Yes	Yes	Yes	Yes	Yes
Save / share lists	Yes	Yes	No	Yes	Yes
Genome Browser	Yes	Yes	Yes	Yes	Yes
Comparative Genomics					
Compare different genomic features e.g.via keyword search	Yes	Yes	Yes	Yes	Yes
Compare sequences via BLAST	Yes	No	Yes	Yes	Yes
Compare genomes based on other tools	No	No	Yes	Yes	Yes
Data access					
Web server based data access	Yes	Yes	Yes	Yes	Yes
Remote access via API (PERL, JAVA, RUBY, PYTHON)	Yes	Yes	No	Yes	No
Bulk Download	Yes	Yes	Yes	Yes	Yes
User selected single feature list based download	Yes	Yes	Yes	Yes	Yes
User integrated feature list based download	Yes	Yes	No	No	Yes, limited.
Genome Annotation					

Table 1 (continued).

	INDIGO	InterMine	Integrated Microbial Genomes	Microbes Online	MicroScope
Public microbial genome annotation	Yes	No	Yes	Limited, uses rast and takes six months	Annotation_editor (manual)
User genome annotation job history	Yes	No	Yes	Yes	Manual
Genome Annotation features					
operon finding	No	No	Yes	Yes	Yes
promoter/terminator finding	No	No	Yes	No	Yes
RNA detection (rRNA/tRNA)	Yes	No	Yes	No	Yes
Protein gene prediction (multiple methods)	Yes	No	Yes	No	Yes
RNA vs. Protein overlap resolution	Yes	No	Yes	No	Yes
HPC BLAST for Proteins to UniProt	Yes	No	No	Yes	Yes
HPC BLAST for Proteins to NCBI NR	Yes	No	No	Yes	No
HPC BLAST for Proteins to NCBI COG	Yes	No	Yes	Yes	Yes
HPC BLAST for Proteins to NCBI CDD	Yes	No	No	No	No
HPC BLAST for Proteins to KEGG	Yes	No	Yes	Yes	Yes
HPC Interproscan domain finding for Proteins	Yes	No	Yes	Yes	Yes
Global Best Taxonomy (GBT) distribution analysis	Yes	No	No	No	No
Annotation data integration to GFF format	Yes	No	Yes	No	No
Annotation data integration to GenBank format	Yes	No	No	Yes	Yes
Annotation data integration to TBL format	Yes	No	No	No	Yes
Annotation data checking using tbl2asn	Yes	No	No	No	No
Annotation data process to NCBI sqn submission format	Yes	No	No	No	No
Annotation data packing into validated xml for data warehouse	Yes	No	No	No	No
Hierarchical classification of COG annotations and visualization	Yes	No	No	Yes	No
Hierarchical classification of GO annotations and visualization	Yes	No	No	No	No
Hierarchical classification of GBT annotations and visualization	Yes	No	No	No	No
Hierarchical classification of InterPro domains annotations and visualization	Yes	No	No	Yes	Yes
Hierarchical classification of ALL annotations and visualization	Yes	No	No	No	No
Immediate access to all data files and visualizations	Yes	No	No, sso accounts	Yes	Yes

doi: 10.1371/journal.pone.0082210.t001

defined with data types and relationships between entities to be stored. Once a genomic model is defined, one can perform a check for the annotation that is to be loaded into the database. Our system first validates the annotation in reference to the defined genomic model using InterMine's Model and Document Perl Modules. It then prepares an xml schema filled with data that is ready to be loaded to the backend Postgres database. InterMine loads data into the database using pre-defined 'sources' for different types of data packed in different formats. For example, to load genes data packed in the gff format, a Java-based data converter is available, but it assumes specific tags and fields. For customized data loading we developed prokaryotic-annots-xml, available as Supplementary material here, which allows loading of our validated annotations packed in xml format. InterMine's build-db setup reads the generated annotation using prokaryotic-annots-xml source and loads the data by defining and populating different annotation tables automatically.

Data integration, post-processing and web-application development

Data integration in the InterMine's framework is a crucial step. It integrates data from sources provided in the project xml file and performs multiple checks (e.g. the absence of empty fields, the absence of duplicate data being stored, etc.). We only provide database identifiers in the annotation xml, for example, for GO or Interpro protein domains (IprD), and InterMine system integrates corresponding detailed

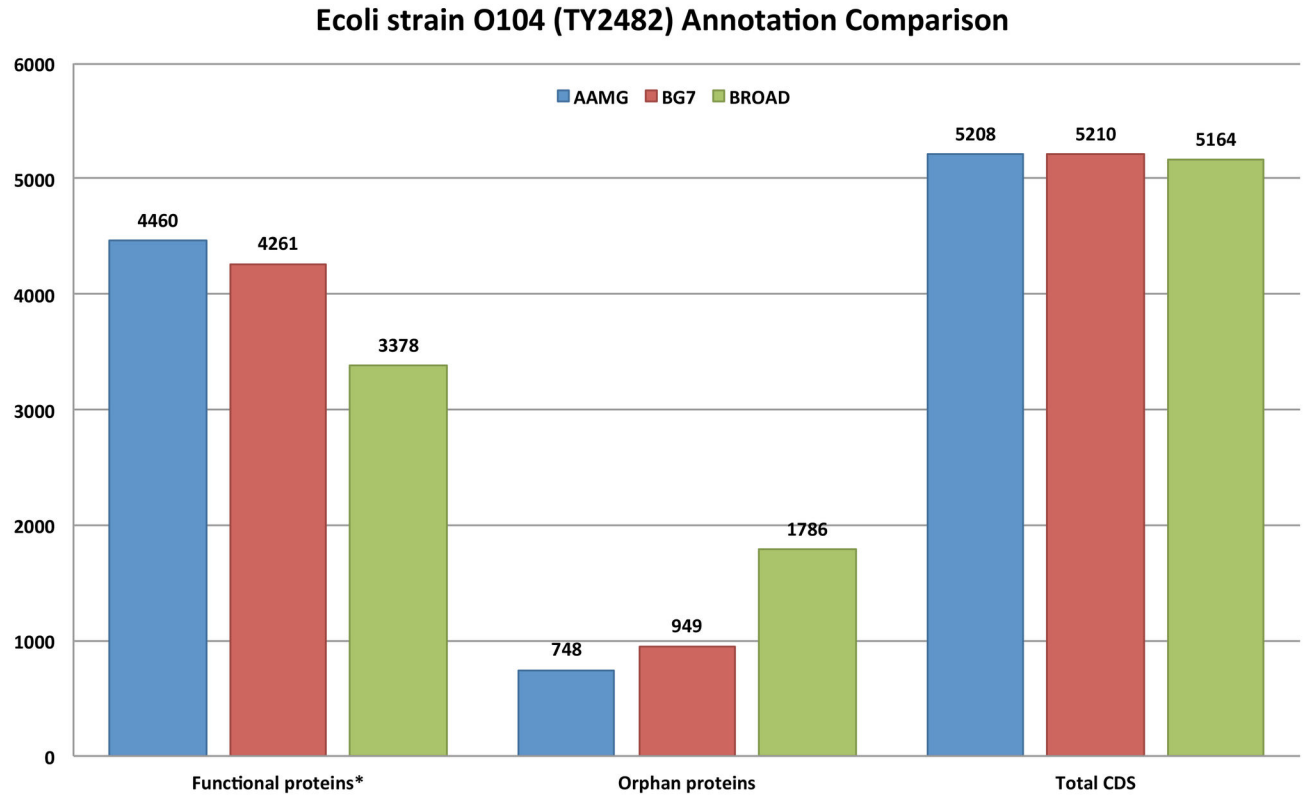

Figure 2. Annotation comparison for *E. coli* O104 (TY2482) among AAMG pipeline, BG7 and reference annotation set from Broad Institute. Regarding the CDS annotation AAMG ranks second (with only 2 CDS region less annotated than BG7), while in annotation of orphan (hypothetical) CDS products (the less the better) and in annotation of functional (non-hypothetical) CDS products (the more the better) AAMG performs the best.

doi: 10.1371/journal.pone.0082210.g002

Table 2. Results of AAMG Annotations compared with NCBI or BROAD institute sets.

	E. coli K12 W3110		*E. coli* TY2482		*C. ruddii* DC	
Gene calls	**AAMG**	**NCBI**	**AAMG**	**BROAD**	**AAMG**	**NCBI**
CDS	4340	4337	5208	5164	190	207
rRNA	22	22	22	22	3	3
tRNA	82	86	97	97	27	28
Total	4444	4445	5327	5288	220	238
False Negatives	235	236	50	11	4	22
Functional genes	3866	3730	4591	3502	182	191
Orphan genes	578	715	736	1786	38	47
Gene calls	**Genes by AAMG**	**% of NCBI genes**	**Genes by AAMG**	**% of BROAD genes**	**Genes by AAMG**	**% of NCBI genes**
Detected* identical	3876	87.20	5172	97.81%	205	86.13
Detected similar**	333	7.49	105	1.99%	11	4.62
Not Detected	236	5.31	11	0.21%	22	9.24
Total	4445		5288		238	

* Genes are identical when both start and stop positions are exactly the same.

** Genes are similar if start or stop positions are in the same region with an offset up to 50 bases.

doi: 10.1371/journal.pone.0082210.t002

annotations from complete GO or IprD source files defined in the project xml.

There are several built-in post-processing steps available in the InterMine framework such as create-search-index, transfer-

Table 3. Data warehouse development stages using InterMine.

INDIGO Steps	Action	Time (seconds)
build database tables	build-db	2
data integration	prokredsea-HLPCO-largexml	59
data integration	prokredsea-HLRTI-largexml	61
data integration	prokredsea-SSPSH-largexml	68
data integration	Sequence ontology	56
data integration	interpro	164
data integration	Gene ontology	1043
post-processing	create-references	28
post-processing	make-spanning-locations	21
post-processing	create-chromosome-locations-and-lengths	40
post-processing	transfer-sequences	89
post-processing	create-bioseg-location-index	15
post-processing	create-attribute-indexes	38
post-processing	summarise-objectstore	31
post-processing	create-autocomplete-index	20
post-processing	create-search-index	59
total time taken		1794

doi: 10.1371/journal.pone.0082210.t003

sequences, etc., that allow for quick indexing of the data. For INDIGO, in order to have all the functionality available, we run all post-processing steps. Table 3 shows different stages in our data warehouse development along with the processing time using InterMine framework.

Web-application templates are available in the InterMine framework and we customized them to fit our requirements. For example, report pages for genome features such as genes, proteins, domains, and pathways are customized according to the data available including hyperlinks to external databases. One of the interesting external links allows for displaying KEGG pathway diagrams showing presence of the KEGG Ortholog ids to which the explored genome is mapped. Such a display shows which elements of the reference pathways are present or missing from the genome being examined. InterMine allows packaging of the web-application as a Web Application Archive or WAR-file that is then deployed on the Tomcat Apache server (http://tomcat.apache.org).

INDIGO Web Interface Organization

INDIGO is equipped with a number of features that allow for the exploration and analysis of the deposited information. INDIGO front-end is organized into different main pages accessible through tabs, namely 'Home', 'Templates', 'Lists', 'QueryBuilder', 'Regions', 'Data', 'API', 'BLAST' and 'MyMine', where each tab provides access to the data in different ways.

The INDIGO 'Home' page presents options for quick keyword search, analysis of a list of genes/proteins and the use of predefined templates to perform queries. The 'Template' tab shows all predefined templates to perform queries such as Organism->Protein which help to obtain all proteins in a genome. The 'Lists' tab provides access of all feature types; for example, selecting a feature type such as gene, protein, protein domain, etc. and providing a list of identifiers, makes a list of items with default attributes that can be saved or further analyzed. The 'QueryBuilder' tab provides the most exhaustive functionality for building queries in INDIGO and it provides more control to include (show option) or limit (constrain option) for different feature types and their attributes. 'Regions' tab provides access to all features present in a given genomic coordinate range. 'Data' tab provides general information about the genomic data sets included in the data warehouse, e.g. genome assembly statistics, counts and links to contigs, ORF sequences, archaeal/bacterial genome completeness statistics based on counts of archaeal/bacterial core COGs [47], minpath-based [32] KEGG pathway association statistics, etc. API provides details on how to access data warehouse using PERL, Python, Ruby and Java programming languages. The 'BLAST' tab allows users to carry out Basic Local Alignment Search Tool (BLAST) based similarity search for a DNA or protein sequence of interest with genes in INDIGO. The result of BLAST search is shown as a list where users can save and select an individual or all hits for further GO, Pathways or protein domain enrichment analysis. Finally, 'MyMine' shows an interface to Automatic Annotation of Microbial Genomes (AAMG) pipeline, user-specific lists and queries performed and saved by a user once the user creates an account on the system. Individual report pages for genomic features provide details and hyperlinks for several related attributes including JBrowse [48] visualization.

Use of INDIGO and its features

When analyzing a new genome, majority of questions can be summarized in 'What', 'Where' and 'How' context. For example, to see whether a gene, protein, protein domain, GO term or a pathway of interest can be found in INDIGO, a search mechanism can help. For 'Where' context questions, the 'Region' search option in INDIGO can list all the genomic

features in a given range of genomic coordinates. For complex questions of the type e.g. 'What is a list of genes involved in pathway X and what are their protein domains and associated GO terms' more control on what is being searched is needed and here it is provided through Query Builder. More details on features of INDIGO, such as a quick and easy keyword search, query builder search, analysis of genomic feature (such as gene, protein, protein domains) lists, genomic region search, and enrichment analysis for GO, protein domains and pathways, are shown in few examples in what follows.

Keyword and Query Builder Search

In the INDIGO system, a keyword search, as well as a more extensive query builder search option, are provided. The keyword search option provides a very simple interface to the underlying annotation data. It is very fast since all the keywords in the database are indexed. Query builder, however, provides more control over annotation classes and attributes to be searched, constrained or viewed. It is possible to combine several queries through constraint logic. Figure 3 shows an example of Query-Builder interface to INDIGO.

Region Search

In order to find out characteristics of a particular genomic region, one can use region search. When coordinates for the specific genomic region are provided, the region search allows for selection of additional upstream and downstream regions, as well as features like gene or intergenic region, etc. (Figure 4). Results can be exported in several different formats. We integrated JBrowser [48] based visualization of our genomic features in the region search results page. In the genome browser users can look up gene names or particular coordinates of genomes to view underlying features. Available tracks are DNA, gene and InterPro domains.

Analysis of Lists

INDIGO makes use of different types of lists. For example, a list could be the list of genes/proteins, or protein domains, etc. Results from keyword search or query builder, can be saved as a list. A click on the saved list link automatically shows GO, protein domain and pathway enrichment, as shown in an example in Figure 5.

The user is also able to save all enriched genes, make sub-lists, view individual gene report pages, or export results. Enrichment analysis provided for a list includes p-values based on hypergeometric distribution with several multiple testing correction options (for further details on the enrichment process, see https://intermine.readthedocs.org/en/1.1/embedding/list-widgets/enrichment-widgets/).

Current content of INDIGO

The King Abdullah University of Science and Technology (KAUST) has in its focus areas the biodiversity and microorganisms of the Red Sea. INDIGO is populated with information from three extremophiles from the Red Sea, whose genomes have been previously reported by our team [11-13]. The details are provided in what follows.

Red Sea environment

The Red Sea is one of the warmest, most saline and most nutrient-poor oceanic water bodies in the world [49,50]. It also hosts several deep-sea anoxic brine lakes, which are considered some of the most remote and extreme environments on Earth [51]. The brines markedly differ from overlying seawater and are unique due to the combination of multiple extremes namely high salinity (7-fold increase), high temperature (up to 70°C), high concentration of heavy metals (1,000- to 10,000-fold increase in concentration), high hydrostatic pressure and anoxic conditions. Despite this combination of multiple environmental extremes, they have been shown to harbor a very high biodiversity, with identification of several new phylogenetic lineages and isolation of several new extremophiles [51].

Three Red Sea extremophiles in INDIGO

Three extremophilic microbes, previously isolated from the deep-sea anoxic brine lakes, were selected as part of a genome-sequencing project due to their phylogenetic position, peculiar features and unique biotope. Analysis of their draft genomes provides us with a first glimpse on some of their unusual characteristics and the ways they cope with living in such a harsh environment [11-13].

Salinisphaera shabanensis

Salinisphaera shabanensis was isolated from the brine-seawater interface of Shaban Deep [52]. It represented a new order within the *Gammaproteobacteria*, and displayed a remarkable physiological versatility. Indeed, *Salinisphaera shabanensis* had quite broad growth ranges for oxygen, temperature, NaCl, pressure, and, to a smaller degree, pH [52].

Haloplasma contractile

Haloplasma contractile was isolated from the brine-sediment interface of Shaban Deep. Phylogenetically it represented a novel lineage within the Bacteria with branching position between the Firmicutes and Tenericutes (Mollicutes), with no close relatives [53]. The most striking feature of *Haloplasma* is its unusual morphology and unique cellular contractility cycle.

Halorhabdus timatea

Halorhabdus tiamatea was isolated from the brine-sediment interface of the Shaban Deep [54] using fluorescence *in situ* hybridization coupled with the "optical tweezers" technique [55,56]. It was described as a new species and is currently still the only member of the Archaea to have been described from a deep-sea anoxic brine.

Features of the three Red Sea extremophiles from INDIGO

In Table 4 we present summary of the basic genomic features associated with the re-assembly of three microorganisms included in INDIGO.

INDIGO provides easy and quick access to genomic annotations of microbial species at the levels of chromosomes, genes, and proteins, as well as to the associated GO and

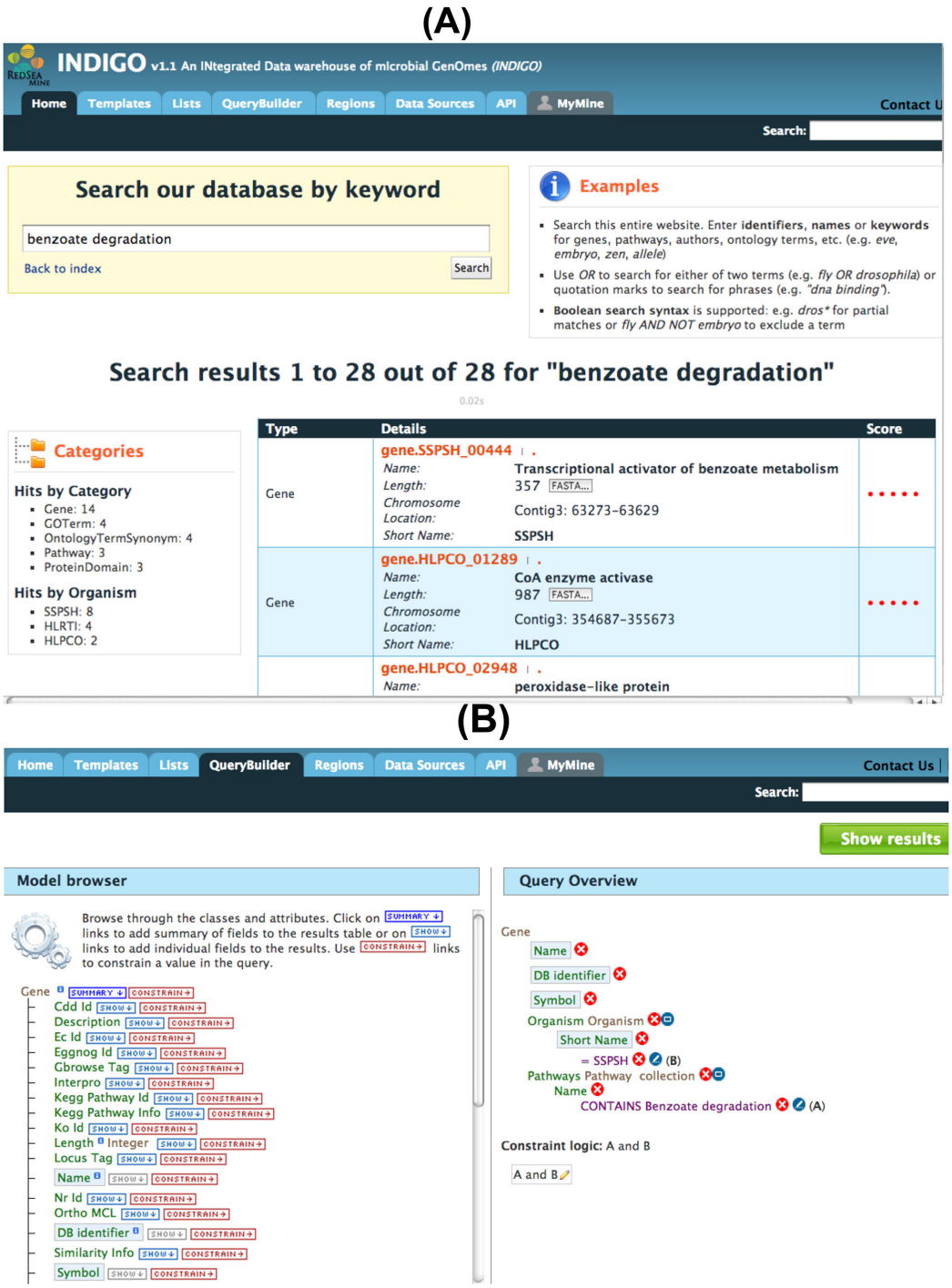

Figure 3. A) Keyword and B) Query builder search interface to INDIGO. The keyword search interface shows an example of the search for "benzoate degradation". Results are categorized on the left side of the resulting page, showing the number of hits found for genes, domains, pathways, etc. These results are further categorized into hits per genome for different organisms. Clicking on any of these categories shows filtered results. The query builder interface has an option to include or constrains an annotation class attribute, e.g. pathway name is constrained for "benzoate degradation", while the organism attribute 'short name' is constrained to "SSPSH". The annotation feature class attributes to be included in the result list here are gene db identifier, symbol, organism's short name and pathway name. User can select any of the available annotation class attributes making it possible to integrate annotation from several different sources. Results of constrained query builder search are shown as a list. There are summary and filter options on the list page that allow a user to further analyze these results.

doi: 10.1371/journal.pone.0082210.g003

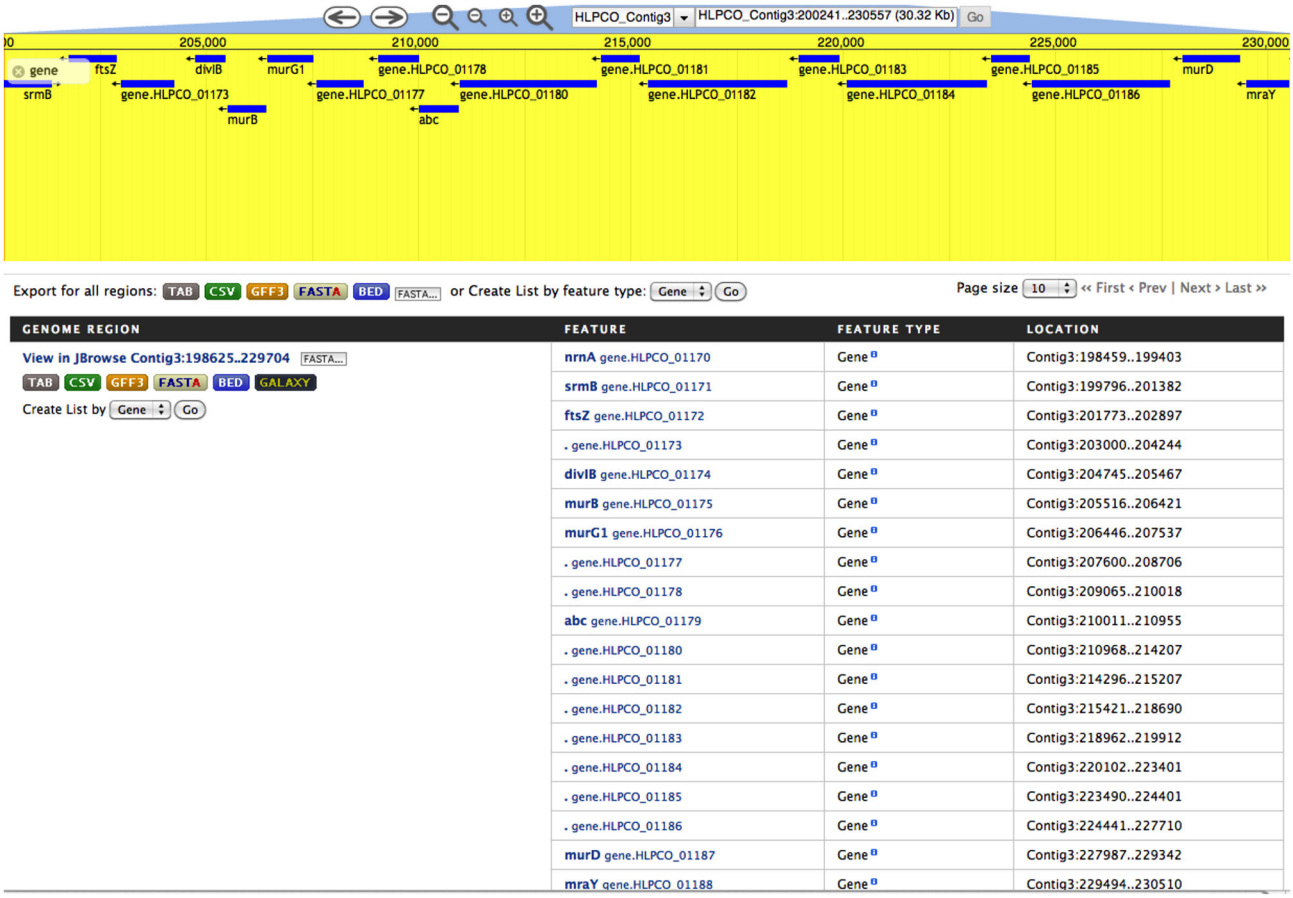

Figure 4. Region search interface. This figure shows features (genes) for a region using coordinates (Contig3:198625-229704) from organism *Haloplasma contractile* (HLPCO). This region shows the cell Division and Cell Wall (DCW) biosynthesis gene cluster. An integrated genome browser view available via Region search results page, shows here the arrangement of genes in this region of the contig from HLPCO . The table below this section shows genome region, data export options, basic details of the feature (genes), type of features and their location on the genome. The create list by feature link saves this gene list in the data warehouse for further analysis. This list stays permanently if the user is logged in.

doi: 10.1371/journal.pone.0082210.g004

pathways. The top 10 pathways based on the number of genes assigned to each of these extremophiles as found by INDIGO are shown in Table 5.

Examples of exploration of Red Sea Extremophiles via INDIGO

Region search: Analysis of the *dcw* gene cluster in *Haloplasma contractile*. The most remarkable features of *Haloplasma contractile* include its unusual morphology and contraction cycle and these provided clear targets for genomic-based exploration. While some aspects of the genetic control of cellular morphology remain unclear, the *dcw* gene cluster seems to play a central role. Gene context is particularly relevant, as morphology is impacted by presence or absence of specific genes, together with relative position and distance within this gene cluster [57,58].

Using the region search of INDIGO we were able to locate the *murD* - one of the central genes of the *dcw* gene cluster. Furthermore, we analyzed the genomic context of *murD* (upstream and downstream regions) and successfully demonstrated multiple gene insertions and disruption of the *murD-ftsW-murG*

gene order, see Figure 4. Such a disruption would justify the atypical morphology of *H. contractile* as they have been previously implicated in all non-rod morphologies currently known [58,59].

Pathway search: Benzoate degradation in *Salinisphaera shabanensis*. Aromatic compounds are abundant, widely distributed and known to constitute some of the most prevalent and persistent pollutants in the environment [60]. Some microbes have evolved complex machinery and metabolic pathways for their degradation [61] with benzoate being widely used as a model compound for studying their catabolism.

Based on previous detection of a variety of complex hydrocarbons in Shaban Deep [62], we looked into genomic-based evidences for possible aromatic compound catabolic

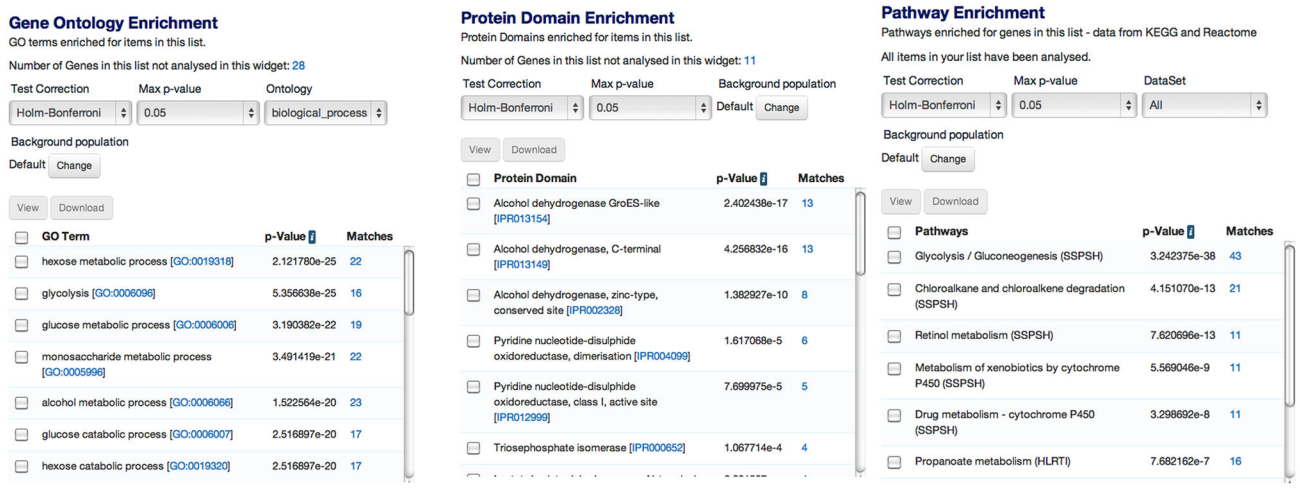

Figure 5. A) Gene Ontology, B) Protein Domain and C) Pathway enrichment analysis. The figure shows a snapshot obtained in case when a term "cell cycle" was searched through the keyword search option and resulting genes were saved in a list that shows enrichment of GO, protein domain and pathways in comparison to the rest of the data in INDIGO. The number of hits shown for reach category can be saved as lists for further analysis.

doi: 10.1371/journal.pone.0082210.g005

Table 4. Basic annotated features of the three Red Sea extremophiles in INDIGO.

Organism	Contigs	N50	ORFs	rRNAs	tRNAs
Haloplasma contractile	34	347868	3036	4	27
Halorhabdus tiamatea	72	58136	3287	3	40
Salinisphaera shabanens	41	129079	3530	3	46

doi: 10.1371/journal.pone.0082210.t004

Table 5. Top 10 pathways from each of the three extremophiles.

Haloplasma contractile	**Genes**	***Salinisphaera shabanens***	**Genes**	***Salinisphaera shabanens***	**Genes**
ABC transporters	115	Two-component system	182	ABC transporters	88
Purine metabolism	69	ABC transporters	131	Purine metabolism	74
Two-component system	67	Purine metabolism	96	Ribosome	64
Pyrimidine metabolism	56	Methane metabolism	78	Pyrimidine metabolism	60
Ribosome	56	Oxidative phosphorylation	75	Oxidative phosphorylation	55
Tyrosine metabolism	52	Butanoate metabolism	73	Amino sugar and nucleotide sugar metabolism	53
Amino sugar and nucleotide sugar metabolism	50	Benzoate degradation	71	Two-component system	50
Starch and sucrose metabolism	49	Fatty acid metabolism	70	Methane metabolism	46
Methane metabolism	46	Arginine and proline metabolism	63	Starch and sucrose metabolism	40
Histidine metabolism	40	Pyruvate metabolism	60	Cysteine and methionine metabolism	39

doi: 10.1371/journal.pone.0082210.t005

capability. The use of the query builder search (Figure 3) in INDIGO and its mapping onto KEGG pathway (Figure 6) led us to promising results, with the identification of an almost complete branch of the benzoate degradation pathway in *Salinisphaera shabanensis*. Such valuable information obtained through the simple use of INDIGO will aid in search for target missing genes and/or design downstream laboratorial experiments to confirm the functionality of this pathway, and explore possible future applications.

Conclusion

The new data warehouse system, INDIGO, enables users to combine information from different sources of annotation for

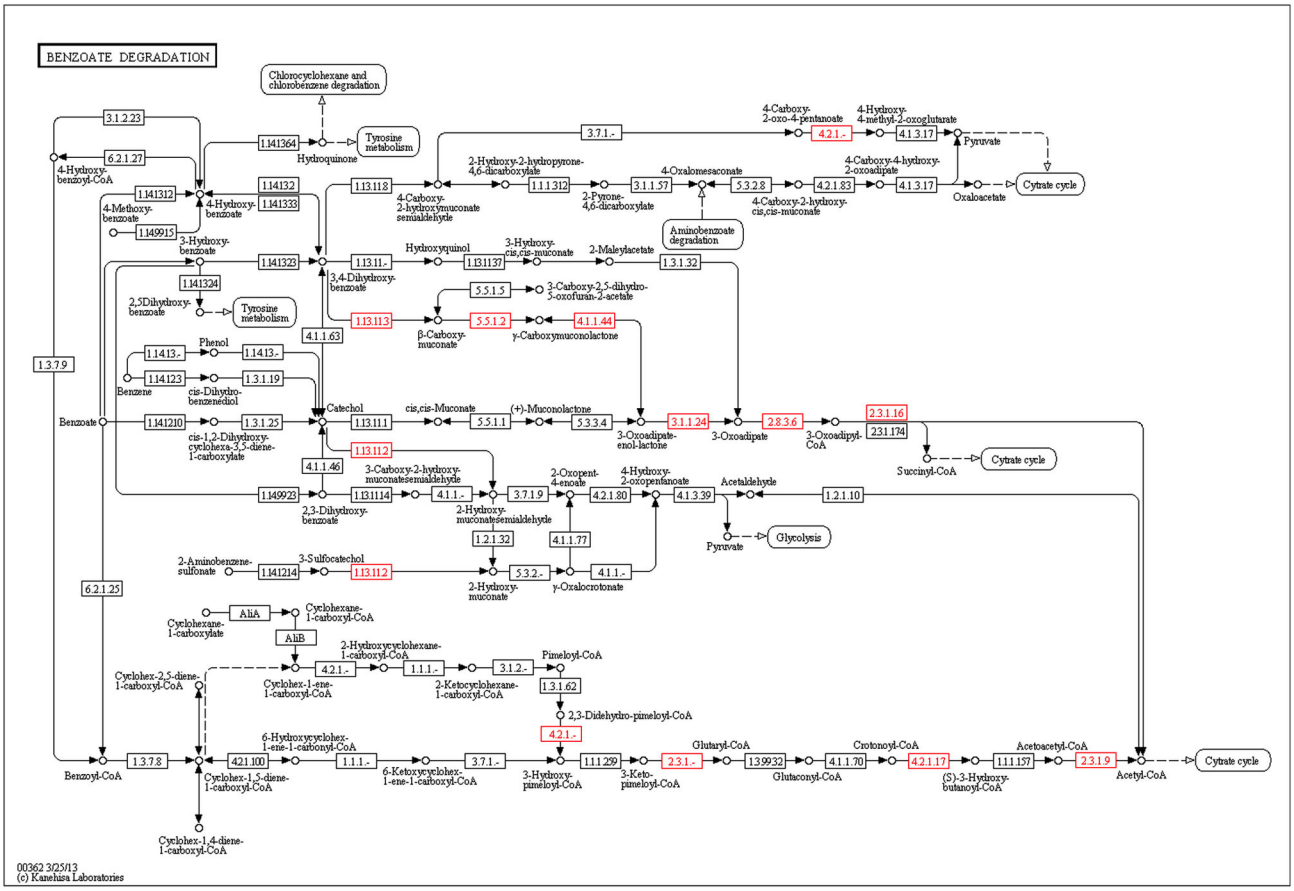

Figure 6. Benzoate degradation in *Salinisphaera shabanensis*. The genes from *Salinisphaera shabanesis* associated with Benzoate degradation pathway by INDIGO are shown in Red. INDIGO developed a functionality, available for all pathways present in INDIGO, that generates a specific URL to automatically display KEGG Orthologs from INDIGO on to pathway diagrams at KEGG webserver.

doi: 10.1371/journal.pone.0082210.g006

further specific or general analysis. This data warehouse of Red Sea microorganisms currently contains information about three genomes (two bacterial and one archaeal). Considering the unique biodiversity present in the Red Sea, KAUST has undertaken a large sequencing effort starting from metagenomes to cultured and uncultured single cell amplified genomes. The plethora of sequencing data produced requires a high throughput assembly, annotation and data warehousing pipelines. This work shows the basic framework through which these pipelines can be used in a high throughput manner to properly warehouse the increasing amount of data for targeted studies. Additional genomes will include both, genomes of pelagic bacteria and archaea, as well as more extremophiles from the brine pools of the Red Sea.

Acknowledgements

Authors thank Julie Sullivan, Alex Kalderimis and intermine.org team for their help in understanding the Intermine system. Authors are also thankful to David Kamanda Ngugi and Mamoon Rashid from the Red Sea Center at KAUST for their help in testing the INDIGO system and to Arturo Magaña Mora for designing the INDIGO logo.

Author Contributions

Conceived and designed the experiments: IA VBB. Performed the experiments: IA WBa MK. Analyzed the data: IA AA US VBB. Contributed reagents/materials/analysis tools: IA. Wrote the manuscript: IA AA US VBB. Developed INDIGO system: IA. Developed web implementation: IA AAK.

References

1. MacLean D, Jones JD, Studholme DJ (2009) Application of 'next-generation' sequencing technologies to microbial genetics. Nat Rev Microbiol 7: 287-296. PubMed: 19287448.
2. Pop M, Salzberg SL (2008) Bioinformatics challenges of new sequencing technology. Trends Genet 24: 142-149. doi:10.1016/j.tig.2007.12.006. PubMed: 18262676.
3. Médigue C, Moszer I (2007) Annotation, comparison and databases for hundreds of bacterial genomes. Res Microbiol 158: 724-736. doi:10.1016/j.resmic.2007.09.009. PubMed: 18031997.
4. Richardson EJ, Watson M (2013) The automatic annotation of bacterial genomes. Brief Bioinform 14: 1-12. doi:10.1093/bib/bbs007. PubMed: 22408191.
5. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57-63. doi:10.1038/nrg2484. PubMed: 19015660.
6. Garcia Castro A, Chen YP, Ragan MA (2005) Information integration in molecular bioscience. Appl Bioinformatics 4: 157-173. doi:10.2165/00822942-200504030-00001. PubMed: 16231958.
7. Stein LD (2003) Integrating biological databases. Nat Rev Genet 4: 337-345. doi:10.1038/ni909. PubMed: 12728276.
8. Triplet T, Butler G (2011) Systems biology warehousing: challenges and strategies toward effective data integration. In Proc. 3rd International Conference on Advances in Databases, Knowledge, and Data Applications, St. Maarten. IARIA. pp 34-40.
9. O'Malley MA, Soyer OS (2012) The roles of integration in molecular systems biology. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 43: 58-68. doi:10.1016/j.shpsc.2011.10.006.
10. Smith RN, Aleksic J, Butano D, Carr A, Contrino S et al. (2012) InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics 28: 3163-3165. doi:10.1093/bioinformatics/bts577. PubMed: 23023984.
11. Antunes A, Alam I, Bajic VB, Stingl U (2011) Genome sequence of Salinisphaera shabanensis, a gammaproteobacterium from the harsh, variable environment of the brine-seawater interface of the Shaban Deep in the Red Sea. J Bacteriol 193: 4555-4556. doi:10.1128/JB.05459-11. PubMed: 21705588.
12. Antunes A, Alam I, El Dorry H, Siam R, Robertson A et al. (2011) Genome sequence of Haloplasma contractile, an unusual contractile bacterium from a deep-sea anoxic brine lake. J Bacteriol 193: 4551-4552.
13. Antunes A, Alam I, Bajic VB, Stingl U (2011) Genome sequence of Halorhabdus tiamatea, the first archaeon isolated from a deep-sea anoxic brine lake. J Bacteriol 193: 4553-4554. doi:10.1128/JB.05462-11. PubMed: 21705593.
14. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ et al. (2013) GenBank. Nucleic Acids Res 41: D36-D42. doi:10.1093/nar/gks1035. PubMed: 23193287.
15. Kodama Y, Shumway M, Leinonen R (2012) The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res 40: D54-D56. doi:10.1093/nar/gkr1263. PubMed: 22009675.
16. Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 40: D130-D135. doi:10.1093/nar/gks463. PubMed: 22121212.
17. Markowitz VM (2007) Microbial genome data resources. Curr Opin Biotechnol 18: 267-272. doi:10.1016/j.copbio.2007.04.005. PubMed: 17467973.
18. Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E et al. (2012) IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res 40: D115-D122. doi:10.1093/nar/gks596. PubMed: 22194640.
19. Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK et al. (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. Nucleic Acids Res, 38: D396–400. PubMed: 19906701.
20. Vallenet D, Belda E, Calteau A, Cruveiller S, Engelen S et al. (2013) MicroScope—an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. Nucleic Acids Res 41: D636-D647. doi:10.1093/nar/gks1194. PubMed: 23193269.
21. Luo R, Liu B, Xie Y, Li Z, Huang W et al. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1: 1-6. doi:10.1186/2047-217X-1-1. PubMed: 23587310.
22. Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821-829. doi:10.1101/gr.074492.107. PubMed: 18349386.
23. Lin S-H, Liao Y-C (2013) CISA: Contig Integrator for Sequence Assembly of Bacterial Genomes. PLOS ONE 8: e60843. doi:10.1371/journal.pone.0060843. PubMed: 23556006.
24. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27: 578-579. doi:10.1093/bioinformatics/btq683. PubMed: 21149342.
25. Boetzer M, Pirovano W (2012) Toward almost closed genomes with GapFiller. Genome Biol 13: R56. doi:10.1186/gb-2012-13-6-r56. PubMed: 22731987.
26. Sommer DD, Delcher AL, Salzberg SL, Pop M (2007) Minimus: a fast, lightweight genome assembler. BMC Bioinformatics 8: 64. doi:10.1186/1471-2105-8-64. PubMed: 17324286.
27. Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6: 31. doi:10.1186/1471-2105-6-31. PubMed: 15713233.
28. Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T et al. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35: 3100-3108. doi:10.1093/nar/gkm160. PubMed: 17452365.
29. Schattner P, Brooks AN, Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res 33: W686-W689. doi:10.1093/nar/gki366. PubMed: 15980563.
30. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW et al. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11: 119. doi:10.1186/1471-2105-11-119. PubMed: 20211023.
31. Besemer J, Lomsadze A, Borodovsky M (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 29: 2607-2618. doi:10.1093/nar/29.12.2607. PubMed: 11410670.
32. Noguchi H, Taniguchi T, Itoh T (2008) MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res 15: 387-396. doi:10.1093/dnares/dsn027. PubMed: 18940874.
33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410. doi:10.1016/S0022-2836(05)80360-2. PubMed: 2231712.
34. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402. doi:10.1093/nar/25.17.3389. PubMed: 9254694.
35. Consortium TU (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40: D71-D75. doi:10.1093/nar/gkr981. PubMed: 22102590.
36. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40: D109-D114. doi:10.1093/nar/gkr988. PubMed: 22080510.
37. Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY et al. (2013) CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Res 41: D348-D352. doi:10.1093/nar/gks1243. PubMed: 23197659.
38. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A et al. (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37: D211-D215. doi:10.1093/nar/gkn785. PubMed: 18940856.
39. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK et al. (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 40: D306-D312. doi:10.1093/nar/gkr948. PubMed: 22096229.
40. Triplet T, Butler G (2013) A review of genomic data warehousing systems. Brief Bioinform: ([MedlinePgn:]) PubMed: 23673292.
41. Zhang J, Haider S, Baran J, Cros A, Guberman JM et al. (2011) BioMart: a data federation framework for large collaborative projects. Database (Oxford). p. bar038. PubMed: 21930506
42. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM et al. (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief Bioinform 11: 40-79. doi:10.1093/bib/bbp043. PubMed: 19955237.
43. Pareja-Tobes P, Manrique M, Pareja-Tobes E, Pareja E, Tobes R (2012) BG7: a new approach for bacterial genome annotation designed for next generation sequencing data. PLOS ONE 7: e49239. doi:10.1371/journal.pone.0049239. PubMed: 23185310.
44. Nakabachi A, Ueoka R, Oshima K, Teta R, Mangoni A et al. (2013) Defensive bacteriome symbiont with a drastically reduced genome.

Curr Biol 23: 1478-1484. doi:10.1016/j.cub.2013.06.027. PubMed: 23850282.
45. Rohde H, Qin J, Cui Y, Li D, Loman NJ, et al. (2011) Open-source genomic analysis of Shiga-toxin–producing E. coli O104: H4. New England Journal of Medicine 365: 718-724.
46. Ondov BD, Bergman NH, Phillippy AM (2011) Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 12: 385. doi: 10.1186/1471-2105-12-385. PubMed: 21961884.
47. Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH et al. (2010) A catalog of reference genomes from the human microbiome. Science 328: 994-999. doi:10.1126/science.1183605. PubMed: 20489017.
48. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH (2009) JBrowse: a next-generation genome browser. Genome Res 19: 1630-1638. doi:10.1101/gr.094607.109. PubMed: 19570905.
49. Ngugi DK, Antunes A, Brune A, Stingl U (2012) Biogeography of pelagic bacterioplankton across an antagonistic temperature–salinity gradient in the Red Sea. Mol Ecol 21: 388-405. doi:10.1111/j.1365-294X.2011.05378.x. PubMed: 22133021.
50. Tragou E, Garrett C (1997) The shallow thermohaline circulation of the Red Sea. Deep Sea Research Part I, Oceanographic Research Papers 44: 1355-1376. doi:10.1016/S0967-0637(97)00026-5.
51. Antunes A, Ngugi DK, Stingl U (2011) Microbiology of the Red Sea (and other) deep-sea anoxic brine lakes. Environ Microbiol Rep 3: 416-433. doi:10.1111/j.1758-2229.2011.00264.x. PubMed: 23761304.
52. Antunes A, Eder W, Fareleira P, Santos H, Huber R (2003) Salinisphaera shabanensis gen. nov., sp. nov., a novel, moderately halophilic bacterium from the brine–seawater interface of the Shaban Deep, Red Sea. Extremophiles 7: 29-34. PubMed: 12579377.
53. Antunes A, Rainey FA, Wanner G, Taborda M, Pätzold J et al. (2008) A new lineage of halophilic, wall-less, contractile bacteria from a brine-filled deep of the Red Sea. J Bacteriol 190: 3580-3587. doi:10.1128/JB.01860-07. PubMed: 18326567.
54. Antunes A, Taborda M, Huber R, Moissl C, Nobre MF et al. (2008) Halorhabdus tiamatea sp. nov., a non-pigmented, extremely halophilic archaeon from a deep-sea, hypersaline anoxic basin of the Red Sea, and emended description of the genus Halorhabdus. Int J Syst Evol Microbiol 58: 215-220. doi:10.1099/ijs.0.65316-0. PubMed: 18175711.
55. Huber R, Burggraf S, Mayer T, Barns SM, Rossnagel P et al. (1995) Isolation of a hyperthermophilic archaeum predicted by in situ RNA analysis. Nature 376: 57-58. doi:10.1038/376057a0. PubMed: 7541115.
56. Huber R, Huber H, Stetter KO (2000) Towards the ecology of hyperthermophiles: biotopes, new isolation strategies and novel metabolic properties. FEMS Microbiol Rev 24: 615-623. doi:10.1111/j.1574-6976.2000.tb00562.x. PubMed: 11077154.
57. Mingorance J, Tamames J, Vicente M (2004) Genomic channeling in bacterial cell division. J Mol Recognit 17: 481-487. doi:10.1002/jmr.718. PubMed: 15362108.
58. Tamames J, González-Moreno M, Mingorance J, Valencia A, Vicente M (2001) Bringing gene order into bacterial shape. Trends in Genetics 17: 124-126. doi:10.1016/S0168-9525(00)02212-5. PubMed: 11226588.
59. Siefert JL, Fox GE (1998) Phylogenetic mapping of bacterial morphology. Microbiology-UK 144: 2803-2808. doi: 10.1099/00221287-144-10-2803. PubMed: 9802021.
60. Carmona M, Zamarro MT, Blázquez B, Durante-Rodríguez G, Juárez JF et al. (2009) Anaerobic catabolism of aromatic compounds: a genetic and genomic view. Microbiol Mol Biol Rev 73: 71-133. doi: 10.1128/MMBR.00021-08. PubMed: 19258534.
61. Valderrama JA, Durante-Rodríguez G, Blázquez B, García JL, Carmona M et al. (2012) Bacterial Degradation of Benzoate CROSS-REGULATION BETWEEN AEROBIC AND ANAEROBIC PATHWAYS. J Biol Chem 287: 10494-10508. doi:10.1074/jbc.M111.309005. PubMed: 22303008.
62. Michaelis W, Jenisch A, Richnow HH (1990) Hydrothermal Petroleum Generation in Red-Sea Sediments from the Kebrit and Shaban Deeps. Applied Geochemistry 5: 103-114. doi:10.1016/0883-2927(90)90041-3.