Conservation of the LexA repressor binding site in *Deinococcus radiodurans*

Feroz Khan[1], S. P. Singh[2], B. N. Mishra[3,*]

[1,2,3] Department of Biotechnology, Institute of Engineering & Technology,
U. P. Technical University, Sitapur Road, Lucknow-226021 (INDIA)
E-mail: [1]ferozkhan306@rediffmail.com, [2]sprakashsingh@gmail.com,
[3]biotechiet@rediffmail.com

Summary

The LexA protein is a transcriptional repressor of the bacterial SOS DNA repair system, which comprises a set of DNA repair and cellular survival genes that are induced in response to DNA damage. Its varied DNA binding motifs have been characterized and reported in the *Escherichia coli*, *Bacillus subtilis*, rhizobia family members, marine magnetotactic bacterium, *Salmonella typhimurium* and recently in *Mycobacterium tuberculosis* and this motifs information has been used in our theoretical analysis to detect its novel regulated genes in radio-resistant *Deinococcus radiodurans* genome. This bacterium showed presence of SOS-box like consensus sequence in the upstream sequences of 3166 genes with >60% motif score similarity percentage (MSSP) on both strands. Attempts to identify LexA-binding sites and the composition of the putative SOS regulon in *D. radiodurans* have been unsuccessful so far. To resolve the problem we performed theoretical analysis with modifications on reported data set of genes related to DNA repair (61 genes), stress response (145 genes) and some unusual predicted operons (21 clusters). Expression of some of the predicted SOS-box regulated operon members then was examined through the previously reported microarray data which confirm the expression of only single predicted operon i.e. DRB0143 (AAA superfamily NTPase related to 5-methylcytosine specific restriction enzyme subunit McrB) and DRB0144 (homolog of the McrC subunit of the McrBC restriction modification system). The methodology involved weight matrix construction through CONSENSUS algorithm using information of conserved upstream sequences of eight known genes including *dinB, tagC, lexA, recA, uvrB, yneA* of *B. subtilis* while *lexA* and *recA* of *D. radiodurans* through phylogenetic footprinting method and later detection of similar conserved SOS-box like LexA binding motifs through both RSAT & PoSSuMsearch programs. The resultant DNA consensus sequence had highly conserved 14 bp SOS-box like binding site.

Keywords: LexA repressor, recA protein, TFBS, *D. radiodurans* SOS-box, *B. subtilis* SOS-box, DNA repair system, genomic predictions, weight matrix, motif discovery.

1 Introduction

One of the great challenges currently facing researchers is to understand the varied and complex mechanisms that regulate gene expression and therefore dissection of regulatory networks that control gene transcription is among the primary goals of the post-genomic era of Biology. Different computational approaches were established to identify the regulatory networks through various algorithms for genomic prediction of conserved regulatory motifs [1]. A promising approach in the discovery of genomic transcriptional networks lies in combining genome wide scanning with weight matrix at different lower threshold scores as well as direct pattern matching of the known consensus sequence [2, 3]. However, weight

* Corresponding author

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

matrices were shown to be quite accurate binding affinity predictors for several transcription factors of bacterial origin in compare to direct pattern matching method [4-9]. The use of weight matrix based DNA binding motif prediction method is justified by the statistical mechanical theory [10]. In the present work, we focus on one important aspect of this challenge, the identification of highly conserved genomic transcription factor binding sites (TFBS) for LexA repressor (Swiss-Prot/ TrEMBL Acc. no. O32506; COG1974; EC:3.4.21.88; GenBank ID: AAF12438.1) involved in the regulation of genes related to SOS response for DNA repair, stress response and operons in *Deinococcus radiodurans*; a Gram-positive, red-pigmented, non-motile bacterium and most radiation resistant organism described to date. As reported exponentially growing cells are 200 times as resistant to ionizing radiation and 20 times as resistant to UV irradiation as *Escherichia coli*. It is expected that resistance may be a side effect of mechanisms that are designed to allow survival of periods of extended desiccation. The radiation resistance makes it an ideal candidate for bioremediation of sites contaminated with radiation and toxic chemicals. It has been isolated worldwide from locations rich in organic nutrients, including soil, animal feces and processed meats, as well as from dry, nutrient-poor environments, including weathered granite in a dry Antarctic valley, room dust and irradiated medical instruments. Its genome is composed of four circular molecules: chromosome I [2,648,638 base pairs (bp)], chromosome II [412,348 bp], a megaplasmid [177,466 bp] and a plasmid [45,704 bp] [11]. However, till date attempts to identify LexA-binding sites and the composition of the putative SOS regulon in *D. radiodurans* have been unsuccessful. In this paper, we combined *in silico* predictions with published data [12] and also explore cross-species comparative genomics as a tool to identify genes whose expression is controlled by a LexA repressor in response to the radiation induction & stress response. The repair of damaged DNA is crucial to cell survival and replication. It is reported that bacterial expression of a number of genes responsible for DNA repair is induced following exposure to agents which cause such damage. This coordinated regulation of many genes at different loci on the genome was first established for *E. coli* and was termed the SOS response [13-15]. The SOS response has been studied in some detail in *E. coli* [16], *Bacillus subtilis* [17] and *Mycobacterium tuberculosis* [18] and the key regulatory components have been shown to be the proteins LexA and RecA. LexA is a repressor protein which normally binds to the SOS-box sequence located in upstream sequences of the genes it regulates and so restricting expression [19, 20]. When DNA becomes damaged, regions of single-stranded DNA arise, either from processing of the damaged region or from blockage of replication [21]. RecA binds to these single-stranded regions, forming a nucleoprotein filament and in this form it stimulates the autocatalytic cleavage of LexA [22]. The cleaved fragments of LexA no longer bind to the SOS-boxes [23], thus relieving repression and leading to increased expression of the genes of the SOS regulon. The basic principles of this regulatory mechanism are found in many other species of bacteria, although the DNA sequence of the LexA binding site or SOS-box varies. Thus, while the SOS-box in *E. coli* and other enterobacteria has the consensus sequence taCTGTatatatatACAGta (bases in lowercase are less conserved than those in uppercase) [24], in rhizobia family the SOS-box is $GAAC(N)_7GTAC$ [25]; in *B. subtilis* the SOS box, originally thought to be $GAAC(N)_4GTTC$ [26, 27], has then refined as CGAACRNRYGTTCG [28] and recently a new SOS-box consensus sequence $TCGAAC(N)_4GTTCGA$ has been reported in *M. tuberculosis* [18]. It has been suggested that in *E. coli* and many other Gram-negative organisms, the SOS-box is a region of 16 bp that displays dyad symmetry, while in several Gram-positive bacteria (e.g., *B. subtilis, D. radiodurans* and *Mycobacteria* sp.), the binding site for LexA is thought to be the previously described Cheo-box, a region of 12 bp with dyad symmetry but no homology to the Gram-negative SOS-box [28]. Due to the huge differences in the TFBS between the *E. coli* and the gram-positive LexA repressor, we hypothsized conservation of the *B. subtilis* SOS-box in

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Gram-positive bacterium *D. radiodurans* instead of Gram-negative *E. coli* SOS-box. In *D. radiodurans* genome conservation of proteins is 69.85% in compare to *B. subtilis* i.e. 69.69%, although the numbers of conserved orthologs are higher in *E. coli K12* genome i.e. 79.86% and lowest in *M. tuberculosis* genome i.e. 65.83% (Table 1).

Organism	Stain type	Total proteins	No. of Proteins in COG database (orthologs)	Percentage of proteins in COG
D. radiodurans	Gram +	3187	2226	69.85
B. subtilis	Gram +	4118	2870	69.69
E. coli K12	Gram -	4275	3414	79.86
M. tuberculosis	Gram +	3927	2585	65.83

Table 1: Details of evolutionary conserved orthologs in the studied bacteria analyzed through Cluster of Orthologous Group (COG) database.

As reported in *B. subtilis*, the inducible DNA repair system (the SOS system) is regulated by two key proteins; RecA and LexA. *B. subtilis* RecA is activated by DNA damage to mediate the proteolytic cleavage of the *B. subtilis* LexA repressor, resulting in derepression of the SOS regulon. We hypothesized that SOS response in *D. radiodurans* progresses in a similar manner, with RecA having an identical role in controlling the SOS regulon together with a cellular repressor LexA protein (DRA0344) that is functionally homologous to the *B. subtilis* LexA repressor. Thus, the basic mechanism of the SOS response seems to be conserved between both *B. subtilis* and *D. radiodurans.* Besides, the protein DRA0074 is suggested to be another LexA in *D. radiodurans*, having similar protein motifs, not involved in RecA induction, but is a regulator of other metabolisms in *D. radiodurans* [12].

In the present work, we identified the genome wide LexA TFBS similar to known *B. subtilis* SOS-box. Later, for experimental validation we performed comparative study on documented genes [12] especially related to (i) SOS response, (ii) Stress response and (iii) Unusual predicted operons. Expression of some of the predicted SOS-box regulated genes then was examined through the previously reported microarray data [29]. A detailed study of the SOS-box would allow the identification of LexA TFBS experimentally in the *D. radiodurans* genome and thus aid in the discovery of other novel LexA-regulated genes. On average, Phylogenetic footprinting along with PSSMs improved the selectivity of TFBS prediction by ~90% which detects most of the known sites in comparison to use weight matrix method (Table S13) or direct pattern matching method (Table S14, S15) alone. All the TFBS detected through RSAT webserver [32] were found similar to the results of PoSSuMsearch program (a Linux plateform based standalone software) [43], thus the results were theoretically validated.

2 Material and Method

We constructed a SOS-box weight matrix based on *B. subtilis* known LexA binding sites (Table 2) [30]. Later it was used to predict novel LexA binding sites in the *D. radiodurans* genome. The weight matrix was also employed to predict LexA binding motifs across closely related bacterial genomes with a goal to identify a common set of LexA regulated genes expected to be conserved.

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Operon	Regulated Gene	Sigma	Regu-lation	Absolute position	Location	Known LexA binding site (*cis*-element)	[Ref. No.] Reference	Year
dinB	dinB	None	-ve	608351. 608364	-26:-13	AGAACTCATGTTCG	[28] Winterling K W, *et al.*	1998
tagC	tagC	None	-ve	3682370. 3682383	-53:-40	AGAACAAGTGTTCT	[28] Winterling K W, *et al.*	1998
tagC	tagC	None	-ve	3682390. 3682423	-34:-1	TATTGAATACCGAACGT ATGTTTGCTTTAATGTA	[28] Winterling K W, *et al.*	1998
lexA	lexA	None	-ve	1917524. 1917537	-39:-26	CGAACCTATGTTTG	[28] Winterling K W, *et al.*	1998
lexA	lexA	None	-ve	1917553. 1917566	-67:-54	CGAACAAACGTTTC	[28] Winterling K W, *et al.*	1998
lexA	lexA	None	-ve	1917590. 1917603	-104:-91	GGAATGTTTGTTCG	[28] Winterling K W, *et al.*	1998
recA	recA	None	-ve	1763831. 1763844	-51:-38	CGAATATGCGTTCG	[28] Winterling K W, *et al.*	1998
recA	recA	None	-ve	1763832. 1763843	None	GAATATGCGTTC	[40] Hamoen LW, *et al.*	2001
uvrBA	uvrB	None	-ve	3614039. 3614052	-55:-44	CGAACTTTAGTTCG	[28] Winterling K W, *et al.*	1998
yneAB-ynzC	yneA	None	-ve	1917553. 1917566	None	GAAACGTTTGTTCG	[41] Kawai Y., *et al.*	2003
yneAB-ynzC	yneA	None	-ve	1917589. 1917603	None	GCGAACAAACATTCC	[41] Kawai Y., *et al.*	2003

Table 2: List of the known LexA TFBS characterized in *B. subtilis* genome retrieved through DBTBS; a database of transcriptional regulation in *B. subtilis*. The SOS-box consensus sequence of *B. subtilis* is CGAACRNRYGTTCG for binding of LexA, defined as negative (-ve) regulator of DNA damage-inducible genes; SOS-like or SOB regulon; analogue of *E. coli* LexA.

(i) Construction of the SOS-box weight matrix

A total of eight gene's upstream sequences from both *B. subtilis* and *D. radiodurans* were used for weight matrix construction and considered as example data set (Table S1 & Table S2). Of those eight, six known promoters with known SOS-box were from *B. subtilis* genes namely, *dinB, tagC, LexA, recA, uvrB & yneA* and two upstream sequences with SOS-box were from *D. radiodurans* genes namely, *DRA0344* (*lexA) & recA*. Following their alignment a matrix was constructed from the relative frequencies of A, T, C or G at each position of the 14 nucleotides (nt) SOS-box motif sequence. This matrix was used to determine an information-based measure of potential binding sites according to the method of Schneider *et al.* (1986) [31]. A 14 nt motif region was moved over the entire genome on both strands and the score (S_i) at each nucleotide position (having base i) was calculated according to following equation:

$$S_i = (1/14)\ \Sigma_j\ [2 + \log2(F_{ij})]$$

Where F_{ij} is the frequency matrix for base *i* at position *j*.

This score, which ranges from −30.761 (the score of the worse match) to 15.199 (the score of the best or exact match with known *B. subtilis* SOS-box consensus sequence), is a measure of the information content of a potential binding site measured against the example data set (Table S3 & Table 3). The lowest example score, that of *DRA0344* (*lexA)*, is 9.79 (88.23%) and thus, a threshold of 9.78 (88.21%) was used to define a 'Genome wide good hits'. On the other hand for our studied data set of genes related to repair mechanism, stress response and unusual predicted operons a threshold or lower cut off score '0' (i.e. 66.93%) was used to avoid filtration of true positives and therefore selected for further SOS-box conservation study. A scan of the entire *D. radiodurans* genome produced about 3166 motif hits on both strands [refers as Direct (D) and Reverse strand (R)]. These motif hits were again filtered to retain only those that were in between −400 to -1 bp on the both strands from an annotated

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

translational start site and having weight score more than & equal to 9.78 (88.21%) using CONSENSUS & PATSER algorithms [32].

Gene's upstream	Strand	Start	End	SOS-box[1] (14 nt)	Score	MSSP[2]	ln(P) value	RMMP[3] (%)
tagC_bsu	D	-68	-55	CGAACGTATGTTTG	13.91	97.19%	-16.89	91.52
lexA_bsu	D	-72	-59	CGAACCTATGTTTG	13.53	96.37%	-16.31	89.02
yneA_bsu	R	-91	-78	CAAACATAGGTTCG	13.53	96.37%	-16.31	89.02
recA_bsu	R	-86	-73	CGAATATGCGTTCG	13.08	95.39%	-15.68	86.06
dinB_bsu	D	-55	-42	AGAACTCATGTTCG	12.86	94.91%	-15.44	84.61
uvrB_bsu	R	-92	-79	CGAACTTTAGTTCG	12.41	93.93%	-14.78	81.65
recA_dra	R	-240	-227	CGACCTCGCGTTCA	10.35	89.45%	-12.47	68.10
lexA_dra	D	-146	-133	CGAACTCACGGAAG	9.79	88.23%	-11.89	64.41

Note:

[1] Predicted SOS-box similar to known *B. subtilis* (refer by 'bsu') binding sites.

[2] MSSP (Motif Score Similarity Percentage) = Similarity of predicted motif score with known SOS-box consensus sequence (CGAACRNRYGTTCG) score. Maximum score of 15.1991 is for exact similarity and minimum score of -30.761 is for poor or worst similarity. Percentage is calculated out of 'Range of scores' i.e. [15.199 - (-30.761)] = 45.960.

[3]RMMP (Relative Motif Matching Percentage) = Nucleotides matching of predicted motif with known SOS-box nucleotides (CGAACRNRYGTTCG). Percentage is calculated out of maximum score i.e. 15.1991. Where, R = A or G (puRines), Y = C or T (pYrimidines) & N = G, A, C or T (aNy) as per Ambiguous nucleotide codes of the IUPAC-IUB commission (http://www.chem.qmw.ac.uk/iupac/misc/naseq.html).

Table 3: Details of SOS-box elements detected in the example dataset using weight matrix method.

(ii) Weight Matrix based Genomic SOS-box prediction

To identify regulated genes, the known example data set was first analyzed and validated with high-scoring matches to the known SOS-box by using frequency matrix based pattern matching tool 'PATSER' implemented at RSAT web server [32]. We considered one nucleotide variation in conserved pattern and set the search parameter at '1' substitution level instead of '0', because at 'zero' substitution level only limited genes were resulted while at '1' mismatch the probability of mutation by one nucleotide is expected. In this method, a position specific frequency matrix representing a consensus sequence was converted to a positional weight matrix or position specific scoring matrix (PSSM), which was later used to score the motif sequence according to the scoring system [33]. Finally we observed significant conservation in non-coding upstream sequences and then used this conservation to improve the LexA TFBS predictions.

(iii) Detection of LexA orthologs

The evidence of LexA TF potential orthologs along with conserved protein motifs or domains were detected with the help of BLASTp [34] and MicroBial Genome Database (MBGD) [35] web resources. Later similar conserved protein domains were also detected through CDD search program [36]. Moreover, detected orthologs of LexA protein were multiple aligned and phylogenetic unrooted tree was constructed as an evidence of amino acid residues conservation, which revealed evolutionary relationship among studied bacteria. Multiple sequence alignment was performed through distance method based tool ClustalW at EBI [37] and the phylogenetic tree was generated by DRAWTREE program of phylip package [38]. Moreover, detected orthologs were again verified & validated through COG database [39].

(iv) Calculations of statistical parameters

To measure the prediction accuracy of newly derived SOS-box matrix, predicted motifs were analyzed in terms of MSSP (Motif Score Similarity Percentage) and RMMP (Relative Motif Matching Percentage). MSSP means similarity of predicted motif score with known SOS-box

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

consensus sequence (CGAACRNRYGTTCG) score in terms of percentage. For RSAT program based predictions, maximum score of 15.1991 revealed best or exact similarity while minimum score of -30.761 for poor or worst similarity. Percentage was calculated out of 'range of scores' i.e. 45.960. For PoSSuMsearch program based analysis, predicted range of score was 0 to 89. Here maximum score indicates exact similarity with known motif while minimum score indicates poor similarity. On the hand RMMP means nucleotides matching of predicted motif with known SOS-box nucleotides (i.e. CGAACRNRYGTTCG). Percentage is calculated out of maximum score i.e. 15.1991 for RSAT program. For PoSSuMsearch program, we computed 'matrix similarity score' (MSS). These MSS scores rescaled to the interval [0, 1] with the minimum reachable PSSM score corresponding to 0 and the maximum reachable PSSM score corresponding to 1. Note that because PSSM thresholds can be derived from similarity without use of probability distributions, they will not be calculated by default. Beside this, PSSM score based predictions were further evaluated by probability value (P-value) and expected number of matches by chance through E-value. A P-value for PSSM based scoring was computed from the score distribution obtained with the weight matrix applied to 1000 randomized sequences with the same length and AT content as the original sequence. On the other hands E-value can be define as the expected number of matches in a given random sequence database. It is widely accepted measure of the significance. Its calculation is based on P-values, it is simply the P-value times database size.

The MSSP was calculated as:

For known LexA TFBS,

Maximum score (S_{max}) = 15.199

Minimum score (S_{min}) = - 30.761

Range of scores (R_{known}) = Maximum score (S_{max}) – Minimum score (S_{min})

= 15.199 - (- 30.761) = 45.960

Range of predicted motif score (Observed score) = Predicted score of motif – Minimum score

$R_{pred} = S_{pred} - S_{min}$

Motif Score Similarity Percentage = (Range of predicted motif score / Range of score) x 100

(MSSP) $P_{\%} = (R_{pred} / R_{known}) \times 100$

The RMMP was calculated as:

Relative Motif Matching Percentage= (Predicted score of motif/ Maximum score) x 100%

(RMMP) $P_{\%} = (S_{pred} / S_{max}) \times 100$

(v) Fast and Sensitive Matching of Matrix using PoSSuMsearch

Using PSSMs for binding motif scoring is a hard problem in Bioinformatics, especially the calculation of p-values is a big problem since the score distribution of a matrix has to be calculated. For the paper 1000 random sequences have been used. For comparative analysis of regularization, p-value calculation, calculation of log-odd-scores etc. we used another program namely, PoSSuMsearch [42, 43], which amongst others provides sound statistics for exact and lazy P-value calculation, instead of using an approximation. To verify & validate the results we evaluated our data set through lookahead scoring method (LAsearch), enhanced suffix arrays method (ESAsearch) and find the appropriate threshold for PSSM searching (LazyDistrib) through PoSSuMsearch program.

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

3 Results and Discussion

Employing weight matrix based genome wide detection of TFBS method, we identified total 3166 genes having SOS-box consensus sequence conservation in their upstream sequences (-1 to -400 bp) at lower cutoff score 'zero' (i.e. 66.93%) on both strands. Total 61 genes coding for replication, repair and recombination (Table 4 & Table S4) and an additional 145 stress response related genes in *D. radiodurans* (some of which were in operons) (Table 5 & Table S7) showed the evidence of varied range of conservation of SOS-box motif on both strands and thus hypothesized to be regulated by LexA. On the other hand stress response related genes with SOS-box showed broad spectrum of proteins that have been associated with various forms of stress responses in other bacteria as well as several proteins that appear to be unique and could contribute to more specific forms of the stress response. Orthologs of almost all known genes involved in different stress responses in other bacteria were present in *D. radiodurans* and showed varied range of SOS-box conservation.

Of the 76 genes coding for replication, repair and recombination identified by Makarova *et al.* (2001) [12], 61 were also detected with our matrix based genomic scanning. However, genes namely, DR2285 (A/G- specific adenine glycosylase), DR0715 (G/U mismatch- specific DNA glycosylase and DRA0074 (hypothetical protein or also predicted as transcriptional regulator, repressor of the SOS regulon autoprotease), were more highly expressed in GY10912 than R1 strain of *D. radiodurans* under normal, unstressed conditions, which revealed that these genes in R1 showing LexA dependent transcription under microarray studies just because of regulation by LexA. Of the 15 genes not identified by our weight matrix, 13 were unusual as they detected by BLAST analysis and were homologs of *B. subtilis dinB* gene encoding uncharacterized family of presumably metal dependent enzymes) and other two were DR1663 (Uracil DNA glycosylase) and DR1819 (UV endonuclease; activity was characterized in Neurospora).

In the LexA regulated operon members predicted here (Table S10), a gene cluster of DRB0143 (AAA superfamily NTPase related to 5-methylcytosine-specific restriction enzyme subunit McrB) and DRB0144 (homolog of the McrC subunit of the McrBC restriction modification system) showed evidence of LexA regulated expression in microarray experiment in response to ionizing radiation [29]. The failure to detect repression of remaining genes expression in microarray experiments is not surprising as the microarray experiments also failed to detect genes in operon, known to be repressed by LexA. The genes namely, recA (DR2340; recombinase; single stranded DNA dependent ATPase, activator of LexA autoproteolysis), LexA (DRA0344; Transcriptional regulator, repressor of the SOS regulon, autoprotease), DR1262 (Ro RNA binding protein; ribonucleoproteins complexed with several small RNA molecules; involved in UV resistance in *Deinococcus*), haml/yggV (DR0179; Xanthosine triphosphate pyrophosphatase, prevents 6-N-hydroxylaminopurine mutagenesis), polA (DR1707; DNA polymerase I), mrr (DR0508, DR0587; MRR-like nuclease; restrictase of the recB archaeal Holliday junction resolvase superfamily), mutS (DR1039; ATPase), recF (DR1089; predicted ATPase; required for daughter strand gap repair), dnlJ (DR2069; DNA ligase), mutS (DR1976; ATPase) etc. appears to be highly regulated upon radiation response and showed SOS-box motif conservation of 89.45% to 80.40% through RSAT webserver program (Table S5, S8, S11).

To verify & validate the RSAT prediction results, we further theoretically analyzed the similar data set through PoSSuMsearch program [43]; a Linux plateform based standalone software meant for weight matrix based motif prediction, which amongst others provides sound statistics for exact and lazy P-value calculation, instead of using an approximation. Through PoSSuMsearch we found similar SOS-box binding motifs as predicted by RSAT webserver under different statistical parameters (Table S6, S9, S12). The prediction

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

performances of both the programs were evaluated using the same PSSM. We found that both the programs predicted similar TFBS but with different statistical scoring parameters (Table 4, 5). The comparative performances of both RSAT & PoSSuMsearch programs are graphically represented in Figure 1 & 2.

In our study we identified conservation of LexA binding sites in all the typical bacterial genes that comprise the basal DNA replication machinery in *D. radiodurans*. The repertoire of DNA-associated proteins in *Deinococcus* is similar to that in other bacteria. Bacterial DNA repair includes several partially redundant pathways and generally shows considerable flexibility [12]. We predicted the putative LexA binding sites in upstream sequences of repair system components of *D. radiodurans*, to detect any possible correlation with its exceptional radioresistant and desiccation-resistant phenotype. Generally, it appears that Deinococcus possesses a typical bacterial system for DNA repair. Studied data sets were: (i) genes related to replication, repair & recombination, (ii) genes related to stress responses and (iii) genes related to unusual predicted operons.

(i) Evidence of SOS-box conservation in genes related to repair mechanism

In this category total 61 genes have been studied and showed 89.45% to 67.41% range of MSSP through RSAT webserver programs. The gene recA coding recombinaseA or Recombinase; single-stranded DNA-dependent ATPase, activator of *lexA* autoproteolysis (or DR2340) showed highest score (10.35) with 89.45% MSSP. It belongs to recombinational repair (RER) and SOS repair (SOS) pathways and showed phylogenetic evidence in all studied bacteria. While gene DR1757 encoding for hypothetical protein or predicted nuclease and zinc finger domain-containing protein; an ortholog is present in *Pseudomonas aeruginosa* showed minimum score of 0.22 and 67.41% MSSP. It was predicted to be member of unknown possible repair pathways and showed phylogenetical pattern in all studied bacteria (Table S4, S5, S6).

The nucleotide excision repair (NER) system that consists of the UvrABC excinuclease and the UvrD and Mfd (transcription-repair coupling factor) helicases showed 79.48 to 70.56% conservation of LexA binding site. The main components of the base excision repair (BER) system including several nucleotide glycosylases and endonucleases, namely, MutM (formamidopyrimidine and 8-oxoguanine DNA glycosylase); MutY (8-oxoguanine DNA glycosylase and apurinic DNA endonuclease-lyase); two paralogous uracil DNA glycosylases (Ung homologs); an additional, recently identified enzyme that has the same activity but is unrelated to Ung (DR1751); endonucleases III (Nth) and V (YjaF); and exonuclease III (XthA) showed 86.14 to 70.06% conservation of LexA TF binding site in their gene upstream sequences. The repertoire of recombinational repair (RER) genes in *Deinococcus* includes orthologs of most of the *E. coli* genes involved in this process namely recF (predicted ATPase, required for daughter strand gap repair); ruvC (endonuclease subunit of the RuvABC); ruvA (holliday junction binding subunit of the RuvABC resolyasome); ruvB (helicase subunit of the RuvABC resolyasome); recQ (helicase, suppressor of illegitimate recombination); recR (required for daughter strand gap repair); recD (helicase/exonuclease); recJ (nuclease); sbcC (exonuclease subunit, predicted ATPase); recN (predicted ATPase); sbcD (exonuclease); recO (required for daughter strand gap repair); recG (holliday junction specific DNA helicase) and recA (recombinase, single stranded DNA-dependent ATPase, activator of LexA autoproteolysis) showed 89.45 to 67.65% conservation of LexA TF binding site, but the RecBCD recombinase was missing. While this complex is not universal in bacteria, it is a major component of recombination systems in most free-living species. In *Deinococcus*, where recombination is thought to be an important contributor to damage-resistance, the absence of this ATP-dependent exonuclease is unexpected. *Deinococcus* does encode an apparent ortholog of one of the helicase-related subunits of this complex, RecD

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

with 76.20% conservation of LexA binding site, but not the other subunits. The methylation-dependent mismatch repair system (mMM) of *D. radiodurans* includes the MutS and MutL ATPases and endonuclease VII (XseA) showed 82.18 to 72.45% conservation of LexA TF binding site. It is reported that orthologs of the site-specific methylases Dcm and Dam, associated with mismatch repair, are not readily detectable and thus appears likely, that other distantly related DNA methylases predicted in *D. radiodurans* could perform similar functions. *D. radiodurans* encodes the LexA repressor-autoprotease (DRA0344), which in *E. coli* and *B. subtilis* controls the expression of the SOS regulon. In addition, unlike any of the other bacterial genomes studied, *D. radiodurans* encodes a second, diverged copy of LexA (DRA0074), which retains the same arrangement of the helix-turn-helix DNA-binding domain and the autoprotease domain.

S.No.	Pathway	No. of Genes	SOS-box motif conservation (%)			
			RSAT		PoSSuMsearch	
			Max.	Min.	Max.	Min.
1.	BER	8	86.14	70.1	76.4	52.81
2.	DR	3	83.49	72.65	70.79	60.67
3.	MM	1	72.45	-	59.55	-
4.	mMM	1	76.04	-	77.53	-
5.	MP	4	81.22	70.65	70.79	52.81
6.	NER	5	78.22	70.56	69.66	53.93
7.	RER	13	81.27	67.65	76.4	47.19
8.	SOS	1	88.23	-	80.9	-
9.	VSP	1	73.69	-	65.17	-
10.	mMM, VSP	2	80.4	78.98	75.28	69.66
11.	BER, DR	1	70.69	-	56.18	-
12.	BER, MMY	1	70.8	-	66.29	-
13.	BER, NER	1	76.65	-	60.67	-
14.	mMM, SOS, NER	1	72.15	-	60.67	-
15.	RER, SOS	1	89.45	-	70.79	-
16.	Unknown	17	83.53	67.41	80.9	56.18
TOTAL =		61				

Table 4: Comparison of SOS-box motif conservation (in %) predicted through both RSAT & PoSSuMsearch programs in *D. radiodurans* genes related to replication, repair and recombination functions.

Through PoSSuMsearch similar SOS-box motifs were predicted. Similarity to known motifs ranges from 80.9% to 47.19% of MSSP. The genes namely, DRA0344 belongs to SOS repair pathways & DR0508 belongs to unknown (?) pathway showed highest score (i.e.72) with 80.9% MSSP. While gene DR1902 (recD) encoding for exodeoxyribonuclease V, subunit RecD, putative belongs to RER pathway showed minimum score of 42 and 47.19% MSSP (Table S6). Detection of similar binding sites through both RSAT and PoSSuMsearch programs indicates localization of conserved SOS-box like motifs in the studied data set sequences, this also revealed conservation of specific regulatory phylogenetic patterns for each metabolic pathway as a sign of phylogenetic footprint in the studied microbial genomes. All the studied genes were arranged systematically under different pathways categories such as Unknown possible repair pathways, Base Excision Repair (BER) pathway, Direct damage Reversal (DR) pathway, Methylation-dependent Mismatch repair (MM) pathway, Multiple Pathways (MP) pathway, Nucleotide Excision Repair (NER) pathway, Recombinational

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Repair (RER) pathway, SOS repair (SOS) pathway and Very-Short-Patch mismatch repair (VSP) pathway (Table S5, S6).

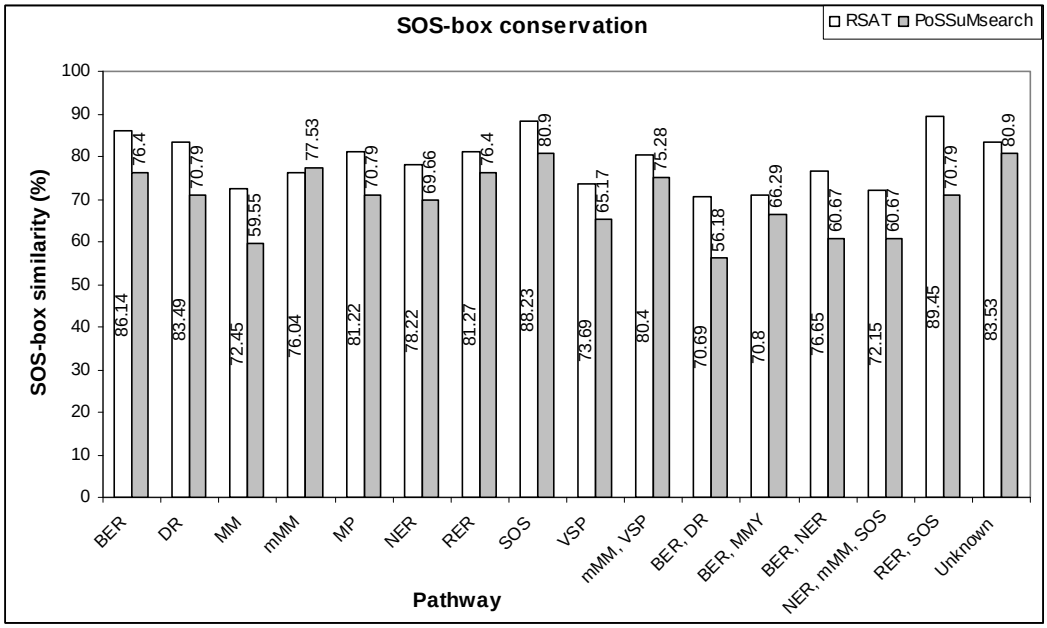

Figure 1: Graphical representation of comparative SOS-box motif conservation (only max. %) predicted through both RSAT & PoSSuMsearch programs in *D. radiodurans* genes related to replication, repair and recombination functions.

(ii) Evidence of SOS-box conservation in the stress response system

The studies of Makarova *et al.*, (2001) [12] revealed that *D. radiodurans* encodes a broad spectrum of proteins that have been associated with various forms of stress response in other bacteria as well as several proteins that appear to be unique and could contribute to more specific forms of the stress response (Table S7, S8, S9). The orthologs of all known genes involved in different stress responses in other bacteria were also found in Deinococcus genus too. On the basis of predicted results of both RSAT and PoSSuMsearch programs, all the studied genes were classified under 16 types of stress response systems and summarized in Table S8 & S9. Results showed that most of the stress response related genes have varied but higher level of SOS-box motifs conservation in their upstream sequences and therefore expected to be a integral component of LexA regulatory network. However, more studies are still required to analyze experimentally the regulation of these genes and the nature of their associated SOS-box motifs. Through PoSSuMsearch program similar SOS-box motif were predicted. Similarity to known motifs ranges from 87.64% to 43.82% of MSSP. The gene namely, DRB0131 showed highest score of 78 with 87.64% MSSP, while gene DR1187 showed minimum score of 39 with 43.82% MSSP (Table S9). Both the genes belong to (terC of Alcaligenes genus) function unknown but encoding for membrane protein belongs to Detoxication stress category. Detection of similar binding sites through both RSAT and PoSSuMsearch programs indicates conservation of SOS-box motif in these sequences. This also indicates conservation of specific regulatory phylogenetic patterns for each metabolic pathway as phylogenetic footprint. This could be a subject of further experimental validation. Besides, on the basis of conservation within upstream non-coding sequences we predicted the missing phylogenetic patterns in unknown category genes (?) by assuming the conservation of phylogenetic pattern of those genes which showed closely relationship in terms of TFBS and MSSP percentage with neighboring genes of the same cluster and on this basis we hypothesized unknown category genes under respective stress type or pathway.

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

S.No.	Stress type	No. of Genes	SOS-box motif conservation (%)			
			RSAT		PoSSuMsearch	
			Max.	Min.	Max.	Min.
1.	Heat/general	8	84.03	69.45	74.16	60.67
2.	Heat	2	79.83	71.02	67.42	55.06
3.	General	35	85.44	67.84	78.65	46.07
4.	Starvation	6	82.53	72.24	78.65	57.3
5.	Osmotic	11	84.88	71.48	78.65	48.31
6.	Phage	1	73.04	-	60.67	-
7.	Alkaline	2	84.44	81.53	75.28	-
8.	Cold	1	77.81	-	56.18	-
9.	Oxidative	13	86.64	67.45	79.78	50.56
10.	Oxidative/detoxication	8	80.09	70.58	69.66	56.18
11.	Detoxication	11	88.79	69.02	87.64	43.82
12.	Toxins/general	4	85.84	77.48	74.16	51.69
13.	Toxins	2	79.59	72.17	74.16	67.42
14.	Desiccation	4	74.96	70.67	66.29	58.43
15.	Drugs	32	85.68	68.62	83.15	51.69
16.	Unknown	5	80.37	69.56	76.4	61.8
TOTAL =		145				

Table 5: Comparison of SOS-box motif conservation (in %) predicted through both RSAT & PoSSuMsearch programs in *D. radiodurans* genes related to different stress types.

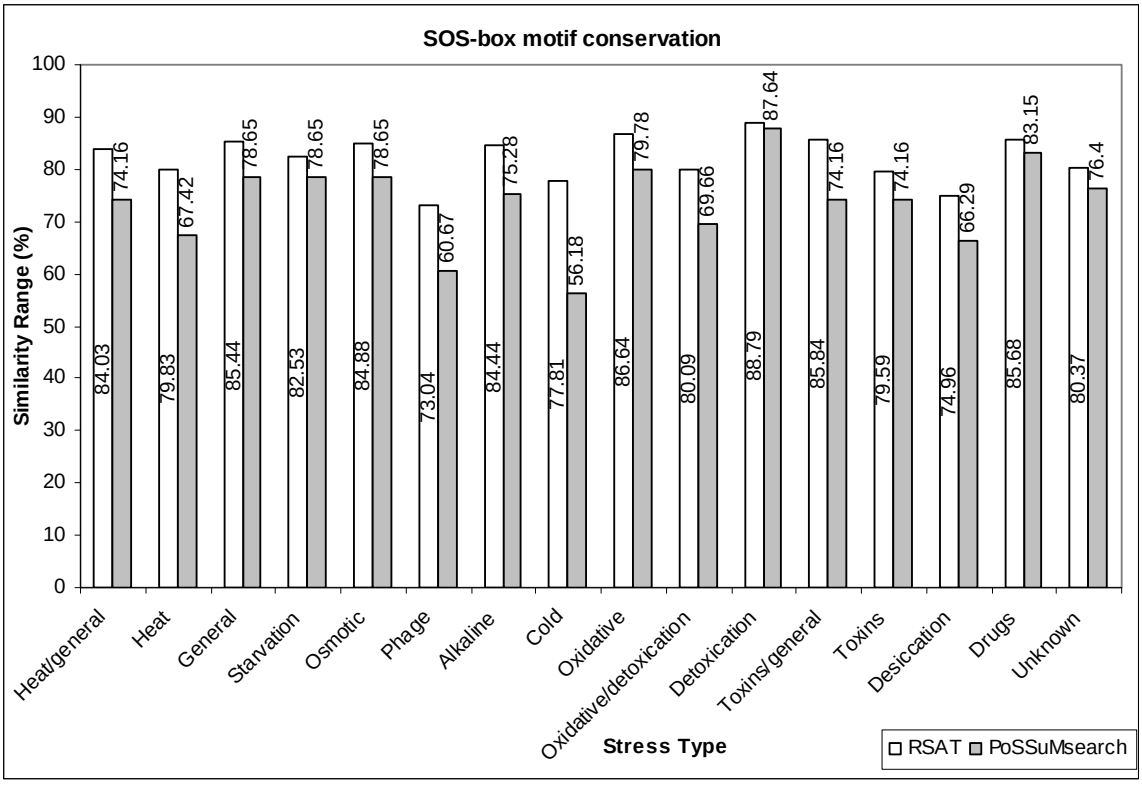

Figure 2: Graphical representation of comparative SOS-box motif conservation (only max. %) predicted through both RSAT & PoSSuMsearch programs in *D. radiodurans* genes related to different stress types.

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

(iii) Evidence of SOS-box motif conservation in the genes of unusual operons

Generally, the genome organization of *D. radiodurans* is similar to that of other bacteria and on the basis of unusual operons [12] many functionally related genes were organized into clusters that were likely to comprise operons e.g., ribosomal protein genes, ATP synthase, NADH dehydrogenase and various ATP-binding cassette (ABC)-type transport systems. Beyond these generic operons, however, several unusual gene clusters were detected showing conservation of *B. subtilis* known SOS-box consensus motif sequence and thus expected to be regulated by LexA *trans*-factor. Some of these were likely to be related to the unique features of Deinococcus. In the present work we studied 21 unusual predicted operons for LexA TFBS which revealed 68.73 to 87.05% conservation of SOS-box consensus sequence in their gene's upstream through RSAT webserver (Table S10, S11).

The first group of such unique gene clusters includes paralogous genes that encode protein families overrepresented in Deinococcus, such as amino-acetyltransferases, Nudix hydrolases and genes of the TerE and DinB/YfiT families. Some of these clusters reported to be evolved by tandem duplication within the Deinococcus lineage, e.g., an acetyltransferase cluster includes genes DR2254 and DR2255 with SOS-box conservation of 74.31% and 68.73% respectively. Other clusters of paralogs clearly resulted from a single horizontal transfer event e.g., the group of tellurium resistance genes *viz*., DR2220 to DR2226 which showed conservation of 72.65% to 85.84%, related to the corresponding gene cluster on the broad-host-range plasmid R478. Finally, some clusters that consist of related genes with apparent phylogenetic affinities to different bacterial lineages (e.g., an acetyltransferase cluster [DR0675 to DR0677]) seem to have originated within the Deinococcus lineage through gene translocation and showed SOS-box conservation of 74.13 to 79.70%.

The second group of unusual predicted operons includes rare gene clusters that probably were acquired by horizontal transfer. Some of these operons could contribute to damage resistance e.g., DNA repair-related functions (deoxypurine kinase operon [genes *viz*., DR0298 and DR0299 with SOS-box conservation of 80.68 and 74.30%]), eukaryotic-type uracil-DNA-glycosylase and topoisomerase IB [genes *viz*., DR0689 and DR0690 with conservation of 73.78% and 76.44%], restriction-modification system [genes *viz*., DRB0143 and DRB0144 with conservation of 76.39% and 70.54%]. However these genes were also detected in the microarray expression data under radiation response, stress response (genes *viz*., DR0398 and DR0390 with SOS-box conservation of 86.34% and 84.44%) and pigment biosynthesis (genes *viz*., DR0861 and DR0862 with SOS-box conservation of 74.02% and 73.46%).

Two operons (genes *viz*., DR0853 to DR0854 with conservation of 84.77% and 70.82% and DR2180 to DR2181 with SOS-box conservation of 76.26% and 74.41%) each consist of a gene for a small GTPase of the Ras/Rab family and a gene coding for a small protein of an uncharacterized family. The orthologous GTPase in Myxococcus was important for gliding motility [21], suggesting a role for these proteins in signaling. Expansion of the uncharacterized protein family encoded by the genes adjacent to the GTPase was seen in Streptomyces and Deinococcus and appears to result from relatively recent duplications (genes *viz*., DR0616, DR0995, and DR1612), with three of these genes forming a cluster in the chromosome (genes *viz*., DR0993 to DR0995 with SOS-box conservation of 73.04 to 72.13%). Juxtaposition of these genes with genes for Ras/Rab-GTPases is frequently observed in other genomes, including Myxococcus and archaeal and bacterial thermophiles, suggesting that they form a mobile operon, with the encoded proteins being functionally coupled.

On the other hand, PoSSuMsearch program also predicted the similar SOS-box motifs in all the operon clusters. Similarity to known motif ranges from 85.39% to 44.94% of MSSP. The gene namely, DRA0232 encoding Flavoprotein dehydrogenase of operon cluster number 19 showed highest score of 76 with 85.39% MSSP, while gene DR1235 encoding Dynamin-like

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

GTPase of operon cluster number 12 showed minimum score of 40 with 44.94% MSSP (Table S12). Highest conservation of SOS-box motif in each cluster indicates existence of regulatory network control within each operon, thus our theoretical analysis revealed existence of regulated operons or regulon in *D. radiodurans* genome, which could be a subject of further experimental validation. Here prediction of similar motif through both RSAT and PoSSuMsearch programs indicates conservation of SOS-box motifs across operonic clusters. This also indicates localization of regulatory evolutionary conserved patterns for each operon as a phylogenetic footprint.

(iv) Genome wide detection of SOS-box motifs through weight matrix

For comparison purpose, we used genome wide scanning method through newly derived matrix (or PSSM) alone at lower cut off score of 9.78 (i.e. 88.21% MSSP), only 33 genes were identified as 'genome wide top hits' which are summarized in Table S13. Mostly all the predicted genes were members of cellular metabolisms other than repair pathway & stress response. These hits showed maximum level of SOS-box conservation of 88.23% to 96.78%. However, genes of DR1083 (hypothetical protein) and DR1084 (methylmalonyl-CoA mutase) were found as a member of predicted operon while genes of DR0943 (hypothetical protein) and DRB0131 (hypothetical protein) were found as a member of two different predicted operons. Interestingly, genes of both recA (recombinase A) and DR0970 (electron transfer flavoprotein, alpha subunit) were also found common with expressed genes under radiation response in earlier microarray studies [29]. However, when compared our predicted results with expression known data, most of the genes were found missing, which suggested that these were either software predicted false positives or due to some experimental error. However, due to higher level of SOS-box conservation it is assumed that these genes could be either induced or repressed under radiation response. Similarly further studies are required to analyze the regulation of these genes and the nature of their associated SOS-box sequences.

(v) Genome wide detection of SOS-box motifs through direct pattern matching

For comparison purpose, we lastly used genomic direct pattern matching method alone in parallel to weight matrix method & combination of phylogenetic footprinting cum weight matrix method. Genome wide detection of known *B. subtilis* SOS-box consensus sequence (CGAACRNRYGTTCG) in the upstream sequences (-1 to -400 bp) of *B. subtilis* genome through direct pattern matching program of RSAT webserver extracted only limited set of genes i.e. 16 genes with score 0.93 and RMMP value of 93% at one nucleotide mismatch. Percentage is calculated out of maximum score 1, which indicates 100% or exact match (Table S14). Similarly, we studied genome wide detection of known SOS-box motifs in the upstream sequences (-400 bp) of *D. radiodurans* genome, we found that only limited set of genes were extracted i.e. 18 genes with score 0.93 and RMMP 93% while 2 genes scored '1' or 100% RMMP *viz*., DR1083 & DR1084 at one nucleotide mismatch (Table S15). Since most of the regulated known genes of *B. subtilis* (as mentioned in Table 2) are missing and thus expecting the same in *D. radiodurans* too, therefore it is now evident that combination of phylogenetic footprinting cum weight matrix (PSSM) method is better prediction approach than others. Al last through this theoretical analysis we concluded that genome wide direct pattern matching method as well as weight matrix method alone is not appropriate for *in silico* TFBS prediction.

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

(vi) Evidence of LexA ortholog in *D. radiodurans*

We collected experimental data of LexA transcription factor for which binding sites and their SOS-regulon has been already determined in *B. subtilis* [30]. LexA repressor is a DNA-binding transcriptional regulator for radiation response. Conserved domain analysis showed a characteristic similar type of protein domains or protein motif information *viz*., COG1974; defined as LexA repressor or SOS-response transcriptional repressors (or RecA-mediated autopeptidases or Transcription/Signal transduction mechanisms) and Pfam00717; Peptidase_S24 (Peptidase family S24) and COG2932; Predicted transcriptional regulator [Transcription] (Figure 3). Multiple sequence alignment of LexA orthologs showed conservation of mostly functionally important amino acid residues. However, in the cases where paralogous LexA proteins from the same species were present, the protein with the highest BLASTp scores was selected as potential orthologs. This was further verified & validated through results of MBGD and COG [39] database search.

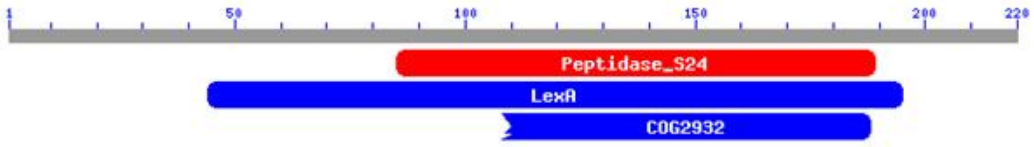

Figure 3: Diagrammatic illustration of conserved LexA transcriptional regulatory protein domains in *D. radiodurans*. Domain belongs to Peptidase_S24 family and showed protein sequence sharing with both LexA and COG2932 family.

4 Conclusion

The power of any TFBS prediction algorithm that uses PSSM depends on the quality of the matrix models that it uses, since the matrices represent an abstraction of experimentally verified TFBS. In the studied work comparative and phylogenetic filtering procedure significantly improves the power of TFBS prediction, as is demonstrated in the paper. The authors compared the LexA promoter sequences with the orthologous sequences from other bacteria; this dramatically reduced the false-positive prediction of TFBS and able to identify successfully a previously documented regulatory site namely SOS-box in *D. radiodurans*. On average, phylogenetic footprinting along with PSSM improved the selectivity of TFBS prediction by ~90%, compared to weight matrix and direct pattern matching method alone. In the studied research work comparative analysis across bacterial genomes revealed indication of conserved LexA regulated DNA binding motifs especially in the genes involved in repair mechanism, stress response & unusual operons. This analysis revealed that SOS-box is more conserved to evolutionary closely related Gram positive bacterium *B. subtilis* than Gram negative bacterium *E. coli*. However, the improvement of weight matrix quality could be beneficial for studying the variety of transcription factors regulation scenarios. An alternative way of understanding such scenarios is a strict molecular modeling of the regulatory networks processes. Our approach allows for significant improvement of existing matrices, thereby presenting an alternative to molecular modeling of the regulatory networks at genomic level. Moreover, we expect that understanding of the regulatory mechanisms related to repair, stress response and operons control, will increases by further comparative genome analysis and prediction driven experiments. The present genomic study yielded many functional predictions that can be tested experimentally and could prove particularly significant in near future. It is becoming evident that comparative genome analysis is very powerful and will be of use not only for genome annotation but also as an adjunct to more traditional disciplines, such as

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

molecular biology and genetics. Just like the sequence-alignment programs that emerged in the early 1990s, TFBS prediction tools will also prove very valuable and timely research tools for the scientific community. Many new research directions are currently being pursued in this area; for example, pairwise sequence comparisons can be expanded to include multiple species and to make use of additional information, such as evolutionary distance and phylogenetic relationships. It can be imagined that, with the emergence of more bacterial genomes in the near future, we can finally identify all the gene regulatory elements in other radiation digestive bacteria and use them as a blueprint for understanding the mysteries of gene regulation.

5 Acknowledgement

We acknowledge the 'Council of Scientific & Industrial Research' (CSIR), New Delhi, for financial support as a SRF (under Bioinformatics; Biotechnology, Engineering discipline) at the 'Institute of Engineering and Technology' (IET), Uttar Pradesh Technical University (UPTU), Lucknow (Uttar Pradesh State), INDIA.

6 References

[1] McGuire A.M. and Church G.M. Predicting regulons and their *cis*-regulatory motifs by comparative genomics. Nucleic Acids Res., 28: 4523–4530, 2000.

[2] Tompa M., Li N., Bailey T.L., Church G.M., De Moor B., Eskin E., Favorov A.V., Frith M.C., Fu Y., Kent W.J., et al. Assessing computational tools for the discovery of transcription factor binding sites. Nature Biotechnology, 23:137–144, 2005.

[3] Fernandez De, Henestrosa A.R., Ogi T., Aoyagi S., Chafin D., Hayes J.J., Ohmori H. and Woodgate R. Identification of additional genes belonging to the LexA regulon in *Escherichia coli*. Mol. Microbiol., 35:1560–1572, 2000.

[4] Khan F., Singh S.P. and Mishra B.N. Conservation of the regulatory *pho*-box in acetoacetyl-CoA reductase promoters across bacterial PHB biosynthetic metabolic pathway. Online Journal of Bioinformatics, 7 (2): 57-68, 2006a.

[5] Khan F., Agrawal S. and Mishra B.N. *In silico* Identification of *cis*-regulatory elements in *Mesorhizobium loti*. Online Journal of Bioinformatics, 6 (2): 129-141, 2005a.

[6] Khan F., Agrawal S. and Mishra B.N. Genomic wide identification of DNA binding motifs of NodD-Factor in *Sinorhizobium meliloti* and *Mesorhizobium loti*. Journal of Bioinformatics & Computational Biology, Vol. 3, No. 4: 773-801, 2005b.

[7] Khan F., Agrawal S. and Mishra B.N. Identification of GltC transcription factor binding DNA motifs and its novel co-regulated genes in nitrogen fixing bacteria. Online Journal of Bioinformatics, 4 (7): 106-114, 2003.

[8] Khan F., Singh S.P., Kumar A., Mehrotra S., Srivastava V. and Mishra B.N. Evolutionary conservation of the regulatory PhoB transcription factor binding sites across eubacteria. Proceedings of All India Seminar on 'Frontier Areas of Chemical Engineering Strategies for Future', 18-19th March, The Institution of Engineers (India), U.P. State Centre, Lucknow and Council of Science & Technology, U.P., Lucknow, India, pg. 39, 2006b.

[9] Khan F., Srivastava A. and Mishra B.N. Identification of regulatory sites controlling expression rate of poly-ß-hydroxybutyrate (PHB) genes under conditions of phosphate

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

starvation by phylogenetic footprinting method. Poster presentation, Proceedings of 7th National Symposium on Biochemical Engineering & Biotechnology, Biohorizon-2005, BMI-26, March 11-12, Indian Institute of Technology Delhi, New Delhi, India, 2005c.

[10] Berg O.G. and Von Hippel P.H. Selection of DNA binding sites by regulatory proteins. Statistical mechanical theory and application to operators and promoters. J.Mol.Biol.,193, 723-750, 1987.

[11] White O., Eisen J.A., Heidelberg J.F., Hickey E.K., Peterson J.D., Dodson R.J., Haft D.H., Gwinn M.L., Nelson W.C., Richardson D.L., Moffat K.S., Qin H., Jiang L., Pamphile W., Crosby M., Shen M., Vamathevan J.J., Lam P., McDonald L., Utterback T., Zalewski C., Makarova K.S., Aravind L., Daly M.J., Minton K.W., Fleischmann R.D., Ketchum K.A., Nelson K.E., Salzberg S., Smith H.O., Venter J.C. and Fraser C.M. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science, 286:1571-1577, 1999.

[12] Makarova K.S., Aravind L., Wolf Y.I., Tatusov, R.L., Minton K.W., Koonin E.V. and Daly M.J. Genome of the Extremely Radiation-Resistant Bacterium *Deinococcus radiodurans* Viewed from the Perspective of Comparative Genomics. Microbiol. Mol. Biol. Rev. 65, 44–79, 2001.

[13] Gudas L.J. and Pardee A.B. Model for regulation of *Escherichia coli* DNA repair functions. Proc. Natl. Acad. Sci. USA 72: 2330-2334, 1975.

[14] Little J.W. and Mount D.W. The SOS regulatory system of *Escherichia coli*. Cell 29: 11-22, 1982.

[15] Radman M. SOS repair hypothesis: Phenomenology of an inducible DNA repair which is accompanied by mutagenesis. Basic Life Sci. 5: 355-367, 1975.

[16] Friedberg E., Walker G. and Siede W. DNA repair and mutagenesis. ASM Press, Washington, D.C., 1995.

[17] Wojciechowski M.F., Peterson K.R. and Love P.E. Regulation of the SOS Response in *Bacillus subtilis*: Evidence for a LexA Repressor Homolog. Journal of Bacteriology, Vol. 173, No. 20, pg. 6489-6498, 1991.

[18] Davis E.O., Dullaghan E.M. and Rand L. Definition of the Mycobacterial SOS Box and Use To Identify LexA-Regulated Genes in *Mycobacterium tuberculosis*. Journal of Bacteriology, Vol. 184, No. 12, pg. 3287-3295, 2002.

[19] Brent R. and Ptashne M. Mechanism of action of the lexA gene product. Proc. Natl. Acad. Sci. USA, 78: 4204-4208, 1981.

[20] Little J.W., Mount D.W. and Yanisch-Perron C.R. Purified lexA protein is a repressor of the recA and lexA genes. Proc. Natl. Acad. Sci. USA 78: 4199-4203, 1981.

[21] Sassanfar M. and Roberts J.W. Nature of the SOS-inducing signal in *Escherichia coli*. The involvement of DNA replication. J. Mol. Biol. 212: 79-96, 1990.

[22] Little J.W. Mechanism of specific LexA cleavage: autodigestion and the role of RecA coprotease. Biochimie 73: 411-421, 1991.

[23] Bertrand-Burggraf E., Hurstel S., Daune M. and Schnarr M. Promoter properties and negative regulation of the uvrA gene by the LexA repressor and its amino-terminal DNA binding domain. J. Mol. Biol. 193: 293-302, 1987.

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

[24] Lewis L.K., Harlow G.R., Gregg-Jolly L.A. and Mount D.W. Identification of high affinity binding sites for LexA which define new DNA damage-inducible genes in *Escherichia coli*. J. Mol. Biol. 241: 507-523, 1994.

[25] Tapias A. and Barbe J. Mutational analysis of the *Rhizobium etli* recA operator. J. Bacteriol. 180: 6325-6331, 1998.

[26] Cheo D.L., Bayles K.W. and Yasbin R.E. Cloning and characterization of DNA damage-inducible promoter regions from *Bacillus subtilis*. J. Bacteriol. 173: 1696-1703, 1991.

[27] Cheo D.L., Bayles K.W. and Yasbin R.E. Elucidation of regulatory elements that control damage induction and competence induction of the *Bacillus subtilis* SOS system. J. Bacteriol. 175: 5907-5915, 1993.

[28] Winterling K.W., Chafin D., Hayes J.J., Sun J., Levine A.S., Yasbin R.E. and Woodgate R. The *Bacillus subtilis* DinR binding site: redefinition of the consensus sequence. J. Bacteriol. 180: 2201-2211, 1998.

[29] Ashlee M. Earl. Global expression analysis of *Deinococcus radiodurans* response to ionizing radiation: IrrE is a novel regulator of this response. Ph.D. Dissertation, Department of Biological Sciences, Graduate Faculty of the Louisiana State University and Agricultural and Mechanical College, 2003.

[30] Makita Y., Nakao M., Ogasawara N. and Nakai K. DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics Nucleic Acids Res., 32, D75-77, 2004.

[31] Schneider T.D., Stormo G.D., Gold L. and Ehrenfeucht A. Information content of binding sites on nucleotide sequences. J. Mol. Biol., 188: 415–431, 1986.

[32] Van Helden J., Andre B. and Collado-Vides J. A web site for the computational analysis of yeast regulatory sequences. Yeast, 16 (2): 177-187, 2000.

[33] Hertz G.Z. and Stormo G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics, 15: 563-577, 1999.

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

[35] Uchiyama I. MBGD: microbial genome database for comparative analysis. Nucleic Acids Res., 31: 58-62, 2003.

[36] Marchler Bauer A., Panchenko A.R., Shoemaker B.A., Thiessen P.A., Geer L.Y. and Bryant S.H. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Research, 30: 281-283, 2002.

[37] Higgins D., Thompson J., Gibson T., Thompson J.D., Higgins D.G. and Gibson T.J. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res., 22: 4673-4680, 1994.

[38] Yang Z. Phylogenetic analysis using parsimony and likelihood methods. J Mol Evol., 42 (2): 294-307, 1996.

[39] Tatusov R.L., Fedorova N.D., Jackson J.D., Jacobs A.R., Kiryutin B., Koonin E.V., Krylov D.M., Mazumder R., Mekhedov S.L., Nikolskaya A.N., Rao B.S., Smirnov S., Sverdlov A.V., Vasudevan S., Wolf Y.I., Yin J.J. and Natale D.A. The COG database: an updated version includes eukaryotes. BMC Bioinformatics, 4 (1): 41, 2003.

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

[40] Hamoen L.W., Haijema B., Bijlsma J.J., Venema G. and Lovett C.M. The *Bacillus subtilis* competence transcription factor, ComK, overrides LexA-imposed transcriptional inhibition without physically displacing LexA. J. Biol. Chem.,16, 276 (46): 42901-7, 2001.

[41] Kawai Y., Moriya S. and Ogasawara N. Identification of a protein, YneA, responsible for cell division suppression during the SOS response in *Bacillus subtilis*. Mol. Microbiol.; 47 (4): 1113-22, 2003.

[42] Rahmann S, Müller T, Vingron M. On the power of profiles for transcription factor binding site detection. Statistical Applications in Genetics and Molecular Biology, 2 (1): Article 7, 2003.

[43] Beckstette M, Homann R, Giegerich R, Kurtz S. PoSSuMsearch: Fast index based algorithms and software for matching position specific scoring matrices. BMC Bioinformatics, 7 (1): 389, 2006.

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Supplementary Tables

Supplementary Table 1 (S1): Details of example data set genes used for construction of improved SOS-box weight matrix.

S. No.	Gene ID	Gene name	Organism	Strand	Left	Right	Description [synonyms]
1.	NP_388444.1	*dinB*	*B. subtilis*	R	607791	608309	nuclease inhibitor [dinB;BSU05630;938057]
2.	NP_391458.1	*tagC*	*B. subtilis*	D	3682468	3683796	polyglycerol phosphate assembly and export (teichoic acid biosynthesis) [tagC;BSU35770;936812]
3.	NP_389668.1	*lexA*	*B. subtilis*	R	1916848	1917465	LexA repressor [lexA;BSU17850;939564]
4.	NP_389576.1	*recA*	*B. subtilis*	D	1763917	1764960	multifunctional SOS repair regulator [recA;BSU16940;939497]
5.	NP_391397.1	*uvrB*	*B. subtilis*	R	3611975	3613960	excinuclease ABC subunit B [uvrB;BSU35170;936663]
6.	NP_389669.1	*yneA*	*B. subtilis*	D	1917615	1917932	hypothetical protein [yneA;BSU17860;936112]
7.	NP_285667.1	*lexA*	*D. radiodurans*	R	379160	379822	lexA repressor [LexA;DRA0344;1799632]
8.	NP_296061.1	*recA*	*D. radiodurans*	D	2337795	2338886	recombinase A [recA;DR2340;1798669]

Supplementary Table 2 (S2): Upstream sequences of example data set genes used in the construction of improved SOS-box weight matrix.

```
>dinB  NP_388444.1; upstream from -400 to -1; size: 400; location: NC_000964.2 608310
608709 R B. subtilis
CGAGCAGCTTTTTCAGTCTGTTAGAGATCGCGCTCTTATGTACATTTTGGTACATGGCCA
GGCTGCCGGGAGAGGTTGGGCCCTTGACTTTTAAAATTTTCAACAGGTCTGCCTGTTCCC
TTGAGATATGCTTGAAGACATCCTGATTGATTTCACGATGCACAAATTCTGTCCCTTCAA
GCAGGACTTCCACAAACAGGTTTATCGCCTGACATAGCTTTTGTTCATTCAATATCTTCC
CCTCCTCATTATAGTTTACCCCGCTAAACTTTATGTTCAATGTCAATTAGTTTATTCCTC
TAAACTATATATGTCAACAAATTTTATTTTCGAGCACTCTACATAAGAACTCATGTTCGT
GTATAATGAAAGCTATACACACGAAAGGGGGAATTTTAAC
>tagC  NP_391458.1; upstream from -400 to -1; size: 400; location: NC_000964.2 3682068
3682467 D B. subtilis
CTGACTATTCTTCTGTTCCCTTTGAATTTTCGATTTTAAATAAACCCATTCTTTTTTATA
CTTACGATTTAAAACTTTATCAACAGAAGCGGGGATTGGTTGATAATTATCTTTCCATAA
TACCTGGAAGAGCCTGCTATGACAGTGAATCATTAATAAATGAAATTCAAACCCCATTTA
ATTATTCCAAAATAAAAGTTTTTTCCGATAGATGGAATAAGTATTCTGATGGGAATTCAA
GCCAAAATTTATTGAATTTCATCGAAAATTTAATAAGCTAAATGATGACACTTGTTCAAA
ACAGAACAAGTGTTCTTTTTTCTATTGAATACCGAACGTATGTTTGCTTTAATGTAATTA
GTTGCTTATGGCAGCAAATAATAATAGAGGTGGTAAATTT
>lexA  NP_389668.1; upstream from -400 to -1; size: 400; location: NC_000964.2 1917466
1917865 R B. subtilis
CCCGGCTGGATATCAGAGGTTTGAAGTTGATTTTTATCAGCTACCCATTCAATAAAATCA
TTTTTGTTTATCTTTTTTGTATCGGCTACCTGATCAGCAATTGACCAGAGTGTGTCGCCT
TGCTGGACTTCTATTTTAACATACTGATTAAGCTCTTGGCCGCTGCTTGTATATGACAGC
ATAAGGATAACCGCGCTCAAAATCACTGTAAACAGACCGACAAAAATAATAGATTCTTTA
CTCATGATCATAACCTCCAACAGGAATGTTTGTTCGCATTTTTATAATTTCAGTATATAC
GAACAAACGTTTCTGTCAATTGTTTTTTCGAACCTATGTTTGTACTGTAAGGAAAATCAT
GCTATAATCTTCTTAAAATCGACGTTTTGAGGTGCGAAAA
>recA  NP_389576.1; upstream from -400 to -1; size: 400; location: NC_000964.2 1763517
```

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

1763916 D *B. subtilis*
AAGGGTGTTCAAAAACTCACTGGCAGCGATATCGGCATTAGCTTTACTGGTGTAGCAGGA
CCTGATGCTCAAGAAGGGCATGAGCCTGGGCATGTGTTTATCGGCATTTCCGCAAATGGT
AAAGAAGAGGTTCACGAGTTTCACTTTGCGGGCTCCAGAACGGGGATCAGAAAACGCGGC
GCTAAATACGGCTGCCATTTAATCTTAAAGCTTTTAGAGCAAAAATAATATTTTCAGCAC
ATTATCCTCCTAAGAAAACATGATTTCTCTGATACATTATGATATTTTGATAGGAATCAC
GCCAAGAAAAAATCCGAATATGCGTTCGCTTTTTTCTTGGCAAATCCCTTCAAACAGGGT
ATAGTATATGTAGTGGTAACATAAAGGAGGAAAAAATAGA
>uvrB NP_391397.1; upstream from -400 to -1; size: 400; location: NC_000964.2 3613961
3614360 R *B. subtilis*
TTTTTGCATTGTTTTTCCCTTTTATGCTTGTTGTTCTATTTACTAGAGTCACCTTTAATC
ATTATGTGGCGATCGCTTTAACAGCTGCATTGCTGTTTGCCTCTTATTTAAAAGGCTATA
CAGAAACGTATTTTATTGTAGGATTGGATGTTGTGTCTCTTGTGGCTGGCGGACTGTATA
TGGCCAAAAAAGCTGCAGAGAAAAAAGAAGAATAAATCGGACATAATGAATATAAAGACT
GAATACCTGCTTTTACGTTTTAAAAGCAGGTTTTTTATACACAAAAACAGCTGGAAATAA
AAAACCACCGAACTTTAGTTCGTATTTTTAGTGATTTTGCTTTCCATTGTGTTACTATAT
CTATAGGAAGATTTCGTTAAAGAAACGGAGGCTTATTTTT
>yneA NP_389669.1; upstream from -400 to -1; size: 400; location: NC_000964.2 1917215
1917614 D *B. subtilis*
GGTACATTAACGACCTGGCTTTGCGGAATGTCTACTTCTTCATCAAGAATTTCTATTGCT
CTTGGTTTTGTCGGATCTCGTCTGATCAGCCCTTTTGTTTCCAAACGGGCCAAATGGCCG
TGGACAGTAGAACTGGACGCAAGCCCGACAGCCTCTCCGATCTCTCTCACGGAAGGCGGA
TATCCTTTTGATTTAACCTCTGCTTTAATAAAACGGAGGATATCAAGTTGCCTTTTTGAT
AGCTTCGTCATTTTTCGCACCTCAAAACGTCGATTTTAAGAAGATTATAGCATGATTTTC
CTTACAGTACAAACATAGGTTCGAAAAAACAATTGACAGAAACGTTTGTTCGTATATACT
GAAATTATAAAAATGCGAACAAACATTCCTGTTGGAGGTT
>lexA DRA0344 NP_285667.1; upstream from -400 to -1; size: 400; location: NC_001264.1
379823 380222 R *D. radiodurans*
GACCTGCTGATGGCCGGGCACCTGCTCGCCATTGCCGAGCGCGAGCAGCGGTGCGGGCCG
ACGCTGGTCTTCGCCCACAACCTGCACCTGCAACGCCCGCTCAGCCGTATGAAGATGGGG
GCCCAGAAGATAGGGGCTCAGACGCTCGACTGGTGGGGCGCCGGGGCGCACGTCAGCCGC
CGCCTAGGCGAGCGCTACGCCTTCATCGCCAGTTATCCGGGCGCCGCAGAGGAGACTTTT
CTGGTGCCTTCTTCCGAACTCACGGAAGCTGCGGCCACCCAGTTGTCAGTCAGGGACGAC
GGTCAGCCCTATTCTTTTCCGTTCAAACTCTCTGACCTCCCGTTTCTGAACGGTGCGCTC
CTCGTAGGCCAATTCTGACGGTTGGGCCGCCGCTTGACCC
>recA dra NP_296061.1; upstream from -400 to -1; size: 400; location: NC_001263.1
2337395 2337794 D *D. radiodurans*
CGCAGGACCTGCCACCGCTGCACGTCAACCTGCGTGGCACTGGCTACTTCCCGAACGAGG
GCAGCCCGCGCGTCTGGTTCGTCAAGACCGAGGCCGAGGGCCTGACCGAACTCGCGGAAA
ACCTGCGGGCCGGCATCCGTGAGCTGGGCATCGGCACCGACGACCTCGCGTTCAAGGCGC
ACATCACCCTCGCCCGCAAGAAGGGGCCCGCGCCCCGCCTCCCTCCCCTCATTTTCGACC
AGAGCTGGACGGCCCCAGGACTGACGCTTTACCGCTCGATCCTGCGTAAGACCGGCCCGA
TCTACGAAGTTCAGAGCACGTTCCGTTTCCGGGGGTCAGCTTCCCAGACCTCCGCCGAAT
CCGCCCCCACTGCTGCGCCGAGCGCGCCCCAGGAGCACCC

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Supplementary Table 3 (S3): Details of the statistical data resulted during construction of matrix using example data set. The data showing information content, motif alignment, position frequency matrix and positional weight matrix.

Construction of Matrix using example data set through 'CONSENSUS' algorithm: (matrix 1 from final cycle)

Information content is calculated using natural logarithms (i.e. Base e) divide by ln(2) = 0.693 to convert to Base 2
Number of sequences = 8
Unadjusted information = 42.1989
Sample size adjusted information = 42.1981
ln(p-value) = -159.481 p-value = 5.47249E-70
ln(expected frequency) = -106.962 expected frequency = 3.52438E-47

Motif Alignment (14 nt):

1|1: 1/346	AGAACTCATGTTCG	sequence 1: dinB_bsu
2|3: 2/333	CGAACGTATGTTTG	sequence 2: tagC_bsu
3|6: 3/329	CGAACCTATGTTTG	sequence 3: lexA_bsu
4|2:-4/315	CGAACGCATATTCG	sequence 4: recA_bsu
5|5:-5/309	CGAACTAAAGTTCG	sequence 5: uvrB_bsu
6|4:-6/310	CGAACCTATGTTTG	sequence 6: yneA_bsu
7|8: 7/255	CGAACTCACGGAAG	sequence 7: DRA0344_lexA_dra
8|7:-8/161	TGAACGCGAGGTCG	sequence 8: recA_dra

Alignment matrix (14 nt): SOS-box motif

A |	1	0	8	8	0	0	1	7	2	1	0	1	1	0
C |	6	0	0	0	8	2	4	0	1	0	0	0	4	0
G |	0	8	0	0	0	3	0	1	0	7	2	0	0	8
T |	1	0	0	0	0	3	3	0	5	0	6	7	3	0
Sum	8	8	8	8	8	8	8	8	8	8	8	8	8	8
Cons.	**C**	**G**	**A**	**A**	**C**	**G/T**	**C**	**A**	**T**	**G**	**T**	**T**	**C**	**G**

Weight matrix (14 nt): SOS-box

A	C	G	T
-0.73	1.24	-2.20	-0.73
-2.20	-2.20	1.52	-2.20
1.12	-2.20	-2.20	-2.20
1.12	-2.20	-2.20	-2.20
-2.20	1.52	-2.20	-2.20
-2.20	0.20	0.58	0.20
-0.73	0.85	-2.20	0.20
0.99	-2.20	-0.41	-2.20
-0.16	-0.41	-2.20	0.67
-0.73	-2.20	1.39	-2.20
-2.20	-2.20	0.20	0.85
-0.73	-2.20	-2.20	0.99
-0.73	0.85	-2.20	0.20
-2.20	-2.20	1.52	-2.20

Matrix based matching & scoring of motif in the example data set through 'PATSER' algorithm:

Information content (base e): 13.233
Sample size adjusted information content (information content minus the average information expected from an arbitrary alignment of random sequences): 10.140
Information content after adding pseudo-counts: 9.652
Maximum score: 15.199

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

```
Minimum score:  -30.761
Range of scores:   15.199 -  -30.761 =   45.960

Minimum score for calculating p-values:    0.000
Maximum ln(numerically calculated p-value):   -4.872
Minimum ln(numerically calculated p-value):  -20.099

ln(cutoff p-value) based on sample size adjusted information content:
-10.140
Numerically calculated cutoff score:    7.852
ln(numerically calculated cutoff p-value):  -10.143
Average score above numerically calculated cutoff:    8.945
```

Supplementary Table 4 (S4): List of studied *D. radiodurans* genes related to replication, repair and recombination functions with their phylogenetic pattern and respective pathway.

Gene name (a)	Gene_ID (b)	Description	Pathway (c)	Phylogenetic pattern (d)
yhdJ	DRC0020	Adenine-specific DNA methylase	mMM?	-m-k--vd-e--huj-------
ogt/ybaZ	DR0248	hypothetical protein	DR	amtkyqvd-ebrhuj---lin-
mutT	DR0261	MutT/nudix family protein	DR	--t----d-ebrhuj---lin-
alkA	DR2074	3-methyladenine glycosidase	DR, BER	-------d--br-----o--nxa-tky--dcebr----------
	DR2584	DNA-3-methyladenine glycosidase II, putative	?	?
mutY	DR2285	A/G-specific adenine glycosylase	BER, MMY	--t----d-ebrhuj---lin-
nth	DR2438	endonuclease III	BER	amtkyqvdcebrhuj--olinx
	DR0289	endonuclease III	?	?
	DR0928	endonuclease III, putative	?	?
mutM/fpg	DR0493	formamidopyrimidine-DNA glycosylase	BER	-------dcebrh--gp-----
nfi (yjaF)	DR2162	hypothetical protein	BER	a--k-qvd-eb----------
polA	DR1707	DNA-directed DNA polymerase	BER	--t--qvdcebrhujgpolinx
ung	DR0689	uracil-DNA glycosylase	BER	----y--d-ebrhujgpo-inx
mug	DR0715	G/U mismatch-specific DNA glycosylase	BER	-------d-e------------
	DR1751	DNA polymerase-related protein	BER	a--k-qvdc-br-----ol--x
	DR0022	hypothetical protein		
xthA	DR0354	exodeoxyribonuclease III	BER	a-t-y--dcebrhuj--ol--x
sms	DR1105	DNA repair protein	NER, BER	-----qvdcebrhuj---linx
mfd	DR1532	transcription-repair coupling factor	NER	------vdcebrhuj--olinx
uvrA	DR1771	excinuclease ABC, subunit A	NER	--t--qvdcebrhujgpolinx
	DRA0188	excinuclease ABC, subunit A	?	?
uvrB	DR2275	excinuclease ABC subunit B	NER	--t--qvdcebrhujgpolinx
uvrC	DR1354	excinuclease ABC, subunit C	NER	--t--qvdcebrhujgpolinx
uvrD	DR1775	DNA helicase II	NER, mMM, SOS	--t-yqvdcebrhujgpolinx

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

	DR1572	helicase-related protein	?	?
mutL	DR1696	DNA mismatch repair protein MutL	mMM, VSP	----yqvdceb-h----olinx--tk-qvdc-b--uj--o----
mutS	DR1976	DNA mismatch repair protein MutS, putative	mMM, VSP	----yqvdceb-h----olinx
	DR1039	DNA mismatch repair protein MutS	?	?
xseA/nec7	DR0186	exodeoxyribonuclease VII large subunit	MM	------vd-ebrhuj----inx
sbcC	DR1922	exonuclease SbcC	RER	amtkyqvdceb------ol---
sbcD	DR1921	exonuclease SbcD, putative	RER	amtkyqvdcebr-----ol---
recA	recA	recombinase A, DR2340	RER, SOS	amtkyqvdcebrhujgpolinx
recD	DR1902	exodeoxyribonuclease V, subunit RecD, putative	RER	-m--y--d-ebrh----o-in-
recF	DR1089	recF protein	RER	-------dcebrh-----li-x
recG	DR1916	DNA helicase RecG	RER	-----qvdcebrhuj--ol--x
recJ	DR1126	single-stranded-DNA-specific exonuclease, putative	RER	amtk-qvdceb-huj--olinx
recN	DR1477	DNA repair protein	RER	-----q-dcebrhuj---l--x
recO	DR0819	hypothetical protein	RER	-------dcebrh-----lin-
recQ	DR2444	nucleic acid-binding protein, putative, HRDC family	RER	----y--dceb-h-----l---
	DR1289	DNA helicase RecQ	?	?
recR	DR0198	recR protein	RER	-----q-dcebrhuj---linx
ruvA	DR1274	Holliday junction binding protein	RER	--t--qvdcebrhujgpolinx
ruvB	DR0596	Holliday junction DNA helicase	RER	------vdceb-hujgpolinx
ruvC	DR0440	Holliday junction resolvase	RER	------vdce-rhuj---linx
dnaE	DR0507	DNA polymerase III, alpha subunit	MP	-----qvdcebrhujgpolinx
dnaQ	DR0856	DNA polymerase III, epsilon subunit, putative	MP	-----qvdcebrhujgpolinx
dnlJ	DR2069	DNA ligase	MP	-----qvdcebrhujgpolinx
ssb	DR0099	single-stranded DNA-binding protein	MP	-----qvdcebrhujgpolinx
lexA	DRA0344	lexA repressor	SOS	-----vdcebrh---------
	DRA0074	hypothetical protein	?	?
ycjD	DR0221	hypothetical protein	VSP?	--t---vd-e-rh---------
	DR2566	hypothetical protein	?	?
ham1/ygV	DR0179	hypothetical protein	DR	amtkyqvdcebrh----olin-
yejH/rad25	DRA0131	hypothetical protein	NER	a--ky--d-e-r------l---
	DR0690	type I topoisomerase, putative	?	----y--d----------
	DR1721	hypothetical protein	?	-------d----------
	DR1262	ribonucleoprotein Ro/SS-A-related protein	?	-------d----------
	DR1757	hypothetical protein	?	-------d----------

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

mrr	DR1877	hypothetical protein	?	------vdc----u--------
	DR0508	mrr restriction system protein	?	?
	DR0587	mrr restriction system protein	?	?

Note:
(a) = The gene names are from *E. coli* whenever an *E. coli* ortholog exists; where no ortholog was detectable in either *E. coli* or *B. subtilis*, no gene is indicated.
(b) = Based largely on reference 12 (Makarova *et al.*, 2001), with modifications.
(c) = Abbreviations of DNA repair pathways: DR, direct damage reversal; BER, base excision repair; NER, nucleotide excision repair; mMM, methylation-dependent mismatch repair; MMY, mutY-dependent mismatch repair; VSP, very-short-patch mismatch repair; RER, recombinational repair, SOS, SOS repair; MP, multiple pathways; ?, unknown possible repair pathways or uncertain assignments.
(d) = Abbreviations in phylogenetic patterns: *a, Archaeoglobus fulgidus; m, Methanococcus jannaschii; t, Methanobacterium thermoautotrophicum; k, Pyrococcus horikoshii; y, Saccharomyces cerevisiae; q, Aquifex aeolicus; v, Thermotoga maritima; c, Synechocystis; e, E. coli; b, Bacillus subtilis; r, Mycobacterium tuberculosus; h, Haemophilus influenzae; u, Helicobacter pylori; j, Helicobacter pylori J99; g, Mycoplasma genitalium; p, Mycoplasma pneumoniae; o, Borrelia burgdorferi; l, Treponema pallidum; i, Chlamydia trachomatis; n, Chlamydia pneumoniae, x, Rickettsia prowazekii*.

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Supplementary Table 5 (S5): Details of SOS-box motifs predicted in the upstream sequences of *D. radiodurans* genes related to replication, repair and recombination functions through RSAT webserver.

Pathway	Gene_ID	RSAT									Phylogenetic pattern#
		Strand	Start	End	Score	ln(P)	Obv. score	MSSP*	RMMP**	Predicted SOS-Box (14 nt)	
BER	DR0715	R	-109	-96	8.83	-10.98	39.59	86.14%	58.10%	TGGACGCGAGGAAG	-------d-e------------
	DR1707	D	-339	-326	7.16	-9.55	37.92	82.51%	47.11%	CGAAGTCGAGGCCG	--t--qvdcebrhujgpolinx
	DR0354	D	-77	-64	5.75	-8.45	36.51	79.44%	37.83%	GGACTTTGAGTTTG	a-t-y--dcebrhuj--ol--x
	DR0493	D	-316	-303	5.04	-7.94	35.8	77.89%	33.16%	AGAACGTGAGTGCG	-------dcebrh--gp-----
	DR2162	R	-339	-326	3.8	-7.08	34.56	75.20%	25.00%	CCAACATGCGCTTT	a--k-qvd-eb----------
	DR2438	D	-39	-26	3.73	-7.04	34.49	75.04%	24.54%	CGAACTGGAGTTCA	amtkyqvdcebrhuj--olinx
	DR0689	R	-395	-382	3.15	-6.67	33.91	73.78%	20.72%	CGCCCCCGAAGACG	----y--d-ebrhujgpo-inx
	DR1751	D	-285	-272	1.46	-5.66	32.22	70.10%	9.61%	TGAACTTGATGTCG	a--k-qvdc-br-----ol--x
DR	DR0179	R	-256	-243	7.61	-9.92	38.37	83.49%	50.07%	CGTAGGCCAGTTCG	amtkyqvdcebrh----olin-
	DR0248	D	-396	-383	4.14	-7.32	34.9	75.94%	27.24%	CGAAAGAATCGTGG	amtkyqvd-ebrhuj---lin-
	DR0261	D	-170	-157	2.63	-6.35	33.39	72.65%	17.30%	CGACTGCGGGTTCA	--t----d-ebrhuj---lin-
MM	DR0186	R	-146	-133	2.54	-6.3	33.3	72.45%	16.71%	CGGATTGGCGCTCG	------vd-ebrhuj----inx
mMM	DRC0020	D	-325	-312	4.19	-7.34	34.95	76.04%	27.57%	CGAACCGAAGTTGT	-m-k--vd-e--huj-------
MP	DR2069	R	-179	-166	6.57	-9.06	37.33	81.22%	43.23%	CAAACAAACGCTGA	-----qvdcebrhujgpolinx
	DR0856	R	-399	-386	5.12	-7.99	35.88	78.07%	33.69%	CGCCTGTCCGTTCG	-----qvdcebrhujgpolinx
	DR0099	R	-210	-197	2.37	-6.19	33.13	72.08%	15.59%	CAAACGGGCGTTGG	-----qvdcebrhujgpolinx
	DR0507	R	-128	-115	1.71	-5.81	32.47	70.65%	11.25%	CAACCGTCGCCTCG	-----qvdcebrhujgpolinx
NER	DRA0131	D	-264	-251	5.19	-8.05	35.95	78.22%	34.15%	TGCGCCCAAGTTCA	a--ky--d-e-r------l---
	DR2275	R	-107	-94	4.52	-7.57	35.28	76.76%	29.74%	AGAACGTAGGGCCA	--t--qvdcebrhujgpolinx
	DR1354	D	-397	-384	3.08	-6.63	33.84	73.63%	20.26%	CGTCTGCGCTTTCG	--t--qvdcebrhujgpolinx
	DR1771	D	-86	-73	2.16	-6.07	32.92	71.63%	14.21%	AGAATCCGCCTTCG	--t--qvdcebrhujgpolinx
	DR1532	D	-68	-55	1.67	-5.78	32.43	70.56%	10.99%	AGTTCGCGTGTTCG	------vdcebrhuj--olinx

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

RER	DR1089	D	-243	-230	6.59	-9.09	37.35	81.27%	43.36%	CGCCCGCGTGGAGG	-------dcebrh-----li-x
	DR0440	D	-331	-318	5.99	-8.63	36.75	79.96%	39.41%	CGCCGGCACGGATG	------vdce-rhuj---linx
	DR1274	R	-382	-369	4.71	-7.7	35.47	77.18%	30.99%	TATCCTCGCGTTCA	--t--qvdcebrhujgpolinx
	DR0596	R	-352	-339	4.6	-7.62	35.36	76.94%	30.26%	CGCACTCTTGTACG	------vdceb-hujgpolinx
	DR0198	D	-242	-229	4.29	-7.42	35.05	76.26%	28.23%	CGAAGGGCCGGGCG	-----q-dcebrhuj---linx
	DR1902	R	-200	-187	4.26	-7.4	35.02	76.20%	28.03%	CGACCATGCGCGAA	-m--y--d-ebrh----o-in-
	DR2444	D	-341	-328	4.07	-7.26	34.83	75.78%	26.78%	CCAACGTGCAGCAG	----y--dceb-h-----l---
	DR1126	R	-133	-120	2.23	-6.11	32.99	71.78%	14.67%	CGCCAAAGATTTCG	amtk-qvdceb-huj--olinx
	DR1922	D	-375	-362	1.74	-5.83	32.5	70.71%	11.45%	CGCGCGTCTGGTCG	amtkyqvdceb------ol---
	DR1477	D	-211	-198	0.83	-5.31	31.59	68.73%	5.46%	CGAAGCGGTGGACG	-----q-dcebrhuj---l--x
	DR1921	D	-233	-220	0.75	-5.26	31.51	68.56%	4.93%	CGTCCTGGGTTTCG	amtkyqvdcebr-----ol---
	DR0819	R	-264	-251	0.5	-5.13	31.26	68.02%	3.29%	CGACCCCGCTTTCA	-------dcebrh-----lin-
	DR1916	D	-204	-191	0.33	-5.04	31.09	67.65%	2.17%	CGAACTTGAAGGCG	-----qvdcebrhuj--ol--x
SOS	DRA0344	D	-146	-133	9.79	-11.89	40.55	88.23%	64.41%	CGAACTCACGGAAG	-----vdcebrh---------
VSP	DR0221	D	-341	-328	3.11	-6.64	33.87	73.69%	20.46%	CGAAGTTGCCGTTG	--t---vd-e-rh---------
mMM, VSP	DR1696	D	-226	-213	5.54	-8.29	36.3	78.98%	36.45%	AGAACTCACGAGCG	----yqvdceb-h----olinx--tk-qvdc-b--uj--o----
	DR1976	R	-247	-234	6.19	-8.77	36.95	80.40%	40.73%	CGACCGGGGGTTGG	----yqvdceb-h----olinx
BER, DR	DR2074	D	-190	-177	1.73	-5.82	32.49	70.69%	11.38%	CGAGTGCGCGGAAG	-------d--br-----o--nxa-tky--dcebr----------
BER, MMY	DR2285	D	-298	-285	1.78	-5.84	32.54	70.80%	11.71%	TGAACGCCATTCTG	--t----d-ebrhuj---lin-
BER, NER	DR1105	R	-349	-336	4.47	-7.54	35.23	76.65%	29.41%	CGAACATGGCCGCG	-----qvdcebrhuj---linx
NER, mMM, SOS	DR1775	D	-168	-155	2.4	-6.21	33.16	72.15%	15.79%	CGAGCGGGCGGAAG	--t-yqvdcebrhujgpolinx
RER, SOS	recA	R	-240	-227	10.35	-12.47	41.11	89.45%	68.10%	CGACCTCGCGTTCA	amtkyqvdcebrhujgpolinx
Unknown	DR1262	R	-81	-68	7.63	-9.93	38.39	83.53%	50.20%	CGGTCTGGGGTTCG	-------d-------------

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

possible repair pathways	DR0508	D	-197	-184	7.12	-9.5	37.88	82.42%	46.84%	TGAACGTGACGGTG	
	DR1039	R	-302	-289	7.01	-9.42	37.77	82.18%	46.12%	GGTCGTCGAGTTCG	
	DR0587	D	-87	-74	6.58	-9.07	37.34	81.24%	43.29%	AAATCGTTGGGTCG	
	DRA0188	R	-17	-4	5.77	-8.46	36.53	79.48%	37.96%	CGGGCAGGCGTTCG	
	DRA0074	D	-322	-309	4.6	-7.62	35.36	76.94%	30.26%	CGAACCGGCGCTCA	
	DR1289	R	-365	-352	4.58	-7.61	35.34	76.89%	30.13%	CGCACGTTGGTTCT	
	DR2584	R	-330	-317	4.56	-7.6	35.32	76.85%	30.00%	CGTCCTTGAAGTCG	
	DR0690	R	-128	-115	4.37	-7.48	35.13	76.44%	28.75%	CTTACTCATGTGCG	----y--d----------
	DR1572	R	-369	-356	3.85	-7.12	34.61	75.30%	25.33%	GGAAGGTGCGTTCC	
	DR2566	R	-393	-380	3.39	-6.81	34.15	74.30%	22.30%	CGAACAGTATTTCT	
	DR0928	R	-344	-331	2.92	-6.53	33.68	73.28%	19.21%	CGACCTCGTGGTGA	
	DR1721	D	-376	-363	2.7	-6.39	33.46	72.80%	17.76%	CGACTTTACGTCCG	-------d----------
	DR1877	R	-208	-195	2.4	-6.21	33.16	72.15%	15.79%	TGGACGCGGGCATG	------vdc----u--------
	DR0289	D	-128	-115	2.28	-6.14	33.04	71.89%	15.00%	GGACCACGCGGTGG	
	DR0022	R	-214	-201	1.44	-5.65	32.2	70.06%	9.47%	CGAACTCACCATCG	
	DR1757	D	-383	-370	0.22	-4.98	30.98	67.41%	1.45%	CGAAACGAAGGTCG	-------d----------

Note:

* = MSSP (Motif Score Similarity Percentage) = Similarity of predicted motif score with known SOS-box consensus sequence (CGAACRNRYGTTCG) score. Maximum score of 15.1991 is for good similarity and minimum score of -30.761 is for poor similarity. Percentage is calculated out of 'Range of scores' i.e. [15.199 - (-30.761)] = 45.960.

** = RMMP (Relative Motif Matching Percentage) = Sequence matching of predicted motif with known SOS-box sequence (CGAACRNRYGTTCG). Percentage is calculated out of maximum score i.e. 15.1991.

= Abbreviations in the phylogenetic patterns are as in Table S4.

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Supplementary Table 6 (S6): Details of similar SOS-box motifs predicted in the upstream sequences of *D. radiodurans* genes related to replication, repair and recombination functions through PoSSuMsearch program.

Pathway	Gene_ID	PoSSuMsearch							Phylogenetic pattern##
		Position	Score	MSS *	P-value **	E-value ***	MSSP#	Predicted SOS-Box	
BER	DR2438	361	54	6.07E-01	9.51E-03	2.32E+02	60.67%	TGGACGCGAGGAAG	amtkyqvdcebrhuj--olinx
	DR0493	84	61	6.85E-01	1.60E-03	3.91E+01	68.54%	CGAAGTCGAGGCCG	-------dcebrh--gp-----
	DR2162	61	56	6.29E-01	5.90E-03	1.44E+02	62.92%	GGACTTTGAGTTTG	a--k-qvd-eb----------
	DR1707	61	67	7.53E-01	2.54E-04	6.22E+00	75.28%	AGAACGTGAGTGCG	--t--qvdcebrhujgpolinx
	DR0689	5	52	5.84E-01	1.48E-02	3.62E+02	58.43%	CCAACATGCGCTTT	----y--d-ebrhujgpo-inx
	DR0715	291	68	7.64E-01	1.83E-04	4.46E+00	76.40%	CGAACTGGAGTTCA	-------d-e----------
	DR1751	115	47	5.28E-01	4.04E-02	9.89E+02	52.81%	CGCCCCCGAAGACG	a--k-qvdc-br-----ol--x
	DR0354	323	63	7.08E-01	8.94E-04	2.19E+01	70.79%	TGAACTTGATGTCG	a-t-y--dcebrhuj--ol--x
DR	DR0248	4	63	7.08E-01	8.94E-04	2.19E+01	70.79%	CGTAGGCCAGTTCG	amtkyqvd-ebrhuj---lin-
	DR0261	230	63	7.08E-01	8.94E-04	2.19E+01	70.79%	CGAAAGAATCGTGG	--t----d-ebrhuj---lin-
	DR0179	144	54	6.07E-01	9.51E-03	2.32E+02	60.67%	CGACTGCGGGTTCA	amtkyqvdcebrh----olin-
MM	DR0186	254	53	5.96E-01	1.19E-02	2.91E+02	59.55%	CGGATTGGCGCTCG	------vd-ebrhuj----inx
mMM	DRC0020	75	69	7.75E-01	1.30E-04	3.18E+00	77.53%	CGAACCGAAGTTGT	-m-k--vd-e--huj-------
MP	DR0507	272	53	5.96E-01	1.19E-02	2.91E+02	59.55%	CAAACAAACGCTGA	-----qvdcebrhujgpolinx
	DR0856	1	53	5.96E-01	1.19E-02	2.91E+02	59.55%	CGCCTGTCCGTTCG	-----qvdcebrhujgpolinx
	DR2069	221	63	7.08E-01	8.94E-04	2.19E+01	70.79%	CAAACGGGCGTTGG	-----qvdcebrhujgpolinx
	DR0099	190	47	5.28E-01	4.04E-02	9.89E+02	52.81%	CAACCGTCGCCTCG	-----qvdcebrhujgpolinx
NER	DR1532	332	56	6.29E-01	5.90E-03	1.44E+02	62.92%	TGCGCCCAAGTTCA	------vdcebrhuj--olinx
	DR1771	314	59	6.63E-01	2.75E-03	6.71E+01	66.29%	AGAACGTAGGGCCA	--t--qvdcebrhujgpolinx
	DR2275	293	48	5.39E-01	3.35E-02	8.19E+02	53.93%	CGTCTGCGCTTTCG	--t--qvdcebrhujgpolinx
	DR1354	3	58	6.52E-01	3.57E-03	8.73E+01	65.17%	AGAATCCGCCTTCG	--t--qvdcebrhujgpolinx
	DRA0131	136	62	6.97E-01	1.20E-03	2.95E+01	69.66%	AGTTCGCGTGTTCG	a--ky--d-e-r------l---
RER	DR1922	25	53	5.96E-01	1.19E-02	2.91E+02	59.55%	CGCCCGCGTGGAGG	amtkyqvdceb------ol---

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

	DR1921	167	50	5.62E-01	2.25E-02	5.51E+02	56.18%	CGCCGGCACGGATG	amtkyqvdcebr-----ol---
	DR1902	200	42	4.72E-01	9.41E-02	2.30E+03	47.19%	TATCCTCGCGTTCA	-m--y--d-ebrh----o-in-
	DR1089	157	68	7.64E-01	1.83E-04	4.46E+00	76.40%	CGCACTCTTGTACG	-------dcebrh-----li-x
	DR1916	196	55	6.18E-01	7.52E-03	1.84E+02	61.80%	CGAAGGGCCGGGCG	-----qvdcebrhuj--ol--x
	DR1126	267	43	4.83E-01	8.04E-02	1.97E+03	48.31%	CGACCATGCGCGAA	amtk-qvdceb-huj--olinx
	DR1477	189	50	5.62E-01	2.25E-02	5.51E+02	56.18%	CCAACGTGCAGCAG	-----q-dcebrhuj---l--x
	DR0819	136	43	4.83E-01	8.04E-02	1.97E+03	48.31%	CGCCAAAGATTTCG	-------dcebrh-----lin-
	DR2444	59	61	6.85E-01	1.60E-03	3.91E+01	68.54%	CGCGCGTCTGGTCG	----y--dceb-h-----l---
	DR0198	158	60	6.74E-01	2.10E-03	5.14E+01	67.42%	CGAAGCGGTGGACG	-----q-dcebrhuj---linx
	DR1274	18	51	5.73E-01	1.83E-02	4.48E+02	57.30%	CGTCCTGGGTTTCG	--t--qvdcebrhujgpolinx
	DR0596	48	55	6.18E-01	7.52E-03	1.84E+02	61.80%	CGACCCCGCTTTCA	------vdceb-hujgpolinx
	DR0440	69	62	6.97E-01	1.20E-03	2.95E+01	69.66%	CGAACTTGAAGGCG	------vdce-rhuj---linx
SOS	DRA0344	254	72	8.09E-01	4.27E-05	1.04E+00	80.90%	CGAACTCACGGAAG	-----vdcebrh---------
VSP	DR0221	59	58	6.52E-01	3.57E-03	8.73E+01	65.17%	CGAAGTTGCCGTTG	--t---vd-e-rh---------
mMM, VSP	DR1696	174	67	7.53E-01	2.54E-04	6.22E+00	75.28%	AGAACTCACGAGCG	----yqvdceb-h----olinx--tk-qvdc-b--uj--o----
	DR1976	153	62	6.97E-01	1.20E-03	2.95E+01	69.66%	CGACCGGGGGTTGG	----yqvdceb-h----olinx
BER, DR	DR2074	210	50	5.62E-01	2.25E-02	5.51E+02	56.18%	CGAGTGCGCGGAAG	-------d--br-----o--nxa-tky--dcebr----------
BER, MMY	DR2285	102	59	6.63E-01	2.75E-03	6.71E+01	66.29%	TGAACGCCATTCTG	--t----d-ebrhuj---lin-
BER, NER	DR1105	51	54	6.07E-01	9.51E-03	2.32E+02	60.67%	CGAACATGGCCGCG	-----qvdcebrhuj---linx
NER, mMM, SOS	DR1775	232	54	6.07E-01	9.51E-03	2.32E+02	60.67%	CGAGCGGGCGGAAG	--t-yqvdcebrhujgpolinx
RER, SOS	recA	160	63	7.08E-01	8.94E-04	2.19E+01	70.79%	CGACCTCGCGTTCA	amtkyqvdcebrhujgpolinx

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Unknown	DR2584	70	58	6.52E-01	3.57E-03	8.73E+01	65.17%	CGGTCTGGGGTTCG	-------d----------
	DR0289	272	55	6.18E-01	7.52E-03	1.84E+02	61.80%	TGAACGTGACGGTG	----y--d----------
	DR0928	56	50	5.62E-01	2.25E-02	5.51E+02	56.18%	GGTCGTCGAGTTCG	----y--d----------
	DR0022	186	51	5.73E-01	1.83E-02	4.48E+02	57.30%	AAATCGTTGGGTCG	----y--d----------
	DRA0188	383	56	6.29E-01	5.90E-03	1.44E+02	62.92%	CGGGCAGGCGTTCG	----y--d----------
	DR1572	31	60	6.74E-01	2.10E-03	5.14E+01	67.42%	CGAACCGGCGCTCA	----y--d----------
	DR1039	98	60	6.74E-01	2.10E-03	5.14E+01	67.42%	CGCACGTTGGTTCT	----y--d----------
	DR1289	35	53	5.96E-01	1.19E-02	2.91E+02	59.55%	CGTCCTTGAAGTCG	----y--d----------
	DRA0074	78	66	7.42E-01	3.53E-04	8.65E+00	74.16%	CTTACTCATGTGCG	----y--d----------
	DR2566	7	56	6.29E-01	5.90E-03	1.44E+02	62.92%	GGAAGGTGCGTTCC	----y--d----------
	DR0690	272	57	6.40E-01	4.60E-03	1.13E+02	64.04%	CGAACAGTATTTCT	----y--d----------
	DR1721	24	59	6.63E-01	2.75E-03	6.71E+01	66.29%	CGACCTCGTGGTGA	----y--d----------
	DR1262	319	61	6.85E-01	1.60E-03	3.91E+01	68.54%	CGACTTTACGTCCG	-------d----------
	DR1757	17	52	5.84E-01	1.48E-02	3.62E+02	58.43%	TGGACGCGGGCATG	------vdc----u--------
	DR1877	192	54	6.07E-01	9.51E-03	2.32E+02	60.67%	GGACCACGCGGTGG	----y--d----------
	DR0508	203	72	8.09E-01	4.27E-05	1.04E+00	80.90%	CGAACTCACCATCG	----y--d----------
	DR0587	313	69	7.75E-01	1.30E-04	3.18E+00	77.53%	CGAAACGAAGGTCG	-------d----------

Note: The PoSSuMsearch program was run on Red Hat Linux release 9.0 (Shrike; Kernel 2.4.20-8 on an i686) Operating System.
* MSS = Matrix Similarity Score cutoff value
** P-value = Probability distribution value
*** E-value = the expected number of matches in a given random sequence database. It is widely accepted measure of the significance. Its calculation is based on p-values, it is simply the p-value times database size.
# MSSP (Motif Score Similarity Percentage) = Similarity percentage of predicted motifs with known SOS-box consensus motif (CGAACRNRYGTTCG). Predicted range of score was 0 to 89 through PoSSuMsearch program. Here maximum score indicates exact similarity with known motif while minimum score indicates poor similarity.
= Abbreviations in the phylogenetic patterns are as in Table S4.

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Supplementary Table 7 (S7): List of *D. radiodurans* genes related to stress response with their phylogenetic pattern and type of stress.

Gene_ID	Gene name	Protein description	Type of stress	Phylogenetic pattern*
DR0607	groL	Hsp10, molecular chaperone	Heat, general	amtkyqvdcebrhujgpolinx
DR0128	grpE	Hsp20, molecular chaperone	Heat, general	--t--qvdcebrhujgpolinx
DR0606	groS	Hsp60, molecular chaperone	Heat, general	----yqvdcebrhujgpolinx
DR0129	dnaK	Hsp70, molecular chaperone	Heat, general	--t-yqvdcebrhujgpolinx
DR0126	dnaJ	Hsp70 chaperone cofactor	Heat, general	--t-yqvdcebrhujgpolinx
DR1424	dnaJ	Hsp70 chaperone cofactor	Heat, general	--t-yqvdcebrhujgpolinx
DR1114	ibpA/ibpB	Small heat shock protein	Heat, general	amtkyqvdcebr---------x
DR1691	ibpA/ibpB	Small heat shock protein	Heat, general	amtkyqvdcebr---------x
DR2056	hslJ	Related to heat shock protein, HslJ; DR1940 contains three repeats of this domain	Heat	-------dce----------
DR1940	hslJ	Related to heat shock protein, HslJ; DR1940 contains three repeats of this domain	Heat	-------dce----------
DR0588	clpA/clpB	ATPase subunit of Clp protease	General	--t-yqvdcebrhujgpolinx
DR1046	clpA/clpB	ATPase subunit of Clp protease	General	--t-yqvdcebrhujgpolinx
DR1117	clpA/clpB	ATPase subunit of Clp protease	General	--t-yqvdcebrhujgpolinx
DR1973	clpX	ATPase subunit of Clp protease	General	----yqvdcebrhuj--olinx
DR0202	clpX	ATPase subunit of Clp protease	General	----yqvdcebrhuj--olinx
DR1972	clpP	ATP-dependent protease with chaperone activity	General	----yqvdcebrhuj--olinx
DR1974	lon	ATP-dependent Lon serine protease	General	----yqvd-eb-hujgpolinx
DR2189	lon	ATP-dependent Lon serine protease	General	----yqvd-eb-hujgpolinx
DR0349	lon	ATP-dependent Lon serine protease	General	----yqvd-eb-hujgpolinx
DR1105	sms	ATP-dependent serine protease	General	-----qvdcebrhuj---linx
DR0327	htrA	Do serine protease, with regulatory PDZ domain	General	--t-yqvdcebrhuj--olinx
DR0745	htrA	Do serine protease, with regulatory PDZ domain	General	--t-yqvdcebrhuj--olinx
DR1599	htrA	Do serine protease, with regulatory PDZ domain	General	--t-yqvdcebrhuj--olinx
DR1756	htrA	Do serine protease, with regulatory PDZ domain	General	--t-yqvdcebrhuj--olinx
DR0984	htrA	Do serine protease, with regulatory PDZ domain	General	--t-yqvdcebrhuj--olinx
DR0300	htrA	Do serine protease, with regulatory PDZ domain	General	--t-yqvdcebrhuj--olinx
DR1308	prc	Tail-specific periplasmic serine protease	General	-----qvdceb-huj--olinx
DR1491	prc	Tail-specific periplasmic serine protease	General	-----qvdceb-huj--olinx
DR1551	prc	Tail-specific periplasmic serine protease	General	-----qvdceb-huj--olinx

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

DR1507	yaeL	Membrane-associated Zn-dependent protease I	General	amtk-qvdcebrhuj--olinx
DR0583	ftsH	ATP-dependent Zn protease	General	----yqvdcebrhujgpolinx
DR1020	ftsH	ATP-dependent Zn protease	General	----yqvdcebrhujgpolinx
DRA0290	ftsH	ATP-dependent Zn protease	General	----yqvdcebrhujgpolinx
DR0190	htpX	Predicted Zn-dependent proteases (possible chaperones)	General	amtkyq-dcebrhuj-------
DR0194	htpX	Predicted Zn-dependent proteases (possible chaperones)	General	amtkyq-dcebrhuj-------
DR1004	sugE	Membrane chaperone	General	-------d-ebr----------
DR1005	sugE	Membrane chaperone	General	-------d-ebr----------
DR1621	hit	Diadenosine tetraphosphate (Ap4A) hydrolase, HIT family, cell cycle regulation	General	amtkyq-dcebrhujgpo-inx
DR0139	hflX	GTPase, protease modulator	General	-m-k-qvdcebrhujgpolinx
DR0646	hflX	GTPase, protease modulator	General	-m-k-qvdcebrhujgpolinx
DR0559	BS_yloA	Fibronectin-binding protein, function unknown	?	amtky-vdc-b--uj--o----
DR1832	BS_ytxJ	General stress protein, related to thioredoxin	General	-------d--b----------
DR0491	thiJ	Protease I, related to general stress protein 18, ThiJ superfamily protein	General	am-kyq-dcebrhujgpol---
DR1199	thiJ	Protease I, related to general stress protein 18, ThiJ superfamily protein	General	am-kyq-dcebrhujgpol---
DR2363	uspA	Universal stress protein, nucleotide-binding	General	amtk-q-dcebrh--------x
DR2132	uspA	Universal stress protein, nucleotide-binding	General	amtk-q-dcebrh--------x
DR1838	spoT	Guanosine polyphosphate (ppGpp) pyrophosphohydrolase/synthetase : no RelA counterpart like in gram-positive bacteria	Starvation	-------dcebrhujgpo----
DRA0065	hupA	Histone-like DNA-binding protein	?	-----qvdcebrhujgpolinx
DRA0243	hmp	Haemoglobin-like flavoprotein	?	----yq-d-eb----------
DR0417	mazF	ppGpp-regulated growth inhibitor	Starvation	-------d-ebr----------
DR0662	mazF	ppGpp-regulated growth inhibitor	Starvation	-------d-ebr----------
DR0416	mazE	Regulatory protein, MazF antagonist	Starvation	-------dce----------
DRA0185	ppx	Phosphatase of ppGpp	?	----yqvdce--huj------x
DR2263	dps	Starvation inducible DNA-binding protein	Starvation	-------dceb-huj-ol--x
DRB0092	dps	Starvation inducible DNA-binding protein	Starvation	-------dceb-huj-ol--x
DR2422	mscL	Large conductance mechanosensitive channel	Osmotic	-------dcebrh--------
DR1995	yggB	Membrane protein	Osmotic	amtk-qvdcebrhuj-ol--x
DR0211	yggB	Membrane protein	Osmotic	amtk-qvdcebrhuj-ol--x
DRB0088	kdpD	Osmosensitive K+ channel histidine kinase sensor domain	Osmotic	-------dce-r----------

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

DR1666	trkA	Potassium uptake system, NAD-binding component	Osmotic	amtk-q-dcebrh-gpol---
DR1667	trkH/trkG	Potassium uptake system component	Osmotic	amtk-qydceb-h-gpol---
DR1668	trkH/trkG	Potassium uptake system component	Osmotic	amtk-qydceb-h-gpol---
DRA0138	proW	Proline/glycine betaine ABC-type transport, permease subunit	Osmotic	a------debr-hj--o----
DRA0139	proW	Proline/glycine betaine ABC-type transport, permease subunit	Osmotic	a------debr-hj--o----
DRA0137	proV	Proline/glycine betaine ABC-type transport, ATPase subunit	Osmotic	a------debr-hj--o----
DRA0135	yehZ	Proline/glycine betaine ABC-type transport, periplasmic binding subunit	Osmotic	a------debr-hj-------
DR1473	pspA	Phage shock protein A, controls membrane integrity	Phage	-----q-dceb----------
DR2068	BS_yloU/BS_yqhY	Alkaline shock protein, function unknown	Alkaline	------vd--b--------in-
DR0389	BS_yloU/BS_yqhY	Alkaline shock protein, function unknown	Alkaline	------vd--b--------in-
DR0907	csp	Cold shock protein, OB fold nucleic acid-binding protein	Cold	-----qvd-ebrh--------x
DR1998	katE	Catalase; DRA0259 has C-terminal proteinase I-like domain	Oxidative	----y--d-eb-huj-------
DRA0259	katE	Catalase; DRA0259 has C-terminal proteinase I-like domain	Oxidative	----y--d-eb-huj-------
DRA0146	catA (S. pombe)	Catalase; eukaryotic type, presumably acquired from nitrogen-fixing bacteria	Oxidative	-------d----------
DRA0145	NA	Peroxidase; Yet present only in plant Polyporaceae spp.	Oxidative	-------d----------
DR1279	sodA	Superoxide dismutase, Mn or Fe dependent	Oxidative	-t--yq-dcebrhuj--o-inx
DR1546	sodC	Superoxide dismutase, Cu/Zn dependent	Oxidative	----yq-d-ebr----------
DRA0202	sodC	Superoxide dismutase, Cu/Zn dependent	Oxidative	----yq-d-ebr----------
DR0865	fur	Ferric uptake regulation protein	?	a----qvdcebrhujgpo----
DR0846	bcp	Antioxidant type thioredoxin fold protein	Oxidative	----yqvdcebrhuj-------
DR1208	bcp	Antioxidant type thioredoxin fold protein	Oxidative	----yqvdcebrhuj-------
DR1538	osmC	Protein involved in alkylperoxide and oxidative stress response, osmotically induced protein	Oxidative	-------d-eb----gp-----
DR1857	osmC	Protein involved in alkylperoxide and oxidative stress response, osmotically induced protein	Oxidative	-------d-eb----gp-----
DR1177	yhfA	Protein involved in alkylperoxide and oxidative stress response, osmotically induced protein	Oxidative	---k-qvd-e----------
DR1849	msrA	Peptide methionine sulfoxide reductase PMSR	Oxidative	-t-y--dcebrhujgp-l---

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

DR2242	ahpC	Thiol-alkyl hydroperoxide reductases	Oxidative/detoxication	amtkyqvdcebr-uj---linx
DR1209	ahpC	Thiol-alkyl hydroperoxide reductases	Oxidative/detoxication	amtkyqvdcebr-uj---linx
DR1982	ahpF/trxB	Thioredoxin reductase/alkyl hydroperoxide reductase	Oxidative/detoxication	amtkyqvdcebrhujgpolinx
DR2623	ahpF/trxB	Thioredoxin reductase/alkyl hydroperoxide reductase	Oxidative/detoxication	amtkyqvdcebrhujgpolinx
DR0412	ahpF/trxB	Thioredoxin reductase/alkyl hydroperoxide reductase	Oxidative/detoxication	amtkyqvdcebrhujgpolinx
DRB0033	ahpF/trxB	Thioredoxin reductase/alkyl hydroperoxide reductase	Oxidative/detoxication	amtkyqvdcebrhujgpolinx
DR2085	grxA	Glutaredoxin	Oxidative/detoxication	a-tky-vdcebrh-----l-x
DRA0072	grxA	Glutaredoxin	Oxidative/detoxication	a-tky-vdcebrh-----l-x
DR2473	BS_cypA or terA of Alcaligenes	Cytochrome P450 (uses O2)	Detoxication	----y--dc-b----------
DR2538	BS_cypA or terA of Alcaligenes	Cytochrome P450 (uses O2)	Detoxication	----y--dc-b----------
DR1723	BS_cypA or terA of Alcaligenes	Cytochrome P450 (uses O2)	Detoxication	----y--dc-b----------
DRA0186	BS_cypA or terA of Alcaligenes	Cytochrome P450 (uses O2)	Detoxication	----y--dc-b----------
DRC0041	BS_cypA or terA of Alcaligenes	Cytochrome P450 (uses O2)	Detoxication	----y--dc-b----------
DRC0001	BS_cypA or terA of Alcaligenes	Cytochrome P450 (uses O2)	Detoxication	----y--dc-b----------
DR2220	terB of Alcaligenes	Function unknown; involved in tellurium resistance response in Alcaligenes	Detoxication	-------d----------
DR2226	terC of Alcaligenes	Function unknown; membrane protein	Detoxication	-------dcebrhuj------x
DR1187	terC of Alcaligenes	Function unknown; membrane protein	Detoxication	-------dcebrhuj------x
DRB0131	terC of Alcaligenes	Function unknown; membrane protein	Detoxication	-------dcebrhuj------x
DR1127	BS_yceH	Toxic anion resistance protein; possibly tellurite resistance	Detoxication	--------d-b----------
DR2225	BS_scp2	Chemical damaging agent resistance; in B. subtilis it is involved in low-temperature and salt stress response	Toxins/general	-------dc-b----------
DR2221	BS_scp2	Chemical damaging agent resistance; in B. subtilis it is involved in low-temperature and salt stress response	Toxins/general	-------dc-b----------
DR2224	BS_scp2	Chemical damaging agent resistance; in B. subtilis it is involved in low-temperature and salt stress response	Toxins/general	-------dc-b----------
DR2223	BS_scp2	Chemical damaging agent resistance; in B. subtilis it is involved in low-temperature and	Toxins/general	-------dc-b----------

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

		salt stress response		
DRA0123	arsC	Arsenate oxidoreductase (arsC-like rodanese protein)	Toxins	-------d-ebh----------
DR0136	arsC	Arsenate oxidoreductase (arsC-like rodanese protein)	Toxins	-------d-ebh----------
DR1372	NA	Desiccation protectant, LEA14 family	Desiccation	am-k---d----------
DRB0118	NA	Desiccation-related protein from Craterostigma plantagineum; found to date only in plants	Desiccation	-------d----------
DR0105	NA	LEA76 family desiccation resistance protein	Desiccation	-------d----------
DR1172	NA	LEA76 family desiccation resistance protein	Desiccation	-------d----------
DRA0345	BS_ybfO	Erythromycin esterase	Drugs	-------d--br----------
DR2257	BS_ybfO	Erythromycin esterase	Drugs	-------d--br----------
DR0454	bacA	BacA bacitracin resistance protein, undecaprenol kinase	Drugs	-----qvdcebr-----o----
DR0455	strA of Streptomyces	Streptomycin resistance protein, streptomycin phosphotransferase	Drugs	-------d----------
DR0066	BS_ycbJ	Antibiotic (aminoglycoside) kinase family protein	Drugs	-------dc-br----------
DRA0194	BS_ycbJ	Antibiotic (aminoglycoside) kinase family protein	Drugs	-------dc-br----------
DR0394	BS_ycbJ	Antibiotic (aminoglycoside) kinase family protein	Drugs	-------dc-br----------
DR0669	BS_ycbJ	Antibiotic (aminoglycoside) kinase family protein	Drugs	-------dc-br----------
DR0842	nimABCD of Bacteroides	5-Nitroimidazole antibiotic resistance protein; distantly related to pyridoxamine phosphate oxidase (PDXH)	Drugs	-------d----------
DR2234	BS_bmrU	Function unknown; involved in multidrug resistance	Drugs	----y-vdcebr----------
DR1363	BS_bmrU	Function unknown; involved in multidrug resistance	Drugs	----y-vdcebr----------
DR1560	BS_bmrU	Function unknown; involved in multidrug resistance	Drugs	----y-vdcebr----------
DR1016	thdR	Thiophen and furan oxidation, predicted GTPase	Drugs	----yqvdceb-hujgpolinx
DR1419	BS_tmrB	Tunicamycin resistance protein, predicted ATPase	Drugs	-------d--b----------
DRA0241	BS_penP	β-Lactamase	Drugs	-------dc-br----------
DR0433	BS_penP	β-Lactamase	Drugs	-------dc-br----------
DRA0284	ybgL	Function unknown, lactam utilization protein	Drugs	---k---d-eb-h----------
DR1695	gloA	Lactoylgluthation lyase, fosphomicin resistance protein	Drugs	amtk---dcebrh----------
DR2022	gloA	Lactoylgluthation lyase, fosphomicin resistance protein	Drugs	amtk---dcebrh----------
DR2104	gloA	Lactoylgluthation lyase,	Drugs	amtk---dcebrh----------

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

		fosphomicin resistance protein		
DR2208	gloA	Lactoylgluthation lyase, fosphomicin resistance protein	Drugs	amtk---dcebrh----------
DR0109	gloA	Lactoylgluthation lyase, fosphomicin resistance protein	Drugs	amtk---dcebrh----------
DRA0224	gloA	Lactoylgluthation lyase, fosphomicin resistance protein	Drugs	amtk---dcebrh----------
DR1619	BS_yoaR	Induced by vancomycin in Enterococcus faecalis	Drugs	-------d--b----------
DR0009	BS_yoaR	Induced by vancomycin in Enterococcus faecalis	Drugs	-------d--b----------
DR0025	BS_yoaR	Induced by vancomycin in Enterococcus faecalis	Drugs	-------d--b----------
DR2034	BS_yokD	Induced by vancomycin in Enterococcus faecalis	Drugs	-------d--b----------
DR0599		Induced by vancomycin in Enterococcus faecalis	Drugs	-------d--b----------
DR2000	BS_yocD	Function unknown; homologs of microcine C7 resistance protein MccF	Drugs	-------dceb--------l--x
DR2164	lytB	Function unknown; penicillin tolerance protein	Drugs	-----qvdcebrhuj---lin-
DR2545	BS_yrpB	2-Nitropropane dioxygenase	Drugs	a---yqvd--br-uj-------
DR1182	BS_ywnH	Phosphinothricin aminoacetyltransferase	Drugs	-------dceb----------

Note:
* = Abbreviations in the phylogenetic patterns are as in Table S4.

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Supplementary Table 8 (S8): Details of SOS-box motifs predicted in the upstream sequences of *D. radiodurans* genes related to stress response through RSAT webserver.

Type of Stress	Gene_ID	RSAT									Phylogenetic pattern
		Strand	Start	End	Score	ln(P)	Obv. score	MSSP	RMMP	Predicted SOS-box	
Heat, general	DR0129	D	-244	-231	7.86	-10.15	38.62	84.03%	51.71%	ggaaCGAACCTGCGGGCGactg	--t-yqvdcebrhujgpolinx
	DR1691	R	-129	-116	6.22	-8.81	36.98	80.46%	40.92%	gcaaCGATCTCGAGTCCGtact	amtkyqvdcebr---------x
	DR0606	D	-158	-145	5.03	-7.94	35.79	77.87%	33.09%	tggaAGCACGTATTGTCGccct	----yqvdcebrhujgpolinx
	DR1114	D	-27	-14	4.34	-7.45	35.1	76.37%	28.55%	tggaTGCACTCAAGTTTTaata	amtkyqvdcebr---------x
	DR0128	R	-103	-90	3.94	-7.17	34.7	75.50%	25.92%	acgcCGACTTTAAGCCCGctga	--t--qvdcebrhujgpolinx
	DR0126	R	-139	-126	3.25	-6.73	34.01	74.00%	21.38%	gtgcCGACCCCGCGCTGGaaga	--t-yqvdcebrhujgpolinx
	DR0607	D	-192	-179	2.89	-6.51	33.65	73.22%	19.01%	ccatGGAAGTCAAGGAAGgcga	amtkyqvdcebrhujgpolinx
	DR1424	D	-164	-151	1.16	-5.49	31.92	69.45%	7.63%	tgtgGGAACCCGGAGCCGtcct	--t-yqvdcebrhujgpolinx
Heat	DR2056	R	-112	-99	5.93	-8.57	36.69	79.83%	39.02%	cggaCGGTCGTTGGTTCGaccc	-------dce------------
	DR1940	R	-300	-287	1.88	-5.91	32.64	71.02%	12.37%	tgctCCCCCCGGGGTTCGgcgg	-------dce------------
General	DR2363	R	-322	-309	8.51	-10.72	39.27	85.44%	55.99%	gcctCGCCCGCGAGTTCGggcg	amtk-q-dcebrh--------x
	DR1046	D	-235	-222	8	-10.26	38.76	84.33%	52.63%	agaaCGAACCAGCATTTGagga	--t-yqvdcebrhujgpolinx
	DR1005	D	-307	-294	7.15	-9.54	37.91	82.48%	47.04%	gagaTGAACGCATGGACAgccc	-------d-ebr----------
	DR0984	R	-243	-230	5.48	-8.26	36.24	78.85%	36.05%	ccaaCTACCACGTGGTCGccgg	--t-yqvdcebrhuj--olinx
	DRA0290	D	-195	-182	5.4	-8.2	36.16	78.68%	35.53%	aacgCAACCGCATAGTCGtcgc	----yqvdcebrhujgpolinx
	DR1973	R	-45	-32	5.27	-8.09	36.03	78.39%	34.67%	tgatCGACCAAGTGATCGacca	----yqvdcebrhuj--olinx
	DR1974	D	-261	-248	5.27	-8.09	36.03	78.39%	34.67%	tgacCGATCTGCTGTTCGaact	----yqvd-eb-hujgpolinx
	DR1308	D	-245	-232	5.25	-8.08	36.01	78.35%	34.54%	gcgcCGAAAGCCTGATCGacaa	-----qvdceb-huj--olinx
	DR0491	R	-78	-65	5.09	-7.97	35.85	78.00%	33.49%	tcgcCGAGCACGGGGTCActtt	am-kyq-dcebrhujgpol---
	DR1551	D	-300	-287	5.07	-7.96	35.83	77.96%	33.36%	tagaCGAACTCGCCCTTGgtga	-----qvdceb-huj--olinx
	DR1491	D	-39	-26	4.84	-7.8	35.6	77.46%	31.84%	aggcTGAAGGCAAGGCCGcccc	-----qvdceb-huj--olinx
	DR0300	D	-395	-382	4.74	-7.72	35.5	77.24%	31.19%	cacccGAACACACCGATGagcg	--t-yqvdcebrhuj--olinx
	DR2189	R	-284	-271	4.73	-7.72	35.49	77.22%	31.12%	cgcaCTACCTTGGGGTGGtagg	----yqvd-eb-hujgpolinx
	DR1105	R	-349	-336	4.47	-7.54	35.23	76.65%	29.41%	aggaCGAACATGGCCGCGtgac	-----qvdcebrhuj---linx
	DR1507	R	-261	-248	4.23	-7.37	34.99	76.13%	27.83%	ttccCTGCCTCGCGCTCGctta	amtk-qvdcebrhuj--olinx
	DR1621	D	-396	-383	4.19	-7.34	34.95	76.04%	27.57%	tcgcCGAACGACTGCGCGgcgt	amtkyq-dcebrhujgpo-inx

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

	DR2132	D	-270	-257	3.97	-7.2	34.73	75.57%	26.12%	tcacCGAACTGATGCAGGacac	amtk-q-dcebrh--------x
	DR1117	R	-306	-293	3.87	-7.13	34.63	75.35%	25.46%	agccCGCCCACGTGGTCGtcgt	--t-yqvdcebrhujgpolinx
	DR1004	D	-272	-259	3.79	-7.08	34.55	75.17%	24.94%	gcggCAAACGTCAGGATGctgt	-------d-ebr----------
	DR0583	D	-275	-262	3.49	-6.89	34.25	74.52%	22.96%	tctcCGGGCGCAATTTTGcgcg	----yqvdcebrhujgpolinx
	DR0139	D	-346	-333	3.24	-6.72	34	73.98%	21.32%	tgccTGAACGTGTCTTTTtacc	-m-k-qvdcebrhujgpolinx
	DR1199	D	-351	-338	3.12	-6.65	33.88	73.72%	20.53%	aggtCGTTCCTAACTTCGagaa	am-kyq-dcebrhujgpol---
	DR1599	D	-191	-178	2.95	-6.54	33.71	73.35%	19.41%	ctggTGATCGTGCGGGCGggga	--t-yqvdcebrhuj--olinx
	DR1020	D	-389	-376	2.87	-6.48	33.63	73.17%	18.88%	gggtTGAGCGCCAGGATGgtcg	----yqvdcebrhujgpolinx
	DR0588	D	-291	-278	2.86	-6.48	33.62	73.15%	18.82%	gcctCGACCGCATCTACCtgca	--t-yqvdcebrhujgpolinx
	DR0745	D	-284	-271	2.81	-6.46	33.57	73.04%	18.49%	gcttCTACCGCGTGGAAGacag	--t-yqvdcebrhuj--olinx
	DR0202	R	-96	-83	2.58	-6.31	33.34	72.54%	16.97%	gtgcCGAACTGTGGAACGtgaa	----yqvdcebrhuj--olinx
	DR0349	D	-102	-89	2.3	-6.15	33.06	71.93%	15.13%	tattTGCACTCAGGTTTTttgg	----yqvd-eb-hujgpolinx
	DR0646	D	-265	-252	1.96	-5.95	32.72	71.19%	12.90%	gagtCGCCCGCGAGGTCCgccg	-m-k-qvdcebrhujgpolinx
	DR1832	R	-167	-154	1.75	-5.83	32.51	70.74%	11.51%	gcacCGGATGCGGGTCCAtttc	-------d--b----------
	DR0327	R	-374	-361	1.7	-5.8	32.46	70.63%	11.18%	ggcgGGACCGTAGGTCAGcagc	--t-yqvdcebrhuj--olinx
	DR0194	D	-310	-297	1.68	-5.78	32.44	70.58%	11.05%	tcggCGGGCGGCTGGTCGtgcc	amtkyq-dcebrhuj-------
	DR1972	D	-291	-278	1.53	-5.7	32.29	70.26%	10.07%	accaCGCTCGCCTGGGCGgccc	----yqvdcebrhuj--olinx
	DR0190	R	-378	-365	0.68	-5.23	31.44	68.41%	4.47%	tgcaGGTCCACGCTCTCGccgc	amtkyq-dcebrhuj-------
	DR1756	R	-356	-343	0.42	-5.09	31.18	67.84%	2.76%	tgggAAACCCAGCGTTAGggcg	--t-yqvdcebrhuj--olinx
Starvation	DR0662	D	-272	-259	7.17	-9.57	37.93	82.53%	47.17%	agccAGAACGTATGAATGctag	-------d-ebr----------
	DR2263	R	-148	-135	6.14	-8.75	36.9	80.29%	40.40%	tgctCAACCGCGCGGTCAtcag	-------dceb-huj-ol--x
	DRB0092	R	-183	-170	4.48	-7.54	35.24	76.68%	29.48%	ctccGGAAAGTGGGTTCAcgct	-------dceb-huj-ol--x
	DR1838	D	-119	-106	4.25	-7.39	35.01	76.17%	27.96%	ggttCGCACCCGCGAACGccgt	-------dcebrhujgpo----
	DR0417	R	-294	-281	2.44	-6.23	33.2	72.24%	16.05%	tatcCGTCATAGTGTTCGctcg	-------d-ebr----------
	DR0416	R	-58	-45	2.44	-6.23	33.2	72.24%	16.05%	tatcCGTCATAGTGTTCGctcg	-------dce-----------
Osmotic	DRA0139	D	-240	-227	8.25	-10.49	39.01	84.88%	54.28%	tgctTGAACCCGAGGACGtgac	a------debr-hj--o----
	DR2422	R	-316	-303	4.84	-7.8	35.6	77.46%	31.84%	gaggCGTCGGTGCGCTCGcgca	-------dcebrh--------
	DR1668	R	-149	-136	4.81	-7.78	35.57	77.39%	31.65%	gggaCCCCCGCAAGTTCGgtgg	amtk-qydceb-h-gpol---
	DR1995	D	-156	-143	4.74	-7.72	35.5	77.24%	31.19%	tggtCGAGCGCCTGGAAGcggc	amtk-qvdcebrhuj-ol--x
	DR1666	D	-357	-344	4.48	-7.55	35.24	76.68%	29.48%	gcggCGAACTCATCTTGTtcga	amtk-q-dcebrh-gpol---
	DRA0138	D	-318	-305	4.46	-7.52	35.22	76.63%	29.34%	tcggCGACCGTATCTGCGtgct	a------debr-hj--o----
	DRB0088	D	-307	-294	4.05	-7.24	34.81	75.74%	26.65%	cccaCCAACGCCAAGTTGcagg	-------dce-r----------

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

	DRA0135	D	-72	-59	2.46	-6.24	33.22	72.28%	16.19%	cggaCGCACCAGGGTCCGagga	a------debr-hj-------
	DR0211	R	-188	-175	2.09	-6.02	32.85	71.48%	13.75%	gcacGGGCCTCGCGTCCGcaga	amtk-qvdcebrhuj-ol--x
	DRA0137	R	-21	-8	1.61	-5.76	32.37	70.43%	10.59%	ctctCCTCCATAACTTCCtcct	a------debr-hj--o----
	DR1667	D	-131	-118	1.43	-5.64	32.19	70.04%	9.41%	tcatCGCGCTGATGTTCCtggg	amtk-qydceb-h-gpol---
Phage	DR1473	R	-46	-33	2.81	-6.46	33.57	73.04%	18.49%	caacCCTCCTCAGATTCGccgc	-----q-dceb----------
Alkaline	DR0389	D	-61	-48	8.05	-10.3	38.81	84.44%	52.96%	acggCGAGCTCACGGACGcggg	------vd--b--------in-
	DR2068	R	-152	-139	6.71	-9.17	37.47	81.53%	44.15%	tcgcCGAACGTGGGCTGGgcga	------vd--b--------in-
Cold	DR0907	R	-195	-182	5	-7.9	35.76	77.81%	32.90%	gacgCGCTGATGGGTTCGgaag	-----qvd-ebrh--------x
Oxidative	DR1177	D	-189	-176	9.06	-11.21	39.82	86.64%	59.61%	taggCGAAGTCGTGGTCGgaca	---k-qvd-e----------
	DR1849	D	-324	-311	7.86	-10.14	38.62	84.03%	51.71%	ccccAGAACCCGAAGTCGaggc	-t-y--dcebrhujgp-l---
	DRA0146	D	-354	-341	4.86	-7.81	35.62	77.50%	31.98%	cctaTGACCCCAAGGCCGacag	-------d----------
	DRA0145	D	-242	-229	4.19	-7.34	34.95	76.04%	27.57%	cgccCGAAGTCAAGGACTttat	-------d----------
	DR0846	D	-192	-179	4.04	-7.24	34.8	75.72%	26.58%	ccatCGAACTCAAGAGCAatca	----yqvdcebrhuj-------
	DR1546	D	-279	-266	3.3	-6.77	34.06	74.11%	21.71%	tattAGAAATTAAGTTTTccgt	----yq-d-ebr----------
	DR1208	D	-144	-131	2.91	-6.52	33.67	73.26%	19.15%	cgtgCGCTCGTATGGTTCgggg	----yqvdcebrhuj-------
	DRA0202	R	-261	-248	2.81	-6.46	33.57	73.04%	18.49%	tccgCAATGACGCCTTCGtgac	----yq-d-ebr----------
	DR1538	D	-301	-288	2.55	-6.3	33.31	72.48%	16.78%	tgccCGAAGGTGTAAACGggga	-------d-eb----gp-----
	DRA0259	D	-364	-351	2.44	-6.23	33.2	72.24%	16.05%	tgctCGAAGGCGGGCACGtcaa	----y--d-eb-huj-------
	DR1279	D	-198	-185	2.18	-6.07	32.94	71.67%	14.34%	cttgCGAAGCAAGGGGTGaggg	-t--yq-dcebrhuj--o-inx
	DR1857	D	-82	-69	1.28	-5.56	32.04	69.71%	8.42%	tcccTGACGGCGGGGTCGggcc	-------d-eb----gp-----
	DR1998	D	-325	-312	0.24	-4.99	31	67.45%	1.58%	tggtCGAAGTTCCAGCCGagcc	----y--d-eb-huj-------
Oxidative/detoxication	DRA0072	D	-90	-77	6.05	-8.66	36.81	80.09%	39.80%	tacaCCAACCTGTGGATGgcgg	a-tky-vdcebrh-----l-x
	DR1209	R	-360	-347	4.32	-7.43	35.08	76.33%	28.42%	cgcgCGACTGTGCGCTGGaggc	amtkyqvdcebr-uj---linx
	DR2623	D	-292	-279	4.3	-7.42	35.06	76.28%	28.29%	acggCGTACTCGTTGTAGaccg	amtkyqvdcebrhujgpolinx
	DR2242	D	-283	-270	3.64	-6.98	34.4	74.85%	23.95%	agtcAGAACCTGCGGACCtcgc	amtkyqvdcebr-uj---linx
	DRB0033	D	-17	-4	3.22	-6.71	33.98	73.93%	21.19%	agcaTGAATCCGAGGTGGcc	amtkyqvdcebrhujgpolinx
	DR0412	R	-291	-278	2.23	-6.1	32.99	71.78%	14.67%	gcgcCAACCCCGAGTGGGgcag	amtkyqvdcebrhujgpolinx
	DR2085	R	-147	-134	2.02	-5.98	32.78	71.32%	13.29%	agaaCGGCAACACCTTCGtttt	a-tky-vdcebrh-----l-x
	DR1982	R	-331	-318	1.68	-5.79	32.44	70.58%	11.05%	ctggCGACTGGCCGTTCCtgag	amtkyqvdcebrhujgpolinx
Detoxication	DRB0131	D	-321	-308	10.05	-12.18	40.81	88.79%	66.12%	tgcaCGAACCATTGTTCGtacc	-------dcebrhuj------x
	DRC0001	D	-328	-315	6.21	-8.8	36.97	80.44%	40.86%	ccttCGAACTTCCGATTGccct	----y--dc-b----------

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

	DRC0041	D	-205	-192	5.4	-8.19	36.16	78.68%	35.53%	cgctCAAACGGATGCTCGaaac	----y--dc-b----------
	DR1723	D	-212	-199	3.94	-7.17	34.7	75.50%	25.92%	ctgcCGAACGTCTTTCTGgcga	----y--dc-b----------
	DR2473	R	-326	-313	3.64	-6.98	34.4	74.85%	23.95%	gatcCAAGCTTGCGCTTGgctt	----y--dc-b----------
	DR2220	D	-140	-127	3.4	-6.82	34.16	74.33%	22.37%	tgcaTGGCCGCATGTGCGaacc	-------d-------------
	DR2226	D	-161	-148	2.63	-6.35	33.39	72.65%	17.30%	ggcgTGAACGTCTGATCTtcca	-------dcebrhuj------x
	DR1127	R	-330	-317	2.5	-6.27	33.26	72.37%	16.45%	cgcgAATACAGCAGTTCGcgtg	--------d-b----------
	DRA0186	R	-37	-24	2.39	-6.21	33.15	72.13%	15.72%	ctgtCGAGCATGACTTTAtcgg	----y--dc-b----------
	DR2538	R	-368	-355	2.11	-6.04	32.87	71.52%	13.88%	tgcgCGGCCAGGACTTCAgcct	----y--dc-b----------
	DR1187	R	-122	-109	0.96	-5.38	31.72	69.02%	6.32%	ctccGAATTACGCGTGCGtcgg	-------dcebrhuj------x
Toxins/general	DR2225	R	-63	-50	8.69	-10.85	39.45	85.84%	57.17%	ctgcCGGACATGCGTCCGcacc	-------dc-b----------
	DR2223	D	-357	-344	6.34	-8.9	37.1	80.72%	41.71%	ggccAGAACGCACAGTGGgctg	-------dc-b----------
	DR2221	R	-308	-295	6.11	-8.73	36.87	80.22%	40.20%	gccgCGACTTCGACTTCGggaa	-------dc-b----------
	DR2224	R	-337	-324	4.85	-7.8	35.61	77.48%	31.91%	acctCATCCACACGGTCTtgat	-------dc-b----------
Toxins	DRA0123	D	-392	-379	5.82	-8.49	36.58	79.59%	38.29%	agcaGGAAGGCGAGTTCGactt	-------d-ebh----------
	DR0136	D	-376	-363	2.41	-6.22	33.17	72.17%	15.86%	cggaCGAAATCGTTTTCTattt	-------d-ebh----------
Desiccation	DRB0118	R	-101	-88	3.69	-7.02	34.45	74.96%	24.28%	taccCGACCTTAAGTCAAgaca	-------d-------------
	DR1372	D	-155	-142	3.61	-6.97	34.37	74.78%	23.75%	gtgcTGAGCTGGAGGTCGgtgg	am-k---d--------------
	DR1172	D	-221	-208	2.74	-6.41	33.5	72.89%	18.03%	ggggCGAGAGCGAGTGCGccgc	-------d--------------
	DR0105	D	-343	-330	1.72	-5.81	32.48	70.67%	11.32%	tgacCGAGGCCGAGGCCGaaga	-------d--------------
Drugs	DR2164	D	-367	-354	8.62	-10.8	39.38	85.68%	56.71%	gttcTGGACGTATGTTTGgaca	-----qvdcebrhuj---lin-
	DRA0345	R	-328	-315	7.24	-9.6	38	82.68%	47.63%	tcacCGTGCAGGCGTTCGtgca	-------d--br----------
	DR0394	R	-206	-193	5.82	-8.49	36.58	79.59%	38.29%	tcggCGGACCTGCGCTCGctga	-------dc-br----------
	DR0433	R	-256	-243	5.66	-8.38	36.42	79.24%	37.24%	tcgcCAACCTTGAGTAAGcgcc	-------dc-br----------
	DR2104	R	-236	-223	5.58	-8.33	36.34	79.07%	36.71%	atggCCAAGATAAGTTCTccag	amtk---dcebrh----------
	DR2022	D	-205	-192	5.36	-8.16	36.12	78.59%	35.27%	ctgcCGACCTTGCGGATGtcca	amtk---dcebrh----------
	DR2234	D	-274	-261	5.19	-8.05	35.95	78.22%	34.15%	acgcTGGTCGCGTGTTCGcggg	----y-vdcebr----------
	DR0066	R	-238	-225	5.06	-7.96	35.82	77.94%	33.29%	gcggCGTCCACCCGGTCGagcg	-------dc-br----------
	DR0454	R	-131	-118	5.01	-7.91	35.77	77.83%	32.96%	gtcgCTGCCGTGGGTGCGaagc	-----qvdcebr-----o----
	DR2257	R	-68	-55	5	-7.9	35.76	77.81%	32.90%	tgagCGAGGAGGAGTTCGccga	-------d--br----------
	DR1182	D	-315	-302	4.66	-7.68	35.42	77.07%	30.66%	ttttCGATGTCGTGTACGgtgc	-------dceb----------
	DR0109	R	-386	-373	4.49	-7.55	35.25	76.70%	29.54%	acacGAAACGTTGGTACGtcgc	amtk---dcebrh----------
	DR2545	D	-197	-184	4.48	-7.55	35.24	76.68%	29.48%	tcgtCGAGCGCGAAATCGagca	a---yqvd--br-uj-------

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

	DR1419	R	-302	-289	3.91	-7.16	34.67	75.44%	25.73%	gtcaCGAACTTCGGTGGGtgcc	-------d--b----------
	DR0025	R	-53	-40	3.58	-6.95	34.34	74.72%	23.55%	ggggCAAAGGTAAGCTCTgctc	-------d--b----------
	DR2208	R	-320	-307	2.93	-6.54	33.69	73.30%	19.28%	gcgcCAACCCCAACTTCAaaga	amtk---dcebrh----------
	DR1560	D	-386	-373	2.81	-6.46	33.57	73.04%	18.49%	ggccCGAGGGCGTGGAAGcggg	----y-vdcebr----------
	DR2034	D	-69	-56	2.74	-6.41	33.5	72.89%	18.03%	gagcAGAACGCCTCTTCAtgag	-------d--b----------
	DR0455	D	-181	-168	2.67	-6.38	33.43	72.74%	17.57%	atgaCGACCTCGAAGGTGtcct	-------d--------------
	DR1695	D	-155	-142	2.67	-6.38	33.43	72.74%	17.57%	acgcCGCCCGTGTGGCCGccgg	amtk---dcebrh----------
	DR0669	D	-187	-174	2.23	-6.11	32.99	71.78%	14.67%	gcctTGAGCGCCTGGGTGccgg	-------dc-br----------
	DR1016	D	-293	-280	2.22	-6.1	32.98	71.76%	14.61%	acccTGAACGGACGGCGGgact	----yqvdceb-hujgpolinx
	DR1619	R	-313	-300	2.21	-6.09	32.97	71.74%	14.54%	gctaTGTCCAGGCGTTTGcccc	-------d--b----------
	DR0842	D	-25	-12	2.16	-6.07	32.92	71.63%	14.21%	gcccTGAACCTGAGACAGgagc	-------d--------------
	DRA0224	D	-299	-286	2.06	-6	32.82	71.41%	13.55%	ccgaCGAACGCCTGCGCCtctt	amtk---dcebrh----------
	DR0599	D	-269	-256	2.05	-6	32.81	71.39%	13.49%	cccaCGAAAGCACCCACGccgg	-------d--b----------
	DRA0241	D	-164	-151	2.01	-5.98	32.77	71.30%	13.22%	cccgAGCACGAGCGGGCGctgt	-------dc-br----------
	DRA0284	R	-121	-108	1.9	-5.91	32.66	71.06%	12.50%	ttaaCAAGTACGTGCTCGaact	---k---d-eb-h----------
	DR0009	R	-189	-176	1.78	-5.84	32.54	70.80%	11.71%	acggGGCACAGGCGCTCGagac	-------d--b----------
	DRA0194	D	-372	-359	1.47	-5.67	32.23	70.13%	9.67%	acgcCGCGCTCGCCTTTGccga	-------dc-br----------
	DR2000	R	-117	-104	1.28	-5.55	32.04	69.71%	8.42%	ctctCGGCTGGGAGATCGtcgc	-------dceb--------l--x
	DR1363	D	-398	-385	0.78	-5.29	31.54	68.62%	5.13%	agCGAACGCTGACGCGcagg	----y-vdcebr----------
Unknown	DR0559	R	-308	-295	6.18	-8.77	36.94	80.37%	40.66%	cgggCGACCTGAAGCTCGaccc	amtky-vdc-b--uj--o----
	DRA0185	D	-179	-166	3.32	-6.78	34.08	74.15%	21.84%	tagcAGAACTCCTGAAAGaaag	----yqvdce--huj------x
	DRA0065	R	-215	-202	2.95	-6.54	33.71	73.35%	19.41%	aaagCAGACGCACGATCTtcac	-----qvdcebrhujgpolinx
	DR0865	D	-169	-156	2.68	-6.38	33.44	72.76%	17.63%	ttccGGACAGCGTGGTCGaggc	a----qvdcebrhujgpo----
	DRA0243	D	-185	-172	1.21	-5.53	31.97	69.56%	7.96%	gcccGCAGCGCGTATTCGgcgg	----yq-d-eb----------

Note: Abbreviations are as in Table S4 & S5. The bases in lowercase are less conserved than those in bold uppercase.

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Supplementary Table 9 (S9): Details of SOS-box motifs predicted in the upstream sequences of *D. radiodurans* genes related to stress response through PoSSuMsearch.

Type of Stress	Gene_ID	PoSSuMsearch							Phylogenetic pattern
		Position	Score	MSS	P-value	E-value	MSSP	Predicted SOS-box	
Heat, general	DR0129	156	66	7.42E-01	3.54E-04	2.06E+01	74.16%	CGAACCTGCGGGCG	--t-yqvdcebrhujgpolinx
	DR1691	271	65	7.30E-01	4.86E-04	2.83E+01	73.03%	CGATCTCGAGTCCG	amtkyqvdcebr---------x
	DR0606	242	64	7.19E-01	6.61E-04	3.84E+01	71.91%	AGCACGTATTGTCG	----yqvdcebrhujgpolinx
	DR1114	373	64	7.19E-01	6.61E-04	3.84E+01	71.91%	TGCACTCAAGTTTT	amtkyqvdcebr---------x
	DR0126	261	60	6.74E-01	2.10E-03	1.22E+02	67.42%	CGACCCCGCGCTGG	--t--qvdcebrhujgpolinx
	DR0607	208	59	6.63E-01	2.75E-03	1.60E+02	66.29%	GGAAGTCAAGGAAG	--t-yqvdcebrhujgpolinx
	DR0128	297	56	6.29E-01	5.90E-03	3.43E+02	62.92%	CGACTTTAAGCCCG	amtkyqvdcebrhujgpolinx
	DR1424	236	54	6.07E-01	9.50E-03	5.53E+02	60.67%	GGAACCCGGAGCCG	--t-yqvdcebrhujgpolinx
Heat	DR2056	288	60	6.74E-01	2.10E-03	1.22E+02	67.42%	CGGTCGTTGGTTCG	-------dce------------
	DR1940	100	49	5.51E-01	2.75E-02	1.60E+03	55.06%	CCCCCCGGGGTTCG	-------dce------------
General	DR1974	139	70	7.87E-01	9.09E-05	5.29E+00	78.65%	CGATCTGCTGTTCG	amtk-q-dcebrh--------x
	DR2132	130	69	7.75E-01	1.30E-04	7.58E+00	77.53%	CGAACTGATGCAGG	--t-yqvdcebrhujgpolinx
	DR1046	165	68	7.64E-01	1.83E-04	1.06E+01	76.40%	CGAACCAGCATTTG	-------d-ebr----------
	DR1308	155	68	7.64E-01	1.83E-04	1.06E+01	76.40%	CGAAAGCCTGATCG	--t-yqvdcebrhuj--olinx
	DR1005	93	66	7.42E-01	3.54E-04	2.06E+01	74.16%	TGAACGCATGGACA	----yqvdcebrhujgpolinx
	DR1621	4	66	7.42E-01	3.54E-04	2.06E+01	74.16%	CGAACGACTGCGCG	----yqvdcebrhuj--olinx
	DR1551	100	65	7.30E-01	4.86E-04	2.83E+01	73.03%	CGAACTCGCCCTTG	----yqvd-eb-hujgpolinx
	DR0300	5	64	7.19E-01	6.61E-04	3.84E+01	71.91%	CGAACACACCGATG	-----qvdceb-huj--olinx
	DR2363	78	64	7.19E-01	6.61E-04	3.84E+01	71.91%	CGCCCGCGAGTTCG	am-kyq-dcebrhujgpol---
	DR1973	355	63	7.08E-01	8.95E-04	5.21E+01	70.79%	CGACCAAGTGATCG	-----qvdceb-huj--olinx
	DRA0290	205	63	7.08E-01	8.95E-04	5.21E+01	70.79%	CAACCGCATAGTCG	-----qvdceb-huj--olinx
	DR0349	298	62	6.97E-01	1.21E-03	7.01E+01	69.66%	TGCACTCAGGTTTT	--t-yqvdcebrhuj--olinx
	DR1491	361	62	6.97E-01	1.21E-03	7.01E+01	69.66%	TGAAGGCAAGGCCG	----yqvd-eb-hujgpolinx
	DR0583	125	62	6.97E-01	1.21E-03	7.01E+01	69.66%	CGGGCGCAATTTTG	-----qvdcebrhuj---linx
	DR0202	304	61	6.85E-01	1.60E-03	9.30E+01	68.54%	CGAACTGTGGAACG	amtk-qvdcebrhuj--olinx
	DR0139	54	61	6.85E-01	1.60E-03	9.30E+01	68.54%	TGAACGTGTCTTTT	amtkyq-dcebrhujgpo-inx

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

	DR1199	49	61	6.85E-01	1.60E-03	9.30E+01	68.54%	CGTTCCTAACTTCG	amtk-q-dcebrh--------x
	DR0588	109	60	6.74E-01	2.10E-03	1.22E+02	67.42%	CGACCGCATCTACC	--t-yqvdcebrhujgpolinx
	DR1117	94	60	6.74E-01	2.10E-03	1.22E+02	67.42%	CGCCCACGTGGTCG	-------d-ebr----------
	DR0984	157	60	6.74E-01	2.10E-03	1.22E+02	67.42%	CTACCACGTGGTCG	----yqvdcebrhujgpolinx
	DR0327	26	59	6.63E-01	2.75E-03	1.60E+02	66.29%	GGACCGTAGGTCAG	-m-k-qvdcebrhujgpolinx
	DR1004	128	59	6.63E-01	2.75E-03	1.60E+02	66.29%	CAAACGTCAGGATG	am-kyq-dcebrhujgpol---
	DR0194	90	58	6.52E-01	3.57E-03	2.08E+02	65.17%	CGGGCGGCTGGTCG	--t-yqvdcebrhuj--olinx
	DR1972	109	55	6.18E-01	7.52E-03	4.37E+02	61.80%	CGCTCGCCTGGGCG	----yqvdcebrhujgpolinx
	DR1020	11	55	6.18E-01	7.52E-03	4.37E+02	61.80%	TGAGCGCCAGGATG	--t-yqvdcebrhujgpolinx
	DR0491	322	55	6.18E-01	7.52E-03	4.37E+02	61.80%	CGAGCACGGGGTCA	--t-yqvdcebrhuj--olinx
	DR1105	51	54	6.07E-01	9.50E-03	5.53E+02	60.67%	CGAACATGGCCGCG	----yqvdcebrhuj--olinx
	DR0745	116	54	6.07E-01	9.50E-03	5.53E+02	60.67%	CTACCGCGTGGAAG	----yqvd-eb-hujgpolinx
	DR1599	209	54	6.07E-01	9.50E-03	5.53E+02	60.67%	TGATCGTGCGGGCG	-m-k-qvdcebrhujgpolinx
	DR2189	116	53	5.96E-01	1.19E-02	6.92E+02	59.55%	CTACCTTGGGGTGG	-------d--b----------
	DR0646	135	52	5.84E-01	1.48E-02	8.60E+02	58.43%	CGCCCGCGAGGTCC	--t-yqvdcebrhuj--olinx
	DR1756	44	51	5.73E-01	1.83E-02	1.06E+03	57.30%	AAACCCAGCGTTAG	amtkyq-dcebrhuj-------
	DR1507	139	49	5.51E-01	2.75E-02	1.60E+03	55.06%	CTGCCTCGCGCTCG	----yqvdcebrhuj--olinx
	DR1832	233	47	5.28E-01	4.04E-02	2.35E+03	52.81%	CGGATGCGGGTCCA	amtkyq-dcebrhuj-------
	DR0190	22	41	4.61E-01	1.09E-01	6.36E+03	46.07%	GGTCCACGCTCTCG	--t-yqvdcebrhuj--olinx
Starvation	DR0662	128	70	7.87E-01	9.09E-05	5.29E+00	78.65%	AGAACGTATGAATG	-------d-ebr----------
	DR1838	281	58	6.52E-01	3.57E-03	2.08E+02	65.17%	CGCACCCGCGAACG	-------dceb-huj-ol--x
	DR0417	106	56	6.29E-01	5.90E-03	3.43E+02	62.92%	CGTCATAGTGTTCG	-------dceb-huj-ol--x
	DR0416	342	56	6.29E-01	5.90E-03	3.43E+02	62.92%	CGTCATAGTGTTCG	-------dcebrhujgpo----
	DRB0092	217	55	6.18E-01	7.52E-03	4.37E+02	61.80%	GGAAAGTGGGTTCA	-------d-ebr----------
	DR2263	252	51	5.73E-01	1.83E-02	1.06E+03	57.30%	CAACCGCGCGGTCA	-------dce------------
Osmotic	DR1666	43	70	7.87E-01	9.09E-05	5.29E+00	78.65%	CGAACTCATCTTGT	a------debr-hj--o----
	DRA0138	82	66	7.42E-01	3.54E-04	2.06E+01	74.16%	CGACCGTATCTGCG	-------dcebrh--------
	DRA0139	160	64	7.19E-01	6.61E-04	3.84E+01	71.91%	TGAACCCGAGGACG	amtk-qydceb-h-gpol---
	DR1668	251	62	6.97E-01	1.21E-03	7.01E+01	69.66%	CCCCCGCAAGTTCG	amtk-qvdcebrhuj-ol--x
	DR1995	244	61	6.85E-01	1.60E-03	9.30E+01	68.54%	CGAGCGCCTGGAAG	amtk-q-dcebrh-gpol---

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

	DR1667	269	61	6.85E-01	1.60E-03	9.30E+01	68.54%	CGCGCTGATGTTCC	a------debr-hj--o----
	DRB0088	93	60	6.74E-01	2.10E-03	1.22E+02	67.42%	CCAACGCCAAGTTG	-------dce-r----------
	DRA0135	328	59	6.63E-01	2.75E-03	1.60E+02	66.29%	CGCACCAGGGTCCG	a------debr-hj-------
	DR0211	212	50	5.62E-01	2.25E-02	1.31E+03	56.18%	GGGCCTCGCGTCCG	amtk-qvdcebrhuj-ol--x
	DR2422	84	48	5.39E-01	3.34E-02	1.94E+03	53.93%	CGTCGGTGCGCTCG	a------debr-hj--o----
	DRA0137	379	43	4.83E-01	8.03E-02	4.67E+03	48.31%	CCTCCATAACTTCC	amtk-qydceb-h-gpol---
Phage	DR1473	354	54	6.07E-01	9.50E-03	5.53E+02	60.67%	CCTCCTCAGATTCG	-----q-dceb----------
Alkaline	DR2068	248	67	7.53E-01	2.55E-04	1.48E+01	75.28%	CGAACGTGGGCTGG	------vd--b--------in-
	DR0389	339	67	7.53E-01	2.55E-04	1.48E+01	75.28%	CGAGCTCACGGACG	------vd--b--------in-
Cold	DR0907	205	50	5.62E-01	2.25E-02	1.31E+03	56.18%	CGCTGATGGGTTCG	-----qvd-ebrh--------x
Oxidative	DR1177	211	71	7.98E-01	6.22E-05	3.62E+00	79.78%	CGAAGTCGTGGTCG	---k-qvd-e----------
	DR0846	208	65	7.30E-01	4.86E-04	2.83E+01	73.03%	CGAACTCAAGAGCA	-t-y--dcebrhujgp-l---
	DR1849	76	64	7.19E-01	6.61E-04	3.84E+01	71.91%	AGAACCCGAAGTCG	-------d----------
	DR1546	121	63	7.08E-01	8.95E-04	5.21E+01	70.79%	AGAAATTAAGTTTT	-------d----------
	DRA0146	46	61	6.85E-01	1.60E-03	9.30E+01	68.54%	TGACCCCAAGGCCG	----yqvdcebrhuj-------
	DRA0145	158	60	6.74E-01	2.10E-03	1.22E+02	67.42%	CGAAGTCAAGGACT	----yq-d-ebr----------
	DR1279	202	60	6.74E-01	2.10E-03	1.22E+02	67.42%	CGAAGCAAGGGGTG	----yqvdcebrhuj-------
	DR1208	256	59	6.63E-01	2.75E-03	1.60E+02	66.29%	CGCTCGTATGGTTC	----yq-d-ebr----------
	DRA0259	36	58	6.52E-01	3.57E-03	2.08E+02	65.17%	CGAAGGCGGGCACG	-------d-eb----gp-----
	DR1538	99	56	6.29E-01	5.90E-03	3.43E+02	62.92%	CGAAGGTGTAAACG	----y--d-eb-huj-------
	DR1857	318	53	5.96E-01	1.19E-02	6.92E+02	59.55%	TGACGGCGGGGTCG	-t--yq-dcebrhuj--o-inx
	DR1998	75	52	5.84E-01	1.48E-02	8.60E+02	58.43%	CGAAGTTCCAGCCG	-------d-eb----gp-----
	DRA0202	139	45	5.06E-01	5.76E-02	3.35E+03	50.56%	CAATGACGCCTTCG	----y--d-eb-huj-------
Oxidative/detoxication	DRA0072	310	62	6.97E-01	1.21E-03	7.01E+01	69.66%	CCAACCTGTGGATG	a-tky-vdcebrh-----l-x
	DR2623	108	61	6.85E-01	1.60E-03	9.30E+01	68.54%	CGTACTCGTTGTAG	amtkyqvdcebr-uj---linx
	DRB0033	383	58	6.52E-01	3.57E-03	2.08E+02	65.17%	TGAATCCGAGGTGG	amtkyqvdcebrhujgpolinx
	DR2242	117	54	6.07E-01	9.50E-03	5.53E+02	60.67%	AGAACCTGCGGACC	amtkyqvdcebr-uj---linx
	DR1209	40	52	5.84E-01	1.48E-02	8.60E+02	58.43%	CGACTGTGCGCTGG	amtkyqvdcebrhujgpolinx
	DR0412	109	52	5.84E-01	1.48E-02	8.60E+02	58.43%	CAACCCCGAGTGGG	amtkyqvdcebrhujgpolinx
	DR2085	253	51	5.73E-01	1.83E-02	1.06E+03	57.30%	CGGCAACACCTTCG	a-tky-vdcebrh-----l-x

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

	DR1982	69	50	5.62E-01	2.25E-02	1.31E+03	56.18%	CGACTGGCCGTTCC	amtkyqvdcebrhujgpolinx
Detoxication	DRB0131	79	78	8.76E-01	3.39E-06	1.97E-01	87.64%	CGAACCATTGTTCG	-------dcebrhuj------x
	DRC0041	195	71	7.98E-01	6.22E-05	3.62E+00	79.78%	CAAACGGATGCTCG	----y--dc-b----------
	DRC0001	72	70	7.87E-01	9.09E-05	5.29E+00	78.65%	CGAACTTCCGATTG	----y--dc-b----------
	DR1723	188	66	7.42E-01	3.54E-04	2.06E+01	74.16%	CGAACGTCTTTCTG	----y--dc-b----------
	DR2226	239	62	6.97E-01	1.21E-03	7.01E+01	69.66%	TGAACGTCTGATCT	----y--dc-b----------
	DR2220	260	61	6.85E-01	1.60E-03	9.30E+01	68.54%	TGGCCGCATGTGCG	-------d-------------
	DR2473	74	55	6.18E-01	7.52E-03	4.37E+02	61.80%	CAAGCTTGCGCTTG	-------dcebrhuj------x
	DRA0186	363	52	5.84E-01	1.48E-02	8.60E+02	58.43%	CGAGCATGACTTTA	--------d-b----------
	DR1127	70	51	5.73E-01	1.83E-02	1.06E+03	57.30%	AATACAGCAGTTCG	----y--dc-b----------
	DR2538	32	42	4.72E-01	9.40E-02	5.47E+03	47.19%	CGGCCAGGACTTCA	----y--dc-b----------
	DR1187	278	39	4.38E-01	1.46E-01	8.48E+03	43.82%	GAATTACGCGTGCG	-------dcebrhuj------x
Toxins/general	DR2223	43	66	7.42E-01	3.54E-04	2.06E+01	74.16%	AGAACGCACAGTGG	-------dc-b----------
	DR2225	337	60	6.74E-01	2.10E-03	1.22E+02	67.42%	CGGACATGCGTCCG	-------dc-b----------
	DR2221	92	57	6.40E-01	4.60E-03	2.68E+02	64.04%	CGACTTCGACTTCG	-------dc-b----------
	DR2224	63	46	5.17E-01	4.84E-02	2.82E+03	51.69%	CATCCACACGGTCT	-------dc-b----------
Toxins	DRA0123	8	66	7.42E-01	3.54E-04	2.06E+01	74.16%	GGAAGGCGAGTTCG	-------d-ebh----------
	DR0136	24	60	6.74E-01	2.10E-03	1.22E+02	67.42%	CGAAATCGTTTTCT	-------d-ebh----------
Desiccation	DR1372	245	59	6.63E-01	2.75E-03	1.60E+02	66.29%	TGAGCTGGAGGTCG	-------d------------
	DRB0118	299	59	6.63E-01	2.75E-03	1.60E+02	66.29%	CGACCTTAAGTCAA	am-k---d-------------
	DR1172	179	57	6.40E-01	4.60E-03	2.68E+02	64.04%	CGAGAGCGAGTGCG	-------d-------------
	DR0105	57	52	5.84E-01	1.48E-02	8.60E+02	58.43%	CGAGGCCGAGGCCG	-------d-------------
Drugs	DR2164	33	74	8.31E-01	1.96E-05	1.14E+00	83.15%	TGGACGTATGTTTG	-----qvdcebrhuj---lin-
	DR1419	98	65	7.30E-01	4.86E-04	2.83E+01	73.03%	CGAACTTCGGTGGG	-------d--br----------
	DR0394	194	63	7.08E-01	8.95E-04	5.21E+01	70.79%	CGGACCTGCGCTCG	-------dc-br----------
	DR2234	126	62	6.97E-01	1.21E-03	7.01E+01	69.66%	TGGTCGCGTGTTCG	-------dc-br----------
	DR2034	331	62	6.97E-01	1.21E-03	7.01E+01	69.66%	AGAACGCCTCTTCA	amtk---dcebrh----------
	DR1016	107	61	6.85E-01	1.60E-03	9.30E+01	68.54%	TGAACGGACGGCGG	amtk---dcebrh----------
	DRA0224	101	61	6.85E-01	1.60E-03	9.30E+01	68.54%	CGAACGCCTGCGCC	----y-vdcebr----------
	DR1182	85	61	6.85E-01	1.60E-03	9.30E+01	68.54%	CGATGTCGTGTACG	-------dc-br----------

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

	DR2545	203	60	6.74E-01	2.10E-03	1.22E+02	67.42%	CGAGCGCGAAATCG	-----qvdcebr-----o----
	DR2022	195	59	6.63E-01	2.75E-03	1.60E+02	66.29%	CGACCTTGCGGATG	-------d--br----------
	DR1363	2	58	6.52E-01	3.57E-03	2.08E+02	65.17%	CGAACGCTGACGCG	-------dceb----------
	DR2104	164	58	6.52E-01	3.57E-03	2.08E+02	65.17%	CCAAGATAAGTTCT	amtk---dcebrh----------
	DR0599	131	58	6.52E-01	3.57E-03	2.08E+02	65.17%	CGAAAGCACCCACG	a---yqvd--br-uj-------
	DR2257	332	57	6.40E-01	4.60E-03	2.68E+02	64.04%	CGAGGAGGAGTTCG	-------d--b----------
	DR0669	213	57	6.40E-01	4.60E-03	2.68E+02	64.04%	TGAGCGCCTGGGTG	-------d--b----------
	DR0842	375	57	6.40E-01	4.60E-03	2.68E+02	64.04%	TGAACCTGAGACAG	amtk---dcebrh----------
	DRA0345	72	56	6.29E-01	5.90E-03	3.43E+02	62.92%	CGTGCAGGCGTTCG	----y-vdcebr----------
	DR0109	14	56	6.29E-01	5.90E-03	3.43E+02	62.92%	GAAACGTTGGTACG	-------d--b----------
	DR0066	162	55	6.18E-01	7.52E-03	4.37E+02	61.80%	CGTCCACCCGGTCG	-------d--------------
	DRA0194	28	55	6.18E-01	7.52E-03	4.37E+02	61.80%	CGCGCTCGCCTTTG	amtk---dcebrh----------
	DR1695	245	55	6.18E-01	7.52E-03	4.37E+02	61.80%	CGCCCGTGTGGCCG	-------dc-br----------
	DR0025	347	55	6.18E-01	7.52E-03	4.37E+02	61.80%	CAAAGGTAAGCTCT	----yqvdceb-hujgpolinx
	DR0455	219	54	6.07E-01	9.50E-03	5.53E+02	60.67%	CGACCTCGAAGGTG	-------d--b----------
	DR1560	14	54	6.07E-01	9.50E-03	5.53E+02	60.67%	CGAGGGCGTGGAAG	-------d--------------
	DR0433	144	54	6.07E-01	9.50E-03	5.53E+02	60.67%	CAACCTTGAGTAAG	amtk---dcebrh----------
	DR2208	80	54	6.07E-01	9.50E-03	5.53E+02	60.67%	CAACCCCAACTTCA	-------d--b----------
	DRA0241	236	52	5.84E-01	1.48E-02	8.60E+02	58.43%	AGCACGAGCGGGCG	-------dc-br----------
	DR0009	211	52	5.84E-01	1.48E-02	8.60E+02	58.43%	GGCACAGGCGCTCG	---k---d-eb-h----------
	DRA0284	279	50	5.62E-01	2.25E-02	1.31E+03	56.18%	CAAGTACGTGCTCG	-------d--b----------
	DR1619	87	50	5.62E-01	2.25E-02	1.31E+03	56.18%	TGTCCAGGCGTTTG	-------dc-br----------
	DR0454	269	46	5.17E-01	4.84E-02	2.82E+03	51.69%	CTGCCGTGGGTGCG	-------dceb--------l--x
	DR2000	283	46	5.17E-01	4.84E-02	2.82E+03	51.69%	CGGCTGGGAGATCG	----y-vdcebr----------
Unknown	DR0559	92	68	7.64E-01	1.83E-04	1.06E+01	76.40%	CGACCTGAAGCTCG	amtky-vdc-b--uj--o----
	DRA0185	221	62	6.97E-01	1.21E-03	7.01E+01	69.66%	AGAACTCCTGAAAG	----yqvdce--huj------x
	DR0865	231	57	6.40E-01	4.60E-03	2.68E+02	64.04%	GGACAGCGTGGTCG	-----qvdcebrhujgpolinx
	DRA0065	185	55	6.18E-01	7.52E-03	4.37E+02	61.80%	CAGACGCACGATCT	a----qvdcebrhujgpo----
	DRA0243	215	55	6.18E-01	7.52E-03	4.37E+02	61.80%	GCAGCGCGTATTCG	----yq-d-eb----------

Note: Abbreviations are as in Table S4 & S6.

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Supplementary Table 10 (S10): List of *D. radiodurans* genes related to some unusual predicted operons with their operon cluster number, best hit details & comments.

Cluster No.	Gene cluster	Protein description	Best hit: species and GI number	Comments
1	DR0298	Deoxypurine kinase subunit, YAAF	Bacillus subtilis (586859)	Clear case of gene exchange with gram-positive bacteria; essential enzymes for biosynthesis of deoxyribonucleotides
	DR0299	Deoxypurine kinase subunit, YAAG	Bacillus subtilis (586860)	
2	DR0398	Alkaline-shock-like protein	Bacillus subtilis (2337812)	Clear case of gene exchange with gram-positive bacteria; possibly stress related; conserved operon in B. subtilis and T. maritima; all three bacteria encode an additional copy of alkaline-shock-like protein
	DR0390	Uncharacterized protein, YloV ortholog	Bacillus subtilis (2337813)	
3	DR0544	Highly conserved membrane transporter	Pyrococcus horikoshii (3256923)	Likely gene exchange with archaea; possibly related to the multidrug resistance system
	DR0545	Small conserved membrane protein, possibly involved in multidrug resistance	Methanobacterium thermoautotrophicum (2621904)	
4	DR0674	Argininosuccinate synthase, ArgG	Thermotoga maritima (4982360)	Acetyltransferase cluster disrupts ArgG/ArgH operon present in some other bacteria
	DR0675	Amino acetyltransferase	Synechocystis (1651699)	
	DR0676	Amino acetyltransferase, related to phosphinothricin acetyltransferase	Escherichia coli (1742360)	
	DR0677	Amino acetyltransferase	Salmonella enterica serovar Typhimurium (586786)	
5	DR0678	Argininosuccinate lyase (ArgH)	Bacillus subtilis (2293243)	DR0681 and DR0683 may have evolved by internal duplication; DR0679 and DR0680 together may comprise a mobile element; this pair of genes is present twice more in the *D. radiodurans* genome
	DR0679	Small nucleotidyltransferase	Synechocystis (1653122)	
	DR0680	Uncharacterized protein next to small nucleotidyltransferases	Synechocystis (1652090)	
	DR0681	Amino acetyltransferase	Aquifex aeolicus (2983780)	

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

	DR0682	Amino acetyltransferase	Salmonella enterica serovar Typhimurium (586786)	
	DR0683	Amino acetyltransferase	Aquifex aeolicus (2983780)	
6	DR0689	Uracil-DNA glycosylase (Ung)	Human (137031)	Likely horizontal transfer from a eukaryote or a eukaryotic virus
	DR0690	Topoisomerase IB	Orf virus (521138)	
7	DR0796	Amino acetyltransferase	Bacillus subtilis (1881232)	Acetyltransferase cluster; DR0796 and DR0797 are possible products of internal duplication
	DR0797	Amino acetyltransferase	Synechocystis (1651699)	
	DR0798	Amino acetyltransferase	Bacillus subtilis (1881232)	
8	DR0853	Rab/Ras family small GTPase	Myxococcus xanthus (94524)	Potential operon conserved also in Thermus, Myxococcus and archaea; involved in gliding motility in Myxococcus
	DR0854	Protein associated with a GTPase	Myxococcus xanthus (94525)	
9	DR0861	Phytoene dehydrogenase, CRTI	Flavobacterium (1842244)	Carotenoid biosynthesis genes; Possibly involved in pigment biosynthesis in Deinococcus
	DR0862	Phytoene synthase, CRTB	Thermus thermophilus (585011)	
10	DR0993	Uncharacterized protein associated with GTPase	Methanococcus jannaschii (1591982)	Cluster of genes that are expanded in Deinococcus and encode uncharacterized small proteins often associated with Ras/Rab family GTPase (see also DR0853-DR0854 and DR2180-DR2181)
	DR0994	Uncharacterized protein associated with GTPase	Distantly related to the family of GTPase-associated proteins	
	DR0995	Uncharacterized protein associated with GTPase	Aquifex aeolicus (2984135)	
11	DR1175	N-terminal CheY family domain +C-terminal histidine kinase	Mycobacterium tuberculosis (2960188)	Signal transduction system; proteins with modified domain architectures compared to M. tuberculosis and Synechocystis
	DR1174	Histidine kinase with 3 PAS +3 PAC + GAF domains	Synechocystis (1652132)	
12	DR1232	Pilin IV-like secreted protein	Psedomonas putida (544344)	Pilin IV cluster also including a GTPase with a possible regulatory role; probably responsible for DNA transformation
	DR1233	Pilin IV-like secreted protein	Legionella pneumophila (3002996)	
	DR1234	Pilin IV-like secreted protein	Klebsiella oxytoca (131598)	
	DR1235	Dynamin-like GTPase	Arabidopsis thaliana	

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

			(4587579)	
13	DR1596	Glucose-6-phosphate 1-dehydrogenase	Synechocystis (2494656)	Clear case of gene exchange with cyanobacteria
	DR1597	OPCA, OxPPCycle gene, involved in assembly of glucose-6-phosphate 1-dehydrogenase	Synechocystis (2498703)	
	DR1641	DinB/YfiT superfamily protein	Both distantly similar to Bacillus subtilis (2633163)	
	DR1642	DinB/YfiT superfamily protein		
14	DR1928	Glycerol kinase (GlpK)	Borrelia burgdorferi (2688136)	Clear case of gene exchange with spirochetes
	DR1929	Glycerol uptake facilitator (GlpK)	Borrelia burgdorferi (2688137)	
15	DR2180	Uncharacterized protein associated with GTPase	Aquifex aeolicus (2984135)	Clear case of gene exchange with thermophiles (See above)
	DR2181	RAB/RAS-like small bacterial GTPase, inactivated	Aquifex aeolicus (2984130)	
16	DR2220	Tellurium resistance protein (TerB)	Plasmid R478 (950680)	Tellurium resistance gene cluster; probable acquisition of a plasmid fragment; stress response-related genes; operon is probably disrupted by a transposon
	DR2221	Tellurium resistance, member of Dictyostelium-type cAMP-binding protein family	Alcaligenes sp. (135597)	
	DR2222	Transposase	No significant similarity	
	DR2223	Tellurium resistance, member of Dictyostelium-type cAMP-binding protein family	Alcaligenes sp. (78205)	
	DR2224	Tellurium resistance, member of Dictyostelium-type cAMP-binding protein family	Plasmid R478 (1181183)	
	DR2225	Tellurium resistance, member of Dictyostelium-type cAMP-binding protein family	Plasmid R478 (950682)	
	DR2226	Tellurium resistance, membrane protein (TerC)	Mycobacterium tuberculosis (2105065)	
17	DR2254	Amino-acetyltransferase	Streptomyces coelicolor (5531439)	Acetyltransferase cluster; these proteins are likely to be a product of internal

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

	DR2255	Amino-acetyltransferase	Streptomyces coelicolor (5531439)	duplication
18	DR2311	Uncharacterized protein, YeiN ortholog	Escherichia coli (465602)	Among bacteria, YeiN orthologs are present only in Deinococcus and gamma proteobacteria; in eukaryotes, their counterparts are fused with the kinase gene; probable gene exchange with proteobacteria
	DR2312	RBSK family ribokinase, YeiI ortholog fused to HTH domain	Escherichia coli (2507177)	
19	DRA0231	Oxidoreductase	Escherichia coli (2495497)	Highly conserved paralogous gene cluster is located next to this one in the chromosome (DRA0235-37), but in the opposite orientation
	DRA0232	Flavoprotein dehydrogenase	Escherichia coli (2495498)	
	DRA0233	Dehydrogenase, iron sulfur protein	Escherichia coli (2495499)	
20	DRA0331	Von Willebrand factor A domain, Mg2+ binding	Synechocystis sp. (2496792)	Serine/threonine protein kinase-based regulatory system
	DRA0332	PKN2 family serine threonine kinase	Anabaena sp. (1709645)	
	DRA0333	Zn-finger and FHA domain-containing protein, ortholog of cyanobacterial FraH	Anabaena sp. (556608)	
	DRA0334	Inactive kinase +PP2C phosphoprotein phosphatase	Mycobacterium tuberculosis (1552573)	
21	DRB0143	AAA superfamily NTPase related to 5-methylcytosine-specific restriction enzyme subunit McrB	Escherichia coli (1790805)	Mcr operon is present only in E. coli and D. radiodurans; a clear case of horizontal gene exchange
	DRB0144	Homolog of the McrC subunit of the McrBC restriction-modification system	Escherichia coli (1790804)	

Note: A gene cluster was considered a likely operon if the genes were localized on the same DNA strand and the distance between them was less than 100 bp. Genes are clustered largely based on reference no. 12, with modifications.

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Supplementary Table 11 (S11): Details of SOS-box motifs predicted in the upstream sequences of *D. radiodurans* genes related to some unusual predicted operons through RSAT webserver.

Cluster No.	Gene cluster	strand	start	end	Predicted SOS-box	score	ln(p)	obv score	MSSP	RMMP
1	DR0298	R	-330	-317	tgcgCGACGCCGAGTTCGacct	6.32	-8.87	37.08	80.68%	41.58%
	DR0299	D	-101	-88	tgctCAAGCGCACGGTGGgcta	3.39	-6.82	34.15	74.30%	22.30%
2	DR0398	R	-395	-382	ccacCAACGACGCGTTCGccaa	8.92	-11.09	39.68	86.34%	58.69%
	DR0390	D	-368	-355	acggCGAGCTCACGGACGcggg	8.05	-10.3	38.81	84.44%	52.96%
3	DR0544	D	-278	-265	acgaCGAACTCAAGGGCAaagg	6.44	-8.96	37.2	80.94%	42.37%
	DR0545	D	-63	-50	ccgcACAACTCAAGCTCTtttt	1.55	-5.71	32.31	70.30%	10.20%
4	DR0674	R	-307	-294	gggaCGAACATCGGTCTTggtg	2.78	-6.44	33.54	72.98%	18.29%
	DR0675	R	-269	-256	ataaCGGCTTCTGGTTCGcccc	5.06	-7.96	35.82	77.94%	33.29%
	DR0676	D	-193	-180	ctggTCAACGTCTGGACGcacc	3.31	-6.77	34.07	74.13%	21.78%
	DR0677	D	-156	-143	acggCGAAATCGTGGGCGgcgc	5.87	-8.55	36.63	79.70%	38.62%
5	DR0678	R	-255	-242	atgaAGATTATGCGTACAgcag	1.04	-5.42	31.8	69.19%	6.84%
	DR0679	D	-190	-177	acgcCGAACTGAAAGCCGccca	4.99	-7.89	35.75	77.79%	32.83%
	DR0680	R	-113	-100	tggcCTTCTTTGAGATCGcgcc	4.6	-7.62	35.36	76.94%	30.26%
	DR0681	R	-79	-66	tcctCGATTGGACGTTCTcctg	3.15	-6.67	33.91	73.78%	20.72%
	DR0682	R	-202	-189	gcggCGACCACACCGTCTacct	3.5	-6.89	34.26	74.54%	23.03%
	DR0683	D	-350	-337	acggCGAAGCGGTGGGCGccgt	2.83	-6.47	33.59	73.09%	18.62%
6	DR0689	R	-395	-382	gcggCCAACATGCGCTTTgccg	3.15	-6.67	33.91	73.78%	20.72%
	DR0690	R	-128	-115	tcagCGAACAGTATTTCTtcgg	4.37	-7.48	35.13	76.44%	28.75%
7	DR0796	D	-302	-289	cccgCGAAACCGAAGGCGacga	2.92	-6.53	33.68	73.28%	19.21%
	DR0797	D	-230	-217	gcaaCGAACGGATGGTTCgcgc	7.15	-9.54	37.91	82.48%	47.04%
	DR0798	D	-290	-277	tcgtCGGGCGTGAGGTCGaagg	5.03	-7.93	35.79	77.87%	33.09%
8	DR0853	D	-215	-202	aactCGAACCCATTCTCGacga	8.2	-10.43	38.96	84.77%	53.95%
	DR0854	D	-173	-160	gctcCGAGCGTCAGGAGGgcag	1.79	-5.85	32.55	70.82%	11.78%
9	DR0861	R	-206	-193	ccggCTACGACAATTTCGggcg	3.26	-6.73	34.02	74.02%	21.45%
	DR0862	R	-271	-258	tctcCGACCTTACGGGCCagcg	3	-6.57	33.76	73.46%	19.74%
10	DR0993	D	-193	-180	cgcaCGACGGAGCGGTTGcaag	2.81	-6.46	33.57	73.04%	18.49%
	DR0994	D	-327	-314	tcacCCAACTTGTGACCGatgt	2.83	-6.47	33.59	73.09%	18.62%
	DR0995	D	-208	-195	gcccAGAAATGATGTTCTgaac	2.39	-6.21	33.15	72.13%	15.72%
11	DR1175	D	-284	-271	gcttCGATCCCAAATACAccga	3.11	-6.64	33.87	73.69%	20.46%
	DR1174	R	-300	-287	ttccCGGCTTCGACTTCGgaca	2.92	-6.53	33.68	73.28%	19.21%
12	DR1232	D	-194	-181	tttcGGAACGTACGGCAGcgga	4.62	-7.65	35.38	76.98%	30.40%
	DR1233	R	-235	-222	gaacCACCTATAGGTTCGacac	7.58	-9.9	38.34	83.42%	49.87%
	DR1234	D	-271	-258	aactTGAACTCGCGTGCCgcgc	3.47	-6.88	34.23	74.48%	22.83%
	DR1235	R	-362	-349	cgaaCAACGACGCGGTCTgggt	3.63	-6.98	34.39	74.83%	23.88%
13	DR1596	R	-216	-203	ccatGCGCCGCGCGTTCGgcgg	2.12	-6.05	32.88	71.54%	13.95%
	DR1597	D	-209	-196	cctgGGAAGGCGCGGACGcggc	3.2	-6.71	33.96	73.89%	21.05%
	DR1641	R	-269	-256	tcctCGCCCTCGACATCGgcgg	1.72	-5.81	32.48	70.67%	11.32%
	DR1642	R	-232	-219	tggcCGCAGGTGGGTGCGggcg	3.65	-6.98	34.41	74.87%	24.01%
14	DR1928	R	-109	-96	gtgtGGACCCCGAGTTCAaccg	4.22	-7.37	34.98	76.11%	27.76%
	DR1929	R	-138	-125	cttaCCTCGGGGCGTTCGtctc	1.93	-5.92	32.69	71.13%	12.70%
15	DR2180	D	-129	-116	cgacCGAACAGGTGATTGgcct	4.29	-7.41	35.05	76.26%	28.23%
	DR2181	R	-271	-258	gtgcCGTCGGCGCGCTCGgtgg	3.44	-6.86	34.2	74.41%	22.63%
16	DR2220	D	-140	-127	tgcaTGGCCGCATGTGCGaacc	3.4	-6.82	34.16	74.33%	22.37%
	DR2221	R	-308	-295	gccgCGACTTCGACTTCGggaa	6.11	-8.73	36.87	80.22%	40.20%
	DR2222	D	-386	-373	gcccCGAAGGAGCCGTCGtcca	3.2	-6.7	33.96	73.89%	21.05%
	DR2223	D	-357	-344	ggccAGAACGCACAGTGGgctg	6.34	-8.9	37.1	80.72%	41.71%
	DR2224	R	-337	-324	acctCATCCACACGGTCTtgat	4.85	-7.8	35.61	77.48%	31.91%
	DR2225	R	-63	-50	ctgcCGGACATGCGTCCGcacc	8.69	-10.85	39.45	85.84%	57.17%
	DR2226	D	-161	-148	ggcgTGAACGTCTGATCTtcca	2.63	-6.35	33.39	72.65%	17.30%
17	DR2254	D	-255	-242	acggCGAAACATTGGTCGgcct	5.69	-8.41	36.45	79.31%	37.44%
	DR2255	D	-46	-33	gcggAGGAATCAGATTCGccta	0.83	-5.31	31.59	68.73%	5.46%
18	DR2311	D	-323	-310	gcgcCGAACTCGTCTGGGtgcg	3.6	-6.96	34.36	74.76%	23.69%
	DR2312	R	-204	-191	acggCGTCCGCAGGGTCAgcag	4.04	-7.24	34.8	75.72%	26.58%
19	DRA0231	R	-336	-323	tgcgCGACCGTTCGTCCTacgc	6.61	-9.1	37.37	81.31%	43.49%
	DRA0232	D	-93	-80	tgcgCGAACGCATGAGCGgtaa	8.96	-11.11	39.72	86.42%	58.95%
	DRA0233	R	-27	-14	ccgcCTAACTTTCGCTCCatga	4.45	-7.52	35.21	76.61%	29.28%
20	DRA0331	R	-333	-320	tgagCGAATACACGCTGGcgac	4	-7.22	34.76	75.63%	26.32%
	DRA0332	R	-153	-140	cactCGACCGCGCCATCGgcga	5.04	-7.94	35.8	77.89%	33.16%
	DRA0333	R	-261	-248	gccgCTATCGTGGTCTCGctca	2.09	-6.02	32.85	71.48%	13.75%
	DRA0334	R	-47	-34	gcgaCGAACTCGCCTTCGgcaa	9.25	-11.37	40.01	87.05%	60.86%
21	DRB0143	D	-294	-281	cgaaTGAACAGAAGGTAGggca	4.35	-7.46	35.11	76.39%	28.62%

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

	DRB0144	R	-398	-385	ccCAGCCCGAGGTTCTgccg	1.66	-5.77	32.42	70.54%	10.92%

Note: Abbreviations are as in Table S4 & S5. The bases in lowercase are less conserved than those in bold uppercase.

Supplementary Table 12 (S12): Details of SOS-box motifs predicted in the upstream sequences of *D. radiodurans* genes related to some unusual predicted operons through PoSSumSearch program.

Cluster No.	Gene cluster	Predicted SOS-box	Position	Score	MSS	P-value	E-value	MSSP
1	DR0298	CGACGCCGAGTTCG	70	63	7.08E-01	8.80E-04	2.19E+01	70.79%
	DR0299	CAAGCGCACGGTGG	299	61	6.85E-01	1.58E-03	3.92E+01	68.54%
2	DR0398	CAACGACGCGTTCG	5	52	5.84E-01	1.46E-02	3.64E+02	58.43%
	DR0390	CGAGCTCACGGACG	32	67	7.53E-01	2.50E-04	6.21E+00	75.28%
3	DR0544	CGAACTCAAGGGCA	122	67	7.53E-01	2.50E-04	6.21E+00	75.28%
	DR0545	ACAACTCAAGCTCT	337	59	6.63E-01	2.71E-03	6.73E+01	66.29%
4	DR0674	CGAACATCGGTCTT	93	57	6.40E-01	4.54E-03	1.13E+02	64.04%
	DR0675	CGGCTTCTGGTTCG	131	53	5.96E-01	1.18E-02	2.93E+02	59.55%
	DR0676	TCAACGTCTGGACG	207	58	6.52E-01	3.52E-03	8.75E+01	65.17%
	DR0677	CGAAATCGTGGGCG	244	64	7.19E-01	6.49E-04	1.61E+01	71.91%
5	DR0678	AGATTATGCGTACA	145	40	4.49E-01	1.26E-01	3.14E+03	44.94%
	DR0679	CGAACTGAAAGCCG	210	65	7.30E-01	4.77E-04	1.19E+01	73.03%
	DR0680	CTTCTTTGAGATCG	287	41	4.61E-01	1.09E-01	2.71E+03	46.07%
	DR0681	CGATTGGACGTTCT	321	57	6.40E-01	4.54E-03	1.13E+02	64.04%
	DR0682	CGACCACACCGTCT	198	55	6.18E-01	7.43E-03	1.85E+02	61.80%
	DR0683	CGAAGCGGTGGGCG	50	59	6.63E-01	2.71E-03	6.73E+01	66.29%
6	DR0689	CCAACATGCGCTTT	5	52	5.84E-01	1.46E-02	3.64E+02	58.43%
	DR0690	CGAACAGTATTTCT	272	57	6.40E-01	4.54E-03	1.13E+02	64.04%
7	DR0796	CGAAACCGAAGGCG	98	54	6.07E-01	9.40E-03	2.34E+02	60.67%
	DR0797	CGAACGGATGGTTC	170	72	8.09E-01	4.17E-05	1.04E+00	80.90%
	DR0798	CGGGCGTGAGGTCG	110	59	6.63E-01	2.71E-03	6.73E+01	66.29%
8	DR0853	CGAACCCATTCTCG	185	75	8.43E-01	1.25E-05	3.12E-01	84.27%
	DR0854	CGAGCGTCAGGAGG	227	56	6.29E-01	5.83E-03	1.45E+02	62.92%
9	DR0861	CTACGACAATTTCG	194	52	5.84E-01	1.46E-02	3.64E+02	58.43%
	DR0862	CGACCTTACGGGCC	129	57	6.40E-01	4.54E-03	1.13E+02	64.04%
10	DR0993	CGACGGAGCGGTTG	207	55	6.18E-01	7.43E-03	1.85E+02	61.80%
	DR0994	CCAACTTGTGACCG	73	61	6.85E-01	1.58E-03	3.92E+01	68.54%
	DR0995	AGAAATGATGTTCT	192	64	7.19E-01	6.49E-04	1.61E+01	71.91%
11	DR1175	CGATCCCAAATACA	116	57	6.40E-01	4.54E-03	1.13E+02	64.04%
	DR1174	CGGCTTCGACTTCG	100	49	5.51E-01	2.73E-02	6.78E+02	55.06%
12	DR1232	GGAACGTACGGCAG	206	64	7.19E-01	6.49E-04	1.61E+01	71.91%
	DR1233	CACCTATAGGTTCG	165	48	5.39E-01	3.32E-02	8.25E+02	53.93%
	DR1234	TGAACTCGCGTGCC	129	59	6.63E-01	2.71E-03	6.73E+01	66.29%
	DR1235	CAACGACGCGGTCT	38	40	4.49E-01	1.26E-01	3.14E+03	44.94%

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

13	DR1596	GCGCCGCGCGTTCG	184	49	5.51E-01	2.73E-02	6.78E+02	55.06%
	DR1597	GGAAGGCGCGGACG	191	55	6.18E-01	7.43E-03	1.85E+02	61.80%
	DR1641	CGCCCTCGACATCG	131	51	5.73E-01	1.81E-02	4.51E+02	57.30%
	DR1642	CGCAGGTGGGTGCG	168	54	6.07E-01	9.40E-03	2.34E+02	60.67%
14	DR1928	GGACCCCGAGTTCA	291	57	6.40E-01	4.54E-03	1.13E+02	64.04%
	DR1929	CCTCGGGGCGTTCG	262	43	4.83E-01	7.99E-02	1.99E+03	48.31%
15	DR2180	CGAACAGGTGATTG	271	69	7.75E-01	1.28E-04	3.17E+00	77.53%
	DR2181	CGTCGGCGCGCTCG	129	49	5.51E-01	2.73E-02	6.78E+02	55.06%
16	DR2220	TGGCCGCATGTGCG	260	61	6.85E-01	1.58E-03	3.92E+01	68.54%
	DR2221	CGACTTCGACTTCG	92	57	6.40E-01	4.54E-03	1.13E+02	64.04%
	DR2222	CGAAGGAGCCGTCG	14	57	6.40E-01	4.54E-03	1.13E+02	64.04%
	DR2223	AGAACGCACAGTGG	43	66	7.42E-01	3.47E-04	8.63E+00	74.16%
	DR2224	CATCCACACGGTCT	63	46	5.17E-01	4.81E-02	1.20E+03	51.69%
	DR2225	CGGACATGCGTCCG	337	60	6.74E-01	2.07E-03	5.15E+01	67.42%
	DR2226	TGAACGTCTGATCT	239	62	6.97E-01	1.19E-03	2.95E+01	69.66%
17	DR2254	CGAAACATTGGTCG	145	66	7.42E-01	3.47E-04	8.63E+00	74.16%
	DR2255	AGGAATCAGATTCG	354	57	6.40E-01	4.54E-03	1.13E+02	64.04%
18	DR2311	CGAACTCGTCTGGG	77	65	7.30E-01	4.77E-04	1.19E+01	73.03%
	DR2312	CGTCCGCAGGGTCA	196	56	6.29E-01	5.83E-03	1.45E+02	62.92%
19	DRA0231	CGACCGTTCGTCCT	64	54	6.07E-01	9.40E-03	2.34E+02	60.67%
	DRA0232	CGAACGCATGAGCG	307	76	8.54E-01	8.33E-06	2.07E-01	85.39%
	DRA0233	CTAACTTTCGCTCC	373	55	6.18E-01	7.43E-03	1.85E+02	61.80%
20	DRA0331	CGAATACACGCTGG	67	64	7.19E-01	6.49E-04	1.61E+01	71.91%
	DRA0332	CGACCGCGCCATCG	247	58	6.52E-01	3.52E-03	8.75E+01	65.17%
	DRA0333	CTATCGTGGTCTCG	139	48	5.39E-01	3.32E-02	8.25E+02	53.93%
	DRA0334	CGAACTCGCCTTCG	353	72	8.09E-01	4.17E-05	1.04E+00	80.90%
21	DRB0143	TGAACAGAAGGTAG	106	67	7.53E-01	2.50E-04	6.21E+00	75.28%
	DRB0144	CAGCCCGAGGTTCT	2	47	5.28E-01	4.01E-02	9.97E+02	52.81%

Note: Abbreviations are as in Table S4 & S6.

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Supplementary Table 13 (S13): Genome wide top hits for LexA TFBS in the upstream sequences of *D. radiodurans* genome at lower threshold cut off 88.21% MSSP using weight matrix alone. Only few genes belongs to DNA repair and stress response family were predicted through this method.

Gene	Function	strand	start	end	Predicted SOS-box	score	ln(P)	MSSP	RMMP
ctaC	cytochrome c oxidase, subunit II	D	-272	-259	ccgcCGAACGCAAGGTCGccgt	13.72	-16.67	96.78%	90.27%
DR1442	hypothetical protein	D	-240	-227	tcgcCGAACTCGTGTTTGccca	12.78	-15.3	94.73%	84.08%
DR1000	hypothetical protein	R	-215	-202	tgacCGAACTCGGGTTCGcgcc	12.59	-15.05	94.32%	82.83%
DRA0093	hypothetical protein	D	-281	-268	ccgtCGAACGCGAGTTTGgcga	12.32	-14.73	93.73%	81.06%
DRA0019	hypothetical protein	D	-383	-370	ttgtCGAACTCGTGGTTGccca	12.13	-14.47	93.32%	79.81%
DR1083	hypothetical protein	D	-30	-17	gaacCGAACGAGCGTTCGcatg	11.14	-13.4	91.17%	73.29%
DR1084	methylmalonyl-CoA mutase	R	-55	-42	catgCGAACGCTCGTTCGgttc	11.14	-13.4	91.17%	73.29%
DR2433	nitrilase-related protein	R	-291	-278	tcctCGAACATGGTTTCGagca	11.11	-13.28	91.10%	73.10%
DRA0100	hypothetical protein	D	-175	-162	gcccCGAGCGCAAGTTCGgcac	11.04	-13.23	90.95%	72.64%
DR2278	amino acid ABC transporter, periplasmic amino acid-binding protein	R	-179	-166	gcctCGATCTTGGGTTCGggaa	10.95	-13.11	90.75%	72.04%
DRA0211	transcriptional regulator, GntR family	D	-122	-109	gcggCGAACGCCTGGTTGaggc	10.71	-12.86	90.23%	70.46%
DRA0212	hypothetical protein	D	-163	-150	gcggCGAACGCCTGGTTGaggc	10.71	-12.86	90.23%	70.46%
DR2042	hypothetical protein	D	-207	-194	gggtCGAACTCAAGTAAGcctg	10.69	-12.84	90.19%	70.33%
DR0951	succinate dehydrogenase	R	-303	-290	cgctCGAACTCGGGTTCAtgct	10.62	-12.76	90.03%	69.87%
DRA0069	cleavage and polyadenylation specificity factor-related protein	D	-378	-365	ggggTGAACTCGAGTTCGagtt	10.62	-12.76	90.03%	69.87%
DR1679	N-methyl-transferase-related protein	D	-363	-350	tcacCGAACGCGAGGACGgcac	10.59	-12.74	89.97%	69.68%
fadD-3	acyl-CoA synthase	D	-238	-225	tgaaCGAACGGATGTTAGcaaa	10.58	-12.71	89.95%	69.61%
DR1993	hypothetical protein	D	-324	-311	cgagCGAACTGGTGTTCGgcta	10.38	-12.49	89.51%	68.29%
recA	recombinase A	R	-240	-227	ccgaCGACCTCGCGTTCAaggc	10.35	-12.47	89.45%	68.10%
DR0970	electron transfer flavoprotein, alpha subunit	D	-333	-320	ccaaCGAACTGAAGGTCGaggg	10.3	-12.41	89.34%	67.77%
DR0270	ribonuclease Z	D	-286	-273	agcgCGAAGTCAAGTTCGcctg	10.28	-12.39	89.30%	67.64%
DR0943	hypothetical protein	R	-359	-346	gcacCAACCATGCCTTCGaccg	10.19	-12.29	89.10%	67.04%
DRA0256	phenylacetyl-CoA ligase	R	-96	-83	caacCAACCAAACGTTCGttag	10.07	-12.19	88.84%	66.25%

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

DRB0131	hypothetical protein	D	-321	-308	tgcaCGAACCATTGTTCGtacc	10.05	-12.18	88.79%	66.12%
DR0781	response regulator, OmpR/PhoB family	D	-393	-380	tggaCGAACTGACGTTTGcccg	10.05	-12.17	88.79%	66.12%
DRA0222	TDP-glucose-4,6-dehydratase-related protein	D	-288	-275	ccgaCGAACGCAAGGTCTggat	10	-12.1	88.69%	65.79%
DR0388	hypothetical protein	R	-333	-320	agcgCGTAGATGCGTTCGgagt	9.89	-12.01	88.45%	65.07%
DR0050	hypothetical protein	R	-108	-95	cctgCAACCGCTCGTTCGgcca	9.85	-11.99	88.36%	64.81%
DR0111	thymidylate kinase	R	-51	-38	gcatCGAAGACGAGTTCGgcat	9.84	-11.99	88.34%	64.74%
DR0238	hypothetical protein	R	-264	-251	agttCGAAGACGAGTTCGccgg	9.84	-11.99	88.34%	64.74%
DR0945	hypothetical protein	R	-234	-221	cctaCGACCACGCGCTCGgcaa	9.83	-11.97	88.32%	64.67%
DR2348	hypothetical protein	R	-375	-362	ccgaCGACTTCGAGTTCGacga	9.83	-11.97	88.32%	64.67%
DRA0344	lexA repressor	D	-146	-133	cttcCGAACTCACGGAAGctgc	9.79	-11.89	88.23%	64.41%

Note: Abbreviations are as in Table S4 & S5. The bases in lowercase are less conserved than those in bold uppercase.

Supplementary Table 14 (S14): Genome wide detection of known *B. subtilis* SOS-box (CGAACRNRYGTTCG) in the upstream sequences (-400 bp) of *B. subtilis* through direct pattern matching method alone. Only 16 genes were identified.

S.No.	Gene ID	Strand	Start	End	Predicted SOS-box	RMMP (%)	Score
1.	pabC	D	-330	-317	CGCACAAGTGTTCG	93	0.93
2.	xkdA	D	-29	-16	AGAACACACGTTCG	93	0.93
3.	xkdA	R	-29	-16	CGAACGTGTGTTCT	93	0.93
4.	ykvR	D	-124	-111	CGAACGTATGTTTG	93	0.93
5.	ykvR	R	-124	-111	CAAACATACGTTCG	93	0.93
6.	yloS	D	-358	-345	CGAACATGTTTTCG	93	0.93
7.	yloS	R	-358	-345	CGAAAACATGTTCG	93	0.93
8.	recA	D	-86	-73	CGAATATGCGTTCG	93	0.93
9.	recA	R	-86	-73	CGAACGCATATTCG	93	0.93
10.	aprX	D	-179	-166	CGAACAAACGTTCT	93	0.93
11.	yqjW	D	-56	-43	CGAACATACTTTCG	93	0.93
12.	yqjW	R	-56	-43	CGAAAGTATGTTCG	93	0.93
13.	yqzH	D	-123	-110	CGAAAGTATGTTCG	93	0.93
14.	yqzH	R	-123	-110	CGAACATACTTTCG	93	0.93
15.	tagC	D	-68	-55	CGAACGTATGTTTG	93	0.93
16.	tagC	R	-68	-55	CAAACATACGTTCG	93	0.93

Note: Abbreviations are as in Table S4 & S5.

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Supplementary Table 15 (S15): Genome wide detection of known *B. subtilis* SOS-box (CGAACRNRYGTTCG) in the upstream sequences (-400 bp) of *D. radiodurans* through direct pattern matching method alone. Only 18 genes were identified.

S.No.	Gene ID	Strand	Start	End	Predicted SOS-box	RMMP	Score
1.	DRB0058	D	-302	-289	CGAAAAGGTGTTCG	93%	0.93
2.	DRC0032	D	-225	-212	CGAAAAGGTGTTCG	93%	0.93
3.	DR0020	D	-246	-233	CGACCAGATGTTCG	93%	0.93
4.	DR0137	R	-123	-110	CGGACGAGCGTTCG	93%	0.93
5.	DR0383	R	-135	-122	CGAGCAAATGTTCG	93%	0.93
6.	DR0992	D	-160	-147	CGATCAGGCGTTCG	93%	0.93
7.	DR1083	D	-30	-17	CGAACGAGCGTTCG	100%	1.00
8.	DR1083	R	-30	-17	CGAACGCTCGTTCG	93%	0.93
9.	DR1084	D	-55	-42	CGAACGCTCGTTCG	93%	0.93
10.	DR1084	R	-55	-42	CGAACGAGCGTTCG	100%	1.00
11.	DR1197	R	-391	-378	CGAAAAGGTGTTCG	93%	0.93
12.	DR1240	D	-32	-19	CCAACAAACGTTCG	93%	0.93
13.	DR1730	R	-302	-289	CGAACAGGCGCTCG	93%	0.93
14.	DR1823	D	-135	-122	AGAACAAACGTTCG	93%	0.93
15.	DR1824	R	-18	-5	AGAACAAACGTTCG	93%	0.93
16.	DR1993	D	-324	-311	CGAACTGGTGTTCG	93%	0.93
17.	DR2386	R	-51	-38	CTAACGAGCGTTCG	93%	0.93
18.	DR2397	D	-109	-96	CGAACAGGCTTTCG	93%	0.93
19.	DRA0168	R	-126	-113	CGAACGGACGTACG	93%	0.93
20.	fadD-3	D	-238	-225	CGAACGGATGTTAG	93%	0.93

Note: Abbreviations are as in Table S4 & S5.

Copyright 2008 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).