*Journal of Cell and Molecular Research (2017) 9 (2), 84-97* DOI: 10.22067/jcmr.v9i2.66166

Research Article

# Systems Biology Analysis of the Key Genes of Surfactin Production in *Bacillus subtilis* MJ01 (Isolated from Soil Contaminated Oil in South of Iran), Spizizenii, and 168 Isolates

Tahereh Deihimi[1], Esmaeil Ebrahimie[1,5]*, Ali Niazi[1], Mansour Ebrahimi [2] , Shahab Ayatollahi[3] , Ahmad Tahmasebi [4] , Touraj Rahimi[1], Moein Jahanbani Veshareh[6]

[1]*Institute of Biotechnology, Shiraz University, Shiraz, Iran.*
[2] *Department of Biology, University of Qom, Qom, Iran*
[3]*School of Chemical & Petroleum Engineering, Sharif University of Technology, Tehran, Iran*
[4]*Department of Crop Production & Plant Breeding, College of Agriculture, Shiraz University, Shiraz, Iran*
[5] *Adelaide Medical School, Faculty of Health and Medical Sciences, The University of Adelaide, Adelaide, Australia*
[6] *Danish Hydrocarbon Research and Technology Center, Kgs. Lyngby, Denmark*

Received 17 July 2017 Accepted 23 Agust 2017

## Abstract

Applying microorganism in oil recovery has attracted attentions recently. Surfactin produced by *Bacillus subtilis* is widely used industrially in a range of industrial applications in pharmecutical and environmental sectors. Little information about molecular mechanism of suffactin compound is available. In this study, we performed promoter and network analysis of surfactin production genes in *Bacillus subtilis* subsp. MJ01 (isolated from oil contaminated soil in South of Iran), spizizenii and 168. Our analysis revealed that *comQ* and *comX* are the genes with sequence alterations among these three strains of *Bacillus subtilis* and are involved in surfactin production. Promoter analysis indicated that *lrp*, *argR*, *rpoD*, *purr* and *ihf* are overrepresented and have the highest number of transcription factor binding sites (TFBs) on the key surfactin production genes in all 3 strains. Also the pattern of TFBs among these three strains was completely different. Interestingly, there is distinct difference between 168, spizizenii and MJ01 in their frequency of TFs that activate genes involve in surfactin production. Attribute weighting algorithms and decision tree analysis revealed *ihf*, *rpoD* and *flHCD* as the most important TF among surfactin production. Network analysis identified two significant network modules. The first one consists of key genes involved in surfactin production and the second module includes key TFs, involved in regulation of surfactin production. Our findings enhance understanding the molecular mechanism of surfactin production through systems biology analysis.

***Keywords*:** Surfactin production gene, Transcription factor (TF), Promoter analysis, Network analysis.

## Introduction

Finding out alternative technologies to increase oil recovery from oil fields around the world has drawn attention for many years. Recently, using microorganisms in this field is popularized (Shibulal et al., 2014). Hence, identification and characterization of novel strains are in demand. *Bacillus subtilis* has been used as a model organism because of its ability to produce surface active compounds (biosurfactants) with highly desirable properties for oil recovery (Anuradha S, 2010). Surfactin shows a remarkable membrane-active and surface-interface properties with a number of biological activities in health care and biotechnology-based processes.

Surfactin draws biotechnologists attention as a potent candidate drug for the resolution of a number of global issues in medicine (Banat et al., 2010; Cao et al., 2010) , industry (Abdel-Mawgoud et al., 2008; Nitschke and Costa, 2007), and environmental protection (Mulligan, 2009). Consequently, an urge has been formed towards understanding the molecular mechanism, gene network and transcriptomic comparison of genes involved in surfactin production.

The dynamics of surfactin production by *Bacillus subtilis* is still ambiguous because of shortage in information on different levels of functional genomics such as promoter activation. While gene function is the result of interactive between upstream non coding promoter region and downstream coding sequence, most of the studies focus on genes (Deihimi et al., 2012). Undoubtedly, examination of

Corresponding authors E-mail:
* esmaeil.ebrahimie@adelaide.edu.au

upstream sequences of genes with similar expression pattern is very important (Yada et al., 1997). The role of transcription factors in controlling the expression of many genes involved in surfactin production in *Bacillus subtilis* is still in paucity and has not been studied in details. Gene network analysis, especially regulatory gene network, is a strong well developed tool in understanding the central genes (hubs) and gene interactions (Shoemaker and Panchenko, 2007). Data mining tools such as decision trees can be used to associate the result of different decisions (Ebrahimi et al., 2015). Moreover, these computational analytical tools discover function of genes and proteins structures and decipher the interactions of genes and also genes with transcription factors more clear (Pashaiasl et al., 2016a).

Herein, comparison of surfactin genes between *Bacillus subtilis* subsp. MJ01, spizizenii and 168 has been performed. In addition, for the first time, *in silico* promoter analysis and prediction of involved TFs and also network analysis of genes associated with surfactin production carried out in order to open a new avenue in molecular mechanism of producing surfactin by *Bacillus subtilis.*

## Materials and Methods

### Sequence Comparison of Key Genes Involved in Surfactin Production and Interacted Genes in *Bacillus subtilis* subsp. *168, Spizizenii* and *MJ01*

Protein and gene sequences of surfactin producing genes were retrieved from NCBI in *B. subtilis* subsp. 168 genome. Genome sequences of *Bacillus subtilis* subsp. 168 (AC: NC_000964), spizizenii (AC: NC_016047) and MJ01 (AC: CP018173) were also extracted from NCBI. The nucleotide and protein sequences of all genes associated with surfactin production isolated from *Bacillus subtilis* subsp. 168 were compared with *B. subtilis* subsp. spizizenii and MJ01 genome using blastn and tblastn respectively using CLC bio Genome Workbench software.

### Gene Interactions, Promoter Analysis and Comparative TF Activation Patterns of Genes Involved in Surfactin Production

In order to identify genes associated with surfactin production and their interactions, we used STRING server (http://string-db.org) (Szklarczyk et al., 2014). The gene sequence of each protein was obtained from NCBI database. We also used NCBI to find the genomic sequence of *Bacillus subtilis* subsp. spizizenii (NC_016047) and *Bacillus subtilis* subsp. Subtilis str. 168 (NC_000964). The genomic locations of genes involved in surfactin production were identified by CLCbio genomic workbench. The potential promoter regions of these genes for all 3 strains (MJ01, 168 and spizizenii) were extracted by selecting the region between each gene (or operon) and the next gene (or operon) as previously described (Mahdi et al., 2014). BPROM algorithm (http://linux1.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb) (Lee and Chen, 2002) was used to confirm the presence of promoter an its -10 and -35 sites for genes associate in surfactin production. TFs were predicted for all promoter sequences in all 3 strains (MJ01, spizizenii and 168) using BPROM.

### Network Analysis

The genes related to surfactin production were selected from the STRING online tool (Szklarczyk et al., 2014) with the cut-off criterion of combined score > 0.4. The relationships of the nodes degree ≤ 5 were abandoned. We added the TF identified by our promoter analysis to the list of genes for network construction. Finally, we visualized the network by Cytoscape 3.4.0 (Shannon et al., 2003). clusterONE algorithm (Nepusz et al., 2012) was employed to identify the significant modules . Moreover, the nodes with high degree and interaction were defined as hub proteins in the network. The node degree ≥3 were selected as the threshold.

### Data Mining

To find the pattern in data of genes and promoters, three sets of data generated:

1. PD200Genes: A promoter dataset of 200 genes which has been selected randomly among all *Bacillus subtilis* MJ01 genes and 5 key surfactine production associating gene containing 53 variables of promoters.
2. PD5NGenes: A promoter dataset of 5 genes which has been selected randomly among 200 random selective gene and 5 key surfactin production associating gene containing 53 variables of promoters. The variables were the number of promoters for each gene (numeric variables).
3. PD5BGenes: Again the same dataset as above (promoters of 5 randomly selected genes) from each group created but each feature set as binomial (the presence or absence of each promoter set as Yes/No values).

All three datasets were imported into RapidMiner Studio software separately (RapidMiner 5.0.001, Rapid-I GmbH, Stochumer Str. 475, 44,227 Dortmund, Germany); the type of gene set as target or labels variable and other 53 of features of promoters appointed as regular variables. The

following data mining algorithms applied on dataset: Attribute Weighting Algorithms and Decision Tree Models

**Attribute Weighting Algorithms**

Attribute weighting algorithms identify the most important attributes or features which differ between two groups of target or labeled attributes (Torkzaban et al., 2015). The models use various statistical approaches to perform the analysis. The following attribute weighting models were applied on datasets: weight by information gain, weight by information gain ratio, weight by rule, weight by deviation, weight by chi squared statistic, weight by Gini index, weight by uncertainty, weight by relief, weight by PCA, and weight by SVM. The algorithms definitions have already been described in our previous paper (Pashaiasl et al., 2016a). Weights were normalized into the interval between 0 and 1 to allow the comparison between different methods.

**Decision Tree Models**

Decision tree algorithms provide visual explanation of the most important features through depicting an inverted tree with the most important feature as root and other variables as leaves. Various decision trees including Random Forest, Decision Stump Decision, ID3, CHAID and Random Tree were applied on dataset. Details of each decision tree model have also been presented before (Pashaiasl et al., 2016b). To calculate the performance of decision tree models in predicting the right class of soil or non-soil group10-times cross validation was applied on dataset. This approach divides data into 10 parts and each time train the model with 9 parts and then test the model by the last part and computes the efficiency of it; repeating it for 10 times and gives the mean performance value.

**Result**

**Analysis of Surfactin Production Genes Between *Bacillus subtilis* Strains (168, Spizizenii, and MJ01 Strain)**

Our previous study identified MJ01 has more desirable properties for oil recovery than other strains (data not published). Here we attempted to investigate genes, promoters and transcription factors (TFs) involved in surfactin production of this strain to explain its importance over the other strain. Hence, we investigated the sequence difference of the key genes associated whith surfactin production between *Bacillus subtilis* MJ01, *Bacillus subtilis* subsp. spizizenii (which has 95% similarity with MJ01 in genome sequence) and *Bacillus subtilis* subsp. Subtilis str. 168 (which is reference genome) to explain this priority (Table 1). The comparative result revealed that there is no significant difference in sequence of the surfactin producing genes (*srfA*, *srfAA*, *srfAB*, *srfAD*, *sfp*) between 168 vs spizizenii and MJ01. Specially, no difference was observed between the surfactin producing genes sequence in spizizenii vs MJ01, and it can be concluded that they are similar for their sequences.

**Table 1.** Percentage identity of nucleotide sequence in genes associated with surfactin production among *Bacillus subtilis* 168 vs spizizenii vs MJ01

| Gene name | % identity of 168 vs spizizenii | % identity of 168 vs MJ01 | % identity of spizizenii vs MJ01 |
|---|---|---|---|
| **srfA** | 92 | 92 | 99 |
| **srfAA** | 92 | 92 | 99.5 |
| **srfAB** | 92 | 92 | 99 |
| **srfAD** | 93 | 93 | 98 |
| **Sfp** | 93 | 93 | 99 |

Hence, by gene network analysis, we tried to find the genes associated with the surfactin producing genes. Sequence similarity assay for all of the genes has been identified by all networks, between 168 vs spizizenii and MJ01 revealed that there is no significant variance between spizizenii and MJ01 nucleotide sequence for the genes associate with surfactin producing networks, but *ppsC*, *ppsA*, *comQ*, *comX* and *YndJ* has less than 90% identity among 168 vs spizizenii and MJ01 (supplementary 1). As a result of protein sequence comparison between 168 vs spizizenii and MJ01, *pksM*, *pksJ*, *pksD*, *pksL*, *pksR*, *ppsC*, *ppsA*, *comQ*, *comX* and *yndJ* showed less than %90 identity. In contrast, *comQ* and *comX* revealed significant variance between spizizenii and MJ01 in their protein sequences (Fig 1.) (Supplementary 2.).Oil activates different pattern of TFs among different strains of *Bacillus subtilis* (168, spizizenii and MJ01)

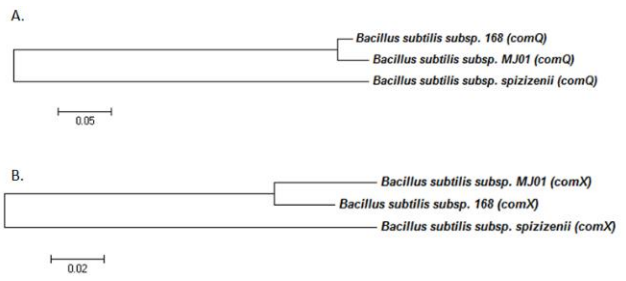


**Figure 1**. Phylogenetic relationship of genes *between Bacillus subtilis* subsp. spizizenii, 168 and MJ01. A. phylogenetic relationships of comQ among *Bacillus subtilis* subsp. spizizenii, 168 and MJ01. B. phylogenetic relationships of comK among *Bacillus subtilis* subsp. spizizenii, 168 and MJ01. The tree was built by using neighbour-joining algorithm.

We sought to identify if there is any difference in TFs capable of binding to genes associate with surfactin production between 3strains of *Bacillus subtilis* (168, spizizenii and MJ01). Generally, the differential TF activation profiles were observed as a result of comparison between 168, spizizenii and MJ01. Among all TFs identified for the promoters of genes associated with surfactin production, the number of *marR* and *ihf* was the same in all three genomes (*Bacillus subtilis* 168, *Bacillus subtilis* spizizeni and *Bacillus subtilis* MJ01).Thirty seven percent difference was observed in the number of TFs of surfactin producing genes promoter by comparing *Bacillus subtilis* MJ01 and *Bacillus subtilis* spizizenii. TFs such as *Lrp*, *argR*, *rpoD*, *purR* and *ihf* have the highest number of TFBSs (at least 5) between all 3 strains. The three highest numbers of TFs in *Bacillus subtilis* 168 gene promoter in surfactin associated genes were *rpoD*, *lrp*, *argR* by 13, 11 and 10, respectively. In *Bacillus subtilis* spizizenii, *rpoD*, *argR* and *purR* with 19, 11 and 8 predicted binding sites were the three highest number of TFs found on promoters of genes involve in surfactin production. The top three number of TFs in *Bacillus subtilis* MJ01 were similar to *Bacillus subtilis* spizizenii but with different numbers (*rpoD*, *argR*, and *purR* with 15, 8 and 8 predicted sites respectively). Our promoter analysis for *srfA*, *srfAA*, *srfAB* and *srfAD* which are the most important genes in surfactin production (Porob et al., 2013) indicates that *rpoH2* has binding site on promoters of these genes in *Bacillus subtilis* subsp. spizizenii but no binding site for *Bacillus subtilis* subsp. MJ01.
Surprisingly, TF analysis revealed that there is a distinct difference between the number of TFs that activate genes associate with surfactin production, between 168 and spizizenii and MJ01 (Table 2).

**Table 2.** Transcription factors (TFs) number of binding sites on the promoter region of surfatin producing genes which were different between all 3 strains (*Bacillus subtilis* 168, spizizenii and MJ01 ) for their TF patterns.

| Gene name | TFs number | | |
|---|---|---|---|
| | **168** | **spizizenii** | **MJ01** |
| **menB** | 2 | 2 | 1 |
| **dhbB** | 7 | 7 | 4 |
| **codY** | 3 | 2 | 0 |
| **ppsC** | 13 | 13 | 4 |
| **ppsA** | 13 | 13 | 2 |
| **dhbF** | 7 | 7 | 4 |
| **ybdZ** | 7 | 7 | 4 |
| **srfAB** | 13 | 10 | 10 |
| **srfAA** | 13 | 10 | 10 |
| **srfAD** | 13 | 10 | 10 |
| **srfA** | 13 | 10 | 10 |

Although the sequences of these genes and their proteins were the same in spizizenii and MJ01, but the number of TFs binding to their promoters are completely different. However, *sfp*, *pksR*, *pksN*, *pksM*, *pksJ*, *pksD*, *comK*, *mecA*, *rapC*, *dltA*, *YndJ* and *pksS* recognized the same TF pattern for spizizenii and MJ01 but completely different for 168. Even *comx* and *comQ* which have different sequence between spizizenii and MJ01, but their TF pattern is completely the same

## Construction of Surfactin Production Gene Regulatory Network in *Bacillus subtilis* subsp. 168

The surfactin production gene regulatory network was generated by combining predicted TFs, target genes and genes associated in surfactin production. *lexA* and *rpoD* with the most edge number among TFs were regulatory hub in this network which play key role in surfactin production. The network of each key surfactin production gene has been constructed (Figure 2.). *Ihf*, *PksM*, *PksD*, *PksL*, *Ybdz* and *arcA* are the joint node between *srfA*, *srfAA*, *srfAB*, *srfAD* and *sfp* network. *Srf* genes and *sfp* have interaction with each other in this surfactin production network which was prospective. *Ihf* and *arcA* are the most common TFs between *srfA*, *srfAA*, *srfAB*, *srfAD* and *sfp* network. *argR2*, *rpoH*, *cytR* and *rpoD* were involved in 80% of key gene networks. The genes involved in surfactin production have enriched functional groups by "Biological Process" term (Table 3).

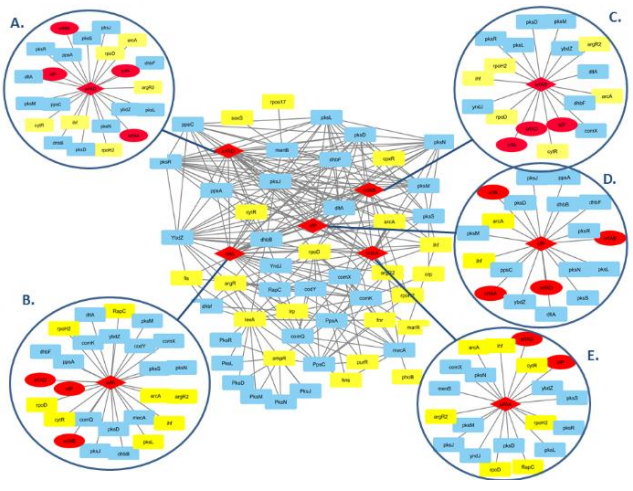


**Figure 2**. Regulatory network of surfactin production. The network has 55 nodes and 262 links. Diamond red nodes represent key surfactin production genes, yellow nodes represents transcription factors (TFs). A., B., C., D. and E. are the subnetwork centred by fAD, srfA, srfAA, sfp and srfAB, respectively.

**Table 3**. Functional classification of gene network of genes involve in surfactin production

| BioProcess | Number of genes involve |
|---|---|
| Antibiotic Biosynthetic Process | 10 |
| Biosynthetic Process | 24 |
| Cellular Biosynthetic Process | 22 |
| Cellular Metabolic Process | 24 |
| Metabolic Process | 28 |
| Transcription DNA template | 10 |
| Regulation of transcription, DNA template | 10 |
| Gene expression | 10 |
| Cellular macromolecule metabolic process | 11 |
| Transcription initiation from bacterial-type RNA polymerase promoter | 2 |
| Transcription from bacterial-type RNA polymerase promoter | 2 |
| primary metabolic process | 11 |
| Positive regulation of transcription, DNA-templated | 3 |

## Subnetwork Analysis of Genes and TFs Involved in Surfactin Production in Reference Genome

Two modules have been constructed using ClusterONE in Cytoscape software. Two modules under the condition of more than 5 nodes were screened and modular significance *P* value less than 0.05 were obtained. Analysis of these modules indicated that module 1 was enriched by the key genes involved in surfactin production. Thirty seven percent of genes consisting module 2 are TFs which directly or indirectly regulate surfactin production (Fig 3).

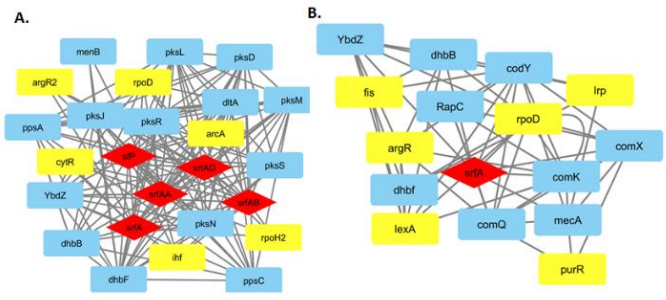


**Figure 3**. The modules identified from the regulatory network of surfectin produvction using ClusterONE. Yellow round rectangular represent transcription factors, and red diamond represent for key surfactin producing genes. A. and B. indicates first and second, respectively.

## Decision Tree and Data Mining Analysis Selected *flHCD, ihf* and Two Members of *rpoD* Groups as the Most Important TF

As mentioned in Materials and Methods, three different datasets of gene promoters were prepared based on the number of samples (in PD200Genes the number of samples were 200; while in two other datasets –PD5NGenes and PD5BGenes – the number of samples was only 5 in each group to balance the number of samples in each group).

### Attribute Weighting Algorithms

PD200Genes: When attribute-weighting algorithms applied on this dataset, *flHCD* attribute was selected as the most important feature by at least 75% of models. Seven algorithms (SVM, Uncertainty, Gini Index, Chi Squared, Rule, Info Gain Ration and Info Gain) generated the highest possible weight of 1.0 to this gene. SVM and Rule algorithms were appointed 1.0 weight to *ihf* and *rpoD17* gene variables; putting them in the next positions. Details of other weights have been shown in Table 4.

### PD5NGenes

When the number of samples in each group balanced and 5 random samples were taken from the main dataset, the attribute weighting algorithms ran on this dataset showed that *ihf* attribute received the best weight of 1.0 by nearly 80% of algorithms; 100% of them gave weights higher than 0.75; confirming the importance of this feature. The other features such as *flHCD*, *rpoD18* and *rpoD17* were also gained more weights here (Table 5).

### PD5BGenes

For this dataset, the presence or the absence of promoters in each gene marked as Yes or No (binomial features created). The results of attribute weighting showed that *ihf* feature gained the best score and marked at the most important feature by 85% of attribute weighting algorithms. *flHCD* and *rpoD18* were among the other features selected by attribute weighting algorithms as second most important (Table 6).

### Decision Trees

Various decision trees were applied on three datasets; the accuracies of each decision tree model calculated based on 10-fold cross validation (dataset divided into 10 equal sets, each time 9 sets used to train the model and tested with the last set, then the model repeated with another 9 and 1 sets and the average of 10 run performances calculated and reported).

**Table 4**. Ten different attribute weighting models applied on all samples showing the most important transcription factors based on various levels of weights assigned by each model.

| SVM | Relief | Uncertainty | Gini Index | Chi Squared | Deviation | Rule | Info Gain Ratio | Info Gain | Attribute | Count 50% | Count 75% | Count 95% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1.0** | .1 | 1.0 | 1.0 | 1.0 | .3 | 1.0 | 1.0 | 1.0 | flHCD | 7 | 7 | 7 |
| **.6** | .1 | .3 | .3 | .3 | .6 | 1.0 | .3 | .6 | rpoD15 | 4 | 1 | 1 |
| **.6** | .0 | .3 | .3 | .3 | .6 | 1.0 | .3 | .6 | rpoD18 | 4 | 1 | 1 |
| **.4** | .7 | .3 | .3 | .3 | .8 | 1.0 | .2 | .7 | ihf | 4 | 2 | 1 |
| **1.0** | .3 | .3 | .2 | .2 | .6 | 1.0 | .2 | .5 | crp | 3 | 2 | 2 |
| **.3** | .2 | .4 | .3 | .4 | .6 | 1.0 | .3 | .6 | purR | 3 | 1 | 1 |
| **.4** | 1.0 | .1 | .0 | .0 | .9 | 1.0 | .0 | .2 | rpoD17 | 3 | 3 | 2 |
| **.0** | .6 | .0 | .0 | .0 | .2 | 1.0 | .0 | .0 | hipB | 2 | 1 | 1 |
| **.0** | .4 | .0 | .0 | .0 | .8 | 1.0 | .0 | .1 | lrp | 2 | 2 | 1 |
| **.0** | .2 | .0 | .0 | .0 | .5 | 1.0 | .0 | .0 | ompR | 2 | 1 | 1 |
| **.0** | .2 | .0 | .0 | .0 | .7 | 1.0 | .0 | .0 | fur | 2 | 1 | 1 |
| **.2** | .3 | .0 | .0 | .0 | .7 | 1.0 | .0 | .0 | arcA | 2 | 1 | 1 |
| **.3** | .3 | .0 | .0 | .0 | .6 | 1.0 | .0 | .0 | phoB | 2 | 1 | 1 |
| **.1** | .4 | .0 | .0 | .0 | .6 | 1.0 | .0 | .1 | tyrR | 2 | 1 | 1 |
| **.3** | .5 | .2 | .2 | .2 | .8 | 1.0 | .1 | .4 | rpoD16 | 2 | 2 | 1 |
| **.3** | .5 | .1 | .1 | .1 | .8 | 1.0 | .1 | .3 | argR | 2 | 2 | 1 |
| **.0** | .3 | .2 | .2 | .2 | .8 | 1.0 | .2 | .4 | argR2 | 2 | 2 | 1 |
| **.3** | .3 | .2 | .1 | .2 | 1.0 | 1.0 | .1 | .3 | lexA | 2 | 2 | 2 |

**Table 5.** Ten different **a**ttribute weighting models applied on five samples showing the most important transcription factors based on various levels of weights assigned by each model.

| PCA | SVM | Relief | Uncertainty | Gini Index | Chi Squared | Deviation | Rule | Info Gain Ratio | Info Gain | Attribute | Count 50% | Count 75 | Count 95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **.8** | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | .8 | 1.0 | 1.0 | 1.0 | ihf | 10 | 10 | 8 |
| **.2** | 1.0 | .1 | .1 | .0 | .1 | .5 | .0 | .2 | .1 | arcA | 1 | 1 | 1 |
| **.3** | 1.0 | .1 | .3 | .2 | .3 | .6 | .4 | .3 | .2 | rpoD17 | 2 | 1 | 1 |
| **.5** | .6 | .1 | .4 | .4 | .4 | 1.0 | .6 | .4 | .4 | lrp | 4 | 1 | 1 |
| **.6** | .1 | .0 | .3 | .1 | .4 | 1.0 | .6 | .2 | .1 | argR | 3 | 1 | 1 |
| **.7** | .0 | .0 | .3 | .1 | .4 | 1.0 | .6 | .2 | .1 | rpoD16 | 3 | 1 | 1 |
| **1.0** | .8 | .6 | .6 | .6 | .7 | .7 | .8 | .6 | .6 | argR2 | 10 | 3 | 1 |
| **1.0** | .8 | .6 | .6 | .6 | .7 | .7 | .8 | .6 | .6 | rpoD18 | 10 | 3 | 1 |
| **1.0** | .8 | .6 | .6 | .6 | .7 | .7 | .8 | .6 | .6 | crp | 10 | 3 | 1 |
| **1.0** | .8 | .6 | .6 | .6 | .7 | .7 | .8 | .6 | .6 | flHCD | 10 | 3 | 1 |

**Table 6.** Ten different attribute weighting models applied on five samples (data transformed into binominal – YES/NO) showing the most important transcription factors based on various levels of weights assigned by each model.

| Relief | Uncertainty | Gini Index | Chi Squared | Rule | Info Gain Ratio | Info Gain | Attribute | Count 50% | Count 75 | Count 95 |
|---|---|---|---|---|---|---|---|---|---|---|
| **.5** | .6 | .7 | .7 | .1 | .6 | .6 | argR2 | 6 | 0 | 0 |
| **.5** | .6 | .7 | .7 | .1 | .6 | .6 | rpoD18 | 6 | 0 | 0 |
| **.5** | .6 | .7 | .7 | .1 | .6 | .6 | crp | 6 | 0 | 0 |
| **.5** | .6 | .7 | .7 | .1 | .6 | .6 | flHCD | 6 | 0 | 0 |
| **1.0** | 1.0 | 1.0 | 1.0 | .0 | 1.0 | 1.0 | ihf | 6 | 6 | 6 |

**PD200Genes**

The best performances in predicting the right genes obtained when Decision Tree model applied on PD200Genes dataset. As seen in Table 7, the accuracy of this model on predicting surfactin gene

was around 80% while the same accuracy for predicting other genes reached at 99.5%. As seen in Fig 4., Decision Tree model drew an inverted tree with *flHCD* feature at the tree root; showing when this feature was less to or equal to 0.5, the class was other genes but when it was higher than 0.5, if *rpoD15* was higher than 0.5, the class was surfacin gene, otherwise other genes.

**Table 7.** Confusion matrix for PD200Genes showing the accuracy of model in predicting the right class of other genes and surfactin genes (accuracy: 99.00% +/- 2.00% (mikro: 99.02%))

| | True other genes | True surfactin gene | Class precision |
|---|---|---|---|
| **Pred. other genes** | 198 | 1 | 99.50% |
| **Pred. surfactin gene** | 1 | 4 | 80.00% |
| **Class recall** | 99.50% | 80.00% | |

## PD5NGenes

The accuracy of Random forest model in predicting the right gene class for this dataset was 90%; with the best possible accuracy for predicting surfactin genes and 83.33% accuracy for predicting the other class (other genes) (Table 8).

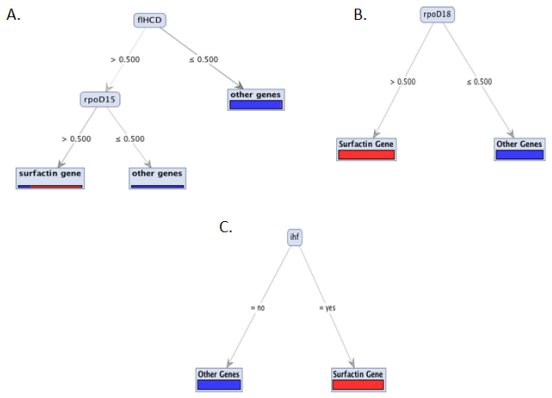


**Figure 4**. Decision trees induced by tree generating algorithms; showing the most important features in separation of surfactin gene. A. A decision tree that did not separate two classes of genes (surfactin and other genes) completely B. A decision tree separated two classes of genes (surfactin and other genes) completely. C. A decision tree confirmed the results of weighting algorithms

The model induced a simple one level tree with the *rpoD18* feature at the top. When this model was higher than 0.5, the gene class was other genes but when it was higher than 0.5, the gene class was surfactin genes.

## PD5BGenes

The best accuracy obtained when Decision Stump algorithm applied on this dataset, the model accuracy reached 90%; the model was perfect in predicting surfactin genes but less accurate in predicting the other genes (83.33%) (Table 8).
Decision Stump algorithm generated a simple tree with just ihf attribute at the root; this tree was capable of predicting the right class with just this feature (ihf), when it was yes, the gene class was surfactin, otherwise other genes (Fig 4).

**Table 8.** Confusion matrix for PD5NGenes and PD5BGene showing the accuracy of model in predicting the right class of other genes and surfactin genes (accuracy: 90.00% +/- 30.00% (mikro: 90.00%)

| | true Other Genes | true Surfactin Gene | class precision |
|---|---|---|---|
| **Pred. other genes** | 5 | 1 | 83.33% |
| **Pred. surfactin gene** | 0 | 4 | 100.00% |
| **Class recall** | 100.00% | 80.00% | |

## Discussion

Bacteria make a wide range of surface active compounds called "biosurfactants ". *Bacillus subtilis* produced surfactin as one of the most popular biosurfactant (Shao et al., 2015). Surfactin can reduce the surface and interfacial tension. Moreover, because of its tremendous potential, it is of great industrial and commercial interest (Yeh et al., 2005). Hence, molecular characterization of surfactin production regulatory elements and network is in demand.
Recently, we have isolated the *Bacillus subtilis* subsp. MJ01 from crude oil contaminated soil in the south of Iran (data not published). In previous study, we sequenced this strain with PacBio RS (Pacific Biosciences) and also annotating of this strain was performed too (data not published). The genome NCBI genbank accession number is CP018173. Laboratory experiments revealed the high amount of surfactin production and also the high Critical Micelle Concentration (CMC) of produced surfactin in distilled water for *B. subtilis subsp. MJ01*. In this study, we focused on promoter analysis and generate

the regulatory gene network for surfactin production for the first time to further investigate about molecular mechanisms of surfactin production. We expected to find some gene sequence alterations in surfactin production key genes between *B. subtilis* subsp. MJ01, spizizenii and 168, in order to explain the difference of surfactin production. Analysis of gene sequences such as *srfA*, *srfAA*, *srfAB*, *srfAD* and *sfp* involved in surfactin production could not reveal any significant variation between MJ01and spizizenii. However, about 7% difference was observed in these gene sequences in MJ01 and spizizenii compare with 168. *comQ* and *comX* were two genes which showed difference in both nucleotide and protein sequences between MJ01 and spizizenii vs 168. In addition, the only difference between MJ01 and spizizenii was observed in protein sequence of *comQ* and *comX*. These genes play the key roles for both competence and surfactin production (Oslizlo et al., 2014; Weinrauch et al., 1991). The findings of previous studies have validated the role of *comQ* and *comX* in surfactin production and our analysis identified them as a hot gene, strongly support the reliability of our results.

It has been suggested that TFs and promoter activating pattern may alter instead of alteration in coding sequences to generate more virulent strain (Mahdi et al., 2014; van Schaik et al., 2007). Hence, the identified potential TFBs and their organization open a new avenue to understand gene expression and regulation during surfactin production. By functional genomics based approach, we could compare TFs activating pattern of *Bacillus subtilis* MJ01 with spizizenii and 168, during surfactin production.

The most amazing result was not observed by difference in gene sequences but in the promoter and TFs patterns of key genes in surfactin production. The number of TFs activated on promoters of key genes involved in surfactin production (*srfA*, *srfAA*, *srfAB* and *srfAD*) was similar for each strain. It is obvious as all of these genes are the most important genes for surfactin production. But the number of TFs promote these genes (*srfA*, *srfAA*, *srfAB* and *srfAD*) was higher in MJ01 and spizizenii compare with 168. Mahdi et al. 2014 suggest that the higher number of TFs can be an index for more active/key genes (Mahdi et al., 2014).

Our findings also revealed that *Lrp*, *argR*, *rpoD*, *purR* and *ihf* were highly activated in all 3 strains. We suggest that these TFs might play a key role in surfactin production pathway based on their numbers. *rpoD* has been identified as the most active TFs among all 3 strains. It promotes transcription by attachment of RNA polymerase to the specific initiation site (Hengge-Aronis, 2002). *Ihf* wraps DNA around the body of the protein to form a higher-order nucleoprotein complex (Pagel et al., 1992; Winkelman and Hatfield, 1990) and facilitates the unwinding of the DNA helix in the -10 hexanucleotide region of the downstream promoter (Parekh and Hatfield, 1996). Also Parekh et al. 1996 suggest that Ihf activates transcription of some genes by forming a higher order protein-DNA complex that change the DNA helix in order to assist opening DNA helix at downstream promoters site (Parekh et al., 1996). *Lrp* has role in global regulation of cellular metabolism (Calvo and Matthews, 1994).

The observed difference between the predicted TF activation patterns among 3 strains suggest that different strains of *Bacillus subtilis* may activate different pattern of regulatory element for producing their surfactin.

However, the pattern of TFs for surfactin producing genes showed completely different among these 3 strains (MJ01, spizizenii and 168). Therefor we can conclude that the both number and pattern of TFs might be important for regulation of surfactin process.

*lexA* has role in transferring of mobile genetic elements and also involve in formation of biofilm. Moreover, it represses the number of genes involved in response to DNA damage (SOS response) (Mo et al., 2014). Also *rpoD* involves in promoting of RNA polymerase to attach to specific initiation site. In addition, it play a key role in transcription of growth related genes(Shimada et al., 2014). The network analysis revealed that these TFs are hub. Their general role in transcription approves our results.

In addition data mining and decision tree revealed that *flHCD*, *rpoD* and *ihf* are the most important TF in order to distinct surfactin producing gene and other genes in *B.subtilis* subsp. MJ01. Promoter analysis in 3 *B. subtilis* subsp. 168, spizizenii and MJ01 also identified these three TFs (*flHCD*, *rpoD* and *ihf*) as the most important and abundant TF among all three strains.

Hence in this research, the result of both data mining and decision tree among general genes and key surfactin producing genes confirmed the result of TF patterns analysis of key surfactin producing genes between spizizenii, 168 and MJ01. However, all analysis identified *flHCD*, *rpoD* and *ihf* as the most important TF for surfactin producing genes.

Also network analysis for surfactin producing genes occurred. Two modules have been constructed. Module 1 and 2 covered 44% and 28% of global network. Module 1 includes surfactin production key genes (*srfA*, *srfAA*, *srfAB*, *srfAD* and *sfp*) with key

TFs that identified by our promoter analysis such as *rpoD*, *ihf*, *arcA*, *cytR*, *rpoH* and *argR2*. In addition, *ppsA* and *ppsC* which are activated in module 1, encode the nonribosomal peptide synthetase (NRPS) subunits (Du and Shen, 2001). However, module 2 consists of just *srfA* and some regulatory genes relate to producing surfactin such as *comQ* and *comX*.
The isoprenyl transferase *ComQ* modifies the signaling peptide *ComX*. The isoprenylated *ComX* is then secreted (Magnuson et al., 1994) and by the time the concentration reached at critical point the auto-phosphorylation of the membrane-bound *ComP* would be activated, that can phosphorylate the transcriptional activator *ComA* (Weinrauch et al., 1990). Phosphorylated *ComA* directly regulates the expression of various genes, such as the *srfA* operon (Oslizlo et al., 2014). Moreover, *rpoD*, *fis*, *purr*, *argR*, *lexA* and *lrp* are the available TFs in module 2 that our analysis confirmed their role in surfactin production.

## Conclusion

In this study, promoter and network analysis, opened for the first time a new avenue for understanding the molecular mechanism of surfactin production in *Bacillus subtilis*. Specially, comparison between 3 strains of *Bacillus subtilis* (MJ01, spizizenii and 168), revealed that *Bacillus subtilis* subsp. MJ01 which we could isolate from south oil contaminated soil of Iran, have potential to be the novel strain, although it requires more studies.

## References

1. Abdel-Mawgoud A. M., Aboulwafa M. M. and Hassouna N. A.-H. (2008) Characterization of surfactin produced by *Bacillus subtilis* isolate BS5. Applied Biochemistry and Biotechnology 150:289-303.
2. Anuradha S N. (2010) Structural and molecular characteristics of lichenysin and its relationship with surface activity. Biosurfactants:304-315.
3. Banat I. M., Franzetti A., Gandolfi I., Bestetti G., Martinotti M. G., Fracchia L., Smyth T. J. and Marchant R. (2010) Microbial biosurfactants production, applications and future potential. Applied Microbiology and Biotechnology 87:427-444.
4. Calvo J. M. and Matthews R. G. (1994) The leucine-responsive regulatory protein, a global regulator of metabolism in *Escherichia coli*. Microbiological Reviews 58:466-490.
5. Cao X.-h., Wang A.-h., Wang C.-l., Mao D.-z., Lu M.-f., Cui Y.-q. and Jiao R.-z. (2010) Surfactin induces apoptosis in human breast cancer MCF-7 cells through a ROS/JNK-mediated mitochondrial/caspase pathway. Chemico-Biological Interactions 183:357-362.
6. Deihimi T., Niazi A., Ebrahimi M., Kajbaf K., Fanaee S., Bakhtiarizadeh M. R. and Ebrahimie E. (2012) Finding the undiscovered roles of genes: an approach using mutual ranking of coexpressed genes and promoter architecture-case study: dual roles of thaumatin like proteins in biotic and abiotic stresses. SpringerPlus 1:1.
7. Du L. and Shen B. (2001) Biosynthesis of hybrid peptide-polyketide natural products. Current Opinion in Drug Discovery & Development 4:215-228.
8. Ebrahimi M., Ebrahimie E. and Bull C. M. (2015) Minimizing the cost of translocation failure with decision-tree models that predict species' behavioral response in translocation sites. Conservation Biology 29:1208-1216.
9. Hengge-Aronis R. (2002) Signal transduction and regulatory mechanisms involved in control of the σS (RpoS) subunit of RNA polymerase. Microbiology and Molecular Biology Reviews 66:373-395.
10. Lee W. and Chen S. L. (2002) Research report genome-tools: A flexible package for genome sequence analysis. Biotechniques 33:1334-1341.
11. Magnuson R., Solomon J. and Grossman A. D. (1994) Biochemical and genetic characterization of a competence pheromone from B. subtilis. Cell 77:207-216.
12. Mahdi L. K., Deihimi T., Zamansani F., Fruzangohar M., Adelson D. L., Paton J. C., Ogunniyi A. D. and Ebrahimie E. (2014) A functional genomics catalogue of activated transcription factors during pathogenesis of pneumococcal disease. BMC Genomics 15:1.
13. Mo C. Y., Birdwell L. D. and Kohli R. M. (2014) Specificity determinants for autoproteolysis of LexA, a key regulator of bacterial SOS mutagenesis. Biochemistry 53:3158-3168.
14. Mulligan C. N. (2009) Recent advances in the environmental applications of biosurfactants. Current Opinion in Colloid & Interface Science 14:372-378.
15. Nepusz T., Yu H. and Paccanaro A. (2012) Detecting overlapping protein complexes in protein-protein interaction networks. Nature Methods 9:471-472.
16. Nitschke M. and Costa S. (2007) Biosurfactants in food industry. Trends in Food Science & Technology 18:252-259.
17. Oslizlo A., Stefanic P., Dogsa I. and Mandic-Mulec I. (2014) Private link between signal and response in *Bacillus subtilis* quorum sensing. Proceedings of the National Academy of Sciences 111:1586-1591.
18. Pagel J. M., Winkelman J. W., Adams C. W. and Hatfield G. W. (1992) DNA topology-mediated regulation of transcription initiation from the tandem promoters of the ilvGMEDA operon of *Escherichia coli*. Journal of Molecular Biology 224:919-935.

19. Parekh B. S. and Hatfield G. W. (1996) Transcriptional activation by protein-induced DNA bending: evidence for a DNA structural transmission model. Proceedings of the National Academy of Sciences 93:1173-1177.
20. Parekh B. S., Sheridan S. D. and Hatfield G. W. (1996) Effects of integration host factor and DNA supercoiling on transcription from the ilvPG promoter of *Escherichia coli*. Journal of Biological Chemistry 271:20258-20264.
21. Pashaiasl M., Ebrahimi M. and Ebrahimie E. (2016a) Identification of the key regulating genes of diminished ovarian reserve (DOR) by network and gene ontology analysis. Molecular Biology Reports 43:923-937.
22. Pashaiasl M., Khodadadi K., Kayvanjoo A. H., Pashaei-asl R., Ebrahimie E. and Ebrahimi M. (2016b) Unravelling evolution of Nanog, the key transcription factor involved in self-renewal of undifferentiated embryonic stem cells, by pattern recognition in nucleotide and tandem repeats characteristics. Gene 578:194-204.
23. Porob S., Nayak S., Fernandes A., Padmanabhan P., Patil B. A., Meena R. M. and Ramaiah N. (2013) PCR screening for the surfactin (sfp) gene in marine Bacillus strains and its molecular characterization from Bacillus tequilensis NIOS11. Turkish Journal of Biology 37:212-221.
24. Shannon P., Markiel A., Ozier O., Baliga N. S., Wang J. T., Ramage D., Amin N., Schwikowski B. and Ideker T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research 13:2498-2504.
25. Shao C., Liu L., Gang H., Yang S. and Mu B. (2015) Structural diversity of the microbial surfactin derivatives from selective esterification approach. International Journal of Molecular Sciences 16:1855-1872.
26. Shibulal B., Al-Bahry S. N., Al-Wahaibi Y. M., Elshafie A. E., Al-Bemani A. S. and Joshi S. J. (2014) Microbial enhanced heavy oil recovery by the aid of inhabitant spore-forming bacteria: an insight review. The Scientific World Journal 2014.
27. Shimada T., Yamazaki Y., Tanaka K. and Ishihama A. (2014) The whole set of constitutive promoters recognized by RNA polymerase RpoD holoenzyme of *Escherichia coli*. PLoS One 9:e90447.
28. Shoemaker B. A. and Panchenko A. R. (2007) Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLOS Computational Biology 3:e43.
29. Szklarczyk D., Franceschini A., Wyder S., Forslund K., Heller D., Huerta-Cepas J., Simonovic M., Roth A., Santos A. and Tsafou K. P. (2014) STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Research:gku1003.
30. Torkzaban B., Kayvanjoo A. H., Ardalan A., Mousavi S., Mariotti R., Baldoni L., Ebrahimie E., Ebrahimi M. and Hosseini-Mazinani M. (2015) Machine learning based classification of microsatellite variation: an effective approach for phylogeographic characterization of olive populations. PloS One 10:e0143465.
31. Van Schaik W., van der Voort M., Molenaar D., Moezelaar R., de Vos W. M. and Abee T. (2007) Identification of the σB regulon of Bacillus cereus and conservation of σB-regulated genes in low-GC-content gram-positive bacteria. Journal of Bacteriology 189:4384-4390.
32. Weinrauch Y., Msadek T., Kunst F. and Dubnau D. (1991) Sequence and properties of comQ, a new competence regulatory gene of *Bacillus subtilis*. Journal of Bacteriology 173:5685-5693.
33. Weinrauch Y., Penchev R., Dubnau E., Smith I. and Dubnau D. (1990) A *Bacillus subtilis* regulatory gene product for genetic competence and sporulation resembles sensor protein members of the bacterial two-component signal-transduction systems. Genes & Development 4:860-872.
34. Winkelman J. and Hatfield G. W. (1990) Characterization of the integration host factor binding site in the ilvPG1 promoter region of the ilvGMEDA operon of *Escherichia coli*. Journal of Biological Chemistry 265:10055-10060.
35. Yada T., Totoki Y., Ishii T. and Nakai K. 1997. Functional prediction of *B. subtilis* genes from their regulatory sequences. *In* ISMB. Vol. 5. 354-357.
36. Yeh M. S., Wei Y. H. and Chang J. S. (2005) Enhanced Production of Surfactin from Bacillussubtilis by addition of solid carriers. Biotechnology Progress 21:1329-1334.

**Supplementary 1**. %sequence identity of genes associate with key surfactin producing genes in their networks between 168 vs spizizenii and MJ01

| Gene name | Gene definition | Blastn result (% Identity) | |
|---|---|---|---|
| | | Spizizenii | MJ01 |
| *pksN* | polyketide synthase; Involved in some intermediate steps for the synthesis of the antibiotic polyketide bacillaene which is involved in secondary metabolism | 94.38 | 93.90 |
| *pksM* | polyketide synthase; Involved in some intermediate steps for the synthesis of the antibiotic polyketide bacillaene which is involved in secondary metabolism | 93 | 93.03 |
| *pksJ* | polyketide synthase; Involved in some intermediate steps for the synthesis of the antibiotic polyketide bacillaene which is involved in secondary metabolism | 96.68 | 96.27 |
| *pksD* | polyketide synthase (EC:2.3.1.-); Probably involved in some intermediate steps for the synthesis of the antibiotic polyketide bacillaene which is involved in secondary metabolism | 94.20 | 94.12 |
| *pksL* | polyketide synthase; Involved in some intermediate steps for the synthesis of the antibiotic polyketide bacillaene which is involved in secondary metabolism | 92.95 | 92.95 |
| *pksR* | polyketide synthase; Involved in some intermediate steps for the synthesis of the antibiotic polyketide bacillaene which is involved in secondary metabolism | 90.79 | 90.92 |
| *ppsC* | plipastatin synthetase; This protein is a multifunctional enzyme, able to activate and polymerize the amino acids Glu and Ala/Val as part of the biosynthesis of the lipopeptide antibiotic plipastatin. The Ala/Val residue is further epimerized to the D-isomer form. The activation sites for these amino acids consist of individual domains | 89.04 | 89.18 |
| *ppsA* | plipastatin synthetase; This protein is a multifunctional enzyme, able to activate and polymerize the amino acids Glu and Orn as part of the biosynthesis of the lipopeptide antibiotic lipastatin. The Orn residue is further epimerized to the D-isomer form. The activation sites for these amino acids consist of individual domains | 89.27 | 89.33 |
| *comQ* | isoprenyl transferase; Involved in the maturation of ComX, part of a major quorum-sensing system that regulates the development of genetic competence | 81.17 | 80.86 |
| *comX* | competence pheromone precursor (pheromone peptide aa 46->55, modified); Part of a major quorum-sensing system that regulates the development of genetic competence. Acts through the activation of the two-component regulatory system ComP/ComA composed of a sensor histidine kinase, ComP, and a response regulator, ComA, that regulates directly the transcription of over 20 genes. Transport through the membrane may involve Spo0K. Under certain conditions plays a role in sporulation | 87.31 | 87.30 |
| *yndJ* | hypothetical protein | 88.26 | 88.49 |
| *ybdz* | hypothetical protein | 96.19 | 96.19 |
| *menB* | naphthoate synthase (EC:4.1.3.36); Converts o-succinylbenzoyl-CoA (OSB-CoA) to 1,4- dihydroxy-2-naphthoyl-CoA (DHNA-CoA) | 93.53 | 93.41 |
| *dltA* | D-alanine--poly(phosphoribitol) ligase subunit 1 (EC:6.1.1.13); Involved in the biosynthesis of D- | 93.77 | 93.85 |

| | | | |
|---|---|---|---|
| | alanyl-lipoteichoic acid (LTA). Catalyzes an ATP-dependent two-step reaction where it forms a high energy D-alanyl AMP intermediate and transfers the alanyl residues from AMP to Dcp | | |
| *dhbB* | isochorismatase (EC:3.3.2.1) | 92.76 | 93.29 |
| *codY* | transcriptional repressor CodY; DNA-binding protein that represses the expression of many genes that are induced as cells make the transition from rapid exponential growth to stationary phase and sporulation. It is a GTP-binding protein that senses the intracellular GTP concentration as an indicator of nutritional limitations. At low GTP concentration it no longer binds GTP and stop to act as a transcriptional repressor | | |
| *RapC* | response regulator aspartate phosphatase | 95.13 | 95.13 |
| *dhbF* | siderophore 2,3-dihydroxybenzoate-glycine-threonine trimeric ester bacillibactin synthetase (EC:2.7.7.-5.1.1.-); Specifically adenylates threonine and glycine, and loads them onto their corresponding peptidyl carrier domains | 91.89 | 91.93 |
| *pksS* | cytochrome P450; Involved in the metabolism of the antibiotic polyketide bacillaene which is involved in secondary metabolism. The substrate is dihydrobacillaene | 92.53 | 92.69 |
| *mecA* | adaptor protein; Enables the recognition and targeting of unfolded and aggregated proteins to the ClpC protease or to other proteins involved in proteolysis. Acts negatively in the development of competence by binding ComK and recruiting it to the ClpCP protease. When overexpressed, inhibits sporulation. Also involved in Spx degradation by ClpC | 98.48 | 98.78 |
| *comK* | competence transcription factor (CTF); Intermediate regulatory gene required for the expression of the late competence genes comC, comE, comG and the bdbDC operon. Receives signals from SrfA, and possibly other regulatory COM genes, and transduces these signals to the late COM genes | 94.82 | 94.99 |
| *YCZE* | integral inner membrane protein regulating antibiotic production | 95.83 | 95.68 |

**Supplementary 2**. %identity of protein sequence of genes associate with key surfactin producing genes in their networks between 168 vs spizizenii and MJ01

| **Gene name** | **Gene definition** | **Tblastn result (% Identity)** | |
|---|---|---|---|
| | | **Spizizenii** | **MJ01** |
| *pksN* | polyketide synthase; Involved in some intermediate steps for the synthesis of the antibiotic polyketide bacillaene which is involved in secondary metabolism | 90.79 | 90.92 |
| *pksM* | polyketide synthase; Involved in some intermediate steps for the synthesis of the antibiotic polyketide bacillaene which is involved in secondary metabolism | 89.04 | 89.18 |
| *pksJ* | polyketide synthase; Involved in some intermediate steps for the synthesis of the antibiotic polyketide bacillaene which is involved in secondary metabolism | 89.27 | 89.33 |
| *pksD* | polyketide synthase (EC:2.3.1.-); Probably involved in some intermediate steps for the synthesis of the antibiotic polyketide bacillaene which is involved in secondary metabolism | 81.17 | 80.86 |

| | | | |
|---|---|---|---|
| *pksL* | polyketide synthase; Involved in some intermediate steps for the synthesis of the antibiotic polyketide bacillaene which is involved in secondary metabolism | 87.31 | 87.30 |
| *pksR* | polyketide synthase; Involved in some intermediate steps for the synthesis of the antibiotic polyketide bacillaene which is involved in secondary metabolism | 88.26 | 88.49 |
| *ppsC* | plipastatin synthetase; This protein is a multifunctional enzyme, able to activate and polymerize the amino acids Glu and Ala/Val as part of the biosynthesis of the lipopeptide antibiotic plipastatin. The Ala/Val residue is further epimerized to the D-isomer form. The activation sites for these amino acids consist of individual domains | 38.98 | 38.91 |
| *ppsA* | plipastatin synthetase; This protein is a multifunctional enzyme, able to activate and polymerize the amino acids Glu and Orn as part of the biosynthesis of the lipopeptide antibiotic lipastatin. The Orn residue is further epimerized to the D-isomer form. The activation sites for these amino acids consist of individual domains | 38.68 | 38.69 |
| *comQ* | isoprenyl transferase; Involved in the maturation of ComX, part of a major quorum-sensing system that regulates the development of genetic competence | 45.59 | 91.30 |
| *comX* | competence pheromone precursor (pheromone peptide aa 46->55, modified); Part of a major quorum-sensing system that regulates the development of genetic competence. Acts through the activation of the two-component regulatory system ComP/ComA composed of a sensor histidine kinase, ComP, and a response regulator, ComA, that regulates directly the transcription of over 20 genes. Transport through the membrane may involve Spo0K. Under certain conditions plays a role in sporulation | 27.08 | 90.57 |
| *yndJ* | hypothetical protein | 89.78 | 90.15 |
| *ybdz* | hypothetical protein | 97 | 97 |
| *menB* | naphthoate synthase (EC:4.1.3.36); Converts o-succinylbenzoyl-CoA (OSB-CoA) to 1,4- dihydroxy-2-naphthoyl-CoA (DHNA-CoA) | 99.26 | 99.26 |
| *dltA* | D-alanine--poly(phosphoribitol) ligase subunit 1 (EC:6.1.1.13); Involved in the biosynthesis of D-alanyl-lipoteichoic acid (LTA). Catalyzes an ATP-dependent two-step reaction where it forms a high energy D-alanyl AMP intermediate and transfers the alanyl residues from AMP to Dcp | 96.42 | 96.42 |
| *dhbB* | isochorismatase (EC:3.3.2.1) | 94.55 | 94.87 |
| *codY* | transcriptional repressor CodY; DNA-binding protein that represses the expression of many genes that are induced as cells make the transition from rapid exponential growth to stationary phase and sporulation. It is a GTP-binding protein that senses the intracellular GTP concentration as an indicator of nutritional limitations. At low GTP concentration it no longer binds GTP and stop to act as a transcriptional repressor | 100 | 100 |
| *RapC* | response regulator aspartate phosphatase | 97.91 | 97.91 |
| *dhbF* | siderophore 2,3-dihydroxybenzoate-glycine-threonine trimeric ester bacillibactin synthetase (EC:2.7.7.-5.1.1.-); Specifically adenylates threonine and glycine, and loads them onto their corresponding peptidyl carrier domains | 94.07 | 93.95 |

| | | | |
|---|---|---|---|
| *pksS* | cytochrome P450; Involved in the metabolism of the antibiotic polyketide bacillaene which is involved in secondary metabolism. The substrate is dihydrobacillaene | 95.06 | 95.31 |
| *mecA* | adaptor protein; Enables the recognition and targeting of unfolded and aggregated proteins to the ClpC protease or to other proteins involved in proteolysis. Acts negatively in the development of competence by binding ComK and recruiting it to the ClpCP protease. When overexpressed, inhibits sporulation. Also involved in Spx degradation by ClpC | 99.08 | 99.08 |
| *comK* | competence transcription factor (CTF); Intermediate regulatory gene required for the expression of the late competence genes comC, comE, comG and the bdbDC operon. Receives signals from SrfA, and possibly other regulatory COM genes, and transduces these signals to the late COM genes | 99.08 | 99.08 |
| *YCZE* | integral inner membrane protein regulating antibiotic production | 95.81 | 95.81 |