Identification, characterization and purification of MSC_0265, a potential immunogenic antigen homologue of *Mycoplasma mycoides* subsp. *mycoides* in *Mycoplasma capricolum* subsp. *capripneumoniae*

M. Orwe[1*], J. Kinyua[2], J. Ngaira[2], H. Wesonga[3], J. Naessens[3]

[1] Production Department, Kenya Veterinary Vaccines Production Institute,

P.O. Box 53260-00200, Nairobi, Kenya.

[2] Biochemistry Department, Jomo Kenyatta University of Agriculture and Technology,

P.O. Box 62000-00200, Nairobi, Kenya.

[3] Bacteriology Department, Kenya Agricultural and Livestock Research Organization,

P.O. Box 32-00902, Kikuyu, Kenya.

ABSTRACT

***In silico* identification and characterization of vaccine antigens has opened up new frontiers in the field of reverse vaccinology to mitigate the effects of livestock diseases by development of new subunit vaccines. This study aims to characterize, express and purifying MSC_0265 for eventual use in immunoassays and inoculation in goats. *Mycoplasma mycoides subs. Mycoides (Mmm)* and *Mycoplasma capricolum capripneumonia* (*Mccp)* are similar pathogens on the genomic level and are the causative agents of Contagious Bovine Pleuropneumonia (CBPP) in cattle and Contagious Caprine Pleuropneumonia (CCPP) in goats respectively. In this study, BLAST was used to identify the homology of MSC_0265 in *Mycoplasma capricolum capripneumonia* genome and the protein it is similar to. Characterization of MSC_0265 was also done using I-TASSER to predict secondary structure, solvent accessibility, normalised B-factor, 3D models and function. With cut off points of 0.0 for E-value, 100% for Query coverage and 90% for Identity, MSC_0265 a pyruvate dehydrogenase enzyme gave a high homology score on tBLASTn and BLASTp. It had earlier been cloned in pGS21a vector before proceeding with expression and purification of the His-tagged protein by Ni-NTA affinity chromatography. This study identified the homologue of MSC_0265 as protein WP_029333261.1 in the *Mycoplasma capricolum capripneumonia* genome (Accession NZ_LN515398.1) using tBLASTn and BLASTp. Additionally, MSC_0265 was characterized and its optimal expression profile and estimated molecular weight verified.**

**Corresponding author: E-mail: msorwe@gmail.com*

Key words: Mmm; CBPP; Mccp; CCPP; BLAST; I-TASSER; Expression; Purification

1.0 INTRODUCTION

Mycoplasma mycoides subsp. *mycoides* (*Mmm*), previously specified as *Mmm* Small Colony (1) and *Mycoplasma capricolum* subsp. *capripneumoniae* (*Mccp*), are two members of *Mycoplasma 'mycoides* cluster', and are responsible for contagious bovine pleuropneumonia (CBPP) in cattle and contagious caprine pleuropneumonia (CCPP) in goats respectively. Both diseases cause significant losses in livestock, particularly in Africa and parts of Asia, besides being threats to disease-free countries (2).

Studies have shown that species of the *Mycoplasma 'mycoides* cluster' can survive in non-primary hosts with bovine pathogens *Mmm* and *M. leachii* having been isolated from goats (3), and caprine pathogens *M. capricolum* subsp. *capripneumoniae* or *M. mycoides* subsp. *capri* being isolated from cattle (4). Isolation of different *Mycoplasma mycoides* cluster members from a single diseased animal has also been reported (5), which has brought forth the idea of development of different variants arising from exchange of pathogen genetic material among animals kept together like in the case of pastoralism in most parts Africa (6).

CBPP and CCPP are respiratory illness characterized by the presence of sero-fibrinous, interstitial pneumonia, interlobular oedema and hepatization giving a marbled appearance of the lung and capsulated lesions termed sequestra in the lungs of affected cattle and goats respectively. The occurrence of subacute, asymptomatic infections and chronic carriers after the clinical phase of the disease are generally believed to create major problems in the control of this disease. CBPP is present in the Middle East, Asia, and is now considered the most significant bacterial disease of cattle in Africa. The last reported occurrence of CBPP in Europe was in Portugal in 1999. CCPP causes high mortality rates of up to 80% (7).

Vaccines play a key role in the control and prevention of infectious diseases. Pasteur's procedure, to 'isolate, inactivate, and inject' the disease-causing agent, has played a key role in production of inactivated and live attenuated vaccines throughout the 20th century (8). Sequencing of the first bacterial genome, *Haemophilus influenzae* in 1995, and the arrival of the era of microbial genomics has led to many advances in vaccinology. The advent of genome sequencing technologies and use of bioinformatics tools has enabled researchers to explore the genomes of microorganism for purposes of antigen discovery and vaccine development.

The first protein based non-recombinant vaccine was a highly active vaccine to hepatitis B consisting of purified hepatitis B surface antigen (HBsAg) from human plasma (9). The first vaccine to be produced using recombinant DNA technology was licensed in 1986 after the successful expression of HBsAg in yeast (10). This new vaccine efficiently elicited protective antibodies after vaccination of chimpanzees (11), thereby replacing the plasma derived hepatitis B vaccine in human use.

The current CBPP and CCPP vaccines are live attenuated and which have a number of shortfalls hence necessitating the need for research and development of better vaccines. The new approach, currently known as Reverse vaccinology (12), was first described in 2000 to identify the novel meningococcal vaccine candidates from the genome sequence of a *Neisseria meningitidis* serogroup B strain. It involves the *in silico* analysis of the whole genome of the pathogen to identify genes that encode proteins with potential as vaccine candidates. A general starting point is the prediction of surface or intracellular proteins. These selected surface proteins are then expressed in *E. coli* or other

expression vehicles and used to perform immunoassays and animal field challenge to analyse immune responses. However, the applications of reverse vaccinology projects have surpassed just the analysis of a single meningococcal genome (13). Potential advantages this technique has over the current live attenuated vaccines include, being safer, have better efficacy, longer shelf life and are thermostable. Subunit vaccine candidates typically consist of surface proteins or polysaccharides (14).

Advanced sequencing technologies and high throughput proteomics have evolved to include the analysis of several genomes within a species like pan-genomic analysis of group B streptococcus (GBS) (15). Additionally, closely related species for examples comparative genomic analysis of GBS, group A streptococcus (GAS) and *Streptococcus pneumoniae* as well as between pathogenic and commensal members of the same species like comparative genomics of *E. coli* have been analysed (16,17).

In silico identification of *Mmm* protein antigens from the complete genome sequence of PG1 (18) and Gladysdale (19) were done at VIDO – InterVac and a reverse vaccinology approach used to determine protein antigen localization, adhesion probability and finally vaccine candidate prediction of the proteins using Vaxign (20). In total, 58 *Mmm* proteins were identified. Additionally, eight predicted *Mmm* surface proteins that had earlier described elsewhere were also included (21, 22), making a sum total of 66 proteins. These proteins were used to inoculate cattle and their immune responses analysed after challenge. 13 proteins gave the best titres (23).

Other innovative and important vaccine research fields have come up, such as immunoproteomics, structural biology and systems biology. These have become instrumental in vaccine development and in aiding researchers in overcoming the limits of conventional approaches in the discovery and development of novel vaccines especially for microorganisms whose culture has proven a challenge (24, 25, 26).

2.0 MATERIALS AND METHODS

2.1 Download and analysis of protein sequences

Sequences of the 13 potential *Mmm* proteins were downloaded from NCBI (www.ncbi.nlm.nih.gov/) (27) on the protein database drop down tab and saved in FASTA format. Kyoto Encyclopedia of Genes and Genomes (KEGG) database was used to verify the protein sequences. KEGG was accessed on www.genome.jp/kegg (28). Analysis of the predicted PDB protein structure of MSC_0265 was done by I-TASSER as explained on **2.4**.

2.2 Homology based analysis of *Mmm* (PG1) proteins to *Mccp* (F38) proteins Using BLAST in the nucleotide collection and Non-redundant databases

tBLASTn portal was used to search for translated nucleotide databases using a protein query which in this case involved the analysis of homology of the 13 *Mmm* proteins with the *Mccp* genome (Accession number NZ_LN515398.1) with cut off points set at 0.0 for e-value, 100% for query coverage and 90% for Identity. This portal was accessed through; www.blast.ncbi.nlm.nih.gov/Blast.cgi on the Nucleotide collection (nr/nt) database.

BLASTp is a program used to search other protein databases using a protein query. It was used to search for homologues proteins in the *Mccp* (F38 strain) genome. It was accessed through; www.blast.ncbi.nlm.nih.gov/Blast.cgi on the Non-redundant protein sequence (nr) database (29).

2.3 Pairwise alignment of MSC_0265 and WP_029333261.1 using EMBOSS Needle

The *Mmm* protein MSC_0265 was aligned to its *Mccp* homologue WP_029333261.1 using EMBOSS Needle accessed on; www.ebi.ac.uk/Tools/psa/emboss_needle/. EMBOSS Needle is a pairwise alignment program that creates an optimal global alignment of two sequences using the Needleman-Wunsch algorithm. Identity is the percentage of identical matches between the two sequences over the reported aligned region (including any gaps in the length). Similarity is the percentage of matches between the two sequences over the reported aligned region (including any gaps in the length) (30).

2.4 I-TASSER analysis of MSC_0265

I-TASSER (Iterative Threading ASSEmbly Refinement) is a hierarchical approach to protein structure and function prediction. Structural templates are first identified from the PDB by multiple threading approach LOMETS. For each target, I-TASSER simulations generate a large ensemble of structural conformations, called decoys. To select the final models, I-TASSER uses the SPICKER program to cluster all the decoys based on the pair-wise structure similarity, and reports up to five models which corresponds to the five largest structure clusters. The confidence of each model is quantitatively measured by C-score that is calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations. C-score is typically in the range of [-5, 2], where a C-score of a higher value signifies a model with a higher confidence and vice-versa. TM-score and RMSD are estimated based on C-score and protein length following the correlation observed between these qualities. Since the top 5 models are ranked by the cluster size, it is possible that the lower-rank models have a higher C-score in rare cases. Although the first model has a better quality in most cases, it is also possible that the lower-rank models have a better quality than the higher-rank models as seen on benchmark tests. If the I-TASSER simulations converge, it is possible to have less than 5 clusters generated; this is usually an indication that the models have a good quality because of the converged simulations (31, 32, 33).

This server was accessed through http://zhanglab.ccmb.med.umich.edu/I-TASSER for prediction of secondary structure. The secondary structure is predicted based on sequence information from the PSSpred algorithm (33), which works by combing seven neural network predictors from different parameters and PSI-BLAST (29).

Solvent accessibility and normalised B-factor, (B-factor is a value to indicate the extent of the inherent thermal mobility of residues/atoms in proteins). In I-TASSER, this value is deduced from threading template proteins from the PDB in combination with the sequence profiles derived from sequence databases. The reported B-factor profile in the figure below corresponds to the normalized B-factor of the target protein, defined by B=(B'-u)/s, where B' is the raw B-factor value, u and s are respectively the mean and standard deviation of the raw B-factors along the sequence (31, 32, 33).

After the structure assembly simulation, I-TASSER uses the TM-align structural alignment program to match the first I-TASSER model to all structures in the PDB library. This section reports the top 10 proteins from the PDB that have the closest structural similarity, i.e. the highest TM-score, to the predicted I-TASSER model. Due to the structural similarity, these proteins often have similar function to the target. PDB files for analysis were selected based on best hits.

2.5 Expression and purification of MSC_0265

The Qiagen Expressionist and Purification manual (Protocol 7) was used for expression.
BL21 STAR *E.coli* strain containing the histidine tagged protein was grown overnight in 5ml of phytone media with 5µl of Ampicillin (100mg/ml), (Sigma) then transferred to 2 litre conical flasks

containing 1 litre phytone media (BD Biosciences) the following day containing 1ml of Ampicilline (100mg/ml). The shaker–incubator used was Innova 4430 set at 37℃ at 221 RPM. Induction with IPTG 1mM (Thermo Scientific) was done after attainment of 0.25 ODs using the Ultrospec 3100 pro spectrophotometer and 2ml samples taken at 0 hrs (before induction), 2hrs, 4hrs, 6hrs, and 16hrs (Overnight) after induction.

For purification, Protocol 17 on the Qiagen QIA*expressionist*™ manual was used.Purification process was done after lysis of the overnight culture with 8M urea lysis buffer and centrifugation. Cleared lysate from 1 litre culture, Ni-NTA resin, Empty columns and buffers (lysis, wash and elution) were of interest in this process. 50% Ni-NTA Agarose was added to lysate in the ratio of 1:4 and mixed gently by shaking (200 rpm on a rotary shaker) for 15–60 minutes at room temperature. The capacity of each of the columns was about 12ml and were soaked in distilled water before use. For each sample, a 50ml corning tubes was used for the Cell lysate, Flow Through, Wash 1, Wash 2 and Wash 3. 15ml corning tubes were also prepared for Elution 1 and Elution 2. The lysate were then loaded – resin mixture carefully into an empty column. The first, second and third wash were done using 8ml of Buffer C and contents collected in the 50ml Wash 1, 2 and 3 tubes respectively. Elution 1 and 2 were done using 4ml of Buffer D and E respectively and contents collected in corresponding 15ml Corning tubes.
10% separating and 4% stacking gels were then prepared for SDS-PAGE analysis. Staining was done using PageBlue protein staining solution (Thermo Fisher Scientific). The marker of choice was the 10-170 kDa PageRuler™ prestained protein ladder (Thermo Scientific).

2.6 Quantification of MSC_0265

Quantification was done using the BIORAD DC Protein assay consisting of Reagent A (alkaline copper tartrate solution), Reagent B (dilute Folin reagent) and reagent S. A standard 96 well plate together with 8 standard tubes (0.1, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4mg/ml) were used. Qiagen Buffer D was used as blank.

5µl of standards and samples were pipetted into a clean dry microtiter plate, 25µl of reagent A added into each well followed by addition of 200µl of reagent B then plate gently agitated and absorbance (750nm) read after 15 minutes.

3.0 RESULTS AND DISCUSSION

3.1 Homology based analysis of *Mmm* (PG1) proteins to *Mccp* (F38) proteins Using BLAST in the nucleotide collection and Non-redundant databases

3.1.1 tBLASTn analysis

When the 13 *Mmm* protein sequences were subjected to a tBLASTn analysis with cut off parameters set as 0.0 for e-value, 100% for query coverage and 90% for Identity, the proteins MSC_0160 and MSC_0265 gave the best homology scores against the *Mccp* (F38 strain) genome (**Table 1**).

MSC_0160 had an E-value of 0.0, Query coverage of 100% and Identity of 99% while MSC_0265 had an E-value of 0.0, Query coverage of 100% and Identity of 90% (**Table 1**). With scoring parameters of BLOSUM62 as the matrix, Gap costs existence as 11 and Extension as 1 with a conditional compositional score matrix adjustment, an inference can be made that there is shared ancestry between *Mmm* and *Mccp* genomes shown by the high homology scores of the *Mmm* proteins in the *Mccp* genome.

Protein	Function	E-value	Query coverage (%)	Identity (%)	Protein Length (AA)
MSC_0136	Hypothetical protein	6e-88	85	59	322
MSC_0957	Prolipoprotein	2e-151	96	63	433
MSC_0499	Prolipoprotein	0.0	100	70	717
MSC_0431	Prolipoprotein	4e-12	37	46	356
MSC_0776	Prolipoprotein	3e-135	77	56	782
MSC_0816	Hypothetical protein	1e-73	75	61	405
MSC_0160	Elongation factor Tu	0.0	100	99	395
MSC_0775	Prolipoprotein	0.0	100	63	727
MSC_0013	Prolipoprotein	4e-56	76	39	548
MSC_0610	Molecular chaperone Dnak	0.0	94	96	591
MSC_0265	Pyruvate dehydrogenase E1	0.0	100	90	370
MSC_0052	Hypothetical protein	0.0	100	67	691
MSC_0240	Prolipoprotein	1e-149	100	49	554

Table 1: tBLASTn homology results of the 13 *Mmm* (PG1 strain) against the *Mccp* (F38 strain) genome

3.1.2 BLASTp analysis

When MSC_0160 and MSC_0265 protein sequences were subjected to the BLASTp portal on the algorithm of BLASTp (protein-protein BLAST), WP_011166362.1 and WP_029333261.1 were identified as the respective homologues in *Mccp* (F38 strain), (**Table 2**). MSC_0160 was shown to have an E-value of 0.0, Query coverage of 100% and Identity of 100% to the *Mccp* protein WP_011166362.1. The protein of interest for this study, MSC_0265 was shown to have an E-value of 0.0, Query coverage of 100% and Identity of 99% to the *Mccp* proten WP_029333261.1 with similar functions and protein lengths (**Table 2**). This therefore infers presence of conserved gene encoding proteins with similar functions and common ancestry among the two members of the *Mycoplasma* '*mycoides* cluster'.

Protein	Function	E-value	Query coverage (%)	Identity (%)	Homologue Accession number	Protein Length (AA)
MSC_0160	Elongation factor Tu	0.0	100	100	WP_011166362.1	395
MSC_0265	Pyruvate dehydrogenase E1	0.0	100	99	WP_029333261.1	370

Table 2 Homologues of MSC_0160 and MSC_0265 in *Mccp* (F38) genome

3.2 Pairwise alignment of MSC_0265 and WP_029333261.1 sequences using EMBOSS Needle program

The pairwise alignment of MSC_0265 (Accession number NP_975264.1) and WP_029333261.1 on the EMBOSS Needle platform generated an Identity of 368/370 (99.5%), similarity of 369/370 (99.7%) with 0 gaps and gap penalty of 10 on the EBLOSUM62 matrix parameter. Mismatches were at positions 38 and 100 only (**Figure 1**). The high degree of similarity and identity therefore leads to prediction of similar characteristics, features and metabolic functions within the respective genomes.

```
# Aligned_sequences: 2
# 1: NP_975264.1
# 2: WP_029333261.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 370
# Identity:     368/370 (99.5%)
# Similarity:   369/370 (99.7%)
# Gaps:           0/370 ( 0.0%)
# Score: 1946.0
#
#
#=======================================

NP_975264.1         1 MTYLGKFDPLKNEKVCVLDKDGKVINPKLMPKISDQEVLEAYKIMNLSRR     50
                      ||||||||||||||||||||||||||||||||||||||:||||||||||||
WP_029333261.       1 MTYLGKFDPLKNEKVCVLDKDGKVINPKLMPKISDQEILEAYKIMNLSRR     50

NP_975264.1        51 QDIYQNTMQRQGRLLSFLSSTGQEACEVAYINALNKKTDHFVSGYRNNAG    100
                      ||||||||||||||||||||||||||||||||||||||||||||||||||.
WP_029333261.      51 QDIYQNTMQRQGRLLSFLSSTGQEACEVAYINALNKKTDHFVSGYRNNAA    100

NP_975264.1       101 WLAMGQLVRNIMLYWIGNEAGGKAPEGVNCLPPNIVIGSQYSQATGIAFA    150
                      ||||||||||||||||||||||||||||||||||||||||||||||||||||
WP_029333261.     101 WLAMGQLVRNIMLYWIGNEAGGKAPEGVNCLPPNIVIGSQYSQATGIAFA    150

NP_975264.1       151 DKYRKTGGVVVTTTGDGGSSEGETYEAMNFAKLHEVPCIFVIENNKWAIS    200
                      ||||||||||||||||||||||||||||||||||||||||||||||||||||
WP_029333261.     151 DKYRKTGGVVVTTTGDGGSSEGETYEAMNFAKLHEVPCIFVIENNKWAIS    200

NP_975264.1       201 TARSEQTKSINFAVKGIATGIPSIIVDGNDYLACIGVFKEVVEYARKGNG    250
                      ||||||||||||||||||||||||||||||||||||||||||||||||||||
WP_029333261.     201 TARSEQTKSINFAVKGIATGIPSIIVDGNDYLACIGVFKEVVEYARKGNG    250

NP_975264.1       251 PVLVECDTYRLGAHSSSDNPDVYRPKGEFEEMAKFDPLIRLKQYLIDKKL    300
                      ||||||||||||||||||||||||||||||||||||||||||||||||||||
WP_029333261.     251 PVLVECDTYRLGAHSSSDNPDVYRPKGEFEEMAKFDPLIRLKQYLIDKKL    300

NP_975264.1       301 WSDEQQAQLEAEQDKFIADEFAWVEQNKNYDLIDIFKYQYDKMDIFLEEQ    350
                      ||||||||||||||||||||||||||||||||||||||||||||||||||||
WP_029333261.     301 WSDEQQAQLEAEQDKFIADEFAWVEQNKNYDLIDIFKYQYDKMDIFLEEQ    350

NP_975264.1       351 YKEAKEFFEKYPESKEGGHH    370
                      ||||||||||||||||||||
WP_029333261.     351 YKEAKEFFEKYPESKEGGHH    370
```

Position 38

Position 100

Figure 1: Pairwise alignment of MSC_0265 and WP_029333261.1 sequences using EMBOSS Needle program

3.3 I-TASSER analysis of MSC_0265

3.3.1 Secondary structure prediction of MSC_0265 by I-TASSER

The sequence of MSC_0265 was submitted to the I-TASSER suite and the predicted secondary structure suggests that this protein is an alpha-beta protein, which contains 16 alpha-helices (in red) and 7 beta-strands (in blue). “H,” “S,” and “C” indicate helix, strand, and coil, respectively (**Figure 2**). Secondary structure prediction is important in the study of protein structure classification and function.

Submitted Sequence in FASTA format

```
>seq
MTYLGKFDPLKNEKVCVLDKDGKVINPKLMPKISDQEVLEAYKIMNLSRRQDIYQNTMQR
QGRLLSFLSSTGQEACEVAYINALNKKTDHFVSGYRNNAGWLAMGQLVRNIMLYWIGNEA
GGKAPEGVNCLPPNIVIGSQYSQATGIAFADKYRKTGGVVVTTTGDGGSSEGETYEAMNF
AKLHEVPCIFVIENNKWAISTARSEQTKSINFAVKGIATGIPSIIVDGNDYLACIGVFKE
VVEYARKGNGPVLVECDTYRLGAHSSSDNPDVYRPKGEFEEMAKFDPLIRLKQYLIDKKL
WSDEQQAQLEAEQDKFIADEFAWVEQNKNYDLIDIFKYQYDKMDIFLEEQYKEAKEFFEK
YPESKEGGHH
```

Predicted Secondary Structure

```
                    20        40        60        80       100       120       140       160
                     |         |         |         |         |         |         |         |
Sequence    MTYLGKFDPLKNEKVCVLDKDGKVINPKLMPKISDQEVLEAYKIMNLSRRQDIYQNTMQRQGRLLSFLSSTGQEACEVAYINALNKKTDHFVSGYRNNAGWLAMGQLVRNIMLYWIGNEAGGKAPEGVNCLPPNIVIGSQYSQATGIAFADKYRKTGGVV
Prediction  CCCCCCCCCCCCCCSSSSCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCHHHHHHHHHHCCCCCCSSSCCCHHHHHHHHCCCCHHHHHHHHCCCCCCCCCCCCCCCCCCCCHHCCHHHHHHHHHHHHHCCCCCSS
Conf.Score  987555788998845788999988886457999999999999999999999999999999977986787899872999999998659998989837036899997799899999998376514789999962788840306746999999999962979889
            H:Helix; S:Strand; C:Coil
                   180       200       220       240       260       280       300       320
                     |         |         |         |         |         |         |         |
VTTTGDGGSSEGETYEAMNFAKLHEVPCIFVIENNKWAISTARSEQTKSINFAVKGIATGIPSIIVDGNDYLACIGVFKEVVEYARKGNGPVLVECDTYRLGAHSSSDNPDVYRPKGEFEEMAKFDPLIRLKQYLIDKKLWSDEQQAQLEAEQDKFIADE
SSSSCCCHHHCCHHHHHHHHHHHHCCCSSSSSSCCCCSSCCCHHHHHCCCCHHHHHHCCCCSSSSCCCCHHHHHHHHHHHHHHHHCCCCCSSSSSSSSSCCCCCCCCCCCCCCCCHHHHHHHHHCCCHHHHHHHHHHCCCCCHHHHHHHHHHHHHHHHH
999567522000799999999996898999995799402235888707976898745459984798995899999999999999976898889999986068877889998888999999888569699999999879999999999999999999999
                   340       360
                     |         |
FAWVEQNKNYDLIDIFKYQYDKMDIFLEEQYKEAKEFFEKYPESKEGGHH
HHHHHHCCCCCHHHHHHHCCCCCCHHHHHHHHHHHHHHHCCCCCCCCCC
999984899799999982324799860999999999999868564577899
```

Figure 2: Sequence and secondary structure prediction of MSC_0265

3.3.2 Prediction of solvent accessibility

The predicted solvent accessibility is presented in 10 levels, from buried (0) to highly exposed (9) and is predicted by the SOLVE program (Y. Zhang, unpublished). Most surface parts of this protein are buried (**Figure 3**) and are therefore minimally accessible to solvents.

Predicted Solvent Accessibility

```
                    20        40        60        80       100       120       140       1
                     |         |         |         |         |         |         |
Sequence    MTYLGKFDPLKNEKVCVLDKDGKVINPKLMPKISDQEVLEAYKIMNLSRRQDIYQNTMQRQGRLLSFLSSTGQEACEVAYINALNKKTDHFVSGYRNNAGWLAMGQLVRNIMLYWIGNEAGGKAPEGVNCLPPNIVIGSQYSQATGIAFADKYRKTGG
Prediction  553346245262310100055152356723260346201400320010030042024003211000200230000000000200337301000000000000000302032001100023302222442101000000000000000000101124644
            Values range from 0 (buried residue) to 9 (highly exposed residue)
                   180       200       220       240       260       280       300       3
                     |         |         |         |         |         |         |
VTTTGDGGSSEGETYEAMNFAKLHEVPCIFVIENNKWAISTARSEQTKSINFAVKGIATGIPSIIVDGNDYLACIGVFKEVVEYARKGNGPVLVECDTYRLGAHSSSDNPDVYRPKGEFEEMAKFDPLIRLKQYLIDKKLWSDEQQAQLEAEQDKFIA
000000100011000000000000130000000000100001033303405300310200301012011000000130033005202767100000010013120021033641147511640474100520240037342024630540164035104
                   340       360
                     |         |
FAWVEQNKNYDLIDIFKYQYDKMDIFLEEQYKEAKEFFEKYPESKEGGHH
163037345140430042014523640440153036106715763646 76
```

Figure 3: Predicted solvent accessibility

3.3.3 Prediction of Normalised B-Factor

The regions at the N- and C- terminals and most of the loop regions are predicted with positive normalized B-factors, indicating that these regions are structurally more flexible than other regions. On the other hand, the predicted normalized B-factors for the alpha and beta regions are negative or close to zero, suggesting these regions are structurally more stable as shown in **Figure 4**.

B-factor is a value to indicate the extent of the inherent thermal mobility of residues/atoms in proteins. In I-TASSER, this value is deduced from threading template proteins from the PDB in combination with the sequence profiles derived from sequence databases.

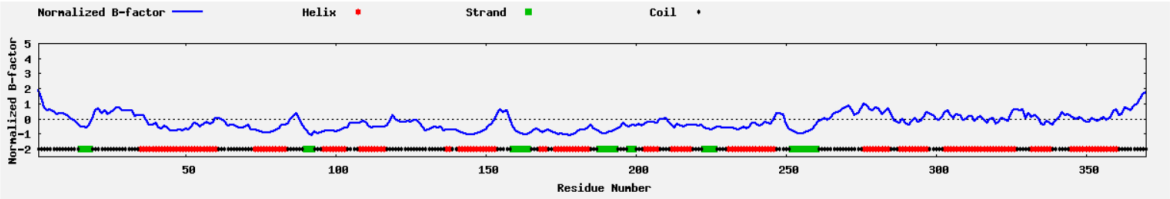

Figure 4: Predicted Normalised B-factor

3.3.4 Predicted 3 dimensional models of MSC_0265

By having the highest C-score of 1.79, **Figure 5** (Model 1) was predicted to be closest to the target protein followed by **Figure 6** (Model 2) with a C-score of -2.99 then finally, **Figure 7** (Model 3) with a C-score of 0.28.

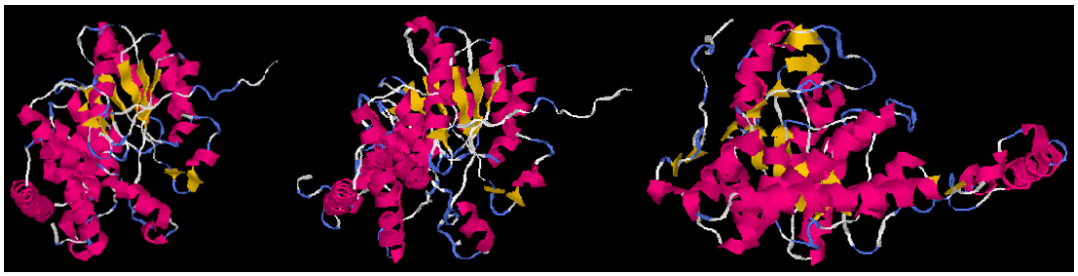

Figure 5: Model 1
C-score=1.79
Estimated TM-score = 0.97±0.05
Estimated RMSD =3.0±2.2Å

Figure 6: Model 2
C-score=-2.99

Figure 7: Model 3
C-score= 0.28

3.3.5 Predicted function using COACH

COACH is based on the I-TASSER structure prediction. COACH is a meta-server approach that combines multiple function annotation results from the COFACTOR, TM-SITE and S-SITE programs.

3.3.5.1 Ligand binding sites

The highest ranked hit below generated a C-score of 0.87, cluster size of 153 denoting number of templates in the cluster with predicted ligand name of TDP. There were 17 ligand binding site residues as shown in **Table 3**

Rank	C-score	Cluster size	PDB hit	Ligand name	Ligand binding site residues
1	0.87	153	1olxA	TDP	94, 95, 96, 135, 136, 137, 165, 166, 167, 168, 171, 193, 195, 197, 198, 199, 264

Table 3: The predicted ligand binding site residues of highest ranked PDB hit

Predicted binding ligands are shown by green yellow sphere and binding residues by blue ball and stick.

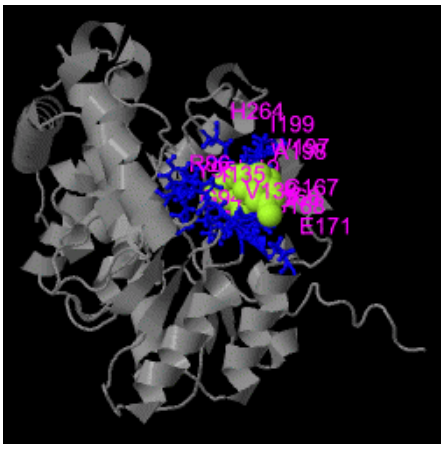

Figure 8: Model of the highest rank predicted ligand binding sites

3.3.6 Enzyme commission numbers (EC) and active sites

The highest ranked PDB hit 1qs0A was generated with a confidence C-score^{EC} of 0.734. TM-score which is the measure of the structural similarity between query and template protein was 0.873. RMSD$^{\alpha}$ is the root mean square division between residues that are structurally aligned by TM-align. The Identity$^{\alpha}$ which was 0.293 is the percentage sequence identity in the structurally aligned region. The Enzyme commission number was 1.2.4.4 and active site residue being 264 (**Table 4**).

Rank	C-score^{EC}	PDB hit	TM-score	RMSD$^{\alpha}$	Identity$^{\alpha}$	Coverage	EC Number	Active site residue
1	0.734	1qs0A	0.873	2.07	0.293	0.932	1.2.4.4	264

Table 4: Predicted enzyme commission numbers (EC) and active site of highest PDB hit

Predicted Enzyme commission active site residues are shown by coloured ball and stick in **Figure 9**

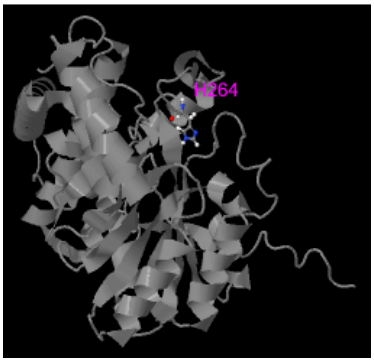

Figure 9: Predicted active site residue of the highest rank model

3.4 Expression and purification of recombinant MSC_0265

3.4.1 Expression of MSC_0265

Figure 10 shows the expression profile of the histidine tagged recombinant protein MSC_0265 on SDS-PAGE of 10% separating gel and 4% stacking gel stained using page blue. There is no leaky expression at 0 hours and thereafter the expression of the protein is consistent over the hours shown. Aliquots of expressed protein at each time zone were loaded on the gel as shown. Phytone media therefore effectively supported the growth of the BL21 STAR (DE3) *E.coli* containing the insert. Its use as a cell factory is well established and studied thereby making it the most popular expression platform. The estimated molecular weight was confirmed by visualisation to be 74kDa.

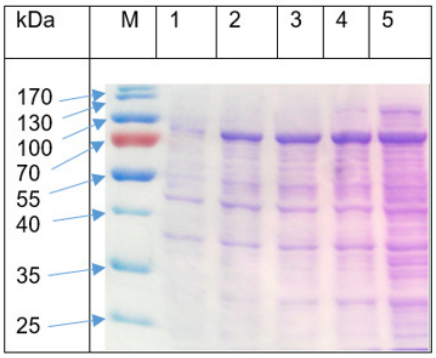

Figure 10: Expression profile of MSC_0265
M: Marker, **1**: 0 hrs, **2**: 2 hrs, **3**: 4hrs, **4**: 6hrs, **5**: 16hrs

3.4.2 Purification of MSC_0265

After expression and lysis of the *E.coli*, an aliquot of the overnight culture was subjected to a purification process**.** Non specific proteins are washed off through the column containing Nickel-NTA Agarose resin during first wash cycles with buffers at higher PH. The histidine tagged protein that attaches to the resin are then eventually eluted at PH 4.5. Efficiency of the process is determined by presence of minimal background on the gel as shown in **Figure 11.**

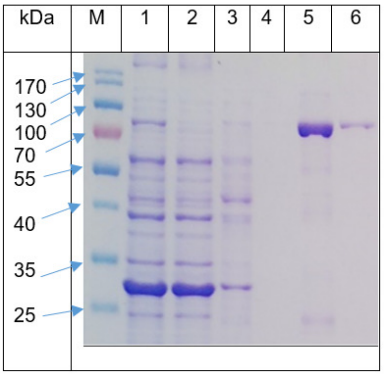

Figure 11: Purification profile of MSC_0265

M: Marker, **1**: Cell lysate, **2**: Flow through, **3**: Wash 1 (PH 6.3), **4**: Wash 2 (PH 6.3), **5**: Elution 1a (PH 4.5), **6**: Elution 1b (PH 4.5)

3.5 Quantification of MSC_0265

Protein quantification is important in the analysis of proteins expression potential especially if a protein would eventually be produced in an industrial setup. It also sheds light on proteins that might need optimization of expression and purification conditions or determination of generally poor expressers. From the yield of a 1 litre culture of MSC_0265, 2.748mg of protein was generated (**Table 5**), which is considered as normal and therefore this infers that this protein's production can be scaled up.

Protein	Concentration (mg/ml)	Volume	Yield/ 1L culture (mg)	Total yield (mg)
His-MSC 0265 E1	0.104	12	1.248	2.748
His-MSC 0265 E1	0.125	12	1.5	-

Table 5: Quantification by use of BioRad Detergent Compatible™ protein assay

4.0 CONCLUSION

MSC_0265, is a pyruvate dehydrogenase E1 protein and first component enzyme of the pyruvate dehydrogenase complex (PDC). The pyruvate dehydrogenase complex facilitates the transformation of pyruvate into acetyl-CoA in a process called pyruvate decarboxylation. Acetyl-CoA may then be used in the citric acid cycle to carry out cellular respiration, therefore pyruvate dehydrogenase contributes to linking the glycolysis metabolic pathway to the citric acid cycle and releasing energy through Nicotinamide adenine dinucleotide (NADH).

MSC_0256 was identified as a potential high score *Mmm* immunogenic homologue in the *Mccp* (F38 strain Accession NZ_LN515398.1) genome as protein WP_029333261. The homologue in *Mccp* was found to have similar sequence length and function. It was characterized and found to have a secondary structure that is alpha-beta in nature with mostly buried solvent accessibility regions. The Histidine-tagged protein, was successfully expressed and purified clearing the way for performing immunoassays, upscale and eventual challenge of goats.

ACKNOWLEDGEMENTS

I would wish to extend my appreciation to technicians Tracy Pryslyak (VIDO), Francis Chuma (ILRI) and Eric Gitonga (KALRO) for their technical input and assistance.

FUNDING

This work was funded by International Development Research Centre (IDRC Project No.107849)

COMPETING INTERESTS

The authors declared that no competing interests exist.

AUTHORS' CONTRIBUTIONS

MO performed the tests and drafting of the manuscript, JK, JN, JN assisted with the structuring and supervision of the activities and manuscript draft. HW assisted with analysis and editing of the manuscript. All authors read and approved the final manuscript.

REFERENCES

1. Manso-Silvan L, Vilei EM, Sachse K, Djordjevic SP, Thiaucourt F, Frey J. Mycoplasma leachii sp. nov. as a new species designation for Mycoplasma sp. Bovine group 7 of Leach, and reclassification of Mycoplasma mycoides subsp. mycoides LC as a serovar of *Mycoplasma mycoides* subsp. *capri*. Int J Syst Evol Microbiol. 2009; 59:1353–1358.
2. Cottew GS, Breard A, DaMassa AJ, Erno H, Leach RH, et al. Taxonomy of the *Mycoplasma mycoides* cluster. Isr J Med Sci. 1987; 23:632–635.
3. Kusiluka LJ, Ojeniyi B, Friis NF, Kokotovic B, Ahrens P. Molecular analysis of field strains of Mycoplasma capricolum subspecies capripneumoniae and Mycoplasma mycoides subspecies mycoides, small colony type isolated from goats in Tanzania. Vet Microbiol. 2001; 82:27–37.
4. Thiaucourt F, Lorenzon S, David A, Breard A. Phylogeny of the Mycoplasma mycoides cluster as shown by sequencing of a putative membrane protein gene. Vet Microbiol. 2000; 72:251–268.
5. DaMassa AJ, Brooks DL, Adler HE. Caprine mycoplasmosis: widespread infection in goats with Mycoplasma mycoides subsp mycoides (large-colony type). Am j vet res. 1983; 44:32-325.
6. Fischer A, Shapiro B, Muriuki C, Heller M, Schnee C, Bongcam-Rudloff E, Vilei EM, Frey J, Jores J. The origin of the 'Mycoplasma mycoides cluster' coincides with domestication of ruminants. PLoS ONE. 2012; 7(4):e36150.
7. OIE Manual of diagnostic tests and vaccines for terrestrial animals, 7th Edition. 2012 (www.oie.int/iternational-standard-setting/terrestial-manual/access-online/.
8. Plotkin SA. Vaccines: The fourth century. Clin Vaccine Immunol. 2009; 16: 1709–1719.
9. Hilleman MR. Vaccines in historic evolution and perspective: A narrative of vaccine discoveries. Vaccine. 2000; 18:1436-1447.
10. Valenzuela P, Medina A, Rutter WJ, Ammerer G, Hall BD. Synthesis and assembly of hepatitis B virus surface antigen particles in yeast. Nature. 1982; Vol.298 No. 22:347-350.
11. McAleer WJ, Buynak EB, Maigetter RZ, Wampler DE, Miller WJ, Hilleman MR. Human hepatitis B vaccine from recombinant yeast. Nature. 1984; 307: 178-180.
12. Rappuoli R. Reverse vaccinology. Curr Opin Microbiol. 2000; 3: 445–450.
13. Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, *et al*. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science. 2000; 287: 1816–1820.
14. Hansson M, Nygren PA, Ståhl S. Design and production of recombinant subunit vaccines. Biotechnol Appl Biochem. 2000; 32 (Pt 2):95-107.
15. Maione D, Margarit I, Rinaudo CD, Margar, Masignani V, Mora M, et al. Identification of a universal Group B streptococcus vaccine by multiple genome screen Science. 2005; 309: 148–150.
16. Moriel DG, Bertoldi I, Spagnuolo A, Marchi S, Rosini R, Nesta B, *et al.* Identification of protective and broadly conserved vaccine antigens from the genome of

extraintestinal pathogenic *Escherichia coli.* Proc Natl Acad Sci. 2010; 107: 9072–9077.
17. Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, *et al.* The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates J Bacteriol. 2008; 190: 6881–6893.
18. Westberg J, Persson A, Holmberg A, Goesmann A, Lundeberg J, Johansson KE, *et al*. The genome sequence of Mycoplasma mycoides subsp. mycoides SC type strain PG1T, the causative agent of contagious bovine pleuropneumonia (CBPP). Genome Res. 2004; 14: 221–227.
19. Wise KS, Calcutt MJ, Foecking MF, Madupu R, DeBoy RT, Roske K, *et al*. Complete genome sequences of Mycoplasma leachii strain PG50T and the pathogenic Mycoplasma mycoides subsp. mycoides small colony biotype strain Gladysdale. J. Bacteriol. 2012; 194: 4448–4449.
20. He Y, Xiang Z, Mobley HL. Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. J.Biomed. Biotechnol. 2010; 297505.
21. Hamsten C, Neiman M, Schwenk JM, Hamsten M, March JB, Persson A. Recombinant surface proteomics as a tool to analyze humoral immuneresponses in bovines infected by Mycoplasma mycoides subsp. mycoides smallcolony type. Mol. Cell. Proteomics. 2009; 8: 2544–2554.
22. Hamsten C, Westberg J, Bolske G, Ayling R, Uhlen M, Persson A. Expression and immunogenicity of six putative variable surface proteins in Mycoplasma mycoides subsp. mycoides SC. Microbiology. 2008; 154: 539–549.
23. Perez-Casal J, Prysliak T, Maina T, Wang Y, Townsend H, Berverov E, *et al*. Analysis of immune responses to recombinant proteins from strains of Mycoplasma mycoides subsp. mycoides, the causative agent of contagious bovine pleuropneumonia Vet. Immunol. Immunopathol. 2015; 168(1-2): 103-10.
24. Bagnoli F, Baudner B, Mishra RP, Bartolini E, Fiaschi L, Mariotti P, *et al.* Designing the next generation of vaccines for global public health OMICS. 2011; 15: 545–566.
25. Rinaudo CD, Telford JL, Rappuoli R, Seib KL. Vaccinology in the genome era J Clin Invest. 2009; 119: 2515–2525.
26. Buonaguro L, Pulendran B. Immunogenomics and systems biology of vaccines Immunol Rev.2011; 239: 197–208.
27. National Center for Biotechnology Information (NCBI). Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 1988. Available from: https://www.ncbi.nlm.nih.gov/
28. Kanehisa M and Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000; 28: 27-30.
29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J. Mol. Biol. 1990; 215: 403–410.
30. Rice P, Longden I and Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16. 2000; 6: 276-277.
31. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008; 9: 40.
32. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nature Protocols. 2010; 5: 725-738.
33. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: Protein structure and function prediction. Nature Methods. 2015; 12: 7-8.