BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 7(6)

Homology modeling, comparative genomics and functional annotation of *Mycoplasma genitalium* hypothetical protein MG_237

Azeem Mehmood Butt[1], Maria Batool[2], Yigang Tong[3]*

[1]Division of Molecular Virology, National Centre of Excellence in Molecular Biology (CEMB), University of the Punjab, Lahore, 53700, Pakistan; [2]Department of Biosciences, COMSATS Institute of Information Technology (CIIT), Islamabad, 44000, Pakistan; [3]State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing, 100071, People's Republic of China; Yigang Tong – Email: tong62035@gmail.com; phone: +86(10)6386 9835; *Corresponding author

Received November 12, 2011; Accepted November 16, 2011; Published November 20, 2011

Abstract:
Mycoplasma genitalium is a human pathogen associated with several sexually transmitted diseases. The complete genome of *M. genitalium* G37 has been sequenced and provides an opportunity to understand the pathogenesis and identification of therapeutic targets. However, complete understanding of bacterial function requires proper annotation of its proteins. The genome of *M. genitalium* consists of 475 proteins. Among these, 94 are without any known function and are described as 'hypothetical proteins'. We selected MG_237 for sequence and structural analysis using a bioinformatics approach. Primary and secondary structure analysis suggested that MG_237 is a hydrophilic protein containing a significant proportion of alpha helices, and subcellular localization predictions suggested it is a cytoplasmic protein. Homology modeling was used to define the three-dimensional (3D) structure of MG-237. A search for templates revealed that MG_237 shares 63% homology to a hypothetical protein of *Mycoplasma pneumoniae,* indicating this protein is evolutionary conserved. The refined 3D model was generated using $(PS)^2$-v2 sever that incorporates MODELLER. Several quality assessment and validation parameters were computed and indicated that the homology model is reliable. Furthermore, comparative genomics analysis suggested MG_237 as non-homologous protein and involved in four different metabolic pathways. Experimental validation will provide more insight into the actual function of this protein in microbial pathways.

Keywords: *Mycoplasma genitalium*; homology modelling; hypothetical proteins; comparative genomics; metabolic pathways;

Background:
Mycoplasma genitalium is a gram-positive disease-causing pathogen. It was first isolated from humans in 1981. It is a common cause of acute and chronic nongonococcal urethritis (NGU), primarily in patients without *Chlamydia trachomatis* infection **[1]**. Studies in non-human primates have clearly demonstrated the pathogenicity of *M. genitalium* in male and female animals. In addition, *in vitro* studies have demonstrated the potential for *M. genitalium* to attach to genital tract epithelial cells using a surface adhesin protein and to enter cells, leading to upregulation of cytokine genes with an associated inflammatory response **[2]**. *M. genitalium* can also attach to spermatozoa, which provides a potential mechanism for spreading to the female upper genital tract. In light of these studies, it has been suggested that *M. genitalium* is a cause of NGU in men, and it may be associated with genital tract inflammatory diseases in women, including cervicitis, pelvic inflammatory disease, and tubal factor infertility **[3]**. The genome of *M. genitalium* consists of 475 protein-coding genes. Among these, there are 94 proteins with no known function and thus referred to as "hypothetical proteins". Recent genome sequencing projects have provided massive amount of data,

however, many of these genomes are still not fully annotated and consist of genes/proteins with unknown function and structure. This is due to several limitations, such as the cost and time required for experimental approaches. An alternative to laboratory-based methods is a bioinformatics approach that utilizes algorithms and databases to estimate protein function. As these algorithms and databases are based on experimental results, they can be an effective means to perform functional and structural annotation of hypothetical proteins. Structure is more evolutionary conserved than sequence; therefore, analysis of three-dimensional (3D) structures holds great potential. Our present study describes the first 3D model of the *M. genitalium* hypothetical protein MG_237 obtained through homology modelling. In addition, primary and secondary sequence structure analysis, functional annotation, comparative genomics, and involvement in bacterial metabolic pathways were also performed.

Methodology:

Sequence analysis and subcellular localization prediction

The amino acid sequence of a *M. genitalium* hypothetical protein MG_237 was retrieved from the SWISS-PROT database **[4]** using the primary accession number P47479 and the entry name, Y237_MYCGE. ProtParam **[5]** was used to predict physiochemical properties. The parameters computed by ProtParam included the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index, and grand average of hydropathicity (GRAVY). Secondary structure predictions (helix, sheets, and coils) were using made using ESPript2.2 **[6]**. Determining subcellular localization is important for understanding protein function and is a critical step in genome annotation. Prediction of subcellular localization of MG_237 was carried out by CELLO v.2.5 **[7]** which is a multi-class support vector machine classification system. Results obtained were also cross-checked with subcellular localization predictions obtained from PSORTb v3.02 **[8]**.

Homology modelling of MG_237

Homology modelling was used to determine the 3D structure of MG_237. A BLASTP **[9]** search with default parameters was performed against the Brookhaven Protein Data Bank (PDB) to find suitable templates for homology modelling. PDB ID: 1TD6_A was identified as the best template based on sequence identity (63%) between query and template protein sequence. The Protein Structure Prediction Server $(PS)^2$-v2 **[10]** was used for homology model construction. $(PS)^2$ is an automated homology modelling server that builds 3D models using the modelling package MODELLER. Multiple sequence alignments between query and template protein sequences made by the $(PS)^2$ server were manually curated before initiating homology modelling of MG_237.

Energy minimization, quality assessment and visualization

Once the 3D model was generated, energy minimization was performed by GROMOS96 force field in a Swiss-PdbViewer **[11]**. Structural evaluation and stereochemical analyses were performed using ProSA-web Z-scores **[12]** and PROCHECK Ramachandran plots **[13]**. Furthermore, Root Mean Squared Deviation (RMSD), superimposition of query and template structure, and visualization of generated models was performed using UCSF Chimera 1.5.3 **[14]**.

Protein structure accession number

The predicted 3D structure of *M. genitalium* hypothetical protein MG_237 was submitted to the Protein Model Database (PMDB) **[15]** and assigned the PMDB ID: PM0077727.

Functional annotation of MG_237

M. genitalium hypothetical protein MG_237 was analysed for the presence of conserved domains based on sequence similarity search with close orthologous family members. For this purpose, three different bioinformatics tools and databases including InterProScan **[16]**, Proteins Families Database (Pfam) **[17]**, and NCBI Conserved Domains Database (NCBI-CDD) **[18]** were used. InterProScan is a tool that combines different protein signature recognition methods native to the InterPro member databases into one resource with look up of corresponding InterPro and GO annotation. Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models. NCBI-CDD is a protein annotation resource that consists of a collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins.

Comparative genomics analysis of MG_237

Workflow of comparative genomics analysis of bacterial genomes as described recently by us **[19]** (accepted, article in press) was used for the analysis of MG_237. At first, BLASTP search was performed between *H. sapiens* proteome and MG_237 to check its similarity to humans. BLASTP search was restricted to proteins from humans only through an option available under BLASTP parameters. Hits were filtered on the basis of expectation value (e-value) inclusion threshold being set to 0.005, and a minimum bit score of 100. Essential proteins of a cellular organism are necessary to live and replicate, and are therefore attractive targets for antimicrobial treatments. Information about essential proteins of *M. genitalium* was retrieved from the Database of Essential Genes (DEG) **[20]**. E-value cut-off of 10^{-10} and a minimum bit score of 100 were used to scan MG_237 against essential proteins listed in DEG from 17 different Gram-positive and Gram-negative bacteria using DEG microbial BLASTP. To check involvement of MG_237 into metabolic pathways of *M. genitalium*, KEGG automatic annotation server (KAAS) was used **[21]**.

Results and Discussion:

The present study focused on sequence, structural and comparative genomics analysis of *M. genitalium* hypothetical protein MG_237. ProtParam was used to analyze different physiochemical properties from the amino acid sequence. The hypothetical protein MG_237 was predicted to be 294 amino acids, with a molecular weight of 34572.1 Daltons and an isoelectric point of 7.69. An isoelectric point above 7 indicates a positively charged protein, and an instability index of 28.33 suggests a stable protein. The negative GRAVY index of -0.235 is indicative of a hydrophilic and soluble protein. The protein sequence was found to be rich in the amino acid leucine, suggesting a preference for alpha-helices in 3D structure. Secondary structure analysis was performed using ESPript **(Figure 1a)** and the protein was predicted to contain several helices, consistent with the ProtParam results **(Figure 1a).** The high percentage of helices in the structure makes the protein more flexible for folding, which might increase protein interactions. Subcellular localization is a key functional attribute

of a protein. Cellular functions are often localized in specific compartments; therefore, predicting the subcellular localization of unknown proteins could be used to obtain useful information about their functions, and to select proteins for further study. Moreover, studying the subcellular localization of proteins is also helpful in understanding disease mechanisms and developing novel drugs **[22]**. The consensus protein subcellular localization predictions suggest that MG_237 is a cytoplasmic protein.

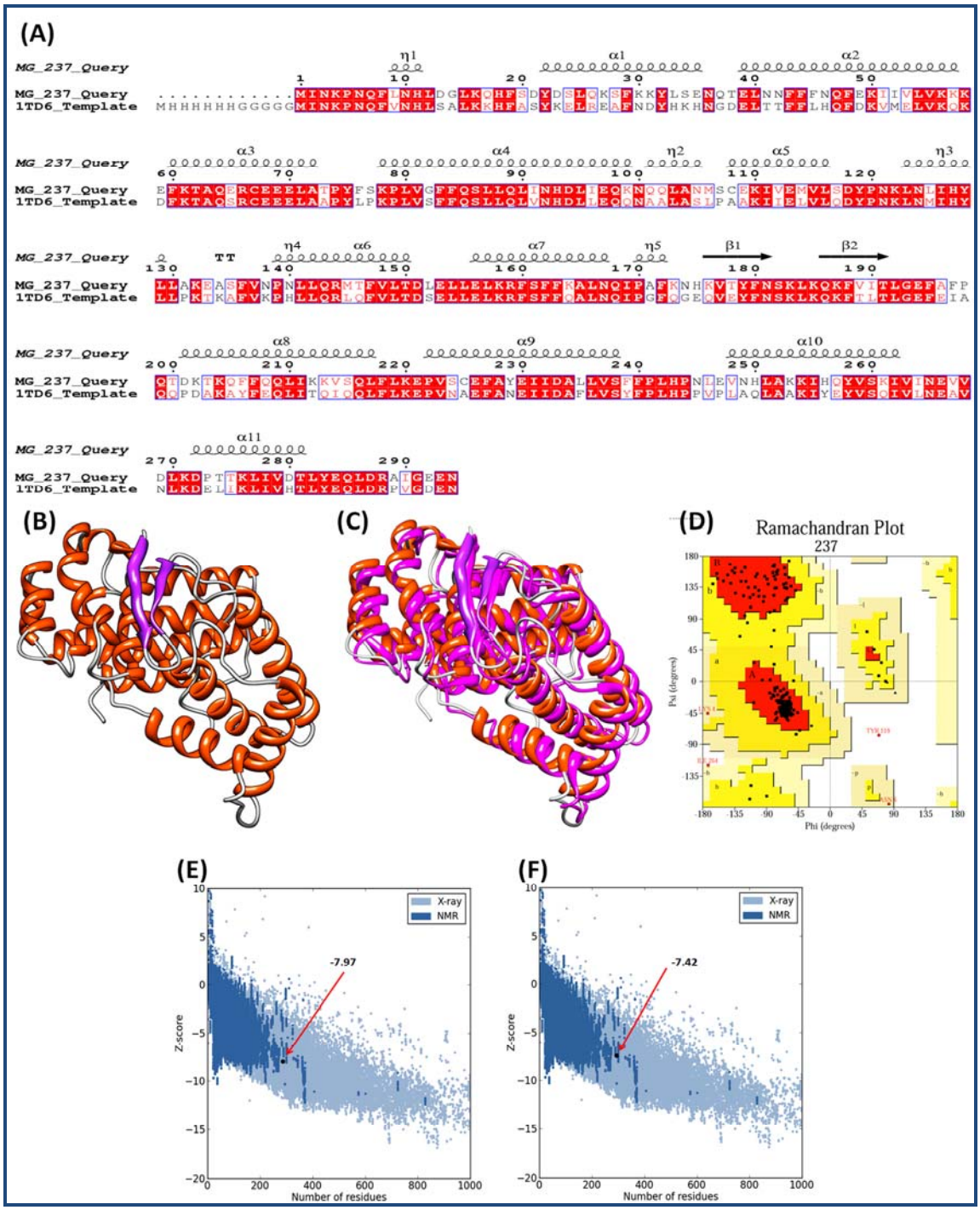

Figure 1: Sequence and structure analysis of *Mycoplasma genitalium* hypothetical protein MG_237. (A) Multiple sequence and structure alignment between protein sequence of MG_237 and the selected template PDB ID: 1td6_A. Conserved regions between template and query are highlighted in red colour. (B) Homology model of MG_237. Alpha helices are shown in orange, beta sheets in purple and coils in grey colour. (C) Superimposed structure of MG_237 and template 1td6_A. Template is shown in pink colour. (D) Ramachandran plot of MG_237 (E) Z-score of MG_237. (F) Z-score of template 1td6_A.

Homology modelling of MG_237

Protein 3D structures can provide us with precise information of how proteins interact and localize in their stable conformation. Homology or comparative modelling is one of the most common structure prediction methods in structural genomics and proteomics. Numerous online servers and tools have become available for homology or comparative modeling of proteins in past years. Despite minimal modifications, one initial step that is common in all modeling tools and servers is to find the best matching template by performing a sequence homology search with BLASTP. Templates are experimentally determined 3D structures of proteins that share sequence similarity with the query sequence. The template sequence and the protein sequence whose structure is to be determined are aligned using multiple sequence alignment algorithms **[23]**. A well-defined alignment is very important for the prediction of a reliable 3D structure. The genome of *M. genitalium* consists of 94 hypothetical proteins without any known function or structure. A BLASTP search was performed for each protein sequence against the PDB to identify templates for homology modeling. MG_237 was selected for homology modeling as it showed maximum identity to its 1td6_A, which is an X-Ray diffraction model of a *M. pneumoniae* hypothetical protein. The query sequence and template ID was then given as input to the $(PS)^2$ server for homology modeling using MODELLER.

Energy minimization, quality assessment and visualization

The predicted 3D structure of MG_237 is shown in **Figure 1b**. Even though there were no steric clashes in the structure generated, it was still subject to energy minimization and assessed for both geometric and energy aspects. The positioning of alpha-helices and beta-sheets was then compared using ESPript2.2. Secondary structure elements were found to be comparable to that of the template **(Figure 1a**). Eleven helices and two beta sheets were predicted in the 3D structure of MG_237, which implies that it is rich in helical structures (Fig. 1a and 1b). Several structure assessment methods including RMSD, Z-scores, and Ramachandran plots were used to check reliability of the predicted 3D model. The RMSD value indicates the degree to which two 3D structures are similar. The lower the value, the more similar the structures. Both template and query structures were superimposed for the calculation of RMSD **(Figure 1c).** The RMSD value obtained from superimposition of MG_237 and 1TD6 in UCSF Chimera was found to be 0.213 Å, suggesting a reliable 3D structure. The Z-score is indicative of overall model quality and is used to check whether the input structure is within the range of scores typically found for native proteins of similar size. Z-scores of the template and query model were obtained from PROSA-web. The template Z score was -7.97 **(Figure 1e)** and for the MG-237 homology model it was -7.42 **(Figure 1f),** suggesting similarity between template and query structure. Finally, the Ramachandran plots were obtained for both the homology model and the template as a quality assessment. PROCHECK displayed 91% of residues in the most favored regions, with 7.6%, 0.7%, and 0.7% residues in additionally allowed, generously allowed and disallowed regions, respectively **(Figure 1d).** This indicated that the backbone dihedral angles, phi and psi, in the MG_237 3D model, were reasonably accurate. The Ramachandran plot for the template structure showed the amino acid residues to be 84.2%, 14.0%, 1.9% and 0.0%, in most favored, additionally allowed, generously allowed and disallowed regions respectively (Data not shown). The comparable Ramachandran plot characteristics, RMSD values, and Z-scores confirm the quality of the homology model of MG_237. The final protein structure was deposited in PMDB and is available under ID: PM0077727.

Functional annotation and comparative genomics analysis of MG_237

Currently, there is no known function of MG_237 is known. In the present study, a systematic workflow consisting of several bioinformatics tools and databases was defined and used with the goal of performing functional annotation of MG_237. Three web tools were used to search the conserved domains and potential function of Mg_237. Based on consensus predictions made by Pfam, NCBI-CDD and InterProScan, it is suggested that MG_237 contains DUF3196 domains and is currently classified as protein of unknown function. Once the functional annotation of hypothetical was performed, we applied comparative genomics approach to further characterize MG_237. This involved search against human proteome, essentiality estimation, and involvement in metabolic pathways. At first, a BLASTP search against human proteome was performed to identify whether MG_237 has any human homologues. It was identified that MG_237 is a unique protein of *M. genitalium* and showed no homology to any of the human proteins. Proteins with no homology to human proteins can effectively be used as drug targets as targeting these proteins will not have any side effects. Identification of proteins that regulate key factors, such as nutrient uptake, virulence and pathogenicity, is of great importance for disruption of pathogen functions and existence. Such proteins are termed as essential for the pathogen. Again, not all essential proteins are non-homologous in nature. Therefore, pathogen proteins that fulfil the criteria of being unique and essential at the same time represent more attractive drug targets. The information about essential genes of *M genitalium* was retrieved from DEG database. Microbial BLASTP search as per selection criteria mentioned in materials & methods section, suggested that it is a non-essential protein. Finally, KEGG was used to identify the involvement of MG_237 in *M. genitalium* metabolic pathways. Based on search performed via KAAS, MG_237 was found to be involved in four metabolic pathways namely; biosynthesis of secondary metabolites, microbial metabolism in diverse environments, glycolysis / gluconeogenesis, and amino sugar and nucleotide sugar metabolism.

Conclusion:

We have used homology modelling and comparative genomics approach to propose the first 3D structure and possible functions for the *M. genitalium* hypothetical protein MG_237. With the assistance of a well-defined structure and annotations, we can predict protein functional and binding sites, which can help in understanding what biological role it fulfils. It is expected that several of these hypothetical proteins may play important roles, including cell survival, pathogenesis, and antibiotic resistance. Additionally, the workflow described in this study can also be applied to other hypothetical proteins.

Acknowledgements:
This work was supported by the Hi-Tech Research and Development (863) Program of China (No. 2009AA02Z111), and the National Natural Science Foundation of China (No. 30872223). We are thankful to Mr. Imran Mehmood Butt for his assistance in preparing graphical illustrations.

Competing interests:
The author(s) declare that they have no competing interests.

References:

[1] Jensen JS. *J Eur Acad Dermatol Venereol*. 2004 **18:** 1 [PMID: 14678525]
[2] Zhang S *et al. FEMS Immunol Med Microbiol*. 2000 **27:** 1 [PMID: 10617789]
[3] Manhart LE *et al. J Infect Dis*. 2003 **187:** 4 [PMID: 12599082]
[4] Boeckmann B *et al. Nucleic Acids Res*. 2003 **31:** 1 [PMID: 12520024]
[5] Wilkins MR *et al. Methods Mol Biol*. 1999 **112:** 531 [PMID: 10027275]
[6] Gouet P *et al. Bioinformatics*. 1999 **15:** 4 [PMID: 10320398]
[7] Yu CS *et al. Proteins*. 2006 **64:** 3 [PMID: 16752418]
[8] Yu NY *et al. Bioinformatics*. 2010 **26:** 13 [PMID: 20472543]
[9] Altschul SF *et al. Nucleic Acids Res*. 1997 **25:** 17 [PMID: 9254694]
[10] Chen CC *et al. BMC Bioinformatics*. 2009 **10:** 366 [PMID: 19878598]
[11] Guex N, Peitsch MC. *Electrophoresis*. 1997 **18:** 15 [PMID: 9504803]
[12] Wiederstein M & Sippl MJ, *Nucleic Acids Res*. 2007 **35:** W407 [PMID: 17517781]
[13] Laskowski RA *et al. Journal of biomolecular NMR*. 1996 **8:** 4 [PMID: 9008363]
[14] Pettersen EF *et al. Journal of computational chemistry*. 2004 **25:** 13 [PMID: 15264254]
[15] Castrignano T *et al. Nucleic Acids Res*. 2006 **34:** D306 [PMID: 16381873]
[16] Zdobnov EM & Apweiler R, *Bioinformatics*. 2001 **17:** 9 [PMID: 11590104]
[17] Finn RD *et al. Nucleic Acids Res*. 2010 **38:** D211 [PMID: 19920124]
[18] Marchler-Bauer A *et al. Nucleic Acids Res*. 2011 **39:** D225 [PMID: 21109532]
[19] Butt AM *et al. Infect Genet Evol*. 2011 [PMID: 22057004]
[20] Zhang R *et al. Nucleic Acids Res*. 2004 **32:** D271 [PMID: 14681410]
[21] Moriya Y *et al. Nucleic Acids Res*. 2007 **35:** W182 [PMID: 17526522]
[22] Wang J *et al. BMC Bioinformatics*. 2005 **6:** 174 [PMID: 16011808]
[23] Butt AM *et al. African Journal of Biotechnology*. 2011 **10:** 38

Edited by P Kangueane
Citation: Butt *et al.* Bioinformation 7(6): 299-303 (2011)