Genome-Wide Protein Structure Characterization of
*Mycoplasma genitalium*

Jun-Wei Hsu¹
b90064@csie.ntu.edu.tw

Chen-hsiung Chan¹
frankch@life.nthu.edu.tw

Cheng-Yan Kao¹,²
cykao@iii.org.tw

¹ Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan
² Institute for Information in Industry, Taipei, Taiwan

**Keywords:** fold recognition, minimum gene set, gene ontology

1 Introduction

*Mycoplasma genitalium* is an organism with a small genome. Only less than 500 ORFs are present in *M. genitalium* genome. Previous studies have attempted to identify a minimal gene set in *M. genitalium* [1]. It has been shown that *M. genitalium* may retain vitality with around 256 genes.

In recent years, the structural genomics initiatives around the world have focused on solving protein structures in an attempt to cover most protein folds. So far, more than 30,000 entries of protein structures and models are available in the Protein DataBank (PDB).

It is interesting to know how complete is the genomic coverage of protein structures in *M. genitalium*. The lack of homology models for some proteins may indicate valuable targets for protein structure determination. The structure coverage also lends possibility for genome-wide modeling of protein-protein interaction networks with protein-protein docking [2].

2 Method and Results

We have mapped protein sequences of *M. genitalium* to protein structures with PSI-BLAST [3]. The genome sequences of *M. genitalium G-37* is downloaded from National Center for Biotechnology and Information (NCBI, http://www.ncbi.nlm.nih.gov/). There are 484 ORFs in *M. genitalium* genome. The sequences of PDB entries are downloaded from Research Collaboratory for Structural Bioinformatics (RCSB, http://www.rcsb.org/). The Gene Ontology (GO) [4] annotations of proteins with or without homologous structures are investigated and analyzed.

2.1 Fold recognition through sequence-profile alignment

The protein sequences of *M. genitalium* genome are searched against PDB entries using PSI-BLAST. Maximum 5 iterations of PSI-BLAST search are performed for each protein sequence. The results of BLAST searches are parsed. Alignments with significant E-values and covering large fraction of the query sequences are considered as matches. Out of the 484 proteins, 56 failed to find any matches with E-values smaller than 10. There are 293 proteins (60.53%) matched to protein structures with significant E-values (< 10⁻⁵). These proteins have homologous structures available. For other proteins, other fold recognition methods like threading may be necessary to identify homologous structures.

2.2 Structure and function characterization

In order to further analyze the structures and functions of *M. genitalium* proteins, the information in the Gene Ontology are used. The proteins are characterized based on level 1 Molecular Function of GO annotations.
These categories include antioxidant activity, binding, catalytic activity, etc. The characterizations are performed against all 484 proteins in the *M. genitalium* genome and proteins with homologous structures, respectively. The results are summarized in Fig. 1.

From Fig. 1 it is clear that the homologous structures for *M. genitalium* are biased toward catalytic activity and the fraction of unknown function is smaller. The structures of proteins with transporter activity are more difficult to determine because many of them are membrane proteins, and this is reflected in our results.

### 3 Discussion

Our analysis on the structure availability of *M. genitalium* has suggested that some of the structures of critical proteins are still missing. However, the current structural coverage of *M. genitalium* proteins is enough for genome-wide modeling of protein-protein interactions using docking. The protein-protein interaction networks of *Mycoplasma genitalium* are likely to provide more information for the essence of life.

### References


