Tuberculosis (TB) still remains the largest killer infectious disease despite the availability of several chemotherapeutic drugs and vaccines. In Iraq the incidence rate was estimated to be 45/100,000 [1]. Antibiotic resistance is very high, in 2008, it has been estimated that 6.6% of isolates were resistant to four drugs in use (Isoniazid, Rifampicin, Streptomycin and Ethambutol) [2], besides the existence of strains which were resistant to mono-, di-, and tri-antibiotics at a high rate, however, WHO estimated MDR (Multi-drug resistance) at 3.4% in 2011 in new TB cases [3]. Iraq has been identified as middle TB burden country in the world and ranks 8th of 22 EMR (Eastern Mediterranean Regions) countries according to estimated incidence of all types of TB [3].

1. Introduction

Despite ever-increasing amounts of biological data from high-throughput experiments and increasing sequencing projects at exponential rates [4], there is a deficiency in functional annotation for many newly sequenced proteins, for example, in bacterial genomes as many as 40% (or more) of proteins labeled "Uncharacterized protein", "Unknown" or "Hypothetical proteins" [5]. The deluge of data needs an efficient computational method to extract information from these data. So, one of the major tasks in post-genomic era of genome annotation is assigning functions of the gene products based mostly on amino acid sequences in order to capitalize on the knowledge gained through these sequencing efforts [6].

Again in Iraq, different studies on TB are carrying on in association with WHO and San Raffaele Scientific Institute / Italy...
2. Materials and methods

2.1 Sequence Retrieval: Hypothetical proteins (List) of M. tuberculosis H37Rv strain were downloaded from database of this strain (http://www.imtech.res.in/raghava/mycoprint/mtb-browse.html), and protein sequence were obtained from Tuberculist database (http://tuberculist.epfl.ch).

2.2 Prediction of Major Functions of Proteins: VICMPred server (http://www.imtech.res.in/raghava/vicmpred/submission.html) was used [10].

2.3 Prediction of Subcellular Localization of Proteins: The Cello available online tool was used in this study “Cello v2.0 http://cello.life.nctu.edu.tw/” [11].

2.4 Annotation via Orthology: The egg NOG database and egg NOG 2.0 software http://eggnog.embl.de/version_2/ were used [12], and groups were classified [13,14].

2.5 Annotation via Gene Ontology: The GO database release (2013) was used via AmiGO v1.8 http://amigo.geneontology.org/cgi-bin/amigo/go.cgi [15, 16].

2.6 UniProt Database: http://www.uniprot.org/ was used to complete some predictions which were not covered by other approaches [17].

3. Results

Database for virulent strain M. tuberculosis H37Rv was used to get the protein sequences, and only protein designated hypothetical were chosen, others labeled conserved hypothetical, putative, unknown proteins were excluded. The numbers of hypothetical proteins were 617 out of total proteins 3918, this means that hypothetical proteins consist 15.75%, while conserved hypothetical proteins were 1533 (23.4%). So, all hypothetical proteins compromised 39.13% (Table S1).

Another batch of hypothetical proteins (Rvs) were exclude as they have been investigated by others [18]. So the work was concentrated on the remaining Rvs (303 proteins), (Supplementary Table S2).

The distribution of studied proteins into major functional groups like: cellular process (Which include cell division, cell envelope biogenesis, cell motility and signal transduction), information and storage molecules (include transcription, translation and DNA replication), metabolic molecules (include energy production, carbohydrates, amino acids, nucleotides, lipid transport and metabolism), in addition to prediction of virulence factors was performed on all protein are shown in Figure 1, using VICMPred [10,19]. Cellular localization of proteins is shown in Figure 2, using Cello software [11]. Distribution among COG groups shown in Figure 3.
“J, translation, including ribosome structure and biogenesis; L, replication, recombination and repair; K, transcription; O, molecular chaperones and related functions; M, cell wall structure and biogenesis and outer membrane; N, secretion, motility and chemotaxis; T, signal transduction; P, inorganic ion transport and metabolism; C, energy production and conversion; G, carbohydrate metabolism and transport; E, amino acid metabolism and transport; H, coenzyme metabolism; I, lipid metabolism; D, cell division and chromosome partitioning; R, general functional prediction only; S, no functional prediction.”

GO annotation of proteins is shown in Table 1. It was possible to annotated 102 proteins (33.66%). Therefore, UniProt database was used to get more annotated proteins as shown in Table 2.
### Table 1: Results of GO prediction

| Rv ID     | Prediction          | GO results                                                                 |
|-----------|---------------------|-----------------------------------------------------------------------------|
| Rv0008c   | Cell wall synthesis | protein CwsA                                                                |
| Rv0011c   | Cell division       | protein CrgA                                                                |
| Rv0203    | Possible exported   | protein                                                                      |
| Rv0240    | Probable ribonuclease | VapC24                                                                      |
| Rv0272c   | Similarity to      | hypothetical protein Rv0272c                                                |
| Rv0298    | Antitoxin           |                                                                              |
| Rv0299    | Toxin               |                                                                              |
| Rv0300    | Antitoxin VapB2     |                                                                              |
| Rv0398c   | Secreted protein    |                                                                              |
| Rv0550c   | Antitoxin VapB3     |                                                                              |
| Rv0582    | Probable ribonuclease | VapC26                                                                      |
| Rv0636    | (3R)-hydroxyacyl-ACP dehydratase subunit HadB |
| Rv0662c   | Putative antitoxin  | VapB7                                                                       |
| Rv0666    | Possible membrane   | protein                                                                      |
| Rv0678    | Conserved protein   |                                                                              |
| Rv0736    | Anti-sigma-L factor | RslA                                                                        |
| fadB      | Probable fatty      | oxidation protein FadB                                                       |
| Rv0860    | Uncharacterized     | membrane protein ArfB                                                        |
| Rv0909    | Antitoxin Rv0909/MT0933 |                                                                             |
| Rv0923c   | Similarity to      | hypothetical protein Rv0923c                                                |
| Rv0948c   | Intracellular       | chorismate mutase                                                            |
| Rv1024    | Cell division       | protein DivIC                                                                |
| Rv1035c   | Transposase         |                                                                              |
| Rv1041c   | Transposase         |                                                                              |
| Rv1103c   | Antitoxin MazE3     |                                                                              |
| Rv1113    | Antitoxin VapB32    |                                                                              |
| Rv1155    | Putative pyridoxine | pyridoxamine 5'-phosphate oxidase                                            |
| Rv1174c   | T-cell antigen      |                                                                              |
| Rv1222    | RNA polymerase sigma-70 factor, ECF subfamily |                              |
| Rv1231c   | Membrane protein    |                                                                              |
| Rv1234    | Transmembrane protein |                                                                             |
| Rv1302    | Decaprenyl-phosphate | N-acetylglucosamine phosphotransferase                                       |
| Rv1390    | DNA-directed RNA polymerase subunit omega |                              |
| Rv1476    | Membrane protein    |                                                                              |
| Rv1825    | UPF0749 protein Rv1825/MT1873 |                              |
| Rv1871c   | Deazaflavin-dependent nitroreductase |                              |
| Rv1885c   | Secreted chorismate mutase |                                                                             |
| Rv1926c   | Immunogenic protein | MPT63                                                                        |
| Rv1955    | Toxin HigB          |                                                                              |
| Rv1957    | SecB-like chaperone | Rv1957                                                                       |
| Rv1960c   | Antitoxin ParD1     |                                                                              |
| Rv2054    | Carboxymethylenbutenolidase |                              |
| Rv2142c   | Toxin ParE2         |                                                                              |
| Rv2146c   | YggT family protein |                                                                              |
| Rv2147c   | Cell division       | protein SepF                                                                  |
| Rv2159c   | Alkylhydroperoxidase | AhpD family core domain-containing protein                                   |
| Rv2235    | Uncharacterized     | SURF1-like protein Rv2235/MT2294                                             |
| Protein ID | Description                                      |
|------------|--------------------------------------------------|
| Rv2253     | Secreted protein                                 |
| Rv2274c    | Putative toxin MazF8                             |
| Rv2275     | Cyclo(L-tyrosyl-L-tyrosyl) synthase               |
| Rv2290     | Putative lipoprotein LppO                         |
| Rv2319c    | Universal stress protein                         |
| Rv2453c    | Probable molybdenum cofactor guanylyltransferase |
| Rv2476c    | NAD-specific glutamate dehydrogenase             |
| Rv2515c    | DNA-binding protein, putative                    |
| Rv2549c    | Probable ribonuclease VapC20                     |
| Rv2550c    | Antitoxin VapB20                                 |
| Rv2553c    | Probable conserved membrane protein              |
| Rv2554c    | Putative Holliday junction resolvasse             |
| Rv2566     | Long conserved protein                            |
| Rv2573     | Putative 2-dehydropantoate 2-reductase            |
| Rv2616     | Conserved protein                                 |
| Rv2617c    | Probable transmembrane protein                   |
| Rv2618c    | Fluoroquinolones export permease protein         |
| Rv2731     | Conserved alanine and arginine rich protein      |
| Rv2819c    | CRISPR type III-associated RAMP protein Csm5     |
| Rv2949c    | Chorismate--pyruvate lyase                       |
| Rv3004     | Low molecular weight protein antigen 6 cfp6      |
| Rv3038c    | Conserved protein                                 |
| Rv3040c    | Conserved protein                                 |
| Rv3091     | Conserved protein                                 |
| Rv3113     | HAD hydrolase, family IA                         |
| Rv3191c    | Transposase                                      |
| Rv3358     | Toxin RelK                                       |
| Rv3378c    | Diterpene synthase                               |
| Rv3386     | Transposase                                      |
| Rv3413c    | Anti-sigma-D factor RsdA                         |
| Rv3427c    | Putative ATP-binding protein Rv3427c in insertion sequence |
| Rv3428c    | Putative transposase Rv3428c                     |
| Rv3437c    | Possible conserved transmembrane protein          |
| Rv3479     | Possible transmembrane protein                   |
| Rv3493c    | Conserved hypothetical Mce associated alanine and valine rich protein |
| Rv3541c    | Conserved protein                                 |
| Rv3552     | Putative CoA-transferase subunit beta Rv3552/MT3656 |
| Rv3587c    | Probable conserved membrane protein              |
| Rv3611     | Hypothetical arginine and proline rich protein   |
| Rv3632     | Possible conserved membrane protein              |
| Rv3669c    | Probable conserved transmembrane protein         |
| Rv3675     | Possible membrane protein                        |
| Rv3689     | Probable conserved transmembrane protein         |
| Rv3690     | Probable conserved membrane protein              |
| Rv3691     | Uncharacterized membrane protein Rv3691          |
| Rv3698     | Conserved protein                                 |
| Rv3705c    | Conserved protein                                 |
| Rv3808c    | Galactofuranosyl transferase GlfT2               |
| Rv3835     | Uncharacterized membrane protein Rv3835/MT3943    |
| Rv3849     | Nucleoid-associated protein EspR                 |
| Rv3879c    | ESX-1 secretion-associated protein EspK          |
| Rv3885c    | ESX-2 secretion system protein EccE2             |
Table 2: Results of UniProt database

| Rv ID   | Description                                                  |
|---------|--------------------------------------------------------------|
| Rv0008c | Cell wall synthesis and cell shape protein A                 |
| Rv0048c | Membrane protein                                             |
| Rv0090  | Membrane protein                                             |
| Rv0240  | Probable ribonuclease VapC                                   |
| Rv0298  | Antitoxin                                                   |
| Rv0300  | Antitoxin                                                   |
| Rv0420c | Transmembrane protein                                        |
| Rv0550c | Antitoxin                                                   |
| Rv0582  | Probable ribonuclease VapC                                   |
| Rv0664  | Antitoxin                                                   |
| Rv0692  | Mycofactocin system RPExFGAL protein                         |
| Rv0887c | PhnB protein                                                 |
| Rv0948c | Chorismate mutase                                            |
| Rv0961  | Membrane protein                                             |
| Rv0962c | Lipoprotein lprP                                             |
| Rv1103c | Antitoxin                                                   |
| Rv1113  | Antitoxin                                                   |
| Rv1324  | Thioredoxin                                                  |
| Rv1494  | Antitoxin                                                   |
| Rv1692  | HAD hydrolase, family IIA                                    |
| Rv1721c | Antitoxin                                                   |
| Rv1744c | Membrane protein                                             |
| Rv1885c | Chorismate mutase                                            |
| Rv1888c | Transmembrane protein                                        |
| Rv2081c | Transmembrane protein                                        |
| Rv2142c | Toxin                                                        |
| Rv2232  | 5'-nucleotidase                                              |
| Rv2253  | Secreted protein                                             |
| Rv2274c | Toxin                                                        |
| Rv2525c | Tat (Twin-arginine translocation) pathway signal sequence    |
| Rv2532c | N utilization substance protein B                            |
| Rv2549c | Probable ribonuclease VapC                                   |
| Rv2550c | Antitoxin                                                   |
| Rv2597  | Membrane protein                                             |
| Rv3387  | Transposase                                                  |
| Rv3428c | Transposase                                                  |
| Rv3445c | ESAT-6 like protein EsxU                                     |
| Rv3645  | Adenylate cyclase                                            |
| Rv3849  | ESX-1 secretion-associated regulator EspR                    |
| Rv3889c | ESX-2 secretion-associated protein EspG2                     |
| Rv3891c | Esat-6 like protein esxD                                    |

4. Discussion
The virulent strain *M. tuberculosis* H37Rv, is one of the pathogens which is worth to be investigated more, this strain was first isolated in 1905 and is the most widely used in tuberculosis researches, the complete genome sequence and annotation was published in 1998 by Welcome Trust Sanger Institute [21]. The genome of *M. tuberculosis* H37Rv contains about 4,000 protein coding genes, of which more than 1/4-1/3 have been annotated as ‘hypothesized’. The genome has a very high 'guanine + cytosine' content that is reflected in the biased amino acids content of the proteins [18]. The genome was re-annotated more than once using different Bioinformatics tools. Genes and its products of this strain designated *Rv*, and upon re-annotation some were changed with extension “c” if they are found on the complementary strand (antisense strand) [4] and these composed about 334 out of 617 (54.13%) of hypothetical proteins. More than one database are specialized for *M. tuberculosis*, such as TubercuList (http://tuberculist.epfl.ch/), which is routinely updated. Different *M. tuberculosis* strains annotated appear to have slight differences in the gene numbers and consequently different numbers of hypothetical proteins [4, 22].

Many attempts were done to characterize and annotate the hypothetical proteins using different Bioinformatics web tools [22], Doerks et al. (2012) carried out a prediction of a large number of hypothetical orfome (497 genes) using different approaches, these gene products (proteins) were associated with different aspects of bacterial activities, by this they raised the functional annotation to about 88% of this medically important bacteria. Although several studies improved functional annotation of open reading frames (ORFs), considerable fractions are still labeled as ‘hypothetical’, ‘conserved hypothetical’ ‘unknown’ or other similar terms imply that there is no functional indication [18], and for this current study was carried out to cover some of these fractions. It is well known that improving the functional annotation is of a great importance for different purposes, such as understanding the pathogenicity process and helps in drug and vaccine design [18].

Therefore, computational approaches for prediction of *M. tuberculosis* proteins can be used to complement the existing wet lab techniques, such predictions provide a method to annotate *M. tuberculosis* Rv proteomes with Subcellular localization and functional information rapidly. After BCG introduction in 1921 till now, we do not have any promising vaccine against tuberculosis, so the membrane proteins predicted (see Figure 2) can be more investigated and exploited as candidate for vaccine development, these proteins was subjected to initial check (hydrophobicity / data not shown) and seems to be promising.

The other approach used in this study was the annotation via orthology use eggNOG this dependent indirectly on GO. The resulted groups seems to be in general agreement with the major functional classification (see Figure 1 and Figure 3). In this approach, proteins in different species can be combined into orthologous groups, which are known to be appropriate for functional analyses and annotation of newly sequenced genomes as the orthologous genes tend to have the same functions [23].

The bulk of hypothetical proteins belong to orthologous group as defined by eggNOG, and thus is amenable to comparative analysis [18,24]

The other approach used in protein functional prediction was Gene Ontology. GO is one of the greatest contributor to the area of functional annotation and play a critical role in modern biology and cover many organisms from different kingdoms. In this study the newest GO database (Release 2013) was used, but most of the proteins still annotated as uncharacterized proteins, OR they are not included in the database yet.

UniProt database was used to get annotation of more hypothetical proteins (as shown in Table 2), some proteins were annotated, which are in agreement with AmiGO prediction (Rv0008c, Rv0298, Rv0300, Rv0550c, Rv0582, Rv0948c, Rv1103c, Rv1113, Rv1885c, Rv2142c, Rv2253, Rv2274c, Rv2549c, Rv2550c, Rv3428c, Rv3849), however, GO annotations were more informative.

For pathogenic organisms, hypothetical proteins hamper the research for new and effective drug targets, and weaken understanding the pathogenicity processes and full understanding the virulence [20]. Now, the global burden of TB has taken a new dimension due to the emergence of drug resistant varieties of *M. tuberculosis* besides synergy of HIV [25], which capitalizes the necessity of more researches.

5. Conclusion

Although plain annotation of hypothetical proteins is beneficial, but, this field still needs more and more efforts, such investigation of conserved, unknown, putative and other proteins. Advanced research needs to characterize the predicted proteins more deeply using sophisticated software and specialized databases such as pdb and performing docking studies [26].

It is strongly indicated that communication mechanisms are existed in cells [27]. These can be exploited to design drugs to disturb the signal transduction pathways and shove the drug resistance away.

Conflict of interest statement

We declare that we have no conflict of interest.

Acknowledgments

The author thanks Institute of Genetic Engineering and Biotechnology for Postgraduate Studies, University of Baghdad, Iraq.

Reference

1. Ministry of Health of Iraq (MOH). National TB program TB (2013). Epidemiological Situation in Iraq, Fifth National Conference of Tuberculosis / Al-Basra.
2. Al-Kareemi K. Effect of Diacetyl on *Mycobacterium tuberculosis* In Vitro. MSc. thesis. Dept. of Microbiology, College of Medicine, University of Baghdad, Iraq 2008.

3. WHO. Global Tuberculosis Control. Geneva, WHO Report, 2012.

4. J. Pryor M., Médigue C., Cole S., Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv, *Microbiology* 2002; 148: 2967-73.

5. Enault F., Suhre K., Claverie J. Phydac “Gene Function Predictor”: A gene annotation tool based on genomic context analysis, *BMC Bioinformatics*, 2005; 6: 247.

6. Mazandu, G.K.; Mulder, N.J. Scoring protein relationships in functional interaction networks predicted from sequence data, *PLoS One* 2011; 19: e18607.

7. Al-Rubayai D. Molecular Detection of Specific Mutation Associated with Isoniazid or Rifampicin resistance among Category II Pulmonary Tuberculosis Using DNA Microarray. PhD thesis, College of Medicine, Al-Nahrain University, Iraq 2011.

8. Al-Himyari A. Role of Human Genetic Polymorphisms in Susceptibility to Tuberculosis in Babil Province-Iraq. PhD thesis, College of Medicine, University of Babylon, Iraq 2012.

9. Ali R. Molecular Study and Genotyping of *Mycobacterium tuberculosis* Complex Isolated in Respiratory Center in Baghdad. PhD thesis, Institute of Genetic Engineering and Biotechnology for Postgraduate Studies, University of Baghdad, Iraq 2013.

10. Saha S., Raghava G., VICMpred : SVM-based method for the prediction of functional proteins of gram negative bacteria using amino acid patterns and composition, *Genomics Proteomics & Bioinformatics* 2006; 4:42-7.

11. Yu C., Chen Y., Lu C., Hwang J. Prediction of protein subcellular localization. *Proteins* 2006; 64: 643–51.

12. Muller J., Szklarczyk D., Julien P., Letunic I., Roth A., Kuhn M., et al eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations, *Nucleic Acids Res.* 2010; 8:190-5.

13. Tatusov R., Koonin E., Lipman D., A genomic perspective on protein families Science, 1997; 278: 631-7.

14. Tatusov R., Galperin M., Natale D., Koonin E. The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 2000; 28:33-6.

15. Clarke E., Loguercio S., Good B., Su A., A task-based approach for Gene Ontology evaluation, *J Biomed Semantics* 2013; Suppl 1:S4.

16. Carbon S., Ireland A., Mungall C., Shu S., Marshall B., Lewis S., AmiGO: online access to ontology and annotation data, *Bioinformatics* 2009; 25:288-9.

17. Magrane M., UniProt Consortium., UniProt Knowledgebase: a hub of integrated protein data. Database 2011; Article ID bar 00.

18. Doersk T., Noort V., Minge P., Peer P. Annotation of the *M. tuberculosis* Hypothetical Orfeome: Adding Functional Information to More than Half of the Uncharacterized Proteins, *PLoS ONE* 2012; 7: e34302.

19. Pillai L., Chauhan U. Prediction of subcellular localization and function of hypothetical proteins of *mycobacterium tuberculosis* H37RA strain *Int. J. Pharma Bio Sci.* 2012; 3: 856–62.

20. Mazandu G., Mulder N., Function prediction and analysis of *Mycobacterium tuberculosis* hypothetical proteins, *Int J Mol Sci.* 2012; 13:7283-302.

21. Cole S., Brosch R., Parkhill J., Garnier T., Churcsher C., Harris D. et al, Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence, *Nature* 1998; 11:393:537-44.

22. Anandakumar S., Shanmughavel P., Computational Annotation for Hypothetical Proteins of *Mycobacterium Tuberculosis*. *J Comput Sci Syst Biol* 2008; 1:50-62.

23. Koonin E. Orthologs, paralogs and evolutionary genomics, *Annu Rev Genet* 2005; 39: 309–38.

24. Muller J., Szklarczyk D., Julien P., Letunic I., Roth A., eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res* 2010; 38: D190–D95.

25. Nunn P., Williams B., Floyd K., Dye C., Elzinga G., Raviglione M., *Tuberculosis control in the era of HIV*, *Nat Rev Immunol* 2005; 5:819-26.

26. Rambabu R., Peri S., Allam A., Computational analysis and function prediction of a hypothetical protein 1RW0. *Int. J. Comp. Bioinformatics and In Silico Modeling*; 2012; 1: 58-62.

27. Karthik K., Chandra N. *Mycobacterium tuberculosis* interactome analysis unravels potential pathways to drug resistance. *BMC Microbiology* 2008; 8: 234-45.

Cite this article as: Zahra M. Al-Khafaji, In *Silico* Investigation of Rv Hypothetical Proteins of Virulent Strain *Mycobacterium tuberculosis* H37Rv. *Indian J. Pharm. Biol. Res.* 2013; 1(4):81-88.