Annotation of Chorismate Mutase from the *Mycobacterium tuberculosis* and the *Mycobacterium leprae* genome

4.1 Introduction ........................................................................................................ 51

4.1.1 Structure of Chorismate Mutase ............................................................... 51

4.2 Structure-Function analysis of *E.coli* CM ..................................................... 54

4.3 Sequence Analysis of CM ............................................................................. 55

4.3.1 Classification and Chain Length of Chorismate Mutases ....................... 57

4.4 Stereochemical Analysis of helix termination in *E.coli* CM ....................... 58

4.5 Complete sequence details of *Rv1885c* from *M. tuberculosis* ................. 59

4.6 Secondary Structure Prediction ..................................................................... 60

4.7 Status of Annotation in the *Mycobacterium tuberculosis* and in the
*Mycobacterium leprae* genome with respect to the aromatic amino acid
biosynthetic pathway .............................................................................................. 61

4.8 Possible CM from *M. leprae* ..................................................................... 62

4.9 Strategy adopted to identify possible CM ....................................................... 63

4.10 References ..................................................................................................... 64
4.1 Introduction

Chorismate Mutase (E.C. 5.4.99.5) is a central enzyme in the shikimate pathway, which is responsible for the production of Tyrosine and Phenylalanine. This is achieved by the conversion of Chorismate to Prephenate, which is then acted upon subsequently by Prephenate Dehydrogenase and Prephenate Dehydratase to form Hydroxy-Phenylpyruvate and Phenylpyruvate. The Aromatic Aminotransferase then converts this into Tyrosine and Phenylalanine. The enzyme provides a $2 \times 10^6$-fold rate acceleration over the unanalyzed rearrangement (Lee, 1995). This reaction is formally a Claisen rearrangement. Chorismate mutase is the only characterized enzyme that catalyzes a pericyclic process and thus has generated considerable interest in the bioorganic circles.

Fig. 1: Chorismate to Prephenate

Since this is located at the branch point of the shikimate pathway, this enzyme in many organisms is an important point of regulation for maintaining the correct balance of aromatic amino acids in the cell. The shikimate pathway is present only in bacteria, fungi and higher plants and hence Chorismate Mutase (CM) exists only in them. This makes Chorismate Mutase an attractive target for developing Herbicides and anti-bacterial products. Furthermore, the low sequence homology amongst known Chorismate Mutase provides a very good potential for developing unique inhibitors targeted to specific microorganisms.

4.1.1 Structure of Chorismate Mutase

Structures have been determined for CM from *Escherichia coli* (Lee, 1995), *Bacillus subtilis* (Chook, 1994) and Yeast (Strater, 1997). One of the interesting observations is that they have a different fold altogether. One class belongs to the bifunctional *E.coli* CM-Prephenate Dehydratase (AroQp), which is an intertwined homodimer of 3-helical subunits. This has been classified under the “Chorismate Mutase II fold” in the SCOP database (Murzin, 1995). The Yeast CM (which is also an all helix protein) although shares a weak sequence similarity with the *E.coli* protein, has got a very good resemblance in the secondary structure. The two structures are super-imposable with an RMSD of 1.51 Å (involving 432 non-bonding atoms Cα atoms). This has also been classified under the “Chorismate Mutase II fold” in the SCOP database. In fact, it has been suggested that the protein with two weak structural repeats resembling subunits of *E. coli* protein could have arisen due to gene duplication.
Fig. 2 The crystal structures of CM from Yeast (2csm) and *E.coli* (1ecm) was obtained from the Protein Data Bank (Bernstein, 1977) and the structures were superimposable with a RMSD of 1.53. The superposition of the structures was done using Swiss-PDB viewer (Guex, 1997) and the final figure was prepared using WEBLAB.
The other class of CM is from *Bacillus subtilis*, named AroH (Chook, 1994 and Kast, 1996). This is a protein with a completely different structure when compared to the AroQ family of CM. This is a trimeric alpha/beta protein, which has been classified under the “Bacillus Chorismate Mutase like fold”.

![Monomeric CM from Bacillus subtilis](1com)

This figure was prepared using WEBLAB.

The basis of classification for the AroQ and the AroH family of CM is that the AroQ family of CM shows a clear sequence similarity with the *Erwinia herbicola* CM which, is encoded by the gene aroQ. AroH class of CMs are ones that do not show sequence homology with the aroQ gene.

Since, CMs occur as both bifunctional and monofunctional enzymes, Gu, 1997 suggested a naming scheme as AroQf, AroQp, AroQt and AroQd for the monofunctional and the bifunctional CM fused with Prephenate Dehydratase, Prephenate Dehydrogenase and the 3-deoxy-D-arabinoheptulosonate-7-phosphate synthase.
4.2 Structure-Function analysis of *E.coli* CM

The *E.coli* crystal structure (1ecm) has been solved with the inhibitor endo-oxabicyclic-dicarboxylic acid. This inhibitor mimics the Transition State of the CM reaction. The *E.coli* structure is a homodimer, which forms two helix bundles with the inhibitors bound to the base of the helix bundles. The contacting residues on the protein (with a distance cut-off of 4 Å) were calculated using a program developed in-house. The program, which is a PERL script, is available for download at www.geocities.com/madanm2.

![Structure of the complex of E.coli CM (1ecm) showing the Inhibitor with the charged contacting residues (within 4 Å)](image)

**Fig. 4:** Structure of the complex of *E.coli* CM (1ecm) showing the Inhibitor with the charged contacting residues (within 4 Å)

From the output it was noted that an Arginine (R11) from the B subunit makes a contact with the inhibitor bound to the A subunit, and vice-versa. The other charged contacting residues from the A subunit were R28, K39, R47, D48, R51, E52, S84 and E88. The residues, which were involved in making contacts with the inhibitor, were noted and these residues were used as constraints when the protein from *Mycobacterium tuberculosis* was pulled out during the BLAST search.

The figure shows the charged residues, which makes contact with the inhibitor. The inhibitor is shown in the centre and is represented in the CPK space-filling model. The contacting residues are represented in the ball and stick model.
4.3 Sequence Analysis of CM

The CM sequence of *E. herbicola* was taken as the query and a PSI-BLAST (Altshul, 1997) was done with the word length option at 2 and E-Value cut-off at 10. The search converged in the second iteration and it picked up a hypothetical protein from *Mycobacterium tuberculosis*. Multiple sequence alignment of the first 8 best matches was carried using ClustalW (Thompson, 1997). The phylogenetic tree was obtained (fig. 5) and was prepared using Treeview (Page, 1996).

![Phylogenetic Tree](image)

**Fig. 5**: The phylogenetic tree for CMs belonging to the AroQ family from *P. aeruginosa, M. jannaschii, M. tuberculosis, etc.*

When the multiple sequence analysis was carried out, absolute conservation was seen only in 4 positions, manual refinement was done to get the best alignment, where most of the contact residues involved were aligned (fig. 6)
4.3 Sequence analysis of CM

Fig. 6: Multiple Sequence Alignment of the 8 CM (the entry from *Mycobacterium tuberculosis* being a hypothetical protein) belonging to the AroQ family. The helix terminating H residue from *E.coli* is marked with a block arrow.

The entry from *Mycobacterium tuberculosis* is of particular interest because the list of proteins annotated in the complete genome does not contain Chorismate Mutase. This may be due to the inherent problem with CM that it does not show a very good sequence homology with other CMs.

A pairwise sequence alignment of the *E.coli* CM and the hypothetical protein with the entry identification Rv1885c was carried out to verify whether all the contacting residues were in place. It was found that all the contacting residues were in place excepting the S84 and Q88, which have been replaced, by T and E (fig. 7). Primary sequence analysis showed that Ala content was the highest with 14.5% followed by Leu with 12% and then Ser with 10%. The protein has a net negative charge of −4.

Fig. 7: Pairwise alignment of the *E.coli* CM with *Mycobacterium tuberculosis* hypothetical protein Rv1885c. The contact residues involved are shown with an arrow. The helix termination residues are shown in block arrow.
4.3.1 Classification and Chain Length of Chorismate Mutase

The known CMs can be classified as monofunctional or bifunctional CM as shown. The length of Chorismate Mutase domain is in the region of 83-90 residues (Table. 1). The actual polypeptide length of monofunctional CM varies from each organism. The domain description is described in the PFAM database (Bateman, 2000), for example the length of the monofunctional CM from *Methanococcus jannaschii* is 99 residues (MacBeath, 1998), whereas the length of the monofunctional CM from *Erwinia herbicola* is 180 residues (precursor) and from Yeast is 246 residues. The polypeptide chain length for the bifunctional CM-Prephenate Dehydrogenase (AroQt) varies from 373 in *E.coli* and 377 in *H. influenzae* to 620 in *Archaeoglobus fulgidus* (which is a multifunctional protein with CM, PDH, PDT and ACT domains). The chain length for the bifunctional CM-Prephenate Dehydratase (AroQp) varies from 387 in *E.herbicola* (Xia, 1993) to 365 in *Pseudomonas stutzeri* and 362 in *A. aeolicus* (which contains CM, PDT and the ACT domains).

**Fig. 8**: Classification of Chorismate Mutase into two broad categories namely the monofunctional and the bifunctional Chorismate Mutase.

<table>
<thead>
<tr>
<th>CM / Source</th>
<th>Length of CM domain</th>
<th>CM / Source</th>
<th>Length of CM domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>AroQp H. inf</td>
<td>87</td>
<td>AroQf M. jan</td>
<td>85</td>
</tr>
<tr>
<td>AroQp E. her</td>
<td>85</td>
<td>AroQf P. aer</td>
<td>84</td>
</tr>
<tr>
<td>AroQp E. col</td>
<td>85</td>
<td>AroQf E. her</td>
<td>85</td>
</tr>
<tr>
<td>AroQp A. aeo</td>
<td>88</td>
<td>AroQf Yeast</td>
<td>113</td>
</tr>
<tr>
<td>AroQp P. stu</td>
<td>89</td>
<td>AroQf Schys</td>
<td>114</td>
</tr>
<tr>
<td>AroQt H. inf</td>
<td>85</td>
<td>AroQf Arath</td>
<td>126</td>
</tr>
<tr>
<td>AroQt E. col</td>
<td>85</td>
<td>AroH B. sub</td>
<td>84</td>
</tr>
</tbody>
</table>

**Table. 1**: This table shows the organism source, length of the CM domain of different types.
4.4 Stereochemical Analysis of helix termination in *E.coli* CM

All the three helices in the A subunit of the *E.coli* CM terminates in the \( \alpha_1 \) conformation (i.e. positive Phi and Psi). The residue in the helix termination position (T residue) for all the three helices was histidine. Since the subunit ends in a helix, the dihedral angles for the last terminating histidine were not identified. When these helix terminating histidines were mapped on to the multiple sequence alignment and on to the pairwise alignment of *E.coli* CM and the hypothetical protein from *M. tuberculosis*, the histidine which terminates the first helix has been either conserved (*E. herbicola*) or replaced by either G (*M. jannaschii* and *P. aeruginosa*), N (*Buchnera*), Q (*M. tuberculosis*, *H. influenzae* and *P. multocida*) or A (*V. cholerae*). The second helix terminating histidine is again either conserved (*M. tuberculosis*) or replaced by N (*E. herbicola*, *M. jannaschii* and *Buchnera*), Q (*P. multocida*, *V. cholerae* and *H. influenzae*) or P (*P. aeruginosa*). The corresponding residues have been shown in triangular arrows in figure 6 and 7.

These observations are in accordance to previously shown results that the helix termination residues could either be G, N, Q or P (Gunasekaran, 1998). Thus this information could be taken as one of the key conformational constraint that can be imposed on the sequence while identifying Chorismate Mutases from other organisms.

**Fig. 9**: Ramachandran plot of the residues in the *E.coli* CM.

The four residues seen in the left-handed helix region of the Ramachandran plot (Ramachandran, 1968) in figure 9 corresponds to the His43 and the His67 terminating residues of the A and the B subunits of the *E.coli* CM. The residues seen in the extended region of the plot corresponds to residues in the turn region and the loop region. All the other residues fall in to the right-handed helix region.
4.5 Complete sequence details of \textit{Rv1885c} from \textit{M. tuberculosis}

<table>
<thead>
<tr>
<th>Gene name:</th>
<th>Class:</th>
<th>Start/End:</th>
<th>Length:</th>
<th>Molecular weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{Rv1885c}</td>
<td>VI</td>
<td>2134276/213472</td>
<td>597 bp / 199 aa</td>
<td>21345 Da</td>
</tr>
</tbody>
</table>

Product: hypothetical protein
Description: \textit{Rv1885c}, \textit{MTU}78033, len 199. Some similarity to Yp12517/\textit{CENM2}, ERWHE NONFUNCTIONAL CHROMATASE MUTASE (151 aa); FASTA score: opt 181 e-score 215.781; 98.6% identity in 133 aa overlap; TBLASTe score 2.926
Translation: MULTIPHEL ATNNGOEKA SLIASKYPEL ANADTOLNA KUPOOAAP ZVAAPYKNE WRACPZIERN ORHEOQALAP ORPAPROQID PUPUNYPOQC QIRANEKET SFQDQVEKP LSAPPEPPEL OARKSIDOL UNRLEQUENT MOLLLOAPAC AGALADKAPD ILOGMLDDOL TQALYTYTQC TKOALPEF

**Fig. 10:** Details of the hypothetical protein \textit{Rv1185c} showing the amino acid chain length, molecular weight, and the complete sequence.
4.6 Secondary Structure Prediction

The primary sequence was subjected to secondary structure analysis by the Gibrat Method (Gibrat, 1987) and by the DPM method (Deleage, 1987). Both the methods predict the protein to be completely helical. This, if true will be in accordance with the expected structure, because the known crystal structures of the AroQ family of CM have only helices and no strands. The figure 11a and 11b summarises the results of the prediction by the two methods.

Fig. 11a: Secondary structure prediction by Gibrat’s method

Fig. 11b: Secondary structure prediction by Double Precision (DPM) method
4.7 Status of Annotation in the *Mycobacterium tuberculosis* and in the *Mycobacterium leprae* genome with respect to the aromatic amino acid biosynthetic pathway

Recently, the complete genome sequence of *M. leprae* (Cole, 2001) and *M. tuberculosis* (Cole, 1998) have been released. It has been noted that there has been a massive gene decay in the *M. leprae* genome. According to the present level of annotation, it is interesting to note that in the aromatic aminoacid biosynthetic pathway from both these organisms (Table. 2) contain all enzymes, but contains no entry for Chorismate Mutase. This may be due to the low level of sequence homology among known CMs (discussed before).

<table>
<thead>
<tr>
<th>Enzyme</th>
<th><em>M. tuberculosis</em></th>
<th><em>M. leprae</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chorismate Synthase</td>
<td>Rv2540c</td>
<td>ML0516</td>
</tr>
<tr>
<td>Chorismate Mutase</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>Prephenate dehydratase</td>
<td>Rv3838c</td>
<td>ML0078</td>
</tr>
<tr>
<td>Prephenate dehydrogenase</td>
<td>Rv3754</td>
<td>ML2472</td>
</tr>
</tbody>
</table>

Table. 2: Annotation status of enzymes in the biosynthesis of Phenylalanine and Tyrosine

<table>
<thead>
<tr>
<th>Organism</th>
<th>Crystallized length</th>
<th>Total length</th>
<th>Resolution</th>
<th>PDBID</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>B. subtilis</em></td>
<td>127</td>
<td>127</td>
<td>2.2</td>
<td>1COM</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>109</td>
<td>386</td>
<td>2.2</td>
<td>1ECM</td>
</tr>
<tr>
<td>yeast</td>
<td>243</td>
<td>256</td>
<td>2.8</td>
<td>2CSM</td>
</tr>
</tbody>
</table>

Table. 3: Status of Chorismate Mutase in the PDB

Thus in this analysis, we present strong evidences to believe that two proteins, one from *Mycobacterium tuberculosis* Rv1885c, which has been annotated as a hypothetical protein and the other from *Mycobacterium leprae* ML2029, which has been annotated as a pseudo gene could be Chorismate Mutase.
4.8 Possible CM from *M. leprae*

When the *Mycobacterium tuberculosis* Rv1885c protein sequence was used as the query for search against the leprae genome, the BLAST search picked up a pseudogene from the leprae genome. The protein ML2029 was labelled as a pseudogene because the 30th residue has been mutated to a stop codon. When compared with the *M. tuberculosis* genome, the 30th position is occupied by Valine and it is understood from the codon usage that it has to be a double mutation to arrive at a stop codon. Whereas if the 30th position was alanine in *M. leprae*, it will be a 1 base mutation to arrive at a stop codon.

Comparisons of the sequence with *M. tuberculosis*, *M. leprae* and *E. coli* CM are shown in the figure below.

**Fig. 12**: Pairwise alignment of the Rv1885c from *M. tuberculosis* with ML2029 from *M. leprae*

**Fig. 13**: Pairwise alignment of ML2029 from *M. leprae* with the CM from *E. coli*. 

62
4.9 Strategy adopted to identify possible CM

Get some preliminary information from the literature

Check for Protein entry in the PDB

Get information on structures from other sources

Check for protein entry in the Sequence database

Get the sequence of homologs from other sources

Perform a BLAST search against the genome of interest with liberal cut-off values

Do a MSA of all the homologs to get information on the conserved residues

Look at noise level hits individually

Analyze the structure to get some stereochemical constraints

Apply

Get the final set of constraints

Meets all or most of the constraints eg. *M. tuberculosis* Rv1885c

Use *M. tuberculosis* Rv1885c as the query against the *M. leprae* genome

Does not meet most of the constraints eg. *M. leprae* genome

Apply

Get contacting residues in case of structures with complexes

Very good sequence match with a Pseudogene from *M. leprae* (ML2029)
4.10 References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J.,
Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.,


Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F.Jr, Brice,M.D., Rogers,J.R.,
Kennard,O., Shimanouchi,T and Tasumi, M., The Protein Data Bank: a computer-based archival


Cole ST, _et.al_, Deciphering the biology of Mycobacterium tuberculosis from the complete
genome sequence, _Nature_, 1998, **393**(6685),537-544.

Cole ST, _et.al_., Massive gene decay in the leprosy bacillus., _Nature_. 2001, **409**(6823), 1007-
1011.

Deleage, G. and Roux, B., An algorithm for protein secondary structure prediction based on class
prediction, _Prot Eng_, 1987, **1**, 289-294

Gibrat, J. F., Garnier, J. et Robson, B. Further developpments of protein secondary structure
prediction using information theory, _J. Mol. Biol._ 1987, **198**, 425-443

Gu W, Williams DS, Aldrich HC, Xie G, Gabriel DW, Jensen RA., The aroQ and pheA domains
of the bifunctional P-protein from _Xanthomonas campestris_ in a context of genomic comparison.
_Microb Comp Genomics_, 1997, **2**(2),141-58.

Guex, N. and Peitsch, M. C., Swiss-Model and and the Swiss-PdbViewer: An environment for

Gunasekaran K, Nagarajaram HA, Ramakrishnan C, Balaram P, Stereochemical punctuation
marks in protein structures: glycine and proline containing helix stop signals, _J Mol Biol_, 1998,
**275**(5), :917-932.
4.10 References


Page RD., TreeView: an application to display phylogenetic trees on personal computers, Comput Appl Biosci 1996, 12(4) .357-8

Ramachandran GN, Sasisekharan V, Conformation of polypeptides and proteins, Adv Protein Chem 1968, 23,283-438

