Domain prediction procedure used in Target Selection

Contact Person: John-Marc Chandonia; e-mail: JMChandonia@lbl.gov

Some *Mycoplasma* ORFs that were filtered out in early selection rounds were multidomain proteins that included tractable domains of unknown structure, but had been eliminated because of homology to a single domain of known structure. Therefore, in the 6th round of target selection, *Mycoplasma* ORFs were divided into domains before entering the target selection filters. The procedure used was the same as that used to identify domains in the ASTEROIDS data set of the ASTRAL database (Chandonia et al., 2004). Hidden Markov models of ASTRAL families and superfamilies were used to predict domains in the *M. pneumoniae* ORFs, using the HMMER tool with a significance cutoff of 10^{-4}. BLAST (Altschul et al., 1990) was also used to compare ASTRAL sequences to all *M. pneumoniae* ORFs, using a significance cutoff E-value of 10^{-4}. Regions of *Mycoplasma* sequence matching one or more ASTRAL sequences or hidden Markov models were annotated as belonging to the same SCOP (Murzin et al., 1995) superfamily as the hit with the most significant E-value produced by either method. Remaining unclassified regions were annotated using Pfam 10.0 (Bateman et al., 2004), using the Pfam_ls model library and the "trusted cutoff" score for each model to determine significance. Significant hits were annotated as Pfam domains. After Pfam annotation, remaining regions of at least 20 consecutive residues were annotated as potential unclassified domains. This procedure is identical to the one documented in the release notes for ASTRAL 1.65.

Putative domains identified by the ASTRAL procedure were further split into two parts at the end of each predicted transmembrane helix, as predicted by TMHMM 2.0a (Krogh et al., 2001). Finally, putative domains shorter than 50 residues were eliminated from further consideration as targets.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403-410.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. 2004. The Pfam protein families database. *Nucleic Acids Res 32 Database issue*:D138-141.

Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. 2004. The ASTRAL Compendium in 2004. *Nucleic Acids Res 32 Database issue*:D189-192.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol 305*:567-580.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol 247*:536-540.