Bioinformatics in the identification of glycosidase genes in the Chlamydia genome

Pamela Greenwell
Sanjiv Rughooputh

School of Biosciences

This is an electronic version of an article published in Molecular and Cellular Proteomics, 2 (9). 826, September 2003. Molecular and Cellular Proteomics is available online at:

http://www.mcponline.org/cgi/reprint/2/9/825

71.6

Bioinformatics in the Identification of Glycosidase Genes in the Chlamydia Genome

P. Greenwell and S. Rughooputh

University of Westminster, London, U.K.

Following the cloning of the 1,042,519-bp Chlamydia trachomatis genome, it was thought that identification of genes, utilizing homology searches, would be simple. However, more than 25% of the DNA sequenced has yet to be assigned function, with over 200 gene products classified as hypothetical proteins. The organism expresses the glycosidases a-mannosidase and b-N-acetylglucosaminidase and therefore must possess the relevant genes. We undertook homology searches using BLASTN and BLASTP to identify regions conserved across a wide range of prokaryotes and interrogated the C.trachomatis genome to locate the relevant genes. Surprisingly little homology was found. Interrogation with a-mannosidase sequences yielded matches with only 2 or 3 proteins, all but one of which had been allocated functions unrelated to glycosidases. Areas of homology lay outside conserved regions for prokaryote a-mannosidase. Interrogation of the genome with the b-hexosaminidase from Neisseria meningitidis revealed a 29% identity (34.25% ungapped) over 165 amino acids with the hypothetical protein CT391. The homology lies within a conserved area of the prokaryotic b-N-acetylglucosaminidase. Nevertheless, this is only scant evidence of identity. Although the use of bioinformatics as a tool for the rapid identification of the genes has not been a complete success, it will prove vital when the enzymes have been purified and N-terminal sequence has been obtained. Interrogation of the genome at this stage will allow us to allocate function to, as yet, uncharacterized genes, and determine why the primary bioinformatics failed to identify the appropriate sequences.