Protein misprediction uncovered by new technique

August 27 2008

A new bioinformatics tool is capable of identifying and correcting abnormal, incomplete and mispredicted protein annotations in public databases. The MisPred tool, described today in the open access journal *BMC Bioinformatics,* currently uses five principles to identify suspect proteins that are likely to be abnormal or mispredicted.

László Patthy led a team from the Institute of Enzymology of the Hungarian Academy of Sciences, Budapest, that developed this new approach. He explained how necessary it is, "Recent studies have shown that a significant proportion of eukaryotic genes are mispredicted at the transcript level. As the MisPred routines are able to detect many of these errors, and may aid in their correction, we suggest that it may significantly improve the quality of protein sequence data based on gene predictions". The MisPred approach promises to save much time and effort that would otherwise be spent in further investigation of erroneously identified genes.

The MisPred approach rates annotations according to five dogmas:

-- Extracellular or transmembrane proteins must have appropriate secretory signals.
-- A protein with intra- and extra-cellular parts must have a transmembrane segment.
-- Extracellular and nuclear domains must not occur in a single protein.
-- The number of amino acid residues in closely related members of a globular domain family must fall into a relatively narrow range.

-- A protein must be encoded by exons located on a single chromosome.

There are some exceptions to these rules, as pointed out by Patthy, "Some secreted proteins may truly lack secretory signal peptides since they are subject to leaderless protein secretion. Similarly, it cannot be excluded at present that transchromosomal chimeras can be formed and may have normal physiological functions. Nevertheless, the fact that MisPred analyses of protein sequences of the Swiss-Prot database identified very few such exceptions indicates that the rules of MisPred are generally valid".

The authors found that the absence of expected signal peptides and violation of domain integrity account for the majority of mispredictions. The authors note that "Interestingly, even the manually curated UniProtKB/Swiss-Prot dataset is contaminated with mispredicted or abnormal proteins, although to a much lesser extent than UniProtKB/TrEMBL or the EnsEMBL or GNOMON predicted entries".

Source: BioMed Central

Citation: Protein misprediction uncovered by new technique (2008, August 27) retrieved 27 April 2024 from https://phys.org/news/2008-08-protein-misprediction-uncovered-technique.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.