Wing-Cheong Wong

Towards a complete sequence homology concept: Limitations and applications

Diss. 2011

Historically, the paradigm of similarity of protein sequences implying common structure, function and ancestry was generalized based on studies of globular domains. The implications of sequence similarity among non-globular protein segments have not been studied to the same extent; nevertheless, homology considerations are silently extended for them. This appears especially detrimental in the case of transmembrane helices (TMs) and signal peptides (SPs) where sequence similarity is necessarily a consequence of physical requirements rather than common ancestry. Since the matching of SPs/TMs creates the illusion of matching hydrophobic cores, the inclusion of SPs/TMs into domain models can give rise to wrong annotations. More than 1001 domains among the 10,340 models of Pfam release 23 and 18 domains of SMART version 6 (out of 809) contain SP/TM regions. As expected, fragment mode HMM searches generate promiscuous hits limited to solely the SP/TM part among clearly unrelated proteins. More worryingly, this work shows explicit examples that the scores of clearly false-positive hits, even in globalmode searches, can be elevated into the significance range just by matching the hydrophobic runs. In the PIR iProClass database v3.74 using conservative criteria, this study finds that at least between 2.1% and 13.6% of its annotated Pfam hits appear unjustified for a set of validated domain models. Thus, false positive domain hits enforced by SP/TM regions can lead to dramatic annotation errors where the hit has nothing in common with the problematic domain model except the SP/TM region itself. A workflow of flagging problematic hits arising from SP/TM-containing models for critical reconsideration by annotation users is provided.

While E-value guided extrapolation of protein domain annotation from libraries such as Pfam with the HMMER suite is indispensable for hypothesizing about the function of experimentally uncharacterized protein sequences, it can also complicate the annotation problem. In HMMER2, the E-value is computed from the score via a logistic function or via a domain model-specific extreme value distribution (EVD); the lower of the two is returned as E-value for the domain hit in the query sequence. We demonstrated that, for thousands of domain models, this treatment results in switching from the EVD to the statistical model with the logistic function when scores grow (for Pfam release 23, 99% in the global mode and 75% in the fragment mode). If the score corresponding to the breakpoint results in an E-value above a user-defined threshold (e.g., 0.1), a critical score region with conflicting E-values from the logistic function (below the threshold) and from EVD (above the threshold) does exist. Thus, this switch will affect E-value guided annotation decisions in an automated mode. To emphasize, switching in the fragment mode is of no practical relevance since it occurs only at E-values far below 0.1. Unfortunately, a critical score region does exist for 185 domain models in the hmmpfam and 1748 domain models in the hmmsearch global-search mode. For 145 out the respective 185 models, the critical score region is indeed populated by actual sequences. In total, 24.4% of their hits have a logistic function-derived E-value<0.1 when the EVD provides an E-value>0.1. Examples of false annotations are provided and the appropriateness of a logistic function as alternative to the EVD is critically discussed. This work shows that misguided E-value computation coupled with non-globular regions embedded in domain model library not only causes annotation errors in public databases but also limits the extrapolation power of protein function prediction tasks.

So far, the preceding work has demonstrated that sequence homology considerations widely used to transfer functional annotation to uncharacterized protein sequences require special precautions in the case of non-globular sequence segments including membrane-spanning stretches from non-polar residues. We found that there are two types of transmembrane helices (TMs) in membrane-associated proteins. On the one hand, there are so-called simple TMs with elevated hydrophobicity, low sequence complexity and extraordinary enrichment in long aliphatic residues. They merely serve as membrane-anchoring device. In contrast, so-called complex TMs have lower hydrophobicity, higher sequence complexity and some functional residues. These TMs have additional roles besides membrane anchoring such as intramembrane complex formation, ligand binding or a catalytic role. Simple and complex TMs can occur both in single-

and multi-membrane-spanning proteins essentially in any type of topology. Whereas simple TMs have the potential to confuse searches for sequence homologues and to generate unrelated hits with seemingly convincing statistical significance, complex TMs contain essential evolutionary information. For extending the homologyconcept onto membrane proteins, we provide a necessary quantitative criterion to distinguish simple TMs in query sequences prior to their usage in homology searches based on assessment of hydrophobicity and sequence complexity of the TM sequence segments.

Theoretical insights from this work were applied to problems of function prediction for specific uncharacterized gene/protein sequences (for example, APMAP and ARXES) and for the functional classification of TM-containing proteins.