Computational Target Discovery: Using HMMs to Identify the Druggable Proteome

Joanne I. Adamkewicz[1], R. Glenn Hammonds[2]

Keywords: Target discovery, HMM, Pfam, protein classification, enzymes, drugs

1 Abstract.

We present a method for the classification of proteins for the purpose of target discovery. Using the functional hierarchy of the Enzyme Commission (EC) classification, with extensions where required, we classified over 9000 publicly available Hidden Markov Models (HMMs) representing evolutionarily conserved protein domains, noting catalytic activity and other properties. We use 4 fungal genomes with differing levels of annotation to illustrate how the method can rapidly identify all potentially "druggable" genes in a genome of interest, using only publicly available software and data. The model collection currently covers 88% of the 52475 Swiss-Prot protein entries with EC assignments. Those enzymatic proteins not hit by the collection will direct building of custom HMMs.

2 Introduction.

Modern target discovery has become a process of sifting wheat from chaff. With full genome sequence available for many organisms, the number of possible target genes is in the thousands or tens of thousands for each species of interest. Validating potential targets via biological experiments is labor-intensive and time-consuming, yet if the product of a gene is not amenable to modulation by drug therapy, prior biological validation work is wasted. As a first step in the drug discovery process, then, it is highly cost-effective to restrict your focus to the fraction of an organism's genes that are considered druggable according to criteria defined for each specific project. Ideally, the filtering process would be rapid and automated; applicable to annotated, unannotated, and even partial genomes; flexible (so each user can define their own criteria for selecting desired targets); use only publicly available software and data; and be applicable to any species of interest.

We present here a curation of publicly available models (or alignments that can be made into models), which has proven useful in target selection and practical to implement using largely off the shelf software. The essence of our curation is a classification of the models into categories of interest to our customers: biologists at biotech or pharmaceutical companies. This classification allows biologists to quickly select candidate targets for a wide variety of projects based on user-defined criteria.

[1] Exelixis, Inc., South San Francisco, CA, USA. E-mail: jadamkew@exelixis.com
[2] Exelixis, Inc., South San Francisco, CA, USA. E-mail: rghammonds@earthlink.net

3 Methods and Results.

Models or alignments were obtained by download from Pfam[1], Interpro[2] for SMART [3] models, TIGR [4], NCBI for COG [5] alignments, and Affymetrix for GPCR models [6]. In addition, custom alignments were built as desired to cover additional protein families. Alignments were converted to models where necessary using either hmmer or SAM, with later conversion of all models to hmmer format. The resulting model collection was curated in collaboration with biologists via automated text classification followed by manual inspection.

We estimate the fraction of known enzymes covered by the current model collection from the fraction of SwissProt sequences annotated with an EC number retrieved by any enzymatic model in the collection: 88% (E value <= 1).

Given the curated collection, the models are then run against a proteome of interest using the hmmpfam or hmmsearch algorithms. The raw output files are processed and stored in a database for easier searching and downstream analysis. Each sequence hit by one or more domains then inherits the classification of those domain hits, according to priorities as set by the user. For newly-sequenced genomes where no protein predictions are available, gene prediction tools such as Orfinder can be used in combination with HMMs calibrated to find partial matches to a domain.

To illustrate the utility of the curated model collection, we analyzed 4 fungal genomes in various states of sequencing and gene annotation: Saccharomyces cerevisae, Schizosaccharomyces pombe, Ustilago maydis, and Phanerochaete chrysosporium. We present comparative results for classes of particular interest, including kinases, proteases, and glycosyltransferases, and show that the method can be used successfully with an unannotated genome by adding an ORF-finding step.

References

[1] Bateman, A. *et al*. 2004 The Pfam protein families database. *Nucleic Acids Research* 32: D138-D141

[4] Haft DH, Selengut JD, White O. 2003 The TIGRFAMs database of protein families. *Nucleic Acids Research* 31:371-3.

[3] Letunic, I., Copley, R.R., Schmidt, S., *et al*. 2004 SMART 4.0: towards genomic data integration. *Nucleic Acids Research* 32: D142-D144

[2] Mulder N.J., Apweiler R., Attwood T.K., *et al*. 2003 The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Research* 31:315-318.

[6] Shigeta R., Cline M., Liu G., and Siani-Rose MA. 2003 GPCR-GRAPA-LIB-a refined library of hidden Markov Models for annotating GPCRs *Bioinformatics* 19:667-668.

[5] Tatusov RL, Natale DA, Garkavtsev IV, *et al*. 2001 The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research* 29: 22-8.