ABSTRACT

Summary: GPCR-GRAPA-LIB is a library of HMMs describing G protein coupled receptor families. These families are initially defined by class of receptor ligand, with divergent families divided into subfamilies using phylogenetic analysis and knowledge of GPCR function. Protein sequences are applied to the models with the GRAPA curve-based selection criteria. RefSeq sequences for Homo sapiens, Drosophila melanogaster, and Caenorhabditis elegans have been annotated using this approach.


Contact: ron.shigeta@affymetrix.com

INTRODUCTION

GPCRs are involved in intercellular signaling, regulate functions ranging from neuronal signaling to lymphocyte trafficking to olfactory and visual processing, and are responsible for activity by over 50% of existing medicines. (Gudermann et al., 1995) They consist of N-terminal and C-terminal domains, and extracellular and intracellular loops linking seven transmembrane helices. Their loops vary in length and combine to form the active sites, sometimes incorporating a chromophore in the helix bundle. Consequently, GPCRs do not subdivide into families easily, making them tricky to model.

Karchin et al. (2002) showed that HMMs can accurately distinguish among GPCR families. While support vector machines proved more accurate than HMMs at GPCR subfamily classification, HMMs performed better at family classification. Karchin has provided a GPCR subfamily classifier at http://www.soe.ucsc.edu/research/compbio/gpcr-subclass/. Unfortunately, some GPCR subfamilies are sparsely populated, and cannot be modeled under this classifier. Thus, complementing Karchin’s methods, we developed a library of HMMs for GPCR family classification.

GPCR FAMILY MODELING

Traditional GPCR classification is based on the receptor’s ligand. Swiss-Prot provides such a classification at http://www.expasy.ch/cgi-bin/lists??7tmrlist.txt. However, GPCRs within the same ligand class can have little in common in function. For example, chemokine receptors, coded on different chromosomes, differ significantly in ultimate physiological response, but are grouped together under the common chemokine ligand. Because traditional groupings do not necessarily capture such subtleties, we subdivided the families further according to our knowledge of receptor function.

When families were more divergent, with excessive gaps in the multiple alignments, we partitioned them into subfamilies manually according to biological knowledge and phylogenetic analysis. Evolutionary trees of the divergent families were estimated using CLUSTALW (Thompson et al., 1994) to identify subpopulations within the alignment. The final family breakdown, shown in Figure 1, is consistent with the GPCRDB (Horn et al., 1998) but more general.

For each GPCR family, an HMM was built with SAMtarget99, (Karplus et al., 1998) using options --family and --tuneup, to generate multiple alignments of pre-selected GPCR sequences. An HMM was estimated from each alignment using Target99’s w0.5 weighted build method.

IDENTIFYING PUTATIVE GPCRS IN THE HUMAN GENOME

We applied this GPCR HMM library to the set of known and predicted genes in the April 2001 Golden Path human genome. The known genes were strict alignments of complete-CDS transcripts to genome. Putative genes were predicted according to mRNA and EST alignments to the genome, and an HMM-based abinitio gene predictor. This gene set is detailed elsewhere (Cline et al., 2002).
Fig. 1. *G protein coupled receptor annotations*. Number of GPCRs found in the April 2001 Golden Path human genome. GPCR families were initially classified via the Swiss-Prot seven-transmembrane scheme, with divergent families subdivided manually according to phylogenetic analysis and biological knowledge.

Each protein was scored against the GPCR HMMs. Significant hits were identified with GRAPA’s graph-based scoring function (Shigeta et al., 2001, 2002). This scoring identifies a distinct e-value cutoff for each model in the library according to the histograms of scores, with the cutoff selected dynamically where the histogram rises sharply. This sharp rise corresponds roughly to the sharp increase in false positives. This dynamic selection process yields strong discriminatory power: 97% selectivity and 80% sensitivity for SCOP family recognition (Shigeta et al., 2002) When one sequence scored more than one putative hit, one model was selected according to the e-values. See Figure 1 for refined families and number of human select genes scored against the models.

We believe the GPCR-GRAPA-LIB collection of GPCR HMMs will provide the drug development and target discovery communities with a tool for accurate identification and alignment of GPCRs.

REFERENCES


