Materials and Methods

Supplemental Online Material for Manning et al (Science 298:1912-1934)

Construction and validation of the ePK Hidden Markov Model

An ePK domain hidden Markov model (HMM) was built from a manually-adjusted alignment of 70 diverse kinase domain sequences from yeast, worm, fly, and human that share <50% sequence identity in the catalytic domain. HMMs have generally been found to be highly sensitive for detection of moderately remote homologs (J. Park et al, JMB (1998) 284:1201-1210). To test the selectivity of the model, it was run against the Swiss Prot release 40 database of 113,434 protein sequences. Using a P value cutoff of 0.1, the model detected 1353 putative ePK domains, all of which were either annotated as kinases or putative kinases, or had convincing sequence similarity to known kinases. The HMM could detect fragments as short as 20 AA for most kinases, allowing it to be used on single-read (~500 nt) genomic sequence containing short or partial exons.

Use of HMM to discover ePK domains

Local and global HMM models were built with the HMMer package (http://hmmer.wustl.edu) and were searched against sequence databases using the Decypher hardware-accelerated HMMer implementation from Time Logic (http://www.timelogic.com). The sequence databases used were the public Genbank, SwissProt and dbEST collections, Celera human genomic sequences (raw reads and assembled sequences), the Incyte LifeSeqGold collection (5.4 million sequences) and sequence collections from internal SUGEN and Pharmacia databases (0.35 million sequences).

Discovery of atypical protein kinases (aPKs)

Profile HMMs were also constructed for the PIKK, RIO, ABC1, PDK and Alpha kinase families, and similarly searched against all sequence sources. Homologs of other atypical kinases were identified using Blast and Psi-Blast against the databases mentioned above as well as against protein sequence predictions from Celera and Ensembl.

Extension of kinase fragments identified by HMM

Sequences which matched the HMM were extracted from their parental sequence and aligned to either Celera or public genomic assemblies. 200-500 kb of flanking genomic sequence were loaded into a custom database for full length gene prediction. Several gene prediction methods were performed on the genomic region, and the results loaded to the database and visualised on a custom genome browser. The Genewise program was used to identify kinase genes within the region: First, the genomic region was compared to the kinase domain HMM using genewise with default parameters, to predict the full ePK domain. The predicted domain was then searched against public and proprietary kinase sequences to discover close homologs. Such homolo gs typically share extensive sequence similarity outside of the kinase domain. Three or more homologs were then used as templates by Genewise to predict a full length protein, and these predictions were added to the database. cDNAs and ESTs were mapped to the genomic region using Blast to discover matching sequences and sim4 to align them. Genscan ab initio gene prediction was also carried out on each locus.

Manual analysis of the various gene predictions was used to assemble a full-length coding region. Where alternative splicing was seen, priority was given to sequences encoding the longest open reading frame, and then to sequences encoding the longest predicted cDNA. Generally, Genewise gave very good predictions for most of the length of a gene, with ESTs and cDNAs filling in at poorly-conserved ends and UTRs (untranslated regions). Many genomic regions contained internal gaps, which often included kinase exons. These regions were filled in with EST or cDNA sequence, or by manually re-assembling the missing sequence from other genomic sources (the sequence of most gaps in both Celera and public contigs was present elsewhere in their databases, but had not been correctly assembled). Genscan predictions were incorporated only where Genewise and ESTs failed, and where the Genscan extension was supported by strong sequence similarity using Blast. Where sequnce similarity was very low, we supplemented Genewise by running TBlastN between a homolgous sequence and the genomic region to identify regions of sequence similarity. This proved to be more sensitive than Genewise, and better able to find homolgous regions when the genomic region was poorly assembled.

Sequence errors and polymorphisms

Fine-scale discrepancies between different sources, such as single-nucleotide polymorphisms, were reanalyzed by comparing the predicted sequences against all available sequence sources, to determine if the sequence was a polymorphism (both alleles present multiple times in different sequence sources) or a likely sequencing error (no confirmation of the minor allele). The 24.2 million raw sequence reads provided by Celera and the ~10.3 million available ESTs gave good sequence coverage of most genes and were particularly useful in identifying potential sequence errors. Where differences were deemed to be due to polymorphisms, the most common allele was used in our final sequence.

Classification of Protein Kinase Domains

Domain sequences were extracted by alignment to HMM and compared by multiple sequence alignment (hmmalign and clustalw) and pairwise comparison (blast, Smith-Waterman). Phylogenetic trees were created using neighbor-joining (clustalw), parsimony (Phylip) or similarity clustering (protomap).

Chromosomal mapping

The chromosomal location of each kinase sequence was determined by alignment to genomic assemblies using Blat (http://genome.ucsc.edu/) and the Celera Discovery System (http://cds.celera.com), as well as using references from the literature and OMIM (http://www.ncbi.nlm.nih.gov/Omim/) for previously mapped genes.

Domain analysis

Final kinase protein sequences were run against the Pfam 7.4 set of domain HMM profiles, using both local and glocal models. All matches with P scores of <0.01 were accepted, and all scores with P values of 0.01-1 were manually evaluated, by comparison with homologous sequences, inspection of the domain alignment, and reference to the literature for description of the domain occurance. Calmodulin binding (CaM) motifs are highly degenerate are are not represented within Pfam. Some CaM motifs were identified from the literature and from database annotations; others were identified by sequence similarity to known CaM motifs, and their position just C-terminal of the kinase domain. Signal peptides were identified with SignalP (http://www.cbs.dtu.dk/services/SignalP/) and transmembrane regions using TM-HMM (http://www.cbs.dtu.dk/services/TMHMM/).

Comparison with other protein sequence databases

Comparisons were made with the Ensembl IPI.1 database (31,780 sequences) and Celera release 25h (39,256 sequences; includes both high-confidence and low-confidence predictions) as well as the NCBI nonredundant Genpept database of April 2, 2002, which was filtered to remove automated genomic ORF predictions. The Smith-Waterman algorithm was used with a match substitution matrix (score 1 for a match, -1 for a mismatch). The sequence difference rate was calculated as the sum of the query and target lengths minus twice the match length, divided by the query length; thus both overprediction and underprediction of sequence are noted.