Fold Recognition using Hidden Markov Models and Secondary Structure

Kevin Karplus and Richard Hughey

Computer Engineering
University of California, Santa Cruz

- **Fold recognition:**
  Starting with a target amino acid sequence of unknown structure, predict its tertiary structure by finding a template of know structure that is homologous, and aligning corresponding residues.

- **Profile Hidden Markov model:**
  A profile Hidden Markov Model (HMM) has a sequence of *match states* corresponding to positions in a protein. Each match state has an *emission probability table* that describes the possible amino acids in that position. There are also parameters for transitions to *insert states* and *delete states,* to handle indels.

- **Multiple alignment of homologs:**
  The SAM-T2K script does an iterative search of NCBI's NR database of proteins using hidden Markov models to find probable homologs. This search is similar to PSI-BLAST, but somewhat more expensive and somewhat more accurate.

- **Template models:**
  We create a profile HMM for each of our *templates* (chains of known structure from PDB). We currently have over 5500 templates, greatly over-representing some parts of structure space.

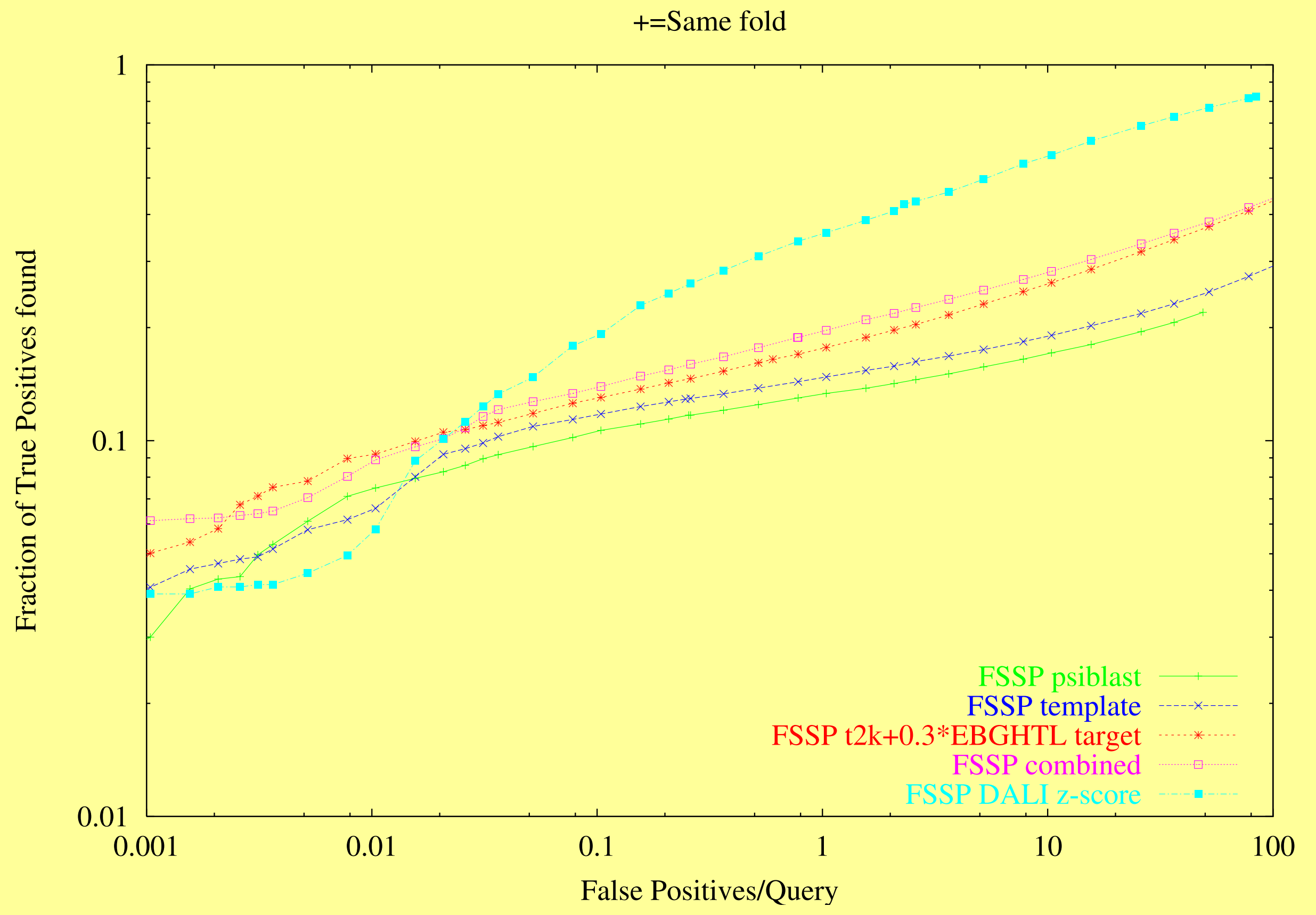
- **Secondary structure prediction:**
  Applying a neural net to the multiple alignment of homologs found by SAM-T2K, we get a probability vector for the 6-letter alphabet EBGHTL in each position. The neural net is trained using STRIDE-labeling of known structures. Our secondary-structure predictor is currently in the top group in the EVA evaluations.

- **Two-track HMM:**
  A two-track HMM has two emission probability tables in each match and insert node: one for amino acids and one for some other alphabet (typically a secondary structure code). A two-track HMM models a pair of sequences: one of amino acid codes and one of secondary structure codes.

- **Target models:**
  We create a two-track HMM for each target sequence whose structure we are trying to predict.

- **Combined approach:**
  We can combine the E-values of target and template models for each target-template pair, reducing the number of false positives.

- **Fold recognition test:**
  Two chains considered correct if they contain domains in same fold in SCOP-1.55. Using 1924 of FSSP representative sequences (x-ray only, excluding not-a-superfamily SCOP classes).



- **Future work on fold recognition:**
  Improved target-template alignments, improved selection among similar templates and alignments, new web server for SAM-T2K (only SAM-T99, which doesn't use secondary structure now up), predictors for other local structure properties, combining hits for multiple templates in same fold class, using keyword matches on sequence annotation to get hints on functional matches, …

- **Future work on tertiary structure prediction:**
  Fragment-packing algorithm implemented in undertaker (which optimizes burial) needs to be upgraded with better score functions. Conformation generator needs some efficiency improvements. Post-processing of predictions (clustering to find common core, for example) needs to be added. Need to combine alignment scores and predicted local structure labels with ab-initio scoring scheme in undertaker.

SAM programs (free for academics, government labs, and non-profits):
http://www.soe.ucsc.edu/research/compbio/sam.html

SAM-T99 web server (free for anyone):
http://www.soe.ucsc.edu/research/compbio/HMM-apps/T99-query.html

UCSC Bioinformatics research and degree programs
http://www.soe.ucsc.edu/research/compbio

Thanks! DOE DE-FG03-99ER62849 and NSF DBI-9808007, EIA-9905322