1 Supplementary material

Supplementary material for the paper PconsC: Combination of direct information methods and alignments improves contact prediction, by Skwark et al.

All the precision (PPV) estimates have been computed excluding short-range contacts (less that 5 residue separation). That is, the shortest range considered is $(i, i + 5)$.

1.1 Supplementary tables

Prediction method	plmDCA	PSICOV	plmDCA+PSICOV
PfamA	0.39	0.31	—
jackhmmer $e = 10^{-40}$	0.31	0.27	0.34
jackhmmer $e = 10^{-10}$	0.47	0.38	0.49
jackhmmer $e = 10^{-4}$	0.48	0.39	0.50
jackhmmer $e = 1$	0.48	0.39	0.50
HHblits $e = 10^{-40}$	0.27	0.27	0.30
HHblits $e = 10^{-10}$	0.46	0.39	0.50
HHblits $e = 10^{-4}$	0.49	0.40	0.52
HHblits $e = 1$	0.47	0.38	0.50
All jackhmmer	0.50	0.42	0.52
All HHblits	0.51	0.44	0.54
ALL	0.52	0.47	0.55

Supplementary Table S1: Prediction precision for combinations of prediction and alignment methods. True positive: Cβ-Cβ distance ≤ 8 Å, False positive: Cβ-Cβ distance > 8 Å

1.2 Alignment method comparison

Both jackhmmer and HHblits appear to produce alignments suitable for contact prediction. The discrepancies in the performance at different e-value thresholds may be partially due to differing sequence coverage and amount of aligned sequences. At e-value of 10^{-4}, both alignment methods produce similar amount of sequences at approximately same coverage. At less stringent cutoffs, HHblits renders more sequences than jackhmmer, but with much lower sequence coverage, whereas at more stringent cutoffs, HHblits renders fewer sequences, at approximately same coverage.

It is worthwhile to observe, that due to design of HHblits, the amount of sequences in the alignment is effectively capped at 65536 (2^{16}) and at very stringent cutoffs, there may be no hits to target sequence below the threshold, resulting in a single sequence in the alignment.

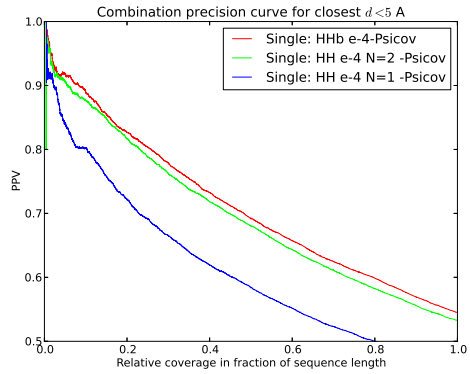
The following table illustrates this (HH: HHblits, JH: jackhmmer)

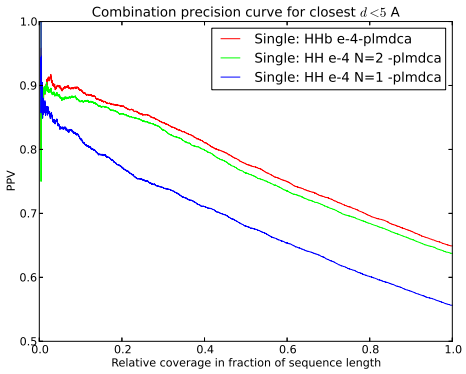
Alignment metod	Mean # seq.	Median # seq.	Mean coverage
JH $e = 10^{-40}$	3998	1169	0.910
JH $e = 10^{-10}$	11791	4265	0.832
JH $e = 10^{-4}$	20824	7343	0.799
JH $e = 1$	21839	7361	0.794
HH $e = 10^{-40}$	1620	648	0.935
HH $e = 10^{-10}$	9046	4147	0.839
HH $e = 10^{-4}$	15443	7198	0.776
HH $e = 1$	19861	10342	0.695

Supplementary Table S2: Statistics for input alignment. Mean and median number of sequences in the alignments as well as mean sequence coverage

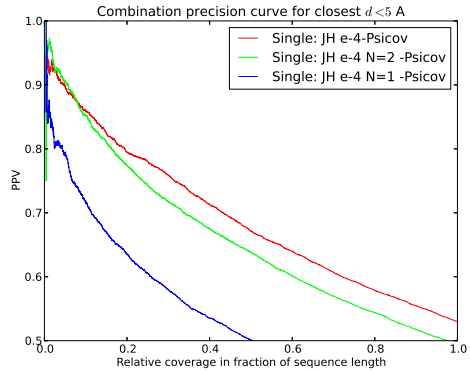
1.3 HHblits dependence on parameters

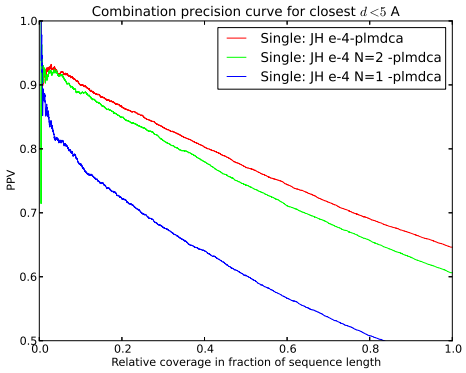
HHblits is characterized by multiplicity of input parameters. One of the main features of HHblits is its time performance, it is capable of producing accurate alignments, even with relatively few method iterations. Internal benchmarks show, that 2 iterations are nearly as accurate as default of 5. One iteration, though, results in significantly decreased prediction performances, both with PSICOV and plmDCA.



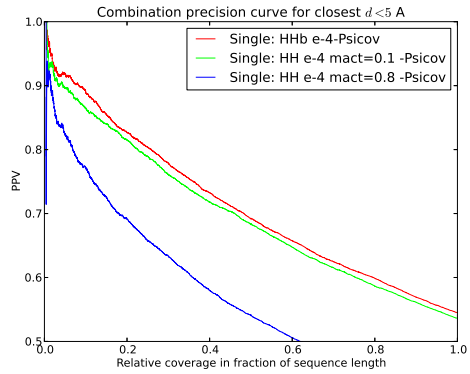


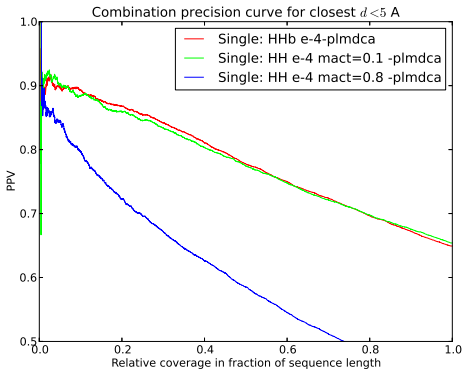
The same does not hold for jackhmmer though, which means that HHblits is capable of identifying more useful homologous sequences than jackhmmer in the span of same number of iterations:





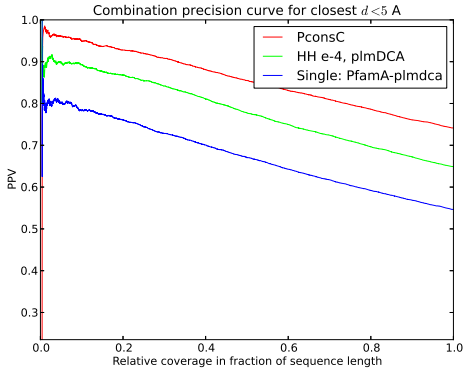
The other parameter — MACT — controls the *greediness* of alignment. It appears that the default value of 0.5 provides most suitable alignments for the purposes of contact prediction. With PSICOV, the value of 0.1 (very global) results in a decreased performance overall, whereas a value of 0.8 (rather local) causes even greater performance drop. While using plmDCA, local alignments – as expected – result in a drop in performance, but the global ones cause a slight prediction accuracy increase, which we believe is due to plmDCA's inference model being able to account better for the inherent noisyness of the data and thus benefit from the increased amount of information in the alignment.

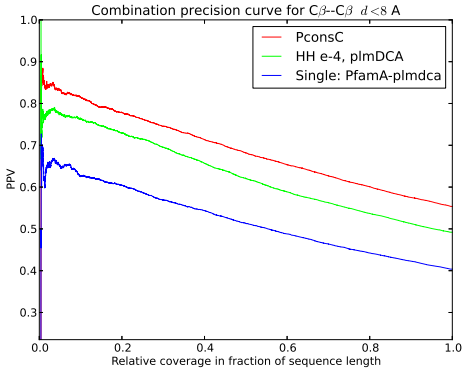




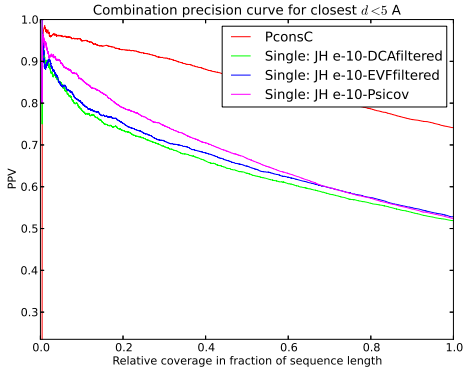
1.4 Other alignment sources and direct information methods

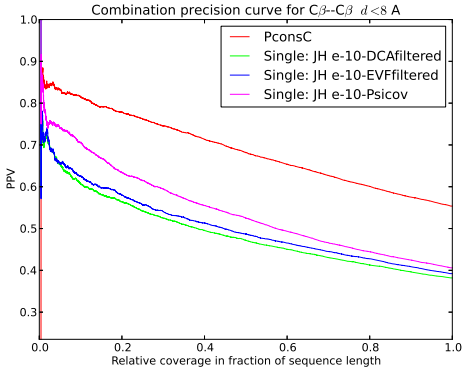
We have explored the possibility of using PfamA alignments as inputs, but that resulted in lower overall prediction accuracy.



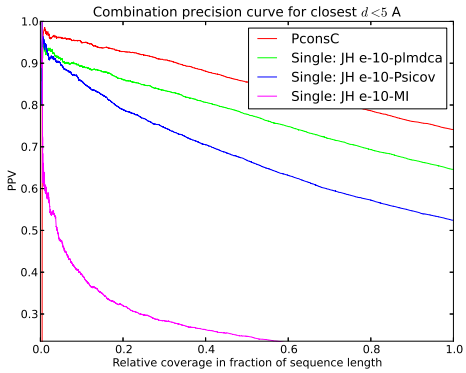


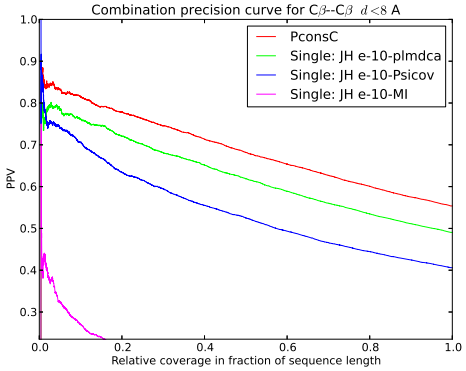
We have explored also other direct information methods (mfDCA and EVC of EVFold suite), based on the internal benchmarks, their prediction performance was on par with PSICOV and significantly less than plmDCA, as illustrated by the precision plot below (predictions with the same set of jackHMMER-based alignments, with e-value threshold of 1e-4).





Nevertheless, all of the presented methods are significantly more precise than widely used mutual information ones (MI).





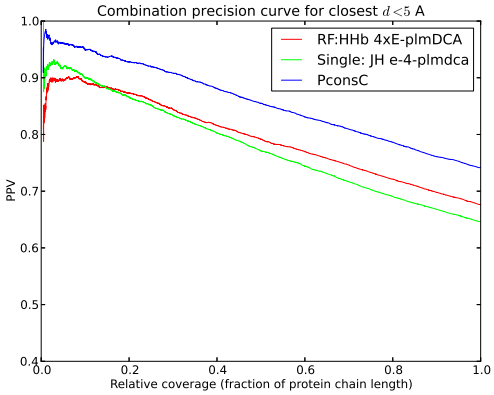
1.5 Supplementary figures

Here we present the precision (PPV) plots using additional combinations of methods and alignments than in the manuscript. In each plot a baseline from the best indivuidual method is plotted as a green line. The performance of an indidividual method is plotted in red and in blue PconsC is shown.

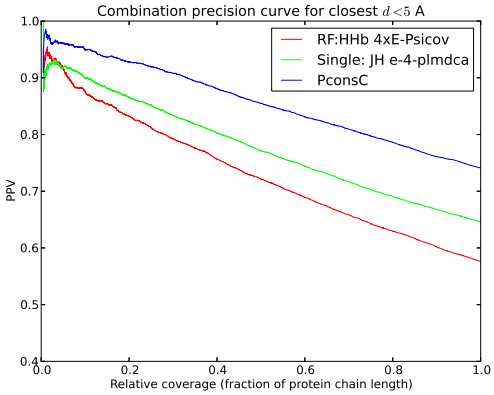
First results using the same contact definitions as in the paper are plotted. The second section shows performance of the same methods using a strict 8 Å Cβ-Cβ contact definition.

1.5.1 Closest contacts

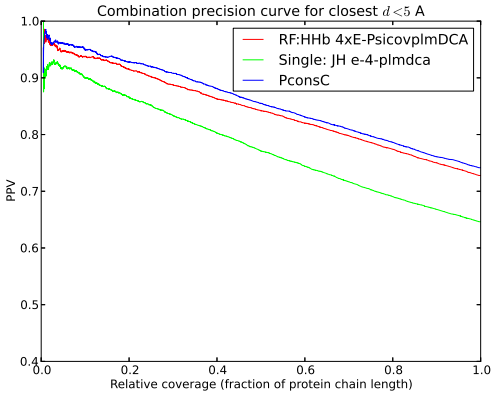
Random Forest, HHblits at 4 e-value cutoffs, plmDCA



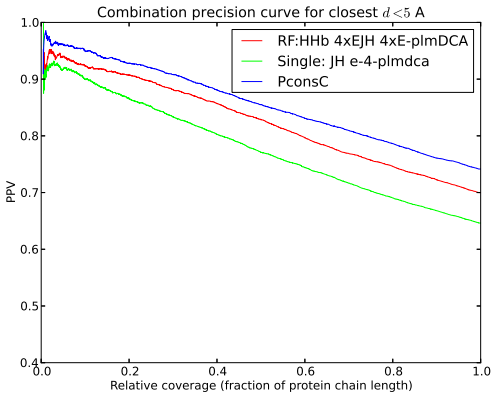
Random Forest, HHblits at 4 e-value cutoffs, PSICOV



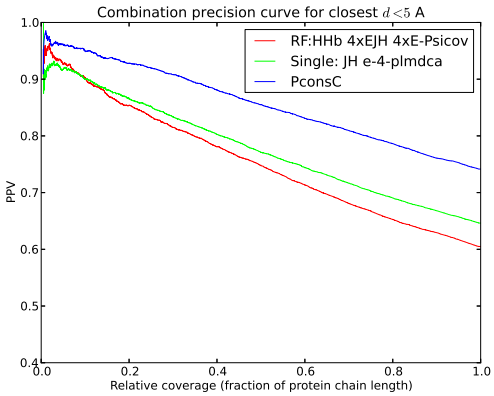
Random Forest, HHblits at 4 e-value cutoffs, PSICOV, plmDCA



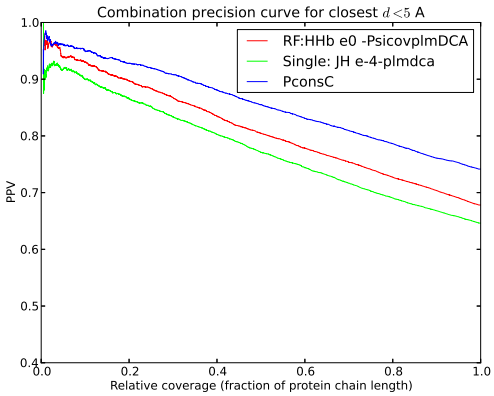
Random Forest, HHblits at 4 e-value cutoffs, jackhmmer at 4 e-value cutoffs, plmDCA



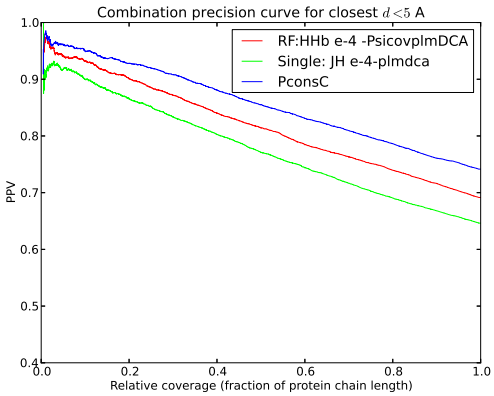
Random Forest, HHblits at 4 e-value cutoffs, jackhmmer at 4 e-value cutoffs, PSICOV



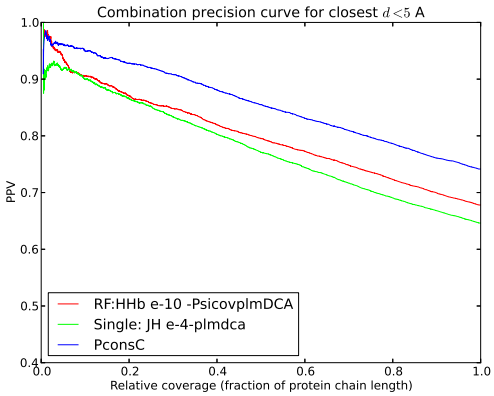
Random Forest, HHblits e-value cutoff=1, PSICOV, plmDCA



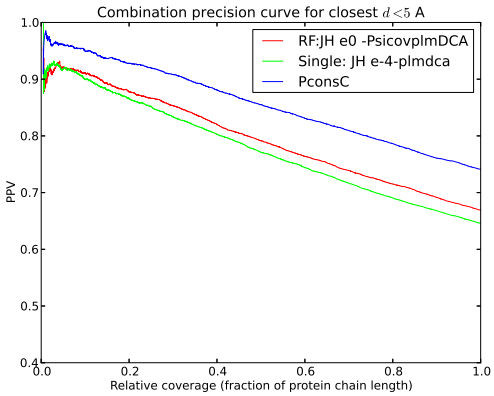
Random Forest, HHblits e-value cutoff=1e-4, PSICOV, plmDCA



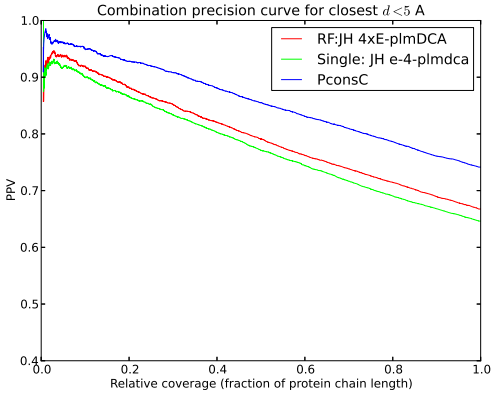
Random Forest, HHblits e-value cutoff=1e-10, PSICOV, plmDCA



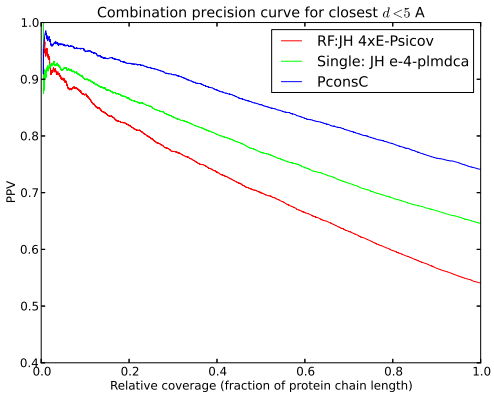
Random Forest, HHblits e-value cutoff=1e-40, PSICOV, plmDCA



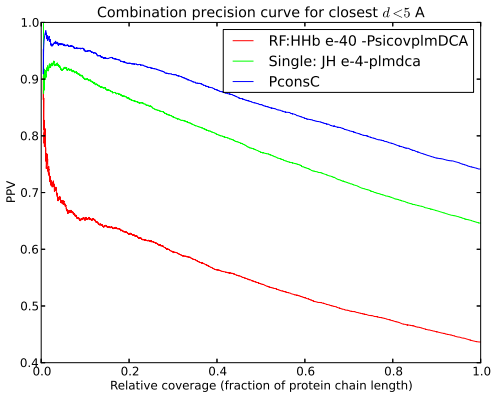
Random Forest, jackhmmer at 4 e-value cutoffs, plmDCA



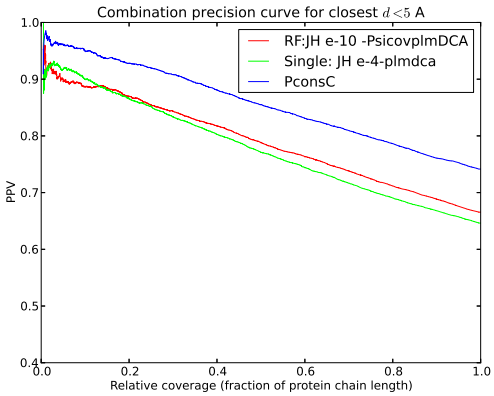
Random Forest, jackhmmer at 4 e-value cutoffs, PSICOV



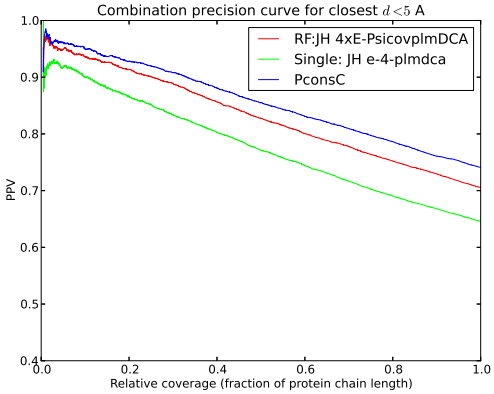
Random Forest, jackhmmer at 4 e-value cutoffs, PSICOV, plmDCA



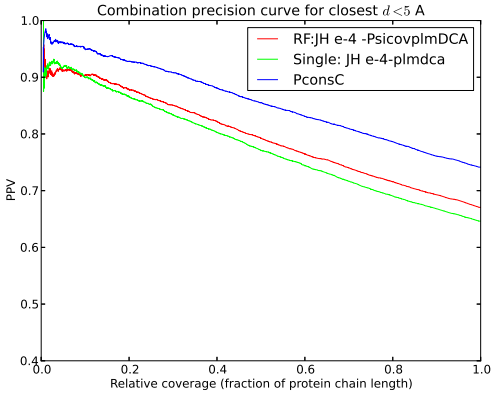
Random Forest, jackhmmer e-value cutoff=1, PSICOV, plmDCA



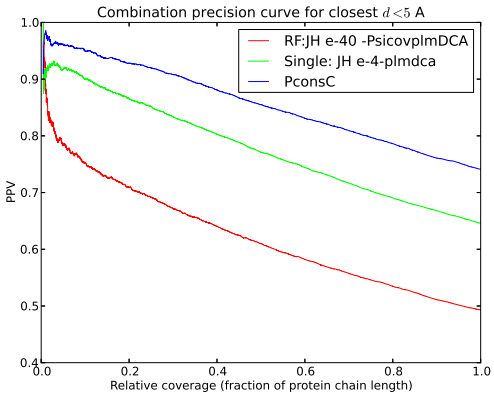
Random Forest, jackhmmer e-value cutoff=1e-4, PSICOV, plmDCA



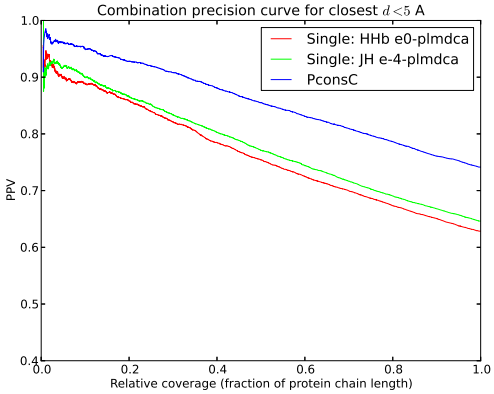
Random Forest, jackhmmer e-value cutoff=1e-10, PSICOV, plmDCA



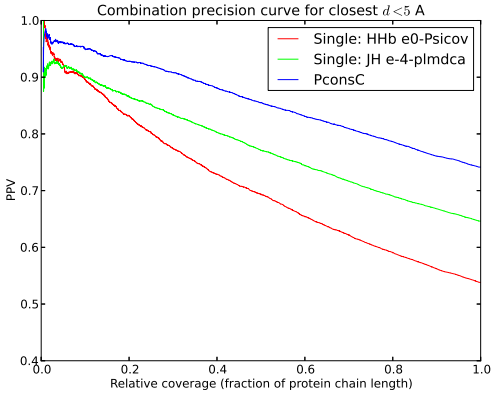
Random Forest, jackhmmer e-value cutoff=1e-40, PSICOV, plmDCA



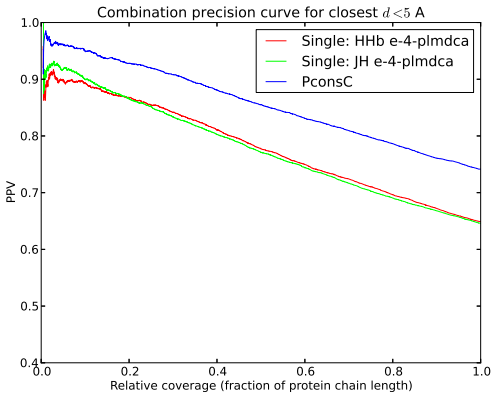
Single: HHblits e-value cutoff=1, plmDCA



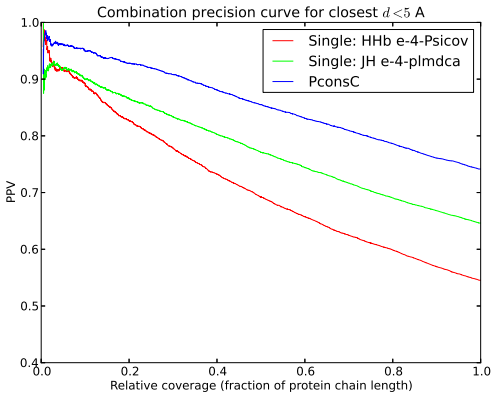
Single: HHblits e-value cutoff=1, PSICOV



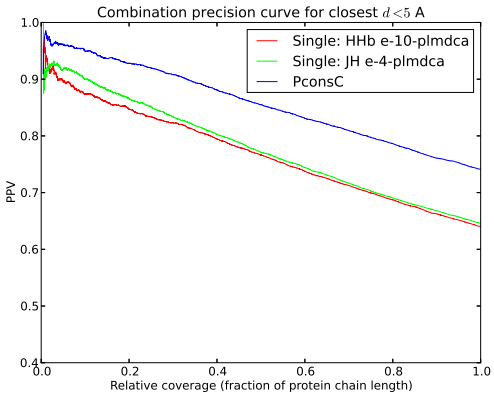
Single: HHblits e-value cutoff=1e-4, plmDCA



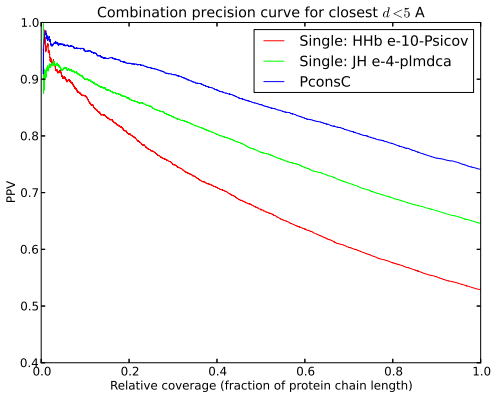
Single: HHblits e-value cutoff=1e-4, PSICOV



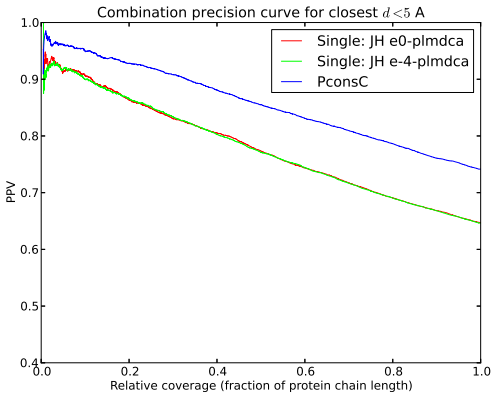
Single: HHblits e-value cutoff=1e-10, plmdCA



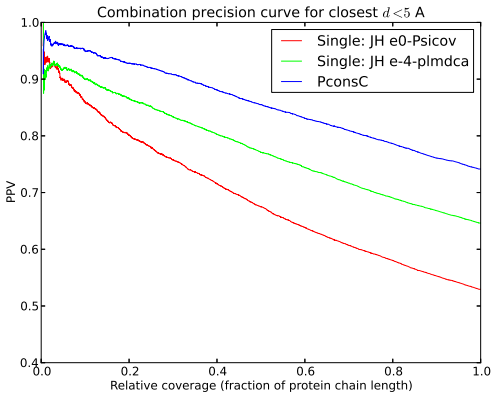
Single: HHblits e-value cutoff=1e-10, PSICOV



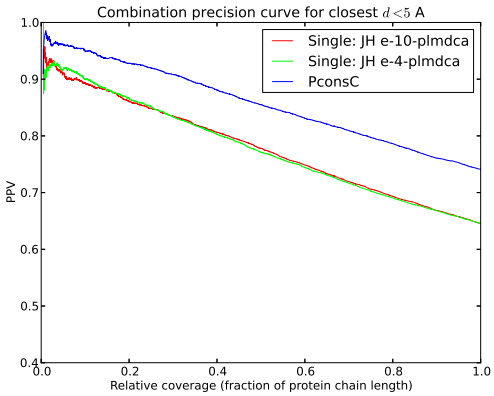
Single: jackhmmer e-value cutoff=1, plmDCA



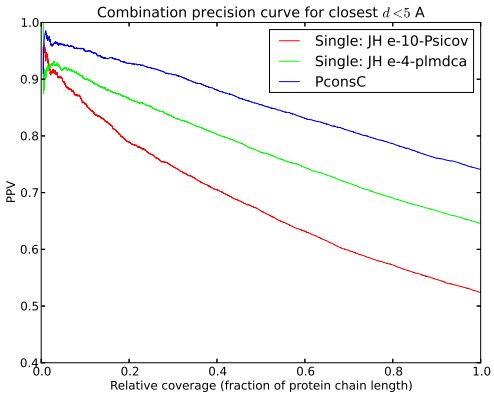
Single: jackhmmer e-value cutoff=1, PSICOV



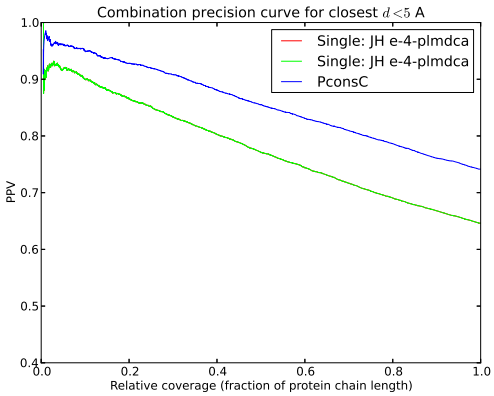
Single: jackhmmer e-value cutoff=1e-10, plmDCA



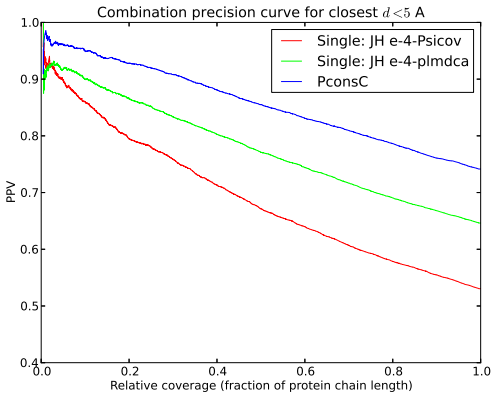
Single: jackhmmer e-value cutoff=1e-10, PSICOV



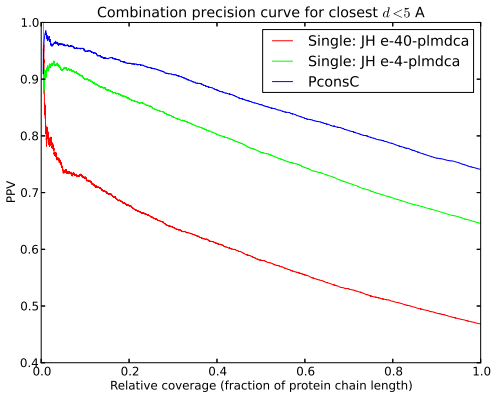
Single: jackhmmer e-value cutoff=1e-4, plmDCA



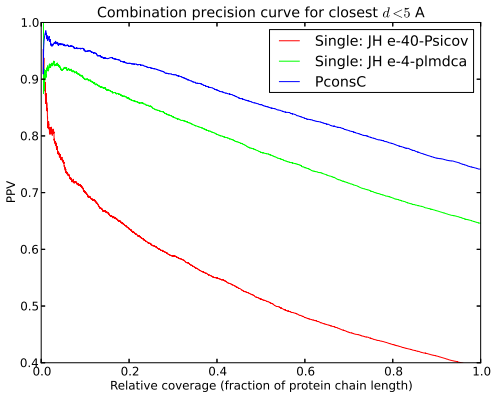
Single: jackhmmer e-value cutoff=1e-4, PSICOV



Single: jackhmmer e-value cutoff=1e-4, plmDCA

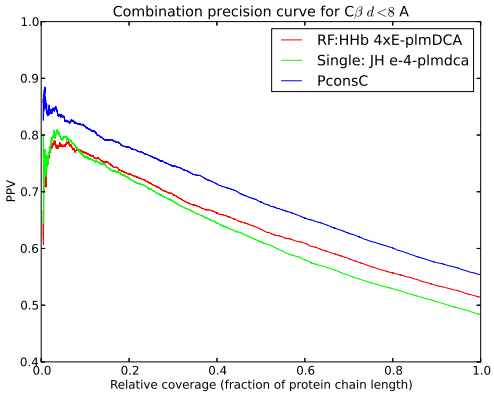


Single: jackhmmer e-value cutoff=1e-4, PSICOV

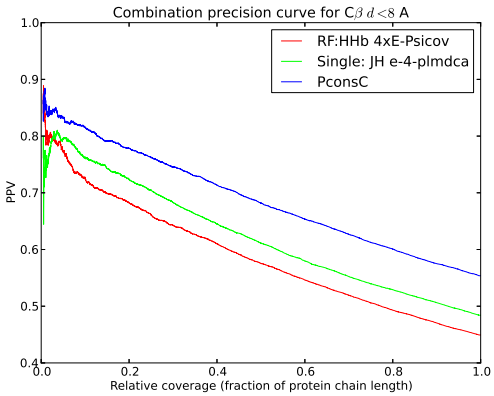


2 Cβ-Cβ contacts

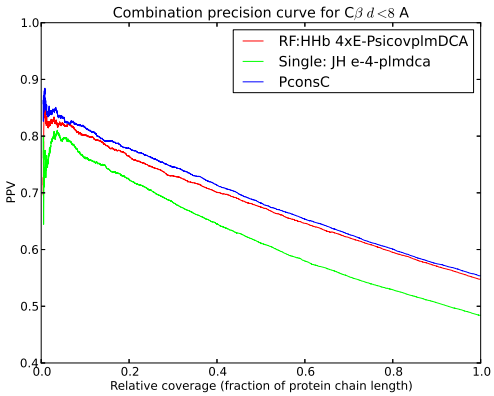
Random Forest, HHblits at 4 e-value cutoffs, plmDCA



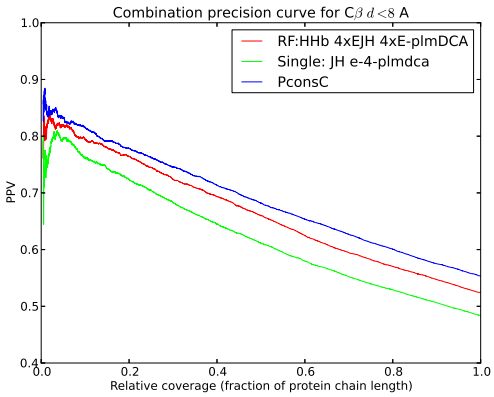
Random Forest, HHblits at 4 e-value cutoffs, PSICOV



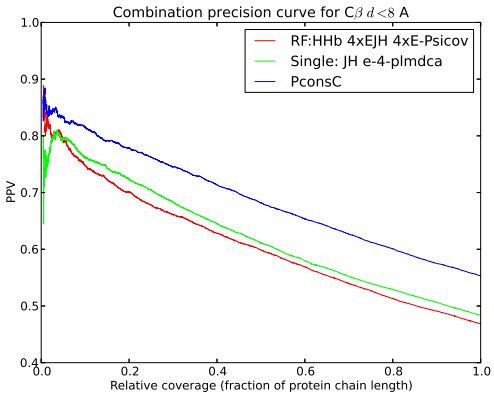
Random Forest, HHblits at 4 e-value cutoffs, PSICOV, plmDCA



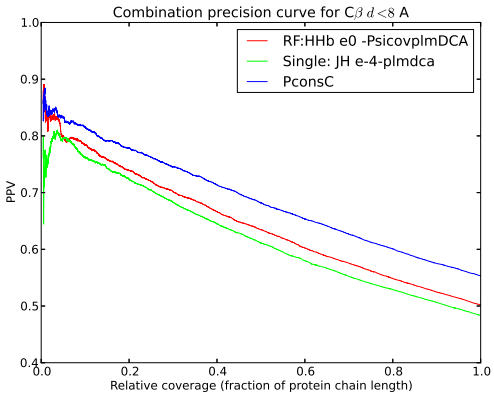
Random Forest, HHblits at 4 e-value cutoffs, jackhmmer at 4 e-value cutoffs, plmDCA



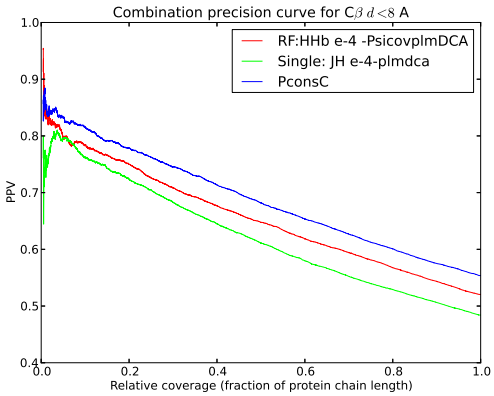
Random Forest, HHblits at 4 e-value cutoffs, jackhmmer at 4 e-value cutoffs, PSICOV



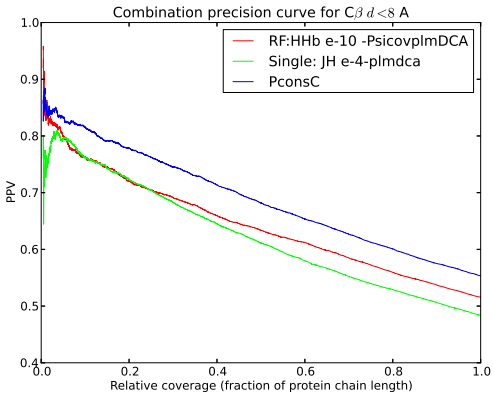
Random Forest, HHblits e-value cutoff=1, PSICOV, plmDCA



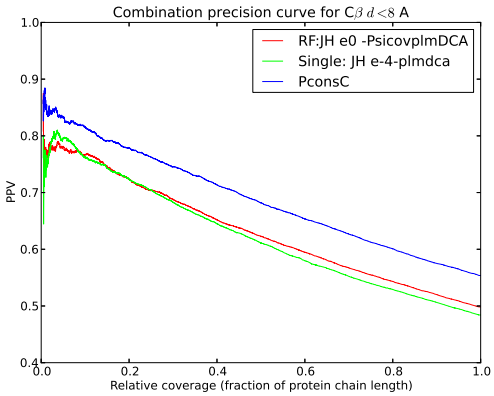
Random Forest, HHblits e-value cutoff=1e-4, PSICOV, plmDCA



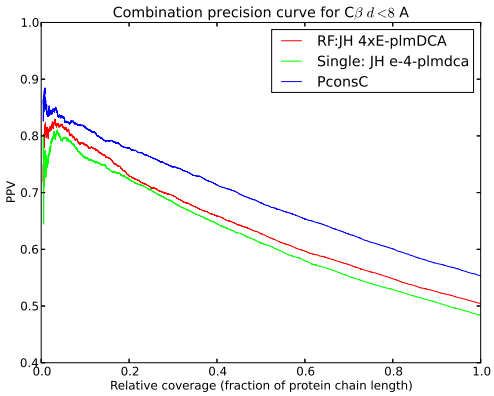
Random Forest, HHblits e-value cutoff=1e-10, PSICOV, plmDCA



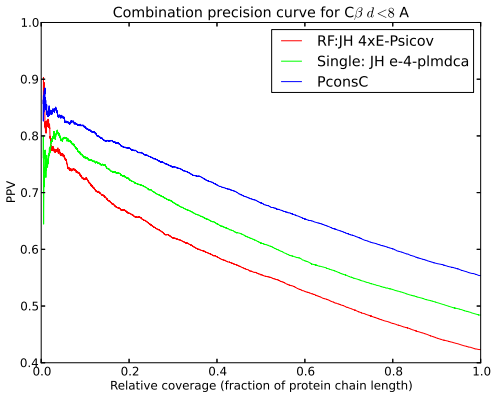
Random Forest, HHblits e-value cutoff=1e-40, PSICOV, plmDCA



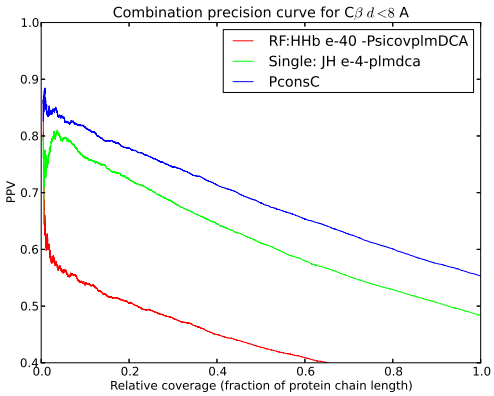
Random Forest, jackhmmer at 4 e-value cutoffs, plmDCA



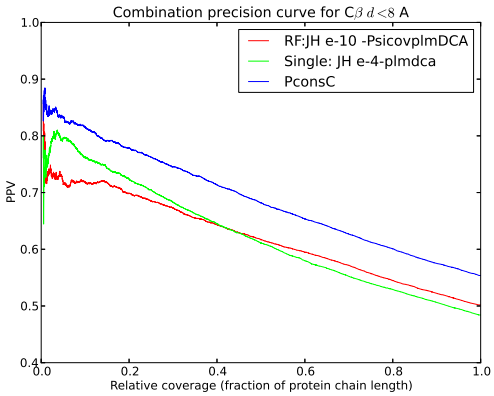
Random Forest, jackhmmer at 4 e-value cutoffs, PSICOV



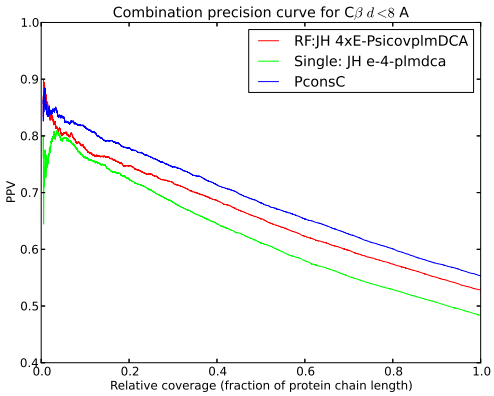
Random Forest, jackhmmer at 4 e-value cutoffs, PSICOV, plmDCA



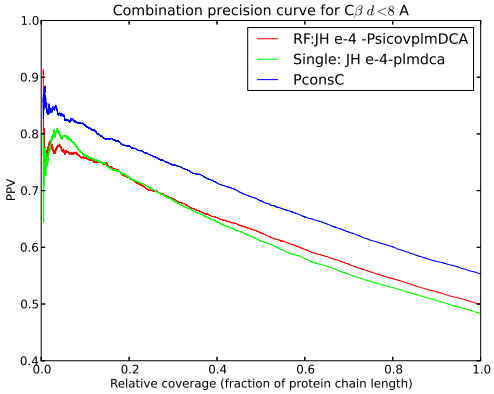
Random Forest, jackhmmer e-value cutoff=1, PSICOV, plmDCA



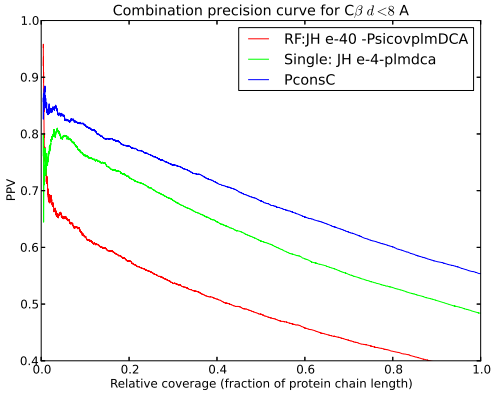
Random Forest, jackhmmer e-value cutoff=1e-4, PSICOV, plmDCA



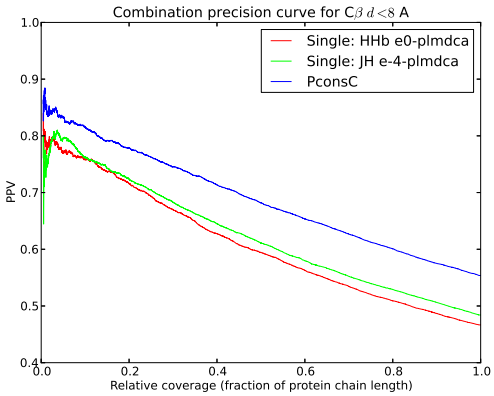
Random Forest, jackhmmer e-value cutoff=1e-10, PSICOV, plmDCA



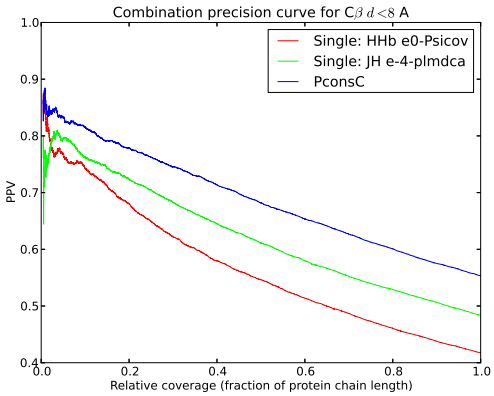
Random Forest, jackhmmer e-value cutoff=1e-40, PSICOV, plmDCA



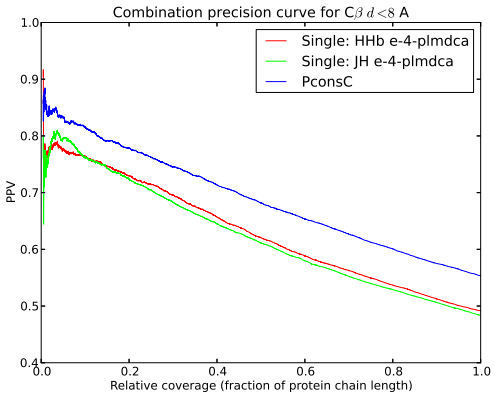
Single: HHblits e-value cutoff=1, plmDCA



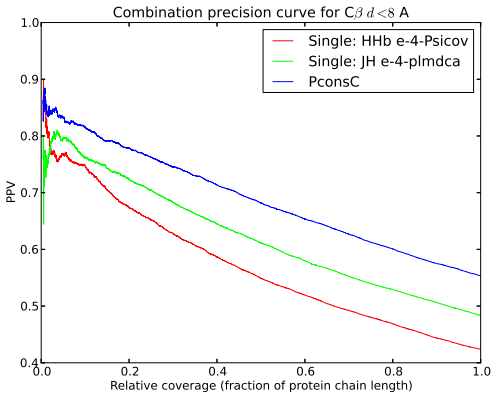
Single: HHblits e-value cutoff=1, PSICOV



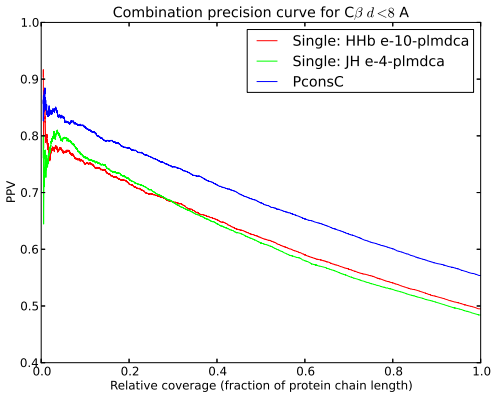
Single: HHblits e-value cutoff=1e-4, plmDCA



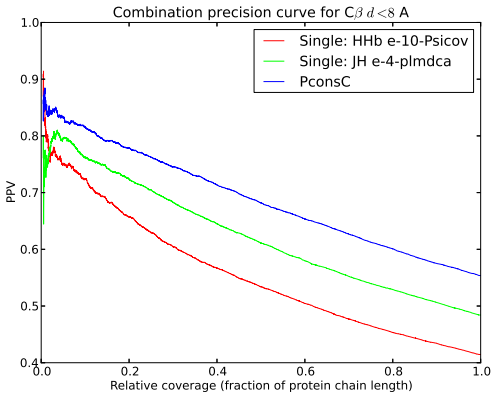
Single: HHblits e-value cutoff=1e-4, PSICOV



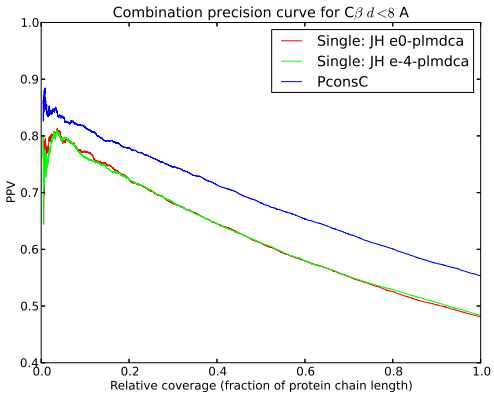
Single: HHblits e-value cutoff=1e-10, plmdCA



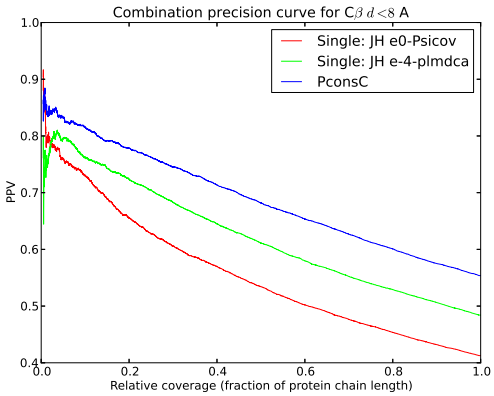
Single: HHblits e-value cutoff=1e-10, PSICOV



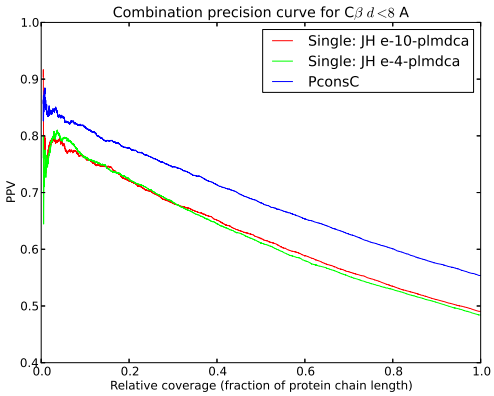
Single: jackhmmer e-value cutoff=1, plmDCA



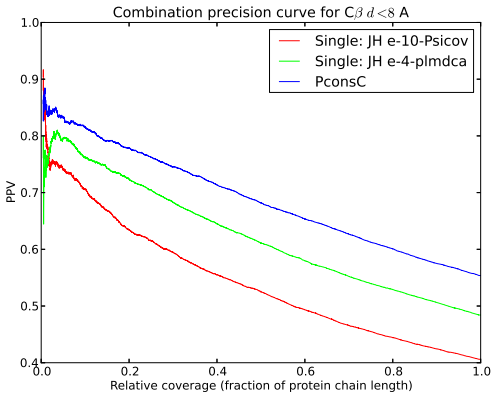
Single: jackhmmer e-value cutoff=1, PSICOV



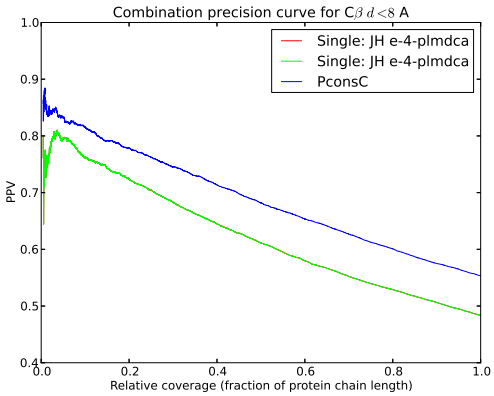
Single: jackhmmer e-value cutoff=1e-10, plmDCA



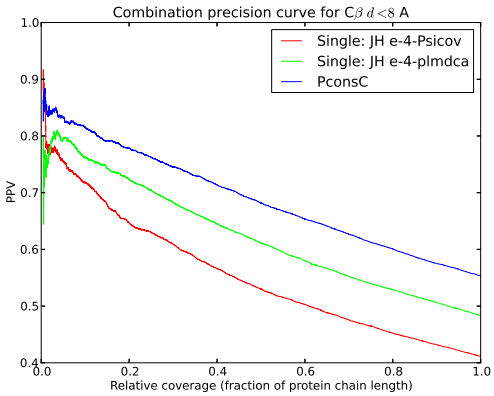
Single: jackhmmer e-value cutoff=1e-10, PSICOV



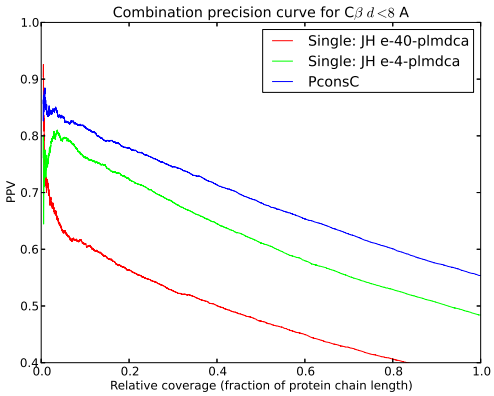
Single: jackhmmer e-value cutoff=1e-4, plmDCA



Single: jackhmmer e-value cutoff=1e-4, PSICOV



Single: jackhmmer e-value cutoff=1e-4, plmDCA



Single: jackhmmer e-value cutoff=1e-4, PSICOV