PROCAIN: PROTEIN PROFILE COMPARISON WITH ASSISTING INFORMATION

APPROVED BY SUPERVISORY COMMITTEE

Nick Grishin

Peter Antich

Dawen Zhao

Jean Gao

Zbyszek Otwinowski

DEDICATION

I would like to thank the members of my Graduate Committee members for their time and their advices on various aspects of this research project. Thank you, Dr. Nick Grishin for guiding me through this project and for always being there whenever I needed your help. Thank you, Dr. Peter Antich, Director of UTSW's BME graduate program, for recruiting me to the graduate program and always being helpful and supportive. Thank you, Drs Dawen Zhao, Zbyszek Otwinowski and Jean Gao for serving my committee and offering me great advices regarding my project.

I would like to offer thanks to everyone else in Dr. Nick Grishin's Lab for their great help and their support.

I would also like to take this chance to thank all my family members, my father, my mother, my sisters and my little brother for their care and their support during all these years.

PROCAIN: PROTEIN PROFILE COMPARISON WITH ASSISTING INFORMATION

Yong Wang

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

April, 2009

PROCAIN: PROTEIN PROFILE COMPARISON WITH ASSISTING INFORMATION

Yong Wang, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, 2009

Nick Grishin, Ph.D.

Detection of remote sequence homology is essential for the accurate inference of protein structure, function, and evolution. The most sensitive detection methods involve the comparison of evolutionary patterns reflected in multiple sequence alignments of protein families. We present PROCAIN, a new method for MSA comparison based on the combination of 'vertical' MSA context (substitution constraints at individual sequence positions) and 'horizontal' context (patterns of residue content at multiple positions). Based on a simple and tractable profile methodology and primitive measures for the similarity of horizontal MSA patterns, the method achieves the quality of homology detection comparable to a more complex advanced method employing hidden Markov models and secondary structure prediction. Adding secondary structure information further improves PROCAIN performance beyond the capabilities of current state-of-the-art tools. The potential value of the method for structure/function predictions is illustrated by the detection of subtle homology between evolutionary distant yet structurally similar protein domains. ProCAIn, relevant databases and tools can be downloaded from http://prodata.swmed.edu/procain/download. The web server can be accessed at http://prodata.swmed.edu/procain/procain.php.

Table of Contents

Prior Publication

1. Yuan Qi, Ruslan Sadreyev, Yong Wang, Bong-Hyun Kim and Nick Grishin, "*A comprehensive system for evaluation of remote sequence similarity detection", BMC Bioinformatics*, 2007, 8:314.
2. Donald W. Hilgemann, Alp Yaradanakul, Yong Wang and Daniel Fuster, "*Molecular Control of Cardiac Sodium Homeostasis in Health and Disease", Journal of Cardiovascular Electrophysiology*, Vol. 17, No. 5, S47-56.
3. Hao Che, Yong Wang and Zhijun Wang, "*A Rule Grouping Technique for Weight-based TCAM Coprocessors", in the Proceedings of IEEE Hot Interconnects*, 2003.

List of Figures

List of Tables

CHAPTER 1:

Introduction

1.1 Significance of Protein Structure Prediction

1.1.1 Protein structure prediction.

One of the main goals for bioinformatics has always been protein structure prediction. The aim here is to predict the 3D structure of a protein starting from its known amino acid sequence. In other words, for proteins with known sequences, different algorithms or methods are applied to predict their 3D structures before using real experimental methods (such as crystallography (Ealick 2000) or NMR spectroscopy (Tyszka, Fraser et al. 2005)) to actually solve their structures.

In recent years, protein structure prediction is becoming more and more important because of the massive amounts of protein sequence data produced by large-scale DNA sequencing projects such as the Human Genome Project (Barnhart 1989). Despite the huge efforts of structural biologists, the structures of most of these proteins are still unsolved. There are about 12 millions proteins within the NR database and only about 50k of these proteins has experimentally solved protein tertiary structures. This means only about 0.4% of the proteins have solved structures. The reason for this is simply because both of the above mentioned experimental methods have their limitations. X-ray crystallography normally can solve protein structures with better precision, but it is very difficult to get stable crystals for some proteins. And NMR spectroscopy is only limited to small proteins. Also both these two methods are very

expensive and time consuming. All these facts point to the importance of genome-wide structure prediction.

However, despite a lot of research by bioinformaticians, it is still an unfinished task to precisely predict a protein's 3D structure. There are mainly two reasons for this. The first reason is that a protein can potentially form a huge amount of different structures. A protein structure prediction method has to efficiently search through all these possible structures to find the correct structure for the protein. The second reason is that we are still not clear about the physical basis of protein structure folding and stability. And this makes it difficult to identify the correct structure of the protein.

1.1.2 Two main protein structure prediction methods

1.1.2.1 De novo or ab initio structure prediction

The de novo- protein prediction method predicts a protein's structure directly based on physical principles. This method tries to predict how a protein folds or search through all possible structures of the protein and find the one with minimal energy, both of which require huge amount of computation times. Although this protein structure prediction method has only limited success so far, it is still a very interesting research area and it attracts a lot of researchers because of its potential. And for proteins for which it is difficult to find any homologues, this method is the only method a bioinformatician can resort to.

1.1.2.2 Comparative structure prediction

Proteins which share similar structures and functions are called homologous proteins. If a protein (called target protein) with unknown structure is predicted as a homologous protein with proteins with known structure (called subject or template proteins), the target protein will share similar structure with the template proteins. The comparative protein structure prediction method takes advantage of this structure similarity between the target protein and the template proteins.

The structure similarity between the target protein and the template protein could be a global structure similarity or a local structure similarity. For the latter case, the comparative structure prediction method will have to find all the similar structures for each regions of the target protein and assemble all these structure pieces together. If the correct template or templates for a target protein can be detected, the accuracy of the comparative structure prediction method can be high.

Obviously the accuracy of this protein structure prediction method depends on two factors. The first factor is whether the correct template(s) can be detected (homology detection accuracy). The second factor is whether the sequence alignment between the target protein and the template protein is accurate or not (alignment accuracy). Unsurprisingly, when the target protein and template protein share similar amino acid sequences (high sequence identity), protein comparative structure prediction method will give the best prediction accuracy (Nayeem, Sitkoff et al. 2006).

So far the most successful automated protein structure prediction method is the mixture of the de novo- protein prediction method and the comparative protein structure prediction method.

The ROSETTA (Das, Qian et al. 2007) program from Dr. David Baker's lab and TASSER (Zhang, Arakaki et al. 2005) program from Dr. Yang Zhang's lab are the representatives of this research direction. Both programs start the prediction by assembling the structure pieces found by protein homology detection programs. If for these regions no similar structures are detected, de novo- protein prediction method will be used to predict their structures.

Most successful protein structure predictions by human structural biology experts also require the correct detection of the homologous templates for the target proteins. All these make protein homology detection the starting point and the most critical part for a successful protein structure prediction attempt.

1.2 History of Protein Homology Detection

Three methods are available for protein homology detection.

1.2.1 The first method is protein sequence alignment.

The BLAST (Basic Local Alignment Search Tool)(Altschul, Gish et al. 1990) program from Dr. Altschul's Lab is one of the most widely used protein sequence alignment program. It is a protein sequence-sequence comparison method. BLAST performs very well for proteins with high sequence identity.

1.2.2 The second method is protein profile-sequence comparison.

In 1997, Dr. Altschul's lab developed PSI-BLAST (Position Specific Iterated BLAST)(Altschul, Madden et al. 1997), a program which uses protein profile-sequence comparison method to detect protein homology. PSI-BLAST firstly uses BLAST to search the query sequence against

one or several sequence databases and then uses the resulted multiple sequence alignment to derive a protein profile (also called position specific scoring matrix). A protein profile is a position-specific numerical representation of the residue content of the multiple sequence alignment. For an alignment of length n, the profile is a matrix of $n \times 21$. Each column of the matrix corresponds to a position in the alignment and includes 20 numbers for each type of amino acid residue, plus one number for gap symbols(Sadreyev and Grishin 2003).

PSI-BLAST then searches this profile against the sequence databases to find more homologous sequences to incorporate into the multiple sequence alignment. PSI-BLAST iterates the search process until the required iteration round is reached or the program converges. Protein profile-sequence comparison method is proven to have a better homology detection performance because a protein profile derived from protein multiple sequence alignment incorporates much more information than a single protein sequence.

1.2.3 The third method is protein profile-profile comparison.

Instead of searching the query sequence against the sequence databases, protein profile-profile comparison methods search the profile of the query sequences against the databases of profiles. It was proved to be one of the better methods for protein homology detection.

Several profile-profile comparison methods, such as COMPASS (Sadreyev and Grishin 2003) and HHsearch (Soding 2005) are available now. COMPASS uses the scheme of log-odds rations. HHsearch uses HMM (Hidden Markov Model)(Eddy 1998). However, even the best profile-profile comparison methods (such as COMPASS or HHsearch) still lack good detection accuracy, especially for hard targets, such as Free Models (protein structures which are not found in

nature) or protein pairs whose pair-wise sequence identities are below 20%. Further research is required to improve the performance of protein profile-profile comparison method in order to provide better protein templates to make genomic wide protein structure prediction possible.

HHsearch incorporates secondary structure (predicted or real) information into the algorithm to assist with homology detection and statistical analysis. This method is proved to be helpful for protein homology detection, especially for the detection of remote protein homologues.

All these evidence point to the possibility that adding more assisting information can help improve the performance of a protein homology detection algorithm. PSI-BLAST gains a better performance by adding query family evolution information; COMPASS gains a further improvement by adding the evolution information of both the query and the subject family; HHsearch adds SS information and makes it even more sensitive. The goal of this project is to add more types of assisting information into protein profile-profile comparison process to test whether they are helpful.

1.3 Overview of Dissertation Work

My dissertation work developed a Protein profile-profile Comparison method with Assisting Information: ProCAIn. After numerous evaluations, ProCAIn is proved to be more sensitive than the best homology detection methods currently existing and also be able to provide better alignment qualities.

Figure 1 is the flowchart for ProCAIn. ProCAIn can be generally divided into two steps. The first step is similarity comparison. For the query protein sequence, a Multiple Sequence Alignment

(MSA) is built by running PSI-BLAST. A MSA is also built for the subject sequence. Then sequence similarity score (3.a), sequence motif score (3.b), amino acid conservation score (3.c) and secondary structure score (3.d) are derived from the pair of MSAs. All these scores are added together to get an all positions against all positions similarity score matrix. This score matrix is feed to Smith-Waterman algorithm(Smith and Waterman 1981) to produce the final optimal score and sequence alignment for this pair of query sequence and subject sequence.

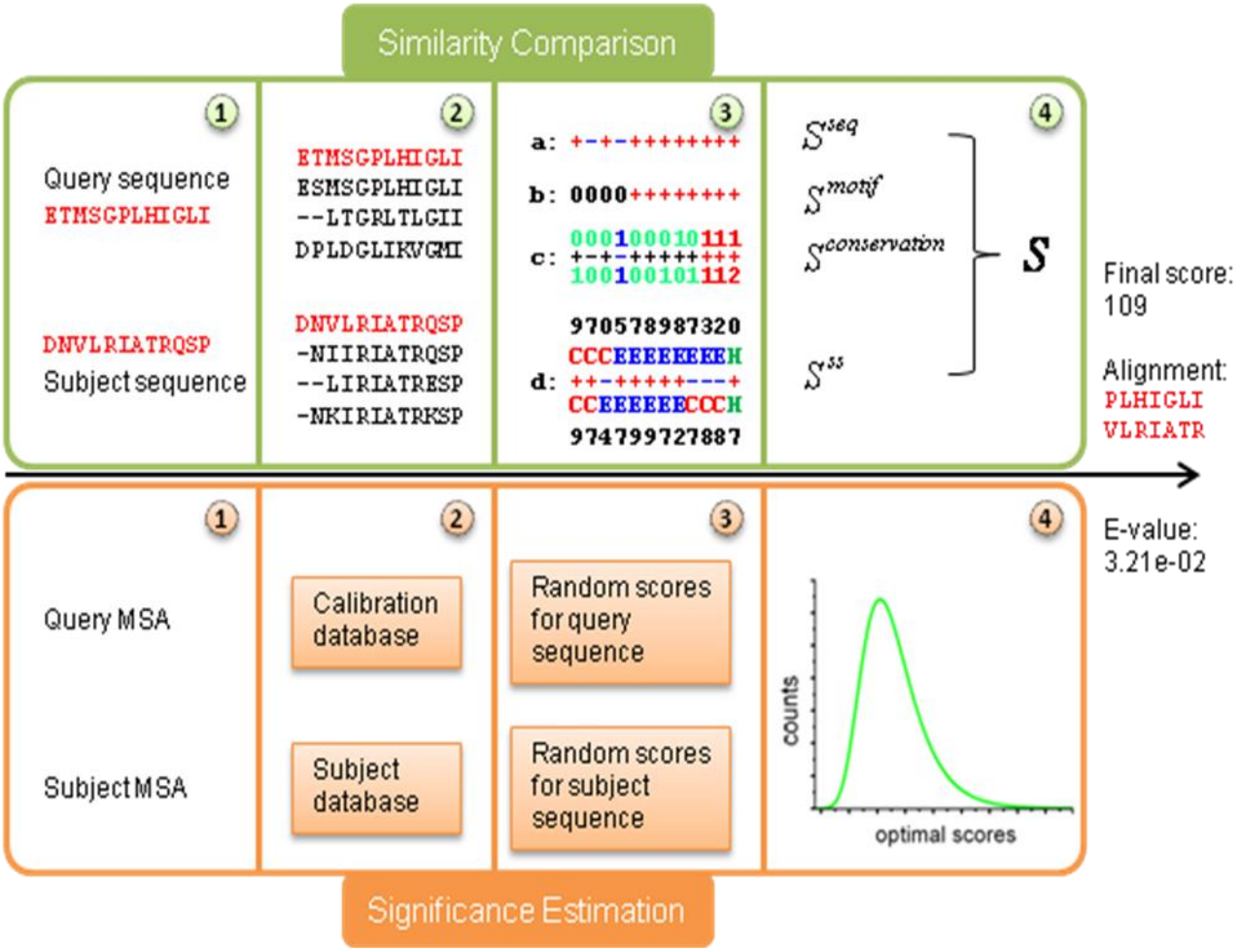

Figure 1 Flowchart of ProCAIn

The second step is statistical significance estimation. ProCAIn compares the query MSA against a calibration database of MSAs to get random scores for the query sequence. ProCAIn then compares the subject MSA against all the other MSAs within the subject database to get random scores for the subject sequence. These two sets of random scores are put together and fit with Extreme Value Distribution (EVD) using the maximum likelihood method(Eddy 1997) to get the two statistical parameters: k and λ. These two parameters as well as the optimal score are used in the Gumbel extreme value distribution equation (Gumbel 1958; Karlin and Altschul 1990), $E = kmne^{-\lambda S}$ to calculate E-value, which is a representative of the similarity significance between the query sequence and the subject sequence. Here *m* and *n* are length of the two profiles and *S* is the optimal score.

I benchmarked ProCAIn together with profile-profile comparison tools HHsearch and COMPASS in an all-to-all comparison of a SCOP(Murzin, Brenner et al. 1995) representative database of 4147 protein domains(Qi, Sadreyev et al. 2007) to evaluate its homology detection performance and its alignment quality. In order to find the weakness of ProCAIn, I used a lot of evaluation methods. Receiver operator characteristics (ROC) curves(Schaffer, Aravind et al. 2001) and protein family pair-wised statistical methods are used to evaluate homology detection sensitivity. Sequence alignment quality is evaluated with methods such as accuracy, coverage and average global distance total test score (GDT_TS)(Zemla 2003).

The SCOP database we used is a well-classified database. All-to-all structural and sequence comparison were applied to the database. Results are feed to Support Vector Machine (SVM) (Joachims 1999) to get a SVM value. The higher the SVM value is, the more similar the protein

pair is. We used the SVM value together with SCOP hierarchy relationship as our gold standard to evaluate homology detection ability. A protein pair from the same SCOP superfamily is considered as close homologous proteins. A protein pair from different SCOP superfamily but with a SVM value bigger than 0.6 is considered as remote homologues. A protein pair not belonging to the same SCOP superfamily and with a SVM value between -0.6 and 0.6 is considered as uncertain and is not counted during evaluation. A protein pair not belonging to the same SCOP superfamily and with a SVM value less than -0.6 is considered non-homologues. We believe a good homology detection method should not only be able to detection close homologues and remote homologues, but also be able to present users close homologues first, then remote homologues. This ranking is important for protein structure modeling, function prediction or protein evolution.

CHAPTER 2:

Adding Sequence Motif Score

2.1 Biological Observation

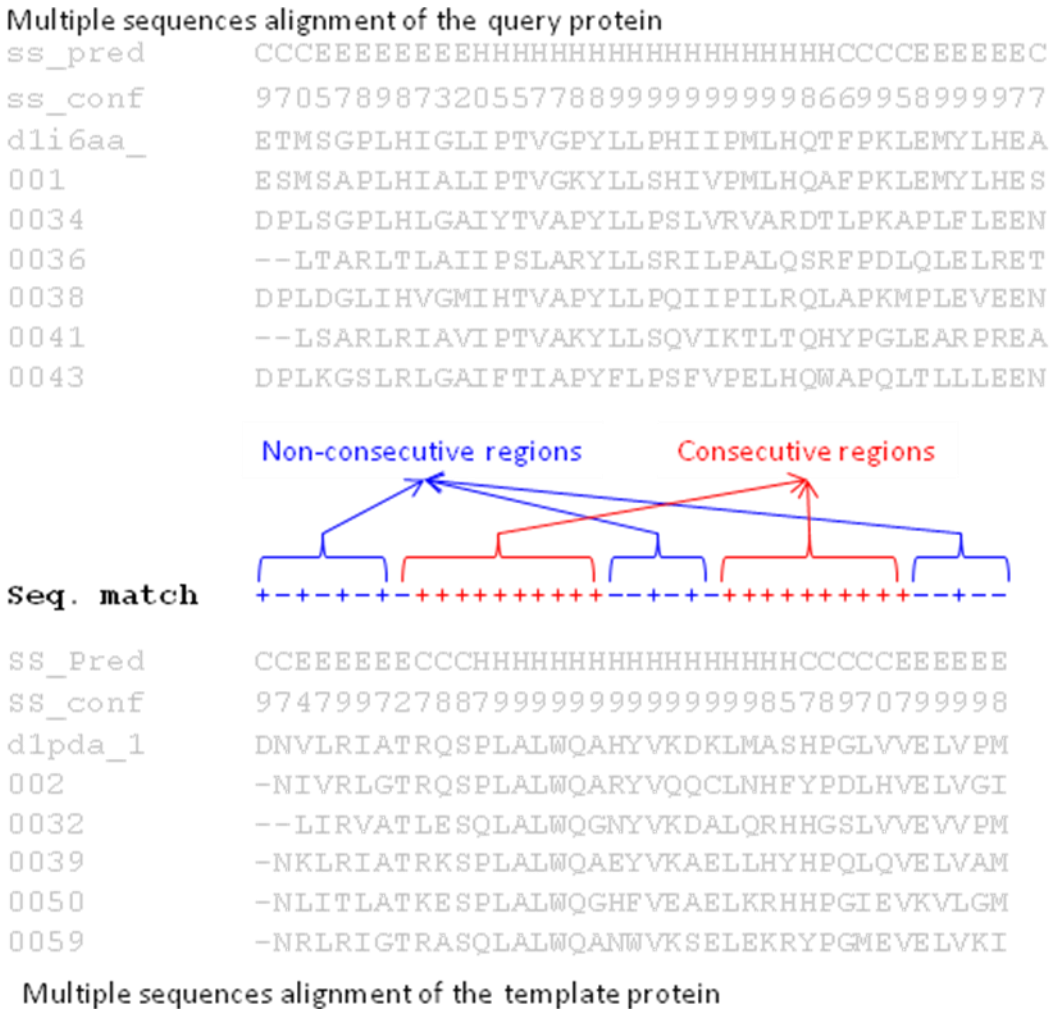

Figure 2 Adding Sequence Motif Information

It was shown by Pei et al. that in alignments of homologous sequences conserved columns tend to occur in clusters along the sequence(Pei and Grishin 2001). The reason for this might be because clustered matches indicate sequence motif matches, which is a good indication of functional matches(Siddharthan, Siggia et al. 2005), hence homologues. This observation is included by ProCAIn. For the above alignment segment, "+" means the corresponding residue pairs are similar and "-" means the corresponding residue pairs are not similar. The red segments have at least three continuous residue matches and more possibly are sequence protein motif matches. The sequence matches within the blue segments are not continuous and more possibly meaningless random matches. Motif matches indicate functional matches and should be rewarded. I designed an algorithm to take advantage of this property to improve the homology detection performance of protein profile-profile comparison method.

2.2 Algorithm

I used the following programming code to execute the algorithm, where S_{ij} is the matrix of all-to-all similarity scores between the query protein and the template protein, w is the weight parameter. For a position pair between the query profile and the subject profile, if its sequence similarity score is positive, and both its previous and next neighboring position pairs have positive sequence similarity scores, then S^{motif} is the sum of these three sequence similarity scores multiplied by a sequence motif score weight w^{motif}. w^{motif} is trained with a testing database and is a constant for all query sequences. I tried different weight parameters from 0.2, 0.3, to 0.8 and found that weight parameter 0.6 gives me the best performance.

$$S^{motif} = w^{motif}(S_{ij}^{seq} + S_{(i-1)(j-1)}^{seq} + S_{(i+1)(j+1)}^{seq})$$

When $S_{ij}^{seq} > 0$ and $S_{(i-1)(j-1)}^{seq} > 0$ and $S_{(i+1)(j+1)}^{seq}$ are all true. Here S_{ij}^{seq} is the all-to-all similarity score matrix is:

$$S^{seq} = c_1 \sum_i n_i^1 \ln \frac{Q_i^2}{p_i} + c_2 \sum_i n_i^2 \ln \frac{Q_i^1}{p_i}$$

$$c_1 = \frac{\sum_i n_i^2 - 1}{\sum_i n_i^1 + \sum_i n_i^2 - 2}$$

$$c_2 = \frac{\sum_i n_i^1 - 1}{\sum_i n_i^1 + \sum_i n_i^2 - 2}$$

Here i represents the 20 residues. n_i^1 and n_i^2 are the effective residue counts(Sunyaev, Eisenhaber et al. 1999) in columns 1 of the query profile and columns 2 of the subject profile. Q_i^1 and Q_i^2 are estimated target residue frequencies(Tatusov, Altschul et al. 1994) of the two columns. p_i is the background residue frequency. c_1 and c_2 are scales to balance the contribution of the two columns.

2.3 Statistical Significance Estimation

E-value is used to estimate the significance of each resulting alignments. First, we compare the query protein profile with all the profiles of the calibration database to get optimal scores for each protein profiles. Because the calibration database is composed of protein domains from each SCOP folds, so the number of homologous scores for the query profile is minimized. Then each score is subtracted by the pre-calculated score average of each calibration profiles. This step is to make the resulted score independent on the calibration profile properties and only dependent on the properties of the query protein profile. For the subject profile, we use the

pre-calculated non-homologous scores as its random scores and each score is also subtracted by the corresponding subject protein profiles to make the scores only dependent on the properties of the subject protein profile. The subtracted random scores are collected together to fit the Extremely Value Distribution to calculate parameters k and λ (Eddy 1997).

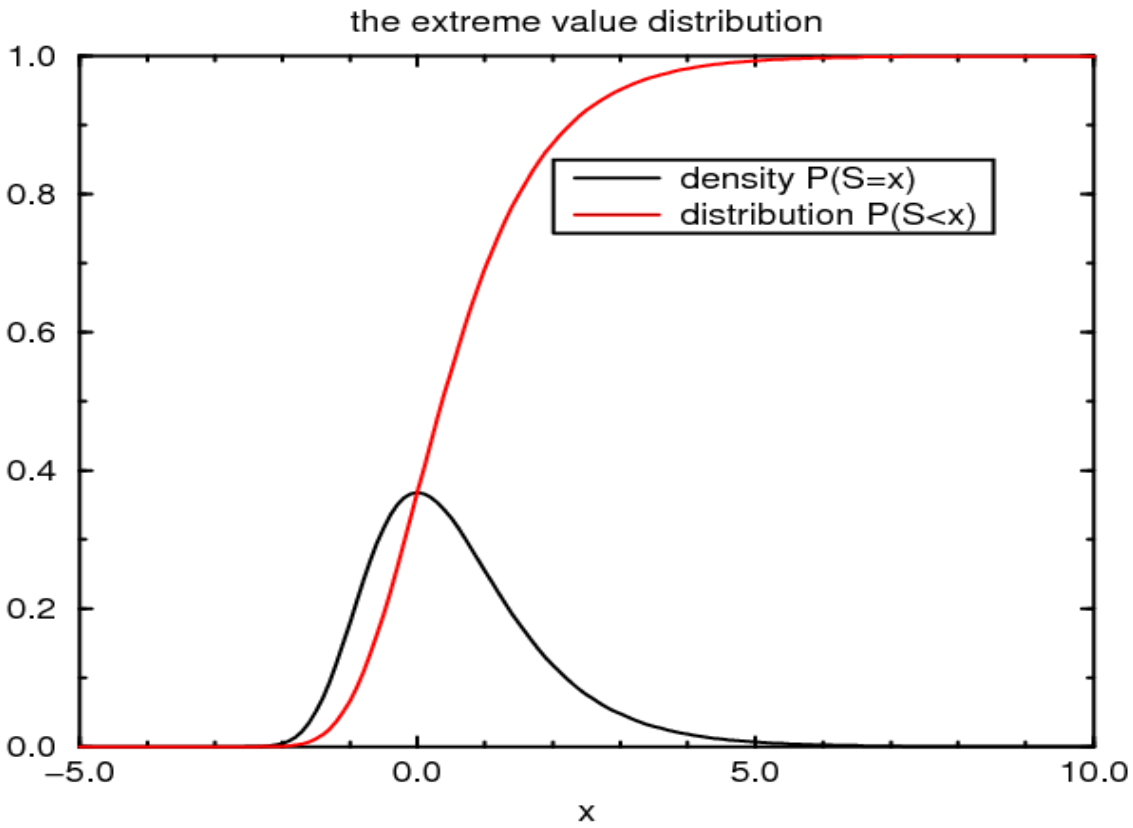

Figure 3 the PDF and CDF of Extreme Value Distribution

Probability density function (PDF) of EVD is:

$$P(x) = \lambda \exp\left[-\lambda(x-\mu) - e^{-\lambda(x-\mu)}\right]$$

Cumulative distribution function (CDF) of EVD is:

$$P(S < x) = \exp\left[-e^{-\lambda(x-\mu)}\right]$$

The μ and λ parameters are location and scale parameters:

$$k = e^{-\mu\lambda}$$

These two parameters as well as the optimal score are then used in the Gumbel extreme value distribution equation proposed by Karlin and Altshul(Karlin and Altschul 1993) for EVD to calculate E-value.

$$E = kmne^{-\lambda s}$$

Where k and λ are statistical parameters of EVD, m and n are the lengths of the query profile and the template profile.

Instead of fitting only the target family scores to calculate k and λ, here I also use the subject family scores. The reason is because the subject proteins are from the protein structure database (SCOP), so the property (structure similarity) and relation (same superfamily or not) between the entire subject proteins are known. I use these known properties and relations to get rid of the homologous protein scores to better estimate k and λ. I also adjust the optimal scores by the average value of the query family scores. All these techniques make the results statistical significance estimation able to detect more true positives and less false positives, in other words, to be more sensitive.

A calibration database of 935 protein SCOP domains is formed by picking a representative protein domain from each SCOP fold(Soding 2005). The subject database is

composed of 4147 SCOP protein domains. MSAs are formed for all the protein sequences for both databases by running *buildali.pl* and then preprocessed into corresponding profiles using ProCAIn profile extraction process. Secondary structures are also predicted for all the proteins for both databases. An all profile-to-all profile comparison is done for all protein profiles within the subject database using ProCAIn to get optimal scores and then, for each protein domain, the average of its non-homologues is calculated. Each protein profile of the calibration database is compared with all the profiles of the subject database using ProCAIn to calculate the average scores for each calibration database protein profile. The average scores of the protein profiles for both calibration database and subject database are a good indication of the properties of these profiles. A protein profile with long length tends to get big scores, even when compared with the profile of a totally random protein.

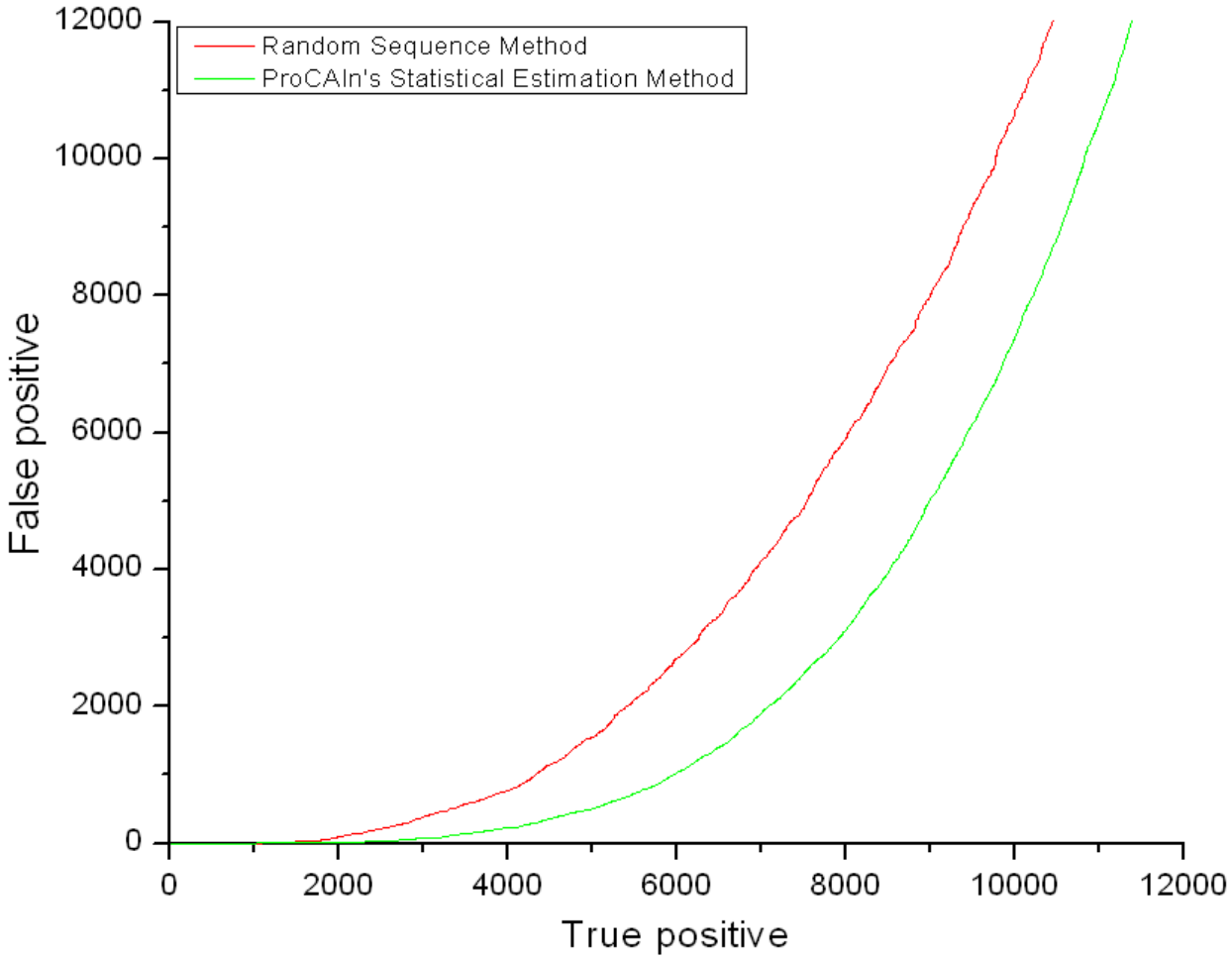

Figure 4 Comparison between ProCAIn's Statistical Estimation Method and The Random Sequence Method

The above ROC shows the comparison between the result of ProCAIn's statistical significance estimation method and the result of random sequence method, which is popularly used by many homology detection programs such as COMPASS and PSI-BLAST. It is very clear that ProCAIn's statistical significance estimation method performs much better. This is because ProCAIn's statistical significance estimation method examines the properties of both target

protein and subject protein. More information is involved in ProCAIn's statistical significance estimation method and this provides better homology detection sensitivity.

2.4 Results

I used SCOP (Structural Classification of Proteins) database(Murzin, Brenner et al. 1995) as the gold standard to evaluate my method's homology detection ability. Protein pairs which belong to the same SCOP super-family are usually believed to be homologous proteins and protein pairs which belong to different SCOP classes are normally believed to be non-homologous proteins.

I used Dali (Holm and Sander 1996) (a protein structure alignment program) structural alignment results as the gold standard to evaluate my method's alignment quality. If the sequence alignment produced by my algorithm match with the corresponding sequence alignment produced with Dali structural alignment, then the first alignment is believed to be a correct alignment.

Numerous evaluation methods are used to test ProCAIn results against the results of other homology detection methods to find the possible weakness of ProCAIn. Most of these methods are from one of my published works (Qi, Sadreyev et al. 2007) and the rest are designed specifically for this project.

The following is a flowchart of the evaluation process. The evaluation methods will be explained one by one in the following sections. I tested ProCAIn with all available evaluation

methods, but not all testing results are shown in this report because results are very consistent with difference evaluation methods.

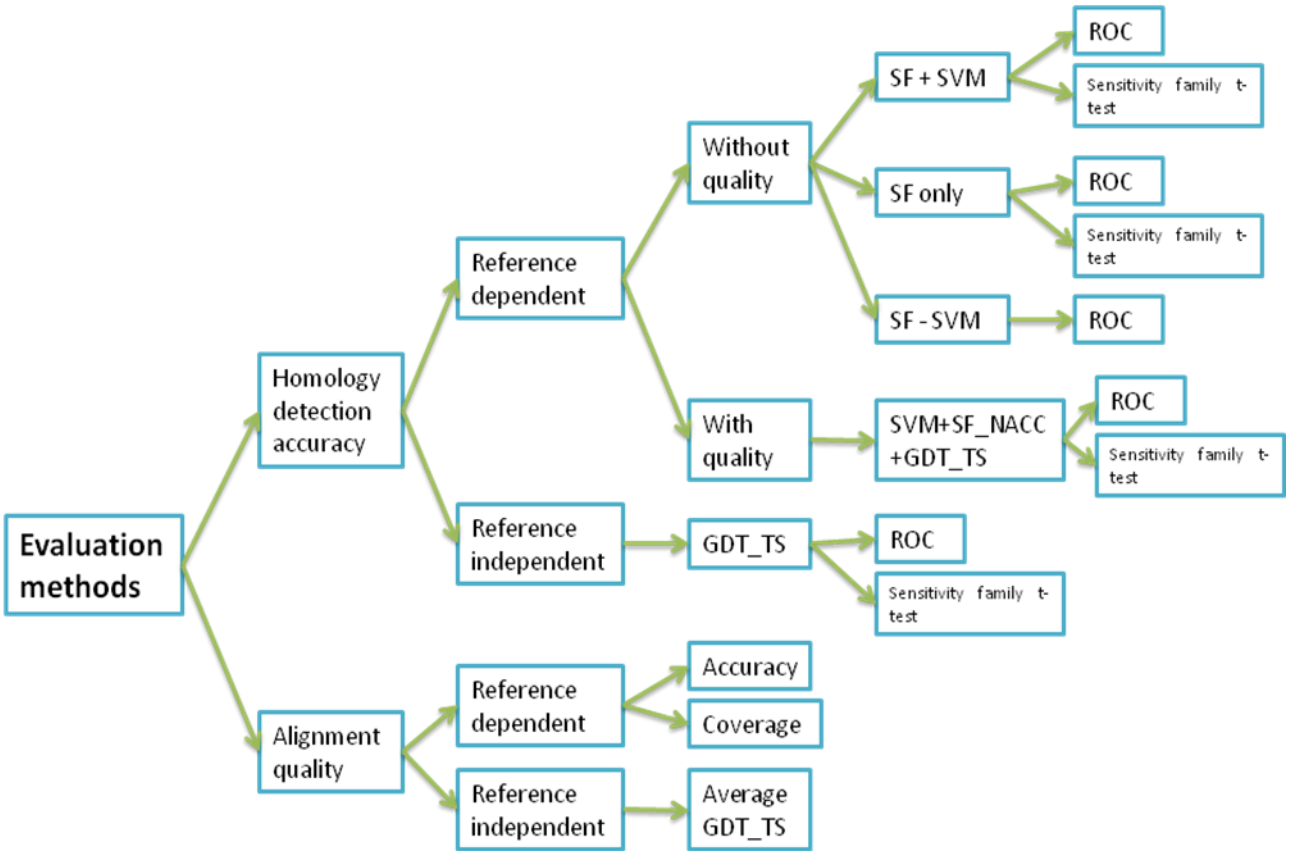

Figure 5 the Classification Tree of Evaluation Methods Used

2.4.1 Protein homology detection sensitivity evaluation

I used ROC (Receiver Operating Characteristic) curve to visualize the homology detection performance of the algorithm. This is because ROC provides a visual as well as numerical summary of an algorithm's behavior and it is the predominant method in bioinformatics applications (Sonego, Kocsor et al. 2008).

Since differentiating homologues by whether they belong to the same SCOP superfamily or differentiating non-homologues by whether they belong to different SCOP class are very crude methods, so Yuan et al (Qi, Sadreyev et al. 2007) ran all-to-all structural and sequence alignment of the whole SCOP database with methods like Dali and HHsearch, then feed the alignment results to SVM (Support Vector Machine) (Byvatov and Schneider 2003) to calculate SVM scores, then use Superfamily relationship and/or SVM score together to decide whether a protein pair is homologous with each other. I will use the SCOP superfamily relationship and the SVM scores together to evaluation the performance of ProCAIn.

2.4.1.1 Reference dependent evaluation with SCOP superfamily relationship only

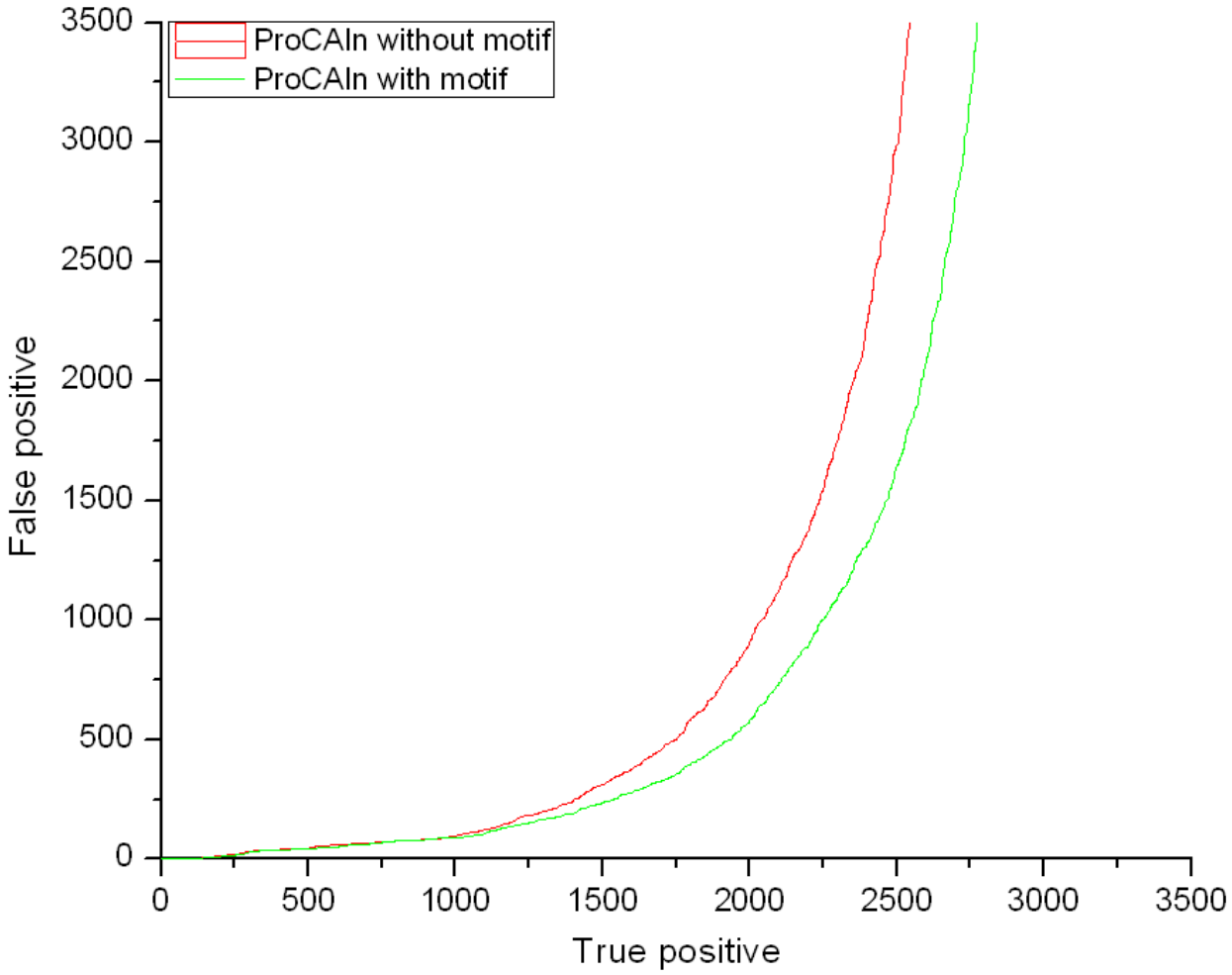

Figure 6 the Result of Reference Dependent Evaluation with SCOP Superfamily Relationship Only

For this evaluation method, protein pairs in the same SCOP superfamily are counted as true positives and all other protein pairs are viewed as false positive. Protein pairs which are in the same SCOP superfamily are very close homologues. The following ROC curve shows that adding motif information helps ProCAIn differentiate close homologues from remote homologues or non-homologues. This is consistent with the observation that the alignments of close

homologues tend to have much more regions with consecutive positive sequence similarity scores. When adding part of the scores of the previous and next position of these regions to the current position scores, close homologues are rewarded more than remote homologues and non-homologues. This gives close homologue a bigger optimal alignment score, and hence a more significant e-value after statistical analysis.

2.4.1.2 Reference dependent evaluation with SCOP superfamily relationship and SVM score

For this evaluation method, proteins pairs in the same SCOP superfamily (close homologues) or proteins pairs with a SVM score larger than or equal to 0.6 (remote homologues) are counted as true positive. Proteins pairs not in the same SCOP superfamily and with a SVM score between -0.6 and 0.6 are seen as uncertain proteins and are discarded from the evaluation. All other protein pairs are viewed as false positive. Close homologues are proteins which share significant evolution relationship and significant structural similarity. Remote homologues are proteins which are proven to have significant structural similarity.

Combined with results 2.4.1.1, the following ROC curve proves that adding motif information can help differentiate close homologues and remote homologues from non-homologues. This is extremely important because a lot of proteins, such as free model proteins, don't have many close homologues. In order to predict the structures of these proteins, it is very critical to detect remote homologues for these proteins.

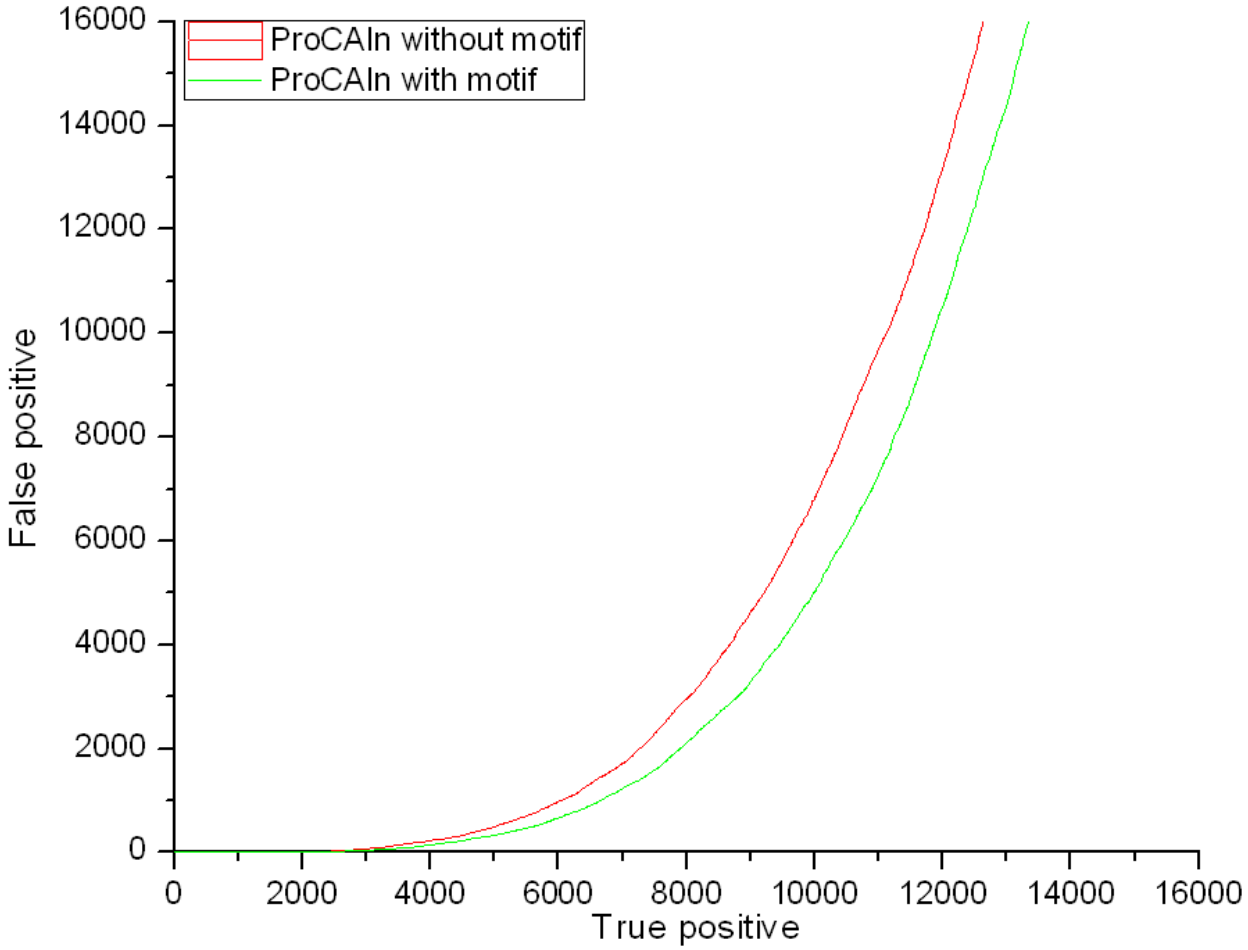

Figure 7 the Results of Reference Dependent Evaluation with SCOP Superfamily Relationship and SVM Score

2.4.1.3 Reference dependent evaluation with SCOP superfamily relationship, SVM score and alignment quality

The difference between this evaluation method and 2.4.1.2 is the true positives in 2.4.1.2 will be further tested whether ProCAIN can produce a good alignment. In order to be considered as true positive by this evaluation method, the protein pairs have to be either in the same SCOP superfamily or have a SVM score larger than or equal to 0.6, and at the same time have a alignment with a NACC (number of correctly aligned positions) (Sadreyev and Grishin 2003)

larger than or equal to 5, or a GDT_TS (global distance test total score) (Zemla 2003) larger than or equal to 0.15.

From 2.4.1.1 to 2.4.1.2 to 2.4.1.3, the evaluation methods are getting tougher and tougher. Evaluation 2.4.1.1 checks whether ProCAIn with motif information can differentiate close homologues better, evaluation 2.4.1.2 checks whether it can differentiate close and remote homologues better, and then evaluation 2.4.1.2 further checks whether it can produce a better alignment at the same time.

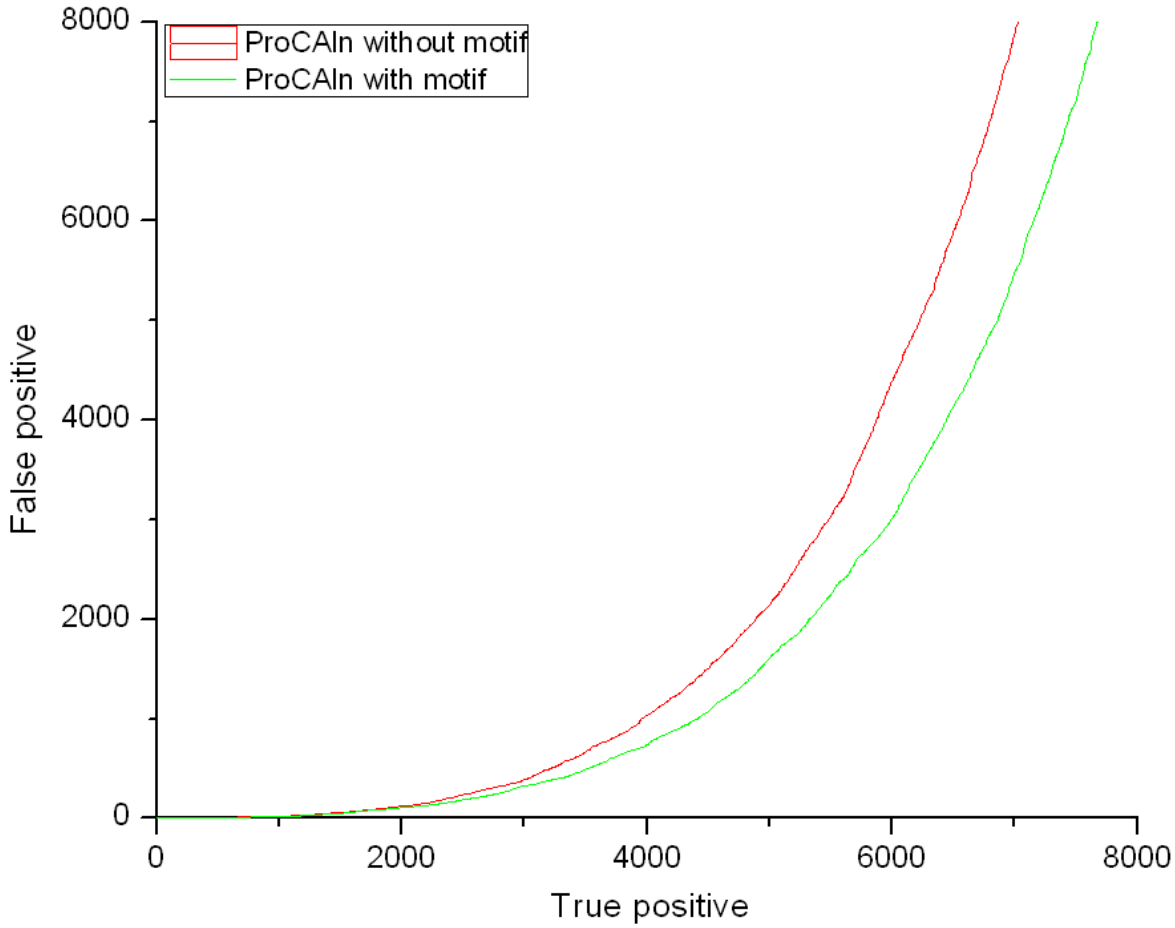

Figure 8 Reference dependent evaluation with SCOP superfamily relationship, SVM score and alignment quality

The result clearly shows that ProCAIn with motif information can also produce better alignments than ProCAIn without motif information.

2.4.1.4 Reference independent global evaluation with GDT_TS

Evaluation methods 2.4.1.1, 2.4.1.2 and 2.4.1.3 all require SCOP database superfamily relationship as a reference to judge whether a protein pair is homologue or not. But it is very common that some users may want to search a query protein against a protein structure database which doesn't have clear superfamily definition. This is why I also tested ProCAIn with a reference independent evaluation method.

This evaluation method doesn't depend on the SCOP superfamily relationship to decide whether a tested protein pair is homologous or not. In this method, a protein pair which has a global GDT_TS larger or equal to 0.15 is counted as true positive, and false positive otherwise. The global GDT_TS is calculated using the following equation.

$$GDT\_TS = \frac{n_1 + n_2 + n_4 + n_8}{4} / len_{query}$$

n_1, n_2, n_4, n_8 are number of aligned residues within 1, 2, 4, 8 angstroms, respectively (Zemla 2003). len_{query} is the sequence length of the query protein. GDT_TS is an inter-molecule structure scores. The structures of a pair of proteins are super-imposed with each other according to their sequence alignment and the distance between corresponding residues is measured.

The result shows ProCAIn with motif information performs only slightly better than ProCAIn without motif information. The reason is ProCAIn with motif information gives longer alignment (shown below by average coverage), and the GDT_TS calculation method penalize longer alignment a lot when it tries to superimpose the alignment. In spite of this, because ProCAIn with motif information gives better performance with other evaluation methods, it still proved that adding motif information is very helpful.

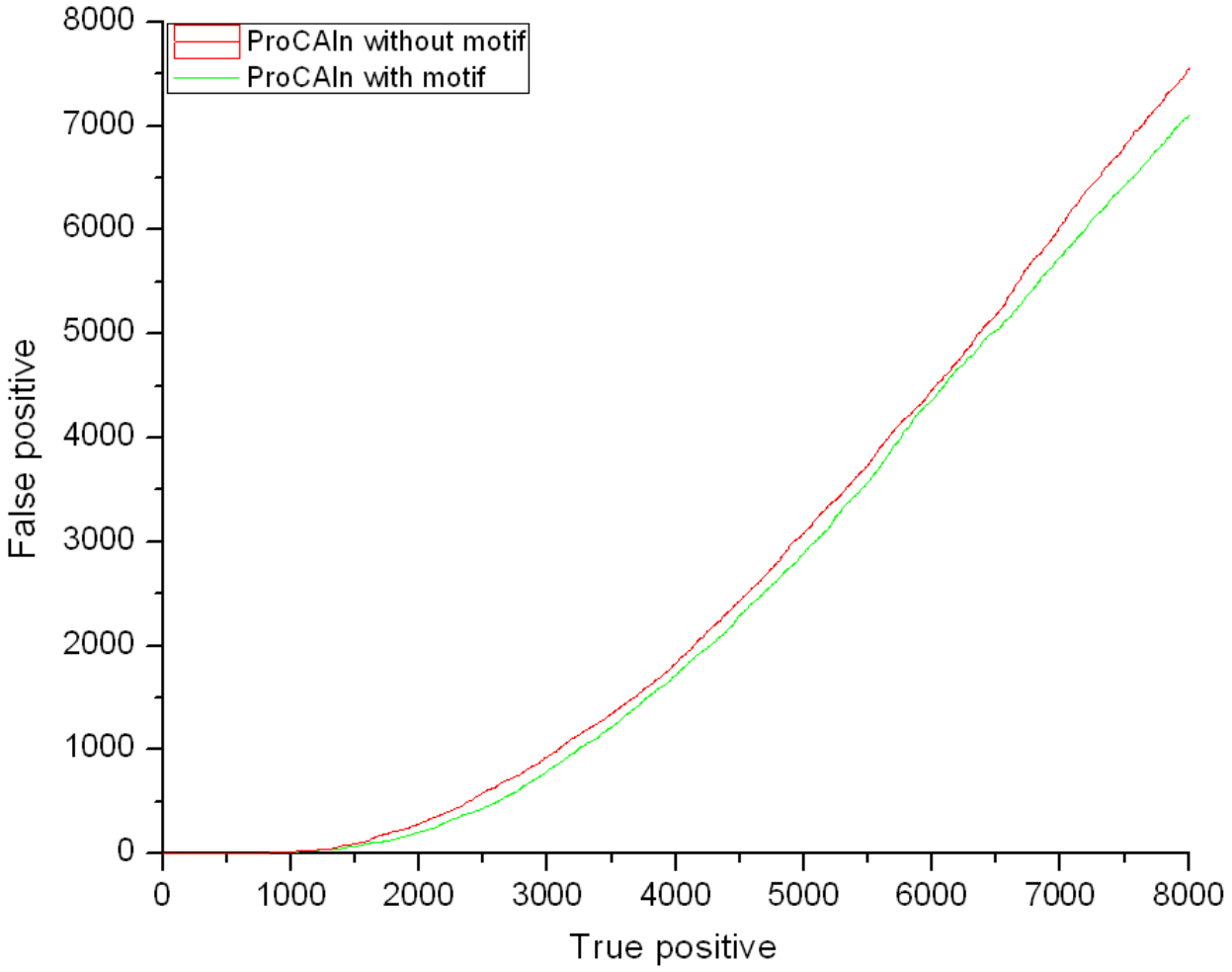

Figure 9 Reference independent global evaluation with GDT_TS

2.4.1.5 10%, 25% and 50% sensitivity family t-test

ROC curve is good at visualizing the homology detection performance when all-to-all comparison is conducted for the whole testing database. However, users in reality rarely do all-to-all comparison. Normally users compare a query protein against the whole database, so it is very important to test the performance of ProCAIn under this kind of circumstance. This sensitivity family t-test method conducts this test.

Corresponding sensitivity (10%, 25%, and 50%) for each query proteins is calculated, and then pairwise student t-test is performed between ProCAIn with or without motif information. The p values are shown in the following table. Negative p value means the left method (ProCAIn without motif) is worse than the right method (ProCAIn with motif). The first number in each row is the p value for evaluation method 2.4.1.1. The second number is for 2.4.1.2, the third for 2.4.1.3 and the forth for 2.4.1.4.

	ProCAIn with motif		
	10%	**25%**	**50%**
ProCAIn without motif	-7.54e-01 -7.27e-01 -7.65e-01	-2.51e-03 -2.54e-02 -1.59e-02	-1.12e-03 -6.75e-03 -9.78e-05

	-2.72e-01	-6.68e-01	-1.91e-01

Table 1 10%, 25% and 50% sensitivity family t-test

The difference for 10% sensitivity is very trivial. This is because most proteins ranked there are very close homologues, so both methods did a good job differentiating these proteins. You can also see this from the ROC curves.

The difference for 25% and 50% sensitivity gets bigger and bigger. This is because the proteins are getting more and more diverse, so it is more and more difficult to differentiate without help from other assisting information. This is why ProCAIn with motif performs better and better.

2.4.2 Protein sequence alignment quality evaluation

There are two factors which affect the accuracy of a protein structure modeling attempt. The first factor is whether the correct homologue can be detected. The second factor is whether the alignment quality between the query protein and its homologue is good or not.

Results in 2.4.1 already proved that adding motif information can help improve protein homology detection. This section tests whether adding motif information can help improve alignment quality.

2.4.2.1 Accuracy

Dali structure alignment is used as gold standard here. The definition of accuracy is the ratio of the number of correctly aligned positions (NACC) to the length L of the region in the structural alignment that includes the pairs of profile positions from the alignment under evaluation

(Sadreyev and Grishin 2003). The results are clustered according to different sequence identities: 0~5%, 5~10%, 10~15% and 15~20%. This is because the protein pairs with less sequence identities are more difficult to align. Clustering the results can show how ProCAIn performs under different difficulty level.

The following graph shows that the accuracy improvement is quite trivial. This is because the alignments produced by ProCAIn with motif are much longer (shown in the average coverage result) and the extended parts of each alignment are much more diverse and hence more difficult to align. Same accuracy with bigger coverage means more positions are correctly aligned, which is an alignment quality improvement.

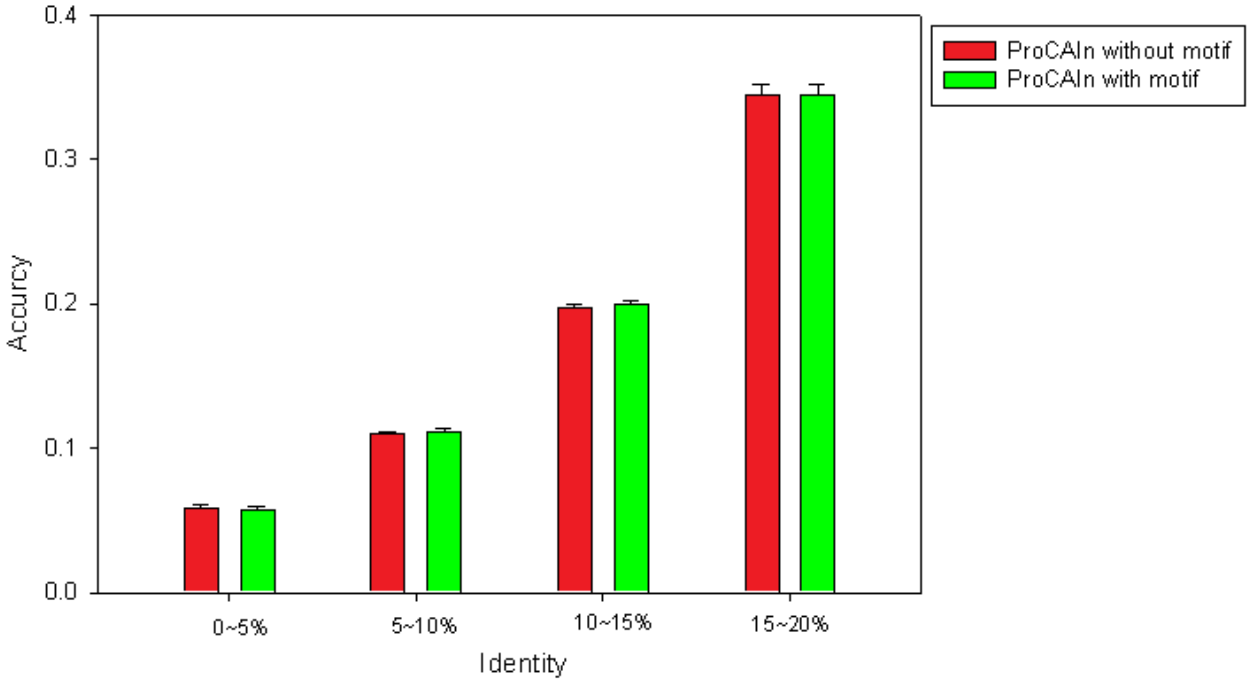

Figure 10 Accuracy of ProCAIn with and without Motif Information

2.4.2.2 Coverage

Dali structure alignment is again used as gold standard here. The definition of coverage is the ratio of the length L of the region in the structural alignment that includes all the positions from the evaluated alignment to the overall length of the structural alignment (Sadreyev and Grishin 2003).

The following graph shows ProCAIn with motif information averagely gives much longer alignment. Combining this result with the result in 2.4.2.1 shows adding motif information improved the alignment quality.

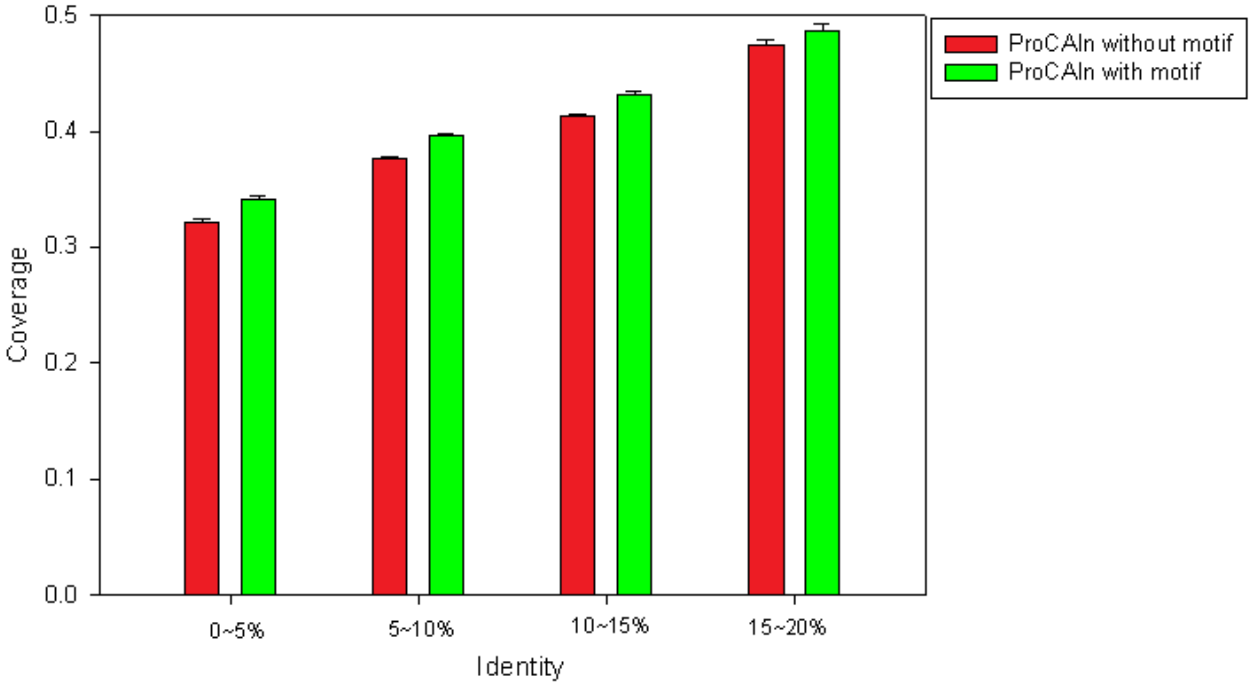

Figure 11 Coverage of ProCAIn with and without Motif Information

2.4.2.3 Average GDT_TS

Method 2.4.2.1 and 2.4.2.2 are reference dependent evaluations. They both used Dali structure alignments as reference to decide whether the sequence alignments produced by ProCAIn are

correct or not. Average GDT_TS is a reference independent evaluation and it is also kind of a mixture of accuracy and coverage. Better accuracy will give bigger GDT_TS. Bigger coverage will also give bigger GDT_TS.

The following graph shows ProCAIn with motif information performs better than ProCAIn without motif information, especially for protein pairs with lower sequence identity. This is very important because normally protein pairs with lower sequence identity (< 10%) are extremely difficult to align.

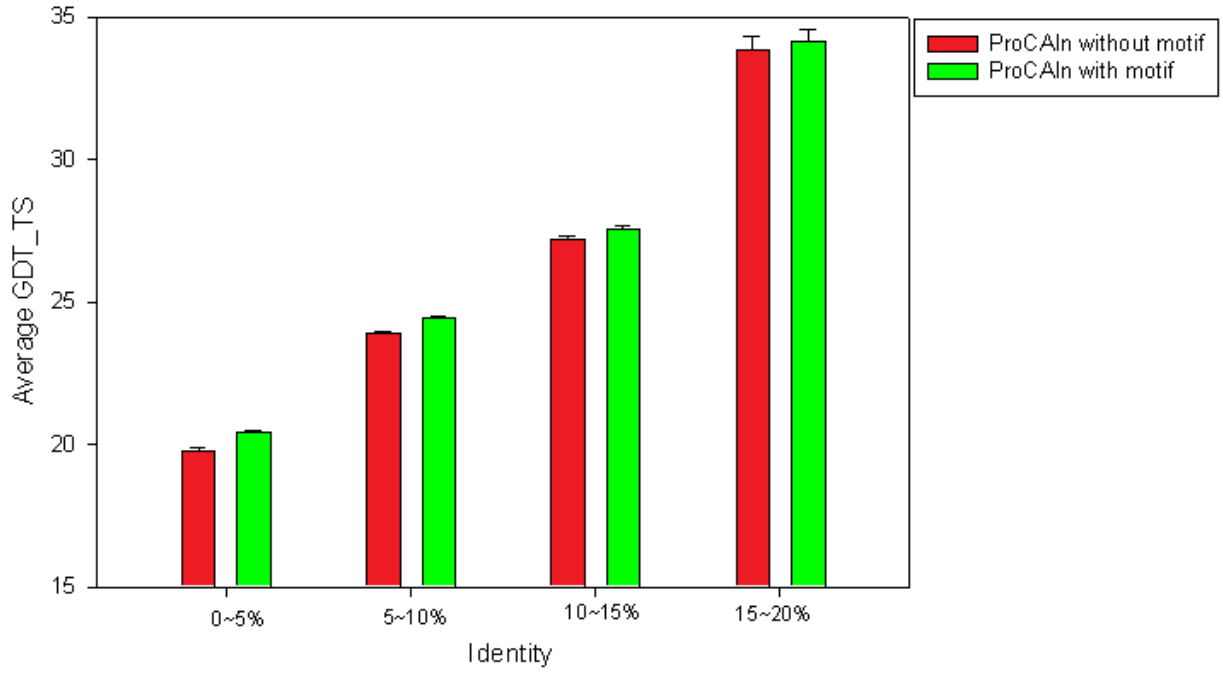

Figure 12 Average GDT_TS of ProCAIn with and without Motif Information

2.5 Conclusion

The results in 2.4 proved the hypotheses that adding motif information can improve ProCAIn with better protein homology detection and better alignment quality.

CHAPTER 3:

Adding Residue Conservation Score

3.1 Biological Observation

Homologous proteins usually share the same protein fold and possess related functions. These structural and functional constraints are reflected in the alignment conservation patterns. Positions of functional and/or structural importance tend to be more conserved (Sonnhammer and Durbin 1994). For the following example protein, the positions marked red are binding positions and also conserved positions.

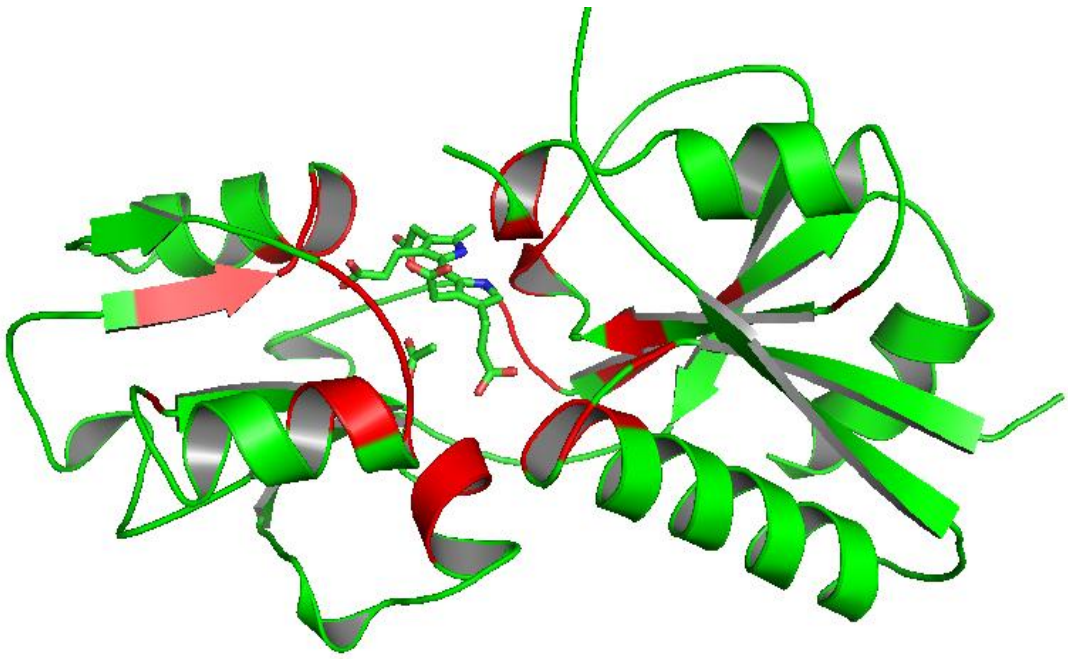

Figure 13 the Structure of an Example Protein with Conserved Regions Marked Red

The degree of sequence conservation at each position will be taken into account and matches at highly conserved positions will be rewarded more than matches at variable positions.

Figure 14 Adding Conservation Information

For the above alignment segment, "+" means the corresponding residue pairs are similar and "-" means the corresponding residue pairs are not similar. The "Conservation" rows are calculated residue conservation values. The red segments are conserved matches, which mean

the corresponding positions within these segments are not only matches but also highly conserved. These segments indicate important functional matches and should be rewarded. The blue segments are conserved mismatches, which mean the corresponding positions within these segments are highly conserved but are not similar. These segments indicate functional mismatches and should be punished.

3.2 Algorithm

We use the entropy method(Pei and Grishin 2001) to calculate residue conservation for columns 1 of the query profile and columns 2 of the subject profile and then normalized it to $0 \sim 1$. 0 means the position is not conserved and 1 means the position is highly conserved.

$$CR = \left(\sum_i f_i \ln(f_i) + 2.9958 \right) \Big/ 2.9958$$

Here f_i is the total residue frequency of columns 1 of the query profile and columns 2 of the subject profile. This conservation value is then combined with sequence similarity score by the following equation to get residue conservation score.

$$S^{conservation} = S^{seq} \times CR \times w^{conservation}$$

Here $w^{conservation}$ is the weight for conservation score. It is trained with the testing dataset and is also a constant for all query sequences.

For positions which are highly conserved, hence a big CR value, and also share sequence similarity, hence a positive S^{seq}, $S^{conservation}$ will be a big positive value. This means these positions are highly rewarded. For positions which are highly conserved, but don't share sequence similarity, hence a negative S^{seq}, $S^{conservation}$ will be a big negative value. This means these positions are highly punished.

This is consistent with the observation that highly conserved positions are normally functional positions, such as binding sites, so highly conserved sequence matches means function matches and should be rewarded and highly conserved sequence mismatches means function mismatches, hence should be punished(Durbin 1998).

3.3 Results

I also tested this idea with all evaluation methods available and all results are very consistent, so I will just show some main results in the following.

3.3.1 Protein homology detection sensitivity evaluation

3.3.1.1 Reference dependent evaluation with SCOP superfamily relationship and SVM score

For this evaluation method, proteins pairs in the same SCOP superfamily (close homologues) or proteins pairs with a SVM score larger than or equal to 0.6 (remote homologues) are counted as true positive. Proteins pairs not in the same SCOP superfamily and with a SVM score between -0.6 and 0.6 are seen as uncertain proteins and are discarded from the evaluation. All other protein pairs are viewed as false positive. Close homologues are proteins which share significant evolution relationship and significant structural similarity. Remote homologues are proteins which are proven to have significant structural similarity.

With this evaluation method, it is very clear that conservation score helps ProCAIn's performance a lot and the difference between ProCAIn with or without conservation score starts from the very beginning. The reason for this is because protein pairs in the top of the ranking are mostly close homologues; proteins which share not only structure similarity but

also functional similarity. Matching of conserved positions is a strong indication of functional matching. This is why conservation can help ProCAIn differentiate between close homologues from remote homologues.

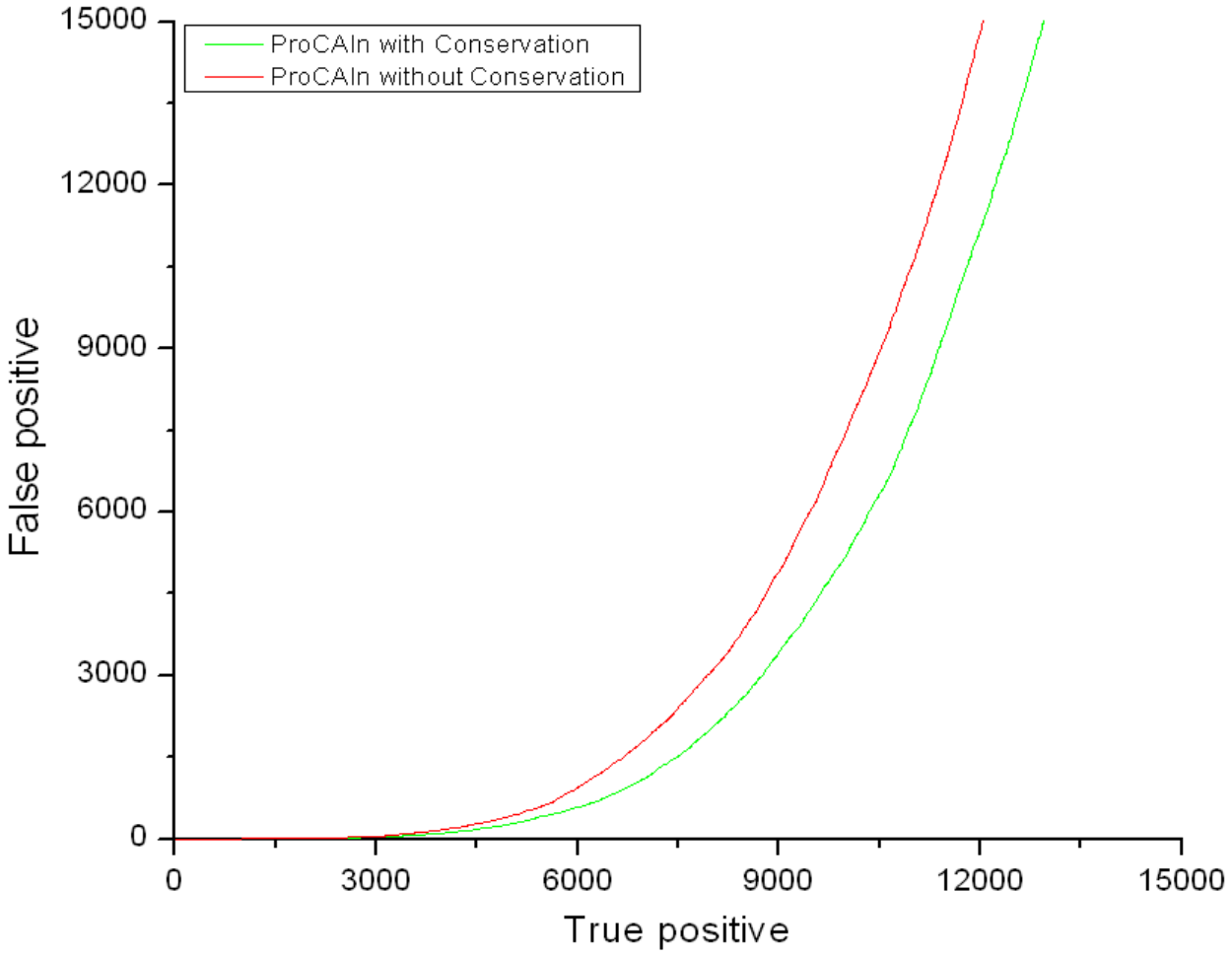

Figure 15 the Results of Reference dependent evaluation with SCOP superfamily relationship and SVM score

3.3.1.2 Reference dependent evaluation with SCOP superfamily relationship, SVM score and alignment quality

The difference between this evaluation method and 3.3.1.1 is the true positives in 3.3.1.1 will be further tested whether ProCAIN can produce a good alignment. In order to be considered as true positive by this evaluation method, the protein pairs have to be either in the same SCOP superfamily or have a SVM score larger than or equal to 0.6, and at the same time have a alignment with a NACC (number of correctly aligned positions) larger than or equal to 5, or a GDT_TS (global distance test total score) larger than or equal to 0.15.

With this evaluation method, ProCAIn with conservation still performs better than ProCAIn without conservation. But compared with the results of the last evaluation method, the difference between ProCAIn with or without conservation gets smaller. This might be because firstly the number of true positives gets smaller when the evaluation method is more restricted. The second reason could be because conservation doesn't help improve alignment quality that much since normally only a few positions is highly conserved for a protein.

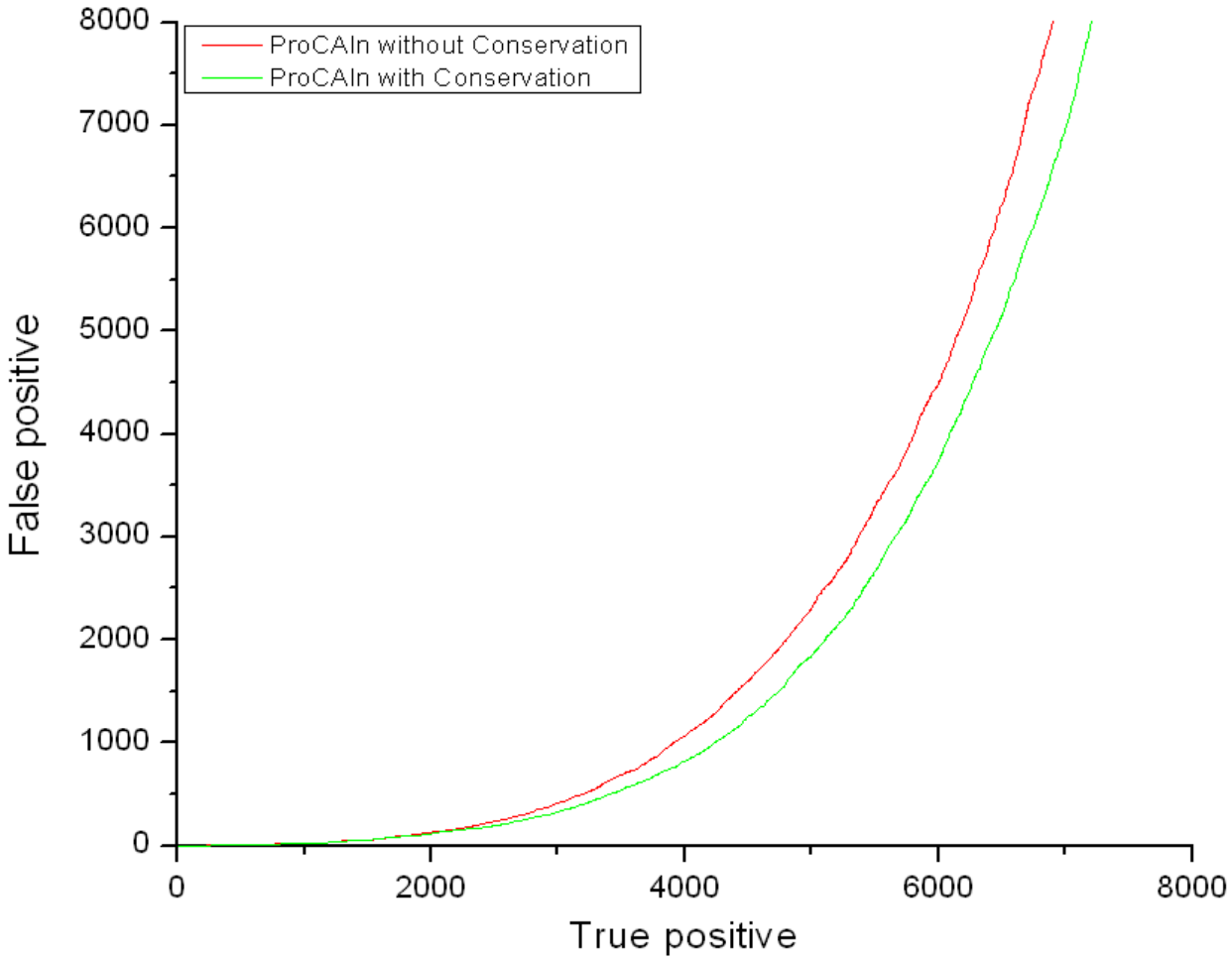

Figure 16 the results of reference dependent evaluation with SCOP superfamily relationship, SVM score and alignment quality

3.3.1.3 Reference independent global evaluation with GDT_TS

This evaluation method doesn't depend on the SCOP superfamily relationship to decide whether a tested protein pair is homologous or not. In this method, a protein pair which has a global GDT_TS larger or equal to 0.15 is counted as true positive, and false positive otherwise.

With this evaluation method, there is almost no difference between ProCAIn with or without conservation. The reason might be because, just like I explained in the above section, that only a few positions in a protein are highly conserved.

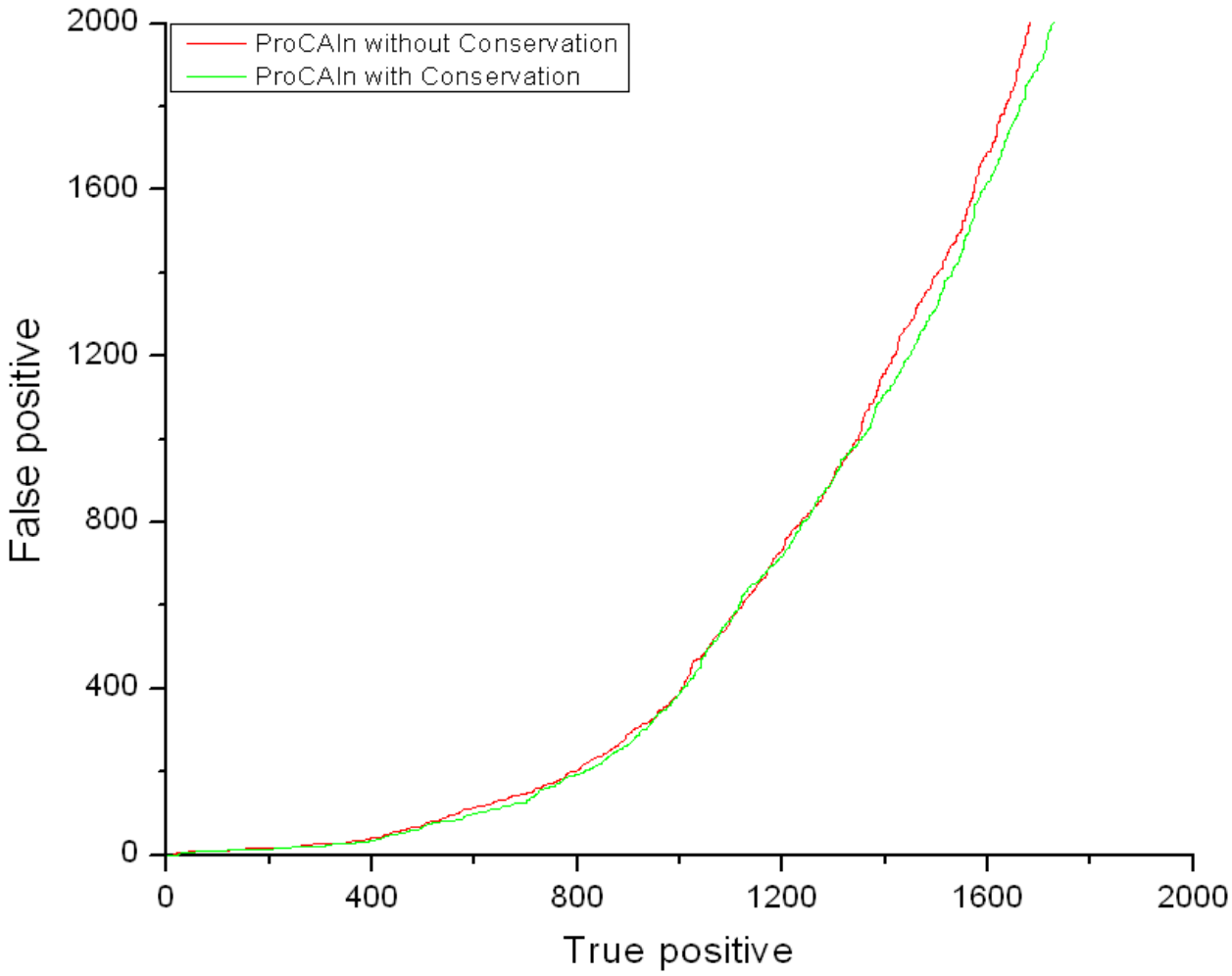

Figure 17 the results of reference independent global evaluation with GDT_TS

3.3.2 Protein sequence alignment quality evaluation

3.3.2.1 Accuracy

The following graph shows that conservation score improves accuracy for proteins with higher sequence identity (5%-20%) but decreases accuracy for protein with very low sequence identity (0-5%). The reason for the accuracy decreasing for proteins with very low sequence identity may be because very few positions are conserved for these proteins. So adding conservation scores will be similar with adding background noises and thus decreases alignment quality. However for proteins with highly conserved positions, adding conservation scores is clearly helpful.

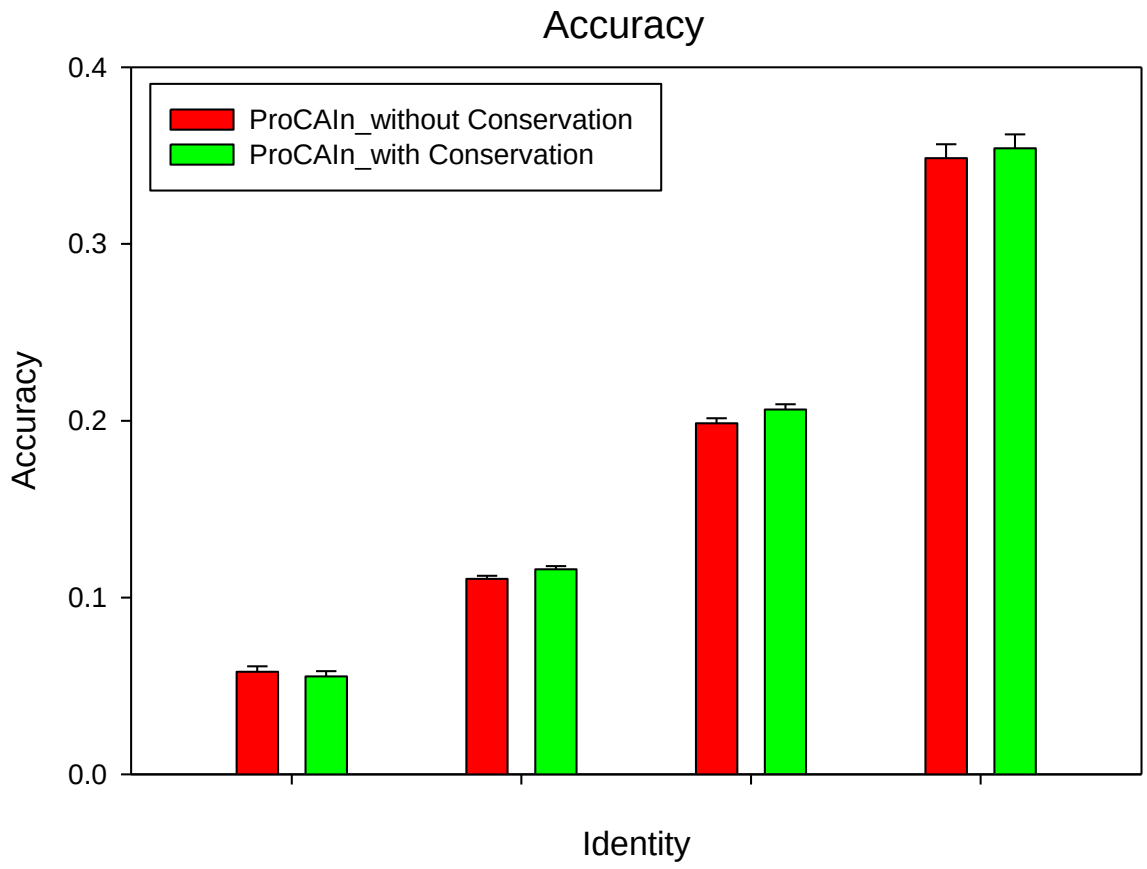

Figure 18 Accuracy of ProCAIn with and without Conservation Information

3.3.2.2 Accuracy

It is very clear that adding conservation scores can decrease coverage. This may be because adding conservation scores is similar with adding a constraint to aligning process and makes it more difficult to get long alignments.

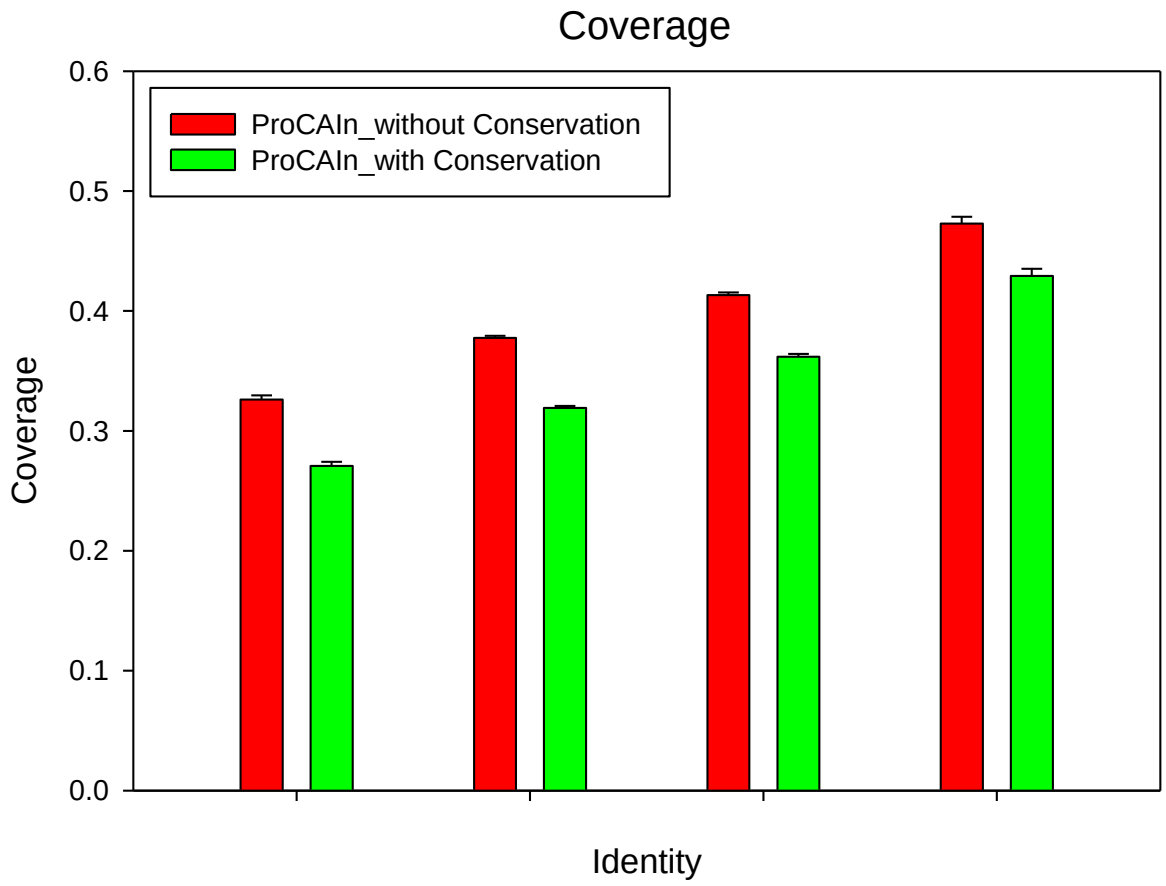

Figure 19 Coverage of ProCAIn with and without Conservation Information

3.3.2.3 Average GDT_TS

The following graph clear shows that adding conservation scores can improve the average GDT_TS value of ProCAIn sequence alignments. Since GDT_TS reflects both accuracy and coverage of a sequence alignment, it is a better parameter to reflect alignment quality of a homology detection method. The following result shows that ProCAIn with conservation can improvement alignment quality.

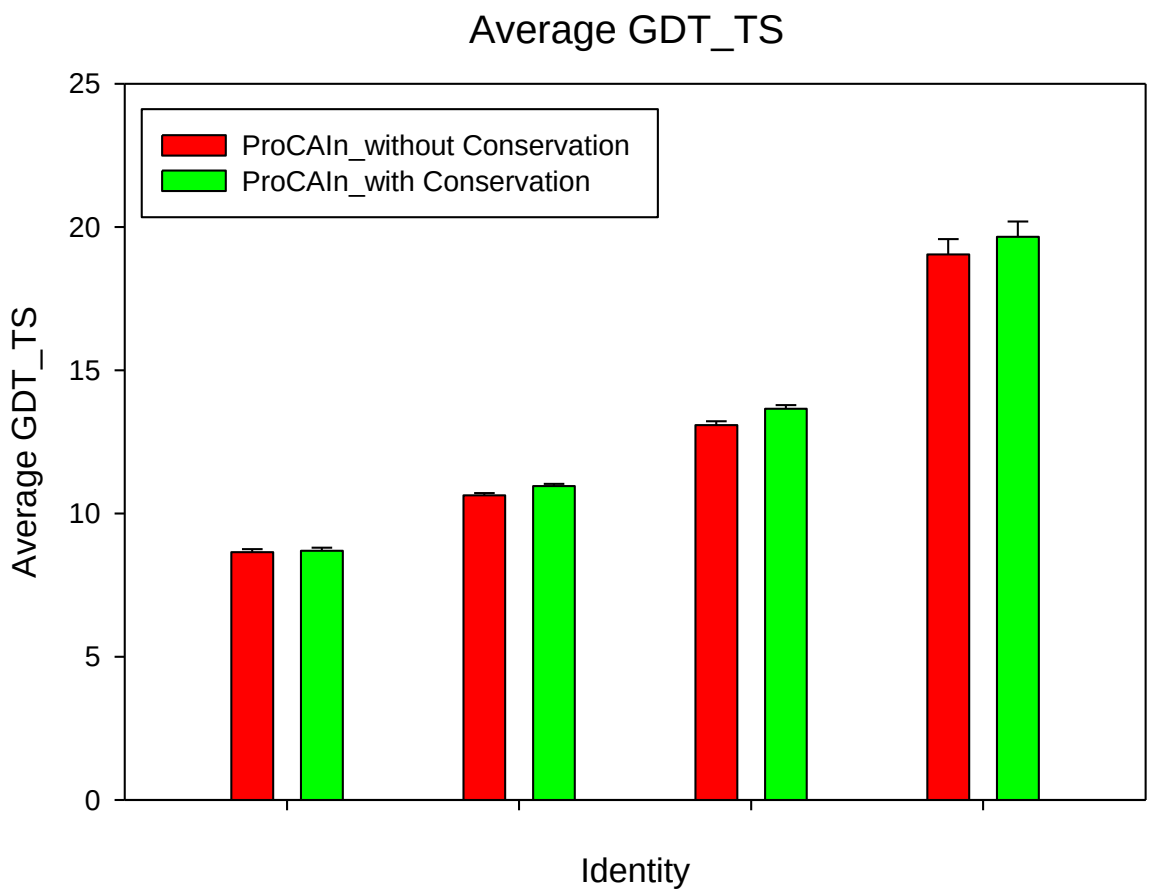

Figure 20 Average GDT_TS of ProCAIn with and without Conservation Information

3.4 Conclusion

The results of both homology detection sensitivity and alignment quality shown above prove that conservation score is helpful information and adding conservation information to ProCAIn can improve its performance.

CHAPTER 4:

Adding Secondary Structure Score

4.1 Biological Observation

Secondary structure is the general three-dimensional form of local segments of proteins. The most common secondary structures are alpha helices, beta sheets and coils. Because protein tertiary structure is more conserved than protein sequence during evolution, a pair of homologous protein tends to have very similar secondary structures, as shown by the following example. So adding predicated secondary structure information can help with homology detection. There are several homology detection algorithms (such as HHsearch (Soding 2005)) available which successfully exploit this idea.

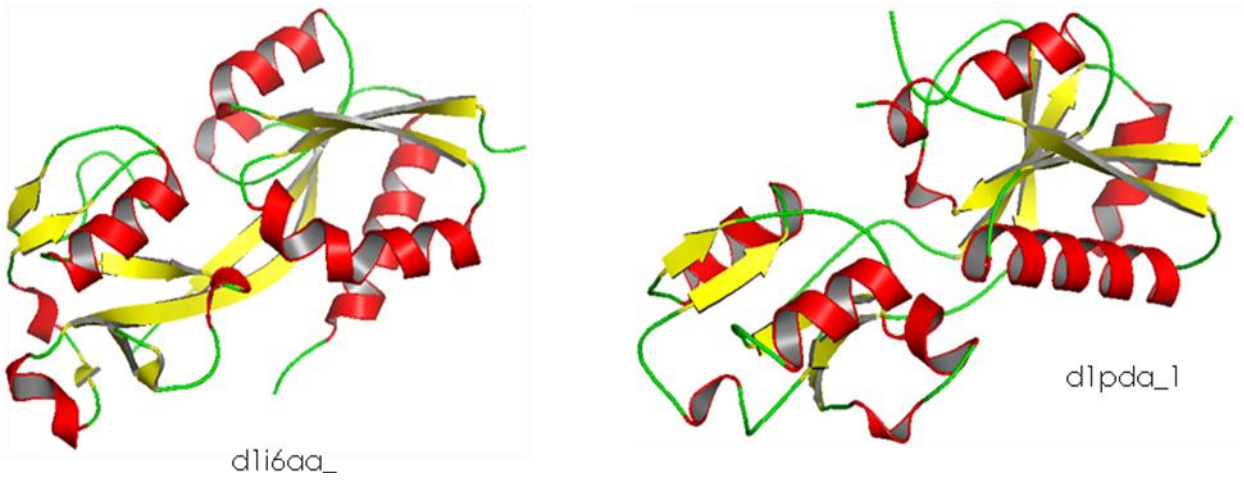

Figure 21 A homologous protein pair which shares very similar predicted secondary structure composition. Here alpha helices are red segments. Beta sheets are yellow segments and coils are green segments.

I here choose PsiPRED(McGuffin, Bryson et al. 2000) to predict secondary structures for my algorithm. PsiPRED is one of the most accurate secondary structure prediction methods at present. It used a two stage neural network and the input for PsiPRED is PSI_BLAST generated position specific scoring matrix (checkpoint file). This makes it very easy to use and fast. And the prediction accuracy of PsiPRED is about 80%. The following graph shows the flowchart of adding secondary structure information.

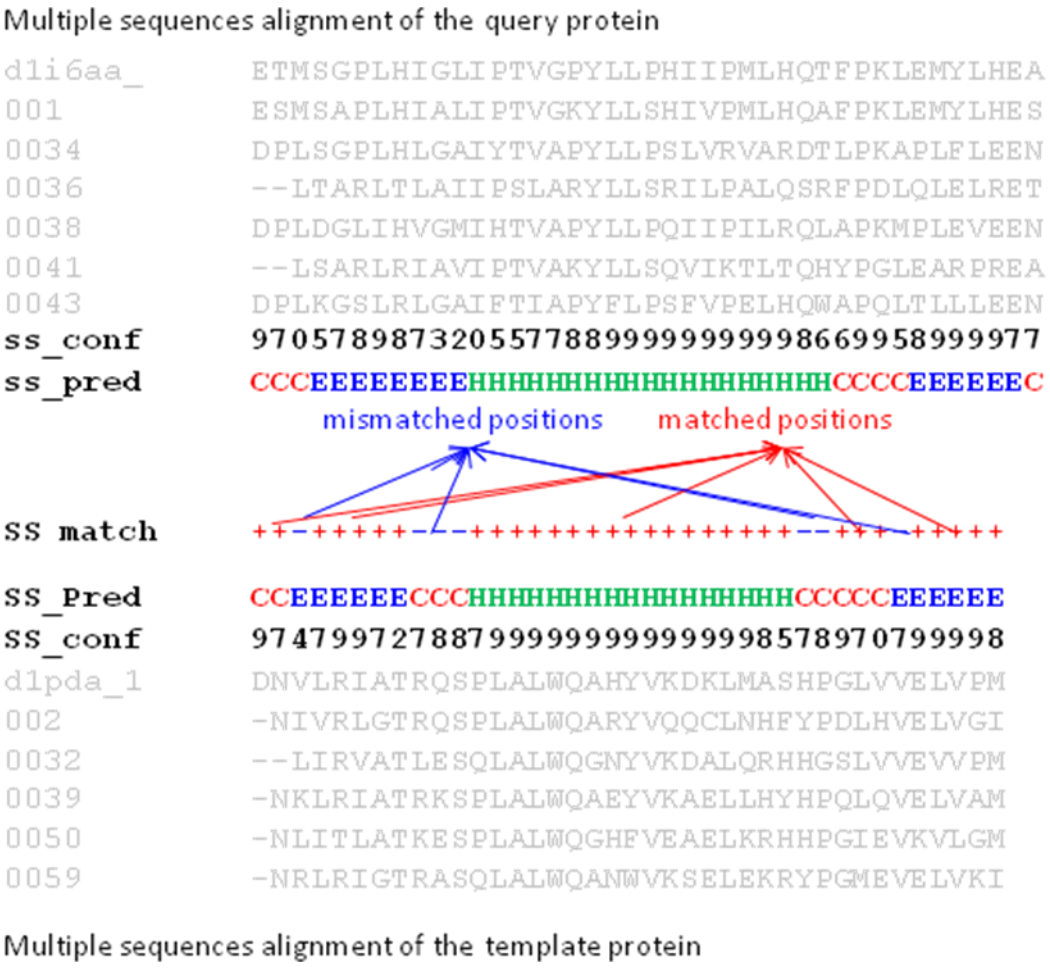

Figure 22 Adding Secondary Structure Information

In the above flowchart, “ss_pred” rows are predicted secondary structures for each multiple sequence alignments (MSAs). “E” here means alpha helix. “H” refers to beta sheet and “C” refers to coil. “ss_conf” rows are secondary structure prediction confidence values. Prediction confidence is from 0 to 9. “0” means no confidence with the prediction and “9” means highly confident. “SS” match row is the secondary structure matching result. “+” here means secondary structure match and “-“means secondary structure mismatch.

4.2 Algorithm

I will use the following equation to code predicted secondary structure information into my method. Again S^{seq} is the all position-to-all position sequence similarity scores between the query protein and the template protein, w^{ss} is the weight parameter. S_{mean} is the average of S^{seq}. SS_{matrix} is the 3 by 3 secondary structure substitution matrix we derived using SCOP structural alignments (shown in the following table), which represents the evolution frequency between each secondary structures. $SS_{num1}[i]$ is the secondary structure type (H,E or C) at column i of the query protein and $CD_{num1}[i]$ is its secondary structure prediction confidence level (from 0 to 9). $SS_{num2}[j]$ is the secondary structure type (H,E or C) at column j of the template protein and $CD_{num2}[j]$ is its confidence level.

$$S^{ss} = S_{mean} \times 0.01 \times w^{ss} \times SS_{matrix}[SS_{num1}[i]][SS_{num2}[j]] \times CD_{num1}[i] \times CD_{num2}[j]$$

	H	E	C
H	0.932	-2.147	-1.186
E	-2.147	1.544	-0.489
C	-1.186	-0.489	0.852

Table 2 Secondary Structure Substitution Matrix

This equation incorporates secondary structure confidence value (CD) into the algorithm. Secondary structure matches (SS_{matrix} > 0) with high confidence level will be rewarded more than secondary structure matches with low confidence level. And secondary structure mismatches (SS_{matrix} < 0) with high confidence level will be penalized more than secondary structure mismatches with low confidence level.

4.3 Results

I also tested this idea with all evaluation methods available and all results are very consistent, so I will just show some main results in the following.

4.3.1 Protein homology detection sensitivity evaluation

4.3.1.1 Reference dependent evaluation with SCOP superfamily relationship and SVM score

With this evaluation method, it is very clear that secondary structure score helps ProCAIn's performance a lot and the difference between ProCAIn with or without secondary structure score doesn't start from the very beginning. Compared with motif score or conservation score,

secondary structure score brings the most improvement to homology detection sensitivity. However, since secondary structure evolution lags behind sequence evolution, so both protein close homologues and remote homologues share significant secondary structure similarity. This is why secondary structure is not very helpful with differentiating close homologues from remote homologues. Thus the performance difference between ProCAIn with or without secondary structure doesn't start from the very beginning of the ROC.

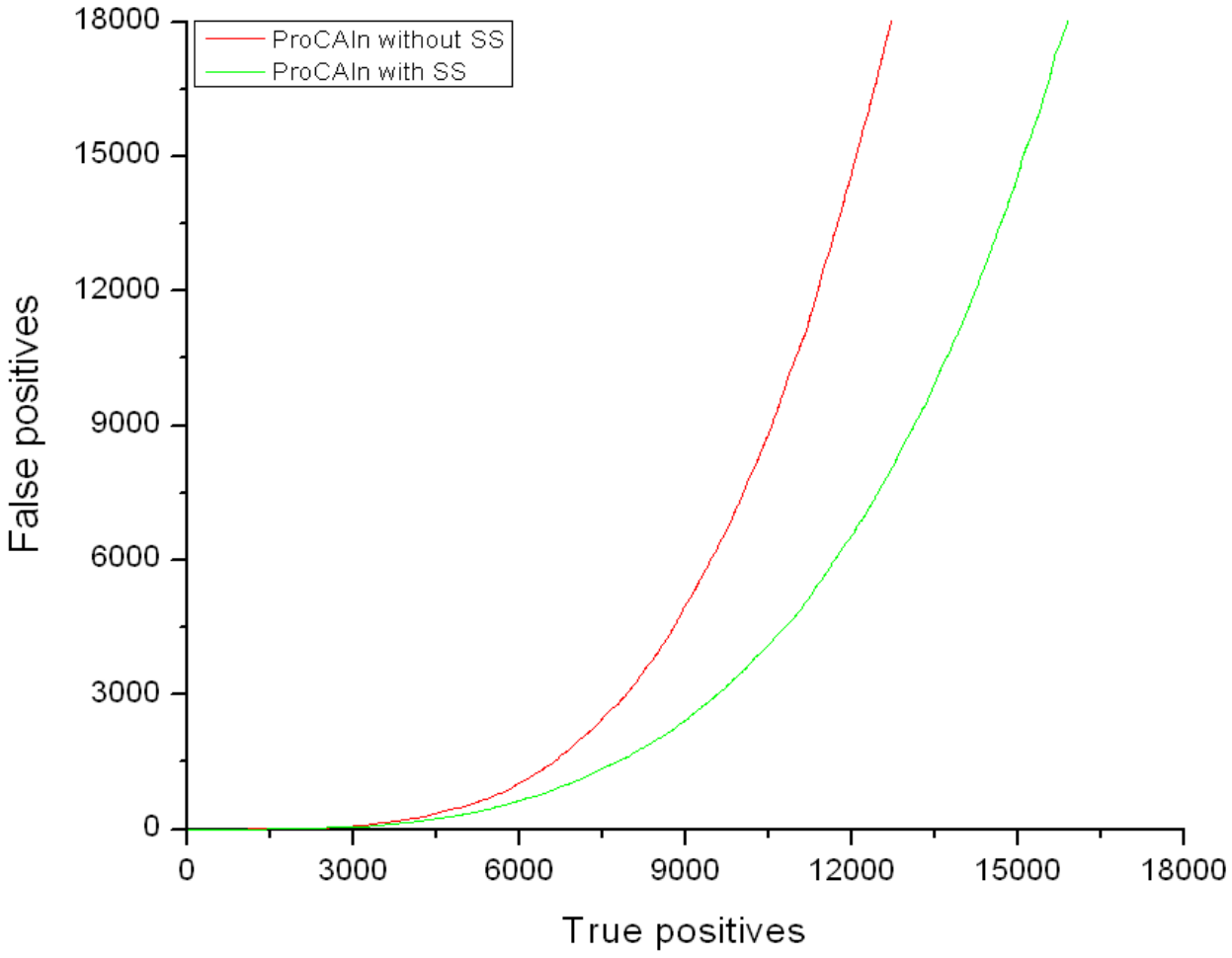

Figure 23 the result of reference dependent evaluation with SCOP superfamily relationship and SVM score

4.3.1.2 Reference dependent evaluation with SCOP superfamily relationship, SVM score and alignment quality

With this evaluation method, secondary structure can also clearly improve ProCAIn's performance and the difference between ProCAIn with or without secondary structure doesn't start from the very beginning either, for same reasons stated above.

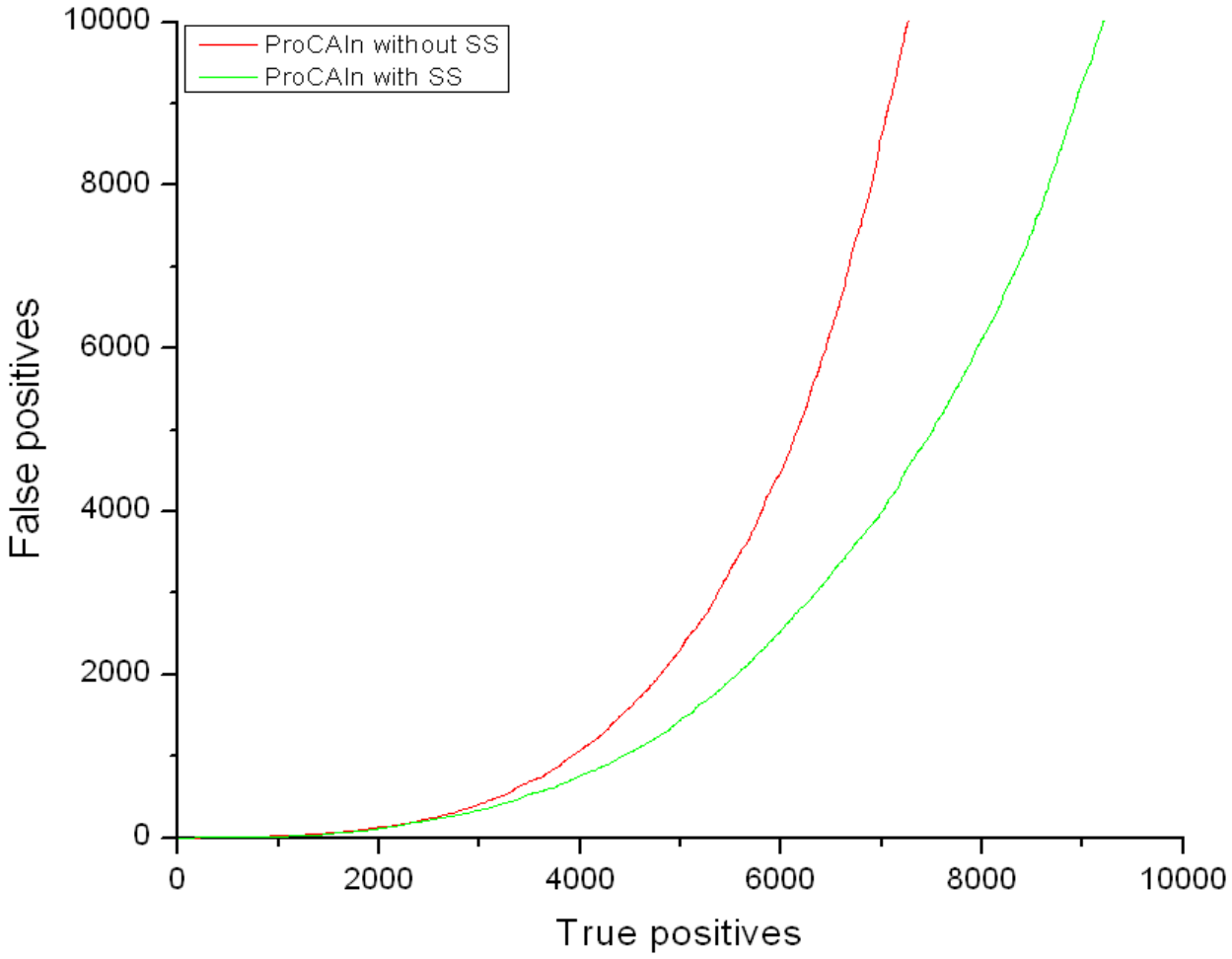

Figure 24 the result of reference dependent evaluation with SCOP superfamily relationship, SVM score and alignment quality

4.3.1.3 Reference independent global evaluation with GDT_TS

With this evaluation, it is also very clear that secondary structure helps a lot. Compared with motif score or conservation score, the difference here between ProCAIn with or without secondary structure is huge. So combined with the results of the last two evaluation methods, it is safe to conclude that among the three types of assisting information, secondary structure is the most significant one to help improve ProCAIn's homology detection sensitivity.

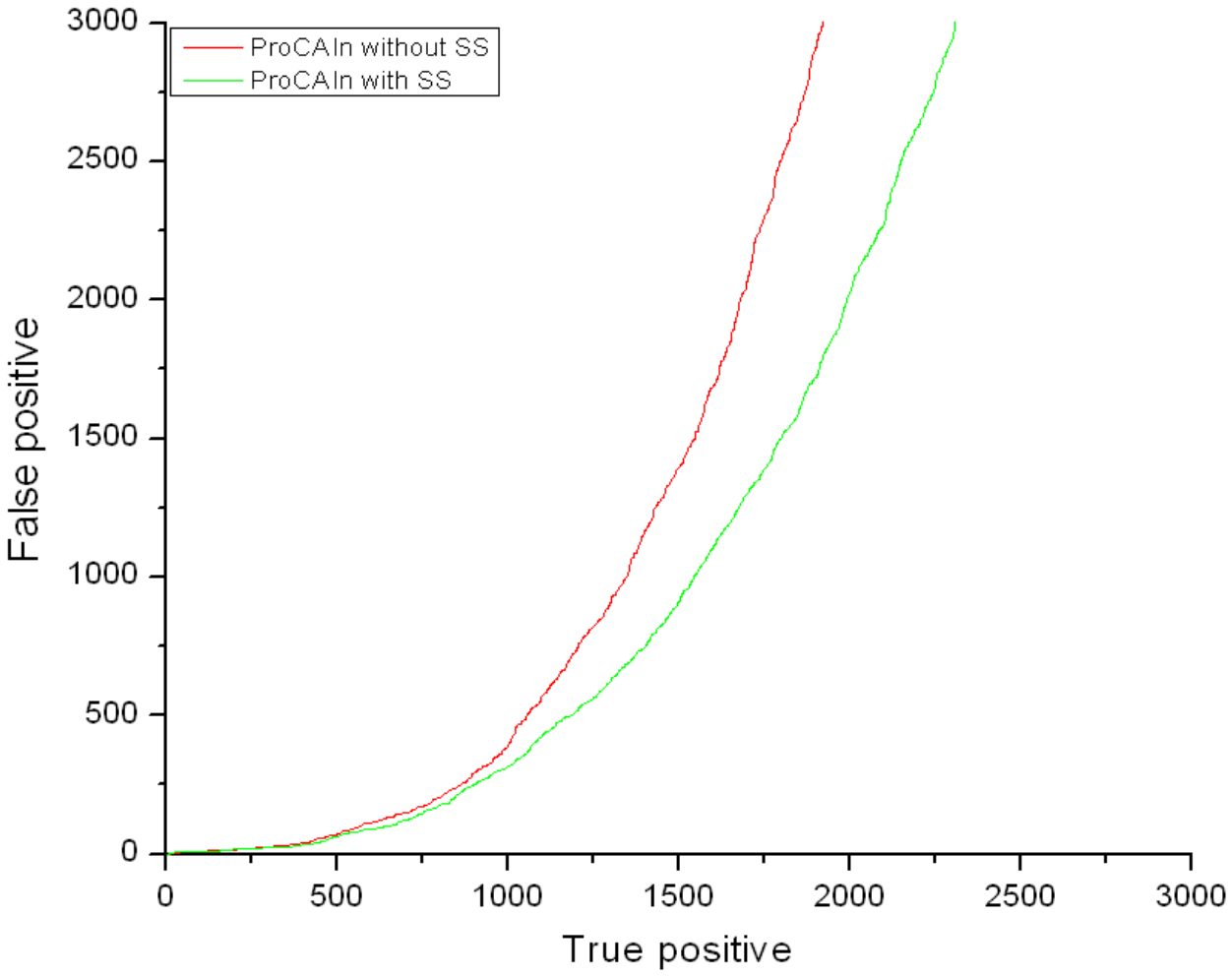

Figure 25 the result of reference independent global evaluation with GDT_TS

4.3.2 Protein sequence alignment quality evaluation

4.3.2.1 Accuracy

The following graph shows that secondary structure score improves accuracy for proteins with sequence identity from 0%-10% but decreases accuracy for protein with very high sequence identity (15-20%). The reason for the accuracy decreasing for proteins with very high sequence identity may be because the sequence alignments for these proteins are already very long, adding secondary structure score makes the alignments even longer (this can be seen from the coverage results in the following pages). Adding secondary structure improves the accuracy and coverage for proteins with sequence identity 0~15%, this definitely shows that adding secondary structure scores improves the alignment quality for this these proteins.

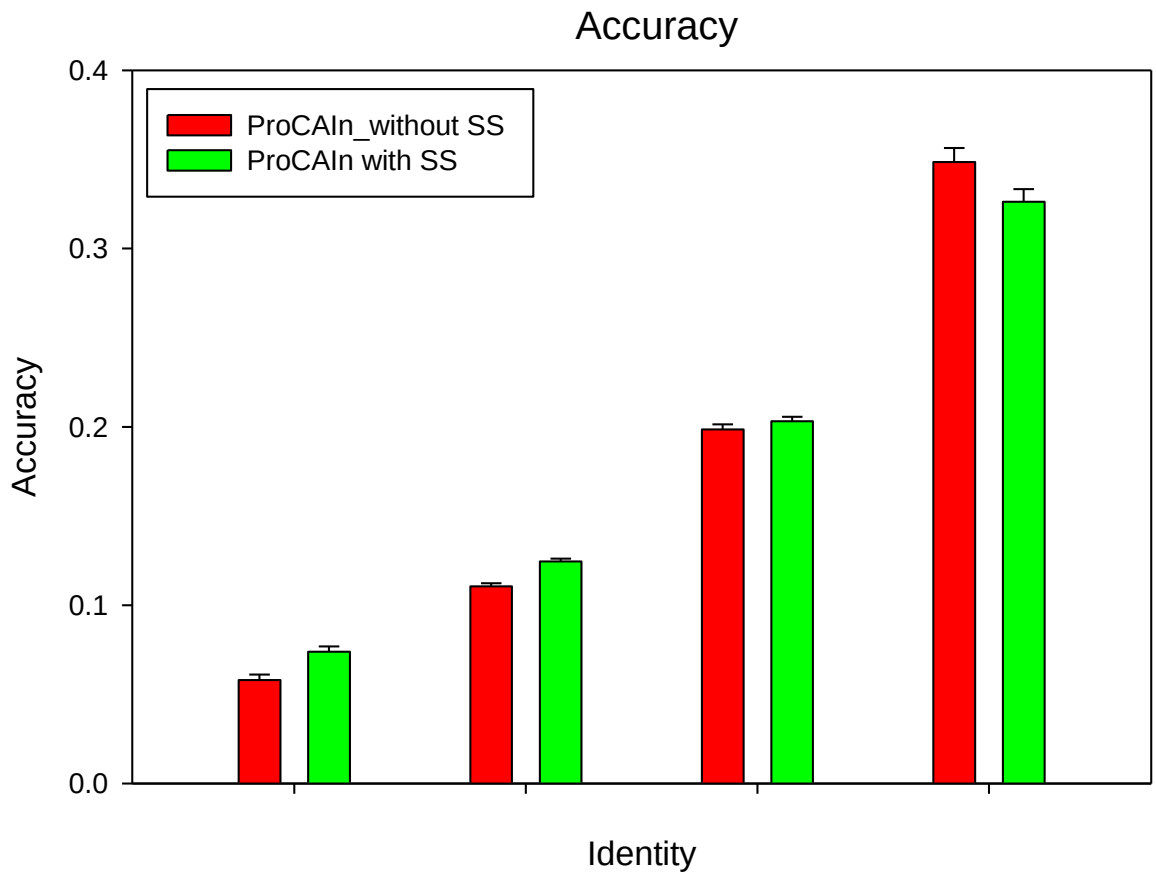

Figure 26 Accuracy of ProCAIn with and without Secondary Structure

4.3.2.2 Coverage

The following graph shows that adding secondary structure scores greatly increases sequence alignment coverage for proteins with all sequence identities. This result is not surprising. Even remote homologues have high secondary structure matches, so adding secondary structure scores make it easier to get long alignments, hence bigger coverage.

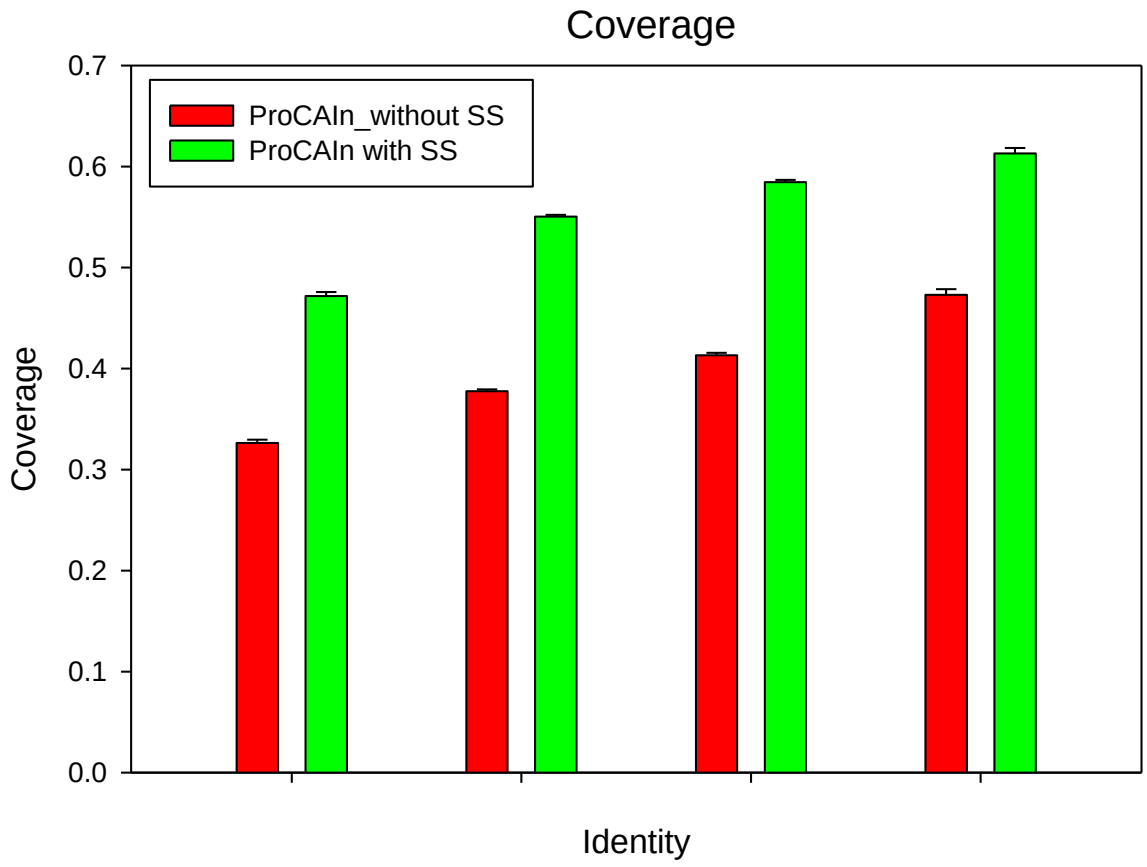

Figure 27 Coverage of ProCAIn with and without Secondary Structure

4.3.2.3 Average GDT_TS

The following graph shows the average GDT_TS of sequence alignments produced by ProCAIn with or without secondary structure scores. ProCAIn with secondary structure provides sequence alignments with higher average GDT_TS values, for proteins with all different sequence identity levels. Compared this result with accuracy and coverage results, it is clear to see that secondary structure improves sequence alignment quality. And among the three types of assisting information, sequence motif, residue conservation and secondary structure, secondary structure is the most significant information to improve ProCAIn's performance with alignment quality.

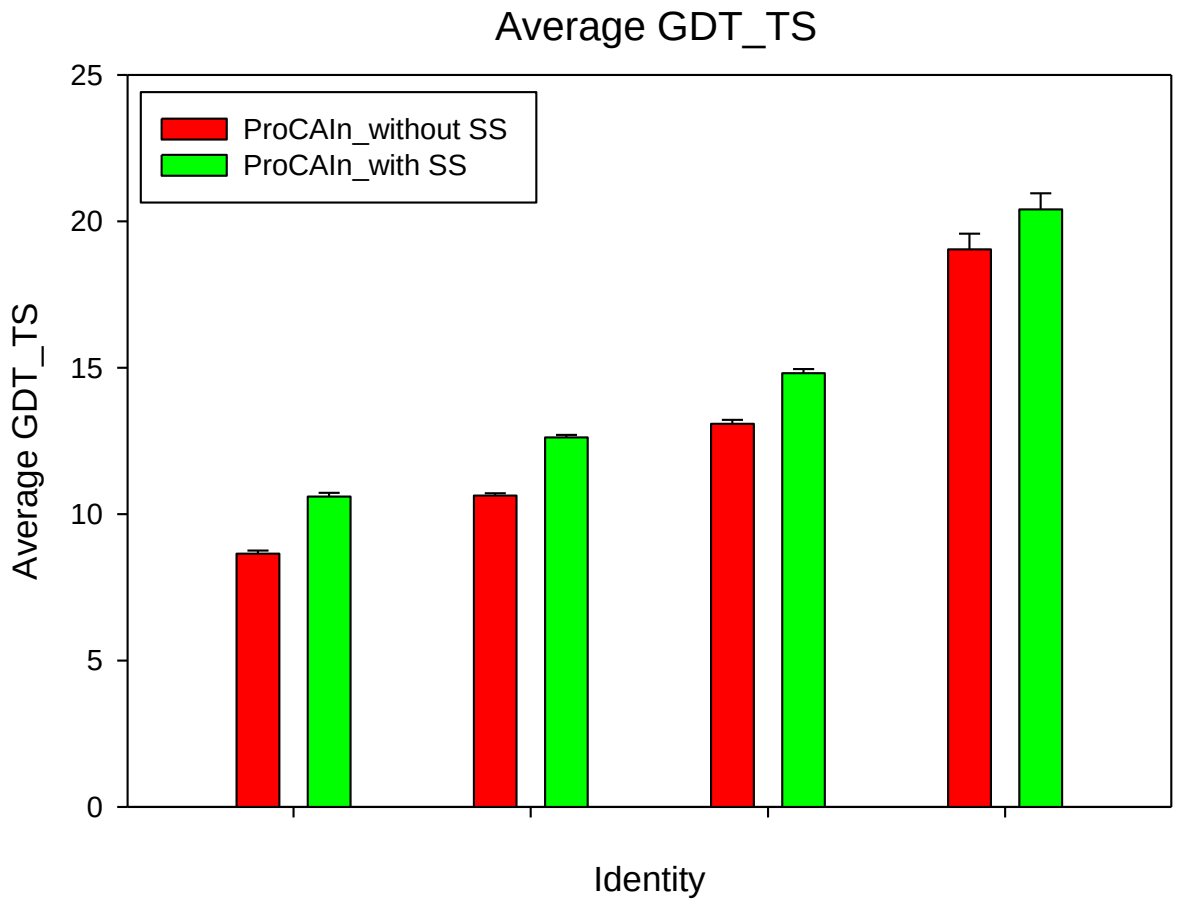

Figure 28 Average GDT_TS of ProCAIn with and without Secondary Structure

4.4 Conclusion

The above results of homology detection sensitivity and alignment quality clearly demonstrate that secondary structure is a very helpful type of assisting information and adding secondary structure scores greatly improves ProCAIn's performance.

The last three chapters introduce three types of assisting information: sequence motif, residue conservation and secondary structure. Results from each chapter demonstrate that incorporating these three types of assisting information with ProCAIn improves ProCAIn's performance with homology detection sensitivity and alignment quality. In the following

chapters, I will explore the relations between these three types on information and test whether combining these three types of information together with ProCAIn can further improve ProCAIn's performance.

CHAPTER 5:

ProCAIn with Three Types of Assisting Information

5.1 Correlation Between Assisting Information and Sequence Similarity Score

I extract three types of assisting information from the same multiple sequence alignment (MSA). Since protein sequence similarity score is also derived from the same MSA, it is reasonable to wonder the relation between these three types of assisting information and sequence similarity score. Are they the same thing or are they totally unrelated?

The results of the last three chapters demonstrate that adding these three types of information helps improve ProCAIn's performance with protein homology detection and sequence alignment quality, so it is unlikely that they are the same thing. I further calculated the correlation coefficient between these three types of assisting scores and protein sequence similarity scores. The next graph shows the correlation between sequence similarity scores and sequence motif matching scores. The Pearson correlation coefficient for this pair of scores is 0.8732. So the correlation between sequence similarity scores and sequence motif matching scores is high but they are not exactly the same.

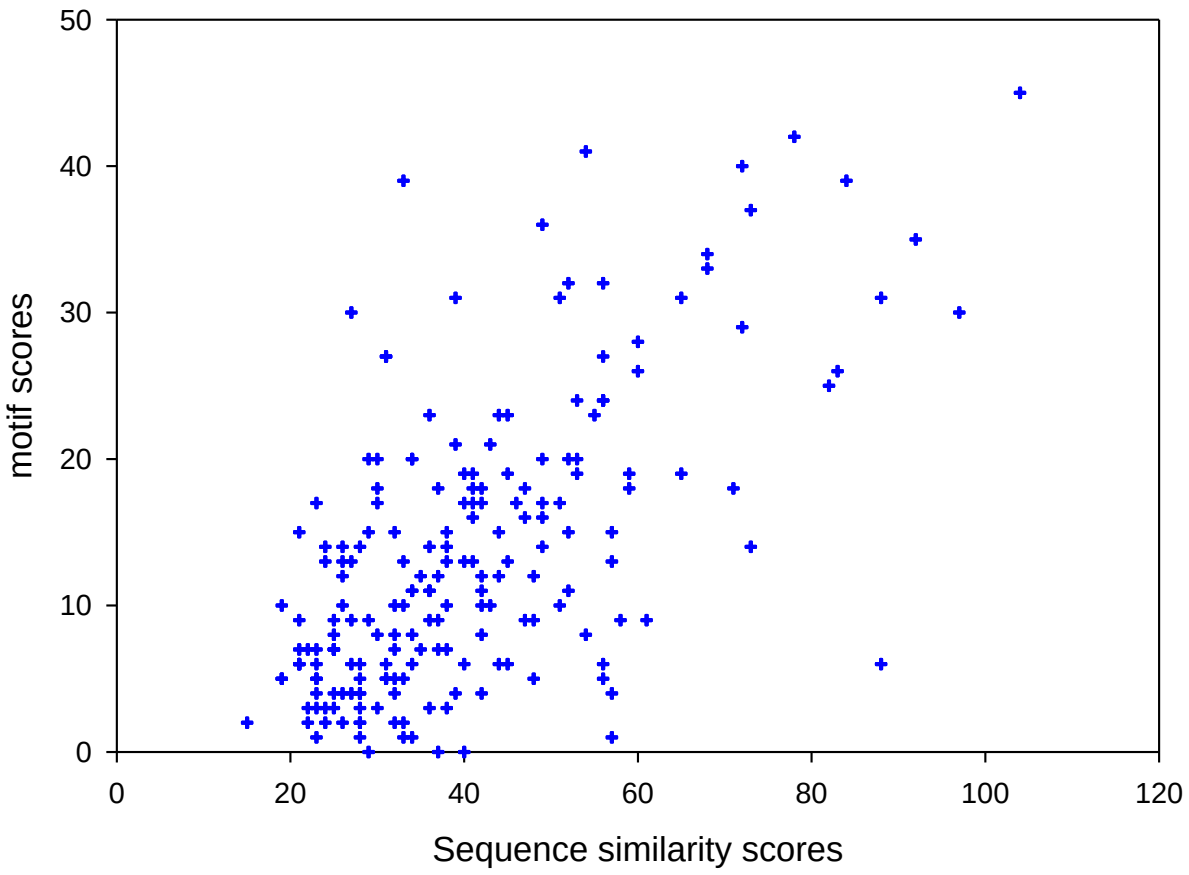

Figure 29 Correlation between Sequence Motif Score and Sequence Similarity Score

The following graph shows the correlation plot between sequence similarity scores and residue conservation scores. The Pearson correlation coefficient for this pair of scores is 0.43062, which is much lower compared to the correlation coefficient for the sequence similarity scores and sequence motif matching scores. This is understandable. Protein pairs which have high sequence similarity don't necessarily have highly conserved positions. A good example for this phenomenon is remote homologues. Some remote homologous protein pairs share high sequence similarity and structure similarity, but they have very different functions, hence low conservation scores. This correlation plot demonstrates that residue conservation score is a different type of information from sequence similarity

scores, so combining these two scores together can help improve protein homology detection performance.

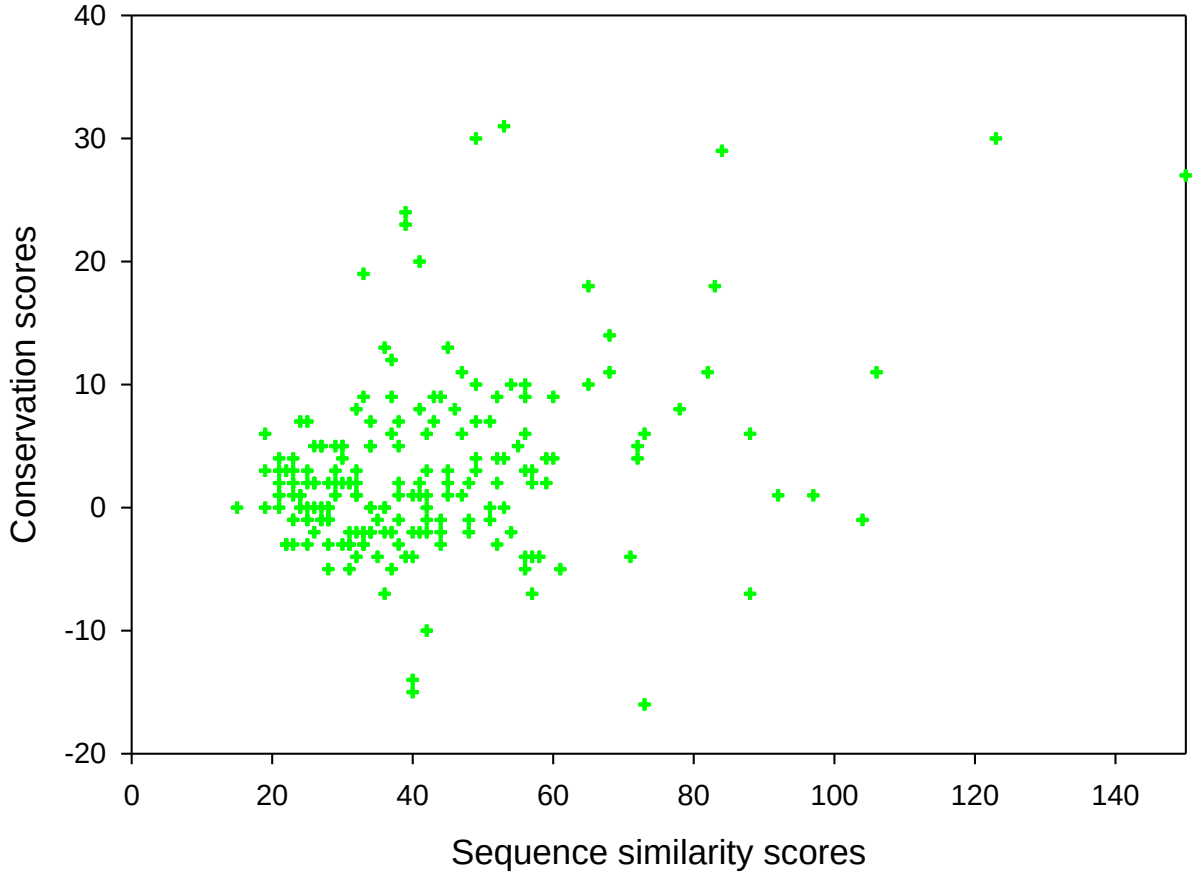

Figure 30 Correlation between Conservation Score and Sequence Similarity Score

The following graph shows the correlation plot between sequence similarity scores and secondary structure scores. The Pearson correlation coefficient for this pair of scores is 0.43577, which is much lower compared to the correlation coefficient for the sequence similarity scores and sequence motif matching scores, and is almost the same as the correlation efficient between the sequence similarity scores and the residue conservation scores. The evolution of protein 3D structure and secondary structure lags behind the evolution of protein sequence. So protein pairs which have high sequence

similarity almost always have high secondary structure similarity, but protein pairs with high secondary structure similarity don't always have high sequence similarity. This makes secondary structure score perfect to detect non-homologues. Proteins which have low secondary structure similarity are unlikely to be homologues.

This correlation plot demonstrates that secondary structure score is a different type of information from sequence similarity scores, so combining these two scores together can help improve protein homology detection performance.

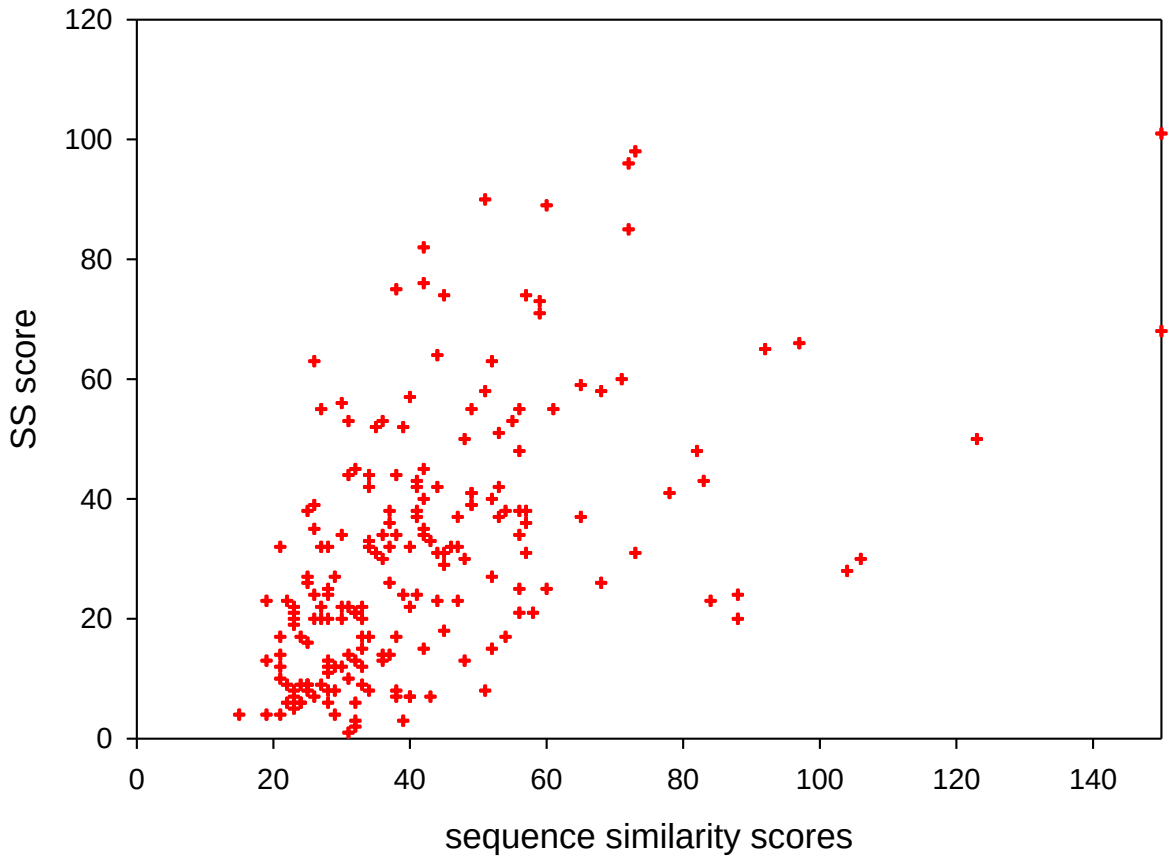

Figure 31 Correlation between Secondary Structure Score and Sequence Similarity Score

I tested ProCAIn with different combination of the three types of assisting information. Some of the results are shown in the following plot. Among these three types of assisting information, secondary structure is the most significant one and it provides the most sensitivity improvement. Combining any two of the three types of assisting information further improves homology detection performance. When all three are combined together, ProCAIn obtains the most sensitivity improvement.

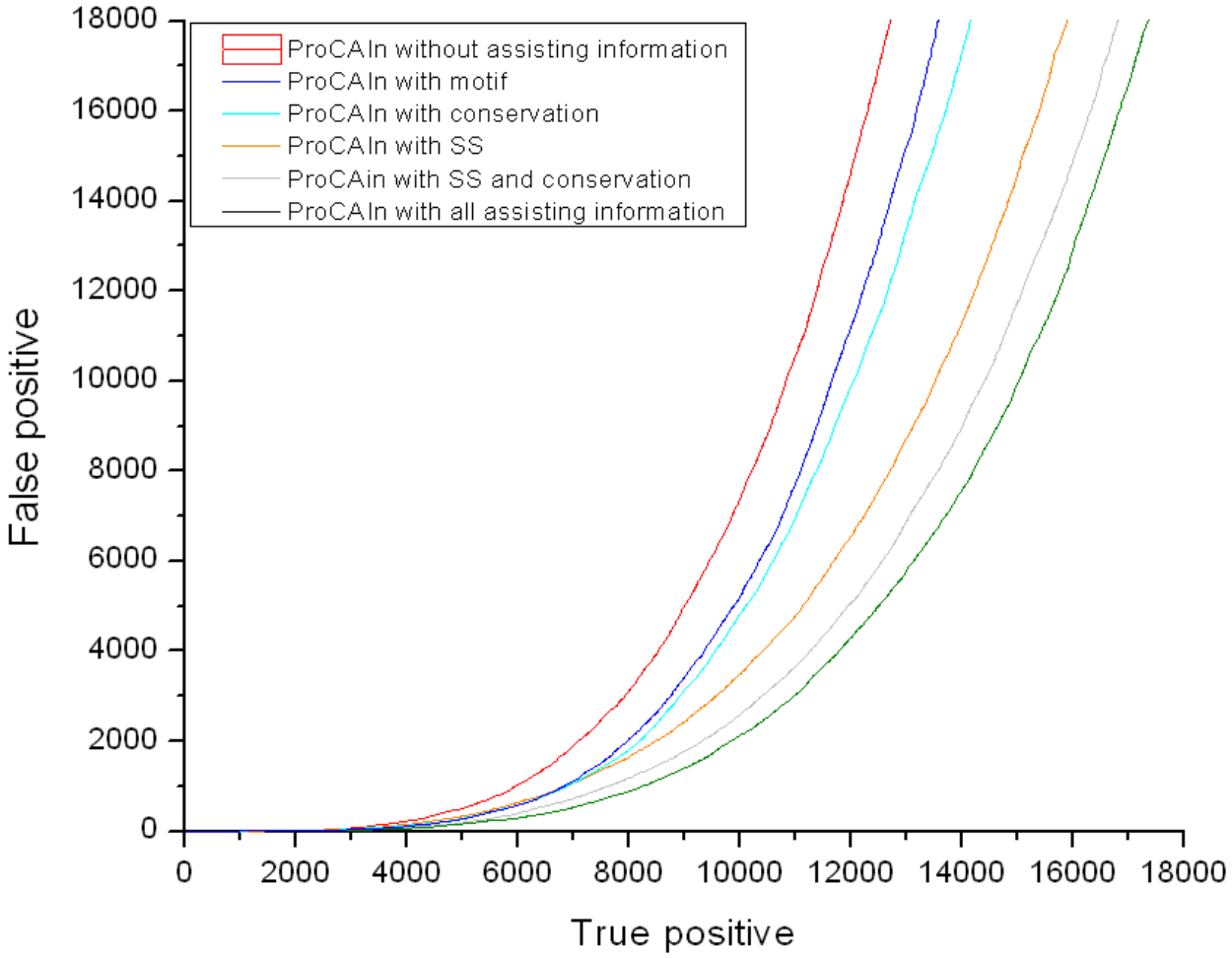

Figure 32 the result of different types of combinations of the three types of assisting information

Above results demonstrate that these three types of information are not the same as sequence similarity scores. They are able to assist sequence similarity score to improve homology

detection performance and when combined together, they can further improve the performance. They are assisting information.

5.2 Results with the training dataset

I randomly picked 1500 protein domains from the whole dataset of 4147 proteins domains to form a training dataset. I trained the weight parameters of the three types of assisting information with this training dataset and benchmarked ProCAIn together with HHsearch and COMPASS to evaluate ProCAIn's performance with protein homology detection and sequence alignment quality. I present these results one by one in the following sections.

5.2.1 Protein homology detection

5.2.1.1. Reference dependent evaluation with SCOP superfamily relationship and SVM score

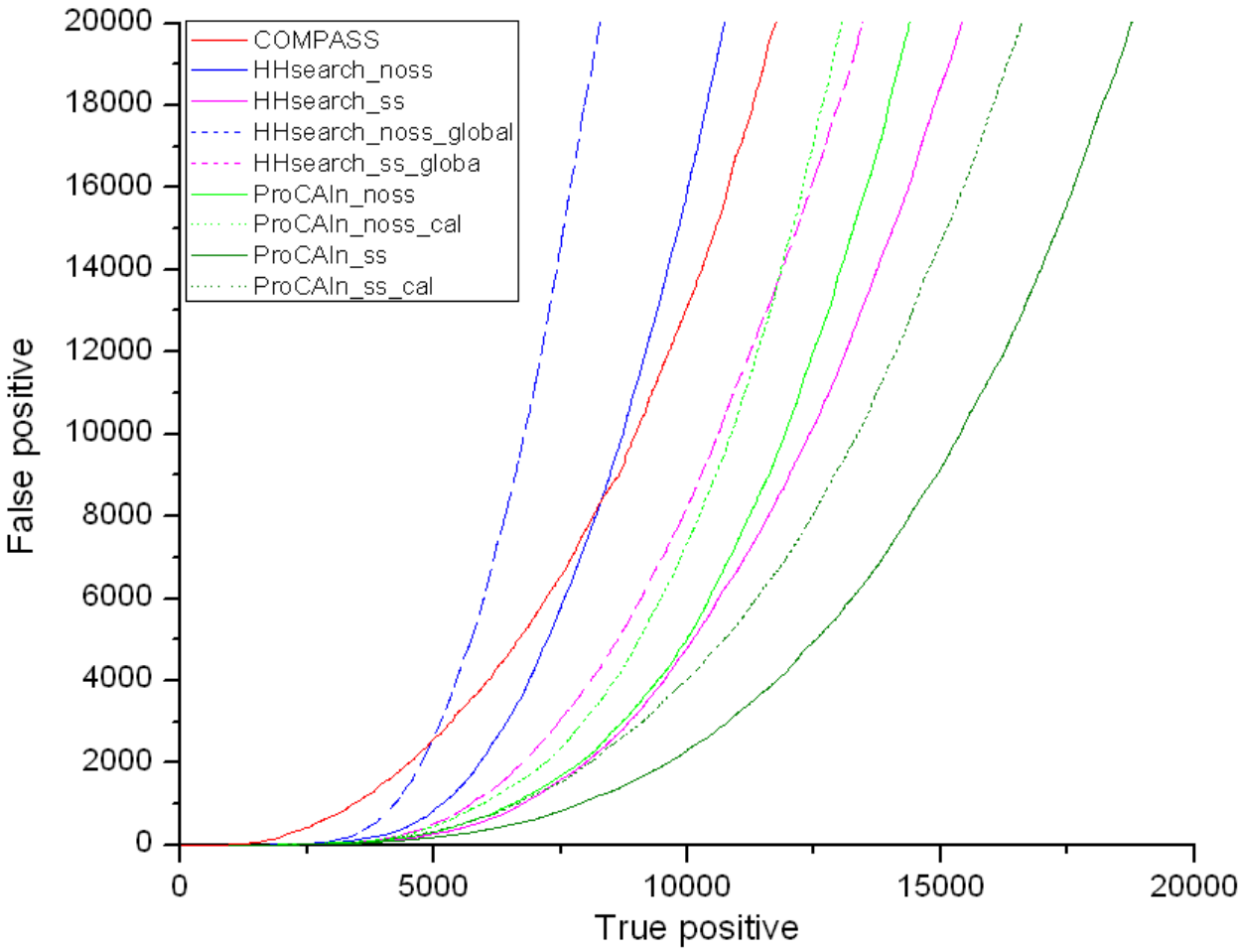

Figure 33 the result of reference dependent evaluation with SCOP superfamily relationship and SVM score

This method has been explained in the above sections. With this method, proteins pairs belong to the same SCOP super family or protein pairs with a SVM score larger than 0.6 are considered homologues and hence true positives. Protein pairs not belong to the same SCOP and with a

SVM score less than -0.6 are considered non-homologues and hence false positives. All other proteins pairs are considered uncertain and discarded from the evaluation process.

Here COMPASS is the results of the latest version of COMPASS. A new version of statistical significance estimation method (Sadreyev and Grishin 2008) is developed and applied to COMPASS. This version of statistical significance estimation method is demonstrated to be more sensitive than the statistical method used by last version of COMPASS.

Just like usual, HHsearch_noss and HHsearch_ss are the results of HHsearch without or with predicated secondary structure information. HHsearch has a global version. Although this version is named HHsearch global, it is still a local sequence alignment program, however it produces much longer alignments. The alignments are so long that they are close to the length of global alignment. This version of HHsearch also produces slightly different optimal scores and hence different probabilities. I also benchmarked this version of HHsearch together with ProCAIn to evaluate its homology detection sensitivity and alignment quality.

ProCAIn_noss is ProCAIn with motif and conservation information but without secondary structure information. This is to test whether ProCAIn_noss performs better than HHsearch_noss. ProCAIn_ss is ProCAIn with all three types of assisting information. ProCAIn uses two slightly different statistical significance estimation methods. One method compares the query protein against the calibration database to get random scores for the query protein and compares the subject protein against the subject database to get random scores for the subject protein. This method is the default method for ProCAIn and is used by ProCAIn unless specified otherwise. Another method of statistical significance estimation for ProCAIn compares

both the query protein and subject protein against the calibration database to get random scores. This method is used for comparisons where protein relationship among the subject proteins is unknown, for example when a database other than SCOP is used as the subject database. This method is labeled as ProCAIn_noss_cal or ProCAIn_ss_cal. Since the protein relationship within the subject database is unknown, this method performs slightly worse than the previous method.

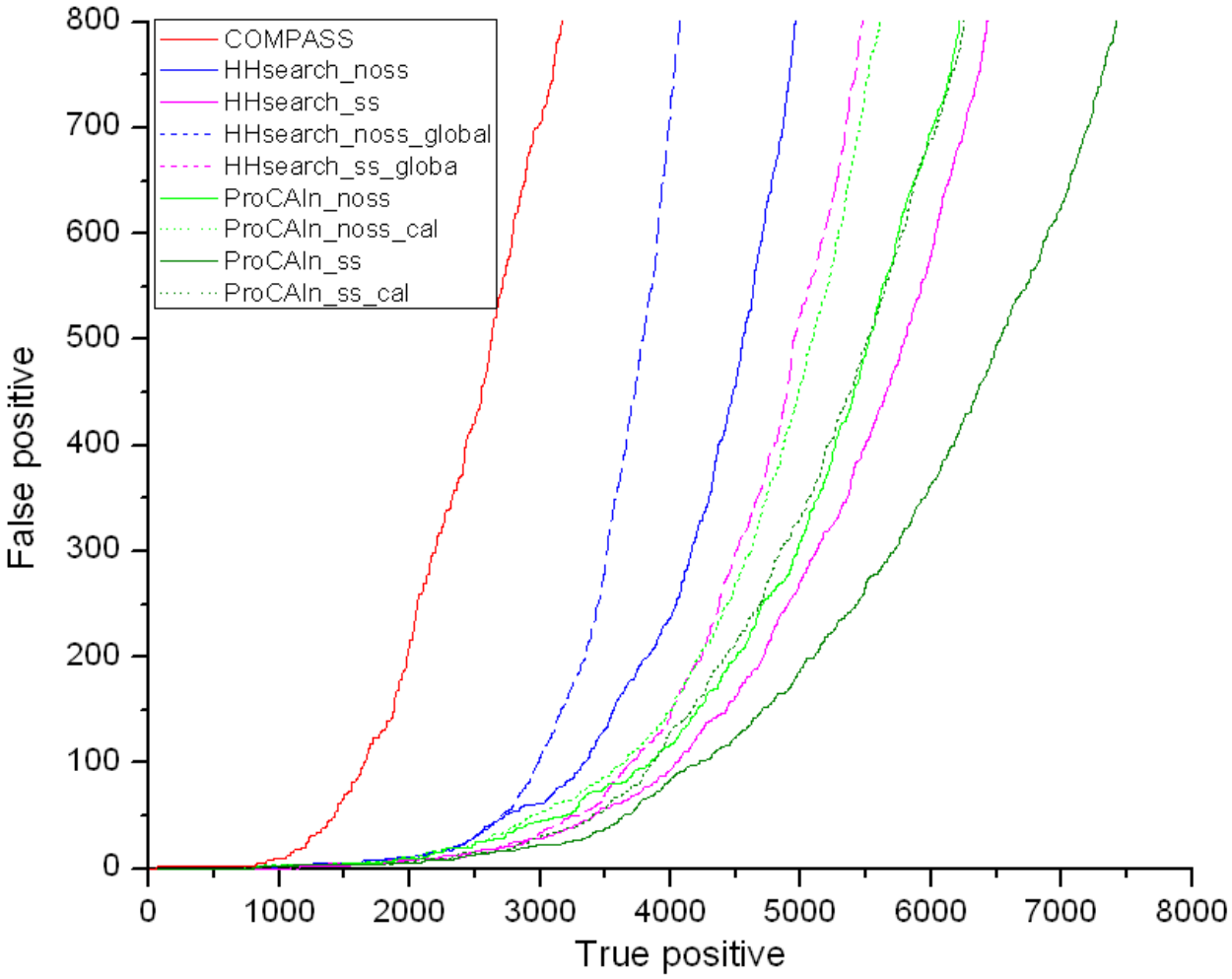

Figure 34 a zoom-in plot of the result of reference dependent evaluation with SCOP superfamily relationship and SVM score

The above plot is a zoom-in of the first plot, so we can see clearly how these protein homology detection methods perform in the beginning.

From the above plots, it is very clear that ProCAIn with secondary structure performs the best. This performance difference starts from the very beginning. This demonstrates that the three types of information are helpful with protein homology detection and secondary structure information is the most helpful one among these three types of assisting information. The results show that HHsearch global always performs worse than regular version of HHsearch. Among these three protein homology detection programs, COMPASS, HHsearch and ProCAIn, COMPASS obviously lags behind.

5.2.1.2 Reference dependent evaluation with SCOP superfamily relationship only

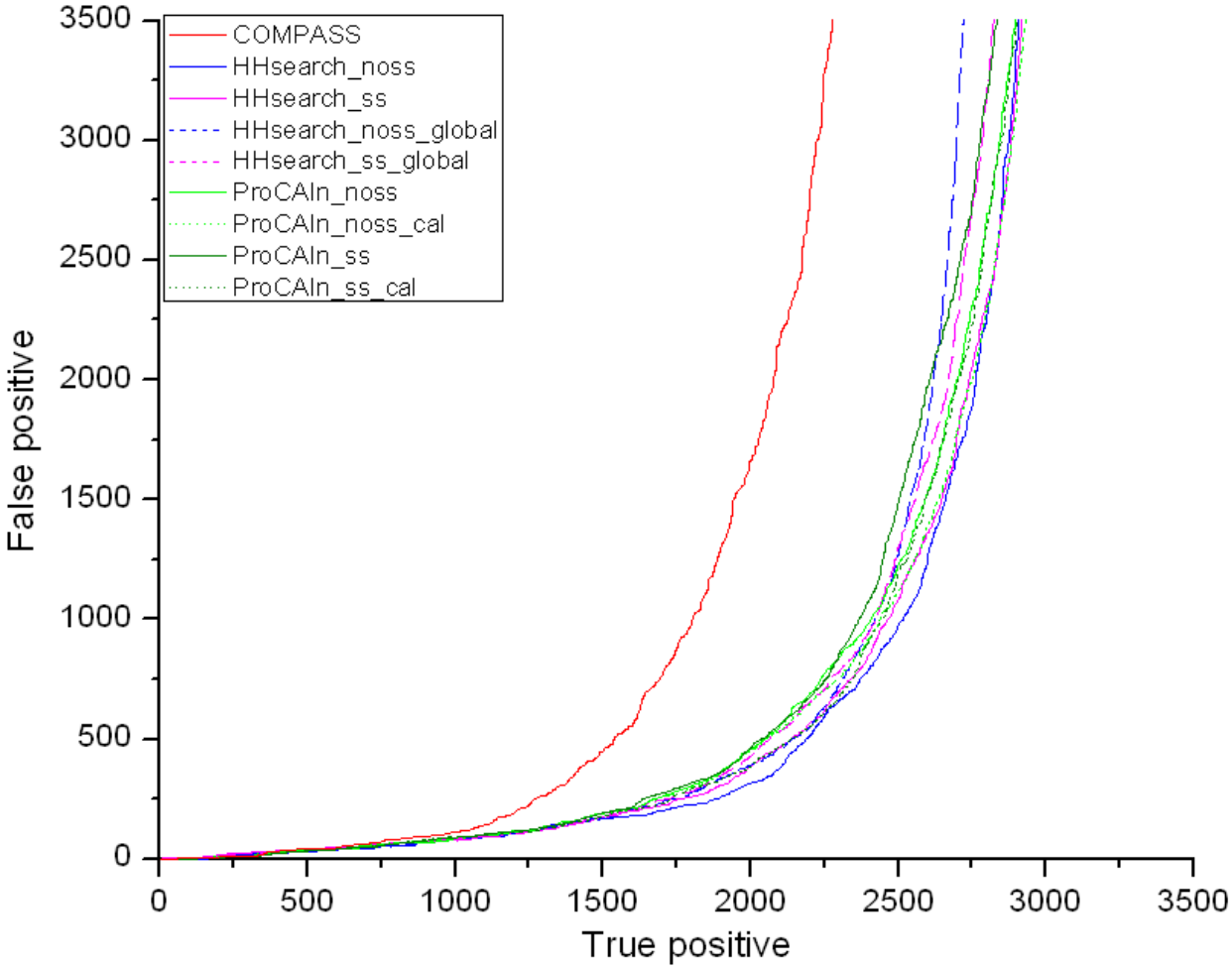

Figure 35 the result of reference dependent evaluation with SCOP superfamily relationship only

This evaluation method has also already been used. With this method, protein pairs belonging to the same SCOP super family are considered as homologues and hence true positives. All other proteins are considered as false positives. This method is to evaluate how a protein homology detection program differentiates close homologues from the rest of proteins.

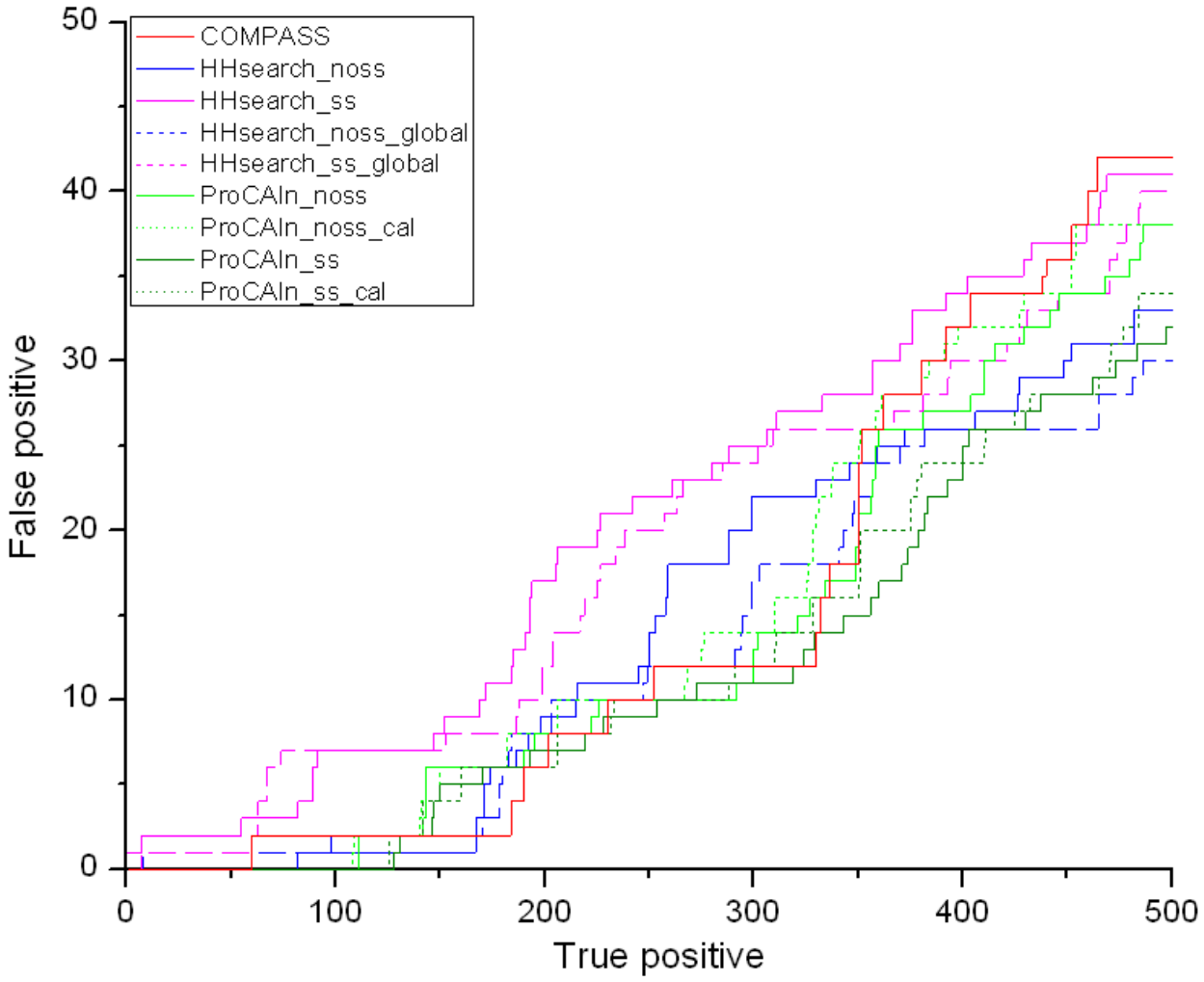

Figure 36 a zoom-in plot of the result of reference dependent evaluation with SCOP superfamily relationship only

The second graph is a zoom-in of the beginning part of the first graph. The above plots demonstrate that ProCAIn and HHsearch have similar performance with this evaluation method. In the beginning, ProCAIn and COMPASS perform slightly better. HHsearch_ss has false positives from the very beginning, this is very disturbing. The reason for this might be because HHsearch simply adds sequence similarity scores together with secondary structure scores. For protein pairs from the same SCOP class, especially for protein pairs from all alpha or all beta

class, sequence similarity scores are always high even for totally unrelated proteins. For these proteins, adding secondary structure scores together with sequence similarity scores may totally overwhelm sequence similarity scores. ProCAIn uses a different method of incorporating assisting information, every type of assisting information is tied together with sequence similarity score.

5.2.1.3 Reference dependent evaluation with SCOP superfamily relationship and SVM score

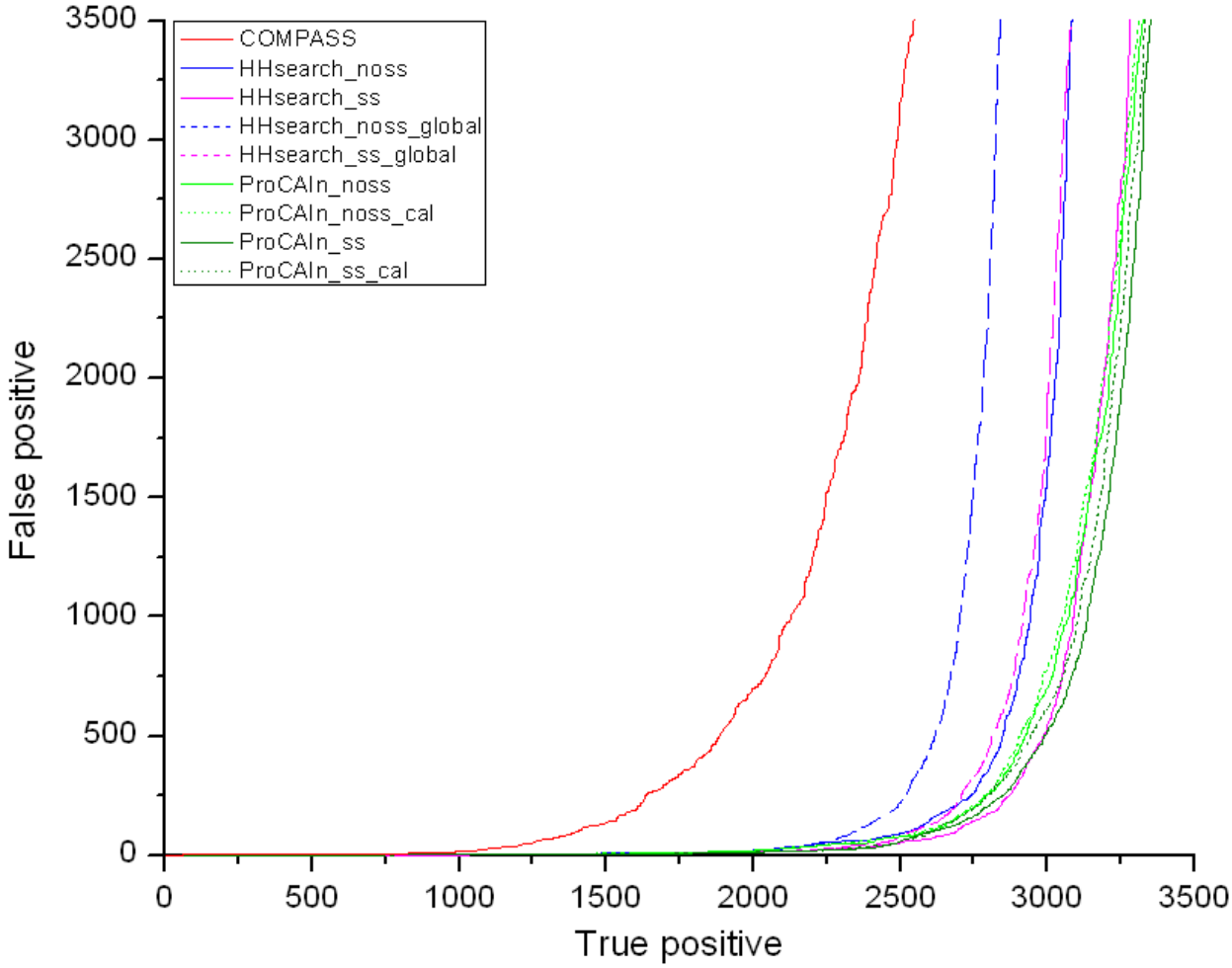

Figure 37 the result of reference dependent evaluation with SCOP superfamily relationship and SVM score

With this evaluation method, protein pairs from the same SCOP super family are considered as homologues and hence true positives. Protein pairs not belonging to the same SCOP super family and also with a SVM score less than -0.6 are considered as false positives. All other proteins are considered uncertain and not counted. This method, combined with the previous two evaluation methods, is to evaluate how a protein homology detection method differentiates close homologues, remote homologues and non-homologues.

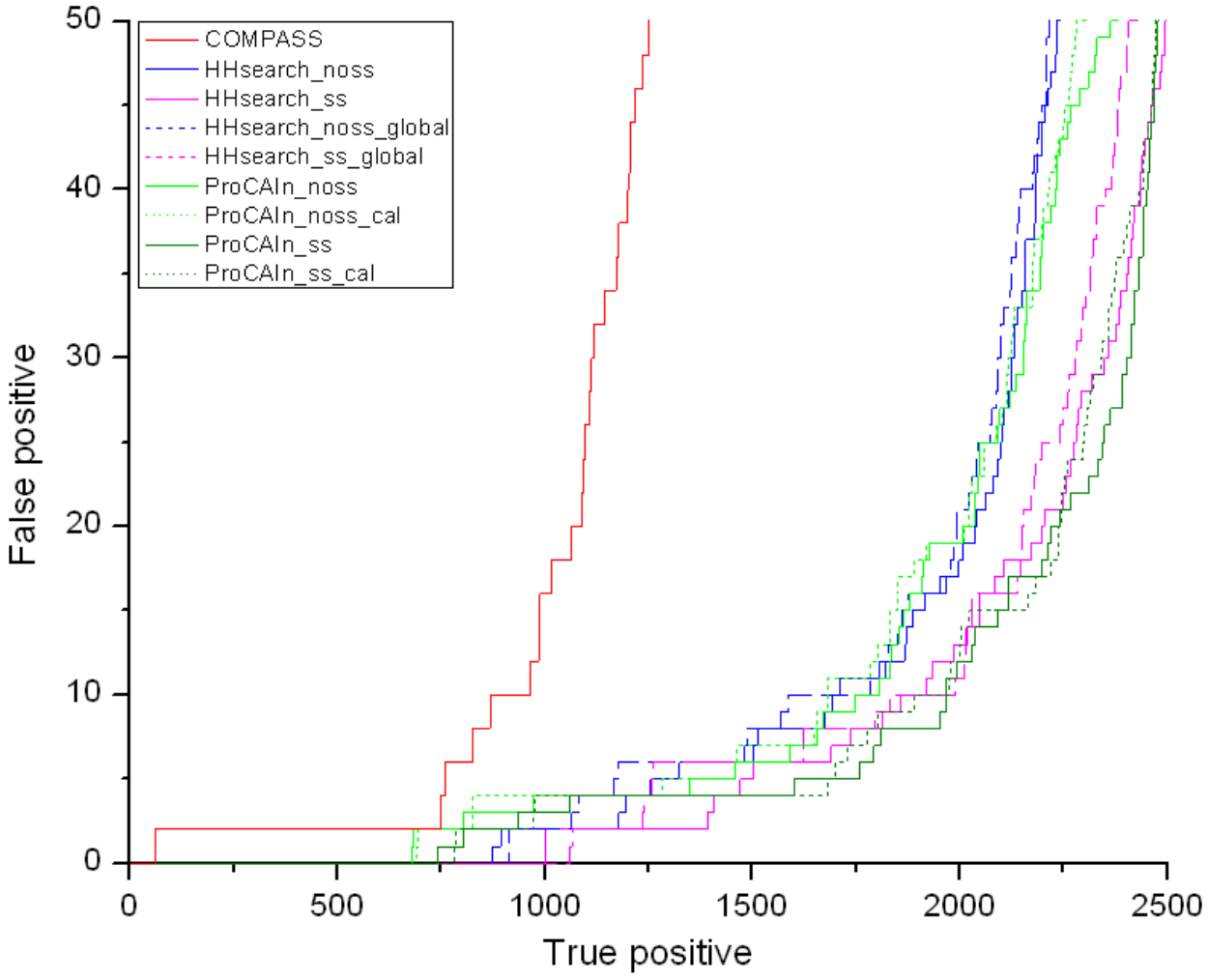

Figure 38 a zoom-in plot of the result of reference dependent evaluation with SCOP superfamily relationship and SVM score

With this evaluation method, ProCAIn_ss, ProCAIn_noss and HHsearch_ss all perform similar and slightly better than HHsearch_noss. COMPASS lags behind all other methods.

The results of the above three evaluation methods demonstrate that ProCAIn and HHsearch perform similar when they are used to detect close homologues. However, ProCAIn performs much better than HHsearch with remote homology detection. This is very significant since all three programs are remote homology detection programs.

5.2.1.4 Reference dependent evaluation with SCOP superfamily relationship, SVM score and alignment quality

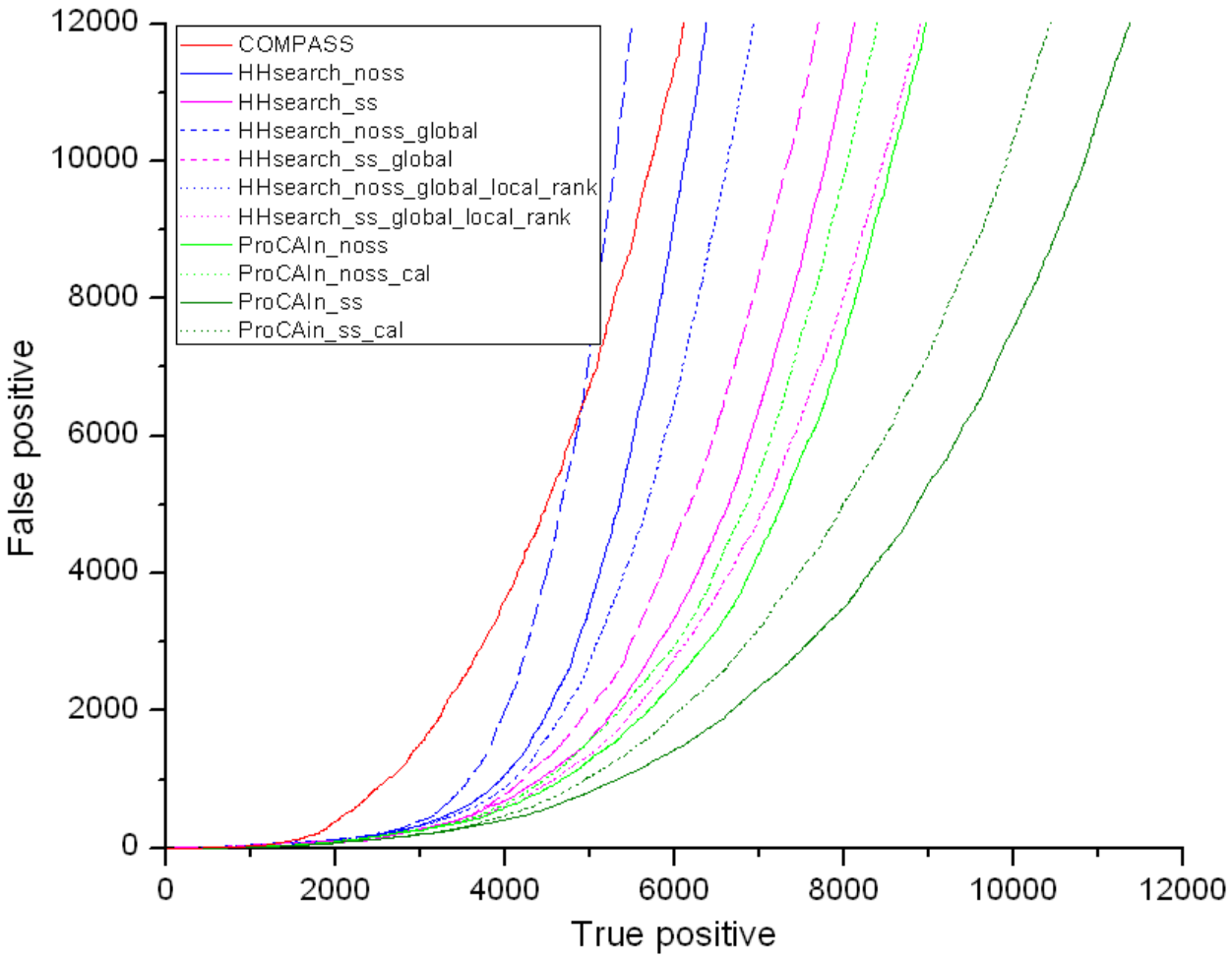

Figure 39 the result of reference dependent evaluation with SCOP superfamily relationship, SVM score and alignment quality

This method is to evaluate a protein homology detection program's sensitivity and alignment quality. Protein pairs which are considered as true positives in evaluation methods 1 are further evaluated to see whether their sequence alignments have a NACC (number of correctly aligned positions) larger than 5 or a GDT_TS larger than 0.15. If they succeed this further test, they will

be considered as true positives stills, otherwise they will be considered as false positives. Here HHsearch_noss_global_local_rank means the result of using the sequence alignment of HHsearch global version without secondary structure and the probability (hence ranking) of HHsearch regular version without secondary structure.

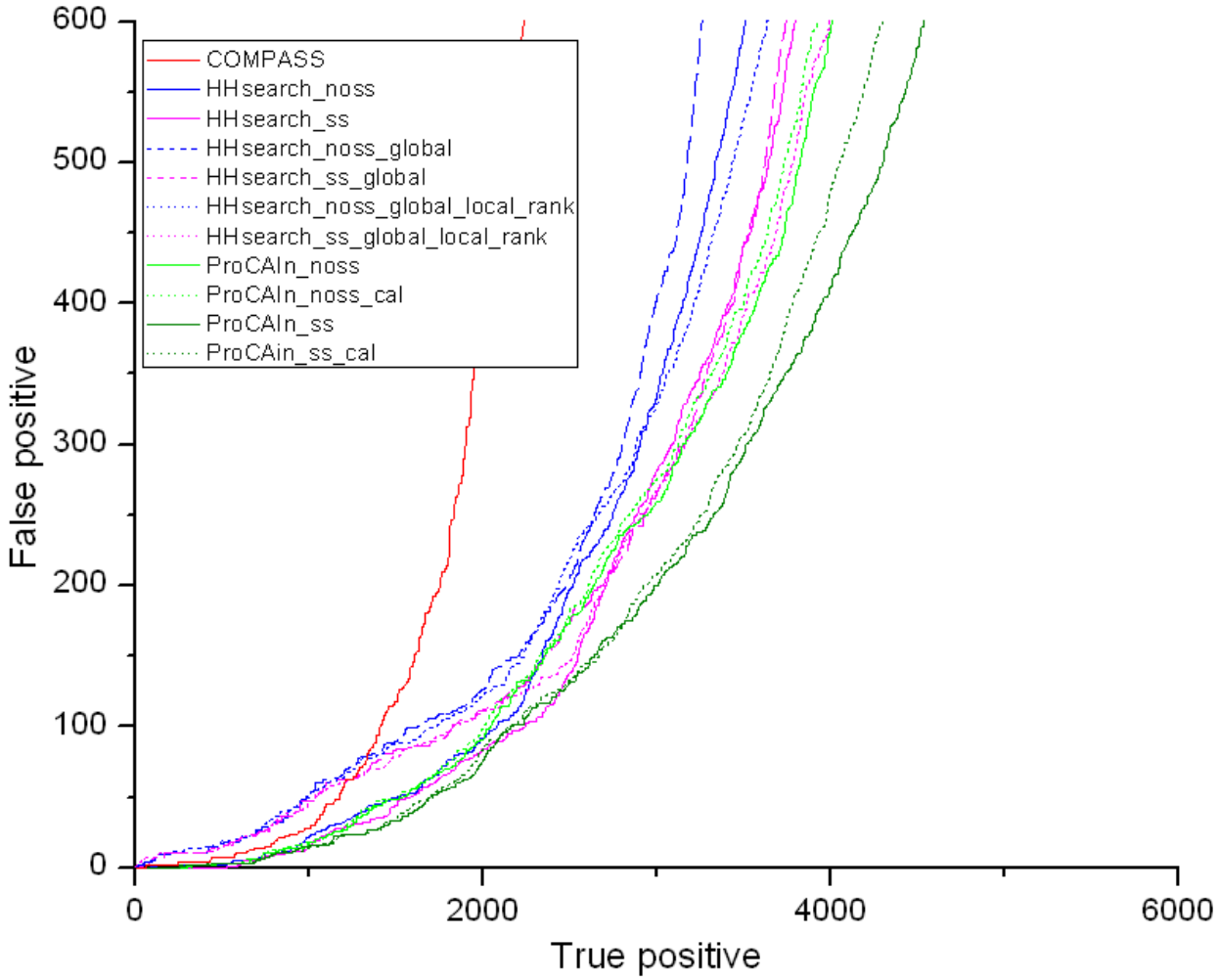

Figure 40 a zoom-in of the result of reference dependent evaluation with SCOP superfamily relationship, SVM score and alignment quality

The above is a zoom-in of the previous plot. Combined the above two plots, we can see: 1. ProCAIn with all three types of assisting performs the best with this evaluation method. It is

much better than HHsearch and COMPASS. 2. Secondary structure helps improve homology detection sensitivity and alignment quality. 3. HHsearch performs better than COMPASS with this evaluation method.

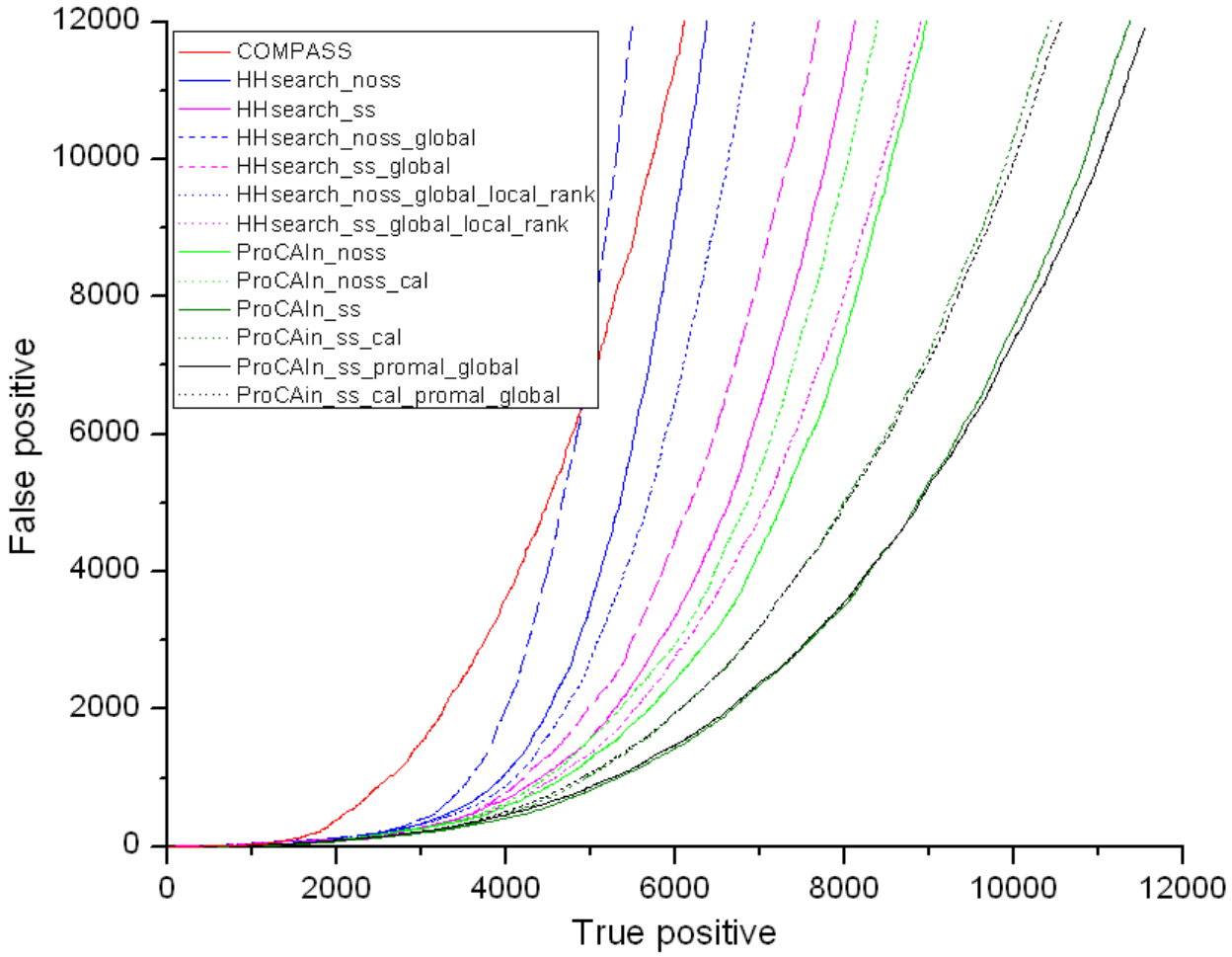

Figure 41 the result of reference dependent evaluation with SCOP superfamily relationship, SVM score and alignment quality (with global results)

This plot adds ProCAIn_ss_promal_global and ProCAIn_ss_cal_promal_global. This is the global version of ProCAIn with the E-value (hence ranking) of regular version of ProCAIn.

5.2.1.5 Reference independent global evaluation with GDT_TS

This evaluation method is not dependent any reference like SCOP to decide whether a pair of proteins are homologous or not, hence it is called reference independent evaluation method.

$$GDT\_TS = \frac{n1 + n2 + n4 + n8}{4} / query\_len$$

n1, n2, n4, n8 are number of aligned residues within 1, 2, 4, 8 angstroms, respectively (Zemla 2003). And *query-len* is the length (amino acid number) of the query protein. Here proteins are superimposed according to the corresponding sequence alignments. This method is similar with the method used by CASP (Critical Assessment of Techniques for Protein Structure Prediction) (Kinch, Wrabl et al. 2003). Only difference is that CASP superimpose proteins by their optimized structure alignments. With this method, proteins with a GDT_TS larger than or equal to 0.15 are considered as true positives and false positives otherwise.

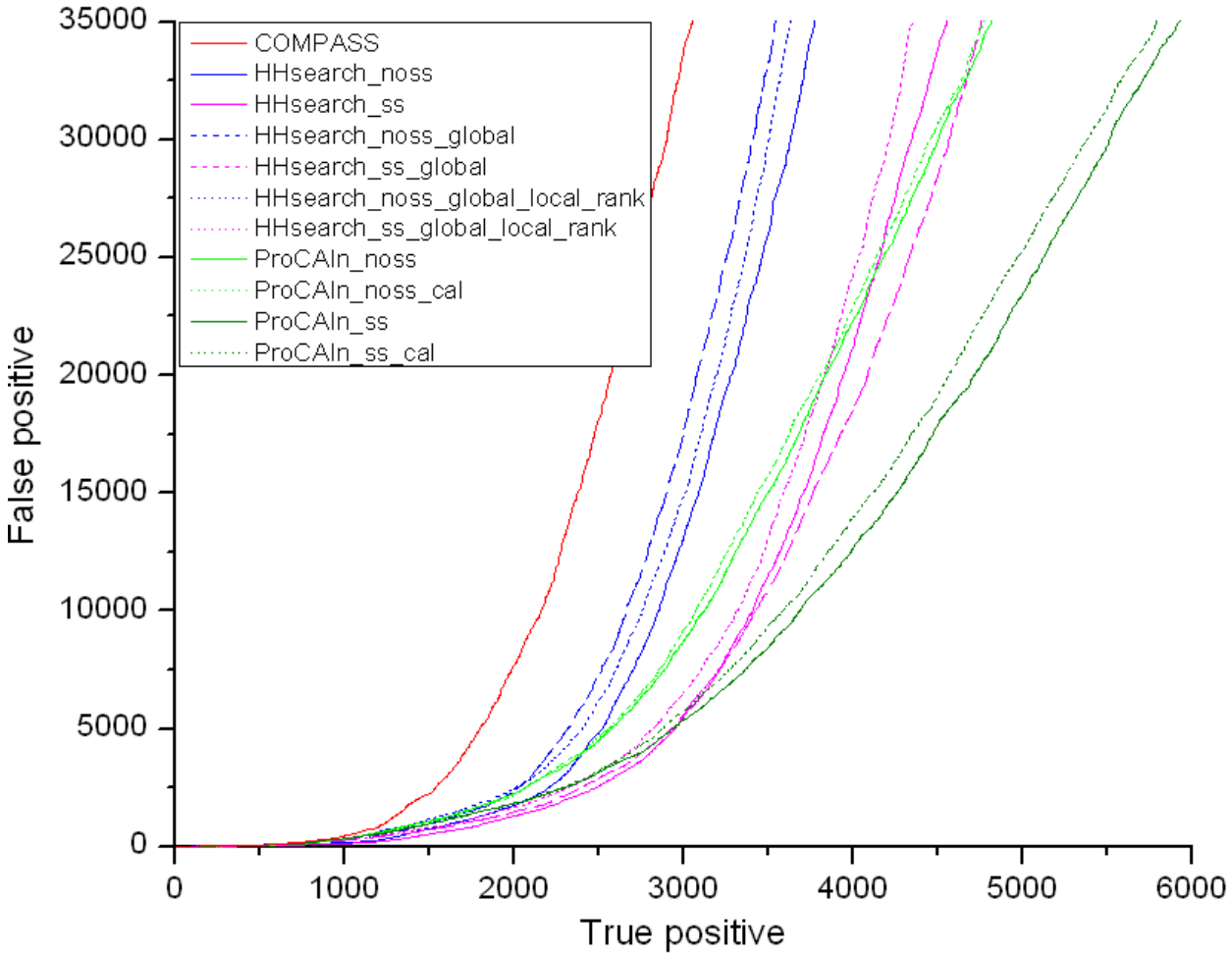

Figure 42 the result of reference independent global evaluation with GDT_TS

Next is a zoom-in of the previous plot. From these two plots, we can see that HHsearch performs the best in the beginning. This is because HHsearch produces much shorter alignments comparing with ProCAIn and COMPASS. And this GDT_TS calculation method favors short alignments since it superimposes protein structures by their corresponding sequence alignments. Longer alignments usually have bigger scores and are normally ranked high in the beginning. Secondly, ProCAIn catches up HHsearch's performance very quickly and outperform HHsearch soon.

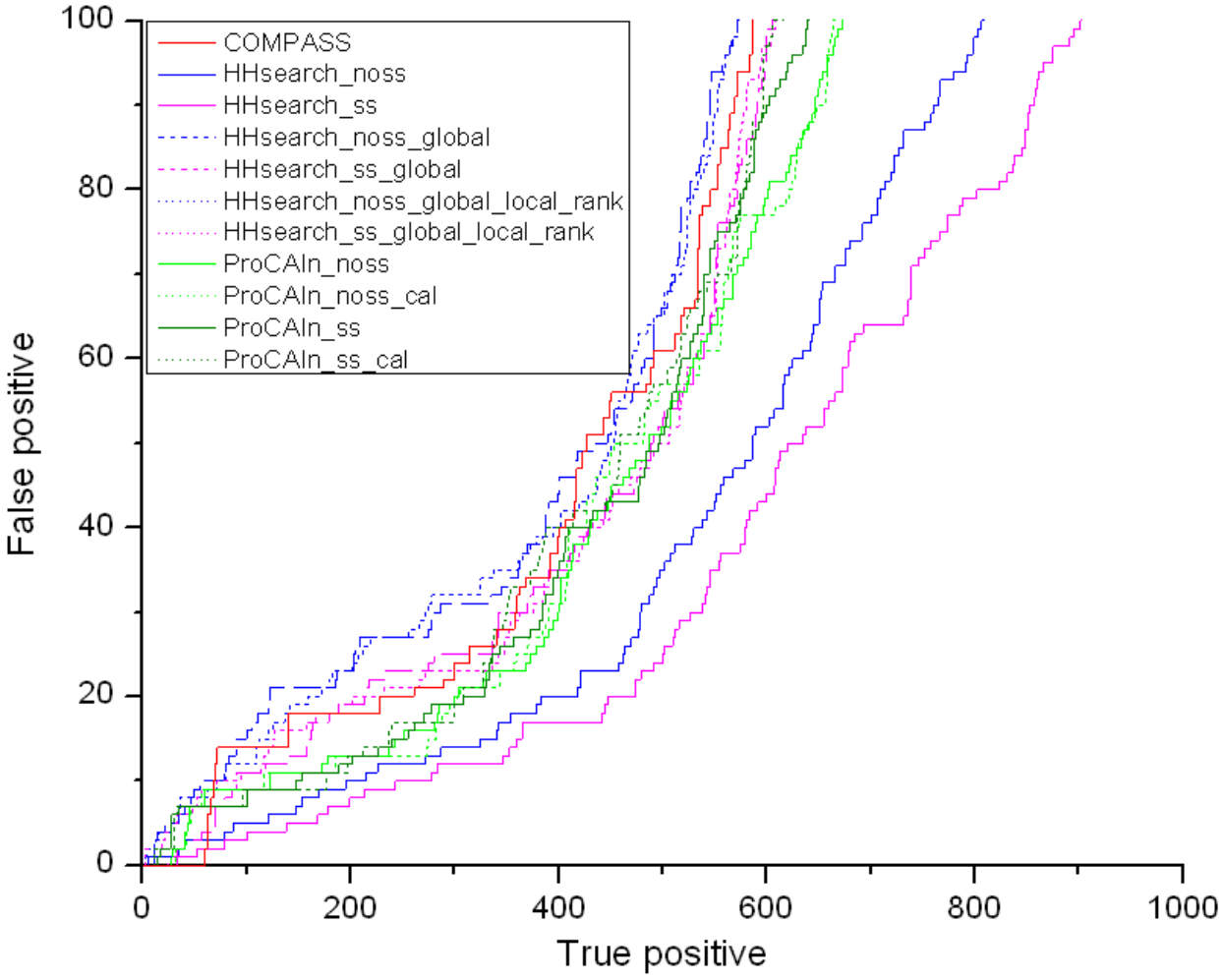

Figure 43 a zoom-in plot of the result of reference independent global evaluation with GDT_TS

The next plot includes ProCAIn global alignments. This version produces global sequence alignments, which is even longer. Since this evaluation method favors short alignments, so you can see ProCAIn global performs worse than ProCAIn regular.

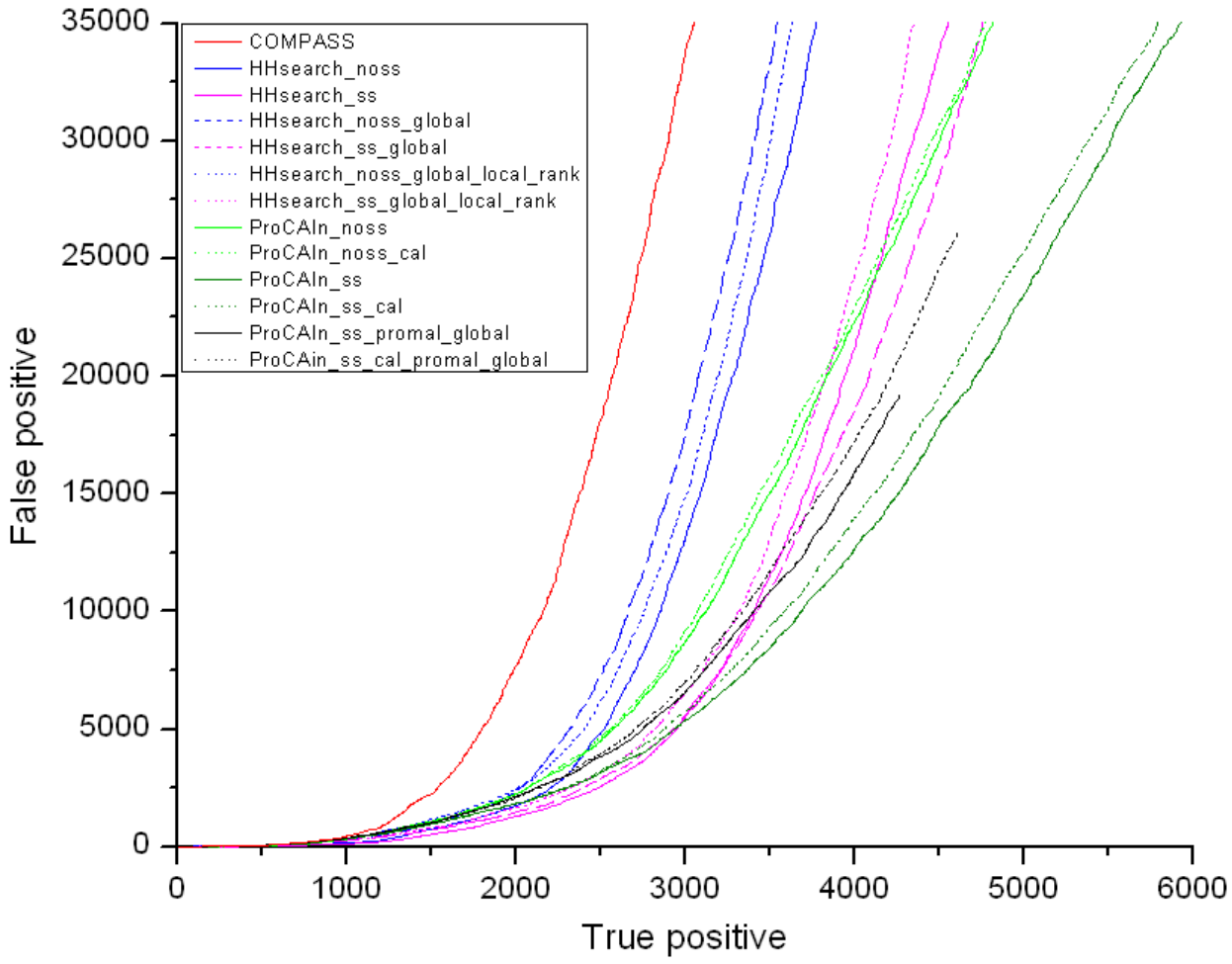

Figure 44 the result of reference independent global evaluation with GDT_TS (with global results)

5.2.1.6 Reference independent global evaluation with LGA GDT_TS

This evaluation method uses the same GDT_TS calculation as the previous one.

$$GDT\_TS = \frac{n1 + n2 + n4 + n8}{4} / query\_len$$

However, here protein structures are superimposed according to their corresponding optimized structure alignments, not sequence alignments. So this method favors longer alignments, since

longer alignments will have more correctly aligned positions even by random. Again, if a sequence alignment has a GDT_TS score larger than 0.15, it will be considered as true positives and false positives otherwise.

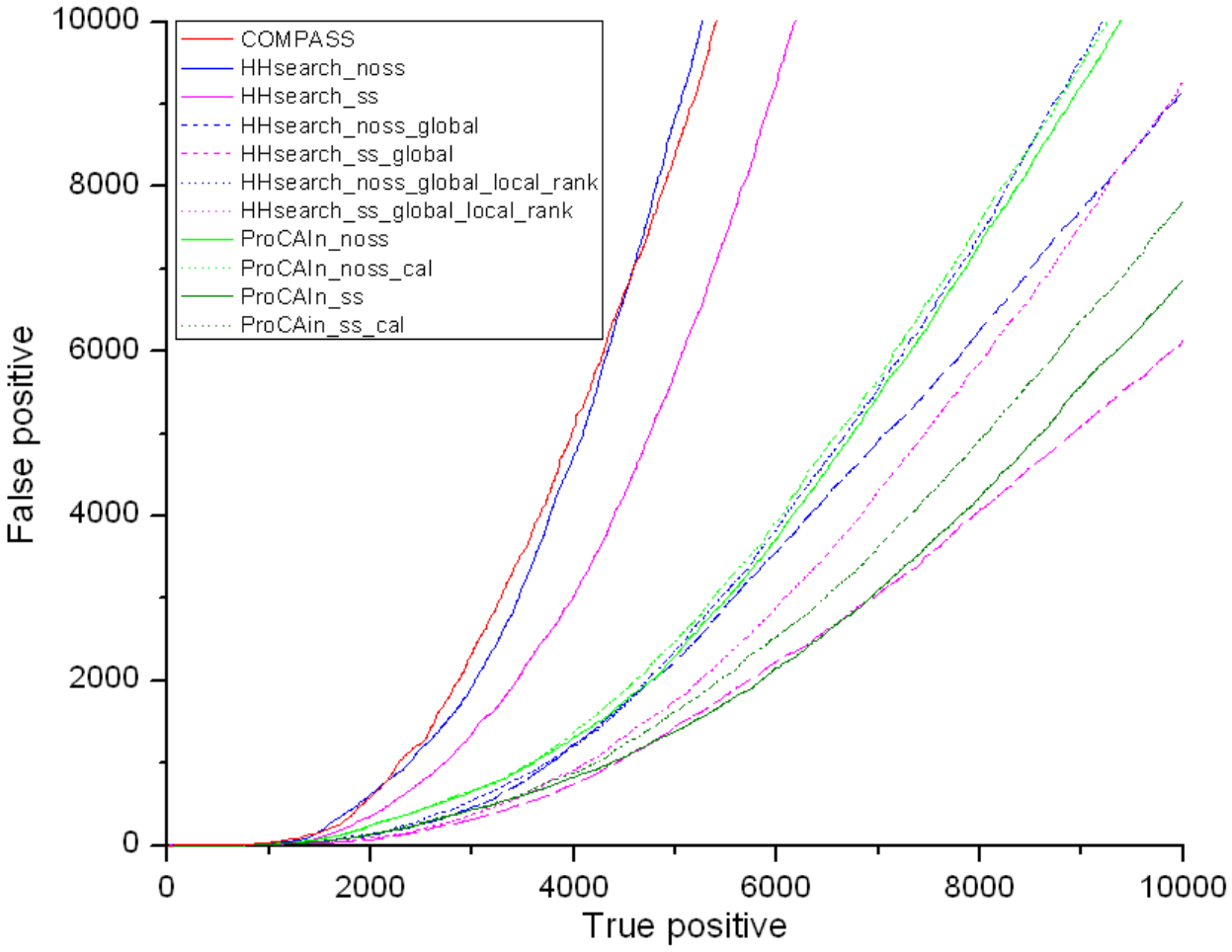

Figure 45 the result of reference independent global evaluation with LGA GDT_TS

The next plot is a zoom-in of the previous plot. The results of these two plots show us that: 1. ProCAIn outperforms both HHsearch and COMPASS. This is because ProCAIn's sequence alignments are generally longer and have more correctly aligned positions. 2. HHsearch global is

almost the same as ProCAIn. HHsearch global version produces even longer alignments and their alignment quality is also good.

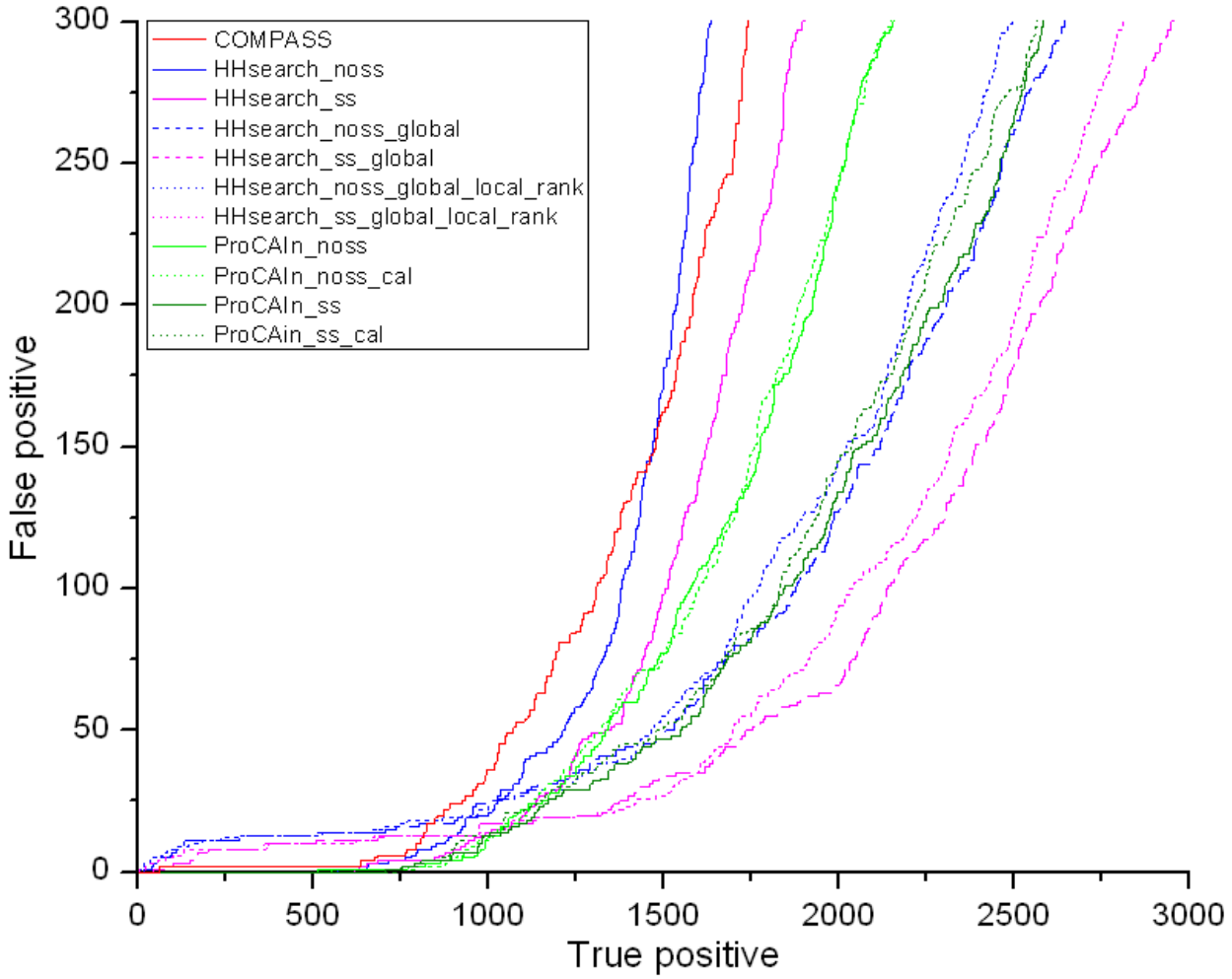

Figure 46 a zoom-in plot of the result of reference independent global evaluation with LGA GDT_TS

Next plot includes ProCAIn global sequence alignment. This result is even better since it is a global sequence alignment, hence much longer. And this evaluation method favors longer sequence alignments.

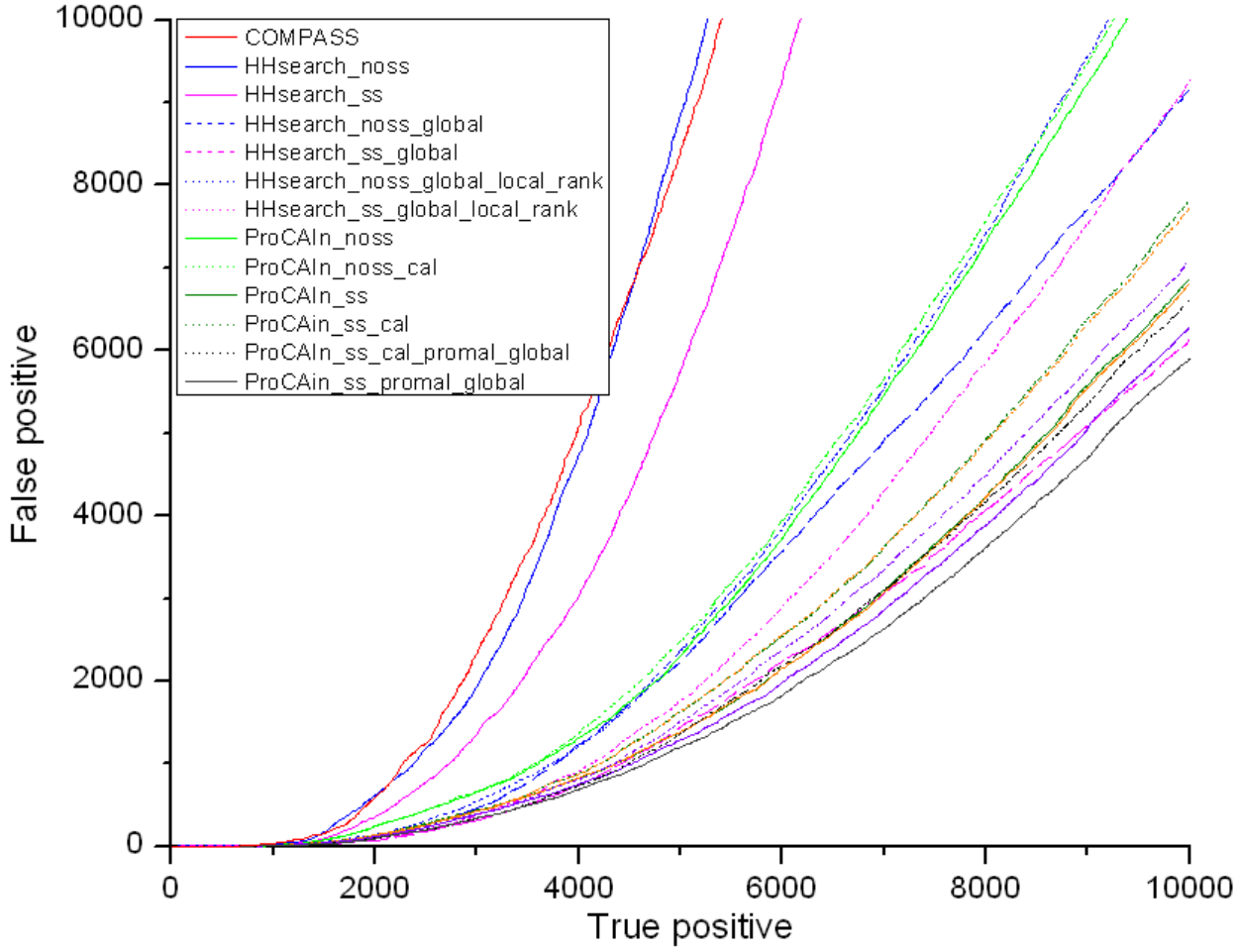

Figure 47 the result of reference independent global evaluation with LGA GDT_TS (with global results)

5.2.1.7 Reference independent global evaluation with Live Bench Contact-a

Live Bench contact-a was developed in the Live Bench experiments (Rychlewski, Fischer et al. 2003). And its equation is the following:

$$LBcontacta = \sum_{i=1}^{L_{aligned}} \frac{\sum_{j=1}^{L_{aligned}} \min\left(D\left(d_{ij}^{1}\right), D\left(d_{ij}^{2}\right)\right)}{\frac{1}{2}\left(\sum_{j=1}^{L_{aligned}} D\left(d_{ij}^{1}\right) + \sum_{j=1}^{L_{aligned}} D\left(d_{ij}^{2}\right)\right)}$$

$$D(d_{ij}) = \begin{cases} \exp(-\ln 2 * d_{ij}), if \ |i-j| \geq 6 \\ 0, otherwise \end{cases}$$

The difference between Live Bench contact and GDT_TS is Live Bench contact doesn't superimpose protein structures. So this method is faster to calculate and it is less biased with protein sequence alignment length. The following plot shows the results. And you can see ProCAIn outperforms HHsearch regular version and COMPASS. And HHsearch global version performs similarly as does ProCAIn.

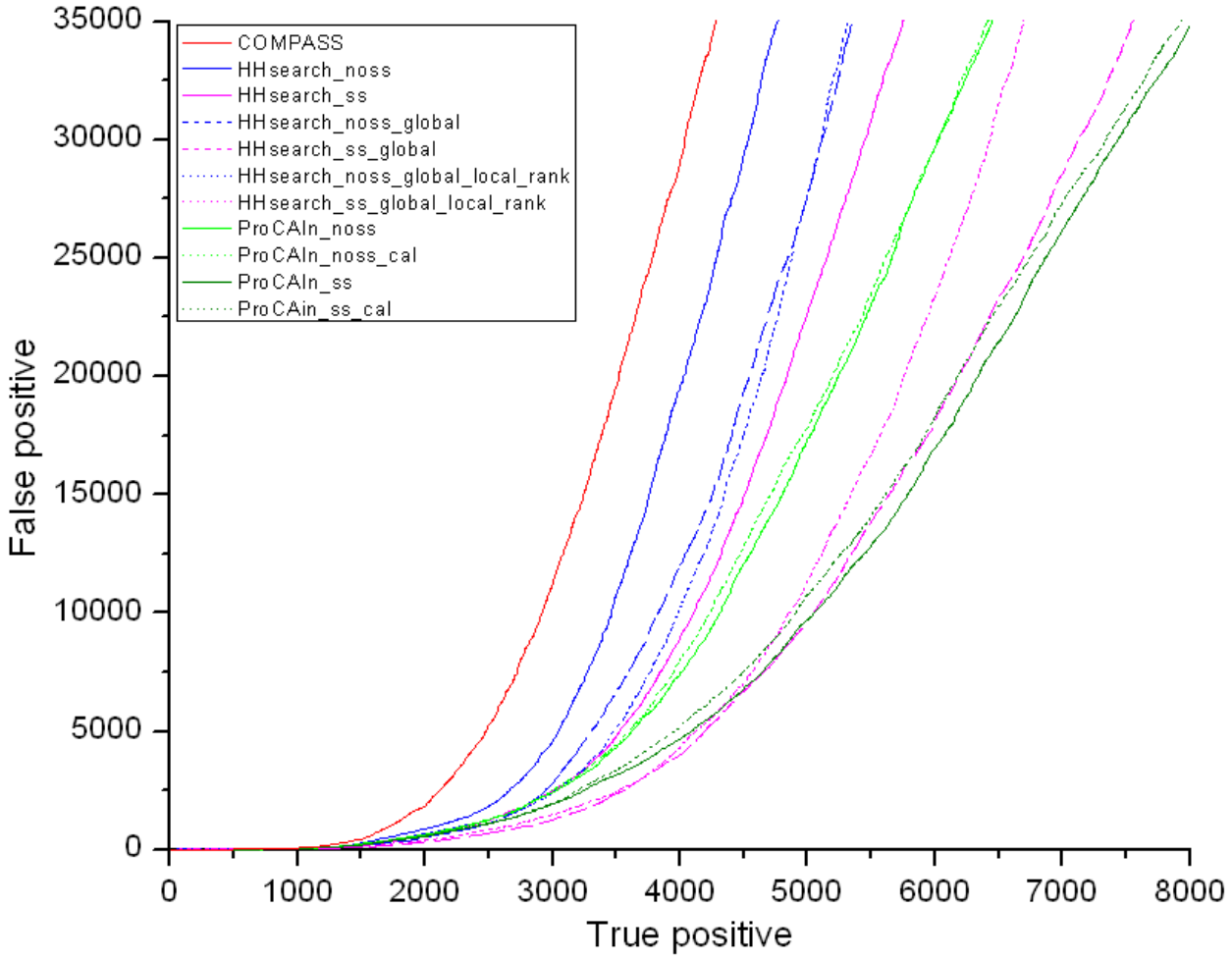

Figure 48 the result of reference independent global evaluation with Live Bench Contact-a

5.2.1.8 Reference independent global evaluation with Live Bench Contact-b

Live Bench contact-b was also developed in the Live Bench experiments (Rychlewski, Fischer et al. 2003). And its equation is the following:

$$LBcontactb = \frac{\sum_{i=1}^{L_{aligned}} \sum_{j=1}^{L_{aligned}} \min\left(D\left(d_{ij}^{1}\right), D\left(d_{ij}^{2}\right)\right)}{\frac{1}{2}\left(\sum_{i=1}^{L_{aligned}} \sum_{j=1}^{L_{aligned}} D\left(d_{ij}^{1}\right) + \sum_{i=1}^{L_{aligned}} \sum_{j=1}^{L_{aligned}} D\left(d_{ij}^{2}\right)\right)} * L_{aligned}$$

$$D\left(d_{ij}\right) = \begin{cases} \exp\left(-\ln 2 * d_{ij}\right), if \ |i-j| \geq 6 \\ 0, otherwise \end{cases}$$

The difference between contact-a and contact-b is that contact-a counts the number of contacts between two proteins and contact-b counts the number of contacts within a protein itself.

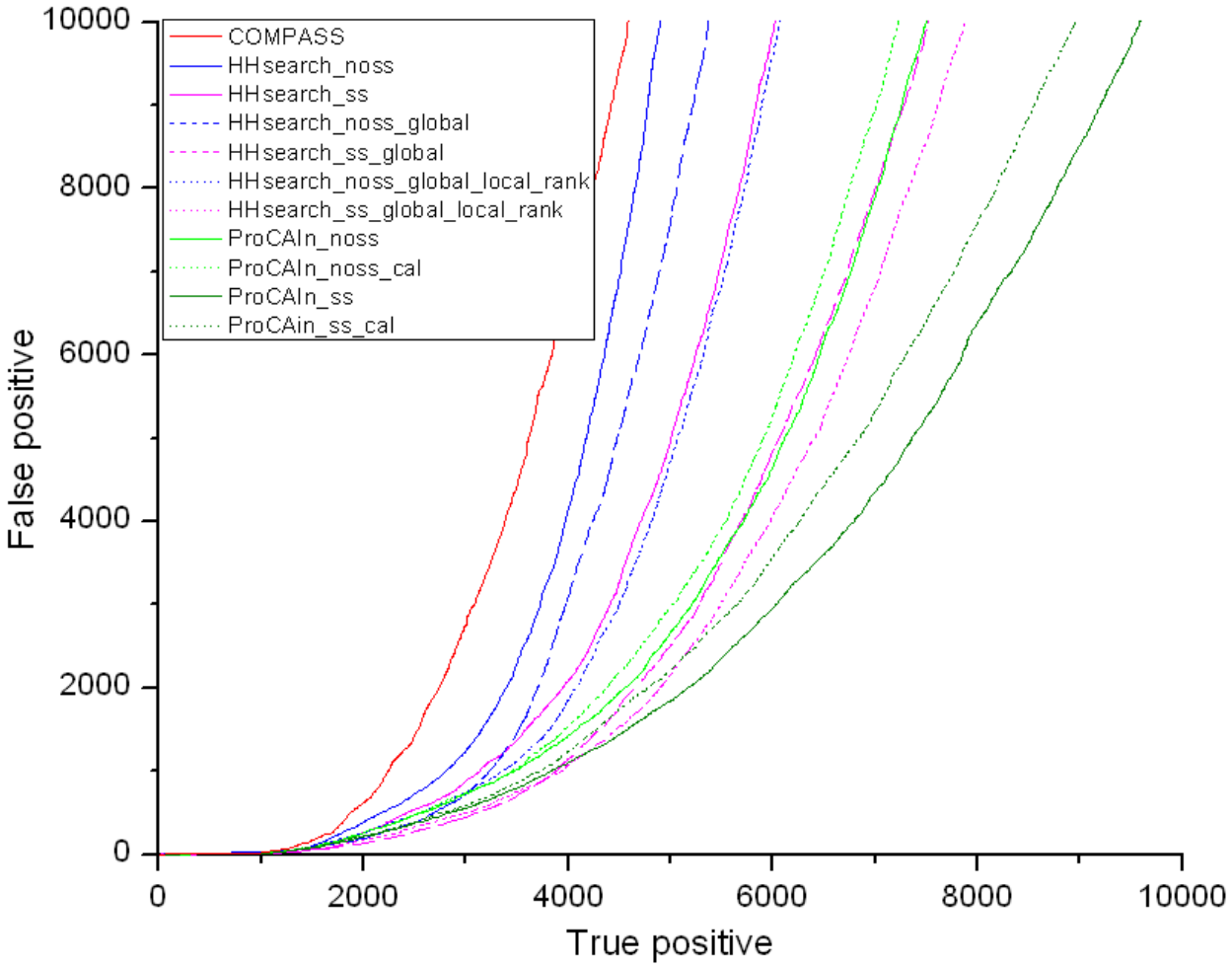

Figure 49 the result of reference independent global evaluation with Live Bench Contact-b

From the previous plot, you can see that ProCAIn outperforms HHsearch regular and global, and COMPASS.

5.2.2 Query family sensitivity student t-test

Evaluation based on all-to-all comparisons might be biased if a subset of queries produces many highly significant hits that dominate the beginning of the ROC curve. To control for such a bias, we compare the performance of the methods query by query. For each query in our set, we consider the sorted list of hits and calculate sensitivity at a given level of selectivity (10%, 25% or 50%). For a pair of methods,

sensitivity values for each query are compared using paired t-test. The following first table shows t-test P-values for sensitivity at 10% selectivity; the second table shows t-test P-values for sensitivity at 25% selectivity and the third table for sensitivity at 50% selectivity. Consistent with the results of all-to-all comparisons, at the level of individual queries PROCAIN performs significantly better than other methods.

From the following three tables we can see:

1. HHsearch and ProCAIn is more sensitive than COMPASS, and the significance of this improvement gets bigger when selectivity gets bigger.
2. Secondary structure helps a lot. ProCAIn_ss outperforms ProCAIn_noss. HHsearch_ss outperforms HHsearch_noss.
3. ProCAIn is better than HHsearch. ProCAIn_ss is better than HHsearch_ss and ProCAIn_noss is better than HHsearch_noss.
4. The statistical estimation method with both calibration databases is better than the method with only calibration database. ProCAIn_noss outperforms Pro_CAIn_noss_cal and ProCAIn_ss outperforms ProCAIn_ss_cal.

5.2.2.1 Ten percent sensitivity t-test

10% sensitivity	**COMPASS**	**ProCAIn_noss_cal**	**ProCAIn_noss**	**ProCAIn_ss_cal**	**ProCAIn_ss**	**HHsearch_noss**
ProCAIn_noss_cal	-6.45e-01 -4.39e-01 5.30e-01 3.32e-03 -1.20e-02 -3.51e-02					

ProCAln_noss	-3.16e-05 -4.71e-02 -7.62e-01 3.69e-02 -2.8e-03 -2.7e-02	-3.06e-25 -9.73e-08 -5.03e-05 -2.29e-03 -3.32e-03 -7.83e-01				
ProCAln_ss_cal	-2.7e-14 -4.15e-12 -2.73e-05 -2.12e-14 -1e-60 -5.85e-08	-1.64e-21 -4.67e-18 -2.47e-10 -1.06e-29 -3.8e-62 -6.84e-05	-8.56e-09 -7.48e-12 -2.23e-09 -4.77e-27 -1.46e-54 -3.5e-05			
ProCAln_ss	-1.68e-30 -4.46e-15 -7.01e-06 -3.1e-14 -1.58e-57 -7.09e-07	-1.03e-37 -8.3e-19 -3.44e-08 -1.48e-26 -5.32e-56 -9.19e-03	-2.37e-24 -7.55e-15 -2.45e-06 -2.53e-24 -2.93e-50 -2.76e-03	-2.13e-14 -2.08e-07 -5.94e-01 2.56e-01 2.83e-01 4.47e-01		
HHsearch_noss	7.02e-17 1.04e-06 1.08e-01 1.28e-06 1.37e-10 -5.41e-01	4.26e-17 4.88e-08 -8.53e-01 1.26e-01 9.74e-12 2.36e-02	4.12e-33 8.02e-11 6.65e-01 1.89e-02 1.46e-13 4e-03	1.48e-46 3.9e-28 7.4e-10 2.77e-22 5.69e-72 1.58e-06	6e-62 5.79e-31 1.44e-07 4.23e-22 2.22e-64 2.74e-06	
HHsearch_ss	-1.85e-15 -2.19e-02 -6.33e-10 -4.86e-12 -2.07e-23 -4.98e-03	-6.29e-16 -2.73e-01 -1.27e-10 -1.67e-14 -8.54e-11 -9.78e-01	-6.68e-07 -9.79e-01 -6.2e-09 -1.15e-12 -1.14e-09 8.87e-01	-7.47e-01 2.69e-06 -1.98e-01 4.49e-01 1.06e-12 6.79e-03	3.26e-04 2.73e-09 -2.74e-02 3.86e-01 4.28e-10 3.25e-02	-1.34e-79 -2.93e-22 -1.26e-15 -1.3e-26 -1.56e-52 -1.97e-05

Table 3 the result of 10% sensitivity t-test

5.2.2.2 Twenty-five percent sensitivity t-test

25% sensitivity	**COMPASS**	**ProCAln_noss_cal**	**ProCAln_noss**	**ProCAln_ss_cal**	**ProCAln_ss**	**HHsearch_noss**
ProCAln_noss_cal	-1.43e-02 -2.98e-08 -2.42e-01 -8.64e-01 -1.08e-04 -7.24e-05					

ProCAln_noss	-2.56e-12 -8.39e-13 -1.41e-01 -4.37e-01 -8.15e-05 -2.36e-05	-7.56e-35 -4.52e-12 -1.51e-03 -8.93e-02 -9.22e-03 6.31e-01				
ProCAln_ss_cal	-1.19e-23 -2.41e-35 -1.11e-11 -3.81e-22 -8.66e-80 -1.19e-09	-4.09e-27 -2.8e-27 -1.24e-11 -1.35e-32 -8.47e-108 -1.16e-04	-4.79e-07 -1.33e-16 -5.93e-10 -1.29e-30 -1.25e-96 -5.84e-04			
ProCAln_ss	-1.13e-46 -1.98e-41 -1.22e-14 -8.33e-23 -9.4e-75 -6.93e-09	-2.14e-56 -3.72e-32 -2.82e-14 -6.69e-32 -3.23e-102 -4.48e-03	-1.09e-32 -8.78e-27 -6.18e-13 -2.31e-31 -2.46e-97 -1.35e-03	-1.21e-34 -2.77e-09 -7.45e-04 -4.51e-01 3.48e-01 -8.95e-01		
HHsearch_noss	2.21e-24 9.68e-05 6.19e-02 2.59e-09 6.83e-30 -1.96e-03	7.81e-35 1.85e-16 2.33e-02 1.56e-08 8.26e-28 2.38e-02	6.74e-51 1.61e-20 7.05e-03 7.83e-11 9.81e-30 3.27e-02	6.25e-66 1.58e-45 6.11e-13 2.59e-44 6.61e-117 2.35e-04	4.44e-82 1.19e-53 7.55e-15 5.94e-41 5.47e-111 1.11e-03	
HHsearch_ss	-8.57e-21 -1.17e-08 -3.01e-12 -2.88e-05 -2.35e-09 -6.42e-07	-8.5e-14 -1.24e-01 -1.56e-13 -5.54e-10 -1.02e-05 -8.14e-01	-9.7e-04 -9.09e-01 -2.05e-12 -4.4e-07 -3.58e-05 -4.13e-01	4.01e-01 2.72e-15 -7.85e-01 3.24e-06 3.16e-38 1.59e-01	6.32e-09 1.75e-20 7.31e-01 2.96e-06 4.17e-35 2.16e-01	-1.16e-103 -2.95e-35 -6.9e-29 -7.84e-53 -4.15e-74 -6.13e-06

Table 4 the result of 25% sensitivity t-test

5.2.2.3 Fifty percent sensitivity t-test

50% sensitivity	**COMPASS**	**ProCAln_noss_cal**	**ProCAln_noss**	**ProCAln_ss_cal**	**ProCAln_ss**	**HHsearch_noss**
ProCAln_noss_cal	-7.98e-04 -1.24e-19 -1.23e-04 -1.15e-01 -2.23e-03 -6.2e-07					

ProCAln_noss	-5.15e-16 -4.28e-26 -1.18e-04 -6.88e-02 -1.53e-03 -1.63e-07	-7e-49 -1.49e-16 -8.13e-03 -2.96e-01 -2.05e-03 2.5e-01				
ProCAln_ss_cal	-3.55e-40 -8.17e-66 -1.31e-18 -4.02e-33 -2.16e-104 -3.1e-15	-2.23e-31 -5.93e-42 -1.23e-14 -1.01e-35 -6.46e-134 -2.62e-05	-1.91e-10 -8.32e-28 -2.41e-15 -5.47e-35 -3.72e-129 -7.01e-06			
ProCAln_ss	-1.26e-65 -5.9e-79 -1.06e-18 -5.38e-32 -5.68e-95 -4.78e-14	-2.09e-64 -2.71e-57 -1.11e-14 -1.81e-33 -2.22e-121 -9.95e-05	-7.9e-48 -1.09e-46 -2.49e-15 -2.11e-33 -7.56e-120 -5.01e-05	-1.9e-65 -2.77e-26 -6.27e-05 -3.76e-01 -1.75e-01 1.22e-01		
HHsearch_noss	5.41e-32 1.54e-04 5.22e-01 8.09e-09 2.65e-67 -1.84e-07	3.95e-43 2.53e-29 3.42e-05 2.77e-19 6.15e-71 9.5e-01	5.59e-54 1.08e-32 2.74e-05 4.72e-20 4.55e-67 -8.08e-01	4.1e-69 2.06e-69 7.86e-20 8.59e-56 2.56e-159 5.88e-03	7.12e-79 9.11e-73 1.82e-19 1.25e-54 6.2e-149 1.8e-02	
HHsearch_ss	-3.1e-21 -3.26e-14 -4.02e-14 -6.06e-03 8.1e-01 -2.71e-13	-1.25e-10 -8.23e-01 -5.86e-08 -1.31e-01 2.23e-01 -1.58e-02	-7.31e-02 2.76e-01 -4.79e-07 -1.15e-01 1.94e-01 -1.57e-02	5.72e-05 1.52e-22 2.19e-01 2.02e-19 7.23e-83 9.98e-01	1.12e-21 5.62e-33 3.78e-01 2.01e-18 4.21e-76 6.6e-01	-1.24e-117 -6.13e-49 -5.72e-29 -2.27e-44 -2.61e-84 -4.43e-09

Table 5 the result of 50% sensitivity t-test

5.2.3 Alignment Quality

Similar to the evaluation of homology detection, I use both reference-dependent and –independent criteria for the assessment of alignment quality.

5.2.3.1 Accuracy

Accuracy with respect to the reference alignment is defined as the ratio of the number of correctly aligned positions (NACC) to the length L of the region in the structural alignment that includes the pairs of profile positions from the alignment under evaluation.

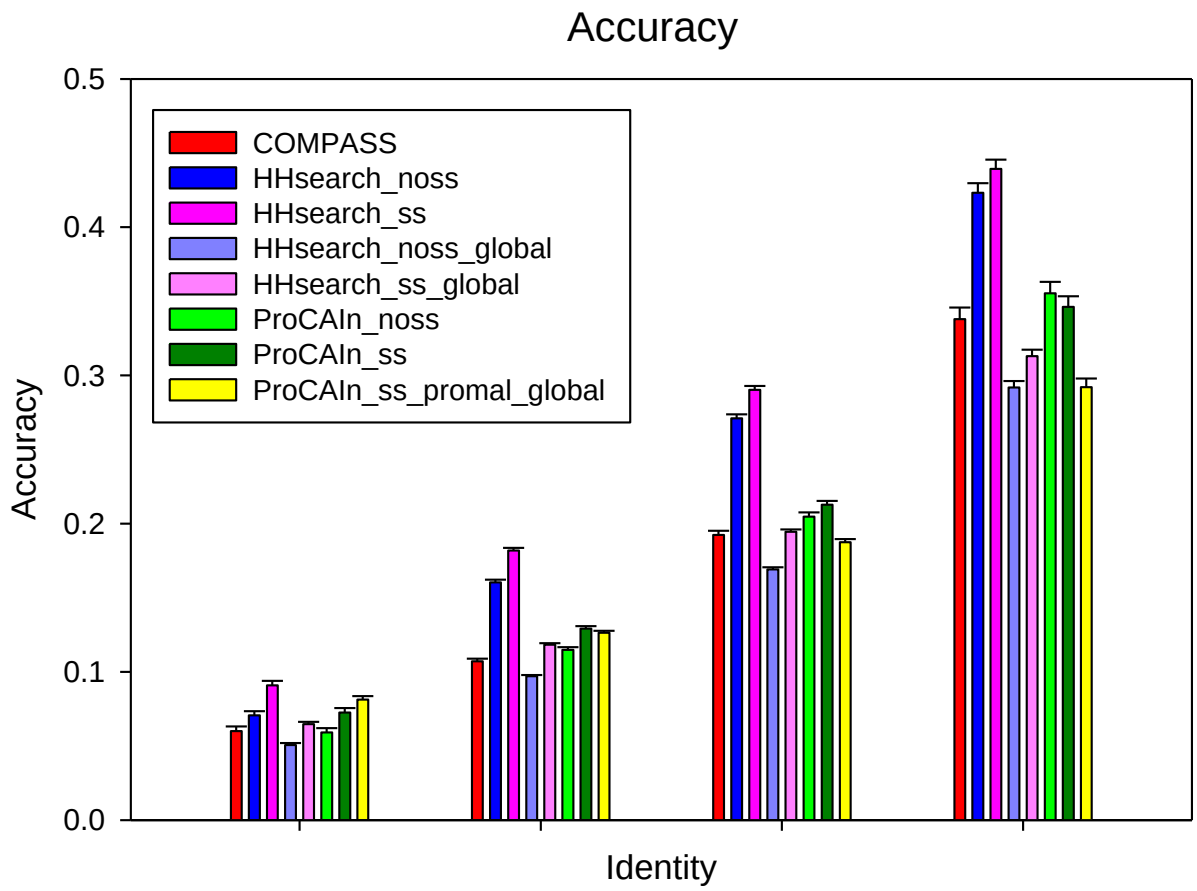

Figure 50 Accuracy of the Benchmarked Methods

PROCAIN generally produces much longer alignments with coverage 40% larger than COMPASS and almost twice larger than HHsearch (next figure). Manual inspection of alignments suggests that PROCAIN aligns the same relatively easy sequence segments as HHsearch or COMPASS, and additionally extends the alignment in both directions. These extended regions often have lower similarity and are harder to align. Lower accuracy in these regions reduces the overall alignment accuracy (previous

figure). However, the less accurate alignments that include more divergent protein parts may better reflect structural and functional protein similarities. Such alignments may be especially beneficial in structure modeling, being more informative than clear-cut yet short alignments covering only a few SS elements.

5.2.3.2 Coverage

Coverage is the ratio of the length L of the region in the structural alignment that includes all the positions from the evaluated alignment to the overall length of the structural alignment.

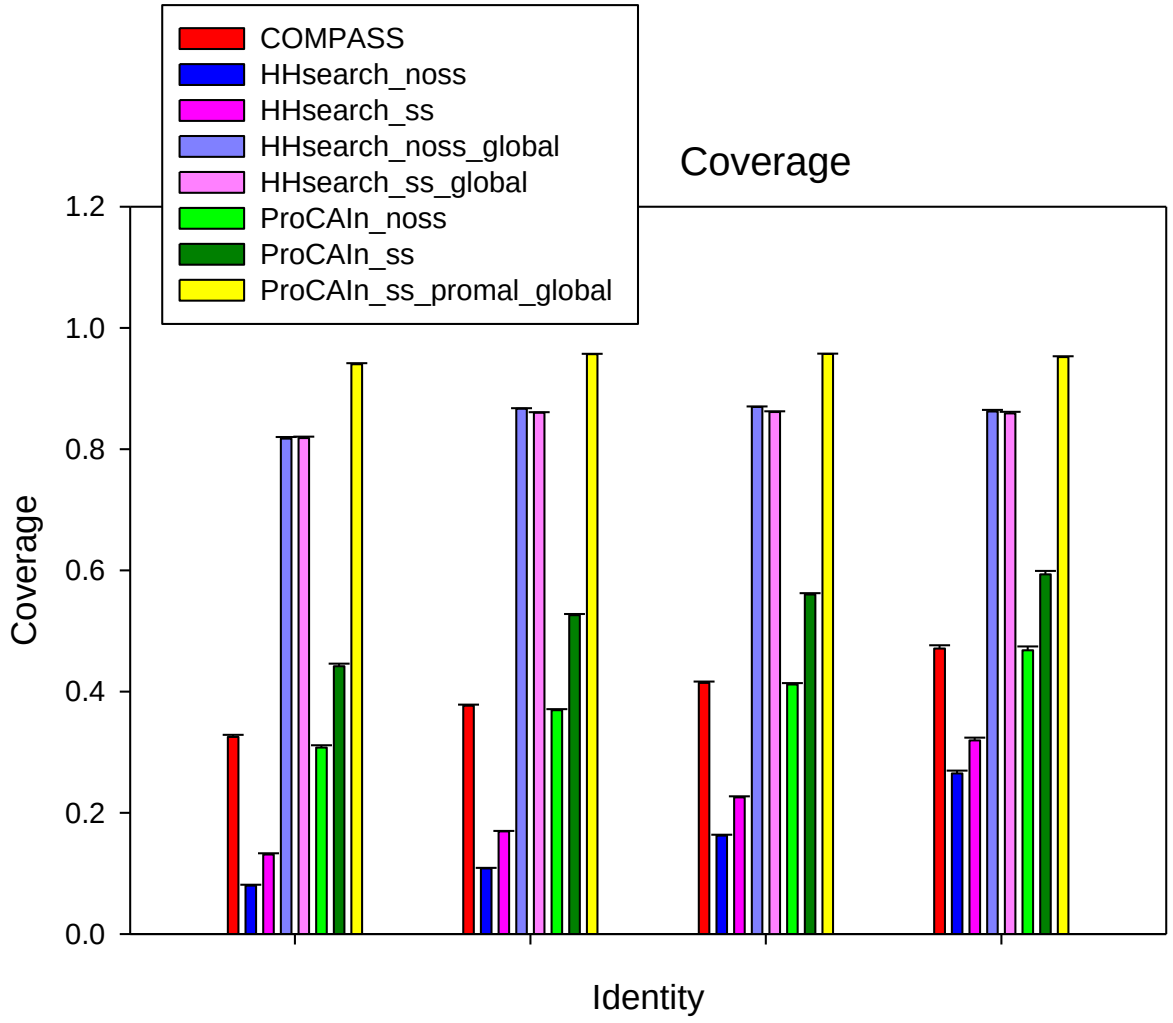

Figure 51 Coverage of the Benchmarked Methods

5.2.3.3 Q-modeler

Q-modeler is the ratio of the number of correctly aligned positions to the total number of positions in the evaluated alignment. This is to evaluate the alignment quality from the protein modeler's point of view. This measurement is close to accuracy.

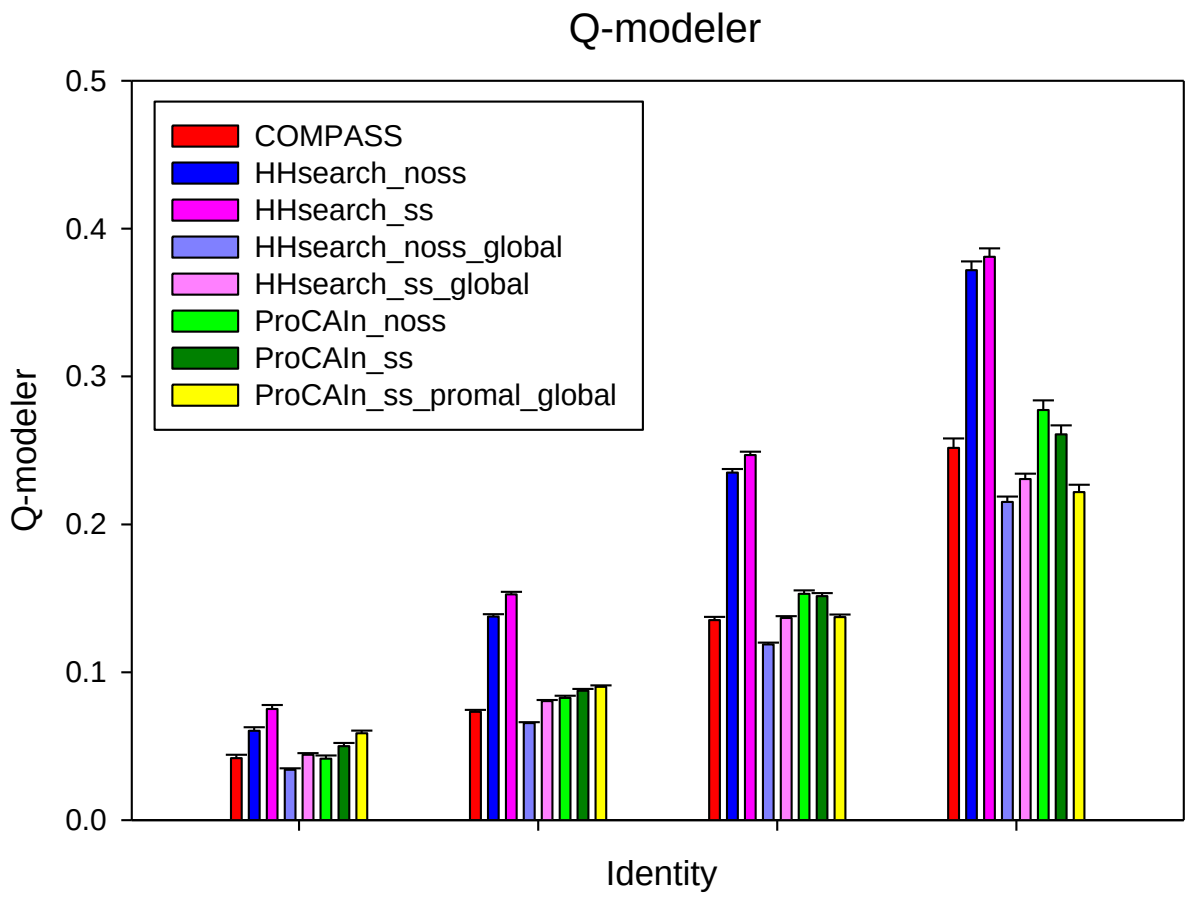

Figure 52 Q-modeler of the Benchmarked Methods

5.2.3.4 Q-developer

Q-developer is the ratio of the number of correctly aligned positions to the total number of positions in the structural alignment. This measurement evaluates the alignment quality from the protein developer's point of view.

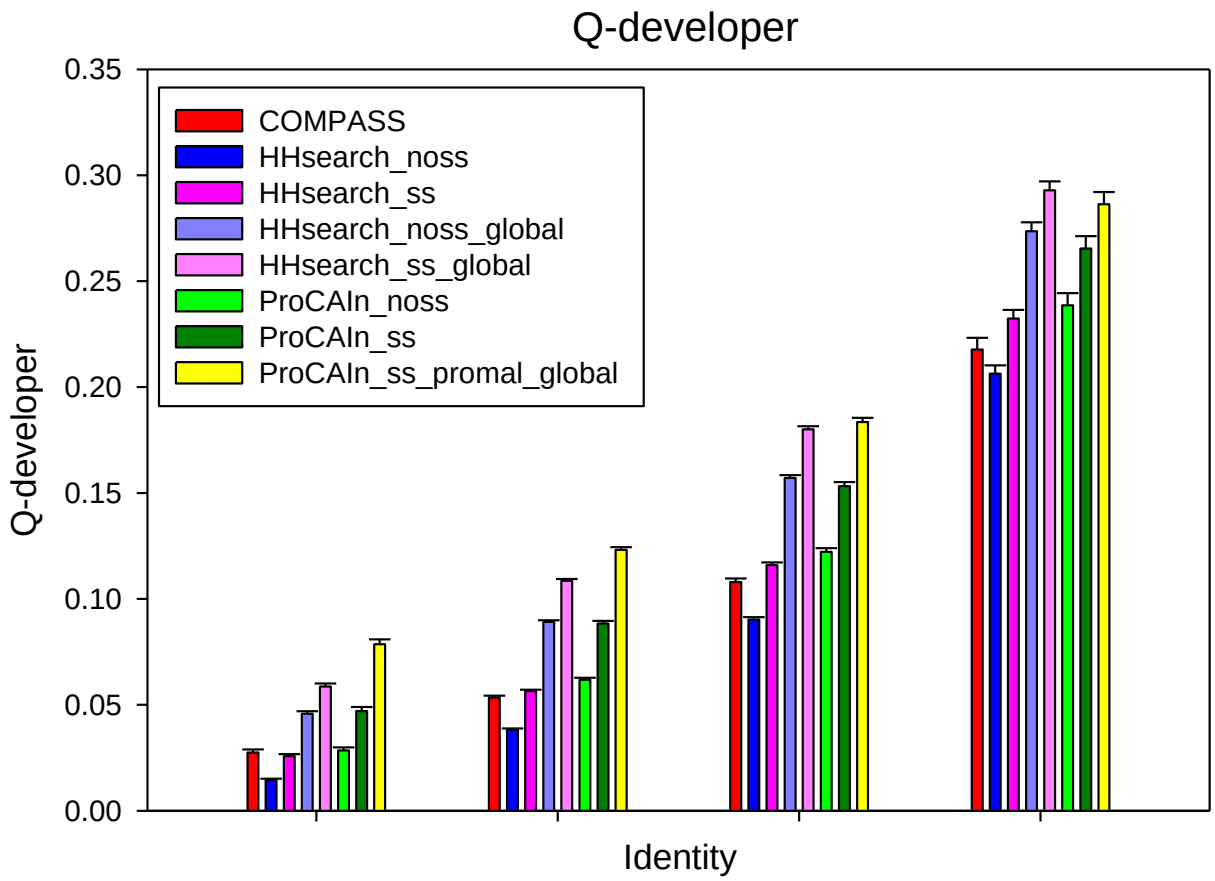

Figure 53 Q-developer of the Benchmarked Methods

5.2.3.5 Q-combined

Q-combined is the ratio of the number of correctly aligned positions to the total number of positions in the structural alignment. This measurement is the combination of Q-modeler and Q-developer.

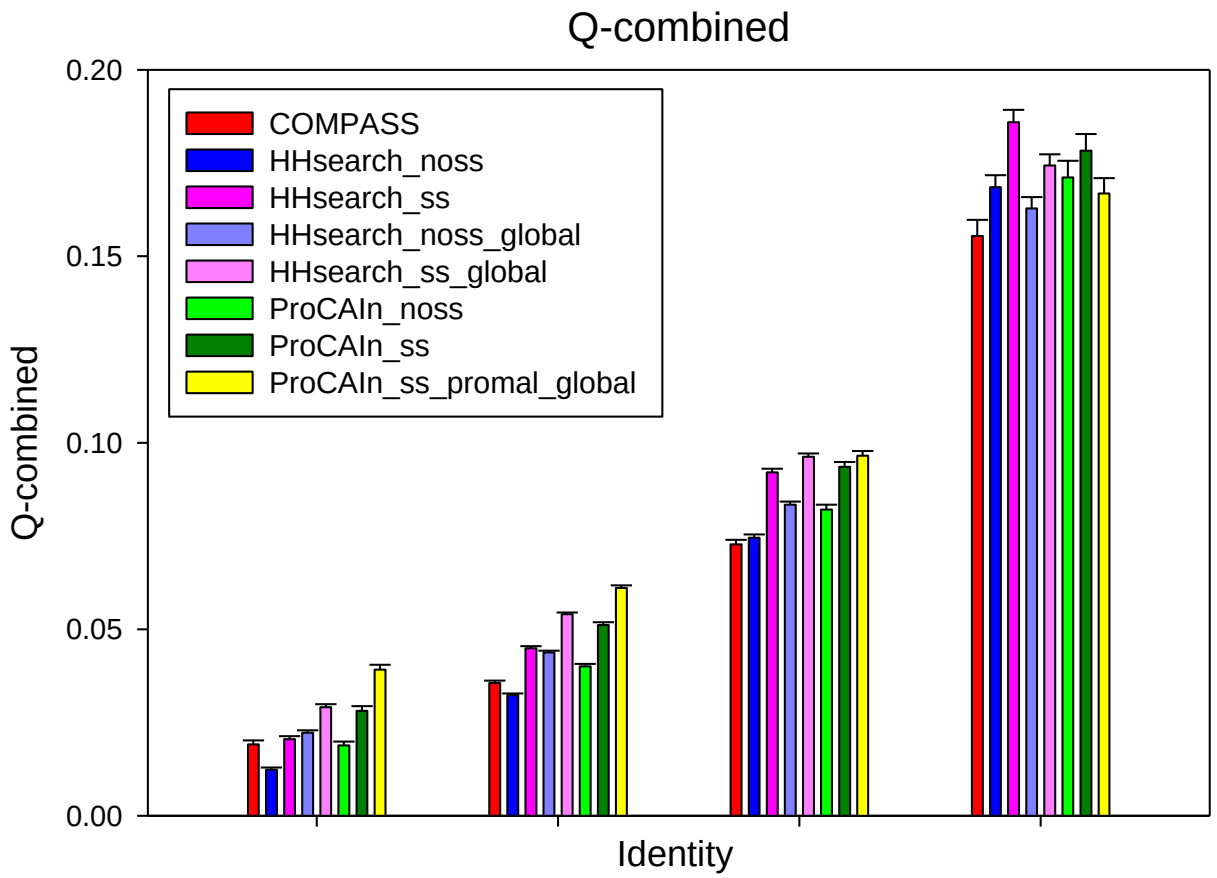

Figure 54 Q-combined of the Benchmarked Methods

5.2.3.6 Average global GDT_TS

As a reference-independent measure, I use GDT_TS of the structural superposition guided by the alignment under evaluation. I also use two slightly different ways of GDT_TS calculation. The first way calculate GDT_TS by super-imposing protein structures according to their corresponding sequence alignments and the results are shown here. The second method calculates GDT_TS by optimized protein structure alignments and the results will be shown later.

Just like I discussed, the first method favors short sequence alignment. However, ProCAIn_ss still has the best global GDT_TS values among all evaluated methods. This proves the alignment quality improvement of ProCAIn is significant.

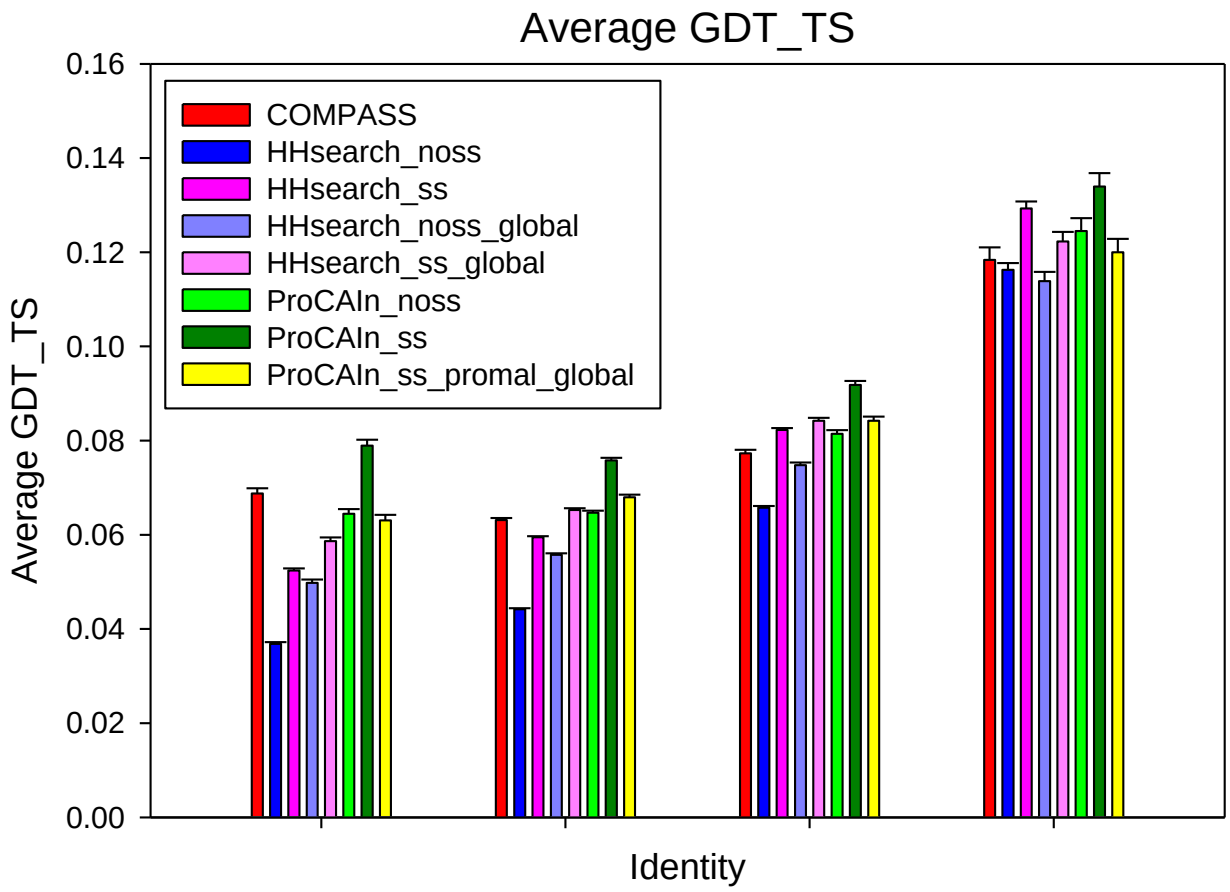

Figure 55 Average GDT_TS of the Benchmarked Methods

5.2.3.7 Average LGA global GDT_TS

This is the second method of GDT_TS calculation. This method calculates GDT_TS by optimally super-imposing protein structures. Protein sequence alignments only provide information of which segments of protein structures will be super-imposed. So this method favors longer sequence alignments, since longer sequence alignments provides longer structure segments and longer structure segments will give more correctly aligned positions even by randomness. The result of next figure also shows this trend. ProCAIn_ss_promal_global, HHsearch_noss_global and HHsearch_ss_global all scores very high with this evaluation

method. However ProCAIn_ss has similar values although ProCAIn is a local sequence alignment method. This proves that ProCAIn aligns very well.

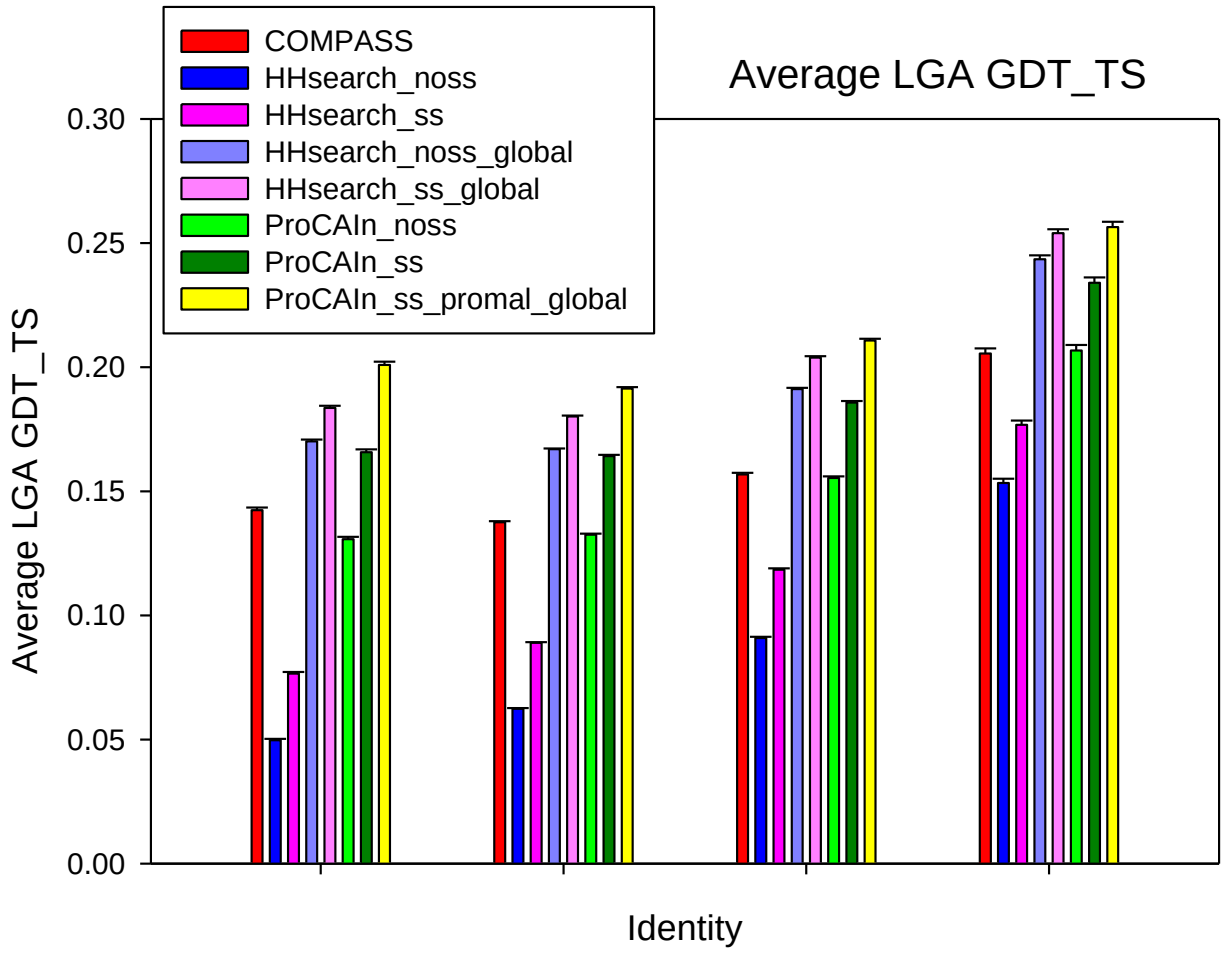

Figure 56 Average LGA GDT_TS of the Benchmarked Methods

5.3 Results with the whole dataset

ProCAIn performs very well with the testing dataset, so I proceeded to run ProCAIn and all other methods with the whole database with 4147 protein domains, to make sure that the performance improvement of ProCAIn is not a result of over optimization. The results of all tested methods with the whole dataset are very consistent with the results of these same methods with the testing dataset, this demonstrate that ProCAIn's performance improvement

comes from the fact that more types of assisting information has been involved and these types of information are helpful with protein homology detection and sequence alignment.

Since these results are very similar with the results of the testing database, so here I will not explain these results one by one.

5.4 Conclusion

It is demonstrated during the previous three chapters that the three types of assisting information: sequence motif, amino acid conservation and secondary structure, are able to improve protein homology detection performance and sequence alignment quality. I combined these three types of assisting together in this chapter and firstly proved that these three types of information are exactly the same so that it is appropriate to combine them together, then I used various evaluation methods (ROC, query family student t-test and bar graphs) to demonstrate that adding these three types of information are able to further improve ProCAIn's homology detection sensitivity and sequence alignment quality.

CHAPTER 6:

Intricate Homology Relations Detected by ProCAIn

I consider distant homology relations between SCOP domains that belong to different superfamilies but are structurally similar (GDT_TS>0.15), being confidently detected by PROCAIN (E-value < 0.01) and missed by HHsearch (HHsearch probability < 0.20). I find 405 such domain pairs in our SCOP dataset. On the other hand, approximately three times less distant relations (129 domain pairs) are detected by HHsearch (probability > 0.91, which corresponds to PROCAIN E-value of 0.01) and missed by PROCAIN (E-value > 2.13, which corresponds to HHsearch probability of 20). Full lists of these similarities are included in the end of this thesis as “List 1 ProCAIn_ss outperforms HHsearch_ss” and “List 2 HHsearch_ss outperforms ProCAIn_ss”. The considerable amounts of remote homologs uniquely detected by either of the methods reflect conceptual differences between PROCAIN and HHsearch. Thus, as is often the case in sequence analysis, a user searching for distant protein similarities would benefit from combining both methods.

Next figure shows one example of intricate homology relationships detected by PROCAIN. The nitrilase Nit domain of NIT-FHIT fusion protein from *C. elegans* (PDB ID 1emsA, domain 2, Fig. 57a) is similar to the mre11 nuclease from achreon *Pyrococcus furiosus*. (PDB ID 1ii7A, Fig. 57b), with a highly significant PROCAIN E-value of $9.90e10^{-3}$. Mre11 is a central component of a protein complex responsible for homologous recombination, telomere length maintenance, and DNA double-strand break repair in eukariotes(D'Amours and Jackson 2002).

NIT-FHIT protein is involved in purine metabolism(Pace and Brenner 2001). In vertebrates, Nit and Fhit homologs are expressed as two separate interacting proteins. Fhit is a nuleotide-binding domain strongly associated with carcinogenesis and tumor suppression (Pace and Brenner 2001), whereas the substrate and cell biology of Nit are unknown. SCOP assigns mre11 and Nit to different superfamilies within metallo-dependent phosphatase fold of α+β class (carbon-nitrogen hydrolases and metallo-dependent phosphatases, respectively), noting that these superfamilies share "some topological similarities" in structure but not establishing homology. The detected sequence similarity should have significant implications for the evolution and biology of both double-strand DNA repair and purine metabolism in eukaryotes.

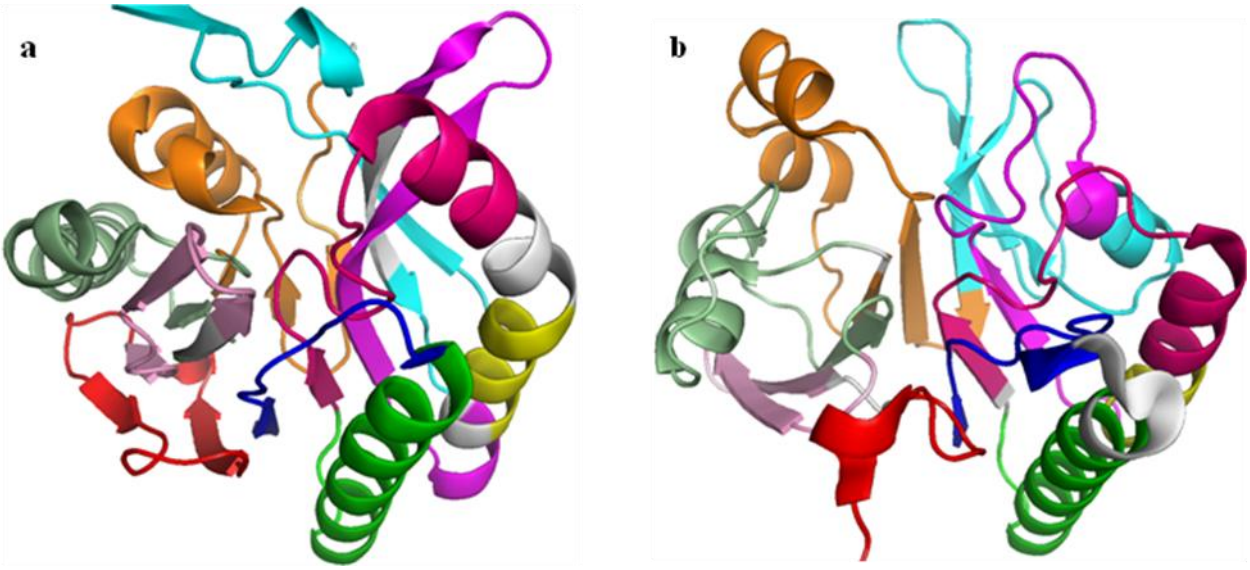

Figure 57 the First Example of Homology Relation Detected by ProCAIn

ProCAIn evalue = 9.90e-03 score = 216.25 LGA GDT_TS = 0.1779

HHsearch probability = 18.17 score = 9.21 LGA GDT_TS = 0.0698

DaliLite Z-score = 8.2

The following is the sequence alignment of this pair of protein domains, produced by ProCAIn_ss. Segments predicted as beta strand are colored in light blue and segments predicted as alpha helix are colored in red. “+” means amino acid matches and mismatches otherwise.

```
d1emsa2: Carbon-nitrogen hydrolase d1ii7a_: Metallo-dependent phosphatases

E-value = 9.90e-03 GDT_TS = 0.18

                EEEEEEEECCCCC=====HHHHHHHHHHHHHHHHHCCCC=EEEECHHHHCCCCCHHHHHH
d1emsa2  6      HFIAVCQMTSDND=====LEKNFQAAKNMIERAGEKKCE=MVFLPECFDFIGLNKNEQID
                ++++ +++++++       ++++++++++++++++++++ +++++ ++    ++++++++
d1ii7a_  4      AHLADIHLGYEQFHKPQREEEFAEAFKNALEIAVQENVDFILIAGDLFHSSRPSPGTLKK
                EEEECHHCCCCCCCCHHHHHHHHHHHHHHHHHHHHHCCCCEEEEECCEEECCCCCHHHHHH


                HHHHHCCHHHHHHHHHHCCCCEEEEECCEECCCCCCCCEEEEEEEEECCCCCCCCCCCCC
d1emsa2         LAMATDCEYMEKYRELARKHNIWLSLGGLHHKDPSDAAHPWNTHLIIDSDGVTRAEYNKL
                ++     ++++++++  ++++++++++   +++++++ +      +++  + ++   ++
d1ii7a_         AI=====ALLQIPKE==HSIPVFAIEGNHDRTQRGPSVLN=====LLEDFGLVYVIGMRK
                HH=====HHHHHHHHH==CCCEEEEEECCCCCCCCCCCHHH=====HHHHCCCEEEECCCC


                EECCCCCCCCCCCEEEEEEECCCCCCCEECCCCCCCCCCHHHHHHHHHHHHHHHHCCCCE
d1emsa2         HLFDLEIPGKVRLMESEFSKAGTEMIPPVDTPIGRLGLSICYDVRFPELSLWNRKRGAQL
                  ++++++++++++++++++  +++ +++++++      ++++ +++++++ ++++++
d1ii7a_         EKVENEYLTSERLGNGEYLVKG==VYKDLEIHGMKYMSSAWFEANKEILKRLFRPTDNAI
                CCCCCCCCCCCEEECCEEEEEEC==CCCCCCCCCCCCCCHHHHHHHHHHHHHHHHCCCCCEE


                EEE=ECCCCCCCCHHHHHHHHHHHHHHCCCEEEEECCCCCCCCCCEEECCCEEEE===CCC
d1emsa2         LSF=PSAFTLNTGLAHWETLLRARAIENQCYVVAAAQTGAHNPKRQSYGHSMVV===DPW
                +++  +  ++++++++++ + ++ +  ++ +++++ +  ++++++ ++++ ++   ++
d1ii7a_         LMLHQGVREVSEARGEDYFEIGLGDLPEGYLYYALGHI==HKRYETSYSGSPVVYPGSLE
                EEEECCCCCCCCCCCCCCCCHHHHHHCCCCCCEEEECCC==CCCEEECCCCCEEEECCCCC


                CCEEEECCCCCEEEEEEECHHH
d1emsa2         GAVVAQCSERVDMCFAEIDLSY
                +  +++++++++++++++ ++
d1ii7a_         RWDFGDYEVRYEWDGIKFKERY
                CCCCCHHCCCCCEEEEEEECCC
```

As another example, PROCAIN predicts homology (with E-value = $2.97e10^{-3}$) between two bacterial all-α proteins: processive endocellulase CelF from *Clostridium cellulolyticum* (PDB ID 1g9gA, Fig. 58a) and squalene-hopene cyclase from *Alicyclobacillus acidocaldaris* (PDB ID 2sqcA, domain 1, Fig. 58b). These domains share a significant structure similarity (DALI Z-score = 16.7) yet belong to different SCOP superfamilies: six-hairpin glycosidases and terpenoid cyclases/protein prenyltransferases, respectively. CelF is a component of cellulosome, protein complex responsible for the degradation of cellulose and similar substrates outside the cell. Squalene-hopene cyclase is a membrane protein with the active site located in a large central cavity (Wendt, Poralla et al. 1997; Full and Poralla 2000). The detected homology between these domains may suggest a similar functional role of internal cavity in enzymatic activity of CelF.

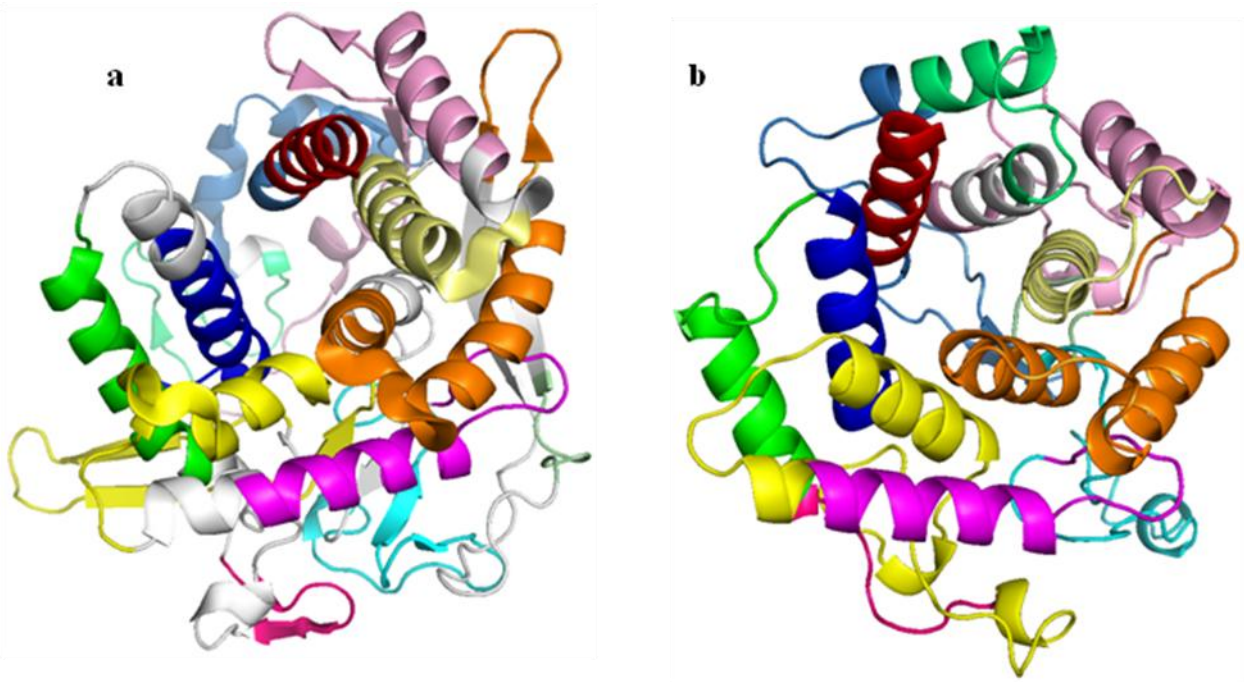

Figure 58 the Second Example of Homology Relation Detected by ProCAIn

ProCAIn e-value = 2.97e-03 score = 186.68 LGA GDT_TS = 0.2911

HHsearch probability = 3.34 score = -1.73 LGA GDT_TS = 0.1282

DaliLite Z-score = 16.7

The following is the sequence alignment of this pair of protein domains, produced by ProCAIn_ss.

```
d1g9ga_: alpha/alpha toroid d2sqca1: alpha/alpha toroid
E-value = 2.97-03 GDT_TS = 0.29

                 CCEEEECCCCCHHHCCCCCCCCCHHHHHHHHHHHHHHHHHHHCCCHHHHHHHHHHHHHH
d1g9ga_ 208      FLDLFTKDTGTPAKQFKYTNAPDADARAVQATYWADQWAKEQGKSVSTSVGKATKMGDYL
                 ++ +  +++++  +++  +   ++ + +++++  +         ++ +  ++++++++++
d2sqca1 14       YLLSCQKDEGYWWGPLXISPVWDT=GLAVLALRAA=========GLPADHDRLVKAGEWL
                 HHHHHCCCCCCCCCCCCCCHHHHH=HHHHHHHHHC=========CCCCCHHHHHHHHHHH


                 HHHHHHHHHHHHCCCCCCCCCHHHHHHHHHHHHHHHCCCCCCCEEEEECCCCCCCCCCCCH
d1g9ga_          RYSFFDKYFRKIGQPSQAGTGYDAAHYLLSWYYAWGGGIDSTWSWIIGSSHNHFGYQNPF
                 +        +++ ++++             +        ++++++  ++++        +
d2sqca1          L=======DRQITVPGD=============WAVKRPNLKPGGFAFQFDNVYYPDVCDTAV
                 H=======HCCCCCCCC=============CCCCCCCCCCCCCCCCCCCCCCCCCCHHHH


                 HEEEECCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCC
d1g9ga_          AAWVLSTDANFKPKSSNGASDWAKSLDRQLEFYQWLQSAEGAIAGGATNSWNGRYEAVPS
                 +  ++   +  +++        ++++++++++ +++++++++     ++  + + +
d2sqca1          VVWALNTLRLPDERRR======RDAMTKGFRWIVGMQSSNGGWG=====AYDVDNTSDLP
                 HHHHHHHCCCCCCCHH======HHHHHHHHHHHHHHCCCCCCCC=====CCCCCCCCCCC


                 CCCCHHCCCCCCCCCCCCCCCCEEEEECCCHHHHHHHHHHCCCHHHHHHHHHHHHHHC
d1g9ga_          GTSTFYGMGYVENPVYADPGSNTWFGMQVWSMQRVAELYYKTGDARAKKLLDKWAKWING
                 +   + ++     + + +             + ++++++ + + +++++++++ ++++++
d2sqca1          NHIPFSDFGEVTDPPSED============VTAHVLECFGSFGYDDAWKVIRRAVEYLKR
                 CCCCHHHHHHCCCCCHHH============HHHHHHHHHHCCCCCHHHHHHHHHHHHHHH


                 CCEECCCCCEEECCCCEECCCCCCCCCCCCCCCCCCCEEEEEECCCCCHHHHHHHHHHHH
d1g9ga_          EIKFNADGTFQIPSTIDWEGQPDTWNPTQGYTGNANLHVKVVNYGTDLGCASSLANTLTY
                 +   ++++ +                      ++ +++         + ++++++++++
d2sqca1          EQ==KPDGSWF====================GRWGVN========YLYGTGAVVSALKA
                 HH==CCCCCCC====================CCCCCC========HHHHHHHHHHHHH


                 HHHHCCCHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCHHHHHCCCCEEECCC=======
```

```
d1g9ga_         YAAKSGDETSRQNAQKLLDAMWNNYSDSKGISTVEQRGDYHRFLDQEVFVPAG=======
                +    ++    ++++++++ +++  +++ + +   +     ++++++   +
d2sqca1         VGIDTRE====PYIQKALDWVEQHQNPDGGWGEDCRSYEDPAYAGKGASTPSQTAWALMA
                HCCCCCC====HHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHH


                =CCCCCCCCCCCCCCCCEEECCHHHHCCCCHHHHHHHHHCCCCCE===EEEEEHHHHHHH
d1g9ga_         =WTGKMPNGDVIKSGVKFIDIRSKYKQDPEWQTMVAALQAGQVPT===QRLHRFWAQSEF
                   + +++ +++  ++ +  +    +++ +++  ++ + +   +    ++++ +  ++
d2sqca1         LIAGGRAESEAARRGVQY==LVETQRPDGGWD==EPYYTGTGFPGDFYLGYTMY=RHVFP
                HHHCCCCCCHHHHHHHHH==HHHHCCCCCCCC==CCCCEECCCCCCEEECCCCC=HHHHH


                HHHHHHHHHH
d1g9ga_         AVANGVYAIL
                ++++++++++
d2sqca1         TLALGRYKQA
                HHHHHHHHHH
```

CHAPTER 7:

Discuss and Future Research

7.1 Contribution of SS prediction

Similar to others (Chung and Yona 2004; Soding 2005), we find that considering SS prediction leads to significant improvement in both similarity detection and alignment accuracy. As expected, this improvement is more pronounced for extremely distant homologs, where direct sequence signals are weak yet SS is conserved. SS prediction itself(McGuffin, Bryson et al. 2000) involves the analysis of various types of information derived from sequence profiles: periodic patterns of hydrophobicity, residue propensities for occurrence in SS elements, specific sequence motifs etc. Thus, for the purposes of homology detection, similarity between SS predictions, regardless of their accuracy, may be considered as a simplistic representation of 'horizontal' sequence patterns in the compared protein families. After testing different ways of including SS predictions in profile comparison, we find that the best performance results from a simple addition of weighted substitution score for SS types. The optimal weight value , w_{ss} = 0.1, appears to be similar to that used in HHsearch(Soding 2005), suggesting that this might be a general optimal ratio of mixing residue and SS information.

7.2 Contribution of additional non-SS features

Although the comparison of SS predictions is a major contributor to the increased quality of homology detection, it does not dominate the improvement as much as reported for

HHsearch, a conceptually similar method based on the comparison of hidden Markov models (HMM) (Soding 2005). Interestingly, inclusion of simple profile features (positional conservation and the presence of ungapped segments in profile alignment), as well as the new protocol of statistical estimation, results in the performance comparable to that of HHsearch with SS included. HHsearch(Soding 2005) is based on HMM-HMM comparison allowing for flexible gap penalties in alignment construction, and is considered among the best performing methods for homology detection. We find that a similar detection quality can be achieved by a simpler profile aligner with fixed gap penalties and no SS consideration. Addition of SS improves the quality of PROCAIN detection further, beyond the previously achievable levels. The simplicity of profile-profile comparison makes it more tractable for analyzing contribution of different score terms and procedures, providing potentially easier platform for finding directions of major improvement. However, evaluation of the effects of additional PROCAIN procedures on HMM comparison would be extremely interesting.

An important PROCAIN feature that differs from previously reported methods is the score that rewards clusters of positive matches in continuous motifs but does not penalize for their absence. In such a cluster, each positional match receives additional score input from neighboring matches. This scheme boosts the importance of longer stretches of similar sequence positions, which are typical in homologs, and smooth the scores within a stretch, so that the signals from extremely conserved positional matches are additionally distributed over their closest neighbors.

7.3 E-value estimation based on symmetrical calibration

A significant contribution to PROCAIN performance comes from the new approach to the estimation of statistical significance of detected similarities. In our symmetrical calibration scheme, the background score distributions are derived for both query and its database counterparts. When used as queries, different profiles are known to differ in the heaviness of the tail of random score distribution: the same score value may be quite significant for one query and marginal for another. These differences are caused by variations in profile properties, some of which are easier to model separately (length, sequence diversity), whereas others are more difficult (residue composition, SS content, etc.) In the same fashion, profiles in the searching database have different propensity to appear as highly scored matches when compared to an unrelated query. Thus, a random model of individual comparison between a query and a database profile would be more accurate if the background distributions for both query and subject are considered. Our scheme does not affect computational speed of the search, since all distributions for the database profiles are pre-computed and analytically approximated in advance. Given the power of today's computational resources, building distributions based on comparisons of unrelated entries in the search database is feasible and may be beneficial for various other search applications.

7.4 Homology detection in protein classes

PROCAIN performs differently in different major protein classes. Results of evaluation of homology detection quality within the main SCOP classes (all α, all β, α/β, and α+β) can be found in next plots. PROCAIN performance in the α/β class is very similar to the overall

performance, whereas other three classes show significant differences. Similar yet somewhat smaller differences are observed for HHsearch (see SI Figure S3, S4, S5 and S6). We hypothesize that these differences may reflect the composition of training set that is used to optimize the weights (w_c, w_{ss}, w_m) of additional terms in PROCAIN score. This set consists of domains randomly chosen from the total evaluation set, and therefore shows a similar distribution of representatives among the main classes. As the protein world in general, this set is dominated by the homologs from α/β class (47.9%), whereas all α, all β, and α+β classes are less represented (17.6%, 9.6%, and 8.9%, respectively).

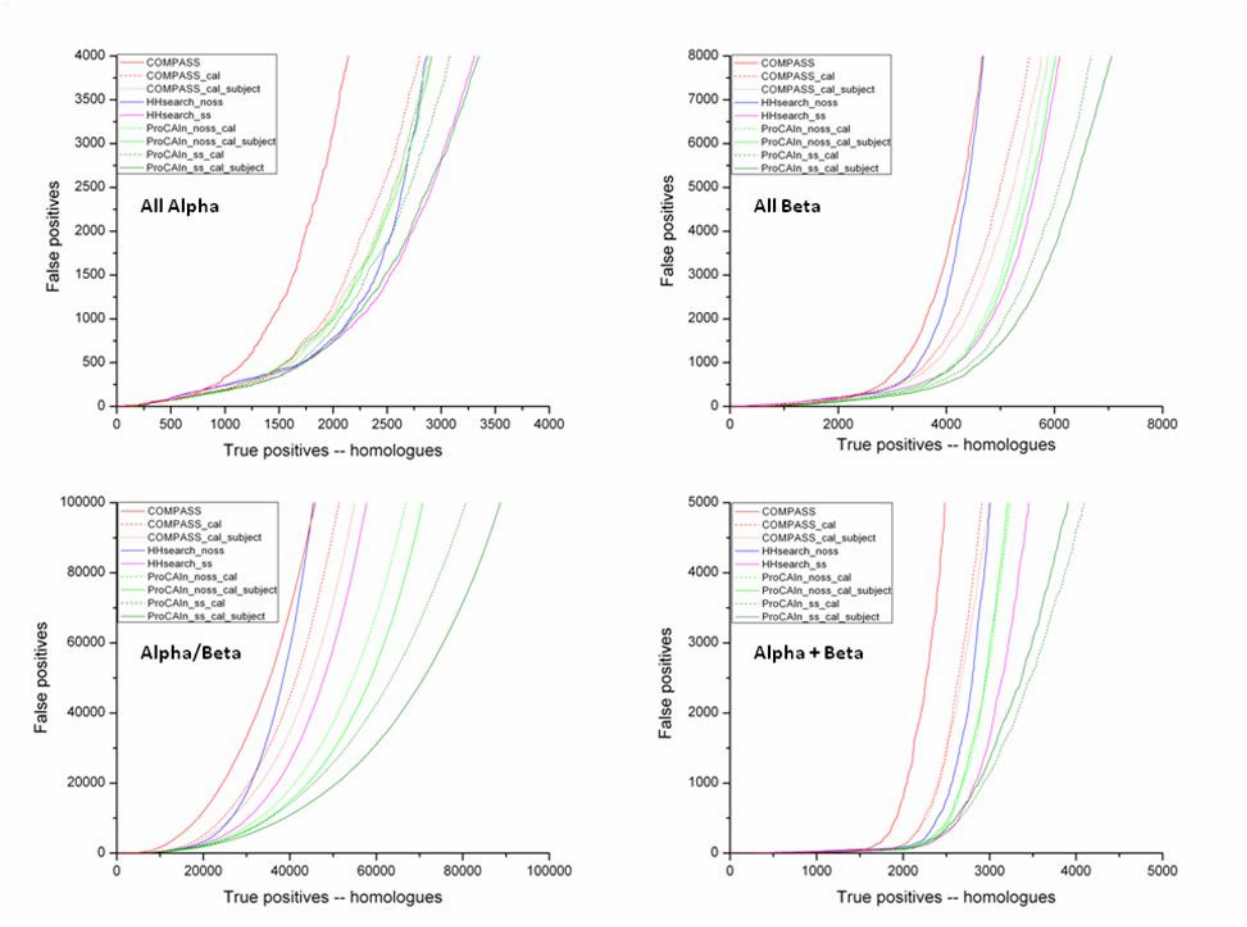

Figure 59 Protein Homolog Detection Performance in Protein Class

7.5 Future research

The observed difference in performance suggests that adjustment of scoring parameters according to the query's class may be a plausible further direction to increase the detection quality. For example, for all α or all β proteins, the improvement introduced by considering SS are smaller compared to the whole set . Indeed, a SS prediction string that consists mainly of a single SS type bears less additional information for an aligner than a string with clearly delimited SS elements of different types. Therefore, in all α and all β proteins, using lower relative weight for SS score may put more emphasis on the direct amino acid similarity, which might be more important to detect.

Results with the Whole Database

i. Protein homology detection

1. Reference dependent evaluation with SCOP superfamily relationship and SVM score

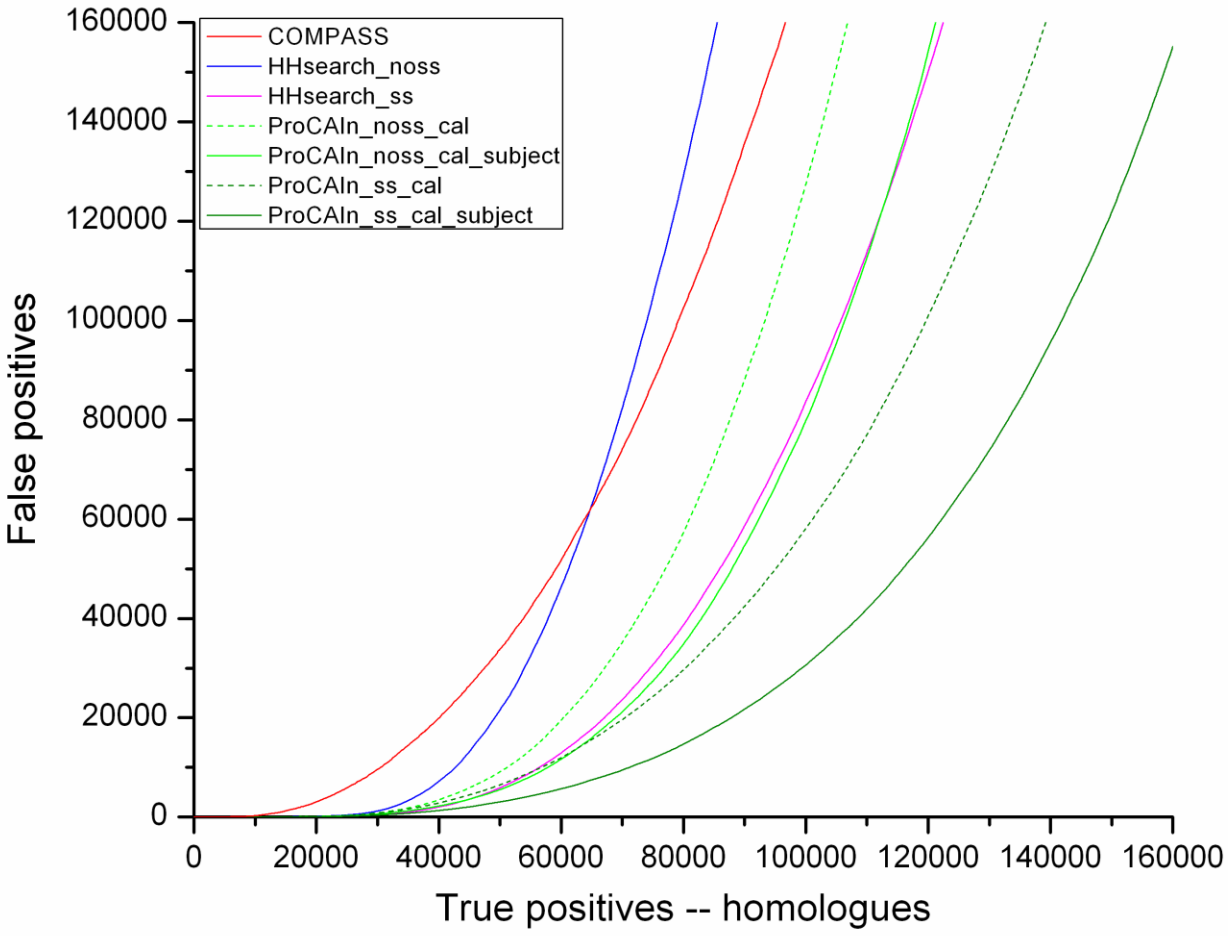

Figure 60 the whole dataset result of reference dependent evaluation with SCOP superfamily relationship and SVM score

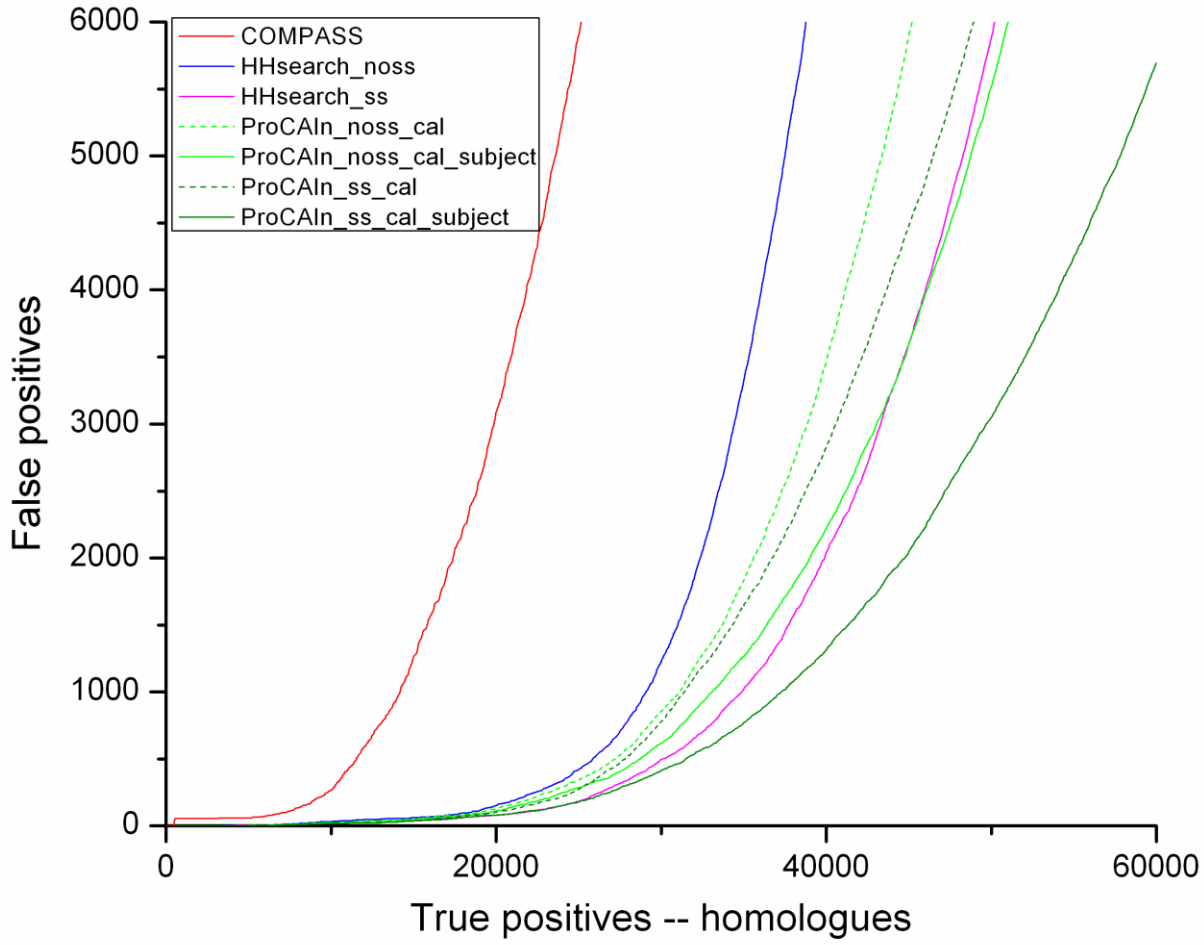

Figure 61 a zoom-in plot of the whole dataset result of reference dependent evaluation with SCOP superfamily relationship and SVM score

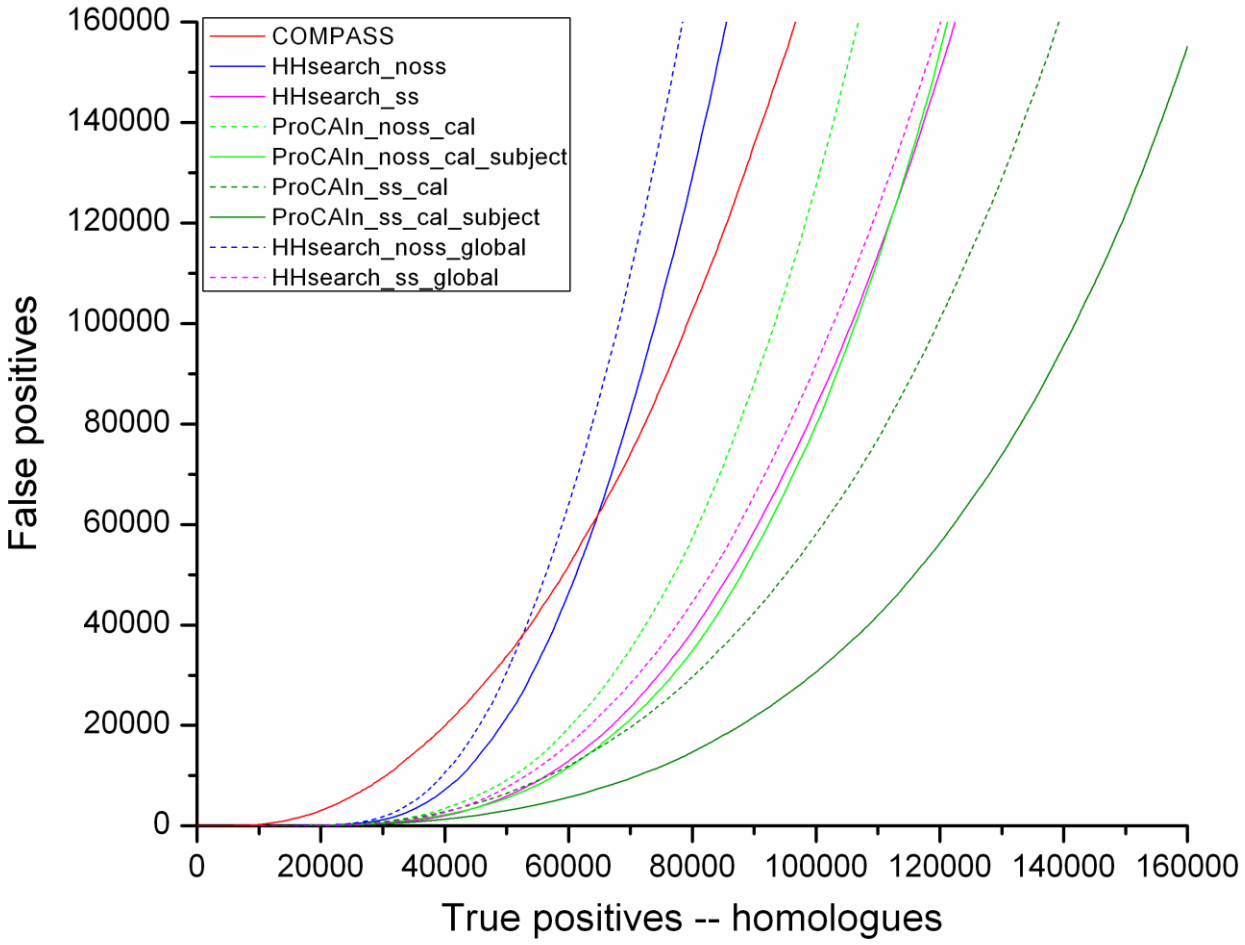

Figure 62 the whole dataset result of reference dependent evaluation with SCOP superfamily relationship and SVM score (with global results)

2. Reference dependent evaluation with SCOP superfamily relationship only

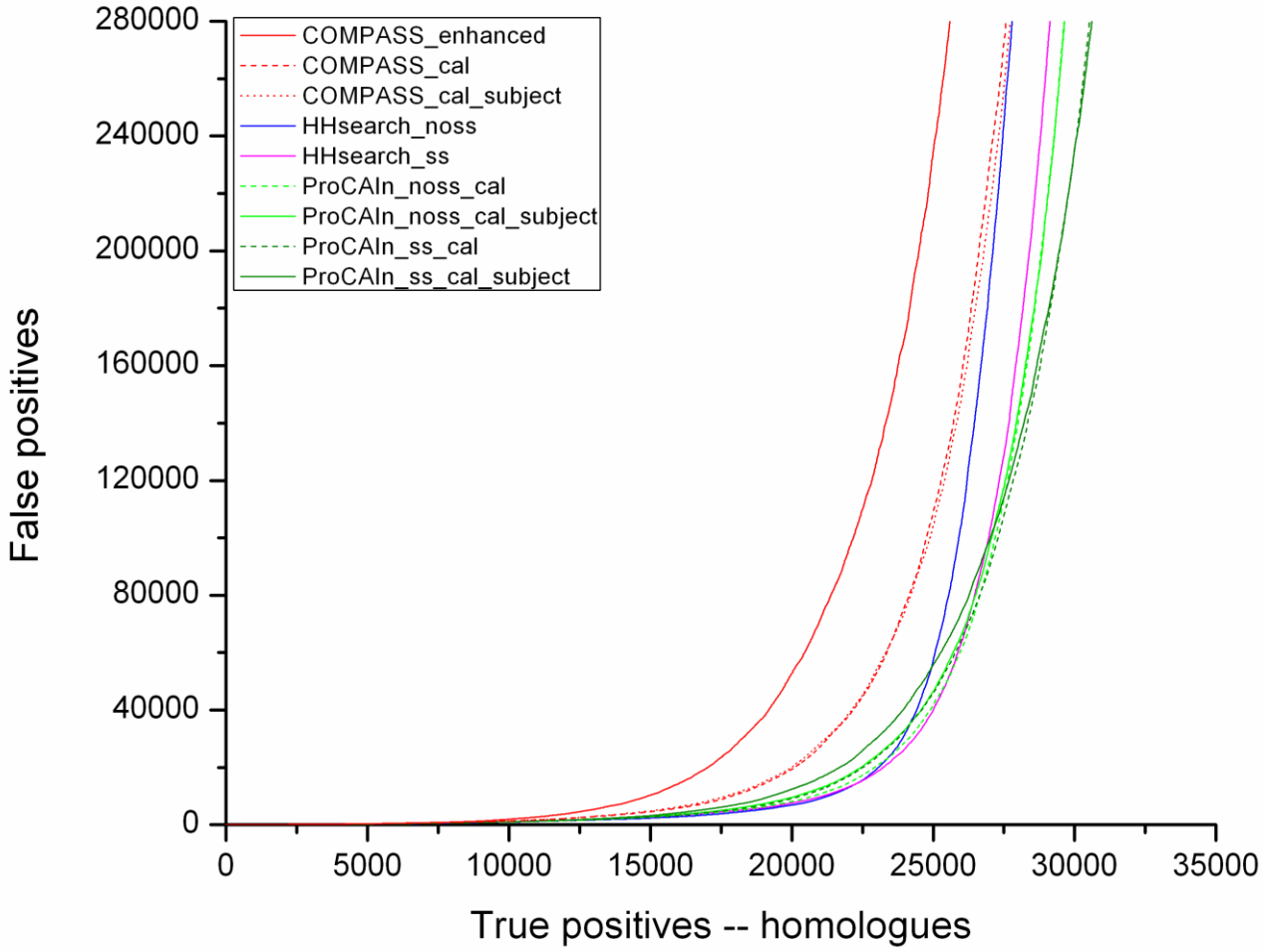

Figure 63 the result of reference dependent evaluation with SCOP superfamily relationship only

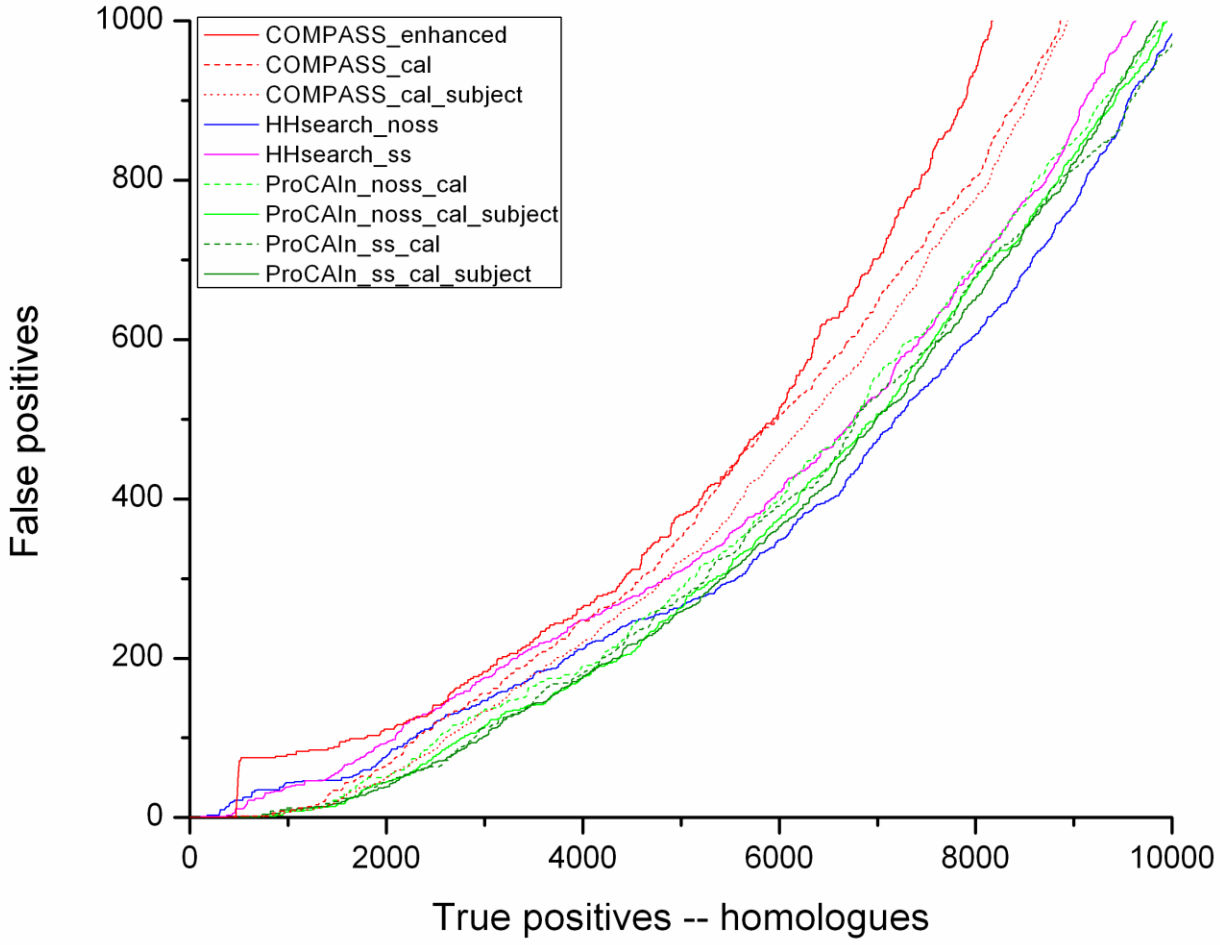

Figure 64 a zoom-in plot of the result of reference dependent evaluation with SCOP superfamily relationship only

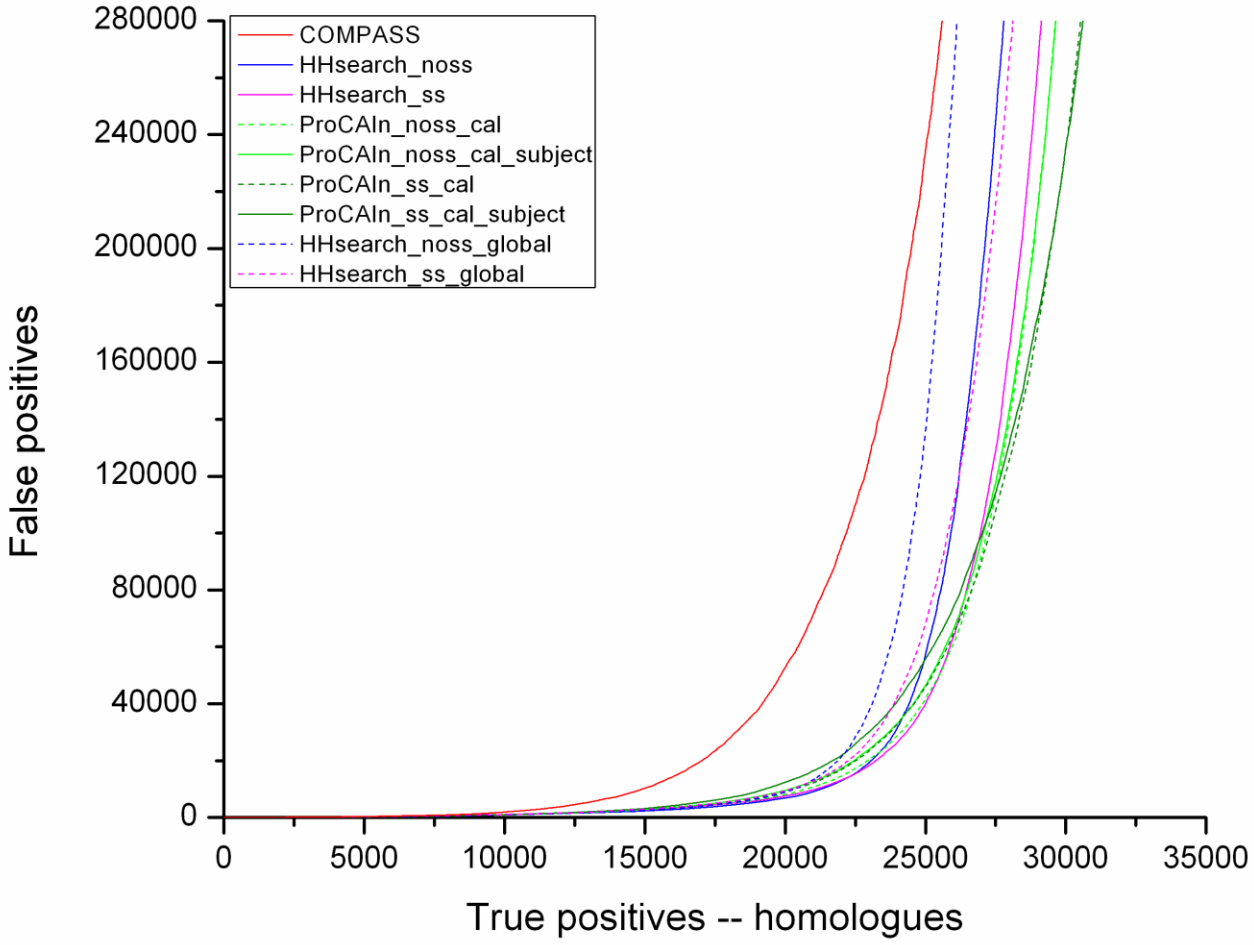

Figure 65 the result of reference dependent evaluation with SCOP superfamily relationship only (with global results)

3. Reference dependent evaluation with SCOP superfamily relationship and SVM score

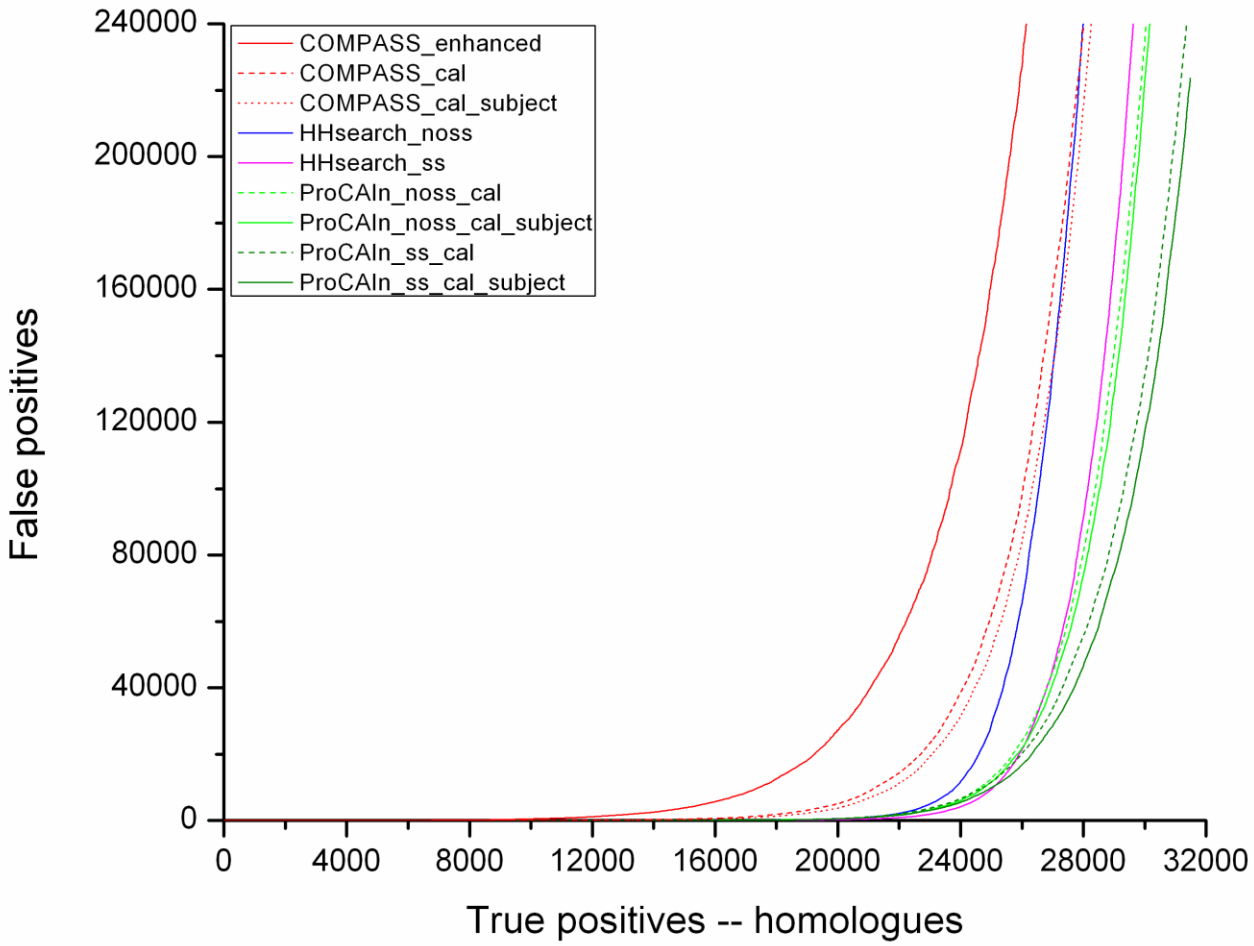

Figure 66 the result of reference dependent evaluation with SCOP superfamily relationship and SVM score

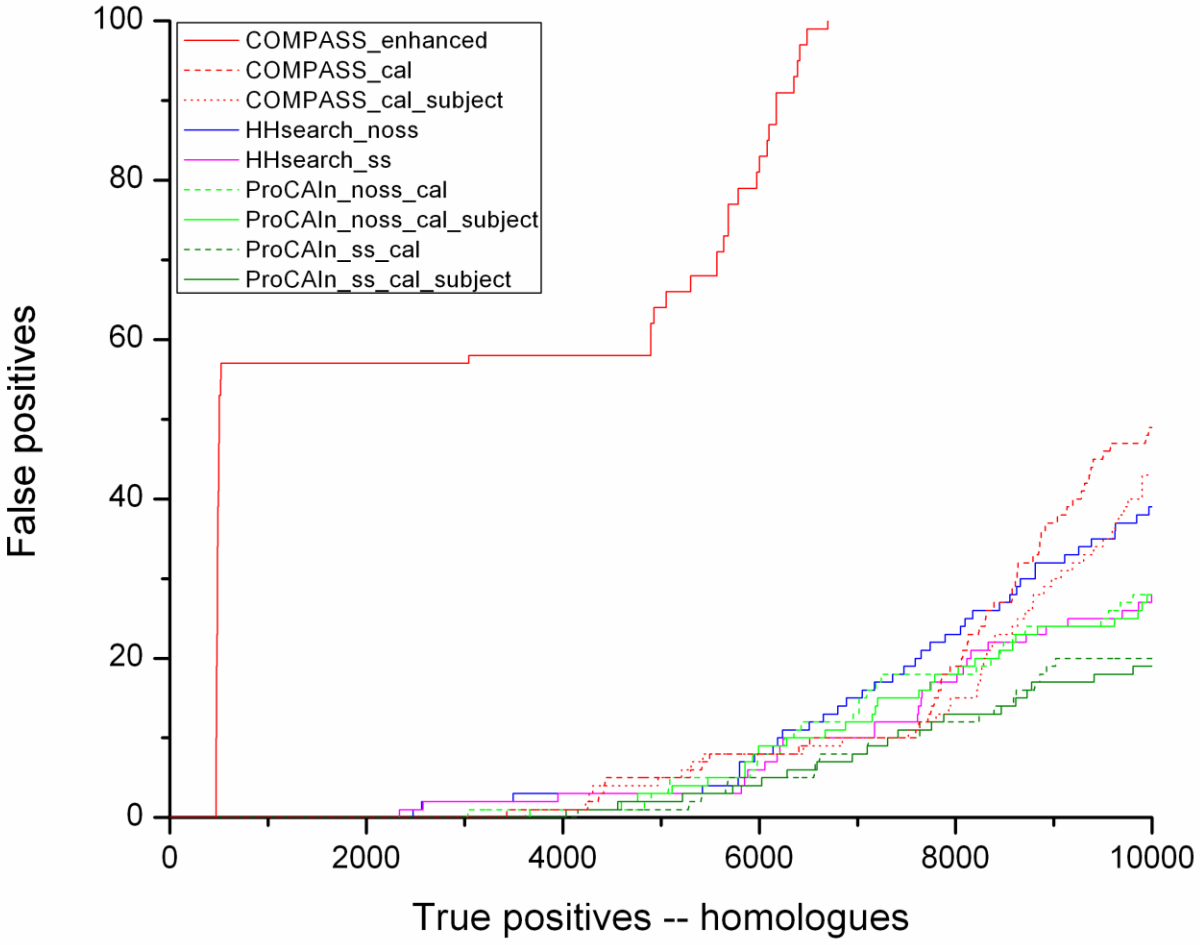

Figure 67 a zoom-in of the result of reference dependent evaluation with SCOP superfamily relationship and SVM score

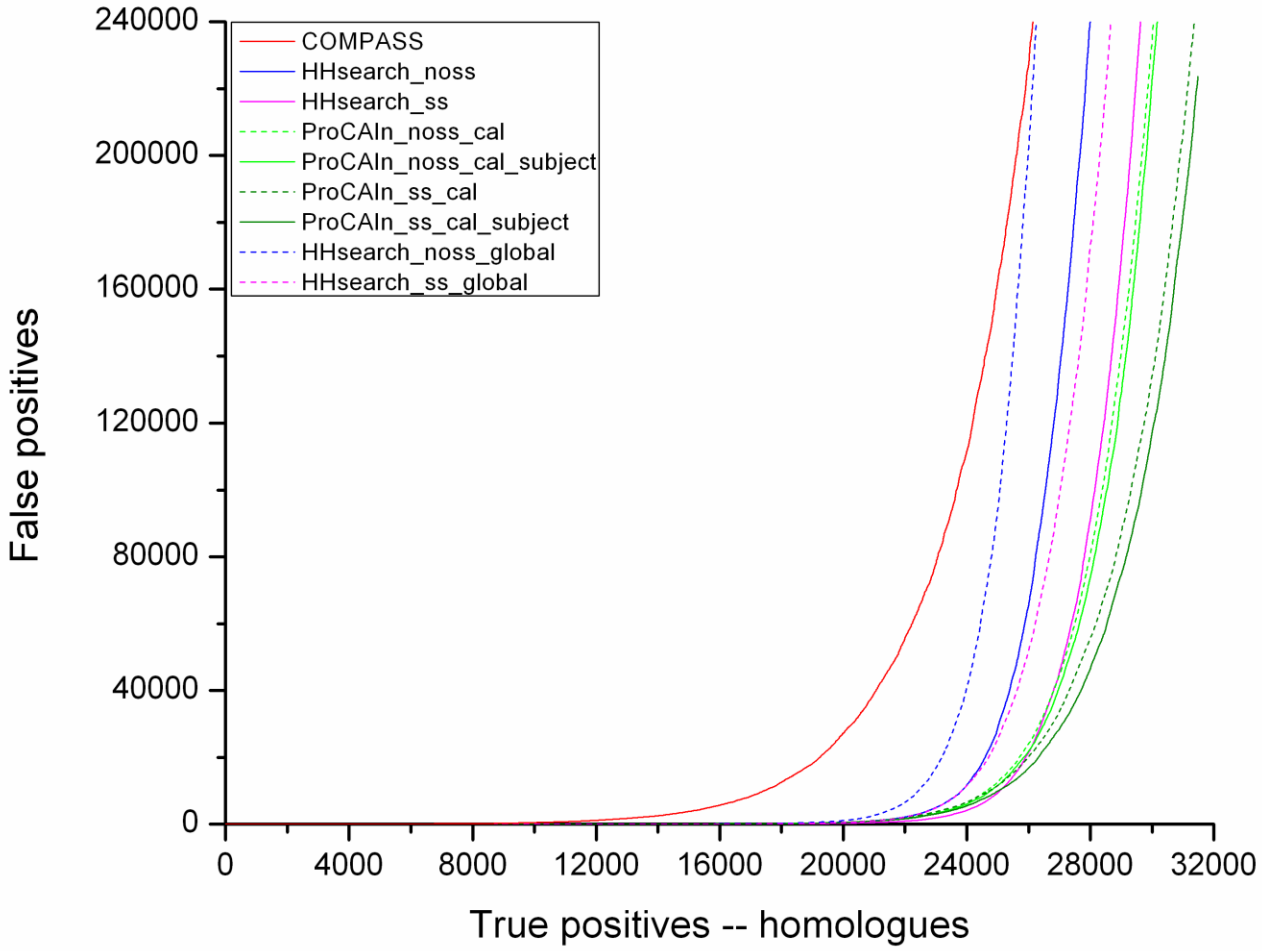

Figure 68 the result of reference dependent evaluation with SCOP superfamily relationship and SVM score (with global results)

4. Reference dependent evaluation with SCOP superfamily relationship, SVM score and alignment quality

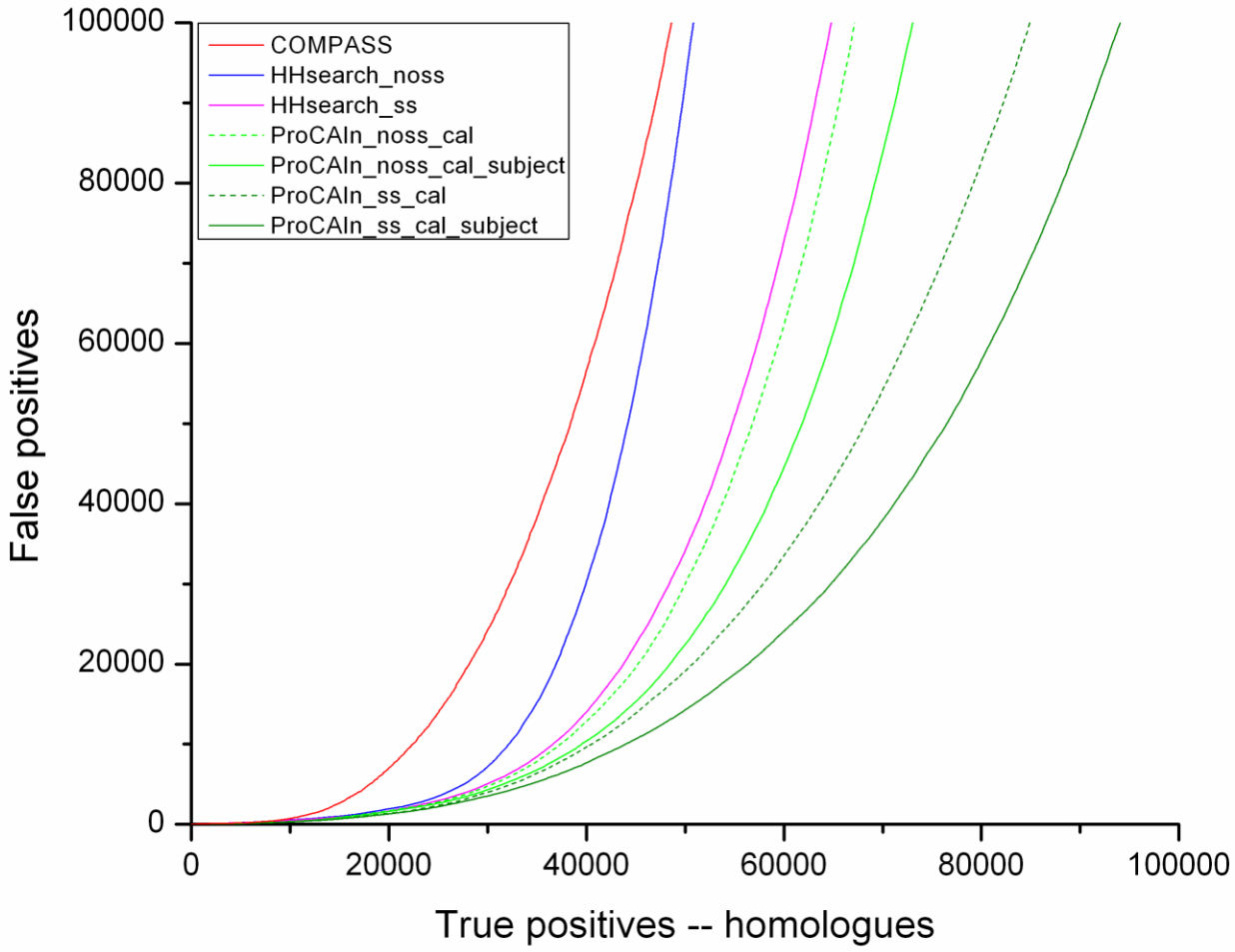

Figure 69 the results of reference dependent evaluation with SCOP superfamily relationship, SVM score and alignment quality

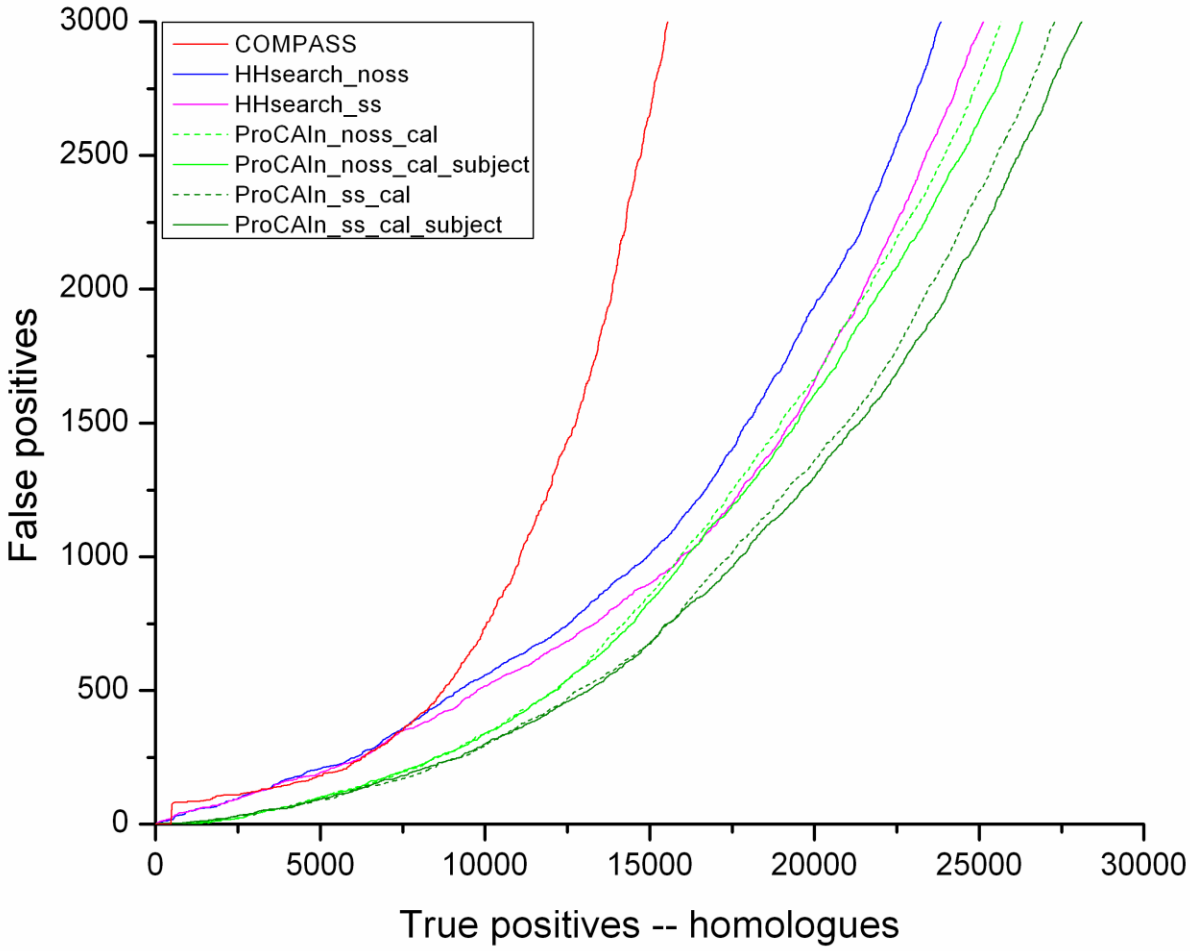

Figure 70 a zoom-in plot of the results of reference dependent evaluation with SCOP superfamily relationship, SVM score and alignment quality

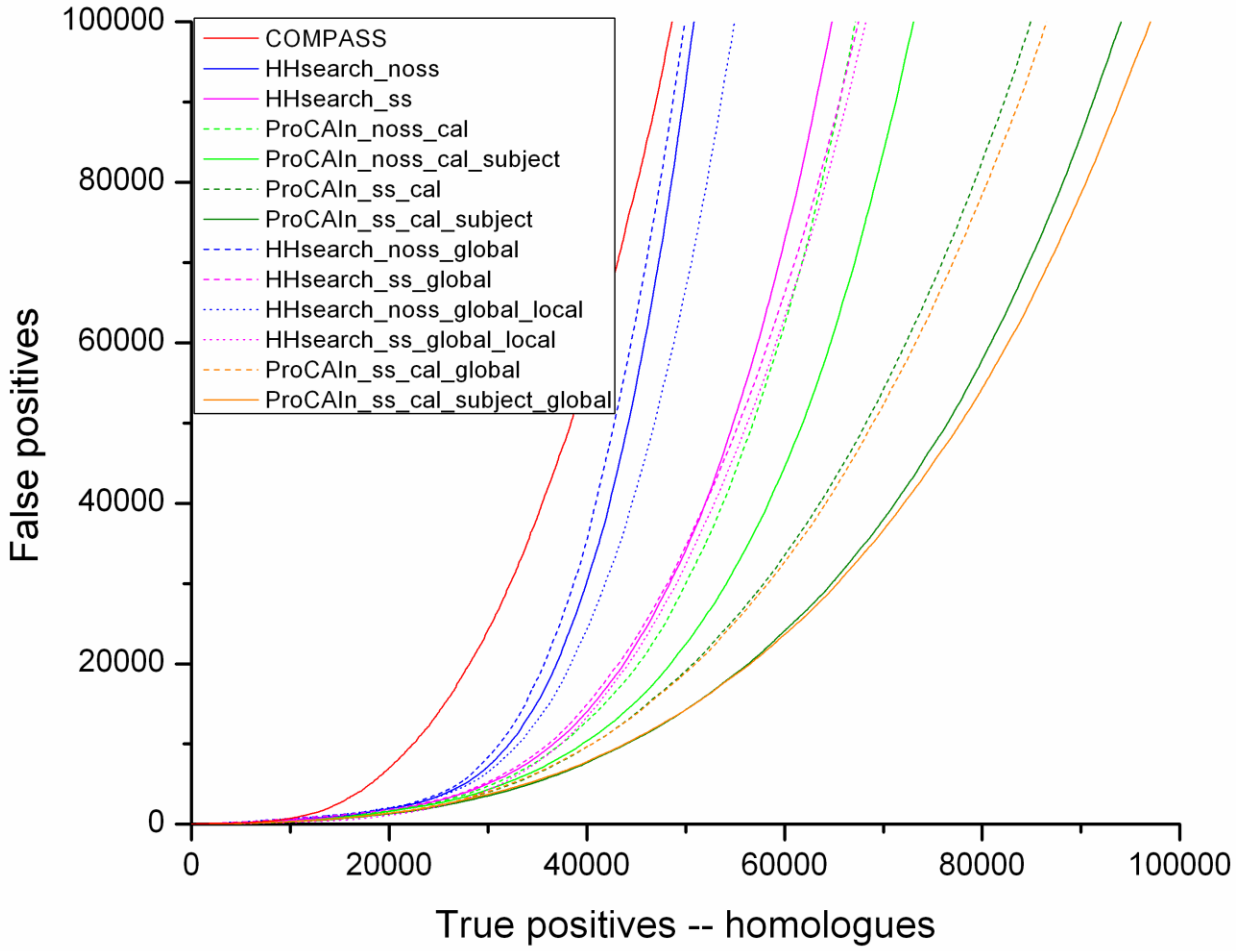

Figure 71 the results of reference dependent evaluation with SCOP superfamily relationship, SVM score and alignment quality (with global results)

5. Reference independent global evaluation with GDT_TS

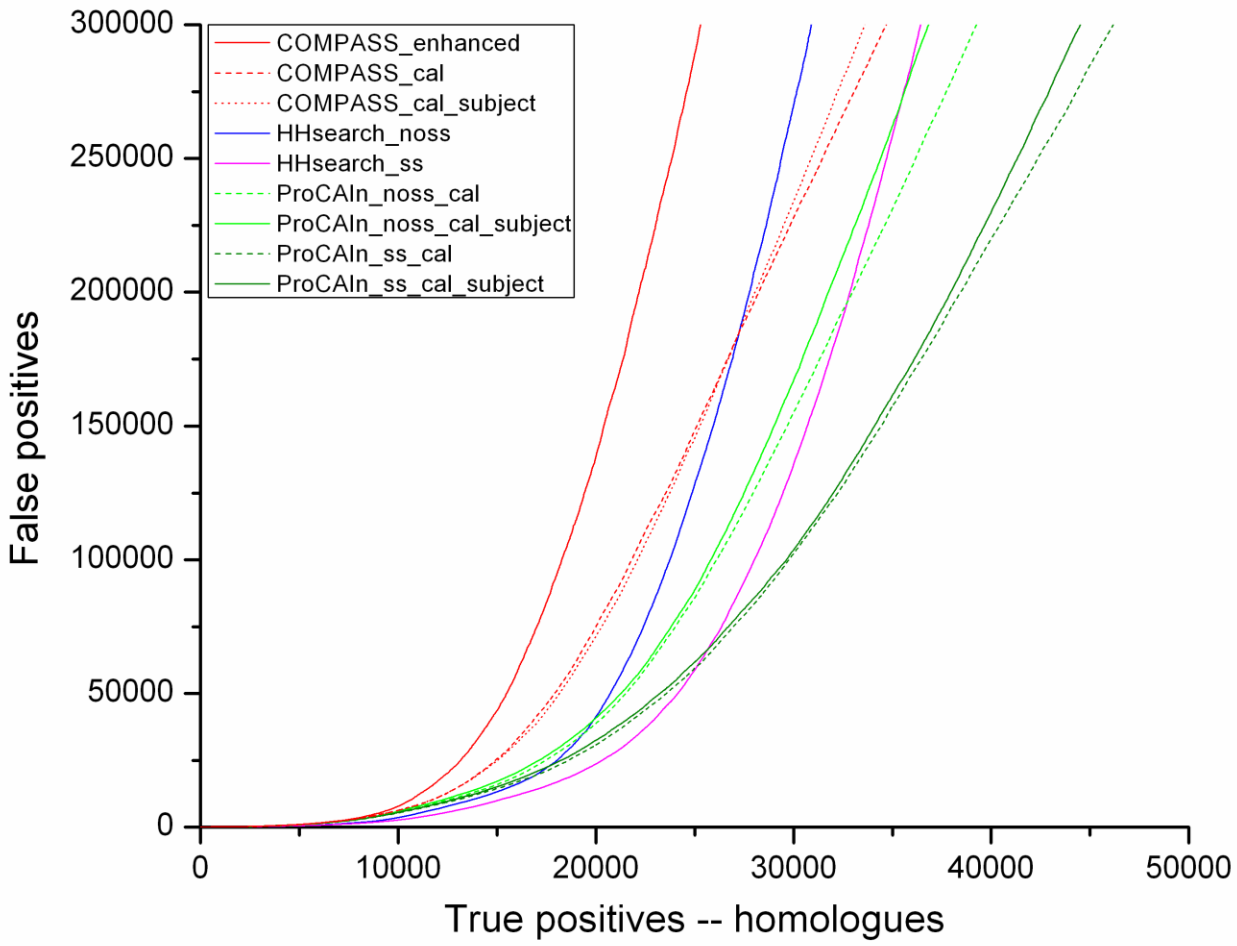

Figure 72 the result of reference independent global evaluation with GDT_TS

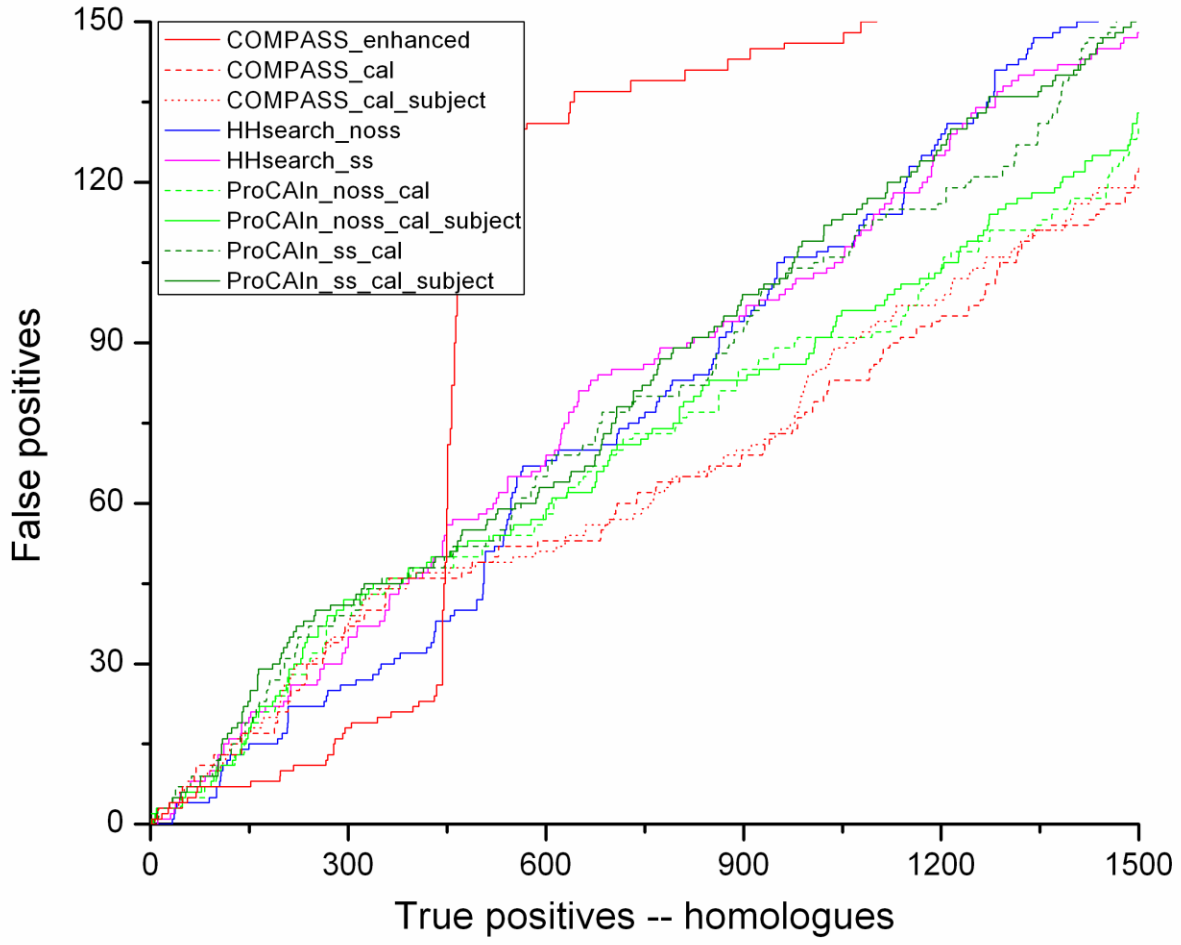

Figure 73 a zoom-in plot of the result of reference independent global evaluation with GDT_TS

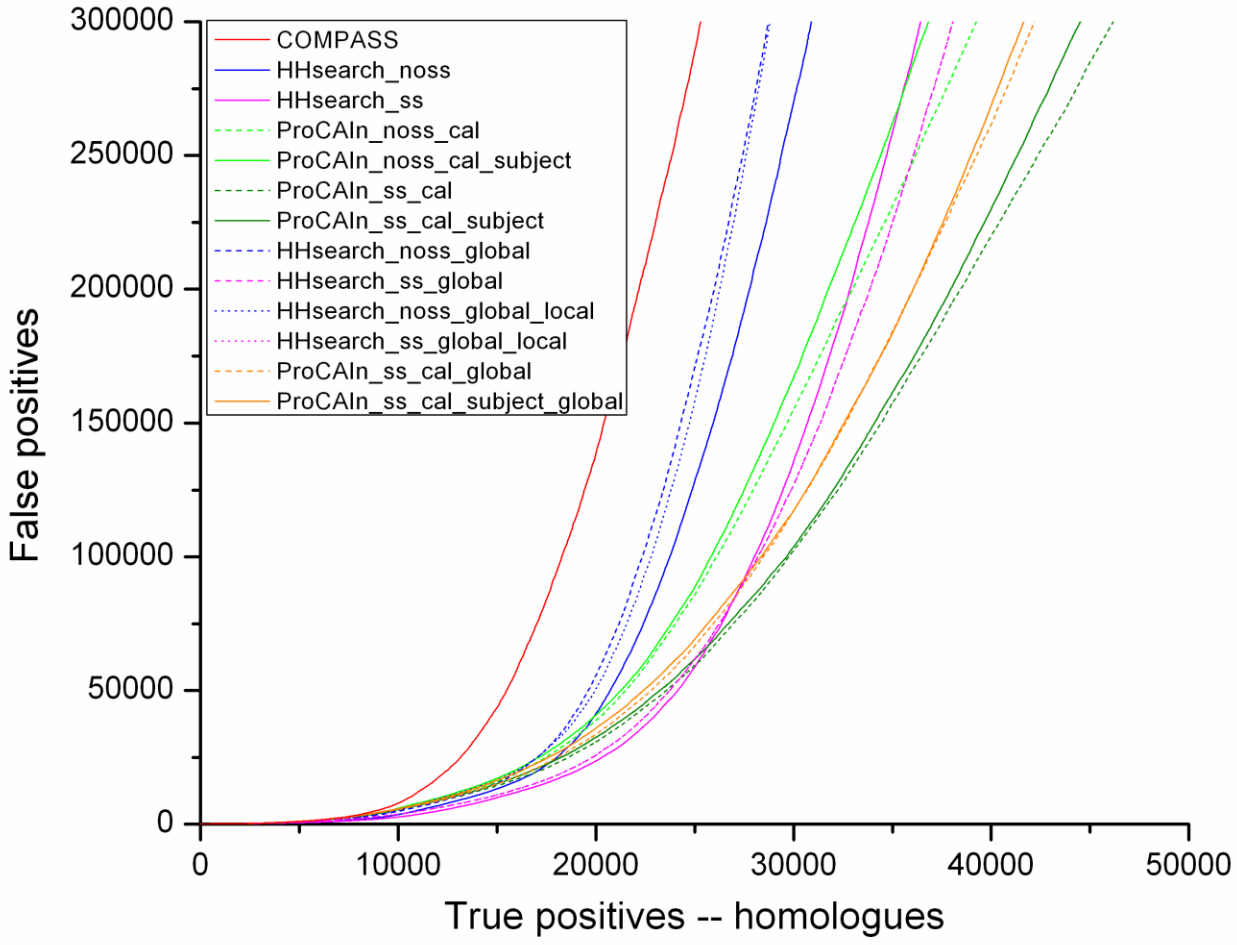

Figure 74 the result of reference independent global evaluation with GDT_TS (with global results)

6. Reference independent global evaluation with LGA GDT_TS

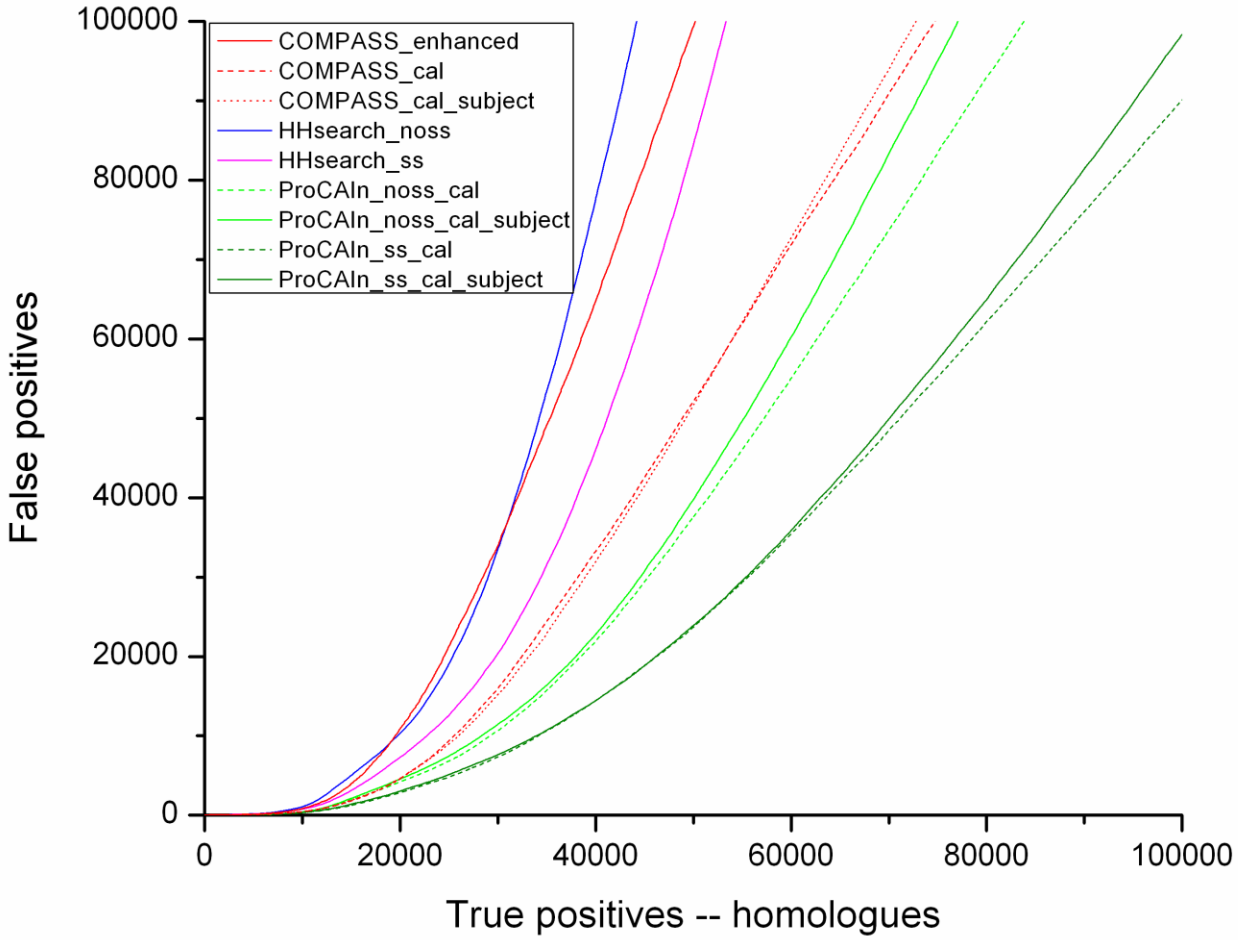

Figure 75 the result of reference independent global evaluation with LGA GDT_TS

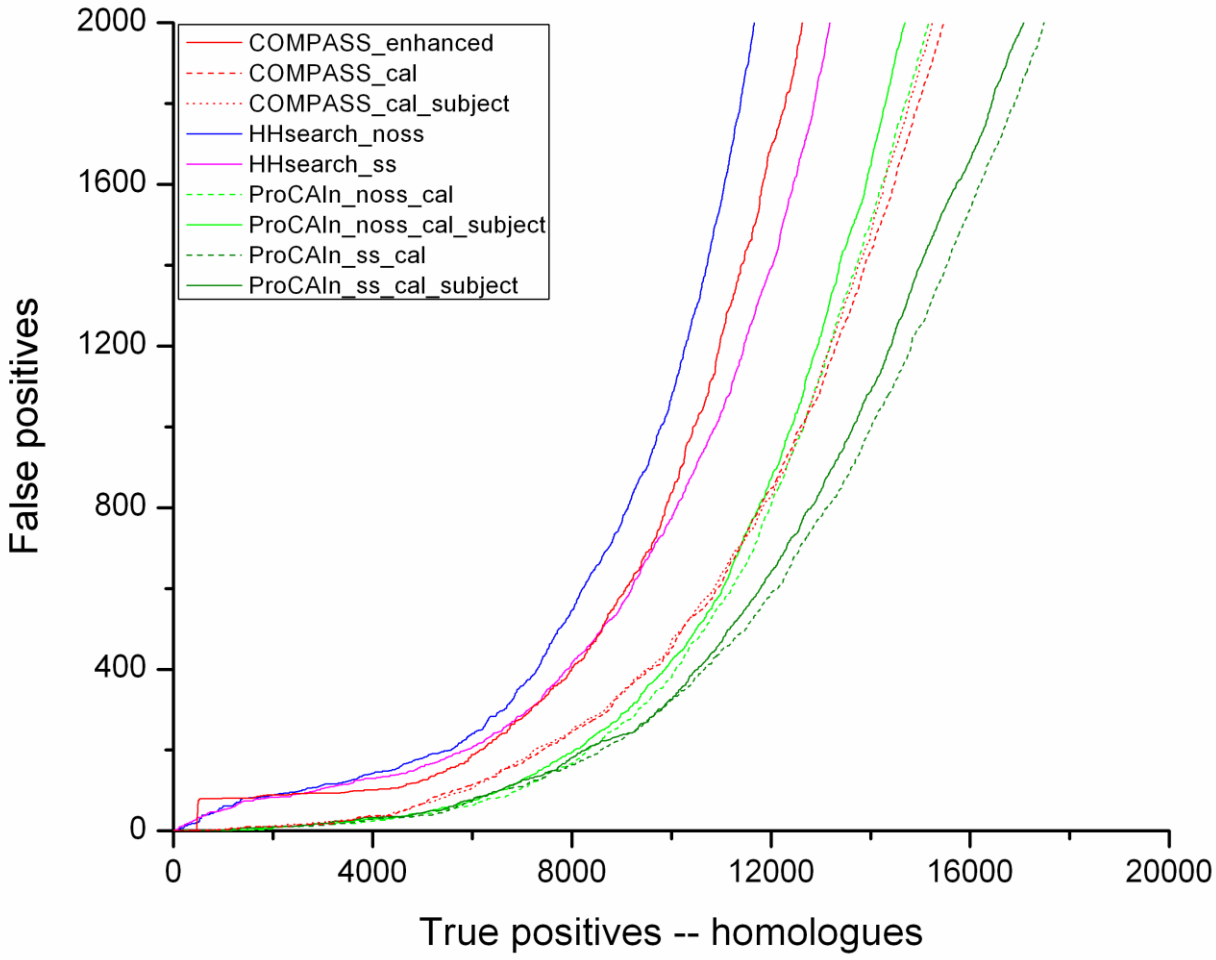

Figure 76 a zoom-in plot of the result of reference independent global evaluation with LGA GDT_TS

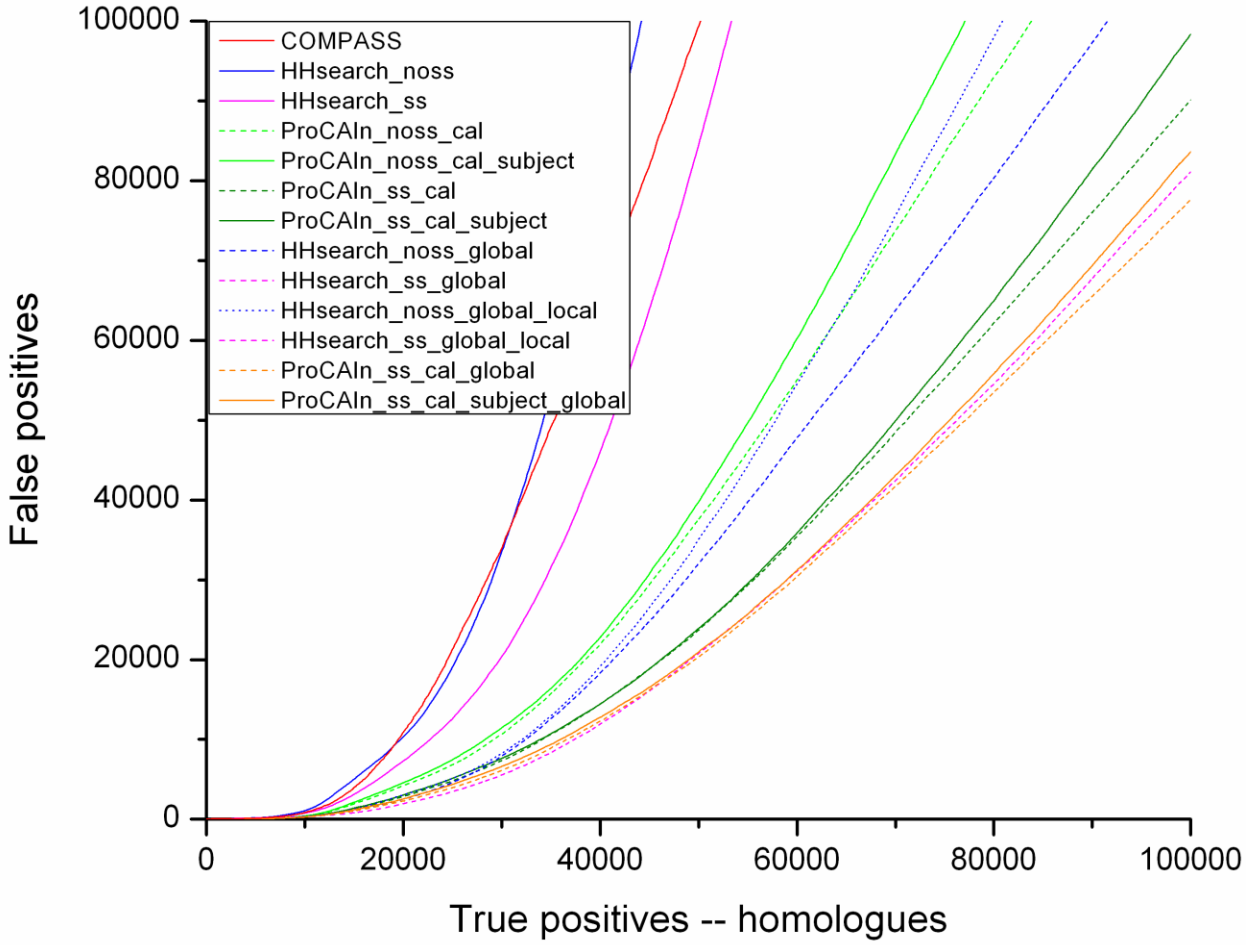

Figure 77 the result of reference independent global evaluation with LGA GDT_TS (with global results)

7. Reference independent global evaluation with Live Bench Contact-a

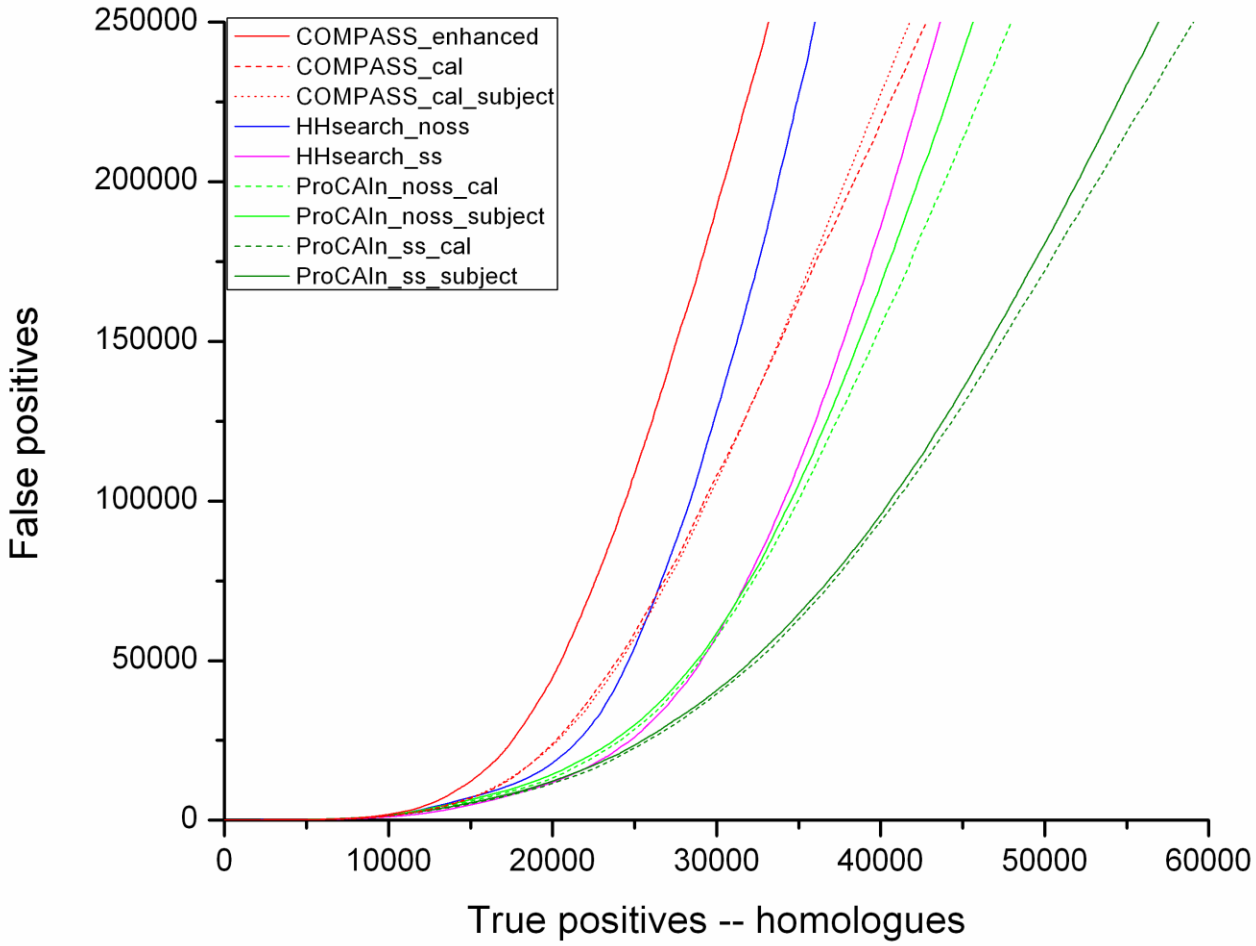

Figure 78 the result of reference independent global evaluation with Live Bench Contact-a

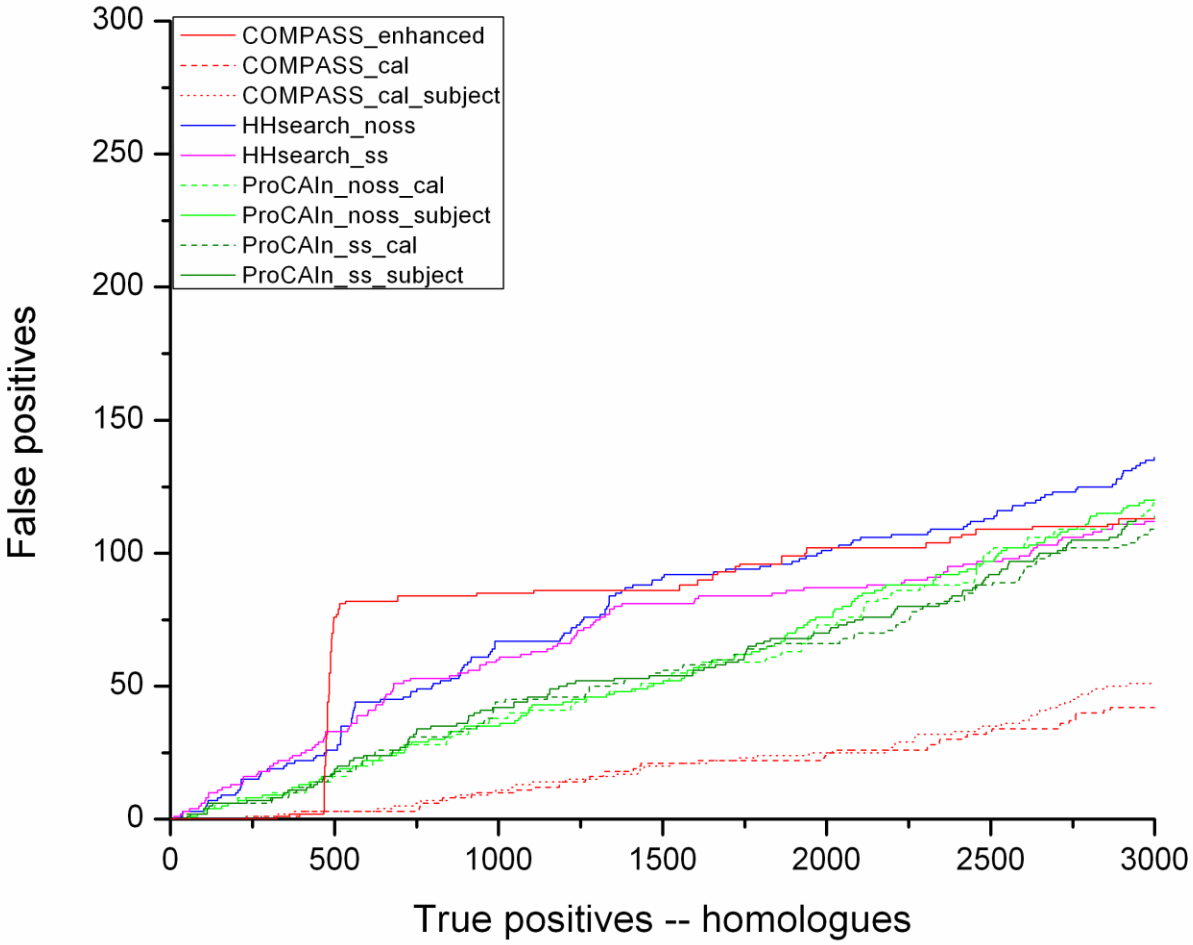

Figure 79 a zoom-in plot of the result of reference independent global evaluation with Live Bench Contact-a

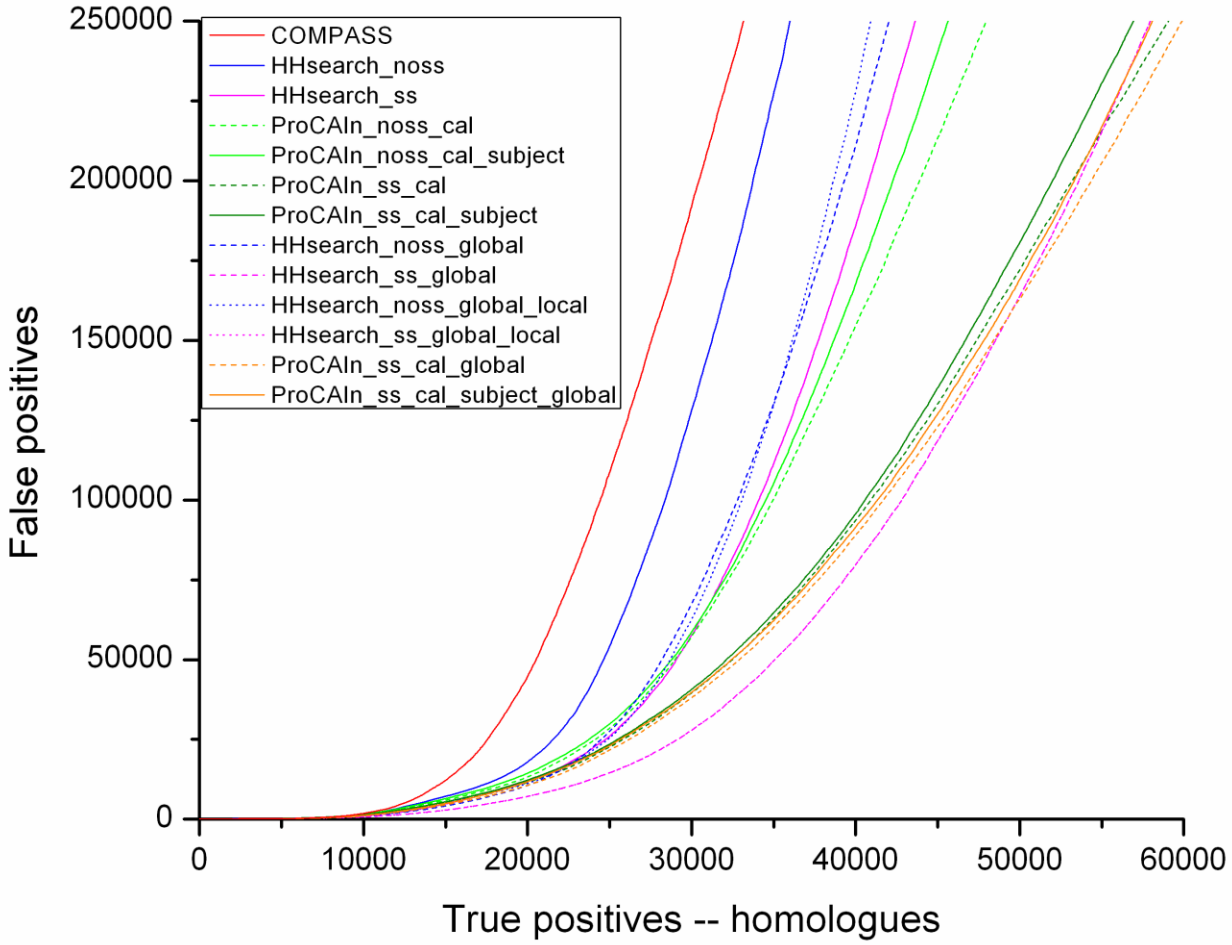

Figure 80 the result of reference independent global evaluation with Live Bench Contact-a (with global results)

8. Reference independent global evaluation with Live Bench Contact-b

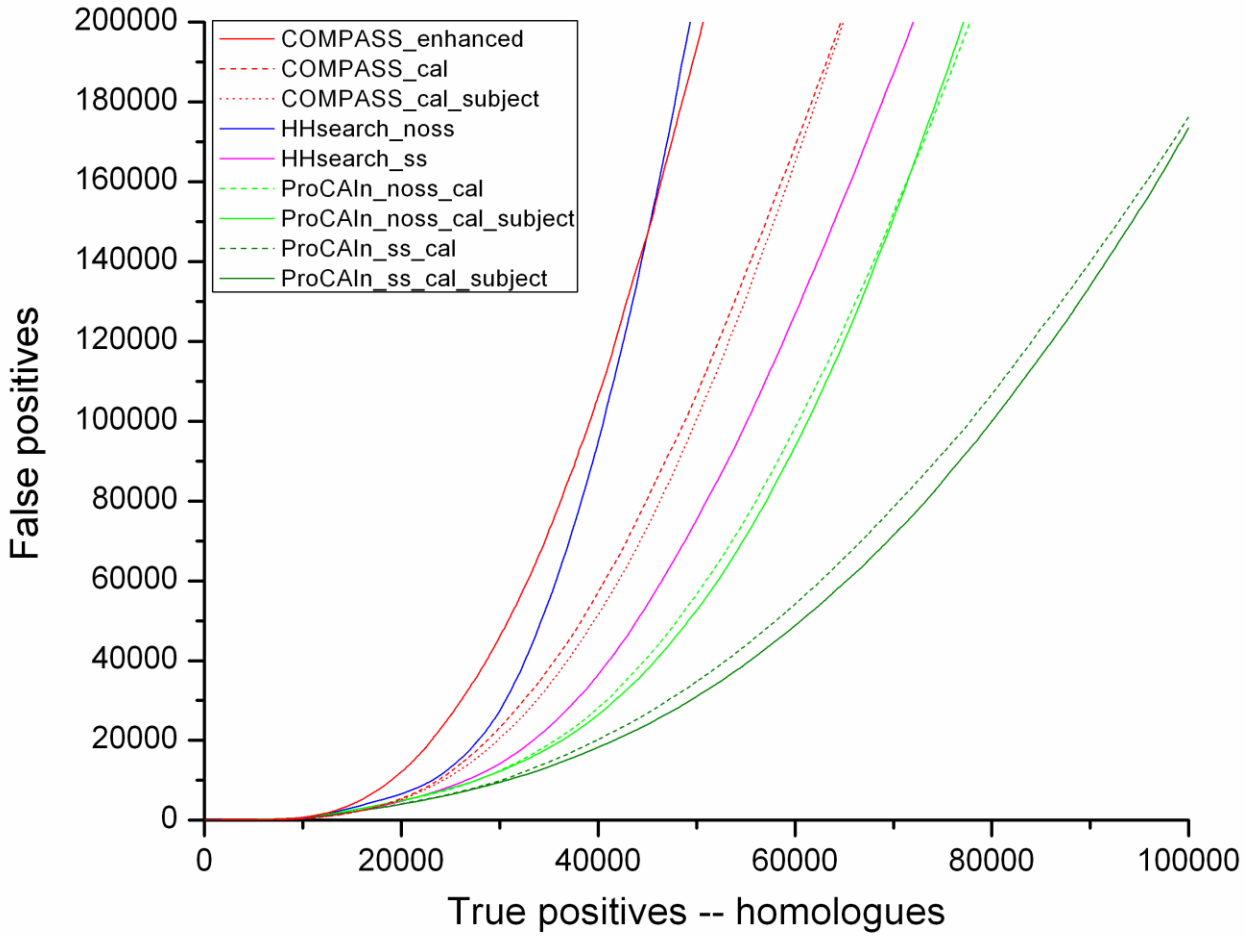

Figure 81 the result of reference independent global evaluation with Live Bench Contact-b

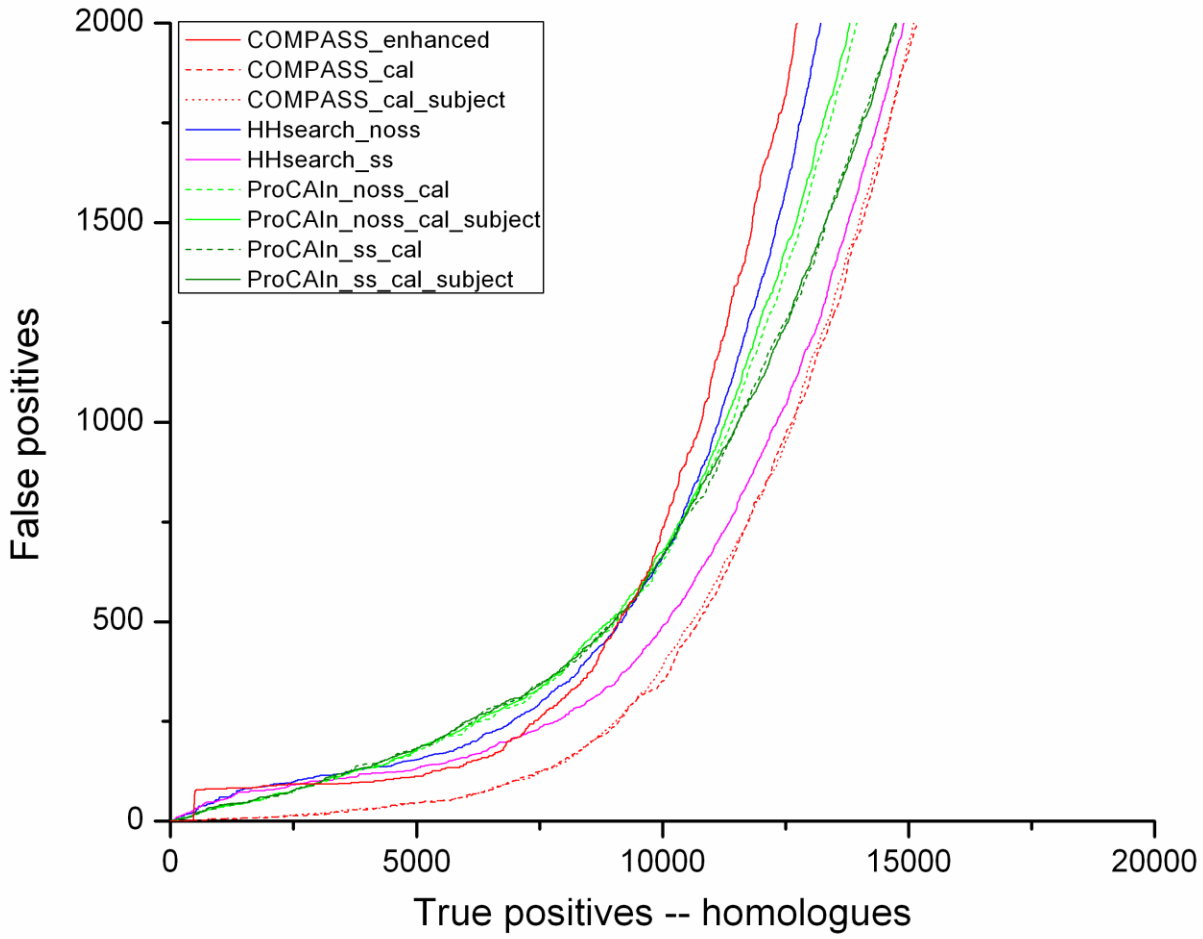

Figure 82 a zoom-in plot of the result of reference independent global evaluation with Live Bench Contact-b

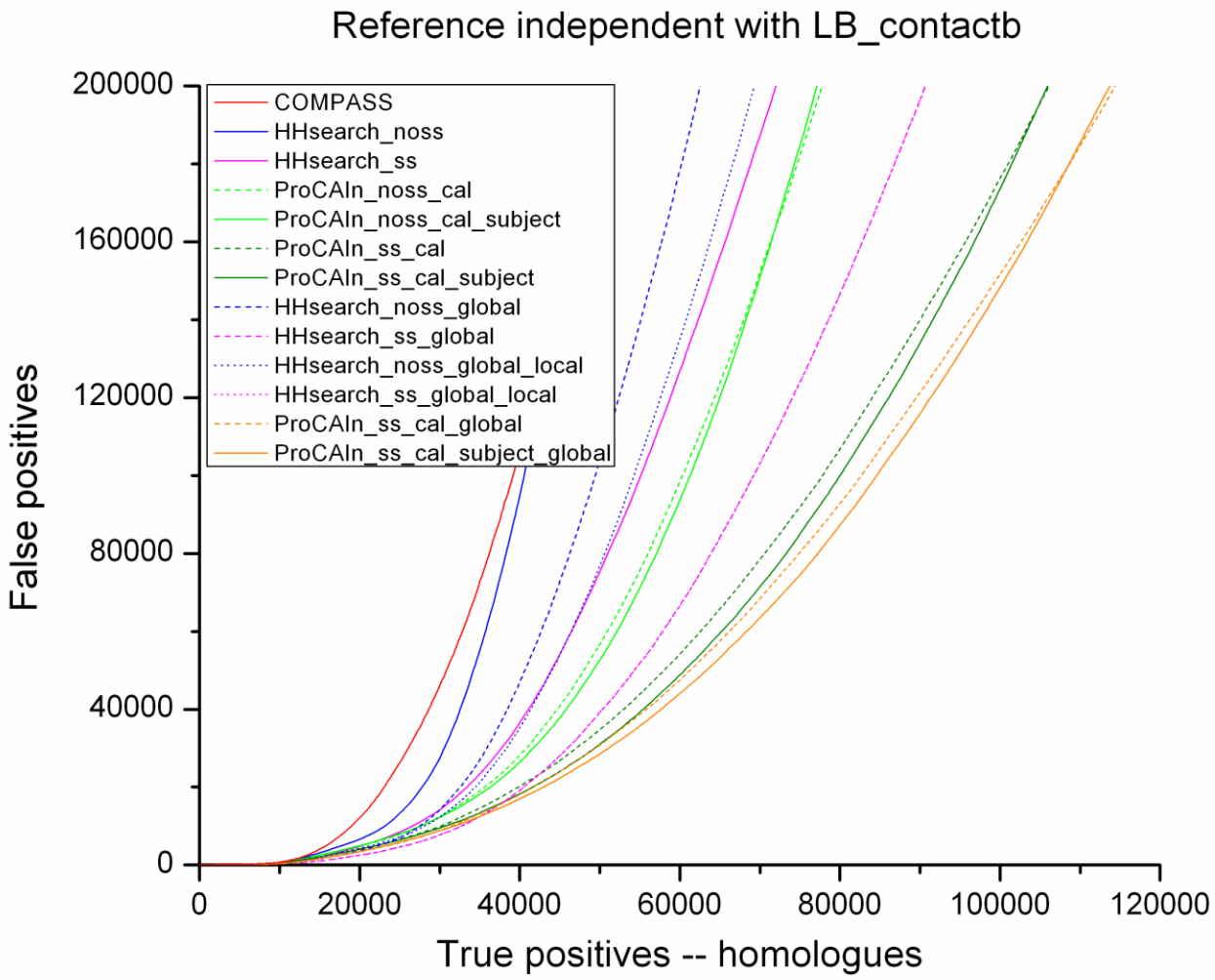

Figure 83 the result of reference independent global evaluation with Live Bench Contact-b (with global results)

ii. Query family sensitivity student t-test

Eight evaluation methods used are listed in the following:

1. Reference dependent evaluation without quality.
2. Reference dependent evaluation with SF only.
3. Reference dependent evaluation with SF only and uncertain (SF true, not SF but SVM>0.6 uncertain, others false).
4. Reference dependent evaluation with quality.
5. Reference independent global evaluation with GDT_TS.
6. Reference independent global evaluation with LGA GDT_TS.
7. Reference independent global evaluation with Live Bench Contact-a.
8. Reference independent global evaluation with Live Bench Contact-b.

1. Ten percent sensitivity t-test

10% sensitivity	COMPASS_en	COMPASS_cal	COMPASS_cal_subj	HHsearch_noss	HHsearch_noss_global	HHsearch_noss_global_local	ProCAln_noss_cal	ProCAln_noss_cal_subj	HHsearch_ss	HHsearch_ss_global	HHsearch_ss_global_local	ProCAln_ss_cal	ProCAln_ss_cal_Promals	ProCAln_ss_cal_subj
COMPASS_cal	-1.61e-64 -9.87e-25 -1.06e-21 -1.72e-45 -1.01e-02 1.16e-13 -9.25e-02 -3.82e-32													
COMPASS_cal_subj	-1.68e-136 -9.46e-24 -4.58e-24 -5.12e-74 -4.98e-02 1.69e-58 -2.03e-01 -3.75e-28	-1.89e-70 -4.57e-01 -4.39e-02 -3.93e-27 -5.94e-02 3.4e-157 1.9e-01 2.37e-01												
HHsearch_noss	1.36e-03 -1.29e-12 -2.13e-11 9.02e-01 5.85e-01 0e+00 4.81e-09 7.39e-06	1.25e-46 1.97e-01 1.78e-01 1.52e-15 8.55e-02 0e+00 2.1e-14 7.04e-47	5.56e-100 2.32e-01 7.93e-02 2.81e-25 1.59e-02 0e+00 1.67e-17 7.86e-44											

HHsearch_noss_global	3.45e-16 -4.76e-08 -1.5e-05 1.02e-03 5.98e-27 -4.74e-48 2.88e-19 9.06e-16	8.78e-83 1.6e-03 5.57e-05 3.36e-33 5.81e-39 -2.73e-96 3.09e-38 3.11e-40	5.37e-138 2.65e-03 6.2e-06 3.21e-46 3.01e-40 -1.98e-152 1.24e-38 1.36e-40	1.14e-17 2.99e-03 3.11e-04 3.36e-13 2.29e-43 -0e+00 1.36e-11 -2.24e-02										
HHsearch_noss_global_local	1.36e-03 -1.29e-12 -2.13e-11 -6.41e-01 9.93e-25 -1.81e-27 4.83e-17 -2.64e-01	1.25e-46 1.97e-01 1.78e-01 1.51e-15 4.23e-33 -1.6e-65 6.46e-34 1.56e-05	5.56e-100 2.32e-01 7.93e-02 2.22e-25 8.24e-35 -8.54e-117 7.38e-35 1.12e-05	NA NA NA 5.67e-01 7.89e-41 -0e+00 3.58e-08 -1.49e-27	-1.14e-17 -2.99e-03 -3.11e-04 -1.56e-20 -2.91e-03 2.29e-37 -1.96e-01 -2.08e-172									
ProCAln_noss_cal	-6.94e-48 -2.09e-25 -6.07e-21 -8.83e-31 1.83e-02 2.68e-129 1.74e-04 -8.42e-11	-6.61e-05 -1.2e-06 -1.27e-04 -2.88e-03 7.63e-05 7.04e-187 5.92e-10 4.39e-02	3.17e-09 -4.61e-07 -2.82e-04 4.73e-01 4.86e-05 8.33e-128 5.68e-10 3.3e-02	-6.99e-82 -7.67e-06 -4.07e-06 -2.06e-29 8.58e-02 -0e+00 -5.03e-01 -1.78e-23	-1.93e-118 -1.11e-09 -1.5e-10 -1.12e-63 -8.21e-28 1.31e-256 -1.17e-13 -2.39e-48	-6.99e-82 -7.67e-06 -4.07e-06 -1.17e-39 -1.31e-22 1.61e-241 -2.32e-11 -1.75e-09								
ProCAln_noss_cal_subj	-4.48e-108 -1.43e-26 -9.16e-25 -4.05e-48 6.98e-02 5.83e-144 1.4e-03 -1.07e-12	-1.13e-44 -1.91e-07 -8.22e-07 -5.14e-12 6.44e-04 1.15e-202 3.09e-07 3.37e-01	-2.61e-06 -2.19e-09 -1.36e-06 -3.01e-04 2.65e-03 3.43e-152 4.61e-08 2.05e-01	-2.13e-145 -2.6e-08 -1.04e-08 -1.84e-51 4.91e-01 -0e+00 -9.47e-02 -1.67e-25	-1.38e-186 -1.89e-14 -5.95e-16 -7.44e-86 -2.05e-31 8.89e-266 -2.76e-17 -4.59e-52	-2.13e-145 -2.6e-08 -1.04e-08 -2.18e-59 -3.94e-25 1.1e-254 -1.26e-13 -4.14e-11	-5.79e-123 -2.88e-03 -6.41e-03 -7.82e-41 -3.43e-04 9.72e-32 -1.13e-03 -6.85e-05							
HHsearch_ss	-2.91e-142 -1.34e-24 -7.49e-25 -1.13e-36 -3.09e-37 1.29e-234 -7.14e-34 -5.63e-46	-1.59e-82 -1.66e-05 -1.16e-06 -2.51e-09 -1.67e-30 2.41e-291 -2.26e-36 -1.28e-18	-3.22e-29 -3.93e-06 -9.98e-07 -3.02e-02 -4.84e-30 6.89e-257 -6.08e-35 -3.75e-18	-5.68e-270 -4.79e-18 -5.12e-15 -2.61e-83 -2.8e-52 -7.35e-238 -1.17e-115 -2.69e-126	-1.82e-301 -3.46e-19 -9.78e-22 -2.17e-117 -5.14e-129 0e+00 -1.23e-137 -4.67e-228	-5.68e-270 -4.79e-18 -5.12e-15 -6.6e-103 -1.88e-124 0e+00 -2.85e-129 -4.41e-110	-1.07e-55 -6.16e-01 -1.95e-01 -9.17e-04 -5.3e-49 3.32e-115 -4.73e-63 -6.1e-11	-3.91e-12 9.13e-01 -5.49e-01 1.82e-01 -4.82e-44 5.58e-94 -3.97e-58 -7.85e-09						

HHsearch_ss_global	-1.46e-105 -2.04e-21 -2.99e-19 -8.46e-19 2.64e-03 -4.91e-260 -5.27e-12 -1.03e-37	-1.04e-50 -1.22e-03 -2.48e-03 -8.89e-02 1.07e-06 -0e+00 -5.6e-08 -4.61e-21	-2.75e-14 -2.03e-03 -1.03e-01 2.44e-01 2.03e-07 -0e+00 -4.89e-08 -2.92e-22	-2.11e-157 -1.23e-05 -8.49e-07 -7.48e-21 3.78e-04 -0e+00 -5.18e-38 -2.98e-101	-1.16e-287 -2.66e-23 -1.38e-21 -2.12e-127 -3.59e-48 -1.03e-307 -4.81e-142 -0e+00	-2.11e-157 -1.23e-05 -8.49e-07 -3.59e-45 -1.59e-29 -6.28e-319 -2.6e-98 -1.23e-151	-1.19e-30 7.06e-01 -9.12e-01 6.79e-05 5.73e-02 -0e+00 -2.62e-20 -3.54e-20	-5.66e-04 1.36e-01 7.56e-02 9.12e-15 1.11e-03 -0e+00 -2.44e-16 -2.98e-17	3.34e-07 2.7e-02 2.73e-02 6.14e-17 8e-67 -0e+00 4.15e-11 -1.11e-01					
HHsearch_ss_global_local	-1.46e-105 -2.04e-21 -2.99e-19 -9.69e-39 2.64e-03 -4.91e-260 -5.27e-12 -1.03e-37	-1.04e-50 -1.22e-03 -2.48e-03 -7.07e-10 1.07e-06 -0e+00 -5.6e-08 -4.61e-21	-2.75e-14 -2.03e-03 -1.03e-01 -1.14e-03 2.03e-07 -0e+00 -4.89e-08 -2.92e-22	-2.11e-157 -1.23e-05 -8.49e-07 -9.09e-47 3.78e-04 -0e+00 -5.18e-38 -2.98e-101	-1.16e-287 -2.66e-23 -1.38e-21 -4.25e-146 -3.59e-48 -1.03e-307 -4.81e-142 -0e+00	-2.11e-157 -1.23e-05 -8.49e-07 -2.71e-68 -1.59e-29 -6.28e-319 -2.6e-98 -1.23e-151	-1.19e-30 7.06e-01 -9.12e-01 -4.76e-01 5.73e-02 -0e+00 -2.62e-20 -3.54e-20	-5.66e-04 1.36e-01 7.56e-02 3.53e-04 1.11e-03 -0e+00 -2.44e-16 -2.98e-17	3.34e-07 2.7e-02 2.73e-02 8.86e-03 8e-67 -0e+00 4.15e-11 -1.11e-01	NA NA NA -3.25e-22 NA NA NA NA				
ProCAln_ss_cal	-7.13e-162 -2.91e-36 -3.1e-35 -1.36e-119 -5.13e-14 -3.05e-57 -1.04e-22 -6.07e-173	-9.38e-104 -5.2e-15 -4.5e-16 -3.62e-80 -1.11e-12 -1.96e-128 -6.25e-25 -7.85e-156	-2.07e-35 -1.27e-14 -6.96e-13 -9.57e-56 -4.33e-11 -1.45e-210 -2.64e-26 -5.07e-161	-4.61e-176 -2.15e-16 -4.93e-17 -2.84e-107 -4.49e-15 -0e+00 -5.57e-52 -4.57e-207	-2.9e-206 -9.35e-22 -7.72e-23 -1.45e-153 -1.38e-67 -6.42e-06 -4.43e-94 -2.24e-254	-4.61e-176 -2.15e-16 -4.93e-17 -1.7e-129 -1.51e-58 -4.69e-16 -6.28e-89 -7.09e-183	-1.72e-96 -1.58e-08 -9.18e-13 -6.04e-71 -5.91e-33 -0e+00 -9.41e-103 -1.56e-267	-1.1e-20 -2.3e-06 -7.1e-07 -1.31e-40 -4.53e-27 -0e+00 -2.75e-82 -1.16e-232	-5.16e-02 -1.39e-04 -4.91e-04 -3.75e-30 3.22e-05 -0e+00 1.51e-01 -2.46e-81	-2.08e-04 -9.93e-08 -5.24e-08 -2.72e-67 -1.6e-27 8.89e-85 -8.68e-08 -8.71e-62	-2.08e-04 -9.93e-08 -5.24e-08 -8.72e-46 -1.6e-27 8.89e-85 -8.68e-08 -8.71e-62			
ProCAln_ss_cal_Promals	-7.13e-162 -2.91e-36 -3.1e-35 -9.48e-123 -8.94e-05 -0e+00 -8.7e-18 -4.34e-240	-9.38e-104 -5.2e-15 -4.5e-16 -8.57e-78 -3.36e-02 -0e+00 -1.19e-14 -1.5e-227	-2.07e-35 -1.27e-14 -6.96e-13 -7.23e-56 -1.4e-01 -0e+00 -4.48e-16 -3.35e-240	-4.61e-176 -2.15e-16 -4.93e-17 -2.43e-118 -2.21e-02 -0e+00 -7.23e-50 -2e-286	-2.9e-206 -9.35e-22 -7.72e-23 -3.93e-165 -5.32e-37 -9.8e-301 -2.78e-101 -0e+00	-4.61e-176 -2.15e-16 -4.93e-17 -2.41e-137 -4.29e-34 -0e+00 -3.18e-97 -3.07e-263	-1.72e-96 -1.58e-08 -9.18e-13 -6.85e-63 -8.55e-03 -0e+00 -9.69e-45 -1.54e-255	-1.1e-20 -2.3e-06 -7.1e-07 -2.03e-36 -5.64e-02 -0e+00 -7.68e-39 -1.73e-244	-5.16e-02 -1.39e-04 -4.91e-04 -5.16e-29 8.81e-19 -0e+00 1.17e-01 -5.04e-140	-2.08e-04 -9.93e-08 -5.24e-08 -8e-68 -9e-07 -5.7e-66 -3.82e-05 -7.24e-127	-2.08e-04 -9.93e-08 -5.24e-08 -3.45e-45 -9e-07 -5.7e-66 -3.82e-05 -7.24e-127	NA NA NA 4.3e-01 4.59e-13 -0e+00 1.58e-03 -1.22e-30		
ProCAln_ss_cal_subj	-9.66e-258 -4.26e-35 -4.28e-39 -3.65e-150 -3.94e-13 -4.24e-05 -1.85e-19 -1.79e-140	-7.73e-198 -1.75e-17 -1.02e-21 -1.25e-108 -6.69e-13 -2.85e-18 -1.94e-19 -2.94e-115	-4.96e-127 -9.1e-16 -7.31e-18 -1.8e-86 -5.14e-12 -5.5e-63 -4.93e-22 -1.88e-125	-3.55e-251 -1.21e-19 -7.53e-26 -4.71e-133 -1.67e-14 -0e+00 -1.19e-45 -1.05e-178	-2.73e-277 -3.9e-22 -2.3e-29 -4.41e-178 -2.37e-69 7.41e-14 -6.39e-83 -1.14e-232	-3.55e-251 -1.21e-19 -7.53e-26 -5.81e-154 -1.49e-60 4.26e-06 -1.1e-81 -2.35e-162	-4.93e-197 -7.15e-10 -1.63e-16 -9.38e-102 -1.26e-33 -9.72e-251 -1.1e-91 -1.31e-214	-2.4e-118 -3.56e-08 -2.67e-13 -2.48e-70 -1.82e-29 -5.11e-302 -5.16e-80 -5.81e-212	-6.25e-42 -4.47e-07 -1.71e-08 -2.34e-53 3.59e-05 -0e+00 3.99e-02 -5.55e-69	-2.32e-45 -1.59e-10 -1.86e-15 -3.98e-96 -4.75e-29 1.37e-163 -7.13e-06 -3.34e-49	-2.32e-45 -1.59e-10 -1.86e-15 -4.05e-71 -4.75e-29 1.37e-163 -7.13e-06 -3.34e-49	-7.86e-81 9.68e-01 -8.82e-02 -6.68e-32 -3.76e-01 3.6e-209 1.17e-05 1.75e-15	-7.86e-81 9.68e-01 -8.82e-02 -1.15e-21 -4.35e-13 0e+00 -9.57e-02 4.3e-43	

ProCAIn_ss_cal_subj_Promals	-9.66e-258 -4.26e-35 -4.28e-39 -2.83e-156 -3.36e-06 -2.17e-288 -9.64e-16 -2.34e-200	-7.73e-198 -1.75e-17 -1.02e-21 -7.61e-111 -5.82e-03 -0e+00 -4.17e-12 -7.63e-183	-4.96e-127 -9.1e-16 -7.31e-18 -1.47e-86 -2.39e-02 -0e+00 -7.75e-14 -7.26e-199	-3.55e-251 -1.21e-19 -7.53e-26 -6.22e-146 -1.08e-02 -0e+00 -1.16e-43 -4.34e-259	-2.73e-277 -3.9e-22 -2.3e-29 -6.94e-197 -5.06e-42 -8.25e-169 -1.47e-87 -1.1e-302	-3.55e-251 -1.21e-19 -7.53e-26 -1.03e-172 -2.85e-38 -1.11e-222 -5.43e-84 -3.44e-232	-4.93e-197 -7.15e-10 -1.63e-16 -5.36e-93 -1.13e-04 -0e+00 -5.67e-39 -7.22e-224	-2.4e-118 -3.56e-08 -2.67e-13 -1.41e-62 -3.39e-03 -0e+00 -2.5e-33 -8.05e-219	-6.25e-42 -4.47e-07 -1.71e-08 -1.31e-52 1.09e-15 -0e+00 3.02e-02 -6.05e-121	-2.32e-45 -1.59e-10 -1.86e-15 -8.22e-110 -7.75e-09 -2.53e-06 -5.4e-04 -2.42e-104	-2.32e-45 -1.59e-10 -1.86e-15 -3.94e-84 -7.75e-09 -2.53e-06 -5.4e-04 -2.42e-104	-7.86e-81 9.68e-01 -8.82e-02 -6.5e-12 4.95e-08 -3.54e-262 9.32e-06 -7.75e-10	-7.86e-81 9.68e-01 -8.82e-02 -1.21e-42 -3.43e-11 1.46e-132 3.16e-02 1.51e-24	NA NA NA -9.1e-01 9.08e-09 -0e+00 1.86e-03 -7.49e-22

Table 6 the result of 10% sensitivity t-test for the whole dataset

2. Twenty-five percent sensitivity t-test

25% sensitivity	COMPASS_en	COMPASS_cal	COMPASS_cal_subj	HHsearch_noss	HHsearch_noss_global	HHsearch_noss_global_local	ProCAIn_noss_cal	ProCAIn_noss_cal_subj	HHsearch_ss	HHsearch_ss_global	HHsearch_ss_global_local	ProCAIn_ss_cal	ProCAIn_ss_cal_Promals	ProCAIn_ss_cal_subj
COMPASS_cal	-2.01e-86 -5.3e-37 -3.79e-31 -7.68e-68 -1.93e-04 1.19e-20 -1.04e-03 -2.57e-25													
COMPASS_cal_subj	-1.04e-198 -7.01e-35 -1.67e-36 -2.52e-107 -5.85e-05 1.07e-73 -1.52e-02 -1.07e-20	-9.76e-119 -2e-01 -8.42e-04 -2.47e-35 -2.62e-02 1.16e-184 7.37e-04 3.83e-01												

HHsearch_noss	8.06e-11 -1.18e-21 -4.61e-16 4.22e-01 3.73e-01 0e+00 3.59e-23 1.18e-47	3.65e-61 -3.94e-01 4.17e-01 1.01e-18 2.26e-02 0e+00 1.42e-45 4.3e-144	1.93e-125 -8.48e-01 4.64e-03 7.44e-32 2.88e-03 0e+00 2.88e-42 1.74e-135											
HHsearch_noss_global	1.45e-45 -1.74e-10 -5.01e-07 7.38e-04 2.85e-11 -2.53e-146 8.49e-11 2.32e-11	9.53e-118 1.51e-04 9.63e-07 3.66e-29 1.3e-19 -2.5e-208 1.18e-27 4.76e-17	5.81e-195 7.61e-06 4.7e-11 1.78e-43 5.06e-21 -2.8e-271 1.99e-24 8.36e-15	3e-36 1.2e-09 5.04e-07 3.26e-07 1.93e-24 -0e+00 8.87e-01 -3.42e-40										
HHsearch_noss_global_local	8.06e-11 -1.18e-21 -4.61e-16 -1.15e-01 2.58e-09 -3.37e-101 1.43e-10 -1.46e-01	3.65e-61 -3.94e-01 4.17e-01 8.53e-10 8.75e-18 -3.31e-162 5.48e-27 7.89e-01	1.93e-125 -8.48e-01 4.64e-03 1.36e-18 3.52e-19 -6.63e-231 5.28e-23 -4.93e-01	NA NA NA -4.33e-06 9.82e-21 -0e+00 -6.33e-01 -1.66e-85	-3e-36 -1.2e-09 -5.04e-07 -5.38e-37 -2.49e-06 6.51e-64 -3.09e-01 -5.27e-295									
ProCAln_noss_cal	-1.66e-87 -1.07e-42 -1.2e-38 -1.17e-74 -3.99e-02 6.34e-114 1.18e-01 -2.26e-14	-4.19e-27 -1.65e-09 -6.41e-07 -2.57e-28 7.93e-01 1.17e-175 3.07e-09 -7.55e-01	1.24e-01 -1.56e-09 -5.52e-05 -4.76e-12 6.67e-01 8.37e-96 1.68e-07 -1.62e-01	-1.06e-125 -5.2e-09 -2.74e-11 -1.62e-68 -3.36e-04 -0e+00 -9.76e-16 -5.07e-104	-4.06e-181 -4.19e-19 -2.67e-19 -4.27e-85 -4.17e-29 6.32e-320 -1.36e-12 -8.28e-22	-1.06e-125 -5.2e-09 -2.74e-11 -3.31e-52 -1.34e-23 2.13e-295 -1.79e-13 -3.15e-01								
ProCAln_noss_cal_subj	-8.77e-156 -6.3e-43 -3.02e-41 -5.81e-102 -3.91e-03 7.53e-130 2.3e-01 -6.65e-15	-2.9e-83 -3.96e-12 -2.83e-11 -1.07e-50 -3.57e-01 1.54e-193 1.32e-07 -5.02e-01	-6.27e-25 -6.97e-11 -2.98e-09 -1.81e-30 -6.41e-01 8.54e-124 2.78e-06 -8.31e-02	-9.05e-187 -2.14e-10 -3.43e-16 -3.13e-92 -7.24e-06 -0e+00 -3e-17 -1.9e-101	-2.42e-244 -1.58e-21 -9.96e-26 -4.48e-107 -6.16e-30 0e+00 -9.82e-14 -1.66e-21	-9.05e-187 -2.14e-10 -3.43e-16 -2.41e-73 -8.24e-26 1.35e-308 -1.42e-15 -3.41e-01	-4.63e-140 -6.09e-02 -2.13e-04 -1.4e-65 -3.84e-06 5.55e-81 -1.49e-06 -7.41e-04							

HHsearch_ss	-1.56e-183 -3.8e-41 -1.44e-38 -9.26e-65 -1.9e-37 5.22e-310 -2.27e-22 -3.27e-21	-1.93e-117 -6.82e-12 -1.64e-10 -1.43e-27 -5.17e-34 0e+00 -1.35e-14 -9.14e-11	-1.33e-40 -2.17e-11 -2.66e-08 -1.56e-12 -3.51e-32 0e+00 -2.7e-16 -1.38e-13	-0e+00 -1.4e-32 -5.59e-33 -1.19e-153 -1.1e-83 -0e+00 -4.55e-184 -4.11e-220	-0e+00 -7.99e-45 -9.67e-39 -2.99e-147 -5.34e-140 0e+00 -2.57e-100 -6.04e-162	-0e+00 -1.4e-32 -5.59e-33 -1.83e-116 -3.5e-133 0e+00 -1.82e-110 -2.81e-53	-5.39e-53 -1.47e-01 -2.59e-02 -1e-01 -1.22e-30 1.02e-286 -3.07e-37 -8.04e-10	-1.76e-06 -5.91e-01 -9.69e-01 1.63e-03 -2.72e-28 1.69e-252 -3.6e-33 -6.66e-09					
HHsearch_ss_global	-7.65e-147 -2.01e-25 -4.48e-23 -2.65e-44 -4.21e-02 -0e+00 -2.29e-20 -1.14e-53	-2.45e-76 -1.4e-02 -3.06e-02 -5.79e-15 -7.35e-01 -0e+00 -2.73e-14 -5.79e-51	-8.8e-22 -2.22e-02 -3.45e-01 -5.41e-06 -5.96e-01 -0e+00 -4.39e-17 -1.42e-59	-4.87e-202 -1.05e-04 -1.24e-06 -1.26e-52 -1.49e-01 -0e+00 -1.92e-78 -7.91e-198	-0e+00 -2.46e-31 -5.31e-28 -6.24e-178 -2.42e-62 -0e+00 -2.5e-175 -0e+00	-4.87e-202 -1.05e-04 -1.24e-06 -8.32e-68 -3.19e-45 -0e+00 -9.14e-148 -9.06e-285	-7.77e-29 9.38e-03 8.72e-02 1.98e-02 7.52e-01 -0e+00 -1.29e-31 -8.98e-61	-5.82e-02 2.38e-03 3.84e-03 2.44e-10 4.24e-01 -0e+00 -1.85e-29 -1.25e-61	4.95e-07 6.34e-11 7.87e-09 3.67e-10 5.65e-49 -0e+00 5.64e-02 -3.12e-30					
HHsearch_ss_global_local	-7.65e-147 -2.01e-25 -4.48e-23 -5.41e-83 -4.21e-02 -0e+00 -2.29e-20 -1.14e-53	-2.45e-76 -1.4e-02 -3.06e-02 -5.47e-42 -7.35e-01 -0e+00 -2.73e-14 -5.79e-51	-8.8e-22 -2.22e-02 -3.45e-01 -5.91e-24 -5.96e-01 -0e+00 -4.39e-17 -1.42e-59	-4.87e-202 -1.05e-04 -1.24e-06 -1.1e-93 -1.49e-01 -0e+00 -1.92e-78 -7.91e-198	-0e+00 -2.46e-31 -5.31e-28 -1.1e-177 -2.42e-62 -0e+00 -2.5e-175 -0e+00	-4.87e-202 -1.05e-04 -1.24e-06 -5.98e-86 -3.19e-45 -0e+00 -9.14e-148 -9.06e-285	-7.77e-29 9.38e-03 8.72e-02 -1.96e-03 7.52e-01 -0e+00 -1.29e-31 -8.98e-61	-5.82e-02 2.38e-03 3.84e-03 1.2e-01 4.24e-01 -0e+00 -1.85e-29 -1.25e-61	4.95e-07 6.34e-11 7.87e-09 -6.91e-01 5.65e-49 -0e+00 5.64e-02 -3.12e-30	NA NA NA -6.66e-22 NA NA NA NA				
ProCAln_ss_cal	-1.02e-241 -6.28e-65 -5.37e-63 -3.92e-200 -9.52e-28 -6.23e-97 -1.7e-42 -3.79e-216	-2.8e-164 -5.42e-34 -2.2e-29 -1.41e-150 -1.54e-23 -2.21e-196 -1.02e-41 -3.03e-224	-3.79e-66 -5.45e-31 -8.45e-23 -1.26e-112 -2.24e-23 -9.47e-294 -1.53e-49 -5.07e-238	-2.06e-227 -4.57e-26 -4.48e-28 -6.92e-178 -1.28e-32 -0e+00 -6.94e-114 -0e+00	-6.55e-267 -4.02e-38 -8.31e-40 -3.78e-180 -2.2e-75 1.96e-02 -1.44e-113 -1.22e-225	-2.06e-227 -4.57e-26 -4.48e-28 -1.9e-148 -1.85e-71 -8.57e-02 -2.65e-113 -1.19e-164	-7.7e-118 -7.1e-20 -4.63e-19 -8e-98 -4.4e-37 -0e+00 -1.83e-141 -0e+00	-5.49e-19 -1.81e-16 -1.04e-11 -2.05e-58 -5.92e-31 -0e+00 -1.71e-131 -0e+00	-1.8e-03 -4.09e-07 -3.87e-06 -1.23e-51 6.18e-02 -0e+00 -1.42e-06 -4.68e-114	-6.3e-07 -3.89e-16 -1.39e-15 -5.2e-63 -3e-24 9.67e-138 -2.56e-08 -3.3e-33	-6.3e-07 -3.89e-16 -1.39e-15 -2.57e-41 -3e-24 9.67e-138 -2.56e-08 -3.3e-33			
ProCAln_ss_cal_Promals	-1.02e-241 -6.28e-65 -5.37e-63 -3.74e-218 -4.45e-06 -0e+00 -2.1e-27 -1.77e-261	-2.8e-164 -5.42e-34 -2.2e-29 -9.56e-167 -3.31e-03 -0e+00 -1.63e-22 -7.12e-266	-3.79e-66 -5.45e-31 -8.45e-23 -5.65e-130 -3.1e-03 -0e+00 -5.72e-27 -9.12e-272	-2.06e-227 -4.57e-26 -4.48e-28 -1.27e-192 -4.23e-06 -0e+00 -9.36e-88 -0e+00	-6.55e-267 -4.02e-38 -8.31e-40 -3.28e-205 -9.99e-39 -6.8e-316 -4.25e-104 -7.75e-295	-2.06e-227 -4.57e-26 -4.48e-28 -2.94e-172 -1.44e-33 -0e+00 -1.9e-108 -3.61e-238	-7.7e-118 -7.1e-20 -4.63e-19 -1.7e-92 -3.52e-03 -0e+00 -9.34e-55 -1.27e-308	-5.49e-19 -1.81e-16 -1.04e-11 -5.48e-59 -3.2e-02 -0e+00 -5.91e-49 -2.96e-295	-1.8e-03 -4.09e-07 -3.87e-06 -4.15e-64 1.89e-21 -0e+00 -6.44e-02 -2.02e-154	-6.3e-07 -3.89e-16 -1.39e-15 -1.08e-78 -6.03e-04 -1.18e-83 -5.75e-04 -1.33e-85	-6.3e-07 -3.89e-16 -1.39e-15 -1.63e-56 -6.03e-04 -1.18e-83 -5.75e-04 -1.33e-85	NA NA NA -1.14e-04 8.71e-13 -0e+00 1.93e-04 -9.4e-35		

ProCAIn_ss_cal_subj	-0e+00 -3.87e-61 -3.53e-66 -3.52e-241 -5.95e-27 -3.09e-20 -1.09e-33 -5.57e-174	-1.19e-256 -1.26e-28 -9.06e-35 -9.31e-195 -1.06e-23 -7.7e-62 -1.48e-31 -1.03e-163	-2.66e-188 -3.85e-32 -6.85e-34 -5.78e-159 -4.36e-24 -1.57e-137 -3.86e-38 -1.25e-183	-1.4e-285 -1.12e-21 -5.61e-34 -7.21e-210 -5.6e-34 -0e+00 -3.46e-101 -1.99e-314	-2.47e-323 -2.13e-30 -3.77e-42 -6.13e-215 -5.05e-75 4.09e-39 -6.34e-99 -3.27e-199	-1.4e-285 -1.12e-21 -5.61e-34 -3.06e-183 -2.21e-72 9.84e-23 -3.36e-100 -2.46e-136	-1.25e-254 -1.73e-17 -5.51e-26 -1.13e-155 -5.74e-39 -0e+00 -2.29e-112 -0e+00	-2.93e-169 -1.97e-14 -2.48e-17 -1.37e-120 -7.63e-36 -0e+00 -8.31e-111 -0e+00	-4.33e-59 -2.52e-05 -2.66e-09 -4.41e-85 4.96e-02 -0e+00 -1.39e-04 -3.57e-93	-7.81e-62 -3.17e-13 -4.95e-19 -4.21e-107 -1.2e-23 5.38e-224 -1.81e-05 -4.94e-22	-7.81e-62 -3.17e-13 -4.95e-19 -5.71e-83 -1.2e-23 5.38e-224 -1.81e-05 -4.94e-22	-1.04e-147 -5.47e-01 -2.61e-03 -1.33e-63 -6.47e-02 2.04e-251 5.48e-05 1.68e-20	-1.04e-147 -5.47e-01 -2.61e-03 -3.4e-13 -1.73e-13 0e+00 -1.12e-02 7.7e-48	
ProCAIn_ss_cal_subj_Promals	-0e+00 -3.87e-61 -3.53e-66 -1.77e-254 -4.22e-08 -0e+00 -1.41e-22 -1.34e-209	-1.19e-256 -1.26e-28 -9.06e-35 -1.19e-205 -3.74e-05 -0e+00 -2.54e-17 -1.36e-204	-2.66e-188 -3.85e-32 -6.85e-34 -8.71e-177 -4.99e-05 -0e+00 -7.85e-21 -1.9e-212	-1.4e-285 -1.12e-21 -5.61e-34 -2.09e-225 -1.39e-08 -0e+00 -3.43e-81 -0e+00	-2.47e-323 -2.13e-30 -3.77e-42 -1.27e-243 -5.21e-44 -1.21e-175 -1.44e-97 -4.85e-256	-1.4e-285 -1.12e-21 -5.61e-34 -6.15e-216 -1.08e-37 -2.87e-243 -1.53e-104 -2.18e-196	-1.25e-254 -1.73e-17 -5.51e-26 -1.15e-147 -1.91e-04 -0e+00 -4.73e-46 -1.08e-261	-2.93e-169 -1.97e-14 -2.48e-17 -1.47e-113 -1.81e-03 -0e+00 -7.74e-43 -1.86e-253	-4.33e-59 -2.52e-05 -2.66e-09 -3.81e-102 7.67e-16 -0e+00 -3.23e-01 -3.42e-125	-7.81e-62 -3.17e-13 -4.95e-19 -9.06e-125 -4.9e-06 -3.39e-06 -8.01e-03 -1.16e-59	-7.81e-62 -3.17e-13 -4.95e-19 -3.04e-100 -4.9e-06 -3.39e-06 -8.01e-03 -1.16e-59	-1.04e-147 -5.47e-01 -2.61e-03 -3.55e-37 5.87e-08 -8.18e-300 1.03e-06 -4.62e-11	-1.04e-147 -5.47e-01 -2.61e-03 -6.8e-72 -1.06e-16 1.36e-190 2.13e-02 2.05e-43	NA NA NA -7.3e-07 6.46e-10 -0e+00 1.92e-04 -1.24e-23

Table 7 the result of 25% sensitivity t-test for the whole dataset

3. Fifty percent sensitivity t-test

50% sensitivity	**COMPASS_en**	**COMPASS_cal**	**COMPASS_cal_subj**	**HHsearch_noss**	**HHsearch_noss_global**	**HHsearch_noss_global_local**	**ProCAIn_noss_cal**	**ProCAIn_noss_cal_subj**	**HHsearch_ss**	**HHsearch_ss_global**	**HHsearch_ss_global_local**	**ProCAIn_ss_cal**	**ProCAIn_ss_cal_Promals**	**ProCAIn_ss_cal_subj**
COMPASS_cal	-4.71e-61 -1.29e-47 -8.64e-40 -3.15e-81 -4.85e-06 2.84e-22 -1.15e-05 -7.26e-17													

COMPASS_cal_subj	-2.23e-165 -9.25e-42 -2.89e-46 -2.28e-136 -1.56e-05 6.96e-75 -2.31e-02 -1.24e-13	-3.16e-105 -3.34e-01 -2.05e-07 -4.67e-56 -2.21e-01 1.98e-142 6.67e-06 1.75e-02												
HHsearch_noss	9.01e-46 -4.6e-36 -1.11e-19 8.3e-04 7.63e-01 0e+00 2.82e-38 5.8e-155	2.09e-84 -2.33e-05 -2.97e-01 2.9e-25 5.01e-02 0e+00 8.43e-59 1.74e-232	5.96e-128 -1.08e-05 2.57e-01 8.03e-42 5.12e-02 0e+00 3.47e-53 3.53e-227											
HHsearch_noss_global	3.13e-106 -1.79e-20 -1.03e-09 2.94e-02 4.57e-04 -1.15e-297 1.35e-07 1.27e-14	5.92e-161 5.59e-01 1.11e-04 4.35e-15 1.21e-07 -0e+00 1.4e-15 1.24e-12	3.45e-194 4.67e-01 1.14e-08 2.13e-30 1.19e-07 -0e+00 1.48e-11 7.12e-09	5.14e-45 8.71e-13 8.8e-13 -7.05e-01 3.09e-06 -0e+00 -4.48e-16 -1.66e-130										
HHsearch_noss_global_local	9.01e-46 -4.6e-36 -1.11e-19 -2.7e-08 5.86e-03 -2.4e-265 9.1e-08 3.9e-01	2.09e-84 -2.33e-05 -2.97e-01 8.58e-01 1.22e-05 -0e+00 3.79e-15 8.95e-01	5.96e-128 -1.08e-05 2.57e-01 2.97e-04 1.65e-05 -0e+00 1.49e-11 -1.19e-01	NA NA NA -2.53e-41 1.33e-05 -0e+00 -1.2e-16 -4.83e-181	-5.14e-45 -8.71e-13 -8.8e-13 -1.26e-61 -1.94e-04 1.46e-45 9.89e-01 -0e+00									
ProCAIn_noss_cal	-2.57e-91 -1.1e-56 -1.74e-41 -3.53e-106 -2.68e-05 9.01e-96 -5.86e-01 -8.21e-09	-3.1e-35 -2.95e-22 -2.28e-11 -3.42e-59 -4.05e-03 3.53e-97 8.64e-02 -1.29e-01	-9.24e-02 -1.03e-18 -5.97e-05 -1.21e-31 -8.06e-03 6.17e-48 5.13e-01 -1.69e-02	-1.99e-162 -1.46e-05 -5.72e-08 -2.33e-91 -5.01e-05 -0e+00 -2.15e-46 -1.34e-187	-1.18e-204 -2.41e-13 -2.4e-18 -2.89e-81 -1.03e-15 0e+00 -3.97e-10 -1.05e-05	-1.99e-162 -1.46e-05 -5.72e-08 -3.41e-42 -8.2e-13 0e+00 -4.07e-11 1.03e-03								

ProCAIn_noss_cal_subj	-2.24e-146 -6.23e-54 -5.52e-44 -1.31e-126 -8.51e-06 1.22e-105 -2.07e-01 -1.88e-08	-2.6e-82 -5.84e-21 -4.79e-15 -9.86e-79 -1.07e-03 2.09e-108 3.93e-01 -6.86e-02	-5.63e-31 -6.9e-19 -2.94e-08 -5.23e-54 -2.37e-03 2.06e-64 -8.79e-01 -9.6e-03	-1.28e-194 -8.52e-05 -3.63e-12 -1.94e-113 -5.35e-06 -0e+00 -6.78e-49 -2.56e-176	-8.5e-232 -2.86e-15 -6.9e-24 -1.84e-99 -2.09e-16 0e+00 -1.98e-11 -5.75e-04	-1.28e-194 -8.52e-05 -3.63e-12 -7.42e-59 -2.31e-13 0e+00 -7.92e-12 1.01e-05	-9.93e-118 -5.92e-02 -8.12e-04 -2.66e-60 -8.63e-07 1.13e-61 -2.25e-03 -6.84e-02						
HHsearch_ss	-3.77e-171 -1.03e-60 -2.22e-49 -7.96e-76 -1.7e-36 0e+00 -1.79e-09 -8.91e-02	-1.39e-97 -2e-22 -3.72e-19 -1.3e-37 -1.65e-30 0e+00 -3.87e-05 -2.83e-01	-1.44e-29 -1.07e-20 -2.04e-12 -1.28e-19 -4.96e-31 0e+00 -5.31e-07 -1.29e-02	-0e+00 -7.43e-28 -7.95e-34 -5.69e-242 -1.35e-120 -0e+00 -1.93e-196 -2.62e-312	-0e+00 -1.19e-47 -1.25e-45 -2.41e-121 -1.39e-91 0e+00 -3.11e-45 -5.2e-59	-0e+00 -7.43e-28 -7.95e-34 -6.03e-83 -1.91e-89 0e+00 -1.76e-49 -8.21e-07	-5.8e-24 -1.15e-01 -5.53e-03 2.18e-01 -2.08e-23 0e+00 -1.03e-10 -1.87e-04	8.09e-01 -4.05e-02 -1.14e-01 1.46e-06 -3.43e-21 0e+00 -1.48e-09 -5.02e-06						
HHsearch_ss_global	-2.08e-98 -9.19e-40 -1.07e-30 -2.71e-79 -3.41e-08 -0e+00 -4.47e-23 -1.59e-42	-1.15e-52 -3.43e-06 -1.01e-04 -1.13e-47 -2.61e-06 -0e+00 -8.36e-20 -5.79e-52	-1.09e-11 -4.79e-05 -4.34e-02 -3.24e-29 -3.16e-06 -0e+00 -5.69e-25 -1.16e-61	-4.62e-205 -1.4e-01 -5.95e-06 -2.25e-119 -4.88e-11 -0e+00 -4.21e-134 -2.04e-320	-0e+00 -3.81e-33 -6.73e-35 -2.68e-211 -1.1e-64 -0e+00 -7.19e-162 -0e+00	-4.62e-205 -1.4e-01 -5.95e-06 -1.35e-84 -1.37e-55 -0e+00 -2.47e-148 -0e+00	-6.24e-06 3.15e-02 9.33e-01 -5.92e-02 -1.37e-02 -0e+00 -1.2e-35 -1.18e-93	1.33e-03 3.16e-02 1.12e-01 1.3e-01 -4.38e-02 -0e+00 -2.04e-34 -1.29e-100	5.52e-07 3.6e-17 1.54e-10 -1.06e-03 2.4e-18 -0e+00 -4.03e-09 -5.59e-99					
HHsearch_ss_global_local	-2.08e-98 -9.19e-40 -1.07e-30 -1.45e-122 -3.41e-08 -0e+00 -4.47e-23 -1.59e-42	-1.15e-52 -3.43e-06 -1.01e-04 -1.33e-84 -2.61e-06 -0e+00 -8.36e-20 -5.79e-52	-1.09e-11 -4.79e-05 -4.34e-02 -3.51e-60 -3.16e-06 -0e+00 -5.69e-25 -1.16e-61	-4.62e-205 -1.4e-01 -5.95e-06 -5.08e-151 -4.88e-11 -0e+00 -4.21e-134 -2.04e-320	-0e+00 -3.81e-33 -6.73e-35 -1.53e-183 -1.1e-64 -0e+00 -7.19e-162 -0e+00	-4.62e-205 -1.4e-01 -5.95e-06 -6.78e-90 -1.37e-55 -0e+00 -2.47e-148 -0e+00	-6.24e-06 3.15e-02 9.33e-01 -4.43e-12 -1.37e-02 -0e+00 -1.2e-35 -1.18e-93	1.33e-03 3.16e-02 1.12e-01 -1.01e-03 -4.38e-02 -0e+00 -2.04e-34 -1.29e-100	5.52e-07 3.6e-17 1.54e-10 -2.22e-18 2.4e-18 -0e+00 -4.03e-09 -5.59e-99	NA NA NA -2.57e-16 NA NA NA NA				
ProCAIn_ss_cal	-9.19e-248 -2.9e-81 -1.82e-70 -1.31e-266 -4.35e-44 -1.57e-148 -1.29e-58 -1.09e-239	-2.24e-176 -4.8e-49 -2.97e-41 -7.79e-218 -1.31e-40 -1.33e-259 -8.08e-60 -5.11e-255	-1.02e-91 -7.64e-40 -4.37e-26 -1.24e-176 -2.44e-41 -0e+00 -4.42e-68 -4.11e-275	-7.1e-238 -3.63e-12 -1.26e-19 -3.54e-231 -4.78e-42 -0e+00 -4.42e-166 -0e+00	-6.07e-269 -2.54e-24 -1.96e-30 -3.79e-208 -3.83e-56 1.33e-27 -6.03e-105 -1.1e-199	-7.1e-238 -3.63e-12 -1.26e-19 -1.29e-164 -2.25e-54 1.52e-16 -1.12e-100 -1.47e-149	-1.42e-139 -3.34e-12 -8.43e-14 -4.58e-170 -7.26e-45 -0e+00 -7.91e-161 -0e+00	-5.88e-38 -4.36e-13 -2.58e-10 -5.41e-108 -3.46e-40 -0e+00 -5.53e-156 -0e+00	-1.89e-29 -2.05e-03 -4.28e-03 -1.31e-85 -8.17e-01 -0e+00 -4.37e-33 -6.9e-142	-1.79e-35 -1.53e-11 -1.69e-08 -4.23e-63 -8.39e-13 4.18e-206 -3.98e-09 -1.14e-23	-1.79e-35 -1.53e-11 -1.69e-08 -3e-38 -8.39e-13 4.18e-206 -3.98e-09 -1.14e-23			

ProCAIn_ss_cal _Promals	-9.19e-248 -2.9e-81 -1.82e-70 -2.98e-262 -8.38e-08 -0e+00 -1.24e-29 -9.88e-242	-2.24e-176 -4.8e-49 -2.97e-41 -1.09e-229 -5.81e-06 -0e+00 -8.43e-30 -2.37e-258	-1.02e-91 -7.64e-40 -4.37e-26 -1.83e-190 -4.35e-06 -0e+00 -6.82e-36 -6.11e-267	-7.1e-238 -3.63e-12 -1.26e-19 -8.26e-251 -3.89e-09 -0e+00 -5.52e-139 -0e+00	-6.07e-269 -2.54e-24 -1.96e-30 -7.6e-236 -3e-24 -0e+00 -4.95e-96 -1.84e-271	-7.1e-238 -3.63e-12 -1.26e-19 -2.91e-191 -2.32e-21 -0e+00 -1.13e-93 -1.66e-225	-1.42e-139 -3.34e-12 -8.43e-14 -3.65e-133 -2.04e-03 -0e+00 -3.01e-53 -0e+00	-5.89e-38 -4.36e-13 -2.58e-10 -2.04e-98 -1.33e-02 -0e+00 -6.92e-50 -0e+00	-1.89e-29 -2.05e-03 -4.28e-03 -4.54e-131 5.98e-11 -0e+00 -9.28e-17 -2.07e-180	-1.79e-35 -1.53e-11 -1.69e-08 -7.62e-99 -7.74e-01 -7.04e-125 -5.48e-03 -6.77e-69	-1.79e-35 -1.53e-11 -1.69e-08 -1.7e-78 -7.74e-01 -7.04e-125 -5.48e-03 -6.77e-69	NA NA NA -2.84e-22 2.11e-13 -0e+00 4.47e-04 -8.94e-36		
ProCAIn_ss_cal _subj	-6.43e-300 -8.45e-87 -1.54e-80 -3.32e-298 -1.37e-47 -6.01e-66 -1.26e-49 -3.6e-198	-3.69e-238 -4.33e-49 -1.16e-46 -8.6e-253 -2.86e-44 -5.53e-149 -3.34e-51 -5.54e-208	-1.19e-191 -3.95e-45 -1.1e-35 -1.18e-225 -1.95e-44 -9.05e-227 -5.33e-58 -2.19e-227	-1.27e-261 -1.21e-12 -1.34e-27 -3.6e-252 -1.37e-42 -0e+00 -3.52e-153 -0e+00	-1.69e-283 -2.66e-25 -1.62e-37 -2.88e-233 -3.2e-56 4.99e-75 -3.79e-94 -2.4e-171	-1.27e-261 -1.21e-12 -1.34e-27 -3.16e-193 -1.01e-55 9.71e-59 -2.12e-91 -2.33e-118	-7.62e-254 -5.45e-14 -6.21e-27 -2.4e-231 -7.9e-50 -0e+00 -3.79e-141 -0e+00	-1.63e-189 -3.59e-15 -1.28e-21 -3.47e-191 -1.81e-46 -0e+00 -1.94e-141 -0e+00	-3.73e-94 -1.39e-03 -8.99e-08 -3.73e-121 -5.1e-01 -0e+00 -5.27e-29 -4.54e-116	-2e-95 -1.47e-12 -3.46e-16 -3.53e-101 -2e-13 1.26e-281 -1.49e-06 -9.93e-13	-2e-95 -1.47e-12 -3.46e-16 -1.66e-70 -2e-13 1.26e-281 -1.49e-06 -9.93e-13	-1.28e-179 -1.56e-01 -2.39e-06 -3.57e-84 -4.31e-05 5.23e-194 9.73e-02 1.37e-21	-1.28e-179 -1.56e-01 -2.39e-06 -1e-02 -5.85e-14 0e+00 -5.84e-03 1.19e-47	
ProCAIn_ss_cal _subj_Promals	-6.43e-300 -8.45e-87 -1.54e-80 -2.54e-293 -1.33e-09 -0e+00 -2.79e-24 -1.04e-201	-3.69e-238 -4.33e-49 -1.16e-46 -6.13e-266 -1.15e-07 -0e+00 -1.35e-23 -1.3e-214	-1.19e-191 -3.95e-45 -1.1e-35 -2.34e-240 -6.94e-08 -0e+00 -2.64e-29 -1.88e-224	-1.27e-261 -1.21e-12 -1.34e-27 -1.07e-284 -4.1e-11 -0e+00 -7.84e-126 -0e+00	-1.69e-283 -2.66e-25 -1.62e-37 -4.45e-266 -8.92e-28 -3.48e-263 -2.92e-86 -2.27e-231	-1.27e-261 -1.21e-12 -1.34e-27 -1.05e-226 -7.43e-25 -0e+00 -2.08e-83 -1.6e-180	-7.62e-254 -5.45e-14 -6.21e-27 -1.51e-190 -5.71e-05 -0e+00 -1.75e-45 -6.82e-303	-1.63e-189 -3.59e-15 -1.28e-21 -1.77e-161 -4.03e-04 -0e+00 -2.54e-42 -1.13e-300	-3.73e-94 -1.39e-03 -8.99e-08 -9.68e-177 7.28e-09 -0e+00 -2.99e-13 -7.22e-149	-2e-95 -1.47e-12 -3.46e-16 -1.59e-145 -2.7e-01 -2.69e-42 -1.54e-01 -1.75e-42	-2e-95 -1.47e-12 -3.46e-16 -2.87e-120 -2.7e-01 -2.69e-42 -1.54e-01 -1.75e-42	-1.28e-179 -1.56e-01 -2.39e-06 -4.86e-78 7.83e-11 -0e+00 8.47e-06 -1.04e-14	-1.28e-179 -1.56e-01 -2.39e-06 -5.56e-93 -3.75e-08 5.15e-126 8.72e-04 1.15e-63	NA NA NA -2.03e-26 6.32e-11 -0e+00 8.9e-05 -8.74e-26

Table 8 the result of 50% sensitivity t-test for the whole dataset

iii. Alignment quality

1. Accuracy

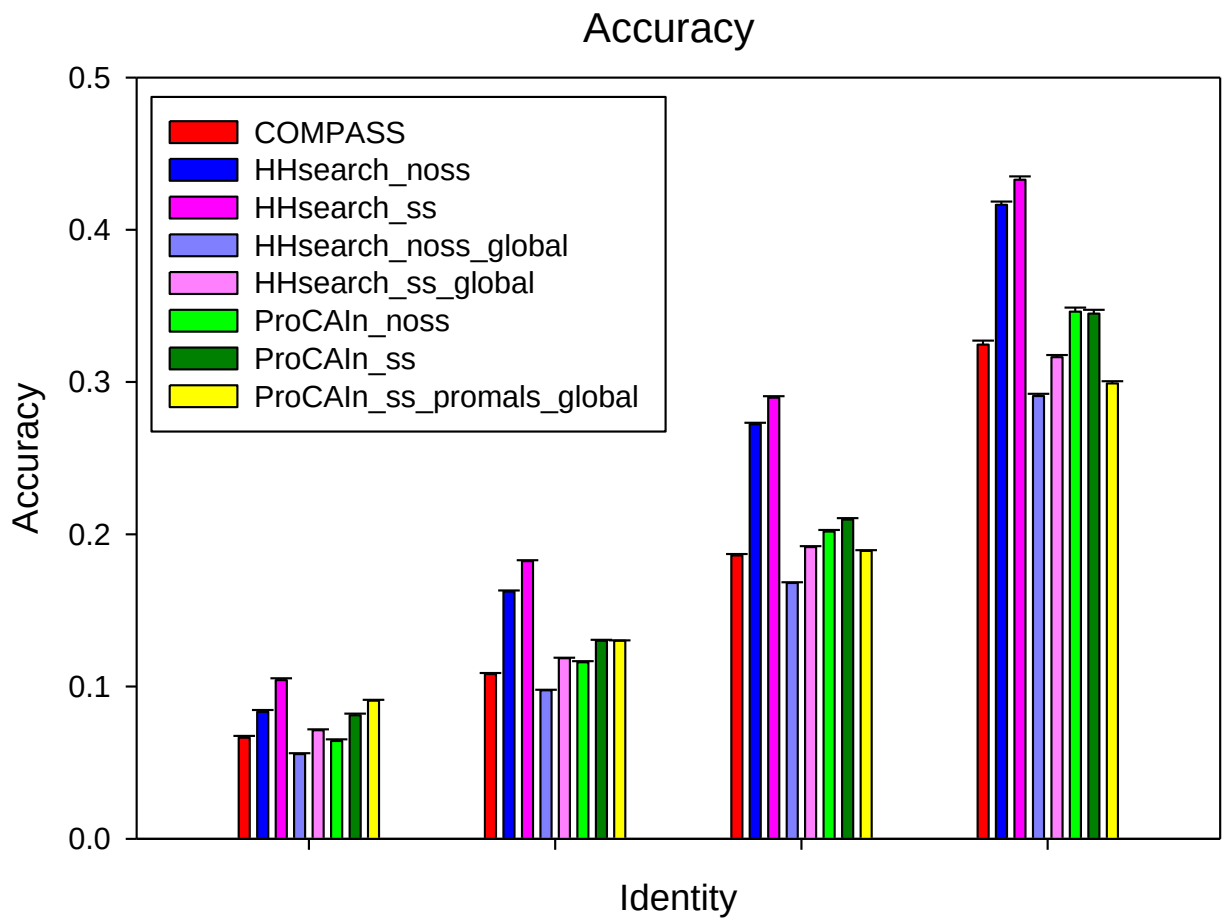

Figure 84 Accuracy of all Bench Marked Methods

2. Coverage

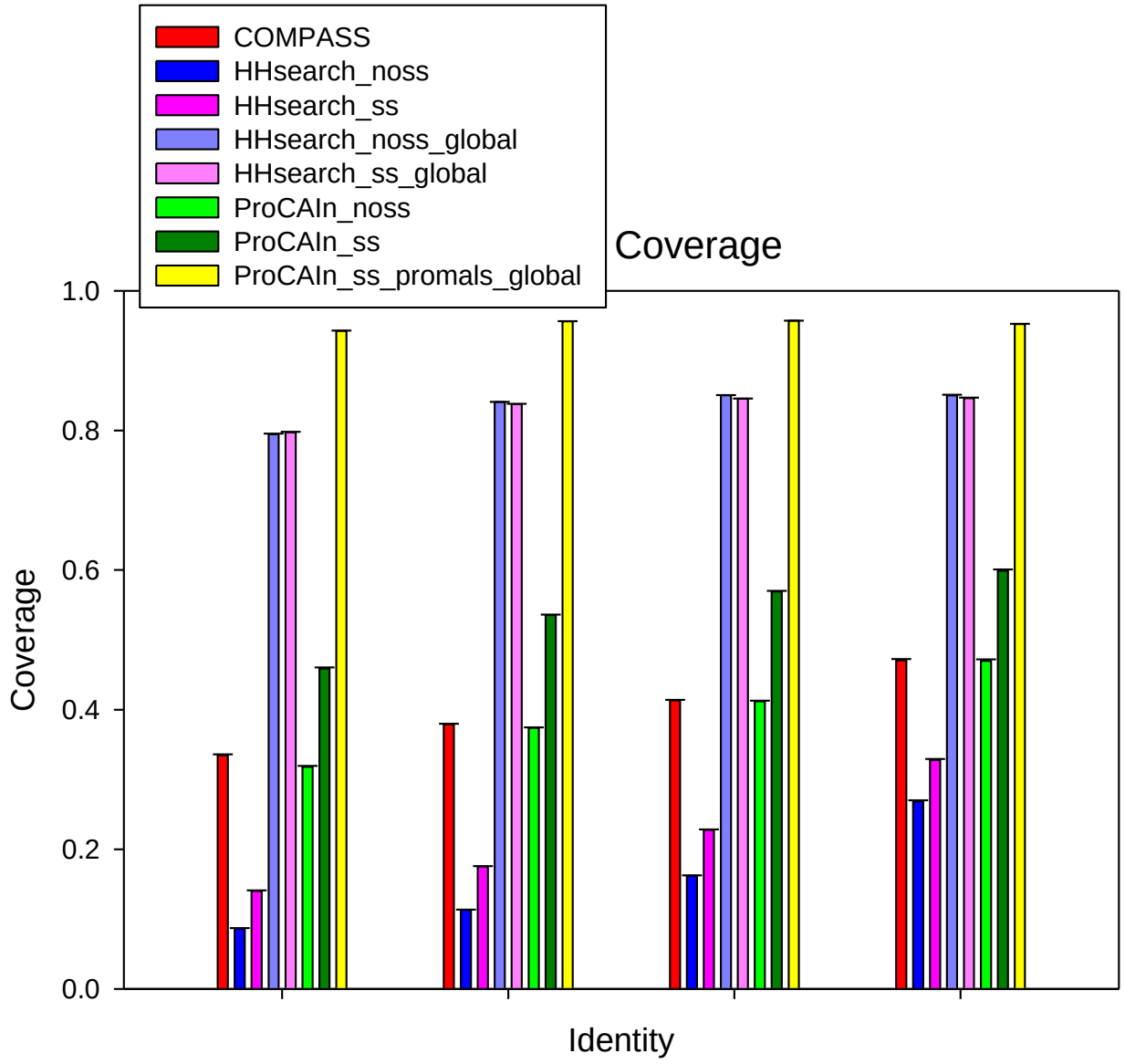

Figure 85 Coverage of all Bench Marked Methods

3. Q-modeler

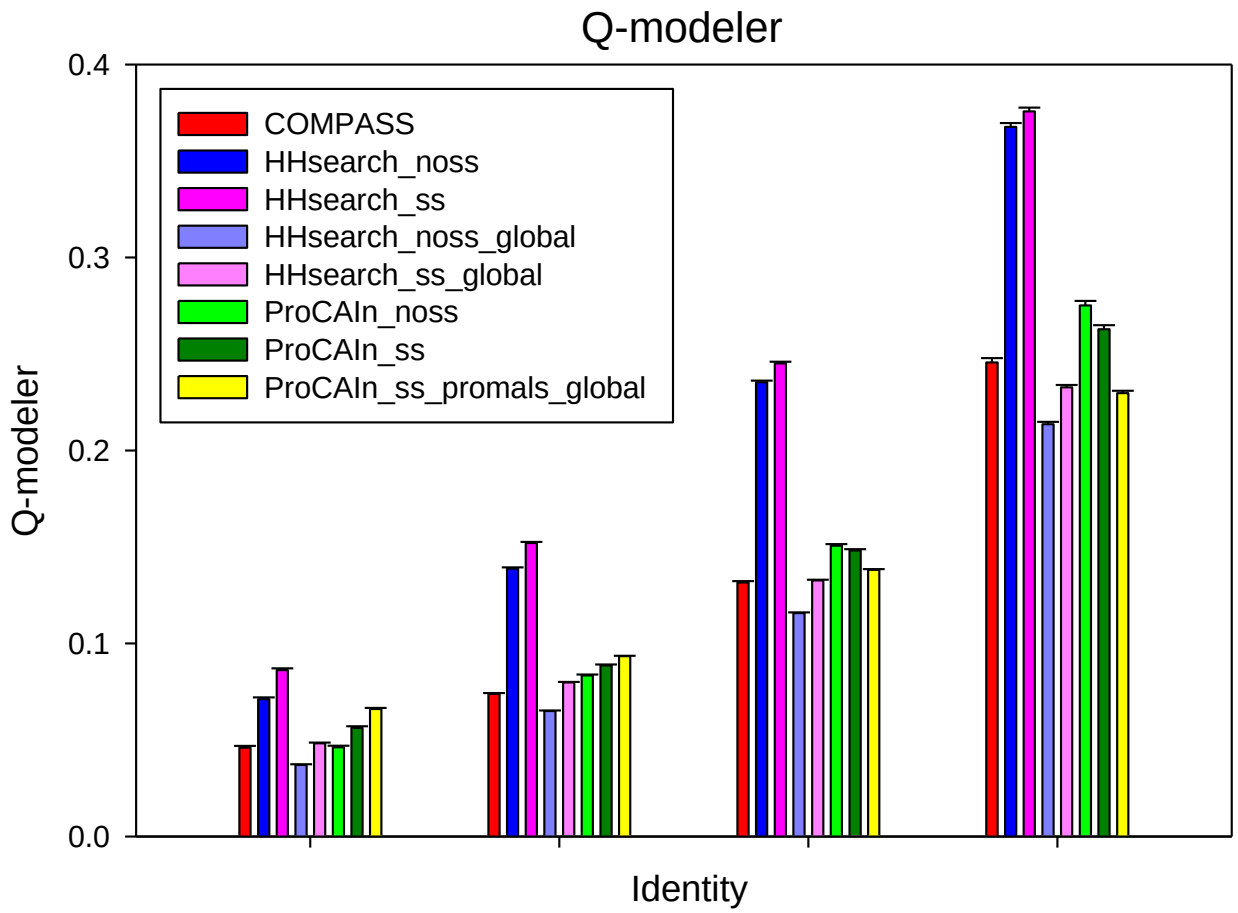

Figure 86 Q-modeler of all Bench Marked Methods

4. Q-developer

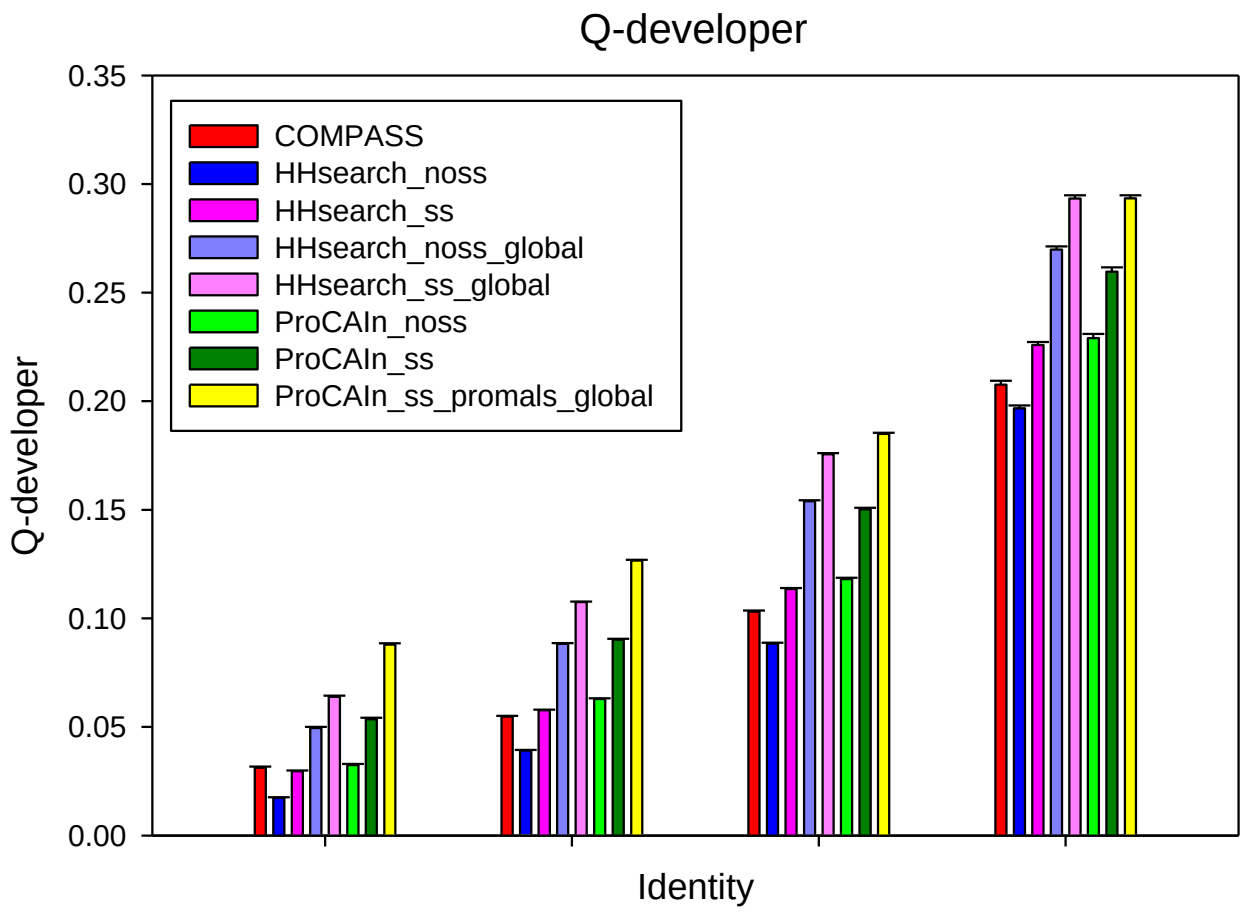

Figure 87 Q-developer of all Bench Marked Methods

5. Q-combined

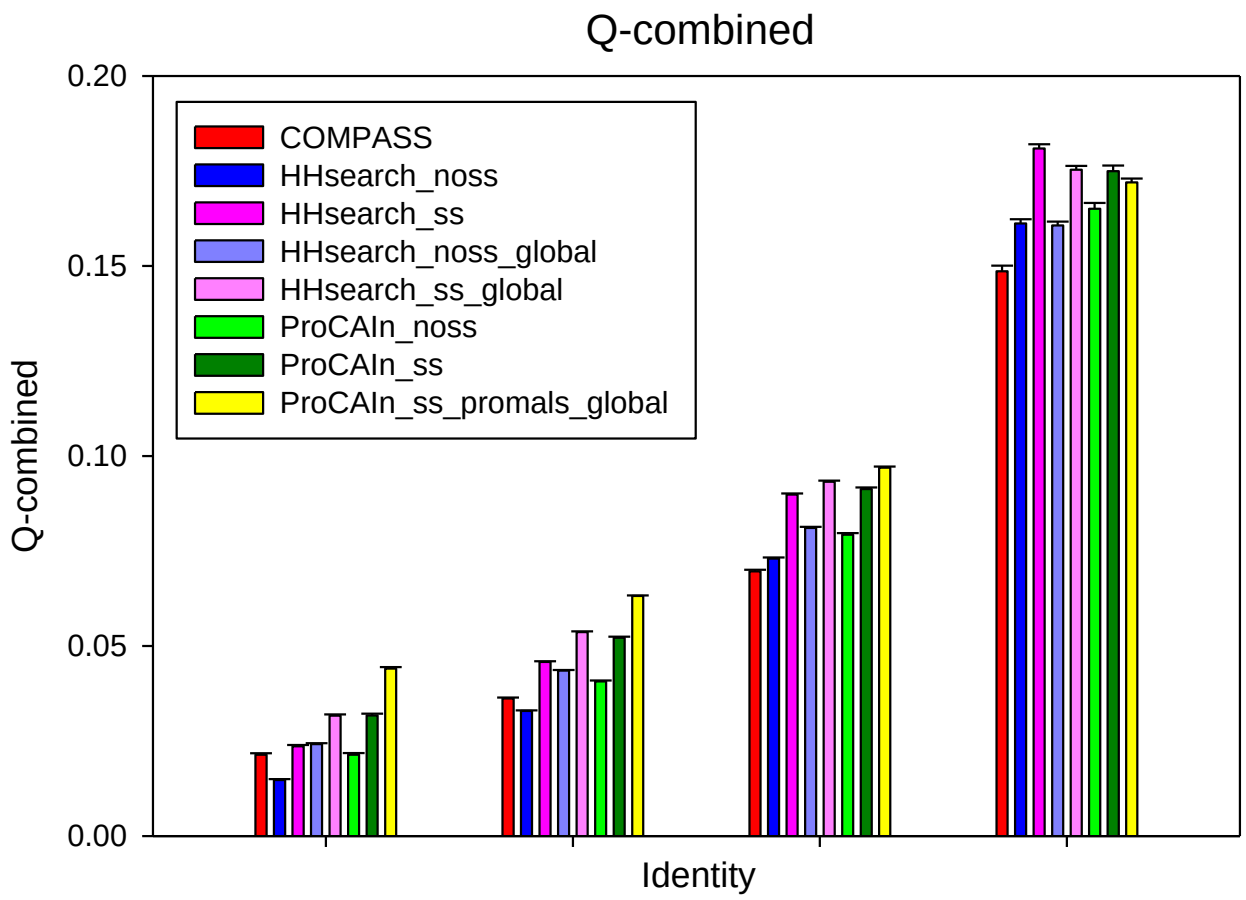

Figure 88 Q-combined of all Bench Marked Methods

6. Average global GDT_TS

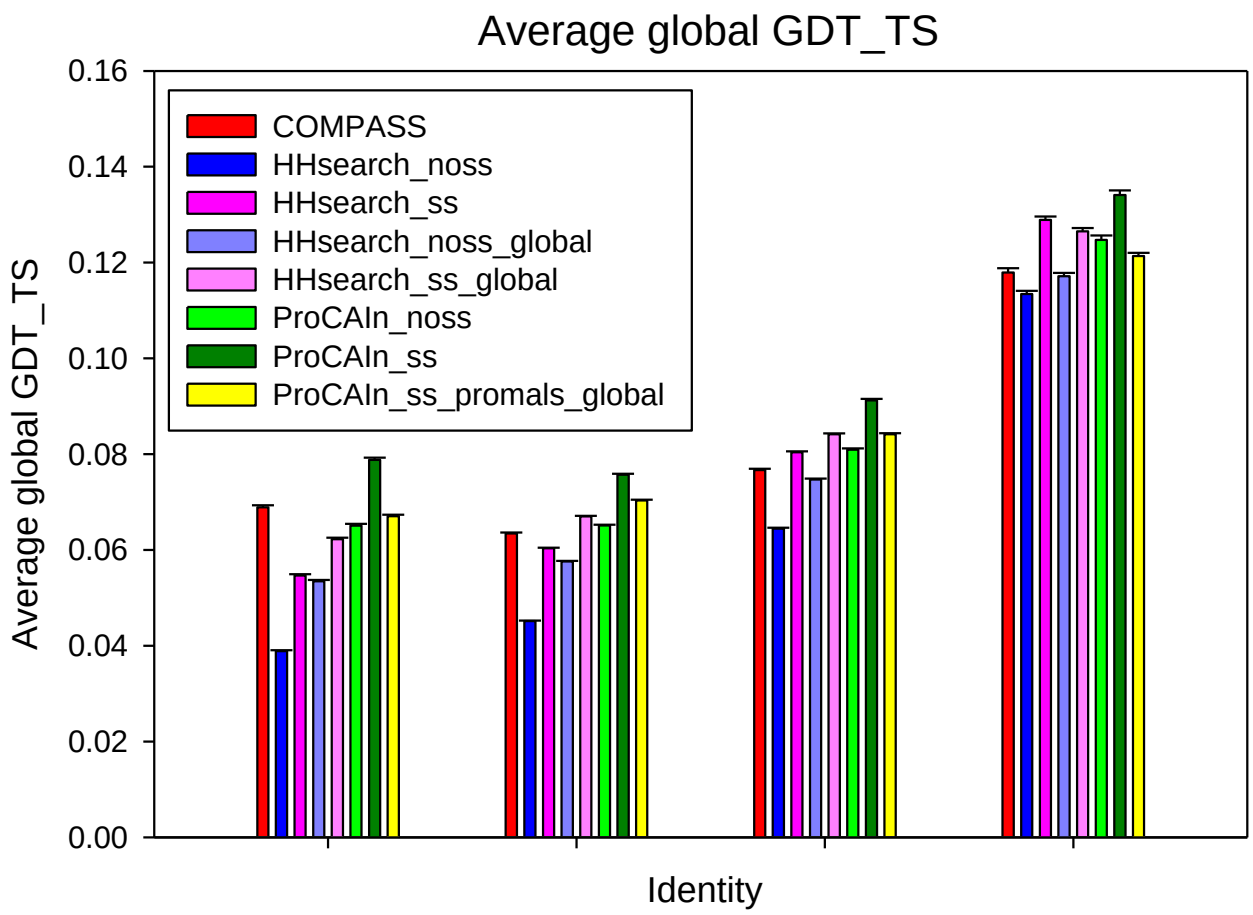

Figure 89 Average GDT_TS of all Bench Marked Methods

7. Average global LGA GDT_TS

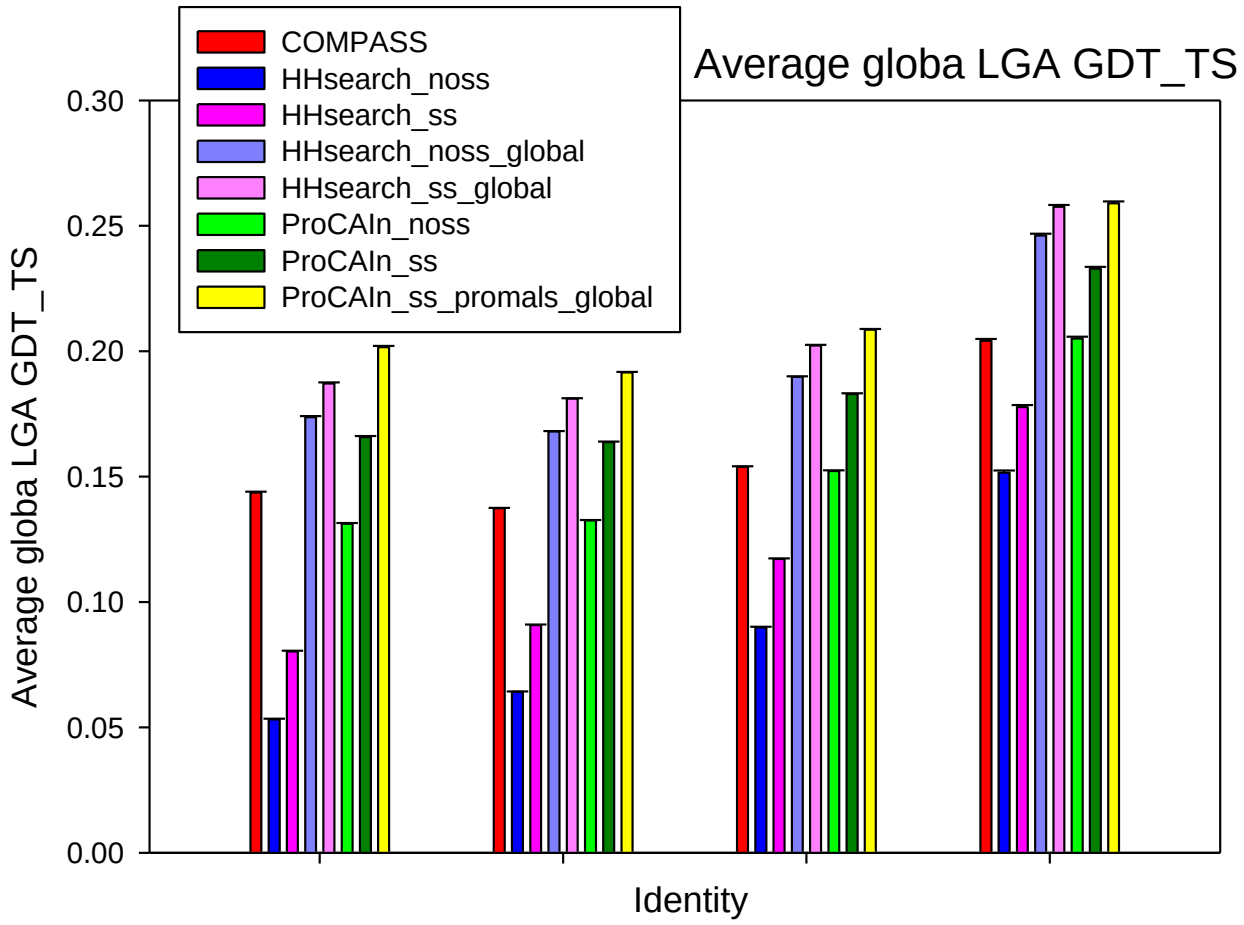

Figure 90 Average LGA GDT_TS of all Bench Marked Methods

8. Average global Live Bench Contact-a

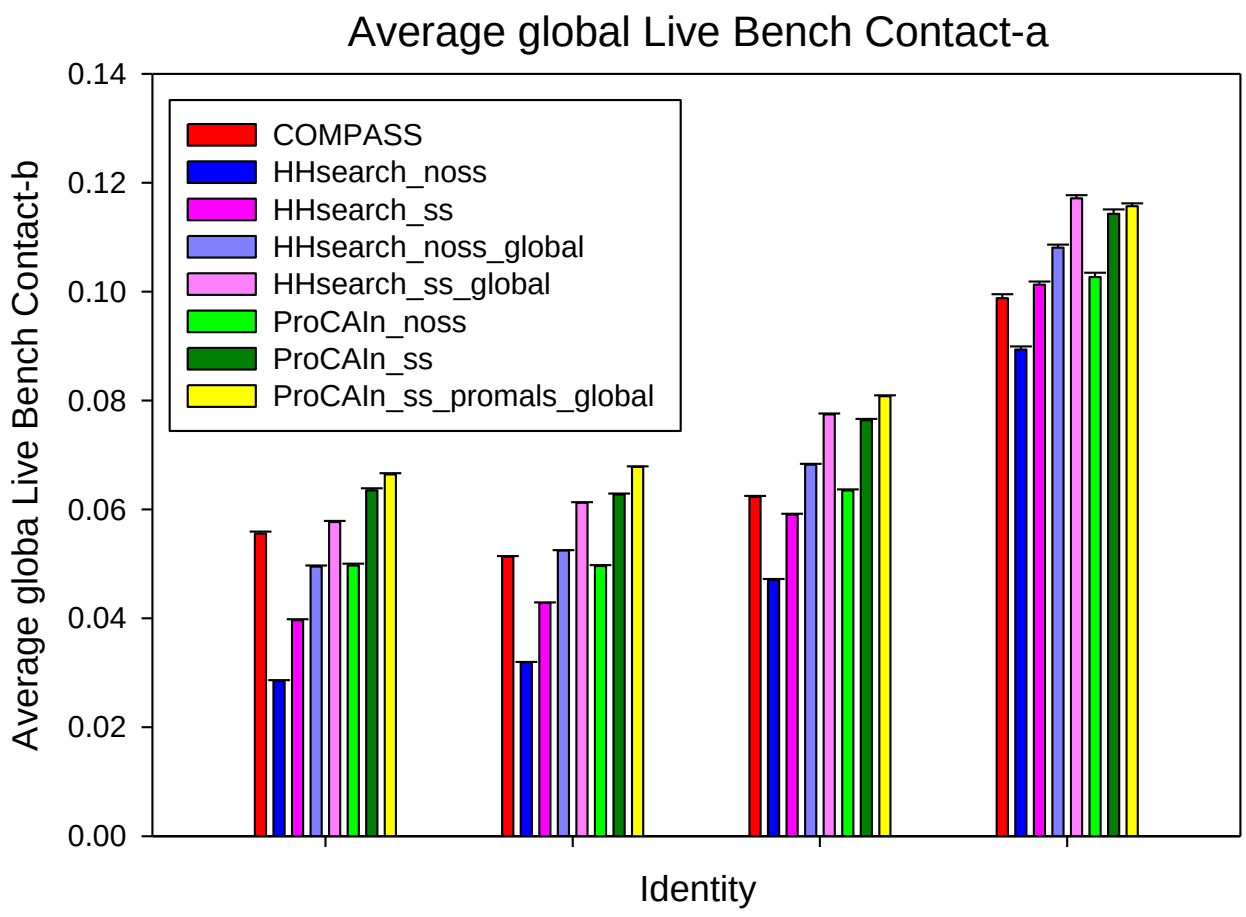

Figure 91 Average Live Bench Contact-a of all Bench Marked Methods

9. Average global Live Bench Contact-b

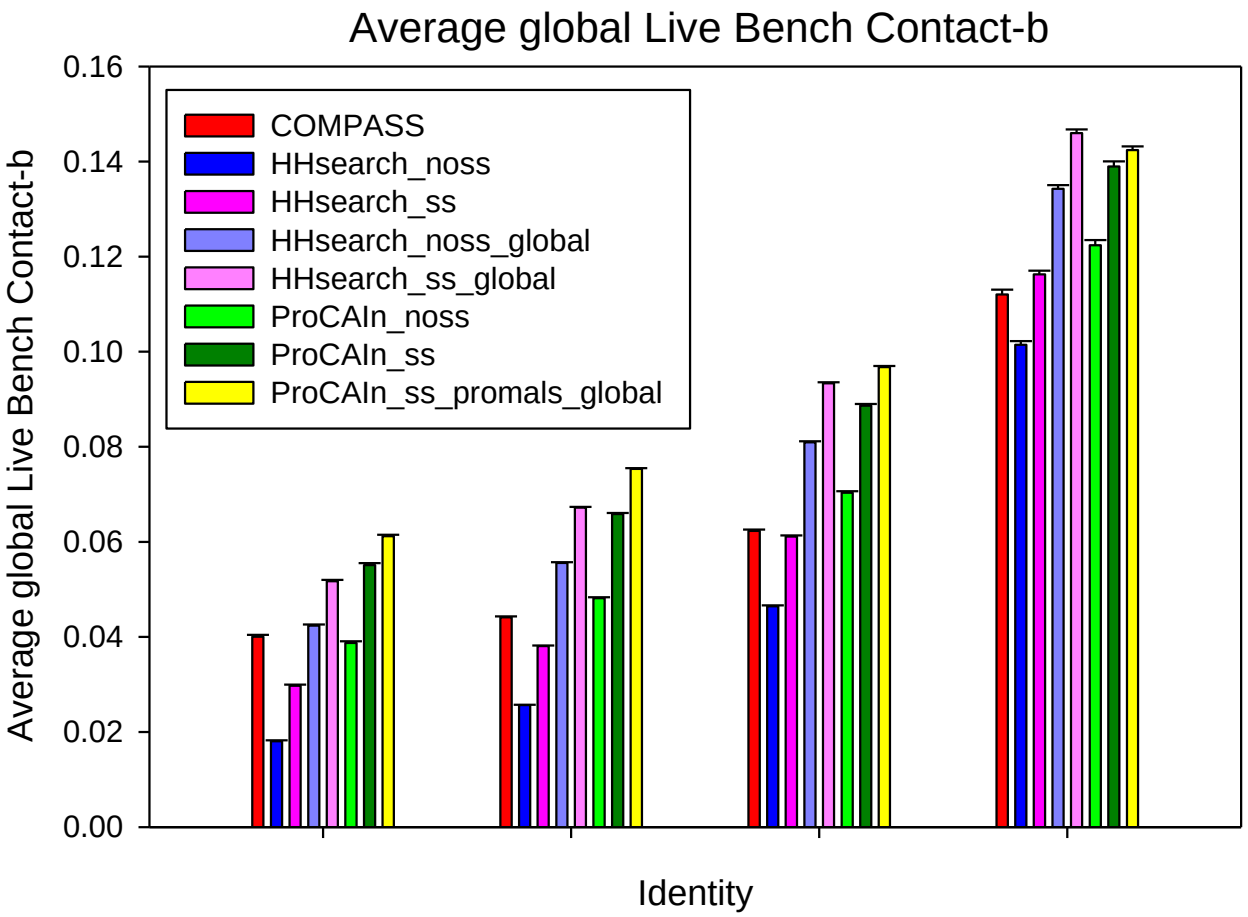

Figure 92 Average Live Bench Contact-b of all Bench Marked Methods

List 1 ProCAIn_ss outperforms HHsearch_ss:

(The first row is ProCAIn results and the second row is the corresponding HHsearch results)

ID1	ID2	SVM	E-value/Prob	Score	GDT_TS	SF	class1	class2
d1qx4a2	d1nrjb_	0	6.40e-03	159.24	0.1769	-1	c	c
d1qx4a2	d1nrjb_	0	0.220000	-0.49	0.0255	-1	c	c
d1rifa_	d1rjda_	1	5.84e-03	218.38	0.18	-1	c	c
d1rifa_	d1rjda_	1	0.290000	-0.82	0.0426	-1	c	c
d1ebfa1	d1e4ea1	1	2.37e-03	161.75	0.247	-1	c	c
d1ebfa1	d1e4ea1	1	0.350000	1.9	0.105	-1	c	c
d1sgja_	d1l6wa_	1	1.58e-03	177.84	0.2965	-1	c	c
d1sgja_	d1l6wa_	1	0.360000	-5.93	0.2251	-1	c	c
d1e4ea1	d1ebfa1	1	4.36e-04	161.75	0.3212	-1	c	c
d1e4ea1	d1ebfa1	1	0.380000	1.9	0.1365	-1	c	c
d1qhxa_	d1p9oa_	1	7.68e-03	182.58	0.198	-1	c	c
d1qhxa_	d1p9oa_	1	0.400000	2.42	0.0674	-1	c	c
d1ddga2	d1nrjb_	1	7.26e-03	161.38	0.2239	-1	c	c
d1ddga2	d1nrjb_	1	0.490000	-2.25	0.0931	-1	c	c
d1ei9a_	d1woha_	0	4.35e-03	197.6	0.1667	-1	c	c
d1ei9a_	d1woha_	0	0.490000	0.11	0.1111	-1	c	c
d1vjga_	d1nrjb_	1	2.69e-03	173.61	0.2251	-1	c	c
d1vjga_	d1nrjb_	1	0.520000	-2.14	0.0423	-1	c	c
d1e5ka_	d1ooea_	0	4.50e-03	156.02	0.1809	-1	c	c
d1e5ka_	d1ooea_	0	0.540000	-0.13	0.1436	-1	c	c
d1nrjb_	d1vjga_	1	3.59e-03	173.61	0.2165	-1	c	c
d1nrjb_	d1vjga_	1	0.540000	-2.14	0.0407	-1	c	c
d1a8p_2	d1h7wa4	1	2.15e-03	184.03	0.1693	-1	c	c
d1a8p_2	d1h7wa4	1	0.560000	3.58	0.1392	-1	c	c
d1twda_	d1keka1	0	6.54e-03	188.58	0.1569	-1	c	c
d1twda_	d1keka1	0	0.580000	2.06	0.169	-1	c	c
d1h7wa4	d1khta_	1	8.81e-03	167.6	0.1671	-1	c	c
d1h7wa4	d1khta_	1	0.630000	0.01	0.1046	-1	c	c
d1nrjb_	d1db3a_	1	6.46e-03	181.15	0.189	-1	c	c
d1nrjb_	d1db3a_	1	0.650000	2.51	0.0813	-1	c	c
d1d1qa_	d1h65a_	1	9.74e-03	172.07	0.2374	-1	c	c

d1d1qa_	d1h65a_	1	0.660000	3.3	0.1132	-1	c	c
d1ryba_	**d1h65a_**	1	9.27e-03	173.27	0.2762	-1	c	c
d1ryba_	d1h65a_	1	0.680000	-4.09	0.1924	-1	c	c
d1mxia_	**d1b7go1**	1	7.07e-03	139.29	0.2468	-1	c	c
d1mxia_	d1b7go1	1	0.730000	1.91	0.2067	-1	c	c
d1nija1	**d1b7go1**	0	1.12e-03	168.19	0.1712	-1	c	c
d1nija1	d1b7go1	0	0.750000	3.38	0.0687	-1	c	c
d1vjra_	**d1rqba2**	0	9.32e-03	221.09	0.1504	-1	c	c
d1vjra_	d1rqba2	0	0.800000	0.27	0.0738	-1	c	c
d1vjra_	**d1booa_**	1	5.06e-03	212.14	0.2021	-1	c	c
d1vjra_	d1booa_	1	0.850000	3.72	0.091	-1	c	c
d1woha_	**d1ei9a_**	0	3.83e-03	197.6	0.1535	-1	c	c
d1woha_	d1ei9a_	0	0.860000	0.11	0.1023	-1	c	c
d1ns5a_	**d1dih_1**	0	5.37e-03	155.25	0.2288	-1	c	c
d1ns5a_	d1dih_1	0	0.860000	4.53	0.0899	-1	c	c
d1l6wa_	**d1sgja_**	1	1.83e-03	177.84	0.3114	-1	c	c
d1l6wa_	d1sgja_	1	0.860000	-5.93	0.2364	-1	c	c
d1f6ba_	**d1e8ca3**	1	3.21e-03	173.93	0.2043	-1	c	c
d1f6ba_	d1e8ca3	1	0.900000	2.94	0.0833	-1	c	c
d1r6wa1	**d1m6ya2**	1	3.47e-03	168.57	0.2014	-1	c	c
d1r6wa1	d1m6ya2	1	0.920000	-1.36	0.1753	-1	c	c
d1bg6_2	**d1f6ba_**	1	6.45e-03	160.6	0.2242	-1	c	c
d1bg6_2	d1f6ba_	1	0.930000	5.11	0.0856	-1	c	c
d1p1ma2	**d1bx4a_**	0	6.50e-03	218.3	0.1655	-1	c	c
d1p1ma2	d1bx4a_	0	0.940000	1.39	0.089	-1	c	c
d1p3da3	**d1svsa1**	1	1.97e-04	188.43	0.1674	-1	c	c
d1p3da3	d1svsa1	1	0.940000	6.33	0.0547	-1	c	c
d1e5ka_	**d1cyda_**	1	5.74e-03	155.94	0.2074	-1	c	c
d1e5ka_	d1cyda_	1	0.960000	1.14	0.0864	-1	c	c
d1booa_	**d1vjra_**	1	7.33e-04	212.14	0.1648	-1	c	c
d1booa_	d1vjra_	1	0.990000	3.72	0.0742	-1	c	c
d1e5ka_	**d1hdoa_**	0	3.72e-05	196.66	0.2194	-1	c	c
d1e5ka_	d1hdoa_	0	1.010000	2.09	0.1051	-1	c	c
d1egaa1	**d1p3da3**	1	1.02e-03	185.65	0.2221	-1	c	c
d1egaa1	d1p3da3	1	1.020000	5.73	0.0894	-1	c	c
d1egaa1	**d1e8ca3**	1	4.34e-04	196.5	0.2612	-1	c	c
d1egaa1	d1e8ca3	1	1.120000	5.9	0.0908	-1	c	c
d1odza_	**d1adoa_**	1	6.06e-03	193.58	0.1831	-1	c	c

d1odza_	d1adoa_	1	1.120000	-1.9	0.1358	-1	c	c
d1vhqa_	d1dlja2	1	8.12e-03	149.68	0.2316	-1	c	c
d1vhqa_	d1dlja2	1	1.210000	1.56	0.0703	-1	c	c
d1ic6a_	d1kyha_	1	7.84e-04	210.21	0.1828	-1	c	c
d1ic6a_	d1kyha_	1	1.230000	1.55	0.138	-1	c	c
d1b7go1	d1nija1	0	6.58e-03	168.19	0.2123	-1	c	c
d1b7go1	d1nija1	0	1.230000	3.38	0.0852	-1	c	c
d1sur__	d1f8fa2	1	3.82e-03	164.01	0.214	-1	c	c
d1sur__	d1f8fa2	1	1.250000	-3.78	0.1756	-1	c	c
d1q74a_	d1g3qa_	1	7.84e-03	179.18	0.154	-1	c	c
d1q74a_	d1g3qa_	1	1.350000	3.34	0.0623	-1	c	c
d1nrjb_	d1eq2a_	1	8.08e-03	172.03	0.1543	-1	c	c
d1nrjb_	d1eq2a_	1	1.370000	4.17	0.0849	-1	c	c
d1e8ca3	d1f6ba_	1	1.31e-03	173.93	0.1624	-1	c	c
d1e8ca3	d1f6ba_	1	1.370000	2.94	0.0662	-1	c	c
d1qhxa_	d1ks9a2	0	9.10e-03	150.82	0.1938	-1	c	c
d1qhxa_	d1ks9a2	0	1.410000	0.63	0.0885	-1	c	c
d1ctqa_	d1b7go1	1	6.60e-03	139.67	0.3012	-1	c	c
d1ctqa_	d1b7go1	1	1.410000	6.9	0.0783	-1	c	c
d1ks9a2	d1bif_1	1	6.69e-03	174.51	0.2051	-1	c	c
d1ks9a2	d1bif_1	1	1.460000	-0.83	0.1437	-1	c	c
d1gsa_1	d1t4ba1	1	9.50e-03	133.13	0.3197	-1	c	c
d1gsa_1	d1t4ba1	1	1.530000	4.89	0.1824	-1	c	c
d1v4va_	d1k6ja_	1	4.46e-03	214.8	0.1542	-1	c	c
d1v4va_	d1k6ja_	1	1.540000	5.27	0.057	-1	c	c
d1k6ja_	d1v4va_	1	1.70e-03	214.8	0.1648	-1	c	c
d1k6ja_	d1v4va_	1	1.560000	5.27	0.0609	-1	c	c
d1svsa1	d1p3da3	1	5.33e-04	188.43	0.1837	-1	c	c
d1svsa1	d1p3da3	1	1.570000	6.33	0.0599	-1	c	c
d1d4aa_	d1b16a_	1	3.89e-03	172.04	0.2006	-1	c	c
d1d4aa_	d1b16a_	1	1.590000	0.42	0.0916	-1	c	c
d1p3da3	d1f6ba_	1	2.67e-04	184.93	0.1767	-1	c	c
d1p3da3	d1f6ba_	1	1.610000	5.81	0.0802	-1	c	c
d1auoa_	d1x9ga_	1	3.78e-03	169.15	0.2259	-1	c	c
d1auoa_	d1x9ga_	1	1.620000	-0.01	0.0321	-1	c	c
d1p3da3	d1egaa1	1	1.04e-03	185.65	0.1849	-1	c	c
d1p3da3	d1egaa1	1	1.670000	5.73	0.0744	-1	c	c
d1h7wa4	d1qx4a2	1	2.20e-03	165.57	0.1722	-1	c	c

d1h7wa4	d1qx4a2	1	1.690000	5.28	0.1224	-1	c	c
d1nrjb_	d1hdoa_	1	5.63e-03	161.77	0.2022	-1	c	c
d1nrjb_	d1hdoa_	1	1.750000	5.13	0.0754	-1	c	c
d1efpa1	d1u0la2	1	6.65e-03	162.36	0.2609	-1	c	c
d1efpa1	d1u0la2	1	1.820000	-0.34	0.2336	-1	c	c
d1hdoa_	d1e5ka_	0	7.72e-04	196.66	0.2012	-1	c	c
d1hdoa_	d1e5ka_	0	1.860000	2.09	0.0963	-1	c	c
d1cqxa3	d1c0pa1	1	6.81e-03	198.68	0.1743	-1	c	c
d1cqxa3	d1c0pa1	1	1.880000	6.25	0.1303	-1	c	c
d1khta_	d1cqxa3	0	6.35e-03	159.89	0.1513	-1	c	c
d1khta_	d1cqxa3	0	1.880000	2.34	0.1	-1	c	c
d1nrjb_	d1es9a_	1	1.99e-03	178.88	0.2667	-1	c	c
d1nrjb_	d1es9a_	1	1.890000	-1.44	0.1435	-1	c	c
d1i9ga_	d1ddga2	1	6.52e-03	182.26	0.1714	-1	c	c
d1i9ga_	d1ddga2	1	1.910000	-4.41	0.0473	-1	c	c
d1f6ba_	d1h7wa4	1	8.26e-03	172.5	0.1828	-1	c	c
d1f6ba_	d1h7wa4	1	1.930000	7.68	0.0887	-1	c	c
d1f6ba_	d1p3da3	1	6.51e-04	184.93	0.2043	-1	c	c
d1f6ba_	d1p3da3	1	1.970000	5.81	0.0927	-1	c	c
d1qo2a_	d1gzga_	1	2.43e-03	194.63	0.3589	-1	c	c
d1qo2a_	d1gzga_	1	1.970000	-0.11	0.3475	-1	c	c
d1gky__	d1p3da3	1	8.77e-03	161	0.1694	-1	c	c
d1gky__	d1p3da3	1	2.040000	6.03	0.1492	-1	c	c
d1khta_	d1ooea_	1	7.09e-03	157.6	0.1763	-1	c	c
d1khta_	d1ooea_	1	2.070000	3.71	0.1158	-1	c	c
d1rqla_	d1e19a_	1	5.91e-03	183.87	0.1518	-1	c	c
d1rqla_	d1e19a_	1	2.080000	2.59	0.0827	-1	c	c
d1m65a_	d1viza_	1	5.66e-03	196.23	0.166	-1	c	c
d1m65a_	d1viza_	1	2.080000	-4.63	0.0707	-1	c	c
d1dcta_	d1dih_1	1	4.41e-03	167.3	0.1782	-1	c	c
d1dcta_	d1dih_1	1	2.130000	3.51	0.0911	-1	c	c
d1obfo1	d1nija1	0	4.86e-03	165.86	0.1868	-1	c	c
d1obfo1	d1nija1	0	2.140000	4.45	0.1451	-1	c	c
d1bif_1	d1ks9a2	1	2.04e-03	174.51	0.1608	-1	c	c
d1bif_1	d1ks9a2	1	2.200000	-0.83	0.1127	-1	c	c
d1cqxa3	d1h7wa4	1	4.44e-03	175.59	0.2324	-1	c	c
d1cqxa3	d1h7wa4	1	2.310000	3.58	0.1673	-1	c	c
d1e6ca_	d1ks9a2	0	6.03e-03	148.67	0.1882	-1	c	c

d1e6ca_	d1ks9a2	0	2.310000	5.93	0.1132	-1	c	c
d1dbta_	**d1x7fa2**	1	1.96e-03	196.97	0.2911	-1	c	c
d1dbta_	d1x7fa2	1	2.390000	2.39	0.173	-1	c	c
d1d1qa_	**d1f6ba_**	1	7.04e-03	149.54	0.1918	-1	c	c
d1d1qa_	d1f6ba_	1	2.410000	8.19	0.1038	-1	c	c
d1dbta_	**d1aw1a_**	1	7.67e-04	222.28	0.3027	-1	c	c
d1dbta_	d1aw1a_	1	2.420000	5.27	0.1508	-1	c	c
d1v8aa_	**d1li4a2**	1	5.73e-03	174.63	0.1525	-1	c	c
d1v8aa_	d1li4a2	1	2.440000	-3.37	0.107	-1	c	c
d1ly1a_	**d1p3da3**	0	7.69e-03	163.81	0.2171	-1	c	c
d1ly1a_	d1p3da3	0	2.450000	3.7	0.1135	-1	c	c
d1nar__	**d1u1ja1**	1	7.14e-03	263.45	0.2751	-1	c	c
d1nar__	d1u1ja1	1	2.460000	6.58	0.2569	-1	c	c
d1qhxa_	**d1f14a2**	0	8.37e-03	164.77	0.1938	-1	c	c
d1qhxa_	d1f14a2	0	2.460000	-2.14	0.1419	-1	c	c
d1b7go1	**d1ctqa_**	1	7.22e-03	139.67	0.2793	-1	c	c
d1b7go1	d1ctqa_	1	2.520000	6.9	0.0726	-1	c	c
d1g2qa_	**d1nria_**	1	9.30e-03	178.82	0.2626	-1	c	c
d1g2qa_	d1nria_	1	2.530000	5.87	0.243	-1	c	c
d1omza_	**d1fmta2**	1	9.20e-03	160.71	0.1623	-1	c	c
d1omza_	d1fmta2	1	2.550000	4.39	0.1462	-1	c	c
d1gzga_	**d1qo2a_**	1	4.99e-04	194.63	0.2629	-1	c	c
d1gzga_	d1qo2a_	1	2.560000	-0.11	0.2546	-1	c	c
d1ju3a2	**d8abp__**	1	5.89e-03	230.21	0.1628	-1	c	c
d1ju3a2	d8abp__	1	2.660000	0.21	0.08	-1	c	c
d1sqsa_	**d1jeyb2**	1	4.97e-03	185.78	0.1853	-1	c	c
d1sqsa_	d1jeyb2	1	2.860000	3.21	0.1293	-1	c	c
d1kyha_	**d1ic6a_**	1	2.59e-04	210.21	0.1855	-1	c	c
d1kyha_	d1ic6a_	1	2.870000	1.55	0.14	-1	c	c
d1khta_	**d1p9oa_**	1	2.74e-03	196.63	0.2026	-1	c	c
d1khta_	d1p9oa_	1	2.900000	-0.99	0.1145	-1	c	c
d1nija1	**d1b16a_**	1	3.75e-03	170.43	0.152	-1	c	c
d1nija1	d1b16a_	1	3.020000	4.58	0.143	-1	c	c
d1e8ca3	**d1egaa1**	1	4.20e-04	196.5	0.1998	-1	c	c
d1e8ca3	d1egaa1	1	3.030000	5.9	0.0694	-1	c	c
d1hdoa_	**d1nrjb_**	1	6.40e-03	161.77	0.2061	-1	c	c
d1hdoa_	d1nrjb_	1	3.050000	5.13	0.0768	-1	c	c
d1knqa_	**d1hdoa_**	1	8.33e-03	158.55	0.329	-1	c	c

d1knqa_	d1hdoa_	1	3.060000	7.73	0.1067	-1	c	c
d1es9a_	d1nrjb_	1	1.75e-03	178.88	0.263	-1	c	c
d1es9a_	d1nrjb_	1	3.070000	-1.44	0.1415	-1	c	c
d1nija1	d1m66a2	0	2.95e-03	172.2	0.1689	-1	c	c
d1nija1	d1m66a2	0	3.070000	2.94	0.0698	-1	c	c
d1h7wa4	d1f6ba_	1	1.29e-03	172.5	0.1735	-1	c	c
d1h7wa4	d1f6ba_	1	3.110000	7.68	0.0842	-1	c	c
d1gky__	d1dih_1	1	6.42e-03	151.83	0.1667	-1	c	c
d1gky__	d1dih_1	1	3.140000	6.01	0.1452	-1	c	c
d1m66a2	d1nija1	0	6.10e-03	172.2	0.1984	-1	c	c
d1m66a2	d1nija1	0	3.160000	2.94	0.082	-1	c	c
d1aw1a_	d1okkd2	1	4.98e-03	174.71	0.1784	-1	c	c
d1aw1a_	d1okkd2	1	3.170000	7.97	0.048	-1	c	c
d1h7wa4	d1cqxa3	1	6.00e-04	175.59	0.1684	-1	c	c
d1h7wa4	d1cqxa3	1	3.170000	3.58	0.1212	-1	c	c
d1p80a1	d1lsua_	1	5.97e-03	139.91	0.3301	-1	c	c
d1p80a1	d1lsua_	1	3.210000	5.83	0.2708	-1	c	c
d1ctqa_	d1hdoa_	1	6.71e-03	152.16	0.3328	-1	c	c
d1ctqa_	d1hdoa_	1	3.270000	5.96	0.1913	-1	c	c
d1vlma_	d1dxy_1	1	9.97e-03	144.65	0.2452	-1	c	c
d1vlma_	d1dxy_1	1	3.330000	8.41	0.2428	-1	c	c
d2sqca1	d1g9ga_	1	2.97e-03	186.68	0.2911	-1	a	a
d2sqca1	d1g9ga_	1	3.340000	-1.73	0.1282	-1	a	a
d1a8p_2	d1af7_2	1	9.05e-03	154.73	0.2468	-1	c	c
d1a8p_2	d1af7_2	1	3.340000	5.31	0.1519	-1	c	c
d1m65a_	d1uf3a_	1	4.94e-03	224.54	0.167	-1	c	d
d1m65a_	d1uf3a_	1	3.360000	5.87	0.0758	-1	c	d
d1rcua_	d1gzga_	0	5.15e-03	187.31	0.2324	-1	c	c
d1rcua_	d1gzga_	0	3.360000	2.92	0.2235	-1	c	c
d1e4ea1	d1b8pa1	1	8.39e-03	133.28	0.2346	-1	c	c
d1e4ea1	d1b8pa1	1	3.430000	3.3	0.2942	-1	c	c
d1li4a2	d1v8aa_	1	3.39e-03	174.63	0.1502	-1	c	c
d1li4a2	d1v8aa_	1	3.440000	-3.37	0.1054	-1	c	c
d1p3da3	d1oywa2	1	5.82e-03	171.47	0.1814	-1	c	c
d1p3da3	d1oywa2	1	3.460000	6.87	0.0791	-1	c	c
d1p3da3	d1h65a_	1	4.75e-04	205.51	0.193	-1	c	c
d1p3da3	d1h65a_	1	3.510000	6.85	0.0872	-1	c	c
d1h65a_	d1k6ja_	1	8.28e-03	198.62	0.1683	-1	c	c

d1h65a_	d1k6ja_	1	3.610000	6.78	0.0749	-1	c	c
d1ps9a3	**d1ni5a1**	1	8.06e-03	166.47	0.2417	-1	c	c
d1ps9a3	d1ni5a1	1	3.710000	0	0.1917	-1	c	c
d1u1ja1	**d1nar__**	1	1.34e-03	263.45	0.2018	-1	c	c
d1u1ja1	d1nar__	1	3.780000	6.58	0.1885	-1	c	c
d1ufka_	**d1cqxa3**	1	3.88e-03	177.08	0.1654	-1	c	c
d1ufka_	d1cqxa3	1	3.790000	7.01	0.1535	-1	c	c
d1e8ca3	**d1n0wa_**	1	8.04e-03	197.77	0.1688	-1	c	c
d1e8ca3	d1n0wa_	1	3.800000	2.94	0.094	-1	c	c
d1bif_1	**d1eq2a_**	0	7.96e-04	202.17	0.1819	-1	c	c
d1bif_1	d1eq2a_	0	3.800000	5.55	0.0927	-1	c	c
d1a3wa3	**d1p1ca_**	1	7.41e-03	147.58	0.3619	-1	c	c
d1a3wa3	d1p1ca_	1	3.840000	0.77	0.3265	-1	c	c
d1u7na_	**d1v4va_**	1	8.53e-03	191.94	0.1603	-1	c	c
d1u7na_	d1v4va_	1	3.950000	2.95	0.06	-1	c	c
d1cjca2	**d1pswa_**	1	3.76e-03	204.27	0.1613	-1	c	c
d1cjca2	d1pswa_	1	3.980000	3.22	0.1115	-1	c	c
d1egza_	**d1d8wa_**	1	5.82e-03	223.78	0.25	-1	c	c
d1egza_	d1d8wa_	1	3.980000	4.64	0.116	-1	c	c
d1e8ca3	**d1nrjb_**	1	5.42e-03	166.52	0.1795	-1	c	c
d1e8ca3	d1nrjb_	1	4.020000	6.74	0.0684	-1	c	c
d1p3da3	**d1f60a3**	1	2.08e-04	190.72	0.236	-1	c	c
d1p3da3	d1f60a3	1	4.020000	8.1	0.0814	-1	c	c
d1e8ca3	**d1svsa1**	1	7.73e-03	160.51	0.1635	-1	c	c
d1e8ca3	d1svsa1	1	4.080000	8.71	0.0566	-1	c	c
d1bgxt2	**d1jeya2**	0	4.64e-03	162.49	0.185	-1	c	c
d1bgxt2	d1jeya2	0	4.090000	-1.24	0.1156	-1	c	c
d1khta_	**d1dlja2**	1	8.19e-03	147.98	0.1671	-1	c	c
d1khta_	d1dlja2	1	4.140000	-2.79	0.0895	-1	c	c
d1i52a_	**d1lu9a1**	1	5.37e-03	157.02	0.18	-1	c	c
d1i52a_	d1lu9a1	1	4.160000	-0.99	0.1489	-1	c	c
d1nbaa_	**d1jeya2**	1	4.48e-03	165.41	0.2026	-1	c	c
d1nbaa_	d1jeya2	1	4.180000	5.42	0.0949	-1	c	c
d1oe0a_	**d1auoa_**	0	2.05e-03	188.7	0.1683	-1	c	c
d1oe0a_	d1auoa_	0	4.270000	5.35	0.0992	-1	c	c
d1u0la2	**d1efpa1**	1	5.50e-03	162.36	0.2122	-1	c	c
d1u0la2	d1efpa1	1	4.310000	-0.34	0.19	-1	c	c
d1f60a3	**d1e8ca3**	1	4.19e-05	219.82	0.2364	-1	c	c

d1f60a3	d1e8ca3	1	4.340000	8.25	0.0628	-1	c	c
d1gzga_	d1dqua_	1	1.94e-03	251.52	0.1679	-1	c	c
d1gzga_	d1dqua_	1	4.430000	4.71	0.0889	-1	c	c
d1lw7a2	d1c0pa1	0	4.65e-03	202.53	0.1536	-1	c	c
d1lw7a2	d1c0pa1	0	4.440000	-0.32	0.099	-1	c	c
d1jeya2	d1nbaa_	1	1.88e-03	165.41	0.233	-1	c	c
d1jeya2	d1nbaa_	1	4.480000	5.42	0.1091	-1	c	c
d1nria_	d1g2qa_	1	4.73e-03	178.82	0.1885	-1	c	c
d1nria_	d1g2qa_	1	4.480000	5.87	0.1744	-1	c	c
d1aw1a_	d1dbta_	1	2.47e-04	222.28	0.2814	-1	c	c
d1aw1a_	d1dbta_	1	4.550000	5.27	0.1402	-1	c	c
d1p9oa_	d1g3qa_	1	9.20e-03	170.5	0.1914	-1	c	c
d1p9oa_	d1g3qa_	1	4.600000	6.23	0.0716	-1	c	c
d1oywa2	d1p3da3	1	5.59e-03	171.47	0.1893	-1	c	c
d1oywa2	d1p3da3	1	4.650000	6.87	0.0825	-1	c	c
d1a1va1	d1m66a2	1	4.26e-03	157.79	0.3364	-1	c	c
d1a1va1	d1m66a2	1	4.710000	5.42	0.1875	-1	c	c
d1f60a3	d1jbwa2	1	9.25e-03	218.34	0.1946	-1	c	c
d1f60a3	d1jbwa2	1	4.720000	10.53	0.0638	-1	c	c
d1bg6_2	d1bif_1	1	5.08e-03	183.69	0.2595	-1	c	c
d1bg6_2	d1bif_1	1	4.740000	3.38	0.1386	-1	c	c
d1nija1	d1k6ja_	1	9.48e-03	194.75	0.1757	-1	c	c
d1nija1	d1k6ja_	1	4.750000	6.72	0.0923	-1	c	c
d1bqca_	d1k6ja_	0	1.89e-03	235.33	0.1714	-1	c	c
d1bqca_	d1k6ja_	0	4.870000	-0.15	0.0654	-1	c	c
d1f14a2	d1qhxa_	0	7.15e-03	164.77	0.1797	-1	c	c
d1f14a2	d1qhxa_	0	4.880000	-2.14	0.1315	-1	c	c
d1m66a2	d1a1va1	1	4.74e-03	157.79	0.2421	-1	c	c
d1m66a2	d1a1va1	1	4.910000	5.42	0.1349	-1	c	c
d1wdua_	d1es9a_	0	2.27e-03	178.47	0.1623	-1	d	c
d1wdua_	d1es9a_	0	4.960000	2.34	0.1349	-1	d	c
d1e8ca3	d1h65a_	1	7.03e-03	181.91	0.1806	-1	c	c
d1e8ca3	d1h65a_	1	5.010000	8.4	0.0609	-1	c	c
d1iu8a_	d1ooea_	1	2.34e-03	167.81	0.2609	-1	c	c
d1iu8a_	d1ooea_	1	5.040000	9.69	0.1735	-1	c	c
d1n0wa_	d1e8ca3	1	2.12e-03	197.77	0.1632	-1	c	c
d1n0wa_	d1e8ca3	1	5.060000	2.94	0.0909	-1	c	c
d1ly1a_	d1ebfa1	0	6.14e-03	150.97	0.2401	-1	c	c

d1ly1a_	d1ebfa1	0	5.120000	6.7	0.097	-1	c	c
d1mnaa_	d1v7za_	1	8.87e-03	196.75	0.1886	-1	c	c
d1mnaa_	d1v7za_	1	5.130000	0.5	0.1263	-1	c	c
d1f60a3	d1p3da3	1	7.82e-04	190.72	0.2123	-1	c	c
d1f60a3	d1p3da3	1	5.130000	8.1	0.0732	-1	c	c
d1k92a1	d1byi__	1	1.58e-04	207.42	0.23	-1	c	c
d1k92a1	d1byi__	1	5.130000	5.28	0.1237	-1	c	c
d1qhxa_	d1eq2a_	1	4.10e-03	174.82	0.1826	-1	c	c
d1qhxa_	d1eq2a_	1	5.160000	8.19	0.1025	-1	c	c
d1k87a2	d1dosa_	1	2.45e-04	281.91	0.2251	-1	c	c
d1k87a2	d1dosa_	1	5.180000	4.29	0.0769	-1	c	c
d1e8ca3	d1f60a3	1	8.23e-06	219.82	0.2415	-1	c	c
d1e8ca3	d1f60a3	1	5.190000	8.25	0.0641	-1	c	c
d1bif_1	d1rq2a1	1	3.97e-03	177.52	0.2136	-1	c	c
d1bif_1	d1rq2a1	1	5.250000	5.84	0.1937	-1	c	c
d1byi__	d1k92a1	1	3.48e-04	207.42	0.1931	-1	c	c
d1byi__	d1k92a1	1	5.280000	5.28	0.1038	-1	c	c
d1nar__	d1dysa_	1	7.71e-03	231.73	0.1791	-1	c	c
d1nar__	d1dysa_	1	5.330000	3.91	0.0952	-1	c	c
d1n0wa_	d1db3a_	1	9.69e-03	189.61	0.1746	-1	c	c
d1n0wa_	d1db3a_	1	5.350000	2.13	0.0837	-1	c	c
d1h65a_	d1e8ca3	1	6.99e-03	181.91	0.1644	-1	c	c
d1h65a_	d1e8ca3	1	5.380000	8.4	0.0554	-1	c	c
d2at2a2	d1nvmb1	1	1.53e-03	144.67	0.3477	-1	c	c
d2at2a2	d1nvmb1	1	5.390000	8.04	0.2583	-1	c	c
d1x7fa2	d1dbta_	1	1.60e-03	196.97	0.2828	-1	c	c
d1x7fa2	d1dbta_	1	5.450000	2.39	0.168	-1	c	c
d1ni5a1	d1g3qa_	1	4.95e-03	165.64	0.1608	-1	c	c
d1ni5a1	d1g3qa_	1	5.480000	5.45	0.1046	-1	c	c
d1dlja2	d1dcta_	1	4.79e-04	196.65	0.3151	-1	c	c
d1dlja2	d1dcta_	1	5.510000	5.77	0.301	-1	c	c
d1dpga1	d1okkd2	1	4.20e-03	164.41	0.2704	-1	c	c
d1dpga1	d1okkd2	1	5.530000	9.44	0.25	-1	c	c
d1e4ea1	d1b7go1	1	1.69e-03	147.8	0.3692	-1	c	c
d1e4ea1	d1b7go1	1	5.560000	6.13	0.1731	-1	c	c
d1p3da3	d1nrjb_	1	1.20e-03	177.29	0.2221	-1	c	c
d1p3da3	d1nrjb_	1	5.560000	9.02	0.0837	-1	c	c
d1ixka_	d1g5qa_	1	1.23e-03	190.55	0.1629	-1	c	c

d1ixka_	d1g5qa_	1	5.570000	5.78	0.1669	-1	c	c
d1bif_1	d1p9oa_	0	1.02e-03	213.56	0.1749	-1	c	c
d1bif_1	d1p9oa_	0	5.590000	0.51	0.1021	-1	c	c
d1egaa1	d1b8pa1	1	2.29e-03	150.83	0.1746	-1	c	c
d1egaa1	d1b8pa1	1	5.610000	9.3	0.1969	-1	c	c
d1dqza_	d1nf9a_	1	3.56e-03	177.61	0.1732	-1	c	c
d1dqza_	d1nf9a_	1	5.690000	3.24	0.0554	-1	c	c
d1nrjb_	d1k6ja_	1	7.97e-03	190.74	0.2105	-1	c	c
d1nrjb_	d1k6ja_	1	5.720000	2.72	0.1292	-1	c	c
d1p3da3	d1hdoa_	1	7.40e-03	159.86	0.214	-1	c	c
d1p3da3	d1hdoa_	1	5.740000	7.7	0.0977	-1	c	c
d1a8p_2	d1c0pa1	0	6.86e-03	196.58	0.1709	-1	c	c
d1a8p_2	d1c0pa1	0	5.800000	2.92	0.1218	-1	c	c
d1l6ra_	d1nf9a_	1	4.49e-03	156.19	0.2011	-1	c	c
d1l6ra_	d1nf9a_	1	5.850000	5.82	0.1267	-1	c	c
d7odca2	d1m5wa_	1	3.20e-03	177.34	0.2854	-1	c	c
d7odca2	d1m5wa_	1	5.910000	4.82	0.2135	-1	c	c
d1jbwa2	d1f60a3	1	2.39e-04	218.34	0.1571	-1	c	c
d1jbwa2	d1f60a3	1	5.930000	10.53	0.0515	-1	c	c
d1h65a_	d1p3da3	1	4.68e-04	205.51	0.1615	-1	c	c
d1h65a_	d1p3da3	1	5.930000	6.85	0.073	-1	c	c
d1bif_1	d1bg6_2	1	5.81e-03	183.69	0.2242	-1	c	c
d1bif_1	d1bg6_2	1	6.050000	3.38	0.1197	-1	c	c
d1nvmb1	d2at2a2	1	6.77e-03	144.67	0.3323	-1	c	c
d1nvmb1	d2at2a2	1	6.050000	8.04	0.2468	-1	c	c
d1p9oa_	d1vcoa2	1	9.97e-03	181.65	0.1569	-1	c	c
d1p9oa_	d1vcoa2	1	6.060000	7.58	0.0793	-1	c	c
d1es9a_	d1h65a_	1	2.37e-03	194.54	0.2182	-1	c	c
d1es9a_	d1h65a_	1	6.130000	-0.29	0.0684	-1	c	c
d1n7ka_	d1okkd2	1	8.60e-03	161.81	0.2511	-1	c	c
d1n7ka_	d1okkd2	1	6.140000	2.03	0.047	-1	c	c
d1ddga2	d1dusa_	1	6.55e-03	163.59	0.2908	-1	c	c
d1ddga2	d1dusa_	1	6.140000	9.26	0.0948	-1	c	c
d1dqza_	d1yaca_	1	5.45e-04	193.94	0.1732	-1	c	c
d1dqza_	d1yaca_	1	6.170000	3.84	0.1348	-1	c	c
d1k7ca_	d1obba1	1	9.88e-03	159.59	0.1942	-1	c	c
d1k7ca_	d1obba1	1	6.230000	5.94	0.0483	-1	c	c
d1vjra_	d1e19a_	1	3.47e-03	186.05	0.1513	-1	c	c

d1vjra_	d1e19a_	1	6.330000	6.06	0.1063	-1	c	c
d1uf3a_	d1qo2a_	1	4.85e-03	179.43	0.261	-1	d	c
d1uf3a_	d1qo2a_	1	6.380000	2.28	0.1502	-1	d	c
d1khta_	d1b8pa1	1	4.34e-03	152.67	0.1803	-1	c	c
d1khta_	d1b8pa1	1	6.380000	1.26	0.0776	-1	c	c
d1aw1a_	d1dosa_	1	1.59e-03	251.64	0.3118	-1	c	c
d1aw1a_	d1dosa_	1	6.460000	-1.05	0.2637	-1	c	c
d1dusa_	d1ddga2	1	9.17e-03	163.59	0.2294	-1	c	c
d1dusa_	d1ddga2	1	6.490000	9.26	0.0747	-1	c	c
d1bif_1	d1hdoa_	1	3.41e-03	174.19	0.1995	-1	c	c
d1bif_1	d1hdoa_	1	6.520000	6.64	0.1056	-1	c	c
d1dosa_	d1k87a2	1	1.54e-04	281.91	0.2207	-1	c	c
d1dosa_	d1k87a2	1	6.520000	4.29	0.0754	-1	c	c
d1pswa_	d1rcua_	1	4.96e-03	178.58	0.1545	-1	c	c
d1pswa_	d1rcua_	1	6.530000	4.2	0.0848	-1	c	c
d1e6ca_	d1dih_1	1	7.24e-03	149	0.2353	-1	c	c
d1e6ca_	d1dih_1	1	6.580000	6.92	0.2177	-1	c	c
d1gkpa2	d1sr9a2	1	1.43e-03	256.15	0.1993	-1	c	c
d1gkpa2	d1sr9a2	1	6.600000	4.68	0.0612	-1	c	c
d1uf3a_	d1dar_2	0	7.02e-03	193.14	0.1853	-1	d	c
d1uf3a_	d1dar_2	0	6.630000	6.63	0.1393	-1	d	c
d1cbua_	d1b8pa1	1	4.43e-03	148.15	0.1986	-1	c	c
d1cbua_	d1b8pa1	1	6.670000	8	0.0903	-1	c	c
d1nija1	d1ff9a1	0	2.96e-03	161.58	0.1577	-1	c	c
d1nija1	d1ff9a1	0	6.680000	2.53	0.1025	-1	c	c
d1np6a_	d1bg6_2	0	3.95e-03	176.69	0.1853	-1	c	c
d1np6a_	d1bg6_2	0	6.770000	2.5	0.125	-1	c	c
d1l6ra_	d1nbaa_	1	5.98e-03	155.23	0.2156	-1	c	c
d1l6ra_	d1nbaa_	1	6.770000	5.52	0.1167	-1	c	c
d1h65a_	d1es9a_	1	1.17e-03	194.54	0.18	-1	c	c
d1h65a_	d1es9a_	1	6.790000	-0.29	0.0564	-1	c	c
d1es9a_	d1svsa1	1	4.05e-03	168.09	0.2724	-1	c	c
d1es9a_	d1svsa1	1	6.820000	0.3	0.2064	-1	c	c
d1ly1a_	d1eq2a_	0	8.88e-03	167.26	0.2072	-1	c	c
d1ly1a_	d1eq2a_	0	6.820000	8.58	0.1447	-1	c	c
d1cjca2	d1r0ka2	1	2.37e-03	175.15	0.1807	-1	c	c
d1cjca2	d1r0ka2	1	6.930000	4.99	0.1147	-1	c	c
d1dcta_	d1dlja2	1	4.81e-05	196.65	0.1906	-1	c	c

d1dcta_	d1dlja2	1	6.970000	5.77	0.1821	-1	c	c
d1nar__	d1gzga_	1	3.84e-03	208.89	0.1869	-1	c	c
d1nar__	d1gzga_	1	7.050000	9.21	0.167	-1	c	c
d1nrjb_	d1p3da3	1	2.55e-03	177.29	0.2285	-1	c	c
d1nrjb_	d1p3da3	1	7.060000	9.02	0.0861	-1	c	c
d1p3da3	d1dar_2	1	9.01e-04	200.17	0.2442	-1	c	c
d1p3da3	d1dar_2	1	7.080000	9.81	0.0826	-1	c	c
d1cjca2	d1uz5a3	1	8.57e-03	156.28	0.171	-1	c	c
d1cjca2	d1uz5a3	1	7.090000	-0.33	0.0974	-1	c	c
d1np6a_	d2pgd_2	0	7.13e-03	152.02	0.1838	-1	c	c
d1np6a_	d2pgd_2	0	7.120000	7.7	0.1279	-1	c	c
d1o14a_	d1ewka_	1	9.70e-03	235.45	0.1536	-1	c	c
d1o14a_	d1ewka_	1	7.130000	-1.32	0.0815	-1	c	c
d1p1ma2	d1m3ua_	1	9.48e-03	201.1	0.2055	-1	c	c
d1p1ma2	d1m3ua_	1	7.130000	-0.66	0.0632	-1	c	c
d1sgja_	d1jr1a1	1	4.68e-03	205.47	0.3831	-1	c	c
d1sgja_	d1jr1a1	1	7.160000	4.34	0.1851	-1	c	c
d1d1qa_	d1obba1	1	7.37e-04	161.4	0.2563	-1	c	c
d1d1qa_	d1obba1	1	7.210000	8.57	0.2201	-1	c	c
d1svsa1	d1es9a_	1	5.29e-03	168.09	0.2946	-1	c	c
d1svsa1	d1es9a_	1	7.260000	0.3	0.2232	-1	c	c
d1kbla1	d1dosa_	1	5.21e-03	237.33	0.1696	-1	c	c
d1kbla1	d1dosa_	1	7.390000	2.87	0.193	-1	c	c
d1p9oa_	d1ihua1	1	2.00e-05	271.12	0.1793	-1	c	c
d1p9oa_	d1ihua1	1	7.470000	7.91	0.0707	-1	c	c
d1f60a3	d1b8pa1	1	6.34e-03	152.04	0.1736	-1	c	c
d1f60a3	d1b8pa1	1	7.500000	4.43	0.1203	-1	c	c
d1q7za1	d1sgja_	1	3.38e-03	181.36	0.306	-1	c	c
d1q7za1	d1sgja_	1	7.540000	1.75	0.1921	-1	c	c
d1ly1a_	d1e8ca3	1	8.68e-03	165.3	0.2138	-1	c	c
d1ly1a_	d1e8ca3	1	7.560000	9.19	0.1118	-1	c	c
d1o14a_	d1onwa2	0	9.56e-03	216.15	0.1505	-1	c	c
d1o14a_	d1onwa2	0	7.580000	5.15	0.0713	-1	c	c
d1eq2a_	d1qj4a_	1	1.62e-03	197.5	0.1596	-1	c	c
d1eq2a_	d1qj4a_	1	7.590000	6.02	0.0806	-1	c	c
d1f6ba_	d1hdoa_	1	4.35e-04	178.36	0.3132	-1	c	c
d1f6ba_	d1hdoa_	1	7.610000	8.17	0.1707	-1	c	c
d1dcta_	d1saya1	1	6.47e-03	161.41	0.1535	-1	c	c

d1dcta_	d1saya1	1	7.650000	6.68	0.1451	-1	c	c
d1nsj__	d1l6wa_	1	3.17e-03	170.89	0.4232	-1	c	c
d1nsj__	d1l6wa_	1	7.680000	9.95	0.3829	-1	c	c
d1ly1a_	d1ks9a2	0	8.20e-03	149.84	0.2188	-1	c	c
d1ly1a_	d1ks9a2	0	7.680000	6.02	0.1316	-1	c	c
d1qj4a_	d1eq2a_	1	1.39e-03	197.5	0.1914	-1	c	c
d1qj4a_	d1eq2a_	1	7.710000	6.02	0.0967	-1	c	c
d1oe4a_	d1es9a_	1	6.82e-03	164.99	0.1531	-1	c	c
d1oe4a_	d1es9a_	1	7.750000	9.64	0.0765	-1	c	c
d1bfd_2	d1rcua_	1	9.88e-03	138.96	0.25	-1	c	c
d1bfd_2	d1rcua_	1	7.780000	8.15	0.2583	-1	c	c
d1q7za1	d1ej0a_	0	3.48e-03	167.69	0.1593	-1	c	c
d1q7za1	d1ej0a_	0	7.810000	9.78	0.1332	-1	c	c
d1e8ca3	d1d2na_	1	7.24e-03	196.83	0.1709	-1	c	c
d1e8ca3	d1d2na_	1	7.880000	9.99	0.0865	-1	c	c
d1a3c__	d1jzta_	1	8.15e-03	176.26	0.2669	-1	c	c
d1a3c__	d1jzta_	1	7.950000	4.52	0.0927	-1	c	c
d1q74a_	d1k6ja_	1	3.45e-03	217.65	0.1532	-1	c	c
d1q74a_	d1k6ja_	1	7.960000	0.45	0.0951	-1	c	c
d1hdoa_	d1ctqa_	1	4.77e-03	152.16	0.2695	-1	c	c
d1hdoa_	d1ctqa_	1	7.980000	5.96	0.1549	-1	c	c
d1ebfa1	d1bif_1	0	9.71e-03	170.32	0.2293	-1	c	c
d1ebfa1	d1bif_1	0	8.070000	6.58	0.0947	-1	c	c
d1egaa1	d1obba1	1	6.83e-03	150.95	0.162	-1	c	c
d1egaa1	d1obba1	1	8.180000	9.6	0.0852	-1	c	c
d1khta_	d1ff9a1	1	9.99e-03	153.36	0.1579	-1	c	c
d1khta_	d1ff9a1	1	8.350000	6.53	0.1079	-1	c	c
d1dpga1	d1ddga2	1	3.63e-03	172.46	0.2704	-1	c	c
d1dpga1	d1ddga2	1	8.370000	3.76	0.2551	-1	c	c
d1npya1	d1k66a_	1	3.25e-03	142.79	0.2904	-1	c	c
d1npya1	d1k66a_	1	8.460000	5.63	0.2739	-1	c	c
d1a8p_2	d1kyqa1	1	4.57e-04	169.16	0.2896	-1	c	c
d1a8p_2	d1kyqa1	1	8.500000	11.69	0.1424	-1	c	c
d1onwa2	d1vk4a_	0	2.78e-03	207.17	0.1673	-1	c	c
d1onwa2	d1vk4a_	0	8.610000	7.52	0.0863	-1	c	c
d1ddga2	d1khta_	1	1.40e-03	181.91	0.201	-1	c	c
d1ddga2	d1khta_	1	8.700000	-0.13	0.1324	-1	c	c
d1a49a2	d1gzga_	1	2.08e-04	233.65	0.3198	-1	c	c

d1a49a2	d1gzga_	1	8.700000	8.62	0.2527	-1	c	c
d1dosa_	**d1aw1a_**	1	7.15e-04	251.64	0.2221	-1	c	c
d1dosa_	d1aw1a_	1	8.720000	-1.05	0.1878	-1	c	c
d1gzga_	**d1a49a2**	1	2.73e-04	233.65	0.2751	-1	c	c
d1gzga_	d1a49a2	1	8.760000	8.62	0.2173	-1	c	c
d1hdoa_	**d1bif_1**	1	9.24e-03	174.19	0.2073	-1	c	c
d1hdoa_	d1bif_1	1	8.760000	6.64	0.1098	-1	c	c
d1sr9a2	**d1gkpa2**	1	7.05e-05	256.15	0.2153	-1	c	c
d1sr9a2	d1gkpa2	1	8.790000	4.68	0.0661	-1	c	c
d1woha_	**d1gca__**	1	9.94e-03	212.5	0.1592	-1	c	c
d1woha_	d1gca__	1	8.820000	6.85	0.0908	-1	c	c
d1f9aa_	**d1jmva_**	1	9.52e-03	143.26	0.3034	-1	c	c
d1f9aa_	d1jmva_	1	8.830000	5.43	0.1479	-1	c	c
d1jsxa_	**d1h7wa4**	1	4.95e-03	188.37	0.1763	-1	c	c
d1jsxa_	d1h7wa4	1	8.860000	10.45	0.1292	-1	c	c
d1ddga2	**d1dpga1**	1	3.83e-03	172.46	0.3464	-1	c	c
d1ddga2	d1dpga1	1	9.000000	3.76	0.3268	-1	c	c
d1rq2a1	**d1ecfa1**	1	9.25e-03	176.98	0.2399	-1	c	c
d1rq2a1	d1ecfa1	1	9.040000	2.87	0.2159	-1	c	c
d1qtwa_	**d1sgja_**	1	5.23e-03	186.59	0.2316	-1	c	c
d1qtwa_	d1sgja_	1	9.210000	3.53	0.057	-1	c	c
d1g3qa_	**d1ks9a2**	1	1.77e-03	167.48	0.1614	-1	c	c
d1g3qa_	d1ks9a2	1	9.300000	9.3	0.0928	-1	c	c
d1ecfa1	**d1rq2a1**	1	2.70e-03	176.98	0.1955	-1	c	c
d1ecfa1	d1rq2a1	1	9.360000	2.87	0.1759	-1	c	c
d1khta_	**d1eq2a_**	1	2.82e-03	182.4	0.1934	-1	c	c
d1khta_	d1eq2a_	1	9.380000	7.62	0.1092	-1	c	c
d1ks9a2	**d1g3qa_**	1	1.65e-03	167.48	0.229	-1	c	c
d1ks9a2	d1g3qa_	1	9.430000	9.3	0.1317	-1	c	c
d1oi7a2	**d1h6da1**	1	9.49e-03	166.01	0.3593	-1	c	c
d1oi7a2	d1h6da1	1	9.450000	9.19	0.3353	-1	c	c
d1obba1	**d1d1qa_**	1	4.86e-03	161.4	0.2383	-1	c	c
d1obba1	d1d1qa_	1	9.550000	8.57	0.2047	-1	c	c
d1egaa1	**d1b7go1**	1	2.03e-03	155.2	0.264	-1	c	c
d1egaa1	d1b7go1	1	9.670000	9.3	0.2514	-1	c	c
d1khta_	**d1ddga2**	1	1.26e-03	181.91	0.1618	-1	c	c
d1khta_	d1ddga2	1	9.750000	-0.13	0.1066	-1	c	c
d1m65a_	**d1sgja_**	1	1.63e-03	194.54	0.2613	-1	c	c

d1m65a_	d1sgja_	1	9.850000	0.55	0.2582	-1	c	c
d1d2na_	d1e8ca3	1	1.69e-03	196.83	0.1626	-1	c	c
d1d2na_	d1e8ca3	1	9.850000	9.99	0.0823	-1	c	c
d1m65a_	d1ii7a_	0	8.14e-03	243.2	0.1865	-1	c	d
d1m65a_	d1ii7a_	0	9.870000	8.92	0.0881	-1	c	d
d1es9a_	d1hdoa_	1	7.86e-03	164.05	0.316	-1	c	c
d1es9a_	d1hdoa_	1	9.880000	8.31	0.2205	-1	c	c
d1dar_2	d1p3da3	1	7.84e-04	200.17	0.1862	-1	c	c
d1dar_2	d1p3da3	1	9.980000	9.81	0.0629	-1	c	c
d1ihua1	d1p9oa_	1	7.95e-06	271.12	0.1757	-1	c	c
d1ihua1	d1p9oa_	1	10.030000	7.91	0.0693	-1	c	c
d1oywa2	d1e8ca3	1	7.04e-03	171.8	0.2124	-1	c	c
d1oywa2	d1e8ca3	1	10.060000	11.47	0.0752	-1	c	c
d1h1na_	d1gzga_	1	1.18e-03	234.99	0.323	-1	c	c
d1h1na_	d1gzga_	1	10.070000	3.65	0.2377	-1	c	c
d1bif_1	d1ebfa1	0	2.98e-03	170.32	0.1819	-1	c	c
d1bif_1	d1ebfa1	0	10.120000	6.58	0.0751	-1	c	c
d1a8p_2	d1fcda1	1	7.42e-03	159.65	0.2421	-1	c	c
d1a8p_2	d1fcda1	1	10.220000	8.56	0.1456	-1	c	c
d1a49a2	d1m5wa_	1	5.86e-03	180.6	0.3092	-1	c	c
d1a49a2	d1m5wa_	1	10.260000	8.34	0.0671	-1	c	c
d1l9ha_	d1q16c_	1	9.61e-03	139.15	0.1566	-1	f	f
d1l9ha_	d1q16c_	1	10.440000	-4.6	0	-1	f	f
d1fcda1	d1eq2a_	1	2.61e-04	200.15	0.1725	-1	c	c
d1fcda1	d1eq2a_	1	10.470000	7.87	0.1484	-1	c	c
d1mwma2	d1hjra_	1	3.74e-03	162.3	0.1841	-1	c	c
d1mwma2	d1hjra_	1	10.750000	6.27	0.1564	-1	c	c
d1nvma2	d1sgja_	1	7.55e-03	188.94	0.2535	-1	c	c
d1nvma2	d1sgja_	1	10.790000	4.19	0.0476	-1	c	c
d1uasa2	d1eyea_	1	8.68e-03	199.87	0.1941	-1	c	c
d1uasa2	d1eyea_	1	10.830000	3.99	0.1731	-1	c	c
d1ctqa_	d1ks9a2	1	1.81e-03	155.62	0.2982	-1	c	c
d1ctqa_	d1ks9a2	1	10.870000	9.79	0.25	-1	c	c
d1jr2a_	d1f8fa2	1	8.64e-03	163.22	0.199	-1	c	c
d1jr2a_	d1f8fa2	1	11.130000	4.32	0.1423	-1	c	c
d1im8a_	d1ddga2	1	7.39e-03	172.87	0.1678	-1	c	c
d1im8a_	d1ddga2	1	11.150000	-0.36	0.0989	-1	c	c
d1gca__	d1woha_	1	8.71e-03	212.5	0.1562	-1	c	c

d1gca__	d1woha_	1	11.240000	6.85	0.089	-1	c	c
d1kyqa1	**d1a8p_2**	1	1.23e-03	169.16	0.305	-1	c	c
d1kyqa1	d1a8p_2	1	11.340000	11.69	0.15	-1	c	c
d1e8ca3	**d1oywa2**	1	7.11e-03	171.8	0.187	-1	c	c
d1e8ca3	d1oywa2	1	11.380000	11.47	0.0662	-1	c	c
d1e6ca_	**d1bg6_2**	0	8.67e-03	164.75	0.2191	-1	c	c
d1e6ca_	d1bg6_2	0	11.390000	7.32	0.1191	-1	c	c
d1k92a1	**d1db3a_**	1	3.41e-03	192.09	0.2354	-1	c	c
d1k92a1	d1db3a_	1	11.410000	6.66	0.1556	-1	c	c
d2at2a2	**d1t4ba1**	1	7.27e-03	140.83	0.2649	-1	c	c
d2at2a2	d1t4ba1	1	11.740000	8.4	0.2334	-1	c	c
d1i52a_	**d1okkd2**	0	8.64e-03	150.29	0.1733	-1	c	c
d1i52a_	d1okkd2	0	11.750000	4.41	0.0678	-1	c	c
d1cjca2	**d1okkd2**	0	2.54e-03	175.77	0.1515	-1	c	c
d1cjca2	d1okkd2	0	11.750000	4.81	0.1461	-1	c	c
d1sur__	**d1m6ya2**	1	6.54e-03	167.24	0.2023	-1	c	c
d1sur__	d1m6ya2	1	11.840000	5.02	0.1546	-1	c	c
d1sr9a2	**d1j79a_**	1	2.61e-03	258.21	0.2282	-1	c	c
d1sr9a2	d1j79a_	1	11.980000	10.94	0.0565	-1	c	c
d1g5qa_	**d1obba1**	1	6.28e-03	147.92	0.2098	-1	c	c
d1g5qa_	d1obba1	1	11.980000	7.55	0.1193	-1	c	c
d1npya1	**d1ej0a_**	1	7.48e-03	147.35	0.3234	-1	c	c
d1npya1	d1ej0a_	1	12.010000	4.51	0.1272	-1	c	c
d1hdoa_	**d1f6ba_**	1	4.43e-04	178.36	0.2841	-1	c	c
d1hdoa_	d1f6ba_	1	12.020000	8.17	0.1549	-1	c	c
d1k92a1	**d1cp2a_**	1	3.86e-03	187.91	0.1955	-1	c	c
d1k92a1	d1cp2a_	1	12.140000	5.66	0.1197	-1	c	c
d1a3c__	**d1j5xa_**	1	5.70e-03	194.34	0.309	-1	c	c
d1a3c__	d1j5xa_	1	12.190000	4.87	0.1208	-1	c	c
d1byi__	**d1p9oa_**	1	1.47e-03	208.57	0.1931	-1	c	c
d1byi__	d1p9oa_	1	12.230000	8.68	0.0904	-1	c	c
d1dosa_	**d1kbla1**	1	1.52e-03	237.33	0.1725	-1	c	c
d1dosa_	d1kbla1	1	12.230000	2.87	0.1962	-1	c	c
d1vhea2	**d1rxya_**	1	6.10e-03	208.38	0.2016	-1	c	c
d1vhea2	d1rxya_	1	12.240000	2.91	0.094	-1	c	c
d1inla_	**d1k6ja_**	1	5.43e-03	208.86	0.2559	-1	c	c
d1inla_	d1k6ja_	1	12.310000	4.91	0.1576	-1	c	c
d1ufka_	**d1a8p_2**	1	8.16e-03	176.93	0.1713	-1	c	c

d1ufka_	d1a8p_2	1	12.330000	10.96	0.1004	-1	c	c
d1tyga_	d1iq8a1	1	6.57e-03	199.83	0.2934	-1	c	c
d1tyga_	d1iq8a1	1	12.400000	11.66	0.375	-1	c	c
d1hdoa_	d1es9a_	1	9.98e-03	164.05	0.3268	-1	c	c
d1hdoa_	d1es9a_	1	12.450000	8.31	0.228	-1	c	c
d1vk4a_	d1onwa2	0	4.86e-03	207.17	0.1655	-1	c	c
d1vk4a_	d1onwa2	0	12.460000	7.52	0.0854	-1	c	c
d1bg6_2	d1e6ca_	0	7.62e-03	164.75	0.2024	-1	c	c
d1bg6_2	d1e6ca_	0	12.520000	7.32	0.1101	-1	c	c
d1np6a_	d1ks9a2	0	7.84e-04	170.91	0.1765	-1	c	c
d1np6a_	d1ks9a2	0	12.740000	5.64	0.1426	-1	c	c
d1cbf__	d1tdj_1	0	6.79e-03	176.7	0.1632	-1	c	c
d1cbf__	d1tdj_1	0	12.810000	2.28	0.092	-1	c	c
d1o1za_	d1g3qa_	0	9.33e-03	155.65	0.1681	-1	c	c
d1o1za_	d1g3qa_	0	13.180000	-0.45	0.1405	-1	c	c
d1f6ba_	d1ks9a2	1	6.82e-03	151.21	0.2554	-1	c	c
d1f6ba_	d1ks9a2	1	13.180000	10.18	0.2204	-1	c	c
d1sgja_	d1tqja_	1	8.32e-03	175.26	0.3409	-1	c	c
d1sgja_	d1tqja_	1	13.210000	6.1	0.2143	-1	c	c
d1k92a1	d1ff9a1	1	3.82e-03	163.36	0.2261	-1	c	c
d1k92a1	d1ff9a1	1	13.270000	8.35	0.1915	-1	c	c
d1g3qa_	d1dlja2	1	5.80e-03	149.63	0.1909	-1	c	c
d1g3qa_	d1dlja2	1	13.290000	10.25	0.0654	-1	c	c
d1jr1a1	d1sgja_	1	4.05e-04	205.47	0.2335	-1	c	c
d1jr1a1	d1sgja_	1	13.320000	4.34	0.1128	-1	c	c
d1egaa1	d1ks9a2	1	5.62e-03	151.83	0.2095	-1	c	c
d1egaa1	d1ks9a2	1	13.330000	14.13	0.2039	-1	c	c
d1sr9a2	d1k6wa2	1	2.10e-03	233.45	0.2564	-1	c	c
d1sr9a2	d1k6wa2	1	13.390000	-3.73	0.0702	-1	c	c
d1h6da1	d1rq2a1	1	7.13e-03	169.55	0.2309	-1	c	c
d1h6da1	d1rq2a1	1	13.410000	7.89	0.2489	-1	c	c
d1knqa_	d1eq2a_	1	7.13e-03	173.11	0.1681	-1	c	c
d1knqa_	d1eq2a_	1	13.490000	10.88	0.1257	-1	c	c
d1geha1	d1o4ua1	1	9.98e-03	174.35	0.193	-1	c	c
d1geha1	d1o4ua1	1	13.510000	9.51	0.1849	-1	c	c
d1fyea_	d1b8pa1	1	8.16e-03	161.2	0.1878	-1	c	c
d1fyea_	d1b8pa1	1	13.520000	5.68	0.1419	-1	c	c
d1fcda1	d1a8p_2	1	5.03e-03	159.65	0.2045	-1	c	c

d1fcda1	d1a8p_2	1	13.540000	8.56	0.123	-1	c	c
d1j79a_	**d1sr9a2**	1	3.05e-03	258.21	0.2063	-1	c	c
d1j79a_	d1sr9a2	1	13.780000	10.94	0.051	-1	c	c
d1cjca2	**d1duvg2**	0	4.36e-03	164.9	0.158	-1	c	c
d1cjca2	d1duvg2	0	14.100000	5.2	0.1591	-1	c	c
d1cjca2	**d1jx7a_**	1	9.67e-03	157.89	0.1937	-1	c	c
d1cjca2	d1jx7a_	1	14.110000	7.37	0.1613	-1	c	c
d1cjca2	**d1b16a_**	1	2.40e-03	179.71	0.2024	-1	c	c
d1cjca2	d1b16a_	1	14.150000	8.26	0.1006	-1	c	c
d1ixka_	**d1b8pa1**	1	6.75e-03	171.21	0.1901	-1	c	c
d1ixka_	d1b8pa1	1	14.170000	2.58	0.1997	-1	c	c
d1k92a1	**d1okkd2**	1	8.14e-04	182.05	0.2726	-1	c	c
d1k92a1	d1okkd2	1	14.240000	8.17	0.1782	-1	c	c
d1gkub1	**d1obba1**	1	8.81e-03	157.1	0.173	-1	c	c
d1gkub1	d1obba1	1	14.270000	5.07	0.159	-1	c	c
d1h7wa4	**d1jsxa_**	1	2.94e-03	188.37	0.1862	-1	c	c
d1h7wa4	d1jsxa_	1	14.420000	10.45	0.1365	-1	c	c
d1cqxa3	**d1npya1**	1	9.74e-03	150.22	0.2975	-1	c	c
d1cqxa3	d1npya1	1	14.480000	10.24	0.1831	-1	c	c
d1obba1	**d1g5qa_**	1	5.15e-03	147.92	0.2135	-1	c	c
d1obba1	d1g5qa_	1	14.600000	7.55	0.1213	-1	c	c
d4kbpa2	**d1gqna_**	0	5.75e-03	213.03	0.1739	-1	d	c
d4kbpa2	d1gqna_	0	14.680000	-0.23	0.1418	-1	d	c
d1h6da1	**d1oi7a2**	1	3.55e-03	166.01	0.2703	-1	c	c
d1h6da1	d1oi7a2	1	14.780000	9.19	0.2523	-1	c	c
d1qf9a_	**d1dpga1**	1	9.59e-03	163.08	0.1972	-1	c	c
d1qf9a_	d1dpga1	1	14.800000	11.95	0.1095	-1	c	c
d1iq8a1	**d1tyga_**	1	1.98e-03	199.83	0.2	-1	c	c
d1iq8a1	d1tyga_	1	14.880000	11.66	0.2556	-1	c	c
d1sr9a2	**d1onwa2**	1	7.64e-03	217.33	0.2073	-1	c	c
d1sr9a2	d1onwa2	1	14.890000	6.11	0.05	-1	c	c
d1k92a1	**d1f8fa2**	1	4.11e-03	161.99	0.3418	-1	c	c
d1k92a1	d1f8fa2	1	14.980000	8.28	0.2287	-1	c	c
d1ks9a2	**d1f6ba_**	1	7.75e-03	151.21	0.2844	-1	c	c
d1ks9a2	d1f6ba_	1	15.000000	10.18	0.2455	-1	c	c
d1bif_1	**d1ff9a1**	0	1.11e-03	179.4	0.2054	-1	c	c
d1bif_1	d1ff9a1	0	15.040000	8.31	0.1033	-1	c	c
d1okkd2	**d1k92a1**	1	5.48e-03	182.05	0.2476	-1	c	c

d1okkd2	d1k92a1	1	15.160000	8.17	0.1618	-1	c	c
d1jzta_	d1dih_1	1	3.31e-03	172.35	0.1955	-1	c	c
d1jzta_	d1dih_1	1	15.190000	6.49	0.0988	-1	c	c
d1dih_1	d1v4va_	1	9.54e-03	176.11	0.365	-1	c	c
d1dih_1	d1v4va_	1	15.210000	8.31	0.1595	-1	c	c
d1pswa_	d1rqla_	1	4.92e-03	198.89	0.166	-1	c	c
d1pswa_	d1rqla_	1	15.290000	4.89	0.1315	-1	c	c
d1obba1	d1xvaa_	1	7.21e-03	192.7	0.3465	-1	c	c
d1obba1	d1xvaa_	1	15.290000	2.32	0.174	-1	c	c
d1ff9a1	d1bif_1	0	4.92e-03	179.4	0.2391	-1	c	c
d1ff9a1	d1bif_1	0	15.340000	8.31	0.1202	-1	c	c
d1gqna_	d1k87a2	1	6.68e-03	225.62	0.3065	-1	c	c
d1gqna_	d1k87a2	1	15.500000	-0.42	0.0724	-1	c	c
d1tjya_	d1woha_	1	6.75e-03	221.99	0.1622	-1	c	c
d1tjya_	d1woha_	1	15.500000	5.47	0.0926	-1	c	c
d1l6ra_	d1e19a_	1	7.41e-03	175.04	0.1656	-1	c	c
d1l6ra_	d1e19a_	1	15.670000	3.1	0.1356	-1	c	c
d1a3c__	d1nria_	1	6.09e-05	227.48	0.257	-1	c	c
d1a3c__	d1nria_	1	15.810000	8.19	0.2486	-1	c	c
d1b74a2	d1h6da1	0	9.31e-03	164.37	0.2789	-1	c	c
d1b74a2	d1h6da1	0	15.820000	6.9	0.2772	-1	c	c
d8abp__	d1cnza_	1	9.39e-03	218.62	0.1631	-1	c	c
d8abp__	d1cnza_	1	15.870000	4.55	0.1197	-1	c	c
d1vjga_	d1h65a_	1	5.90e-03	184.37	0.2562	-1	c	c
d1vjga_	d1h65a_	1	15.870000	5.56	0.0585	-1	c	c
d1qq5a_	d1oi7a2	1	8.94e-03	154.67	0.1735	-1	c	c
d1qq5a_	d1oi7a2	1	15.890000	7.9	0.1622	-1	c	c
d1geha1	d1vlia2	1	7.25e-03	182.65	0.1979	-1	c	c
d1geha1	d1vlia2	1	15.900000	8.23	0.0985	-1	c	c
d1uoua2	d1k6ja_	1	5.09e-03	196.51	0.1612	-1	c	c
d1uoua2	d1k6ja_	1	16.060000	6.04	0.0843	-1	c	c
d1b16a_	d1cp2a_	1	9.96e-03	180.51	0.1545	-1	c	c
d1b16a_	d1cp2a_	1	16.080000	9.8	0.0748	-1	c	c
d1vjga_	d1svsa1	1	2.79e-03	169.62	0.2749	-1	c	c
d1vjga_	d1svsa1	1	16.120000	9.09	0.2425	-1	c	c
d1rlia_	d1g5qa_	1	5.94e-03	148.69	0.2975	-1	c	c
d1rlia_	d1g5qa_	1	16.130000	9.1	0.2346	-1	c	c
d1rqla_	d1pswa_	1	5.16e-03	198.89	0.2247	-1	c	c

d1rqla_	d1pswa_	1	16.140000	4.89	0.178	-1	c	c
d1a1va1	d1bg6_2	1	6.33e-03	169.14	0.3162	-1	c	c
d1a1va1	d1bg6_2	1	16.220000	5.21	0.2151	-1	c	c
d1oboa_	d1q7ra_	1	9.25e-03	164.25	0.2796	-1	c	c
d1oboa_	d1q7ra_	1	16.390000	12.02	0.1746	-1	c	c
d1xvaa_	d1obba1	1	4.35e-04	192.7	0.2029	-1	c	c
d1xvaa_	d1obba1	1	16.630000	2.32	0.1019	-1	c	c
d1gkpa2	d1m5wa_	1	7.54e-03	175.65	0.1522	-1	c	c
d1gkpa2	d1m5wa_	1	16.680000	5.22	0.1545	-1	c	c
d1vjga_	d1f6ba_	1	8.10e-03	159.87	0.2525	-1	c	c
d1vjga_	d1f6ba_	1	16.720000	13.22	0.0535	-1	c	c
d1ks9a2	d1ctqa_	1	1.27e-03	155.62	0.2964	-1	c	c
d1ks9a2	d1ctqa_	1	16.800000	9.79	0.2485	-1	c	c
d1tqja_	d1sgja_	1	3.38e-03	175.26	0.3563	-1	c	c
d1tqja_	d1sgja_	1	16.830000	6.1	0.224	-1	c	c
d1qyra_	d1k6ja_	1	4.17e-03	196.83	0.2341	-1	c	c
d1qyra_	d1k6ja_	1	16.850000	11.58	0.2133	-1	c	c
d1npya1	d1cqxa3	1	5.12e-03	150.22	0.253	-1	c	c
d1npya1	d1cqxa3	1	16.970000	10.24	0.1557	-1	c	c
d1ecfa1	d1jzta_	1	4.28e-03	190.14	0.249	-1	c	c
d1ecfa1	d1jzta_	1	16.990000	3.64	0.2233	-1	c	c
d1ns5a_	d1k66a_	1	6.22e-03	143.97	0.2386	-1	c	c
d1ns5a_	d1k66a_	1	17.020000	7.71	0.1977	-1	c	c
d1pda_1	d1qgoa_	1	9.30e-04	205.63	0.174	-1	c	c
d1pda_1	d1qgoa_	1	17.070000	10.61	0.2028	-1	c	c
d1b74a2	d1dih_1	1	9.62e-03	147.04	0.284	-1	c	c
d1b74a2	d1dih_1	1	17.120000	8.61	0.2976	-1	c	c
d1l7da2	d1duvg2	1	4.41e-03	147.83	0.2333	-1	c	c
d1l7da2	d1duvg2	1	17.230000	9.31	0.2167	-1	c	c
d1ep3b2	d1b7go1	1	3.29e-03	153.54	0.3109	-1	c	c
d1ep3b2	d1b7go1	1	17.270000	8.65	0.2766	-1	c	c
d1fyea_	d1v8aa_	1	3.68e-03	185.55	0.2653	-1	c	c
d1fyea_	d1v8aa_	1	17.380000	7.17	0.238	-1	c	c
d1qx4a2	d1kyqa1	1	4.99e-03	148.91	0.2908	-1	c	c
d1qx4a2	d1kyqa1	1	17.400000	6.39	0.1327	-1	c	c
d1b7go1	d1egaa1	1	8.83e-03	155.2	0.264	-1	c	c
d1b7go1	d1egaa1	1	17.490000	9.3	0.2514	-1	c	c
d1h7wa4	d1ihua1	0	8.55e-03	200.35	0.1543	-1	c	c

d1h7wa4	d1ihua1	0	17.540000	10.38	0.1212	-1	c	c
d1q74a_	d1v4va_	1	8.35e-03	196.51	0.2231	-1	c	c
d1q74a_	d1v4va_	1	17.650000	11.34	0.0724	-1	c	c
d1t5ba_	d1rq2a1	1	5.13e-03	172.73	0.2674	-1	c	c
d1t5ba_	d1rq2a1	1	17.660000	7.13	0.2674	-1	c	c
d1sqsa_	d1ks9a2	1	5.74e-03	171.78	0.209	-1	c	c
d1sqsa_	d1ks9a2	1	17.680000	6.64	0.2188	-1	c	c
d1tf7a1	d1af7_2	1	3.45e-03	186.42	0.1539	-1	c	c
d1tf7a1	d1af7_2	1	17.700000	2.72	0.093	-1	c	c
d1edg__	d1d8wa_	1	3.68e-03	229.6	0.1783	-1	c	c
d1edg__	d1d8wa_	1	17.720000	2.46	0.1263	-1	c	c
d1v4va_	d1dih_1	1	7.83e-03	176.11	0.1595	-1	c	c
d1v4va_	d1dih_1	1	17.760000	8.31	0.0697	-1	c	c
d1cfza_	d1cbua_	1	5.96e-03	161.7	0.2392	-1	c	c
d1cfza_	d1cbua_	1	17.800000	10.18	0.2022	-1	c	c
d1p80a1	d1ks9a2	1	8.29e-03	146.08	0.3173	-1	c	c
d1p80a1	d1ks9a2	1	17.930000	9.09	0.2804	-1	c	c
d1ps9a3	d1eq2a_	1	6.33e-03	173.03	0.2111	-1	c	c
d1ps9a3	d1eq2a_	1	17.950000	10.44	0.1472	-1	c	c
d1k87a2	d1gqna_	1	2.41e-03	225.62	0.2201	-1	c	c
d1k87a2	d1gqna_	1	18.020000	-0.42	0.052	-1	c	c
d1qhxa_	d1hdoa_	1	2.45e-03	165.37	0.2612	-1	c	c
d1qhxa_	d1hdoa_	1	18.090000	13.06	0.0955	-1	c	c
d1ii7a_	d1emsa2	1	9.90e-03	216.25	0.1779	-1	d	d
d1ii7a_	d1emsa2	1	18.170000	9.21	0.0698	-1	d	d
d1k92a1	d1ps9a3	1	9.91e-03	172.29	0.1915	-1	c	c
d1k92a1	d1ps9a3	1	18.170000	6.82	0.1569	-1	c	c
d1dl5a1	d1a3wa3	1	3.46e-03	164.61	0.23	-1	c	c
d1dl5a1	d1a3wa3	1	18.380000	6.88	0.1843	-1	c	c
d1eyea_	d1b5ta_	1	2.50e-03	222.45	0.2444	-1	c	c
d1eyea_	d1b5ta_	1	18.530000	2.48	0.2852	-1	c	c
d1q74a_	d1k92a1	1	1.61e-03	212.55	0.1961	-1	c	c
d1q74a_	d1k92a1	1	18.570000	9.91	0.1094	-1	c	c
d1aw1a_	d1adoa_	1	2.81e-03	198.41	0.4088	-1	c	c
d1aw1a_	d1adoa_	1	18.920000	6.21	0.2235	-1	c	c
d1bg6_2	d1khta_	0	2.51e-04	200.79	0.2038	-1	c	c
d1bg6_2	d1khta_	0	18.980000	7.72	0.1875	-1	c	c
d1j9ja_	d1v4va_	1	4.93e-04	211.36	0.2308	-1	c	c

d1j9ja_	d1v4va_	1	19.040000	9.93	0.2085	-1	c	c
d1a9xa2	**d1bmta2**	1	8.43e-03	143.85	0.2717	-1	c	c
d1a9xa2	d1bmta2	1	19.200000	14.06	0.2065	-1	c	c
d1bif_1	**d1k6ja_**	1	2.54e-03	209.88	0.1749	-1	c	c
d1bif_1	d1k6ja_	1	19.250000	3.39	0.1068	-1	c	c
d1b5ta_	**d1eyea_**	1	2.64e-03	222.45	0.24	-1	c	c
d1b5ta_	d1eyea_	1	19.300000	2.48	0.28	-1	c	c
d1l6wa_	**d1twda_**	1	9.99e-04	209.11	0.3557	-1	c	c
d1l6wa_	d1twda_	1	19.350000	12.47	0.2898	-1	c	c
d1lw7a2	**d1hdoa_**	1	2.98e-03	170.86	0.2643	-1	c	c
d1lw7a2	d1hdoa_	1	19.360000	13.22	0.0964	-1	c	c
d1qyra_	**d1npya1**	1	5.46e-03	164.74	0.2272	-1	c	c
d1qyra_	d1npya1	1	19.450000	7.7	0.2004	-1	c	c
d1bg6_2	**d1a1va1**	1	3.77e-03	169.14	0.2337	-1	c	c
d1bg6_2	d1a1va1	1	19.650000	5.21	0.159	-1	c	c
d1nija1	**d1rq2a1**	1	3.25e-04	193.41	0.1734	-1	c	c
d1nija1	d1rq2a1	1	19.710000	9.77	0.1092	-1	c	c
d1edg__	**d1adoa_**	1	1.95e-03	225.43	0.2079	-1	c	c
d1edg__	d1adoa_	1	19.950000	1.05	0.1684	-1	c	c

List 2 HHsearch_ss outperforms ProCAIn_ss:

(The first row is ProCAIn results and the second row is the corresponding HHsearch results)

ID1	ID2	SVM	E-value/Prob	Score	GDT_TS	SF	class1	class2
d1rlr_1	d1l1la_	1	2.33e+03	76.75	0.0778	-1	a	c
d1rlr_1	d1l1la_	1	93.490000	30.96	0.2889	-1	a	c
d1peqa1	d1l1la_	1	1.52e+03	73.31	0.1296	-1	a	c
d1peqa1	d1l1la_	1	93.830000	30.77	0.358	-1	a	c
d1sf9a_	d1u9da_	0	1.46e+03	31.88	0.0763	-1	b	d
d1sf9a_	d1u9da_	0	93.700000	34.99	0.1568	-1	b	d
d1u9da_	d1sf9a_	-1	1.13e+03	31.88	0.0738	-1	d	b
d1u9da_	d1sf9a_	-1	93.250000	34.99	0.1516	-1	d	b
d1r7ia1	d1hwla2	1	3.12e+02	67.87	0.1591	-1	d	d
d1r7ia1	d1hwla2	1	96.160000	40.4	0.1909	-1	d	d
d1u1ia2	d1p1ja1	-1	2.42e+02	77.87	0.1738	-1	d	c
d1u1ia2	d1p1ja1	-1	97.920000	61.15	0.1619	-1	d	c
d1qba_1	d1jaka1	-1	1.29e+02	93.57	0.2429	-1	b	c
d1qba_1	d1jaka1	-1	91.410000	31.09	0.2262	-1	b	c
d1jmsa3	d1jiha2	-1	1.01e+02	86.63	0.3792	-1	a	e
d1jmsa3	d1jiha2	-1	91.530000	31.04	0.3333	-1	a	e
d1gkpa1	d1j79a_	-1	8.74e+01	97.14	0.1855	-1	b	c
d1gkpa1	d1j79a_	-1	97.660000	47.47	0.1714	-1	b	c
d1dfca3	d1v7wa2	-1	7.82e+01	68.27	0.2053	-1	b	b
d1dfca3	d1v7wa2	-1	92.110000	19.03	0.2073	-1	b	b
d1a53__	d1ofda2	1	3.28e+01	135.22	0.2621	-1	c	c
d1a53__	d1ofda2	1	95.680000	34.11	0.2611	-1	c	c
d1j6ua1	d1pn0a1	1	3.02e+01	128.9	0.2753	-1	c	c
d1j6ua1	d1pn0a1	1	92.490000	28.31	0.2753	-1	c	c
d1kyqa1	d1m1na_	1	2.81e+01	128.52	0.3283	-1	c	c
d1kyqa1	d1m1na_	1	92.790000	24.57	0.18	-1	c	c
d1i36a2	d1d5ta1	0	2.76e+01	127.63	0.1891	-1	c	c
d1i36a2	d1d5ta1	0	93.790000	28.93	0.199	-1	c	c
d1lsua_	d1m1na_	1	2.47e+01	130.77	0.416	-1	c	c
d1lsua_	d1m1na_	1	91.230000	24.68	0.306	-1	c	c
d1b8pa1	d1k0ia1	0	2.22e+01	128.63	0.1699	-1	c	c

d1b8pa1	d1k0ia1	0	94.600000	30.91	0.1635	-1	c	c
d1r7ja_	d1olta_	-1	1.84e+01	123.06	0.4028	-1	a	c
d1r7ja_	d1olta_	-1	94.730000	30.24	0.3722	-1	a	c
d1i36a2	d1w4xa1	1	1.82e+01	123.57	0.1859	-1	c	c
d1i36a2	d1w4xa1	1	92.950000	27.78	0.1941	-1	c	c
d1nkga1	d1dmha_	1	1.81e+01	121.91	0.4741	-1	b	b
d1nkga1	d1dmha_	1	96.700000	35.25	0.4511	-1	b	b
d1qaza_	d1clc_1	1	1.69e+01	133.62	0.1887	-1	a	a
d1qaza_	d1clc_1	1	92.110000	21.03	0.1645	-1	a	a
d1p6qa_	d1dxea_	0	1.66e+01	101.12	0.3411	-1	c	c
d1p6qa_	d1dxea_	0	91.760000	25.49	0.3353	-1	c	c
d1iz0a2	d1k0ia1	1	1.62e+01	134.68	0.1842	-1	c	c
d1iz0a2	d1k0ia1	1	92.970000	30.11	0.1798	-1	c	c
d1j6ua1	d1k0ia1	1	1.48e+01	127.88	0.3258	-1	c	c
d1j6ua1	d1k0ia1	1	95.200000	34.85	0.3202	-1	c	c
d1nzna_	d1dcea1	1	1.47e+01	84.19	0.375	-1	a	a
d1nzna_	d1dcea1	1	93.530000	24.27	0.4057	-1	a	a
d1jmsa3	d1t94a2	1	1.43e+01	115.41	0.5208	-1	a	e
d1jmsa3	d1t94a2	1	94.190000	32.79	0.5	-1	a	e
d1dfca3	d1ttua3	1	1.39e+01	67.32	0.1585	-1	b	b
d1dfca3	d1ttua3	1	92.380000	27.58	0.1768	-1	b	b
d1j6ua1	d1w4xa1	1	1.35e+01	121.09	0.3258	-1	c	c
d1j6ua1	d1w4xa1	1	93.260000	31.09	0.3287	-1	c	c
d1lu9a1	d1d5ta1	1	1.35e+01	140.4	0.1545	-1	c	c
d1lu9a1	d1d5ta1	1	95.430000	33.42	0.161	-1	c	c
d1lu9a1	d1uwka_	1	1.30e+01	139.35	0.2539	-1	c	e
d1lu9a1	d1uwka_	1	91.610000	26.33	0.2448	-1	c	e
d1lu9a1	d1k0ia1	1	1.16e+01	141.37	0.1584	-1	c	c
d1lu9a1	d1k0ia1	1	95.770000	34.37	0.161	-1	c	c
d1i36a2	d1b37a1	1	1.02e+01	162.59	0.1859	-1	c	c
d1i36a2	d1b37a1	1	93.240000	27.82	0.1908	-1	c	c
d1p4xa1	d1ldja3	1	1.02e+01	119.94	0.346	-1	a	e
d1p4xa1	d1ldja3	1	92.860000	29.64	0.354	-1	a	e
d1i1ra1	d1biha1	1	9.21e+00	63.48	0.4575	-1	b	b
d1i1ra1	d1biha1	1	92.710000	25.48	0.4325	-1	b	b
d1k66a_	d1dxea_	1	9.03e+00	110.56	0.3272	-1	c	c
d1k66a_	d1dxea_	1	91.810000	25.51	0.3205	-1	c	c
d1lu9a1	d1pn0a1	1	8.77e+00	160.77	0.1387	-1	c	c

d1lu9a1	d1pn0a1	1	92.600000	27.18	0.1505	-1	c	c
d1gg4a1	d1lt8a_	1	8.68e+00	125.48	0.3241	-1	c	c
d1gg4a1	d1lt8a_	1	91.460000	24.08	0.2333	-1	c	c
d1j6ua1	d1ngva_	1	8.64e+00	138.55	0.5449	-1	c	c
d1j6ua1	d1ngva_	1	93.050000	29.39	0.5758	-1	c	c
d1id1a_	d1d5ta1	1	8.38e+00	147.32	0.201	-1	c	c
d1id1a_	d1d5ta1	1	94.220000	29.39	0.1961	-1	c	c
d2pgd_2	d1d5ta1	1	8.15e+00	146.68	0.196	-1	c	c
d2pgd_2	d1d5ta1	1	95.410000	35.29	0.1847	-1	c	c
d1w0jd1	d1pvoa3	-1	8.12e+00	102.43	0.1859	-1	a	c
d1w0jd1	d1pvoa3	-1	95.770000	36.29	0.1838	-1	a	c
d1vjxa_	d1udda_	1	7.96e+00	98.07	0.4077	-1	a	a
d1vjxa_	d1udda_	1	94.260000	17.51	0.3691	-1	a	a
d1mkma1	d1ldja3	1	7.93e+00	122.2	0.5333	-1	a	e
d1mkma1	d1ldja3	1	94.760000	30.4	0.5467	-1	a	e
d2pgd_2	d1pn0a1	1	7.47e+00	162.46	0.1648	-1	c	c
d2pgd_2	d1pn0a1	1	91.130000	28.69	0.1719	-1	c	c
d1j6ua1	d1d5ta1	1	7.24e+00	140.93	0.2978	-1	c	c
d1j6ua1	d1d5ta1	1	95.400000	35.74	0.3034	-1	c	c
d1f8fa2	d1d5ta1	1	7.17e+00	147.5	0.1753	-1	c	c
d1f8fa2	d1d5ta1	1	91.420000	26.1	0.171	-1	c	c
d1c1da1	d1e5xa_	1	7.10e+00	144.75	0.2127	-1	c	c
d1c1da1	d1e5xa_	1	91.480000	23.32	0.2139	-1	c	c
d1kkoa1	d1lt8a_	1	6.95e+00	137.28	0.2659	-1	c	c
d1kkoa1	d1lt8a_	1	96.400000	24.49	0.2361	-1	c	c
d1i36a2	d1k0ia1	1	6.92e+00	146.24	0.2072	-1	c	c
d1i36a2	d1k0ia1	1	95.310000	33.03	0.2023	-1	c	c
d1eut_1	d1u2ca1	1	6.90e+00	68.24	0.3592	-1	b	b
d1eut_1	d1u2ca1	1	91.100000	22.61	0.3835	-1	b	b
d1bia_1	d1ldja3	1	6.75e+00	124.22	0.6071	-1	a	e
d1bia_1	d1ldja3	1	91.790000	27.42	0.5873	-1	a	e
d1ohzb_	d1auib_	-1	6.68e+00	86.55	0.3795	-1	a	a
d1ohzb_	d1auib_	-1	91.640000	23.99	0.3795	-1	a	a
d1a9xa3	d1f14a2	1	6.43e+00	99.18	0.2736	-1	c	c
d1a9xa3	d1f14a2	1	91.860000	29.78	0.25	-1	c	c
d1biha1	d1i1ra1	1	6.36e+00	63.48	0.4867	-1	b	b
d1biha1	d1i1ra1	1	93.290000	25.48	0.4601	-1	b	b
d2naca2	d1j6ua1	1	6.31e+00	78.23	0.1257	-1	c	c

d2naca2	d1j6ua1	1	92.970000	32.85	0.1604	-1	c	c
d1saya1	**d1m1na_**	1	6.18e+00	151.78	0.256	-1	c	c
d1saya1	d1m1na_	1	92.480000	25.27	0.2708	-1	c	c
d1tl2a_	**d1crua_**	1	5.93e+00	109.67	0.1809	-1	b	b
d1tl2a_	d1crua_	1	97.380000	37.51	0.2809	-1	b	b
d2pgd_2	**d1k0ia1**	1	5.91e+00	150.03	0.179	-1	c	c
d2pgd_2	d1k0ia1	1	94.800000	34.05	0.179	-1	c	c
d1id1a_	**d1w4xa1**	1	5.89e+00	141.07	0.201	-1	c	c
d1id1a_	d1w4xa1	1	94.680000	30.22	0.201	-1	c	c
d1ti6b1	**d1dmha_**	1	5.78e+00	141.24	0.5475	-1	b	b
d1ti6b1	d1dmha_	1	97.300000	39.67	0.5411	-1	b	b
d1xd7a_	**d1ldja3**	1	5.75e+00	126.96	0.3268	-1	a	e
d1xd7a_	d1ldja3	1	93.930000	28.65	0.3228	-1	a	e
d1udda_	**d1jgca_**	1	5.74e+00	94.62	0.2872	-1	a	a
d1udda_	d1jgca_	1	91.290000	13.25	0.2791	-1	a	a
d2pgd_2	**d1w4xa1**	1	5.66e+00	140.47	0.1861	-1	c	c
d2pgd_2	d1w4xa1	1	94.220000	33.1	0.1776	-1	c	c
d1lu9a1	**d1w4xa1**	1	5.49e+00	141.47	0.2552	-1	c	c
d1lu9a1	d1w4xa1	1	94.020000	29.47	0.1623	-1	c	c
d1fsea_	**d1sd4a_**	1	5.46e+00	88.85	0.4851	-1	a	a
d1fsea_	d1sd4a_	1	93.020000	29.26	0.5187	-1	a	a
d1f8fa2	**d1pn0a1**	1	5.36e+00	167.01	0.1739	-1	c	c
d1f8fa2	d1pn0a1	1	91.630000	27.14	0.171	-1	c	c
d2uaga1	**d1m1nb_**	1	5.31e+00	149.08	0.5242	-1	c	c
d2uaga1	d1m1nb_	1	92.250000	25.06	0.5672	-1	c	c
d1ft9a1	**d1nr3a_**	-1	5.29e+00	75.24	0.175	-1	a	d
d1ft9a1	d1nr3a_	-1	91.410000	29.83	0.1531	-1	a	d
d1adr__	**d1nr3a_**	-1	5.24e+00	75.7	0.1875	-1	a	d
d1adr__	d1nr3a_	-1	92.390000	31.42	0.1711	-1	a	d
d1id1a_	**d1pn0a1**	1	5.14e+00	169.27	0.1846	-1	c	c
d1id1a_	d1pn0a1	1	94.840000	29.15	0.1863	-1	c	c
d1dxea_	**d1p6qa_**	0	5.06e+00	101.12	0.1739	-1	c	c
d1dxea_	d1p6qa_	0	91.260000	25.49	0.1709	-1	c	c
d1lt8a_	**d1kkoa1**	1	4.85e+00	137.28	0.1849	-1	c	c
d1lt8a_	d1kkoa1	1	96.290000	24.49	0.1641	-1	c	c
d1j6ua1	**d2naca2**	1	4.83e+00	78.23	0.264	-1	c	c
d1j6ua1	d2naca2	1	93.780000	32.85	0.3371	-1	c	c
d1f8fa2	**d1w4xa1**	0	4.75e+00	141.68	0.1997	-1	c	c

d1f8fa2	d1w4xa1	0	91.960000	26.65	0.181	-1	c	c
d1udda_	**d1vjxa_**	1	4.71e+00	98.07	0.2826	-1	a	a
d1udda_	d1vjxa_	1	93.880000	17.51	0.2558	-1	a	a
d1fsea_	**d1ku9a_**	1	4.63e+00	97.25	0.4851	-1	a	a
d1fsea_	d1ku9a_	1	91.550000	27.24	0.4963	-1	a	a
d1j6ua1	**d1m3sa_**	1	4.49e+00	114.09	0.486	-1	c	c
d1j6ua1	d1m3sa_	1	91.150000	27.7	0.2612	-1	c	c
d1jhga_	**d1p4xa1**	1	4.44e+00	85.67	0.2847	-1	a	a
d1jhga_	d1p4xa1	1	91.260000	23.58	0.2995	-1	a	a
d1nnxa_	**d1gm5a2**	1	4.44e+00	102.59	0.467	-1	b	b
d1nnxa_	d1gm5a2	1	92.430000	25.32	0.4575	-1	b	b
d1geqa_	**d1ofda2**	1	4.32e+00	165.35	0.2429	-1	c	c
d1geqa_	d1ofda2	1	93.070000	25.6	0.2591	-1	c	c
d1nkga1	**d1eo9a_**	1	4.21e+00	88.85	0.3793	-1	b	b
d1nkga1	d1eo9a_	1	94.310000	28.08	0.4167	-1	b	b
d1f14a2	**d1a9xa3**	1	4.14e+00	99.18	0.181	-1	c	c
d1f14a2	d1a9xa3	1	93.670000	29.78	0.1654	-1	c	c
d1jmsa3	**d1gm5a2**	1	4.00e+00	99.75	0.3583	-1	a	b
d1jmsa3	d1gm5a2	1	95.280000	35.99	0.4333	-1	a	b
d1iv0a_	**d1huxa_**	1	3.97e+00	138.83	0.301	-1	c	c
d1iv0a_	d1huxa_	1	93.640000	28.84	0.3087	-1	c	c
d1h8la1	**d1dmha_**	1	3.94e+00	147.37	0.6108	-1	b	b
d1h8la1	d1dmha_	1	97.730000	43.08	0.6171	-1	b	b
d1f8fa2	**d1k0ia1**	1	3.82e+00	155.01	0.1925	-1	c	c
d1f8fa2	d1k0ia1	1	95.730000	34.19	0.1782	-1	c	c
d1j6ua1	**d1el5a1**	1	3.76e+00	128.38	0.3455	-1	c	c
d1j6ua1	d1el5a1	1	94.770000	33.72	0.3371	-1	c	c
d2pgd_2	**d1b37a1**	1	3.67e+00	181.48	0.1747	-1	c	c
d2pgd_2	d1b37a1	1	92.350000	29.29	0.179	-1	c	c
d1j6ua1	**d1b37a1**	1	3.59e+00	174.06	0.3034	-1	c	c
d1j6ua1	d1b37a1	1	95.290000	34.98	0.2978	-1	c	c
d1b8pa1	**d1el5a1**	1	3.56e+00	132	0.1699	-1	c	c
d1b8pa1	d1el5a1	1	92.630000	28.32	0.1635	-1	c	c
d1c1da1	**d1m1na_**	0	3.49e+00	164.06	0.2363	-1	c	c
d1c1da1	d1m1na_	0	91.250000	25.39	0.2102	-1	c	c
d1jx7a_	**d1t5ba_**	1	3.42e+00	105.32	0.2906	-1	c	c
d1jx7a_	d1t5ba_	1	91.750000	24.03	0.3312	-1	c	c
d1ks9a2	**d1rzua_**	1	3.33e+00	161.18	0.241	-1	c	c

d1ks9a2	d1rzua_	1	91.500000	25.36	0.1512	-1	c	c
d1cuk_2	d1pu6a_	1	3.30e+00	108.76	0.4038	-1	a	a
d1cuk_2	d1pu6a_	1	93.360000	30.06	0.2917	-1	a	a
d1ohzb_	d2scpa_	-1	3.24e+00	100.79	0.4375	-1	a	a
d1ohzb_	d2scpa_	-1	93.250000	26.67	0.4152	-1	a	a
d1lsua_	d1uwva2	1	3.19e+00	158.35	0.5261	-1	c	c
d1lsua_	d1uwva2	1	93.280000	24.23	0.4328	-1	c	c
d1r0da_	d1h6ga1	1	3.19e+00	79.93	0.2719	-1	a	a
d1r0da_	d1h6ga1	1	91.750000	13.89	0.317	-1	a	a
d1jhga_	d1fsea_	1	3.17e+00	70.39	0.2822	-1	a	a
d1jhga_	d1fsea_	1	93.090000	29.66	0.2847	-1	a	a
d1k66a_	d1xi3a_	1	3.14e+00	102.14	0.2651	-1	c	c
d1k66a_	d1xi3a_	1	91.650000	21.72	0.297	-1	c	c
d1v5oa_	d1ip9a_	1	3.13e+00	81.22	0.3137	-1	d	d
d1v5oa_	d1ip9a_	1	94.790000	30.64	0.3137	-1	d	d
d1ku9a_	d1ldja3	1	3.12e+00	133	0.2864	-1	a	e
d1ku9a_	d1ldja3	1	94.970000	31.47	0.2864	-1	a	e
d1fsea_	d1jhga_	1	3.06e+00	70.39	0.4254	-1	a	a
d1fsea_	d1jhga_	1	92.930000	29.66	0.4291	-1	a	a
d1klfa1	d1e42a1	1	2.96e+00	73.92	0.1777	-1	b	b
d1klfa1	d1e42a1	1	91.390000	21.44	0.3554	-1	b	b
d1knwa2	d1jaka1	1	2.91e+00	158.08	0.1528	-1	c	c
d1knwa2	d1jaka1	1	93.970000	28.44	0.1569	-1	c	c
d1j6ua1	d1q7ra_	1	2.90e+00	101.9	0.4972	-1	c	c
d1j6ua1	d1q7ra_	1	93.760000	31.15	0.4831	-1	c	c
d1t5ba_	d1jx7a_	1	2.87e+00	105.32	0.1692	-1	c	c
d1t5ba_	d1jx7a_	1	92.930000	24.03	0.1928	-1	c	c
d1i36a2	d1seza1	1	2.82e+00	171.27	0.1941	-1	c	c
d1i36a2	d1seza1	1	93.960000	29.04	0.1957	-1	c	c
d1eo9a_	d1nkga1	1	2.80e+00	88.85	0.1634	-1	b	b
d1eo9a_	d1nkga1	1	94.110000	28.08	0.1795	-1	b	b
d1doi__	d1ep3b2	-1	2.72e+00	105.73	0.168	-1	d	c
d1doi__	d1ep3b2	-1	94.870000	35.72	0.1582	-1	d	c
d1du2a_	d1rrza_	-1	2.70e+00	64.44	0.2072	-1	a	a
d1du2a_	d1rrza_	-1	91.560000	27.81	0.2237	-1	a	a
d1a04a2	d1fq0a_	1	2.63e+00	112.23	0.2862	-1	c	c
d1a04a2	d1fq0a_	1	92.890000	27.48	0.3533	-1	c	c
d1ip9a_	d1v5oa_	1	2.57e+00	81.22	0.3765	-1	d	d

d1ip9a_	d1v5oa_	1	94.760000	30.64	0.3765	-1	d	d
d1npya1	d1d5ta1	1	2.56e+00	161	0.1707	-1	c	c
d1npya1	d1d5ta1	1	92.590000	28.62	0.1766	-1	c	c
d1qwya_	d1e2wa2	1	2.50e+00	81.75	0.15	-1	b	b
d1qwya_	d1e2wa2	1	95.720000	32.71	0.1556	-1	b	b
d1c1da1	d1seza1	1	2.43e+00	173.46	0.1542	-1	c	c
d1c1da1	d1seza1	1	93.420000	29.8	0.1567	-1	c	c
d1o94a1	d1g5aa2	1	2.41e+00	209.08	0.1044	-1	c	c
d1o94a1	d1g5aa2	1	93.190000	20.24	0.1824	-1	c	c
d1lu9a1	d1b37a1	1	2.36e+00	189.53	0.1505	-1	c	c
d1lu9a1	d1b37a1	1	94.780000	31.33	0.1558	-1	c	c
d1id1a_	d1b37a1	1	2.35e+00	190.62	0.2026	-1	c	c
d1id1a_	d1b37a1	1	94.430000	29.8	0.2026	-1	c	c
d2uaga1	d1pn0a1	1	2.35e+00	177.36	0.3226	-1	c	c
d2uaga1	d1pn0a1	1	95.920000	37.47	0.328	-1	c	c
d1buoa_	d1fs1b1	1	2.33e+00	82.11	0.2479	-1	d	a
d1buoa_	d1fs1b1	1	91.230000	27.98	0.2459	-1	d	a
d1rq2a1	d1j5va_	1	2.32e+00	145.18	0.3497	-1	c	c
d1rq2a1	d1j5va_	1	91.440000	17.24	0.3321	-1	c	c
d1p4xa1	d1jhga_	1	2.31e+00	85.67	0.23	-1	a	a
d1p4xa1	d1jhga_	1	92.630000	23.58	0.242	-1	a	a
d1cuk_2	d1vdda_	0	2.27e+00	112.63	0.2853	-1	a	e
d1cuk_2	d1vdda_	0	97.380000	44.55	0.3045	-1	a	e
d1j6ua1	d1pj5a2	1	2.27e+00	133.25	0.3118	-1	c	c
d1j6ua1	d1pj5a2	1	95.140000	35.04	0.2978	-1	c	c
d1b16a_	d1f3la_	1	2.24e+00	160.08	0.19	-1	c	c
d1b16a_	d1f3la_	1	92.710000	26.27	0.1909	-1	c	c
d1m66a2	d1ngva_	1	2.23e+00	166.56	0.2738	-1	c	c
d1m66a2	d1ngva_	1	95.320000	27.95	0.1574	-1	c	c
d2rsla_	d1ny1a_	0	2.22e+00	138.9	0.2418	-1	c	c
d2rsla_	d1ny1a_	0	92.040000	25.61	0.2418	-1	c	c
d1ldja3	d1p4xa1	1	2.18e+00	119.94	0.1567	-1	e	a
d1ldja3	d1p4xa1	1	92.950000	29.64	0.1603	-1	e	a
d1ghk__	d1gpr__	1	2.17e+00	88.93	0.3006	-1	b	b
d1ghk__	d1gpr__	1	94.230000	26.78	0.3133	-1	b	b
d1o4ua1	d1o94a1	1	2.17e+00	140.53	0.3059	-1	c	c
d1o4ua1	d1o94a1	1	94.630000	27.72	0.325	-1	c	c

Bibliography

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Barnhart, B. J. (1989). "The Department of Energy (DOE) Human Genome Initiative." Genomics **5**(3): 657-60.

Byvatov, E. and G. Schneider (2003). "Support vector machine applications in bioinformatics." Appl Bioinformatics **2**(2): 67-77.

Chung, R. and G. Yona (2004). "Protein family comparison using statistical models and predicted structural information." BMC Bioinformatics **5**: 183.

D'Amours, D. and S. P. Jackson (2002). "The Mre11 complex: at the crossroads of dna repair and checkpoint signalling." Nat Rev Mol Cell Biol **3**(5): 317-27.

Das, R., B. Qian, et al. (2007). "Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home." Proteins **69 Suppl 8**: 118-28.

Durbin, R. E., S. Krogh, A. Mitchison, G. (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge, UK, Cambridge University Press.

Ealick, S. E. (2000). "Advances in multiple wavelength anomalous diffraction crystallography." Curr Opin Chem Biol **4**(5): 495-9.

Eddy, S. (1997). Maximum likelihood fitting of extreme value distributions. Maximum likelihood fitting of extreme value distributions.

Eddy, S. F. (1997). Maximum likelihood fitting of extreme value distributions. Maximum likelihood fitting of extreme value distributions.

Eddy, S. R. (1998). "Profile hidden Markov models." Bioinformatics **14**(9): 755-63.

Full, C. and K. Poralla (2000). "Conserved tyr residues determine functions of Alicyclobacillus acidocaldarius squalene-hopene cyclase." FEMS Microbiol Lett **183**(2): 221-4.

Gumbel, E. J., Ed. (1958). Statistics of Extremes. New York, Columbia Univeristy Press.

Holm, L. and C. Sander (1996). "Mapping the protein universe." Science **273**(5275): 595-603.

Joachims, T. (1999). Making Large-Scale SVM Learning Practical. Advances in kernel methods: support vector learning. B. C. Scholkopf B, Smola AJ. Cambridge, Mass., MIT press.

Karlin, S. and S. F. Altschul (1990). "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes." Proc Natl Acad Sci U S A **87**(6): 2264-8.

Karlin, S. and S. F. Altschul (1993). "Applications and statistics for multiple high-scoring segments in molecular sequences." Proc Natl Acad Sci U S A **90**(12): 5873-7.

Kinch, L. N., J. O. Wrabl, et al. (2003). "CASP5 assessment of fold recognition target predictions." Proteins **53 Suppl 6**: 395-409.

McGuffin, L. J., K. Bryson, et al. (2000). "The PSIPRED protein structure prediction server." Bioinformatics **16**(4): 404-5.

Murzin, A. G., S. E. Brenner, et al. (1995). "SCOP: a structural classification of proteins database for the investigation of sequences and structures." J Mol Biol **247**(4): 536-40.

Nayeem, A., D. Sitkoff, et al. (2006). "A comparative study of available software for high-accuracy homology modeling: from sequence alignments to structural models." Protein Sci **15**(4): 808-24.

Pace, H. C. and C. Brenner (2001). "The nitrilase superfamily: classification, structure and function." Genome Biol **2**(1): REVIEWS0001.

Pei, J. and N. V. Grishin (2001). "AL2CO: calculation of positional conservation in a protein sequence alignment." Bioinformatics **17**(8): 700-12.

Qi, Y., R. I. Sadreyev, et al. (2007). "A comprehensive system for evaluation of remote sequence similarity detection." BMC Bioinformatics **8**: 314.

Rychlewski, L., D. Fischer, et al. (2003). "LiveBench-6: large-scale automated evaluation of protein structure prediction servers." Proteins **53 Suppl 6**: 542-7.

Sadreyev, R. and N. Grishin (2003). "COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance." J Mol Biol **326**(1): 317-36.

Sadreyev, R. I. and N. V. Grishin (2008). "Accurate statistical model of comparison between multiple sequence alignments." Nucleic Acids Res **36**(7): 2240-8.

Schaffer, A. A., L. Aravind, et al. (2001). "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements." Nucleic Acids Res **29**(14): 2994-3005.

Siddharthan, R., E. D. Siggia, et al. (2005). "PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny." PLoS Comput Biol **1**(7): e67.

Smith, T. F. and M. S. Waterman (1981). "Identification of common molecular subsequences." J Mol Biol **147**(1): 195-7.

Soding, J. (2005). "Protein homology detection by HMM-HMM comparison." Bioinformatics **21**(7): 951-60.

Sonego, P., A. Kocsor, et al. (2008). "ROC analysis: applications to the classification of biological sequences and 3D structures." Brief Bioinform **9**(3): 198-209.

Sonnhammer, E. L. and R. Durbin (1994). "A workbench for large-scale sequence homology analysis." Comput Appl Biosci **10**(3): 301-7.

Sunyaev, S. R., F. Eisenhaber, et al. (1999). "PSIC: profile extraction from sequence alignments with position-specific counts of independent observations." Protein Eng **12**(5): 387-94.

Tatusov, R. L., S. F. Altschul, et al. (1994). "Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks." Proc Natl Acad Sci U S A **91**(25): 12091-5.

Tyszka, J. M., S. E. Fraser, et al. (2005). "Magnetic resonance microscopy: recent advances and applications." Curr Opin Biotechnol **16**(1): 93-9.

Wendt, K. U., K. Poralla, et al. (1997). "Structure and function of a squalene cyclase." Science **277**(5333): 1811-5.

Zemla, A. (2003). "LGA: A method for finding 3D similarities in protein structures." Nucleic Acids Res **31**(13): 3370-4.

Zhang, Y., A. K. Arakaki, et al. (2005). "TASSER: an automated method for the prediction of protein tertiary structures in CASP6." Proteins **61 Suppl 7**: 91-8.