Analysis of Reliability of Protein Functional Site Prediction Methods: A Comparison of Webservers and a Manual Pipeline.

Aakriti Jain, Bernardo Cervantes, Brian St. Aubin, Michael Ting, Ramya Prathuri, Sharyu Barapatrey, Thomas Chow

BioE C144/C144L Spring 2013 session, Department of Bioengineering, University of California, Berkeley

Abstract:

Reliable functional site prediction is beneficial for multiple applications such as protein engineering, site directed mutagenesis, finding novel functions, understanding specificity, and function prediction. Numerous web servers have been developed to predict functional sites, taking advantage of varied combinations of sequence and structure information. Despite great advances in the field, it is evident that improvements are necessary to better address the needs of the community. Here, we assess the performance of six functional site prediction webservers using three carefully selected, well-characterized proteins. In addition, we define and utilize a manual prediction pipeline to intelligently identify functional sites. Webserver predictions were summarized in a consensus approach, and compared to the manual prediction and the literature accepted data. These comparisons are meant to outline the strengths and weaknesses of state of the art prediction servers. Our results and analysis suggest that manual functional site predictions can be very reliable, and web server performance can fluctuate significantly between query submissions.

Introduction:

Functional site prediction is an important process in the field of bioinformatics due to its application in other endeavors such as protein engineering, drug-target identification, and enzyme function prediction. Functional sites include a variety of annotations that can be linked to particular residues in a sequence. Catalytic annotations (also known as active sites) are the best understood functional residues. These are defined as directly participating in the chemical reaction that an enzyme catalyzes [9]. To account for our growing understanding of enzyme functionality, the bioinformatics community has focused on the development of catalytic site prediction web servers.

Most catalytic site prediction web servers take advantage of sequence information. One of the most valuable pieces of sequence information relies on site-specific evolutionary conservation. Catalytic sites are known to be among the mostly highly conserved residues due to their direct involvement in chemical activity [10]. Mutations at catalytic sites are seldom observed, and they usually indicate a significant change in the function or specificity of the enzyme [10]. Another important piece of sequence information for the prediction of catalytic sites is the residue type. It has been previously shown that polar and charged residues are present in catalytic sites with a higher frequency [11]. Residue type information is very helpful in differentiating residues that have been conserved due to structural importance, from residues that have been conserved due to their participation in catalytic activity.

More recent functional site prediction methods, such as I-Tasser, Evo-Trace, and Discern have been successful in incorporating structural information as part of their prediction protocol.

Structural information can elucidate a vast variety of properties that increase or decrease the catalytic propensity of a site. Catalytic residues are expected to be near each other in space, likely to reside in a cleft (also known as binding pocket), and likely to have some degree of solvent exposure [12, 13, 14]. Inclusion of structure information while predicting catalytic sites has been shown to improve performance of web servers, particularly when coupled with advanced machine learning algorithms [6].

Here we compare the performance of a variety of web servers that span the different existing methodologies. The webservers examined are Consurf [3], Pool [1], I-Tasser [2], Prosite [5], Evo-Trace[4], and Smart [15]. In addition, we assess the plausibility of manually predicting catalytic sites using a thoughtful manual prediction protocol that incorporates structure and sequence information. We aim to expose the strengths and weaknesses of different methodologies by analyzing three different proteins referred to as case studies. Bovine alpha-Chymotrypsin (E.C. 3.4.21.1), Glutathione S-transferase (E.C. 2.5.1.18), and VirB4 (E.C. 2.3.1.129) were carefully selected to represent different challenges commonly observed during catalytic site prediction. Ultimately, we hope to assess the feasibility and accuracy of catalytic site prediction using both manual prediction and a consensus webserver approach.

Materials and Methods

General Pipeline:

A schematic of the protocol used is depicted in Figure 1. Proteins for which structures have been solved and active sites have been experimentally identified were chosen. The three chosen proteins that match these criteria were obtained through literature searches and used to query six webservers: POOL, I-TASSER, CONSURF, SMART, Evolutionary Trace, and Prosite. Simultaneously, the chosen proteins were submitted to a manual catalytic site prediction protocol through which a predicted structure and multiple sequence alignment of distant homologs were generated. The combined results were used to make assertions about catalytic sites in the chosen proteins, and the accuracy of the results was measured against known experimental data for each protein.

Manual Protocol:

A protein was submitted to the database searching program PSI-BLAST [20] to obtain homologs from the non-redundant protein database. Two iterations of the algorithm were used with an E-value cutoff of 0.005. About 500 sequences were initially aligned using the MAFFT [18] algorithm for multiple sequence alignment under automatic conditions. The resulting MSA was masked using Jalview 2.0 [16] or Belvu [17] to limit redundancy to less than 50% sequence identity, remove partial sequences, and remove putative proteins. We did this to ensure that the conserved residues we identified were informative in terms of functional conservation across evolution. The remaining ~400 sequences were inspected to find columns with above 95 – 98% sequence identity. These columns were interpreted to describe functionally important sites. Small variations in the construction of the alignment were observed between each case study.

For Bovine chymotrypsin and glutathione s-transferase the structures was available, but for VirB4 the structure had not been released. To ensure that additional noise was not introduced to the protocol, the solved structures for these two proteins were used. The query proteins were also submitted to the protein structure prediction server Phyre2 [19]. Using a combination of HMMs and threading algorithms, Phyre2 outputs a structure for the query sequence. Fortunately, Phyre2 identified the query proteins as the highest confidence templates. Phyre2's predictive model capabilities were particularly important for obtaining an accurate VirB4 model for active site analysis.

The structures were then used as maps onto which the conserved residues highlighted on the MSA were mapped. Given the placement and identity of the residues as detailed earlier, residues that may be active sites were identified. Of the conserved residues, we considered only the ones that were solvent accessible, either in a binding cleft or next to a residue in a cleft, and those known to have high catalytic propensity. Residues that met these criteria became our manual active site prediction.

Catalytic Site Prediction Webservers analyzed in this paper:

POOL [1], or Partial Order Optimum Likelihood, analyzes the electrostatic properties using THEMATICS and the 3D structure for binding clefts using ConCavity of a submitted protein structure, and outputs a list of residues ranked by importance in function. The INTREPID webserver that utilizes phylogenetic tree information for a protein, is usually tied to POOL, but the former is currently not functioning.

Consurf [3] works by taking in data in the form of a nucleotide/amino acid sequence (required) and an MSA, protein structure, and/or a phylogenetic tree (all optional). Consurf then finds homologs of the original sequence through PSI-BLAST or CSI-BLAST and removes redundant (isoforms or partial sequences) sequences. From here all the sequences are aligned and a phylogenetic tree is constructed. A conservation rate for every position of the MSA is calculated using Rate4Site.

Prosite [5] is a database of protein families and domains which are used to identify the presence of any domains on the query sequence aided by Prorule. Prorule is a set of manually created rules that describe the required characteristics for any particular domain in the database. Prosite serves as a functional site prediction by making use of the annotations in the database.

I-TASSER [2] uses LOMETS (Local Meta-Threading-Server) to perform secondary structure prediction and uses the results to find matches to structures in PDB. These structures are then manipulated and filtered to develop a structure for the query. The PDB is then searched with the resulting structure to find closely aligning structures. Finally, conserved residues, GO terms, and EC classification are analyzed to inform the functional site selection presented to the user.

SMART [15] is a webserver devoted to the identification and annotation of domains as well as the analysis of domain architectures. It indirectly serves as a functional site prediction because the domains in the database are extensively annotated and include functionally important residues.

Evolutionary Trace [4] takes consecutive vertical slices of the tree being analyzed starting at the root. Then identifies the most conserved residues and plots them onto a 3D structure while identifying their proximity to each other and solvent accessibility. Results are generated in the form of a report, which is easy to follow and contains various types of relevant information.

Results:

Each case study is discussed below. Key differences in the proteins and the difficulties they present to active site prediction are highlighted.

Case Study 1:

Bovine chymotrypsin alpha was selected due to the extensive research that has been done on the protein. Such a well characterized sequence serves as the best case scenario for the prediction ability of each method. Chymotrypsin alpha is a serine endopeptidase/ peptidase S1 (Uniprot accession: P00766) that cleaves polypeptides on the carboxyl end of Tyr, Trp, Leu, or Phe as long as the following residue is not proline. The protein consists of a single globular domain known as trypsin like serine protease (pfam PF00089 [21]).

As part of the analysis of results, a web server consensus was generated by summating all

webserver predictions (no penalties or weights were assigned in this consensus). It is evident that the existence of manually curated predictions from Swiss-Prot, and the availability of a fully solved structure aid in the overall performance of web servers (Figure 2). It is possible that some of the web servers have saved annotations for well characterized sequences, and thus the results obtained may not be de-novo predictions.

Using the manual prediction protocol, six potential catalytic residues were identified: Asp 102, His 57, Trp 141, Ser 195, Asp 194, and Ser 214. These conserved residues marked with hydrophobicity index (See Figure 3) satisfied the necessary conditions to be potentially active residues: solvent accessible, found either in a binding cleft or next to a residue in a cleft: members of a list of frequently occurring catalytic residues. The chosen coloring scheme allowed for easy identification of residues that are more likely to be buried in a binding cleft and therefore more likely to be involved in catalytic activity.

The Phyre2 structure predicted with highest confidence for the query was 1GL1. This was one among three other 100% confidence results. The structure used was chosen because it was labeled to be the solved structure of the bovine chymotrypsin. Indeed, a figure showing the predicted active residues mapped onto the structure shows that the residues fit very well into the main binding pocket of the protein.

The confirmed catalytic residues from Sjolander et al. (2009) are His 57, Asp 102, and Ser 195. Thus, manual prediction yields all of the active residues. The additional residues predicted, Trp 141, Asp 194, and Ser 214, are artifacts of the method of analysis used, and further analysis revealed that these residues are also involved in catalytic function. Tryptophan is sometimes classified as a polar residue that can potentially have catalytic activity [6]. Aspartate 194 and Serine 214 are also located in the binding cleft and are also often classified as polar and charged by many webservers, potentially biasing their selection as active residues.

Case Study 2: Glutathione S-transferase (uniprot: P08515, E.C.: 2.5.1.18, PDB: 1M9B)

We selected glutathione s-transferase, a protein that functions in detoxification by GSH conjugation to substrates. There were two reasons for this second selection. First, it falls in a different functional category (indicated by the enzyme commission number) than the first protein, which helps evaluate how generalizable different methods may be. Second, this protein has extensive literature support for its functions so we can compare our analysis results to the true functions of the protein.

Our MSA prediction (See figure 4) alone yielded the following residues: Y7, W8, R18, L21, L55, P56, Y57, N60, Q67, I71, N152, F153, L176 (all conserved with at least 98% sequence identity in our MSA). We mapped these residues onto the structure of glutathione s-transferase verified by X-ray crystallography (PDB ID: 1M9B), and used structural criteria for active sites to prune residues that did not fit this criteria from our set of predictions (See Figure 5). Criteria for catalytic residue prediction included residues that were solvent accessible, found either in a binding cleft or next to a residue in a cleft, and membership in a list of frequently occurring catalytic residues. The set of conserved amino acids was narrowed down to Y7, W8, R18, L21, L55, P56, Y57, N60, Q67, our candidates for potential active site residues.

Like protein 1, an unweighted webserver consensus approach was used to record the total number of appearances of residues listed as functional sites (Figure 6). The webservers ConSurf, POOL, I-TASSER, PredictProtein (which uses Prosite for functional site prediction), Evolutionary Trace, Swiss-Prot, and SMART were queried for functional site information. Notably, SMART returned no hits. Unlisted in the figure is the Catalytic Site Atlas (CSA), which when queried, only listed Tyr7 as a functional residue.

We omitted this result from the table due to lack of information. The sum totals of appearances of functional residues are highlighted in blue. Experimentally determined functional sites from Cardoso et al (2003) were used as a gold standard for this protein.

From our MSA, we were able to obtain predictions for half of the residues with a webserver prediction total greater than or equal to 4 (Y7, W8, Q67). However, it is notable that our MSA prediction yielded many false positives (predicted by our manual MSA but not by the literature gold standard; residues R18, L21, P56, Y57) and false negatives (not predicted by our manual MSA but was predicted by the literature gold standard; residues G12, L13, W41, K45, N54, L55, S68, D101, R103, S107, Y111, Q204, Q207). This may be the result of slight error in our MSA, as masking to produce results from the MSA is very subjective and can easily affect conservation and identity scores of residues, changing our manual prediction of functional residues.

Using additional webservers in our prediction would have been advantageous since we only sampled 8 webservers and removed results from one (CSA) from the data comparison. Our results indicate that manual prediction of functional residues using an MSA can be difficult and error-prone, depending on the skill of whoever is performing masking and analysis of the MSA. A possible solution to this would be to use a consensus manual prediction method in an attempt to mitigate bias from individual maskers.

Case Study #3: VirB4 (Uniprot: P17794)

VirB4 is the final protein that was selected for analysis. The sequence used is from Agrobacterium tumefaciens. This protein is defined in Kegg as a type IV secretion system protein. It belongs in the TrbE/VirB4 family, and its possible function is to provide energy, via ATP hydrolysis, for the translocation of virulence proteins of the transfer of a T-DNA-protein complex across the Agrobacterium membrane. This protein is one of the first in its family to have x-ray crystallography data, which means that it doesn't have many structural homologs that have solved 3D structure. Furthermore, this protein was also picked because it was studied extensively in literature (Middleton et al). The conclusions from the Middleton et al paper were used as the gold standard for the VirB4 case study.

Since the 3D structure is not available for VirB4 in agrobacterium, functional site prediction for it becomes slightly more complicated because the primary amino acid sequence doesn't necessarily give information about where the amino acid falls within the 3D structure. Therefore, while it is easy to see which positions are most conserved, it is difficult to tell which positions are important due to their location on the 3D structure. Therefore, a comparative model was used in place of a solved 3D structure to determine where the most conserved sites fall.

The multiple-sequence alignment was constructed according to the protocol presented earlier in the paper and included both distant homologs and the PDB template that was found to be most similar: 1E9R. The MSA was visualized and masked using Belvu and was similar to the one found in Middleton et al, with respect to the motifs and conserved residues that were found in that paper (WalkerA, WalkerB, and the conserved glutamine at position 668). These motifs are circled in red in the MSA and the conserved glutamine is highlighted in green (Fig. 7). By using the MSA, other conserved residues were identified across other parts of the VirB4 molecule. However, they were narrowed down and identified as being potentially functionally or structurally important by mapping the residues onto the 3D Phyre model. The residues that were thus identified (D465, W573, L578, W580, R657, K658) are circled in black on the MSA and all of the identified motifs and conserved residues are mapped on the 3D Phyre model of VirB4 (Fig. 8). From the 3D structure, we determined that of the 9 conserved residues/motifs, only 5 were most likely

functionally important. Since the Walker A, Walker B, and glutamine residues are present in a cleft, fairly solvent exposed, and clustered together in 3D space, it is likely that these residues are important for protein function. From the analyses and literature, it was determined that the WalkerA and WalkerB motifs are functionally important because they are both part of the AAA+ domain and are thus important for ATP binding. Furthermore, the conserved lysine and arginine residues are probably also important for function since they too are solvent exposed. In fact, upon further literature search, we determined that R657 and K658 are likely part of an arginine finger; this is structurally important because it allows the VirB4 to oligomerize, much like VirD4 does, and helps form a functional ATP binding site. Also important to note about the arginine finger is that it is far from the ATP binding site so the presence of it cannot disrupt the structure of the binding site. [22] Our manual prediction protocol predicted more residues than what the gold standard predicted. It is possible that these residues are important for structure, but not necessarily important for function.

Because VirB4 does not have a known 3D structure and since it is one of the first proteins in its family to have x-ray crystallography data, there hasn't been much work done on functional site prediction for VirB4 by web-server based predictors. In fact, VirB4 is a CASP protein, which means that it is a fairly difficult protein to conduct functional site prediction on. Due to this reason, most of the webservers that were used in this case study failed. SMART was the only webserver that produced possible results. SMART for VirB4 only returned possible domains -- the CagE_TrbE_VirB domain and AAA_10 ATPase domain. These domains were the same as the ones shown in Pfam. The SMART description also said that the Walker A and B motifs were part of the AAA10_ATPase domain which is consistent with the gold standard. These results are summarized in Fig. 8. SwissProt only predicted the Walker A motif as a conserved/functionally important domain.

All other webservers failed at producing any results. I-Tasser said that VirB4 is a CASP sequence, so it did not process the query. Evolutionary trace did not process either PDB ID corresponding to VirB4 (4AG5 or 4AG6), but it did process the close homolog 1E9R. However, this result does not correlate with the protein in question, so that was a failed webserver. CSA did not contain our query. Prosite returned "no hits" even after scanning against the consensus pattern for the AAA protein family. It was interesting, however, that both Interpro and PredictProtein returned results even though they used Prosite. These two webservers possibly use the same database as Prosite, but a different algorithm to find the functional important residues. Both Interpro and PredictProtein predicted the Walker A motifs, and the R657 and K658 sites as functionally important. POOL returned that the job crashed during a pre-processing step using TINKER due to high potential atoms, and Consurf lost the query in the consensus alignment.

Conclusion

Advances in bioinformatics have generated state-of-the-art methods that take advantage of different types of information for catalytic site prediction. This paper has detailed results of several functional site prediction webservers and compared them with the results of our manual functional site prediction pipeline. It is evident from our analysis that methods combining sequence and structure information have the best performance as has been demonstrated before. Despite the complexity of new methods, weaknesses can be identified for most webservers and manual prediction.

Our manual prediction relies on the existence of a solved structure for the given protein. A comparison of accuracy between Case Studies 1 and 2 versus Case study 3 shows that lacking a solved structure can result in many

falsely predicted active sites. Similarly, webservers did not perform as well when there was no solved structure to the query.

Although a consensus webserver approach may identify functional sites, individual webservers tend to exhibit wide variation in their predictions, resulting in a lack of confidence in each prediction.

Manual prediction methods also exhibited variation in results, as seen from the predictions for all three case studies. Manual prediction methods may suffer from biases during the masking process in the form of both pre-emptive knowledge of meaningful sequence patterns and the experience of the person performing the masking.

Future work should include automation of the manual pipeline, improvement of the consensus webserver approach (by applying penalties and weights), and the reactivation of Discern. A caveat of automation of the manual pipeline is that alignment masking can be a very subjective process, and it relies on the knowledge and experience of the masker to produce the best results.

It is important to remember that functional sites include more than just catalytic sites. We should realize that in the future, information may become available for every residue in a particular sequence. Development of new webservers should take this into account when constructing databases and pipelines.

To progress the ability of scientists to correctly predict functional sites and active residues, the availability of high performance prediction servers such as Discern is critical. An automation of the manual protocol we used, which seemed to work better than some web servers may also provide added insight. The development of a meta-server approach could also improve the accuracy and coverage of predictions for most proteins.

References

1. POOL: Tong W., et al. Partial Order Optimum Likelihood (POOL): maximum likelihood prediction of protein active site residues using 3D structure and sequence properties. PLoS Comput. Biol.2009;5:e1000266.
2. I-TASSER:Ambrish Roy, Alper Kucukural, Yang Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. Nature Protocols, vol 5, 725-738 (2010).
3. Consurf:Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, et al. (2003) ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. Bioinformatics 19: 163–164.
4. Evolutionary Trace:Lichtarge, O., H.R. Bourne and F.E. Cohen (1996). "An evolutionary trace method defines binding surfaces common to protein families." J. Mol. Biol. 257(2): 342-58.
5. Prosite:Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. *New and continuing developments at PROSITE* Nucleic Acids Res. 2012; doi: 10.1093/nar/gks1067 PubMed: 23161676
6. Sankararaman S, Sha F, Kirsch JF, Jordan MI, Sjolander K. Active site prediction using evolutionary and structural information. Bioinformatics. 2010;26(5):617–624. doi: 10.1093/bioinformatics/btq008.
7. Cardoso, R. M.F., Daniels, D. S., Bruns, C. M. and Tainer, J. A. (2003), Characterization of the electrophile binding site and substrate binding mode of the 26-kDa glutathione S-transferase from *Schistosoma japonicum*. Proteins, 51: 137–146. doi:10.1002/prot.10345
8. Middleton et al "Predicted hexameric structure of the Agrobacterium VirB4 C terminus suggests VirB4 acts as a docking site during type IV secretion" PNAS 2005.
9. Craig T. Porter, Gail J. Bartlett, and Janet M. Thornton. The Catalytic Site Atlas: a resource of catalytic sites and residues identiÆed in enzymes using structural data. Nucleic Acids Research, 2004, Vol. 32, Database issue D129-D133. DOI: 10.1093/nar/gkh02
10. George Casari, Chris Sander, and Alfonso Valencia. A method to predict functional residues in proteins. Structural biology volume 2 number 2 February 1995. 1996 Nature Publishing Group. http://www.nature.com/nsmb
11. Gail J. Barlett, et. al., Analysis of Catalytic Residues in Enzyme Active Sites. doi:10.1016/S0022-2836(02)01036-7 available online at http://www.idealibrary.com J. Mol. Biol.(2002)324, 105–121
12. Fetrow,J. and Skolnick,J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. J. Mol. Biol.,281, 949–968.
13. Peters,K.P. et al. (1996) The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. J. Mol. Biol., 256, 201–213.
14. Bate,P. and Warwicker,J. (2004) Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods. J. Mol. Biol., 340, 263–276.
15. Schultz, J., Milpetz, F., Bork, P. & Ponting, C.P. SMART, a simple modular architecture research tool: Identification of signaling domains. PNAS 1998; 95: 5857-5864
16. Waterhouse, A.M., Procter, J.B., Martin, D.M.A, Clamp, M. and Barton, G. J. (2009) "Jalview Version 2 - a multiple sequence alignment editor and analysis workbench" Bioinformatics25 (9) 1189-1191 doi: 10.1093/bioinformatics/btp033
17. "Scoredist: A simple and robust protein sequence distance estimator" Erik LL Sonnhammer and Volker Hollich BMC Bioinformatics 6:108 (2005)
18. Katoh, Standley 2013 (*Molecular Biology and Evolution* 30:772-780) MAFFT multiple sequence alignment software version 7: improvements in performance and usability.
19. Protein structure prediction on the web: a case study using the Phyre server. Kelley LA and Sternberg MJE. Nature Protocols 4, 363 - 371 (2009)
20. Kaushik S, et al., Improved detection of remote homologues using cascade PSI-BLAST: influence of neighboring protein families on sequence coverage. PLoS One. 2013;8(2):e56449. doi: 10.1371/ journal.pone.0056449. Epub 2013 Feb 20.
21. The Pfam protein families database: M. Punta, P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E.L.L. Sonnhammer, S.R. Eddy, A. Bateman, R.D. Finn Nucleic Acids Research (2012) Database Issue 40:D290-D301
22. Walldén K, et al., Structure of the VirB4 ATPase, alone and bound to the core complex of a type IV secretion system. http://www.pnas.org/content/early/2012/06/20/1201428109.full.pdf

Figures

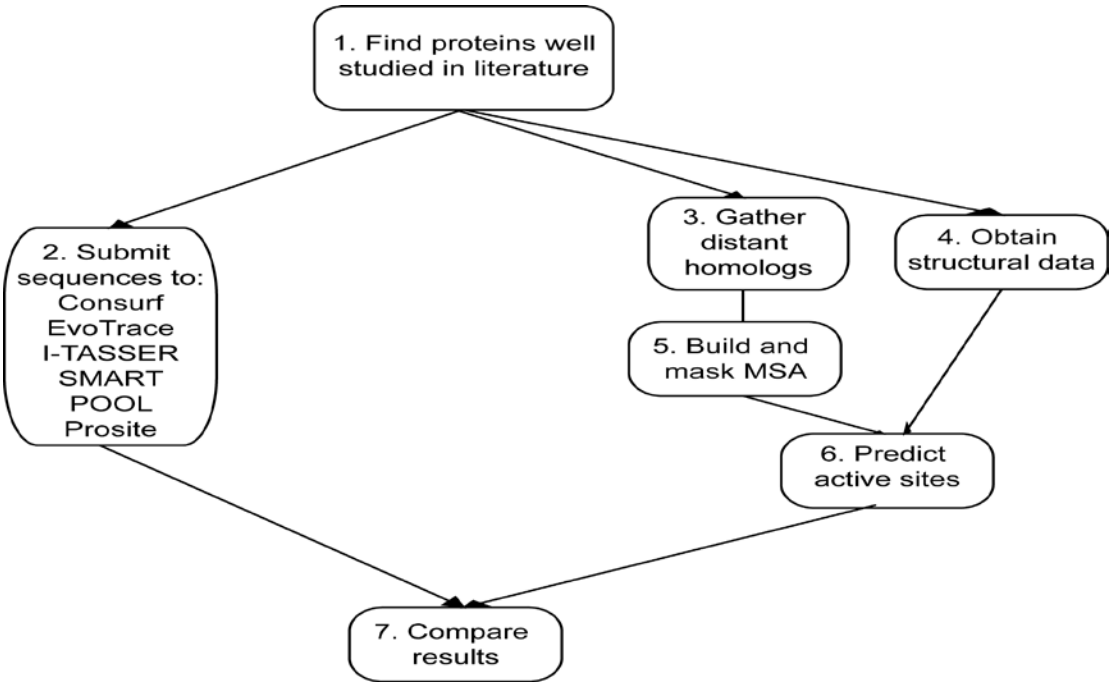

Figure 1. General pipeline of functional site prediction analysis. (1) Search literature for a diverse set of proteins whose functions have been experimentally confirmed, present in CSA or Swissprot. (2) Submit sequences to Consurf, EvoTrace, I-Tasser, SMART, POOL, and Prosite, compile results. (3) Gather distant homologs via PSI-BLAST. (4) Obtain structural data through prediction or crystal structure. (5) Construct and mask MSA. (6) Predict active sites by analyzing MSA and mapping conserved regions onto structure. (7) Compare results of residue predictions.

	H40	H57	D102	W141	G193	D194	S195	G196	S214
Consurf Exposed and Conserved	1	0	1	1	1	1	1	1	0
Pool	1	1	0	0	0	0	0	0	0
I-Tasser	0	1	1	0	1	0	1	1	0
Prosite	0	1	1	0	0	0	1	0	0
Evo-Trace	0	1	0	0	1	1	1	0	1
Smart	0	1	1	0	0	0	1	0	0
Sum of Webservers	2	5	4	1	3	2	5	2	1
Ramya's Prediction	0	1	1	1	0	1	1	0	1
Swissprot	0	1	1	0	0	0	1	0	0

Figure 2: Consolidated prediction results for residues of Bovine Chymotrypsin alpha. This study case's gold standard is shown in blue, and the sum of web servers is shown in pink. Manual prediction followed the protocol outlined in the methods section. Note that in this case study, both a webserver consensus approach, and manual prediction generate very accurate results.

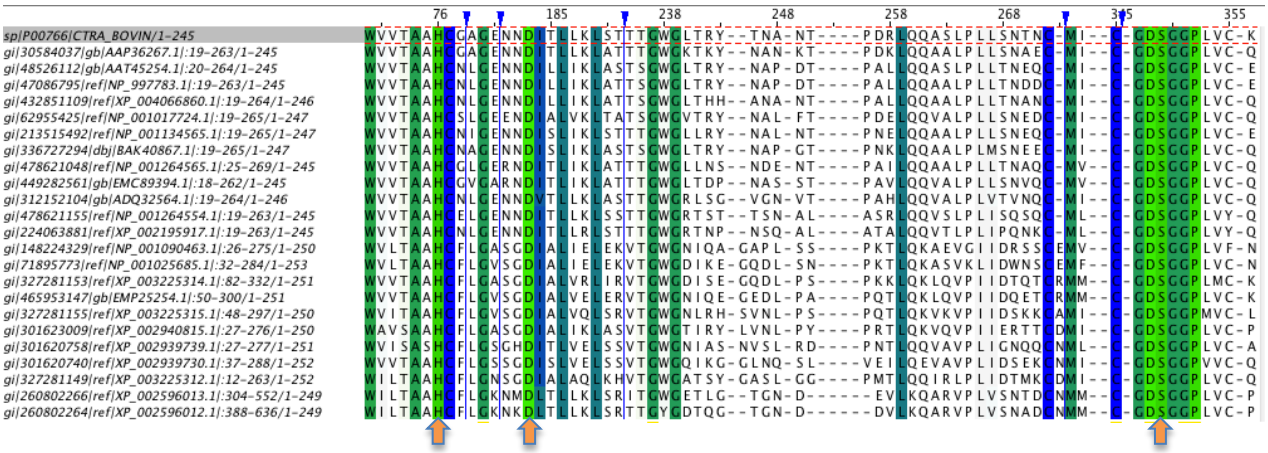

Figure 3. Excerpt of the Bovine Alpha Chymotrypsin MSA with several distant homologs is depicted above. The first protein listed and outlined in red is the query. The coloring scheme used is the Buried Index, indicating residues usually highly exposed in a protein structure in dark blue, and residues that are located in binding pockets or clefts in the interior of a structure in bright green. From this MSA, it is clear that cysteine residues are highly conserved, presumably for effective disulfide binding and structural integrity. Longer sequences of conserved, buried residues often indicate functionality, and this was observed in several places in the MSA, for example, the GDSGGP sequence. This segment of conserved residues was accurately predicted to contain a catalytic residue. Indeed, this MSA shows as highly conserved all three confirmed and predicted catalytic residues. They columns for these residues are indicated with orange arrows.

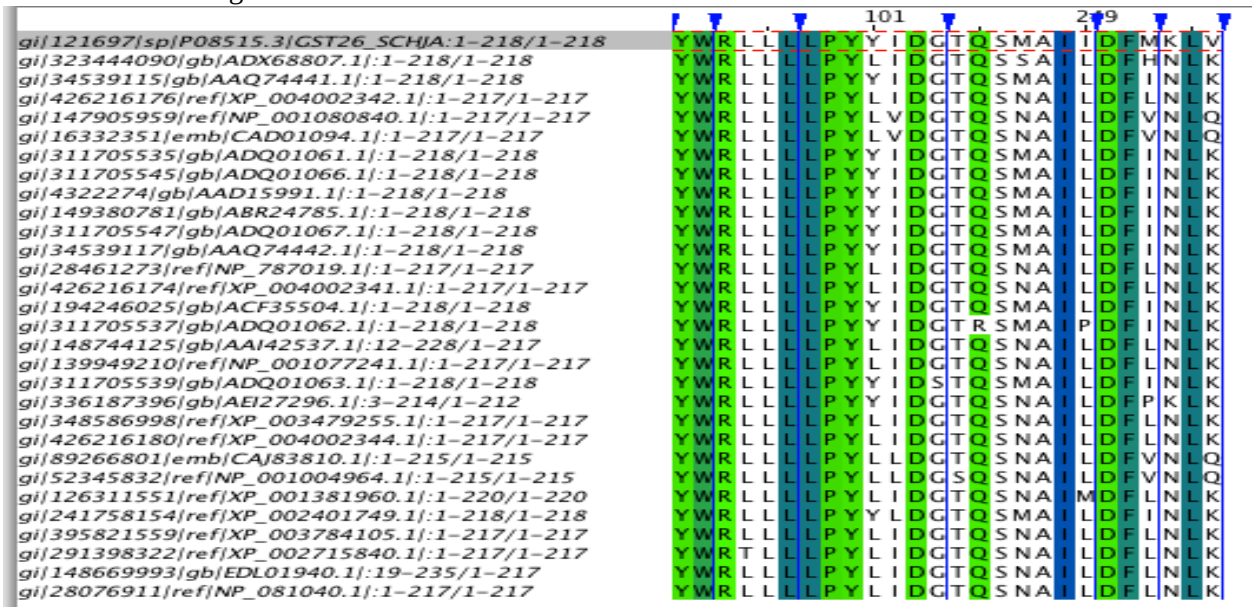

Figure 4. Section of Glutathione S-Transferase MSA. Query sequence appears the top, highlighted in gray. The MSA of about 300 sequences after masking was viewed using Jalvu, and the coloring scheme denotes the buried index of each column. Only the conserved columns in the MSA are shown (13 identified), with 98% sequence identity or above. The Y7 catalytic residue experimentally confirmed in the paper can be seen here to be highly conserved, and it was predicted after structural analysis to be catalytic.

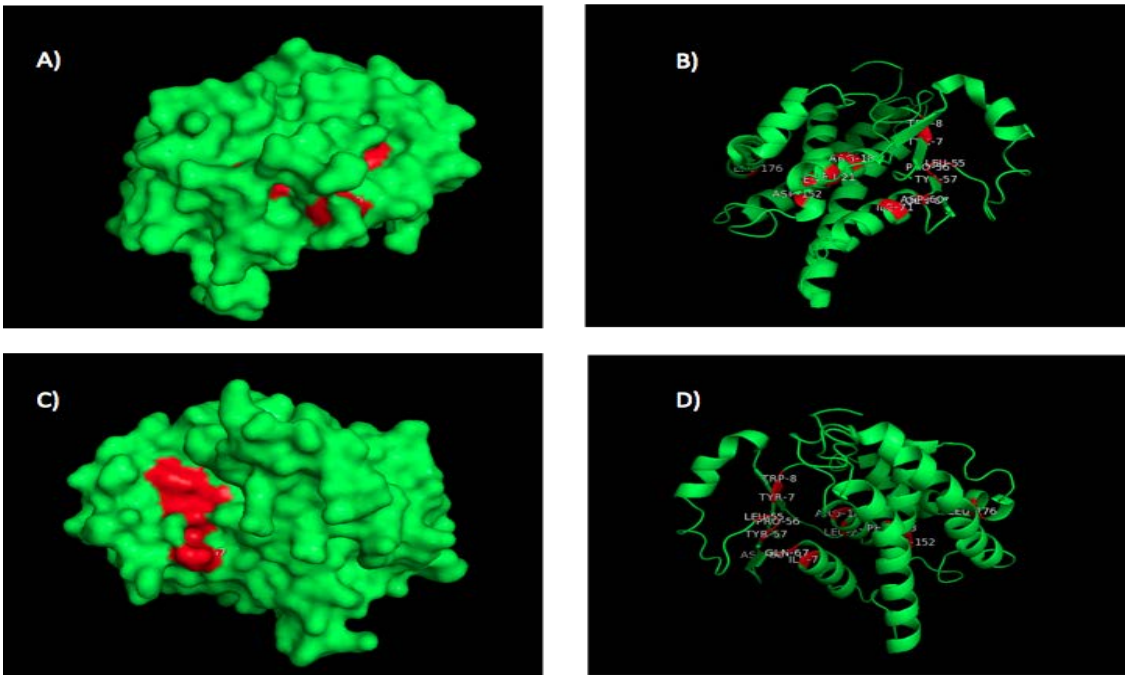

Figure 5. Surface and ribbon plots of Glutathione-S transferase. The red indicates the conserved residues extracted from the MSA. (A) Surface plot of the first cluster of exposed residues (R18, N60, I71). (B) Ribbon structure of figure A, shows that only some of the residues are solvent exposed and others are buried. (C) Surface plot of the second cluster of exposed residues (Y7, W8, L55, P56, Y57, Q67). (D) Ribbon structure of figure C.

	Y7	W8	I10	K11	G12	L13	Q15	R18	L21	H31	Y33	W41	K45	P53	N54	L55	P56
ConSurf	✓	✓		✓				✓	✓			✓	✓	✓	✓	✓	✓
POOL	✓		✓	✓	✓	✓	✓	✓		✓	✓						✓
I-TASSER	✓	✓			✓	✓						✓	✓		✓	✓	
PredictProtein (Prosite)										✓	✓						
Evolutionary Trace	✓	✓		✓	✓		✓	✓	✓					✓	✓	✓	✓
Swiss-Prot	✓	✓										✓	✓		✓	✓	
SMART																	
Sum of Webservers	5	4	1	3	3	2	2	3	2	2	2	3	3	2	4	4	3
MSA prediction	✓	✓						✓	✓								✓
Cardoso et al (2003)	x	x	x		x	x						x	x		x	x	

	Y57	G61	Q67	S68	I71	D101	R103	S107	Y111	H150	D152	F153	Y156	D157	S195	Q204	Q207
ConSurf	✓	✓	✓	✓	✓	✓					✓	✓	✓	✓	✓		
POOL	✓						✓			✓			✓	✓		✓	
I-TASSER			✓	✓			✓	✓	✓							✓	
PredictProtein (Prosite)										✓	✓				✓		
Evolutionary Trace	✓	✓	✓	✓	✓						✓	✓		✓			
Swiss-Prot			✓	✓					✓								
SMART																	
Sum of Webservers	3	2	4	4	2	1	2	1	2	2	3	2	2	3	2	2	0
MSA prediction	✓		✓														
Cardoso et al (2003)			x	x		x	x	x	x							x	x

Figure 6. A comparison of webserver queries for functional site information on glutathione s-transferase. The sum total of webserver hits for residues are listed in blue. Experimentally identified functional residues from Cardoso et al (2003) are listed in gold.

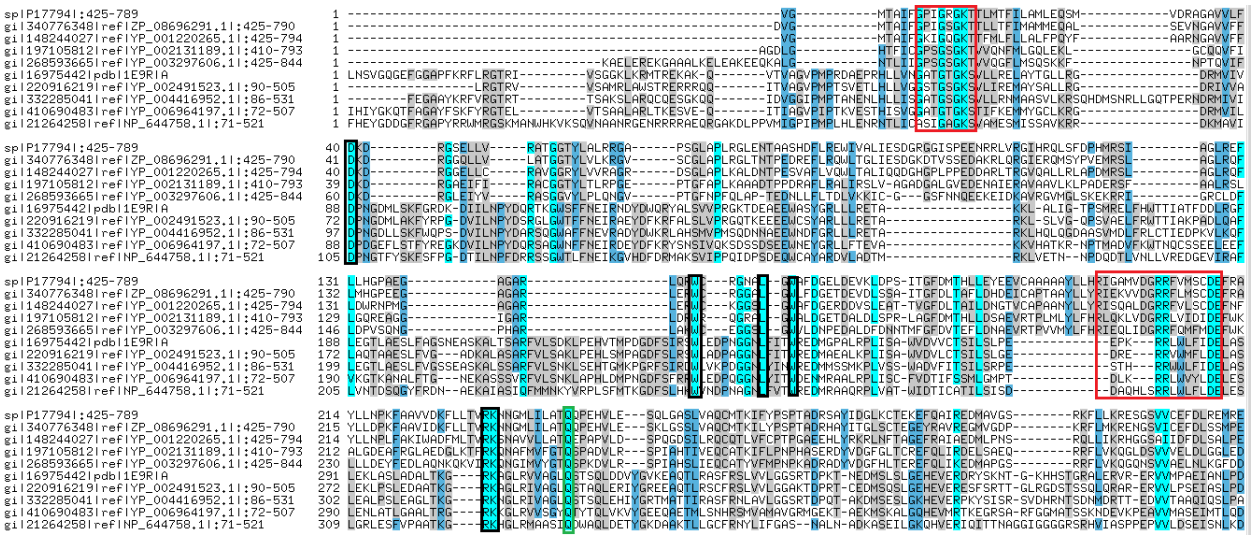

Figure 7. Multiple Sequence Alignment for Protein # 3: VirB4. MSA is in wrap-around format, visualized on Belvu. Boxed in red are the Walker A and Walker B conserved motifs. Boxed in green is the conserved glutamine (Q668). These are the same motifs found in the Middleton et al paper. Additional highly conserved domains found in the MSA are boxed in black (D465, W573, L578, W580, R657, K658).

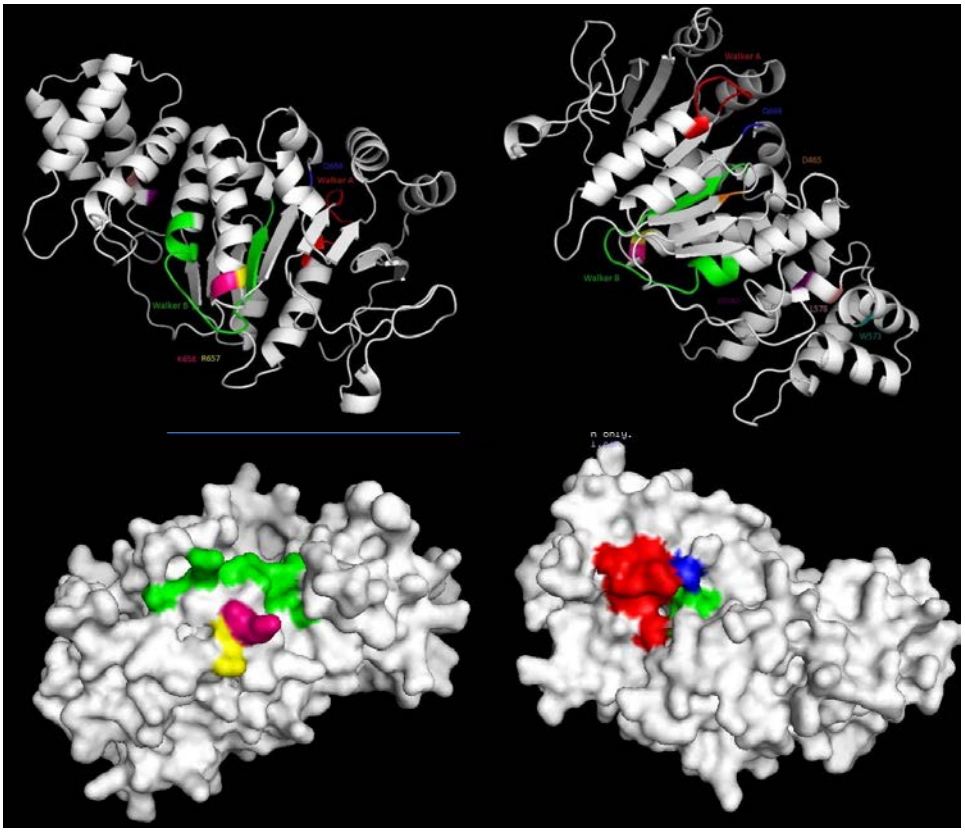

Figure 8. 3D Comparative Model for VirB4 with highlighted residues. The ribbon structure has color-coded residues highlighted and labeled. As seen in the surface plot, however, the D465, W73, L578 and W580 are not very solvent exposed, or present on loops or clefts, as is common for active site residues. The pink and yellow residues are the arginine finger and the red, green, and blue are the Walker A, Walker B, and conserved glutamine active sites.