BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 11 2014, pages 1595–1600
doi:10.1093/bioinformatics/btu073

Databases and ontologies

Advance Access publication February 3, 2014

AMASS: a database for investigating protein structures

Clinton J. Mielke[1,2,3,*], Lawrence J. Mandarino[2] and Valentin Dinu[3]

[1]Biodesign Institute, Arizona State University, Tempe, AZ 85287, USA, [2]The Center for Metabolic and Vascular Biology, Mayo Clinic, Scottsdale, AZ 85259, USA and [3]Department of Biomedical Informatics, Arizona State University, Scottsdale, AZ 85259, USA

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Modern techniques have produced many sequence annotation databases and protein structure portals, but these Web resources are rarely integrated in ways that permit straightforward exploration of protein functional residues and their co-localization.

Results: We have created the AMASS database, which maps 1D sequence annotation databases to 3D protein structures with an intuitive visualization interface. Our platform also provides an analysis service that screens mass spectrometry sequence data for post-translational modifications that reside in functionally relevant locations within protein structures. The system is built on the premise that functional residues such as active sites, cancer mutations and post-translational modifications within proteins may co-localize and share common functions.

Availability and implementation: AMASS database is implemented with Biopython and Apache as a freely available Web server at amass-db.org.

Contact: clinton.mielke@gmail.com

Received on August 21, 2013; revised on January 17, 2014; accepted on January 29, 2014

1 INTRODUCTION

Our ability to read the 'source code' of life has produced large databases of biological sequences, and analysis of sequence data has given much insight into how genes and proteins work. Conservation at the nucleotide or residue level has unveiled the active sites in enzymes and important binding motifs on proteins. Sequencing of tumors has also unveiled specific site mutations that affect the functions of important signaling molecules. Complementing the advances in gene sequencing, mass spectrometry has enabled routine and rapid protein sequencing. This technique enables the high-throughput discovery of protein post-translational modifications (PTMs) such as phosphorylation and acetylation, which can dramatically alter biological function in ways not easily predictable from the genome sequence. These PTMs provide yet another layer of complexity that influences biophysical mechanisms at single residues. These residue-specific annotations will continue to grow in several bioinformatics databases (UniProt, 2012)

Alongside sequence informatics, the field of structural bioinformatics is well established, with projects such as the Protein Data Bank (PDB; Bernstein *et al*., 1978) providing researchers with a large collection of determined protein structures. New projects such as the Protein Structure Initiative (Burley *et al*., 2008) are dramatically increasing the coverage of this dataset by developing high-throughput crystallography techniques. As the number of crystallized proteins has increased, an opportunity has arisen to unify sequence annotation data with protein structure data to gain additional insights into the function of proteins.

Multiple Web portals have been designed that offer analysis and visualization of specific functional residues in proteins. Much of this work focuses on predicting the possible function of non-synonymous single-nucleotide polymorphisms (SNPs). These online databases include SNPs3D (Yue *et al*., 2006), SNPeffect (De Baets *et al*., 2012), PolyPhen (Adzhubei *et al*., 2013), MSV3d (Luu *et al*., 2012) and MutDB (Mooney and Altman, 2003). These databases focus exclusively on SNPs, but recently new databases have been created to address the new need of characterizing the functions of PTMs. Phospho3D (Zanzoni *et al*., 2007) curates phosphosites from the Phospho.ELM (Diella *et al*., 2008) database and provides Jmol (Jmol: an open-source Java viewer for chemical structures in 3D. www.jmol.org) visualizations of these sites on individual pages. Another recent project is PTMfunc, which extracted nearly 200 000 phosphorylation, acetylation and ubiquitinylation sites from several data sources (Beltrao *et al*., 2012). They established an impressive analysis pipeline that attempts to predict the function of these modifications; however, their database does not permit visualization or query of novel sites.

While there is clearly a plethora of structure mapping Web resources, they offer a heterogeneous set of features. These databases solely focus on either variants or modifications, but we believe that unification of these annotations may allow researchers to gain additional insights. Some of these tools offer visualizations of single sites on proteins, but rarely allow exploration of multiple arbitrary residues at once. We believe that a better tool would enable the interactive exploration of any set of residues simultaneously. This would enable investigators to find co-localized features across annotation databases that may share allosteric functions. Such a service would allow users to guess at the 'big picture' of how a protein functions.

We have developed AMASS, a free resource that brings together multiple databases: the UniProt KnowledgeBase (UniProt, 2012), the Catalogue of Somatic Mutations in Cancer (COSMIC) (Forbes *et al*., 2011) and the PhosphoSite database (Hornbeck *et al*., 2012) of PTMs. Alongside this aggregation of these sequence-level annotations, our system matches searched proteins with structures from the PDB. Structures are

[*]To whom correspondence should be addressed.

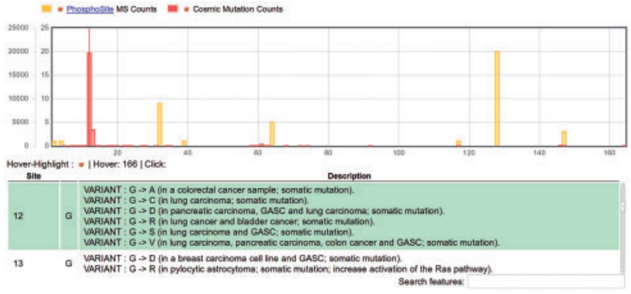
visualized with the Jmol viewer, and sequence annotations can be interactively explored in 3D by clicking links within the interface. With this system, users can explore the annotations and physical location of any arbitrary residue. The system also permits the upload of mass spectrometry data to find possible functional roles of thousands of queried sites by performing a search of annotated neighboring residues. Because our resource 'amasses' together multiple bioinformatics databases and is rooted in the analysis of mass spectrometry data, we have named our resource 'AMASS', and it can be accessed at amass-db.org.

2 FEATURES

The AMASS server allows visitors to both (i) *browse* for proteins to investigate sequence–structure relationships and (ii) to *analyze* modification data derived from mass spectrometry–based experiments. Both usage modes provide similar interfaces.

For each protein viewed on the site, we present a table of residue-level annotations derived from UniProt (Fig. 1). These annotations can include active sites, clinically relevant variants, ligand-binding residues or sites that have been studied in prior mutagenesis experiments. Additionally, we present an interactive bar chart that displays count data derived from two sources. Somatic mutation frequencies in tumor samples are displayed from the COSMIC database. We also present post-translational modification count data derived from the PhosphoSite database, which aggregates modification count data based on prior mass spec datasets and literature publications. This count data serves to complement the textual annotations that we provide. The counts plot and annotations table are both linked, so that exploration of either with the mouse causes automatic highlighting to occur in the other.

A major feature of AMASS is its mechanism for unifying sequence annotation data with protein structures from the PDB. Even if the queried protein lacks an exact match, AMASS automatically performs sequence alignments against homologous structures and displays them with an embedded Jmol applet. Users can interactively explore the structure by clicking on sequence annotations or on peaks in the counts plot. These actions highlight residues in the sequence alignment, showing the user how the mappings are made between the queried protein and loaded structure. Sites are also instantly zoomed in on the model, showing where they are located in 3D. (Fig. 2) The structure viewing panel also contains a contacts widget that uses Jmol to display neighboring residues and ligands that form van der Waals contacts with the most recently selected residue.


🟨 ■ PhosphoSite MS Counts
🟥 ■ Cosmic Mutation Counts
25000
20000
15000
10000
5000
0
25
20
15
10
5
0
20
40
60
80
100
120
140
160
Hover-Highlight : ■ | Hover: 166 | Click:
Search features:

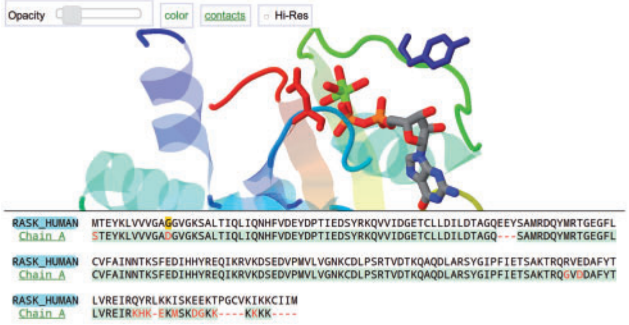
Site		Description
12	G	VARIANT : G -> A (in a colorectal cancer sample; somatic mutation). VARIANT : G -> C (in lung carcinoma; somatic mutation). VARIANT : G -> D (in pancreatic carcinoma, GASC and lung carcinoma; somatic mutation). VARIANT : G -> R (in lung cancer and bladder cancer; somatic mutation). VARIANT : G -> S (in lung carcinoma and GASC; somatic mutation). VARIANT : G -> V (in lung carcinoma, pancreatic carcinoma, colon cancer and GASC; somatic mutation).
13	G	VARIANT : G -> D (in a breast carcinoma cell line and GASC; somatic mutation). VARIANT : G -> R (in pylocytic astrocytoma; somatic mutation; increase activation of the Ras pathway).

Fig. 1. The counts plot in the top panel shows modifications from the PhosphoSite database (yellow/light) and cancer mutations (red/dark) from the COSMIC database. The lower table shows annotations from UniProt that describe the functions of specific residues. Residue G12 (red/dark peak) is a prominent hotspot mutation present in nearly 20 000 tumor samples from the COSMIC data

As an example, we use AMASS to search for human KRAS, a well-studied GTPase mutated in human cancers. Figure 1 illustrates how AMASS represents 1D sequence annotations and counts, and Figure 2 shows a paired protein structure, complete with a visualization using Jmol and a displayed sequence alignment. A common cancer mutation (G12) and commonly observed phosphotyrosine (Y32) are observed near the GTP binding site. Visualizations such as this permit investigators to hypothesize common functional roles between the two residues.

2.1 Mass spectrometry analysis

To upload a dataset for analysis, the input is a list of capitalized peptide or protein sequences with queried sites denoted as lowercase letters. This set of sequences is matched against sequences from the UniProt database. Matching proteins are presented in a table (Fig. 3), with added columns indicating features in the

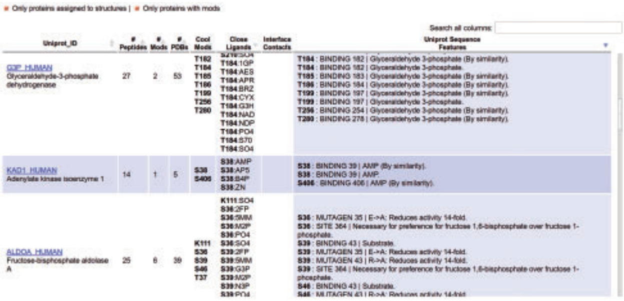


```
RASK_HUMAN  MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFL
 Chain A    STEYKLVVVGADGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQ---SAMRDQYMRTGEGFL

RASK_HUMAN  CVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDLPSRTVDTKQAQDLARSYGIPFIETSAKTRQRVEDAFYT
 Chain A    CVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDLPSRTVDTKQAQDLARSYGIPFIETSAKTRQGVDDAFYT

RASK_HUMAN  LVREIRQYRLKKISKEEKTPGCVKIKKCIIM
 Chain A    LVREIRKHK-EKMSKDGKK----KKKK----
```

Fig. 2. To explore mutations and modifications of human KRAS, a structure from the PDB (4dsu) is loaded and aligned to the UniProt KRAS sequence. Residues of interest in the counts plot (Fig. 1) were clicked to highlight them in the structure. The prominent mutation peak at G12 is associated to a mutated aspartic acid in the structure and alignment (red stick model, left). The phosphorylation peak at Y32 corresponds to a tyrosine (blue stick model, top) that is near the GDP binding site



■ Only proteins assigned to structures | ■ Only proteins with mods

Search all columns:

Uniprot_ID	# Peptides	# Mods	# PDBs	Cool Mods	Close Ligands	Interface Contacts	Uniprot Sequence Features
G3P_HUMAN Glyceraldehyde-3-phosphate dehydrogenase	27	2	53	T182 T184 T185 T186 T199 T256 T280	S210:SO4 T184:1GP T184:AES T184:APR T184:BRZ T184:CYX T184:G3H T184:NAD T184:NDP T184:PO4 T184:S70 T184:SO4		T184 : BINDING 182 | Glyceraldehyde 3-phosphate (By similarity). T184 : BINDING 182 | Glyceraldehyde 3-phosphate. T185 : BINDING 183 | Glyceraldehyde 3-phosphate (By similarity). T186 : BINDING 184 | Glyceraldehyde 3-phosphate (By similarity). T199 : BINDING 197 | Glyceraldehyde 3-phosphate (By similarity). T199 : BINDING 197 | Glyceraldehyde 3-phosphate. T256 : BINDING 254 | Glyceraldehyde 3-phosphate (By similarity). T280 : BINDING 278 | Glyceraldehyde 3-phosphate (By similarity).
KAD1_HUMAN Adenylate kinase isoenzyme 1	14	1	5	S38 S406	S38:AMP S38:AP5 S38:B4P S38:ZN		S38 : BINDING 39 | AMP (By similarity). S38 : BINDING 39 | AMP. S406 : BINDING 406 | AMP (By similarity).
ALDOA_HUMAN Fructose-bisphosphate aldolase A	25	6	39	K111 S36 S39 S46 T37	K111:SO4 S36:2FP S36:5MM S36:M2P S36:PO4 S36:SO4 S39:2FP S39:5MM S39:G3P S39:M2P S39:N3P S39:PO4		S36 : MUTAGEN 35 | E->A: Reduces activity 14-fold. S36 : SITE 364 | Necessary for preference for fructose 1,6-bisphosphate over fructose 1-phosphate. S39 : BINDING 43 | Substrate. S39 : MUTAGEN 35 | E->A: Reduces activity 14-fold. S39 : MUTAGEN 43 | R->A: Reduces activity 14-fold. S39 : SITE 364 | Necessary for preference for fructose 1,6-bisphosphate over fructose 1-phosphate. S46 : BINDING 43 | Substrate. S46 : MUTAGEN 43 | R->A: Reduces activity 14-fold.

Fig. 3. The results of a mass spectrometry dataset analysis. Several proteins are found with interesting modifications, located near important annotated residues

protein structure that are proximal to the queried modification sites. To illustrate the utility of AMASS, we have analyzed a prior mass spectrometry dataset collected from human skeletal muscle (Lefort *et al.*, 2009). AMASS identified several proteins with potentially interesting modifications near active sites.

2.2 Pyruvate dehydrogenase

The pyruvate dehydrogenase (PDH) complex is a central metabolic enzyme that connects glycolysis with the citric acid cycle by converting pyruvate to acetyl-coA. Our mass spectrometry data show that serines 293 and 295 of E1 subunit alpha (UniProt: ODPA_HUMAN) are phosphorylated *in vivo*. Both of these serines are documented in the PhosphoSite database, with S293 in particular having 625 references to external mass spectrometry datasets. Furthermore, S293 is the most studied mutagenesis site, with 13 references documented in PhosphoSite. Most interestingly, these two phosphorylations are located in a cluster of six phosphosites present in a short span of the protein sequence (Fig. 4). Two of these sites (S293 and S300) are annotated in UniProt as targets of pyruvate dehydrogenase kinase 4 (PDK4), which is a kinase that regulates the activity of PDH (Kato *et al.*, 2008).

The protein has seven crystal structures in the PDB, and in one of these structures (3exh) AMASS reported that S293 is proximal to an interesting ligand. On visualizing the alignments, we note that the crystal structure was prepared from the mature enzyme that has a 29-residue mitochondrial transit peptide removed. The phosphosite S293 in the full-length UniProt sequence aligns with S264 in the mature enzyme. Despite this discrepancy, the sequence alignment procedure of AMASS ensured that the correct full-length phosphosites were correctly associated to the chains in the structure. The crystal structure has four chains of this protein, and this residue is actually found phosphorylated in one of these chains. This structure was prepared by phosphorylating PDH with PDK4, and the heterogeneous phosphorylation of the four chains permits comparison of secondary structure.

AMASS reported that in chain C in particular, this phosphoserine is located close to the ligand thiamine pyrophosphate (TPP), a derivative of thiamine (vitamin B1), and a required cofactor of PDH (Fig. 5). Furthermore, we readily observe that the other phosphorylated serine S295 (S266 in the mature form) is also repositioned in the chains where S264 is phosphorylated. It is discussed in the corresponding publication of this crystal model (Kato *et al.*, 2008) that phosphorylation of the S293 residue (via PDK4) causes derangements of this phosphorylation loop region and prevents substrate channeling to the TPP cofactor, thus inactivating the entire PDH complex. In this case, S293 is well studied, yet the annotations within UniProt only discuss PDK4 as the responsible kinase but do not assign a strong functional role of S293 in regulating metabolic activity. Nonetheless, AMASS reported the proximity to TPP, which enabled us to uncover the literature citation associated with this PDB structure, and ultimately provided us with insight into our data. This emphasizes the potential of AMASS to serve as a tool to integrate data and help the research community improve functional annotations of sites.

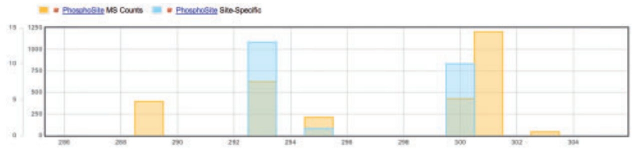


Fig. 4. A cluster of phosphorylation sites in the PDH sequence, documented in the PhosphoSite database. Yellow bars denote phosphosites detected in mass spectrometry experiments. Blue bars represent publication counts that use techniques such as site-directed mutagenesis

2.3 ATP synthase beta

In human skeletal muscle, adenosine triphosphate (ATP) synthase is an important enzyme that provides energy for the cell to use through the synthesis of ATP from adenosine diphosphate (ADP). Peptides from this protein are particularly abundant from skeletal muscle samples. On the beta subunit of this complex, AMASS identified a phosphothreonine modification (T213) that is located near interesting ligands in several structures in the PDB. Human ATP synthase beta has not been crystallized, but several models exist of the bovine (97% sequence identity) and the yeast (78% identity) proteins. AMASS automatically aligned these homologous structures and found that phosphorylated T213 is located nearby magnesium ions and either ATP or ADP molecules. This site was previously investigated in relation to insulin-resistant skeletal muscle, and deduced to likely be located near the ATP binding domain based on sequence analysis (Hojlund *et al.*, 2003).

Using AMASS to directly visualize the ATP synthase beta protein (PDB 4asu), we observe that three beta subunits and three alpha subunits complex to form the F1 subunit. Each of the three beta subunits has an ADP molecule present at the active site. Additionally, magnesium cofactor ions are seen bound to the terminal phosphates of the ADP molecules. The T213 side chain is positioned such that direct contact is made with the magnesium ions (Fig. 6). This finding is interesting, as this phosphorylation of T213 could thus be an autocatalytic mechanism due to the proximity of the ADP phosphates, or perhaps an unknown reaction intermediate. We hypothesize that phosphorylation of this threonine may affect catalytic

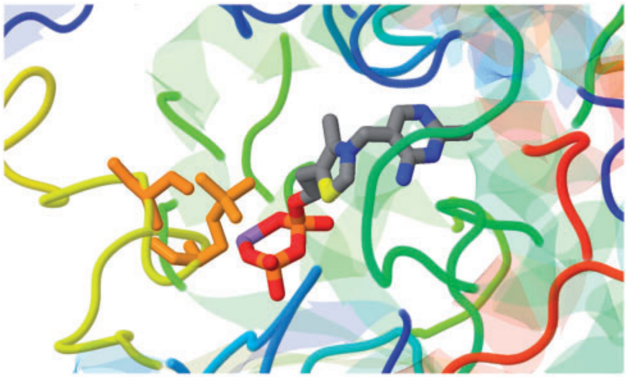
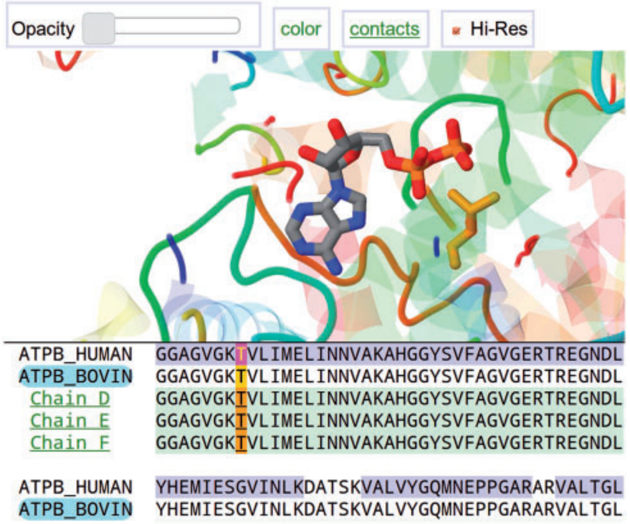


Fig. 5. A serine and phosphoserine residue (orange) are located on a loop region in the PDH crystal structure. This loop is proximal to the TPP ligand, a cofactor of PDH



```
ATPB_HUMAN   GGAGVGKTVLIMELINNVAKAHGGYSVFAGVGERTREGNDL
ATPB_BOVIN   GGAGVGKTVLIMELINNVAKAHGGYSVFAGVGERTREGNDL
 Chain D     GGAGVGKTVLIMELINNVAKAHGGYSVFAGVGERTREGNDL
 Chain E     GGAGVGKTVLIMELINNVAKAHGGYSVFAGVGERTREGNDL
 Chain F     GGAGVGKTVLIMELINNVAKAHGGYSVFAGVGERTREGNDL

ATPB_HUMAN   YHEMIESGVINLKDATSKVALVYGQMNEPPGARARVALTGL
ATPB_BOVIN   YHEMIESGVINLKDATSKVALVYGQMNEPPGARARVALTGL
```

Fig. 6. Bovine ATP synthase with bound ADP and T213 (orange) that binds the magnesium cofactor (green). Below: Multiple sequence alignment showing mass spec coverage (purple) and alignment between human and bovine ATP synthase beta, and alignments to chains in PDB

function by inhibiting binding of either the ADP substrate or the magnesium cofactor.

3 IMPLEMENTATION

Development of AMASS required substantial database preparation and alignment computations. Most of this effort was performed ahead of time to allow rapid query and analysis for end users. We used the UniProt database as the core foundation of our platform. Our pipeline downloads the monthly compressed flat-file releases of Swiss-Prot and TrEMBL and parses them with the Biopython library. A set of python scripts performs the necessary alignments and processing, which altogether take 1 week.

We maintain a mirrored copy of the Worldwide Protein Data Bank (wwPDB) on our server using rsync. The PDB has seen exponential growth over the past decade, but there has been little concerted effort to correlate deposited structures with protein sequence identifiers. Fortunately, the SIFTS consortium (Velankar *et al.*, 2013) has recently established a unified database of UniProt sequence mappings to individual chains within PDB files. This database provided us a reliable source of sequence pairings to which we could perform our own sequence alignments. At the time of writing this article, the SIFTS database links ~28 000 UniProt sequences to ~81 000 structures in the PDB.

We used the Prody (Bakan *et al.*, 2011) library to parse and analyze each PDB file. For each polypeptide chain, sequences of residue names and residue IDs were extracted. Each chain sequence was then aligned to its associated UniProt sequence using Clustal Omega (Sievers *et al.*, 2011) to determine the residue-level mappings. These 'chain alignments' were stored as FASTA files to disk. The residue IDs for each PDB chain were stored into a Python array and serialized to an object-oriented database for rapid retrieval. With both the computed 'chain alignments' and these residue ID mapping arrays we were able to associate any UniProt residue to residue IDs within PDB chains.

We also extracted topological data from each structure. For every non-water residue, the ProDy selection engine was used to find other non-water residues within a 7 Å distance. These protein *zones* were stored in an object-oriented database to allow rapid lookup for later analysis efforts. This approach was largely inspired by the approach taken by Phospho3D in which 3D *zones* were calculated for all phosphosites. Our system performed this neighborhood search for all residues and ligands.

Despite the large number of structures deposited in the PDB, many human proteins have not been crystallized. Fortunately, however, highly similar homologues can be found, either as human paralogs or orthologs in other species. To expand the coverage of our platform, we designed it to pair queried proteins with homologous proteins in the PDB. Similar Web resources, such as MSV3d, instead produce homology models of uncrystallized proteins. This approach is less transparent and more computationally demanding. Our approach is to instead find similar structures and provide sequence alignments to end users, permitting inspection of the similarity between the queried protein and the visualized structure. Inferences made at the single residue level can be sanity-checked based on the local sequence conservation.

This mechanism of pairing queried proteins to homologous structures dramatically increases the utility of the system. For human proteins, for example, the SIFTS database matches ~4600 human UniProt proteins to a structure in the PDB. Matching to homologous sequences with 50% identity, however, ~8800 human proteins can be mapped and visualized to a structure.

To accomplish this sequence–structure pairing, we used all structure sequences documented in the SIFTS database as a query in an all versus all BLAST (Altschul *et al.*, 1997) search against the Swiss-Prot database. Sequences that had at minimum of a 50% match were stored in a relational database. This cutoff was chosen due to the steep rise of non-relevant paired proteins that were observed at lower overlaps. Subsequently, we precomputed all sequence alignments between the SIFTS sequences and matched Swiss-Prot sequences using Clustal Omega. (Sievers *et al.*, 2011) These 'homology alignments' were stored to disk.

To map a queried protein sequence to a homologous structure, the system merges the relevant 'homology alignment' with the 'chain alignment' for the structure. The system joins these two alignments based on the common UniProt sequence defined by the SIFTS database. This merging algorithm is implemented in Python, where sequences are converted to linked-list data structures for efficient manipulation, and then gaps are inserted into each alignment while scanning the common UniProt sequence in each. Once this merged alignment is produced between a queried protein and a homologous protein structure, we can map sequence positions on the query protein to the protein sequence in the PDB structure, and, then, onto single residues in each chain. To perform this mapping rapidly, we implemented a simple data structure in Python that represents the alignment as a set of Python integer arrays. These arrays allow efficient random access mapping between any two sequences.

While our site allows visualization of any protein of interest, AMASS also provides an analysis backend to scrutinize single modifications in submitted peptide sequences, e.g. derived from mass spectrometry experiments. The input format follows a standard in the mass spectrometry field, in which peptide sequences are capitalized and detected modifications are denoted with lowercase letters. The user provides an email address for notification when the job is completed. After submitting the data, the system checks for common errors such as non-amino acid characters, and, then, de-duplicates non-unique peptide sequences while keeping an internal count of coverage and modifications. We encourage users to upload raw peptide data, wherein the same peptide sequence is listed many times. This 'spectral count' data are reported back to the user for each and every residue in unmodified and modified counts.

Given the list of unique peptide sequences, our system finds the set of matching UniProt proteins from the species specified by the user. This search is efficiently implemented with the UNIX fgrep program. After this set of target proteins is found, peptide sequences are aligned. Peptide coverage and modification counts are calculated and stored into a database for long-term aggregation. After validation, the backend analysis pipeline finds similar PDB structures and merges alignments as described earlier. Modifications are then mapped to 3D coordinates, and a neighborhood of structural features is found. Within the neighborhood residue set, our algorithm looks for residues that have important sequence features such as active sites, mutagenesis sites or binding residues as annotated within UniProt. Our system also determines whether the site is located at the binding interface between two distinct proteins in a complex. Finally, the system checks nearby to look for interesting ligands bound to the protein.

4 DISCUSSION

Our platform enables the exploration of proteins in an intuitive manner not possible with existing tools. We hope that AMASS will grow into a valuable resource for the biomedical research community and we shall thus have a substantial roadmap of desired features planned for future releases. We plan to integrate many more biological databases to bring together more functional annotations. For example, the COSMIC database has given us an excellent jumpstart in providing mutational frequencies of single amino acids; however, the recent release of full tumor sequencing data from The Cancer Genome Atlas project (Collins and Barker, 2007) hints that much more mutation data can be integrated in the coming years as sequencing becomes cheaper.

Our analysis system is simplistic, as it only considers proximity of co-localized features to infer significance. Our primary focus for this system was to create an interactive visualization, as we felt that such visualization was not adequately implemented in existing platforms. Nonetheless, existing platforms perform far more sophisticated analysis of the structural consequences of modified sites, typically for SNPs. For example, PolyPhen and SNPs3D both offer computerized classification of whether amino acid changes affect function based on changing biophysical properties such as side-chain hydrophobicity, charge or packing. Adding this deeper level of analysis in the future would improve the utility of our tool. The challenge for post-translational modifications will be to find appropriate algorithms to assess functional consequences of these novel side chains.

We have observed that many proteins in the UniProt database lack comprehensive functional annotations, especially when proteins from less-studied species are queried. Furthermore, highly valuable feature records such as mutagenesis sites tend to be species-specific because those annotations arise from academic publications that focused in a single model organism. We plan to address this issue in the future by integrating a homology clustering system into AMASS, whereby all homologous proteins (and their associated residue annotations) can be visualized in alignments and simultaneously projected onto 3D structures. This improvement poses considerable user interface design challenges.

Many protein structures within the PDB have multiple proteins in complex; however, several distinct structures can be structurally aligned on common shared protein sequences to elucidate other possible complexes that may exist *in vivo*. For example, a structure containing proteins A and B can be structurally aligned with a separate structure consisting of proteins B and C. Such an alignment may hint how proteins A and C can interact *in vivo*. Performing such alignments across all structures may expand the set of possible neighboring features that can be discovered. Furthermore, adopting a technique similar to PTMfunc (Beltrao *et al*., 2012), whereby conserved domains among many distinct proteins are aligned, may allow greater generalization to be made for the functions of single sites.

With the development of AMASS, we have a system that is capable of quickly finding functional neighbors of any queried sites. This analysis, however, is only performed on submission of data, or when the user explores a particular residue through the Web interface. In the future, we plan to implement algorithms that will automatically find clusters of functional residues in each structure. The challenge here is presenting the results in a format that scores and ranks identified clusters by relevance.

Given our developed system and planned features, we envision that AMASS will serve as a valuable tool to the research community at large by both enabling the understanding of datasets and the discovery and annotation of unrecognized functional sites of proteins.

Funding: This work was supported by the National Institutes of Health [R01DK047936 to L.J.M.], Arizona State University, Mayo Clinic, and the Virginia G. Piper Charitable Trust (V.D.).

Conflicts of Interest: none declared.

REFERENCES

Adzhubei,I. *et al.* (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, **Chapter 7**, Unit7 20.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bakan,A. *et al.* (2011) ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics*, **27**, 1575–1577.

Beltrao,P. *et al.* (2012) Systematic functional prioritization of protein posttranslational modifications. *Cell*, **150**, 413–425.

Bernstein,F.C. *et al.* (1978) The Protein Data Bank: a computer-based archival file for macromolecular structures. *Arch. Biochem. Biophys.*, **185**, 584–591.

Burley,S.K. *et al.* (2008) Contributions to the NIH-NIGMS protein structure initiative from the PSI production centers. *Structure*, **16**, 5–11.

Collins,F.S. and Barker,A.D. (2007) Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Sci. Am.*, **296**, 50–57.

De Baets,G. *et al.* (2012) SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res.*, **40**, D935–D939.

Diella,F. *et al.* (2008) Phospho.ELM: a database of phosphorylation sites–update 2008. *Nucleic Acids Res.*, **36**, D240–D244.

Forbes,S.A. *et al.* (2011) COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **39**, D945–D950.

Hojlund,K. *et al.* (2003) Proteome analysis reveals phosphorylation of ATP synthase beta -subunit in human skeletal muscle and proteins with potential roles in type 2 diabetes. *J. Biol. Chem.*, **278**, 10436–10442.

Hornbeck,P.V. *et al.* (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.*, **40**, D261–D270.

Kato,M. *et al.* (2008) Structural basis for inactivation of the human pyruvate dehydrogenase complex by phosphorylation: role of disordered phosphorylation loops. *Structure*, **16**, 1849–1859.

Lefort,N. *et al.* (2009) Proteome profile of functional mitochondria from human skeletal muscle using one-dimensional gel electrophoresis and HPLC-ESI-MS/MS. *J. Proteomics*, **72**, 1046–1060.

Luu,T.D. *et al.* (2012) MSV3d: database of human MisSense variants mapped to 3D protein structure. *Database (Oxford)*, **2012**, bas018.

Mooney,S.D. and Altman,R.B. (2003) MutDB: annotating human variation with functionally relevant data. *Bioinformatics*, **19**, 1858–1860.

Sievers,F. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

UniProt,C. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.

Velankar,S. *et al.* (2013) SIFTS: structure integration with function, taxonomy and sequences resource. *Nucleic Acids Res.*, **41**, D483–D489.

Yue,P. *et al.* (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.

Zanzoni,A. *et al.* (2007) Phospho3D: a database of three-dimensional structures of protein phosphorylation sites. *Nucleic Acids Res.*, **35**, D229–D231.