SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
MOLECULAR BIOLOGY
SYSTEMS BIOLOGY
PROTEOMICS
POST-TRANSLATIONAL MODIFICATIONS

Received
2 August 2011

Accepted
26 August 2011

Published
13 September 2011

Correspondence and requests for materials should be addressed to C.A.F. (floudas@titan.princeton.edu)

Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database

George A. Khoury, Richard C. Baliban & Christodoulos A. Floudas

Department of Chemical and Biological Engineering, A325 Engineering Quadrangle, Princeton University, Princeton, NJ 08544.

Post-translational modifications (PTMs) broadly contribute to the recent explosion of proteomic data and possess a complexity surpassing that of protein design. PTMs are the chemical modification of a protein after its translation, and have wide effects broadening its range of functionality. Based on previous estimates, it is widely believed that more than half of proteins are glycoproteins. Whereas mutations can only occur once per position, different forms of post-translational modifications may occur in tandem. With the number and abundances of modifications constantly being discovered, there is no method to readily assess their relative levels. Here we report the relative abundances of each PTM found experimentally and putatively, from high-quality, manually curated, proteome-wide data, and show that at best, less than one-fifth of proteins are glycosylated. We make available to the academic community a continuously updated resource (http://selene.princeton.edu/PTMCuration) containing the statistics so scientists can assess "how many" of each PTM exists.

Over a decade ago, Apweiler et al.[1] deduced statistics on the frequency of protein glycosylation by analyzing the Swiss-Prot[2] database. Swiss-Prot, updated monthly, provides a high level of manually curated information (i.e., sequence, mutations, function, ontology) about the proteome, as deduced from the literature, and strives to provide a nominal level of redundancy while being highly unified with over 100 other biomolecular databases. As of release 2011_07, there are 530,264 sequence entries. Although other databases exist[3–5], they focus mainly on one specific category of modification, or on one organism[6], and as such we wished to focus our efforts on gathering statistics from Swiss-Prot.

The rate and distribution of glycosylation was identified as important since it was thought to be the most common PTM at the time. Identifying this gap in information, Jung and coworkers proceeded in the significant undertaking of systematically annotating the glycoproteins in Swiss-Prot[7]. With the abundance of PTM data beyond glycosylation generated and submitted to Swiss-Prot each month, Farriol-Mathis and coworkers in 2004 nobly undertook the task to standardize the annotation of PTM features by creating a controlled vocabulary and updating the entries in the Swiss-Prot database[8].

dbPTM[9] was created in 2006 to annotate PTMs from Swiss-Prot, and currently gives the latest global quantification of PTM statistics. dbPTM was subsequently updated online in 2007, but has not been updated since. It was striking to discover that the number of post-translational modifications far exceeded the number of mutations identified and in time is projected to exceed the number of protein sequences contained in Swiss-Prot. Further, the rate of detection of PTM sites is considerably outpacing our biological knowledge of the function of those modifications[10].

Therefore we sought to create a method capable of curating, quantifying, and summarizing the level of each PTM reported from the high-quality information stored in Swiss-Prot. Our intent is to quantify and update these statistics dynamically in line with the monthly updates from the Swiss-Prot database, and more importantly to provide them at large in an easily accessible format to the academic community for further use in studying systems biology, proteomics, protein design, and the origins of life.

Results

Quantification of post-translational modifications levels. We have populated and made available the levels of each PTM (Supplementary Tables I, II, III) as outlined in our methods, and have made a resource (http://selene.princeton.edu/PTMCuration) available to the academic community that will update in line with Swiss-Prot. As of release 2011_07, we found that there were 87,308 experimentally identified post-translational modifications and 234,938 putative modifications on 530,264 proteins. These results nearly triple the number of experimental PTMs and double the putative PTMs previously reported[9]. Figure 1 summarizes the results of the major categories of PTMs.

Based on previous analysis of well-characterized and annotated glycoproteins, the current consensus estimate in the field is that more than half of proteins are glycosylated[1]. Our results reveal that the relative level of glycosylation is perhaps much lower than previously estimated. If every experimental and putative glycosylation event (N-linked, O-linked, C-linked glycosylation) only occurred once per protein, then only 104,011 proteins of the total set of 530,264 would be glycoproteins. Glycosylation events may occur more than once on a protein, and therefore the total level of glycoproteins at best case is less than one-fifth, as opposed to the widely accepted estimate of one-half[1].

Phosphorylation dominates the number of experimental PTMs identified by an order of magnitude, whereas N-linked glycosylation dominates the number of putative PTMs. The experimental distribution is not necessarily an absolute indication of the relative levels of each PTM, as some PTMs may be harder to experimentally elucidate than others, or currently may be of more interest to the academic community. This is why we chose to include the putative statistics. This is instructive as it may be important to focus substantial efforts on developing new methods for identifying PTMs that are not widely annotated (Supplementary Table II, III), as they may have a major undiscovered impact on the cell. The high level of phosphorylation experimentally elucidated may be directly related to the observation that it may be the most revealing and distinguishing factor of cellular status[11]. Considering both experimental and putative PTM sites, phosphorylation is found most frequently at 139,582 instances on 530,264 proteins. In comparison with previously reported results that used four[2, 3, 5, 12] external biological databases related to post-translational modification[9], all counts for these categories increased as expected, with the exception of O-linked glycosylation. This is a direct result of the former method utilizing additional information from ~2700 glycosylation events from O-GLYCBASE[3]. In time it is likely that the other 3 databases specific to phosphorylation[5], O-linked glycosylation[3], and ubiquitylation[12] may be cross-referenced with Swiss-Prot, adding the non-redundant entries to the full dataset. It may be possible that the non-redundant entries may have already been incorporated, as one cannot exhaustively determine this unless it is explicitly cross-referenced.

In Figure 2, we present the top experimental and putative PTMs identified from a starting controlled vocabulary of 431 PTMs. Phosphoserine, phosphothreonine, and N-linked glycosylation are the top 3 PTMs experimentally found, whereas N-linked glycosylation, phosphoserine, and N6-acetyllysine are the top 3 putatively found. Interestingly, the top 15 experimentally validated and putative PTMs correspond to 88% of those reported. The full raw and summarized datasets are available for download at http://selene.princeton.edu/PTMCuration and are included in Supplementary Table II and III.

It is noteworthy that the D-alanine isomer is among the top 15 experimentally found PTMs. This prompted us to explore the relative abundance of D-isomers further. Table 1 presents the relative frequencies of D-isomers experimentally found. There were just 837 D-isomers found out of a dataset of 187,941,074 amino acids! D-alanine, the most frequently occurring, was found to be 6-fold more frequent than the 2nd most, D-serine. Using these statistics

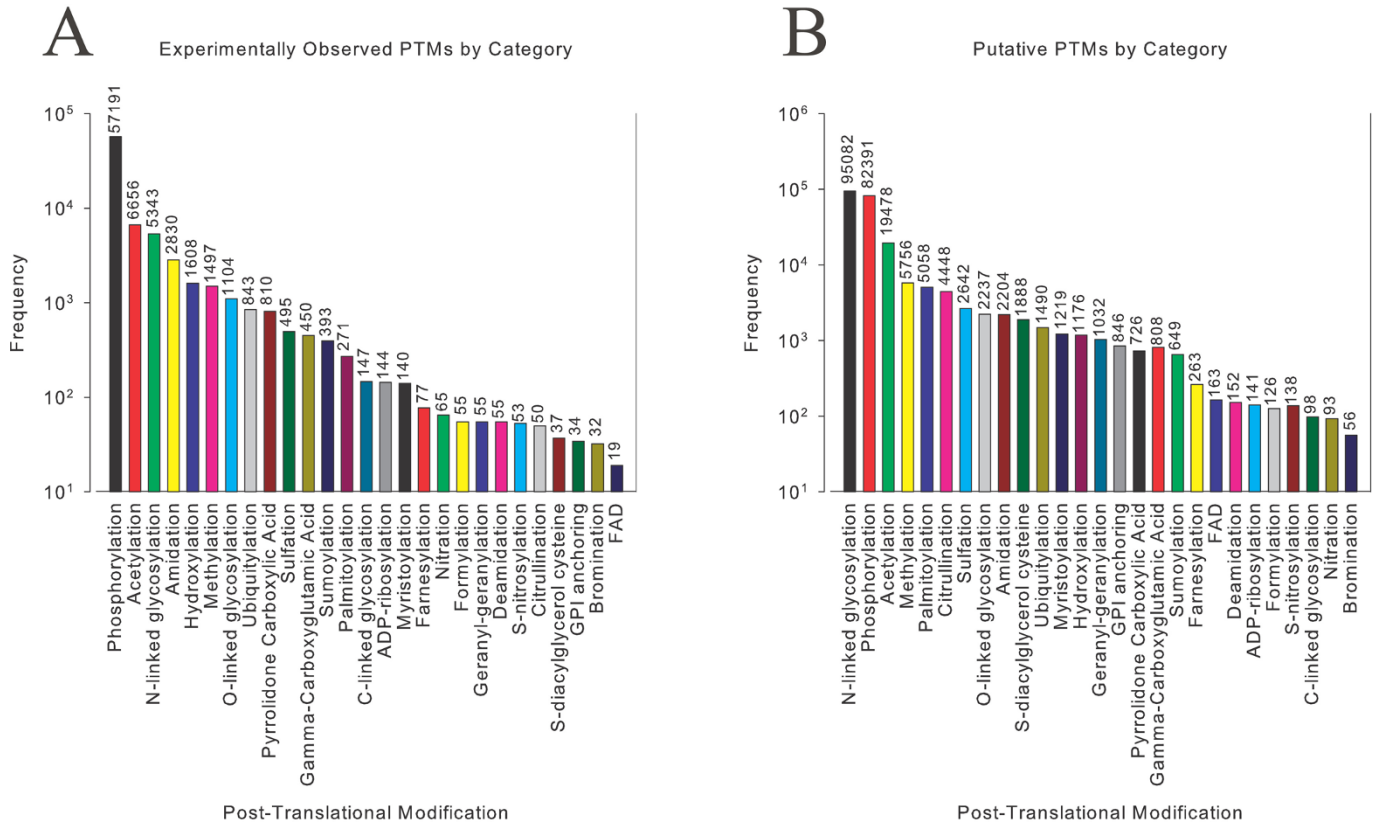

Figure 1 | Summary of (A) Experimental and (B) Putative Post-Translational Modifications Curated from Swiss-Prot.

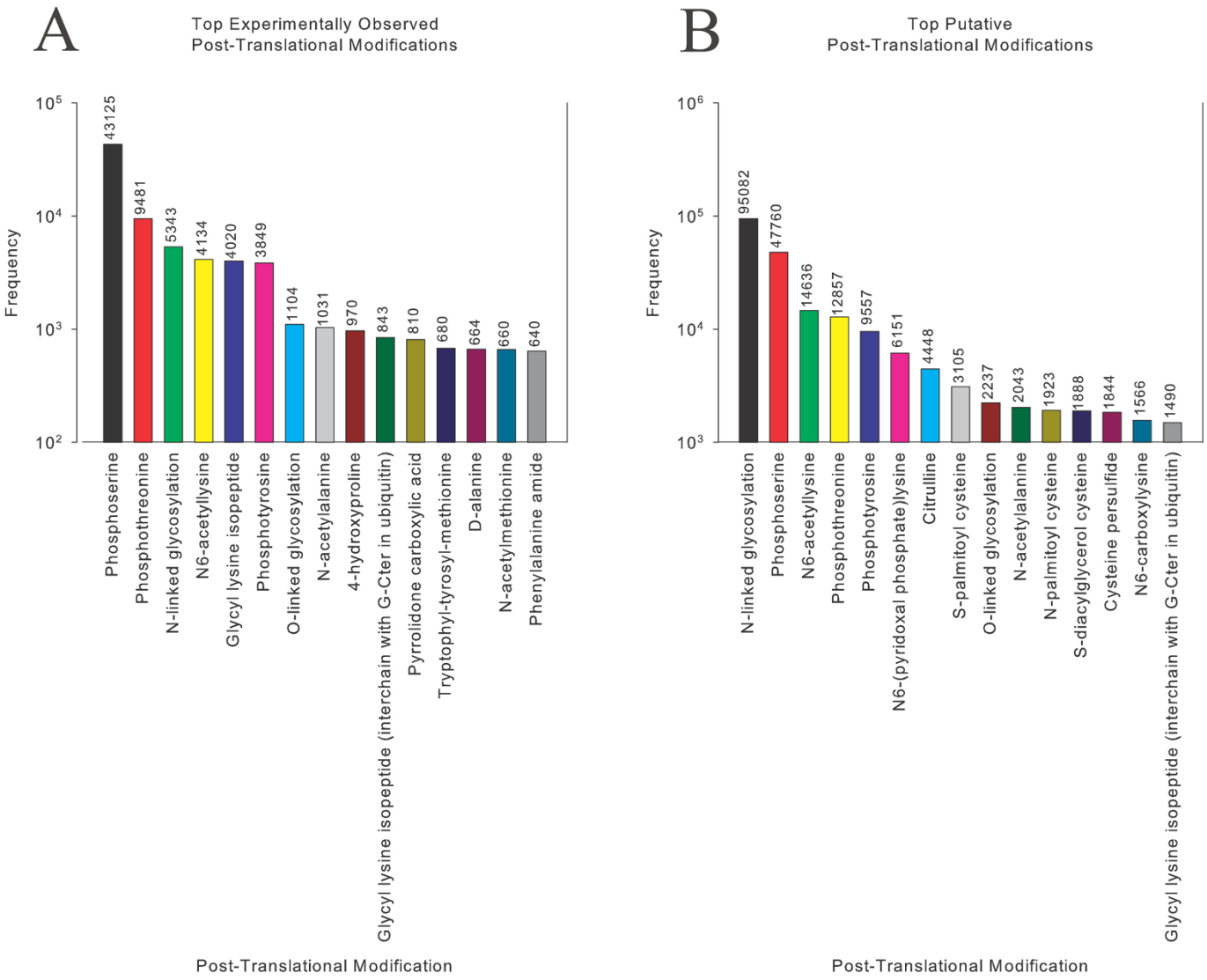

Figure 2 | Top (A) Experimental and (B) Putative Post-Translational Modifications Curated from Swiss-Prot.

uncovered by the algorithm, one may pursue rationalizations as to why D-alanine is so frequently found, if this has any implications on the origin of life, and how alanine's status as the "default amino acid" translated into this result. Specifically, knowing these relative frequencies may help researchers studying the origins of life by hinting towards what amino acids may be more important targets for theoretical and experimental analysis.

Table 1 | Statistics of experimentally validated D-isomers.

PTM Type	*# Experimental Sites*
D-alanine	664
D-serine	114
D-methionine	19
D-phenylalanine	15
D-valine	8
D-tryptophan	7
D-leucine	6
D-asparagine	2
D-threonine	2
Total	**837**

Discussion

The frequency of occurrence for each PTM we present can be translated into probabilistic form that represents the possibility of identification of a PTM on a peptide or protein. In the field of proteomics, it is often desirable to determine the false-discovery rate for the identification of a peptide, protein, or post-translational modification from tandem mass spectrometry. Algorithms often use a "decoy" database comprised of reverse protein sequences to determine the rate at which the method will identify an incorrect sequence by random chance. However, when attempting to perform this task using post-translational modifications, the number of database entries grows exponentially with the number of PTMs and can significantly hinder search times. The statistics uncovered by the methodology outlined in this manuscript can help facilitate the development of a probabilistic method that can determine the expectation value associated with a PTM identification for an individual protein or a complete cellular sample. In addition, this information will be of utmost importance for untargeted[13] and targeted[14] PTM studies that seek to uncover novel protein modification types and sites. Specifically, the assignment of a new modification to a protein can be validated using a probability of occurrence derived using the existence of that modification on other proteins. Also, the identification of a modification type to a new amino acid residue can utilize the

statistics surrounding how frequently the modification type appears and how frequently that amino acid is modified to help validate the assignment.

Recently, there have been many successes reported designing/redesigning proteins for novel functions using computational methods. Utilizing this information, researchers creating force-fields may focus in on those PTMs most frequently occurring to add to the repertoire of possible computational designs. This will aid in filling the gap of unexplored computational design space by including post-translationally modified side chains. Additionally, it may lead researchers to understand how PTMs affect structural changes in folding. Understanding the PTM's relative abundances may help us understand more fully the combinatorial histone code and its influence on diseases such as cancer[15]. Further work may include the decomposition of the PTM statistics by taxonomy to uncover which species are lacking in their list of annotated PTMs, and therefore require further investigation. The statistics/probabilities can also be cross-referenced into the BioNumbers database[16].

The rate-limiting step for continuously updating the statistics will be the manual curation rate, but 1691 sequences have been added since the previous release (2011_6), and 129,194 have been revised. As the Swiss-Prot database is continually updated, we believe this will not be diminishingly limiting. The data inputted is also subject to error itself (to date there are 71,424 entries with at least one sequence correction), but this error is not a major cause for concern as further manual annotation in time will increase the accuracy of data contained in the database.

For the field of proteomics to continue to expand and to garner a full picture of cellular function, open access to proteome-wide results are necessary[11]. Global datasets such as the ones we present are necessary for a systems-level understanding of biology; we must know what mechanisms exist that can control the state of a protein and therefore the cell to minimize malfunctions and provide therapeutics for the treatment of disease[17]. Ideally no absence of post-translational modification information should be overlooked by the academic community, and for this reason, we have created this tool. There is still a lot to learn about post-translational modifications, but scientists now have a way to assess "how many" PTMs exist.

Methods

Generation and curation of post-translational modification statistics. The workflow used to generate the data is presented in Figure 3. Curation and population of statistics are performed on the high-performance PC cluster Sesame (http://www.princeton.edu/researchcomputing/computational-hardware/machine-4/) using a 2.66 GHz Nehalem CPU with 3GB of RAM. All calculations involved in the compilation of statistics were completed within 5 CPU hours per run.

The workflow utilizes 6 major steps: (1) Accessing the most recent release of the Swiss-Prot database and its controlled vocabulary of PTMs; (2) Preprocessing of the PTM IDs; (3) Populating the experimental and putative statistics using logical conditions; (4) Sorting and categorizing the results; (5) Manually checking results to update any Problem IDs; (6) Sending the results to the web interface.

In step (1), we access, download, and decompress the most recent release of Swiss-Prot from ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.dat.gz. Subsequently, we extract the current controlled vocabulary of PTMs used in Swiss-Prot from http://www.uniprot.org/docs/ptmlist. Release 2011_07 contained 530,264 sequence entries and 187,941,074 amino acids. The vocabulary is passed to step (2), where the initial PTM IDs are automatically

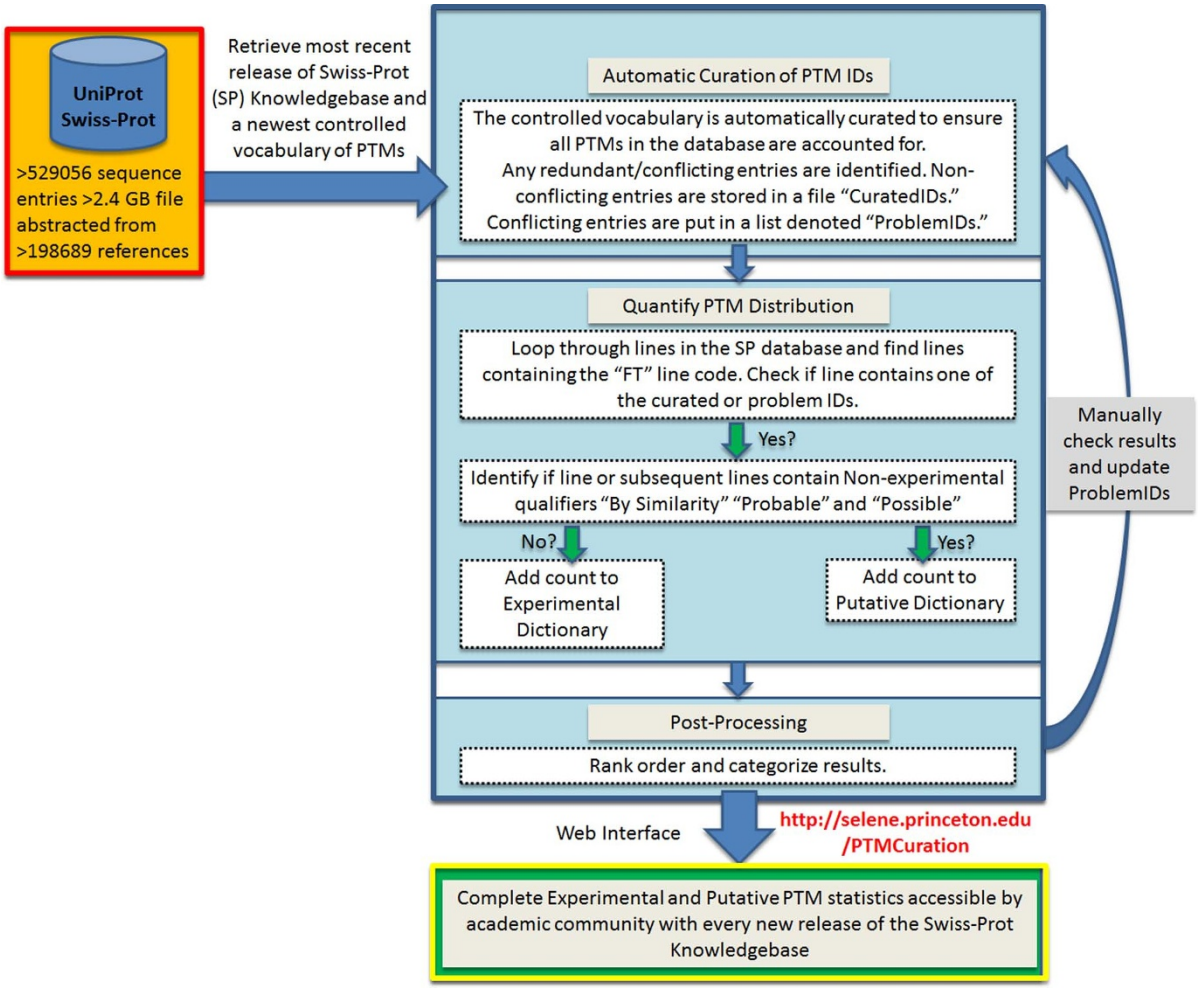

Figure 3 | Workflow for proteome-wide collection of post-translational modification statistics.

curated in a way to minimize any redundancy, as well as incorporate IDs identified as problematic in step (5). Problematic IDs that caused false negatives are stored in a separate file called "problemids.txt." The curated and problem ids are passed to step (3), where a program loops through the > 2.4 GB database line by line, finds lines containing the "Feature" line code, denoted by "FT" and checks if the line contains one of the curated or problematic IDs. The algorithm uses logical conditions to identify and separate experimentally validated PTMs and putative PTMs. Putative PTMs are sites that contain a non-experimental qualifier (Potential, Probable, By similarity) as defined by Swiss-Prot[2]. Following the exact definitions from Swiss-Prot, the term 'Potential' indicates there is some logical or conclusive evidence that the given annotation could apply. This non-experimental qualifier is often used to present results from protein sequence analysis software tools, which are only annotated if the result makes sense in the biological context of a given protein[2]. The term 'Probable' indicates stronger evidence than the qualifier Potential, as it implies that there must be at least some experimental evidence that indicates that the information is expected to be found in the natural environment of a given protein. The last non-experimental qualifier is 'By similarity'. This qualifier tags biological information that was obtained experimentally for a given protein and is thought to be propagated within a certain taxonomic range to other proteins[2]. Experimentally validated PTMs are those that do not have a non-experimental qualifier, as defined in previous analysis[1, 9].

Step (3) generates two raw statistics files that are post-processed in Step (4) by rank ordering and categorizing the results, as well as preparing them to be sent to the webpage. In Step (5), we manually check the list of PTMs from the controlled vocabulary that contained either zero experimental or zero non-experimental hits from the Swiss-Prot database to ensure that these particular PTMs were not present in the database. In fact, a PTM may be included in the controlled vocabulary even if it has not yet been observed in the literature or if it is part of the TrEMBL (unannotated) portion of the UniPort database. Using the controlled vocabulary, "zero" hits may arise as there are PTMs added that may have been included in the "Comments" section of a given entry, even though it has not been actually observed yet. An example of this is the PTM "diiodotyrosine" where the "Comments" section of the sequence entry Iodotyrosine dehalogenase 1 from PIG (Accession ID: Q6TA49) includes a reference to the PTM diiodotyrosine as its substrate, but not as a PTM. Also, "zero" hits may arise if there are PTMs that have been added to the controlled vocabulary as a result of being included in the TrEMBL (unannotated) portion of the UniProt database. TrEMBL currently has ~30X as many entries as Swiss-Prot does as of release 2011_07 (16,014,672 sequences), and as entries are manually annotated, they are moved from TrEMBL to Swiss-Prot. Furthermore, incorrect "zero" hits were mitigated by adding respective entries from the controlled vocabulary separately to a Problem IDs file in a way that ensured they were detected by the logical conditions, to ensure all PTMs were being accounted for.

Finally, in Step (6), the algorithm completes by compiling and compressing the raw and summarized experimental and putative PTM statistics, and sends them to an Apache webserver to the webpage http://selene.princeton.edu/PTMCuration. The webpage will be automatically updated monthly to keep accurate the PTM statistics contained in the Swiss-Prot database. We envision our method, its derivative, or the statistics generated can be incorporated/cross-referenced into the UniProt knowledgebase to maximize its utility to the academic community.

1. Apweiler, R., Hermjakob, H. & Sharon, N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta, Gen. Subj.* **1473**, 4–8 (1999).
2. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
3. Gupta, R., Birch, H., Rapacki, K., Brunak, S. & Hansen, J. E. O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res.* **27**, 370–372 (1999).
4. Gnad, F. *et al.* PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.* **8**, R250 (2007).
5. Diella, F. *et al.* Phospho.ELM: A database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics* **5**, 79 (2004).
6. Peri, S. *et al.* Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.* **32**, D497–D501 (2004).
7. Jung, E., Veuthey, A.-L., Gasteiger, E. & Bairoch, A. Annotation of glycoproteins in the SWISS-PROT database. *PROTEOMICS* **1**, 262–268 (2001).
8. Farriol-Mathis, N. *et al.* Annotation of post-translational modifications in the Swiss-Prot knowledge base. *PROTEOMICS* **4**, 1537–1550 (2004).
9. Lee, T.-Y. *et al.* dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res.* **34**, D622–D627.
10. Naegle, K. M. *et al.* PTMScout: A web resource for analysis of high-throughput post-translational proteomic studies. *Mol. Cell. Proteomics*, 2558–2570 (2010).
11. Tyers, M. & Mann, M. From genomics to proteomics. *Nature* **422**, 193–197 (2003).
12. Chernorudskiy, A. *et al.* UbiProt: a database of ubiquitylated proteins. *BMC Bioinformatics* **8**, 126 (2007).
13. Baliban, R. C. *et al.* A Novel Approach for Untargeted Post-translational Modification Identification Using Integer Linear Optimization and Tandem Mass Spectrometry. *Mol. Cell. Proteomics* **9**, 764–779 (2010).
14. DiMaggio, P. A., Young, N. L., Baliban, R. C., Garcia, B. A. & Floudas, C. A. A Mixed Integer Linear Optimization Framework for the Identification and Quantification of Targeted Post-translational Modifications of Highly Modified Proteins Using Multiplexed Electron Transfer Dissociation Tandem Mass Spectrometry. *Mol. Cell. Proteomics* **8**, 2527–2543 (2009).
15. Young, N., DiMaggio, P. & Garcia, B. The significance, development and progress of high-throughput combinatorial histone code analysis. *Cell. Mol. Life Sci.* **67**, 3983–4000 (2010).
16. Milo, R., Jorgensen, P., Moran, U., Weber, G. & Springer, M. BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Res.* **38**, D750–D753 (2010).
17. Kitano, H. Systems Biology: A Brief Overview. *Science* **295**, 1662–1664 (2002).

Acknowledgements

CAF gratefully acknowledges funding from the National Institutes of Health, National Library of Medicine under grant number 5R01LM009338. GAK gratefully acknowledges financial support from a National Science Foundation Graduate Research Fellowship under grant number DGE-0646086 and from Princeton University. We thank Eric First for help making the web tool.

Author contributions

GAK wrote the programs and collected the data. GAK, RCB, and CAF designed the algorithm, analyzed the data, and wrote the paper. All authors discussed and commented on the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

License: This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/

How to cite this article: Khoury, G.A., Baliban, R.C. & Floudas, C.A. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci. Rep.* **1**, 90; DOI:10.1038/srep00090 (2011).