Exploration of the Computational Predicted Internal Tagging Preferred Protein Properties

Ruoyu Chen*[1], Lukas Gerasimavicius[2]

[1] School of Biological Sciences, University of Edinburgh
[2] Institute of Genetics and Cancer, University of Edinburgh

Open Access

Received
23 Sep 2024

Revised
03 Oct 2024

Accepted
15 Oct 2024

Published
24 Oct 2024

Abstract

Protein tags are commonly used in many biological experiments, and adding a tag to an intolerant sequence position can significantly damage the functions of a protein and affect the outcome of an experiment. Recently, an in-house computational prediction method, TagScore, was developed which uses sequence homology to identify non-conservative regions of proteins permissive of tagging. The properties of the proteins that are predicted by TagScore to prefer internal tagging were explored using gene enrichment strategies. Proteins that prefer internal tagging were found to be related to GTPase-associated proteins and sequence features with disordered and polar regions, as predicted by TagScore.

DOI: 10.2218/esjs.9999 **ISSN** 3049-7930

Introduction

Gene fusion techniques are commonly used in tag insertion experiments, such as the use of green fluorescent protein tags for cell localisation, polyhistidine-tags for protein purification, and FLAG peptide tags (a synthetic polypeptide tag consisting of eight amino acids) for protein detection and purification, which all involve the construction of fusion proteins. In the majority of scientific literature concerning protein tags, these tags are placed at the N- or C-terminus of the protein of interest (POI). Some of the reasons may be conventional, for instance, if a given tag was already previously used at a certain terminus, it is likely that new proteins will be tagged at the same site with the same tag, following established practices from the past. Other reasons are theoretical, for example, that the end of the protein rarely includes active sites (Osuna 2021).

Most protein labels in contemporary research are added haphazardly, and have a few drawbacks because tagging a protein will make it different from its native form as a fusion protein (Yofe *et al.* 2016; Weill *et al.* 2019) or unable to perform their proper biological functions (Yofe *et al.* 2016; Ki *et al.* 2020). In protein engineering, internal tagging is often necessary, especially when the N- and C-termini of the protein cannot tolerate a tag (van Zwam *et al.* 2024). In addition, internal labelling is critical for several other important reasons when: the termini of the protein are buried or are functionally associated (Zordan *et al.* 2015); there is a need for multiple tagging (Dhar *et al.* 2020); internal tagging is resistant to proteolytic degradation for some proteins (Bäckström *et al.* 1994); the peptide needs to be structurally and functionally stabilised (Barthelmes *et al.* 2011); and when some functions of specific peptide are considered (Park *et al.* 2014).

Our as yet unpublished computational method, which we refer to as the TagScore method, was developed for predicting the best place to put a tag within a protein sequence using evolutionary information, which predicts the tagging sites both for human proteins and mouse proteins. This method is based on a simple principle of searching for regions of non-evolutionary conservatism by multiple sequence alignment (MSA) that may represent adaptive changes in a particular species or in a particular environment. Unlike conserved regions, which often contain critical active or binding sites, the non-conserved regions are more tolerant to sequence modifications, such as the insertion of protein tags, due to their lower functional importance.

*Student Author

Linkers of a protein are supposed to be a proper place to tolerate tag sequence insertions, which often connect two adjacent functional domains in a protein (George *et al.* 2002). These regions, because of their flexibility, can tolerate or adapt to the insertions of different gene sequences without disrupting the basic function of the protein, sometimes even acquiring new functions such as enhanced stability (Coyote-Maestas *et al.* 2020; Ford *et al.* 2020; Zane *et al.* 2023). For example, Lanthanide-binding tags (LBTs) could be incorporated into three different loops into the interleukin-1β (IL01β) protein without any impact of the overall fold of the protein or binding affinity of the LBTs tag (Barthelmes *et al.* 2011). In the case of Ras GTPase activating protein p120-RasGAP, the catalytic domain in its C-terminus promotes guanosine triphosphate (GPT) hydrolysis and the SH2, SH3, PH and CalB/C2 domains in the N-terminus allow functions such as cell migration and proliferation (Pamonsinlapatham *et al.* 2009).

Our TagScore method, using evolutionary conservation from MSA to infer protein positions tolerant of insertions, was run for every human and mouse protein, generating a dataset of 19,708 unique human protein-coding genes with computational predictions for tag tolerability. Based on the observed alignments and insertions in homologous proteins, each protein residue position was annotated with a TagScore that is scaled in the range 0-1 and represents the probability of tolerating an insertion at the given position. Additional features, such as relative solvent accessibility (RSA) and the AlphaFold modelling quality metric predicted local distance difference test score (pLDDT) were also annotated for comparison. The score was utilised both at the per-residue level and at gene-level. At the gene level, the per-residue TagScores were used to compare tag tolerability across three distinct protein sequence locations — at the N-terminus, the C-terminus, or internally. For each location class, a residue position with the highest TagScore was chosen to represent the gene tag tolerance at that location and, for each protein, the location class (N-terminus, C-terminus or internal) with the highest TagScore was chosen to annotate the protein as being the most tolerant of insertions at that location.

In this study, to gain a deeper understanding of the potential application of TagScore method in specific biological contexts, gene enrichment analyses were performed to explore whether proteins with high tag tolerance are clustered in certain specific biological processes or pathways. This analysis helped reveal which biologically functional proteins might be more suitable for tag insertion, thus providing more precise biological information for protein engineering.

Method

Based upon whether tags are predicted to be the most favourable according to the highest TagScore across the three classes: N-score, C-score, and internal score, the proteins were classified into three groups.

To clarify the biological function and signalling pathways associated with internal tagging preferred tagging genes, gene ontology (GO), Reactome pathway (Gillespie *et al.* 2022) and sequence feature enrichment analysis were conducted by the online resource DAVID v2024q2, accessed on 5th September 2024 (Sherman *et al.* 2022). GO provides comprehensive and computable knowledge concerning gene functions and products (Aleksander *et al.* 2023), which includes three functional categories: biological process (BP), cellular components (CC) and molecular function (MF).

Input data consisted of a list of gene identifiers of genes of the proteins that prefer internal tagging in 'UNIPROT_ACCESSION' format which was the accepted standard of DAVID. The 'Gene_Ontology' functional annotation tool within DAVID was selected and three main GO categories ('GOTERM_BP_DIRECT', 'GOTERM_CC_DIRECT', 'GOTERM_MF_DIRECT') were selected to retrieve comprehensive functional data for the input genes. For pathway analysis, 'Pathways' options were selected, and from the list of available databases, 'REACTOME_PATHWAY' was chosen to explore the specific signalling pathways related to our input gene set. 'UP_SEQ_FEATURE' was selected for sequence feature enrichment analysis. All the analyses used the default parameters. The results were generated through the 'functional annotation clustering' option. A p-value of $P < 0.05$ was considered statistically significant and terms were selected using an Benjamini-Hochberg false discovery rate threshold of 0.05. Enrichment plots were generated according to fold enrichment, count read, and $-\log_{10}(P)$ through 'ggplot2' 3.5.1. All the code and enrichment results in DAVID online resource are provided at the end of this paper.

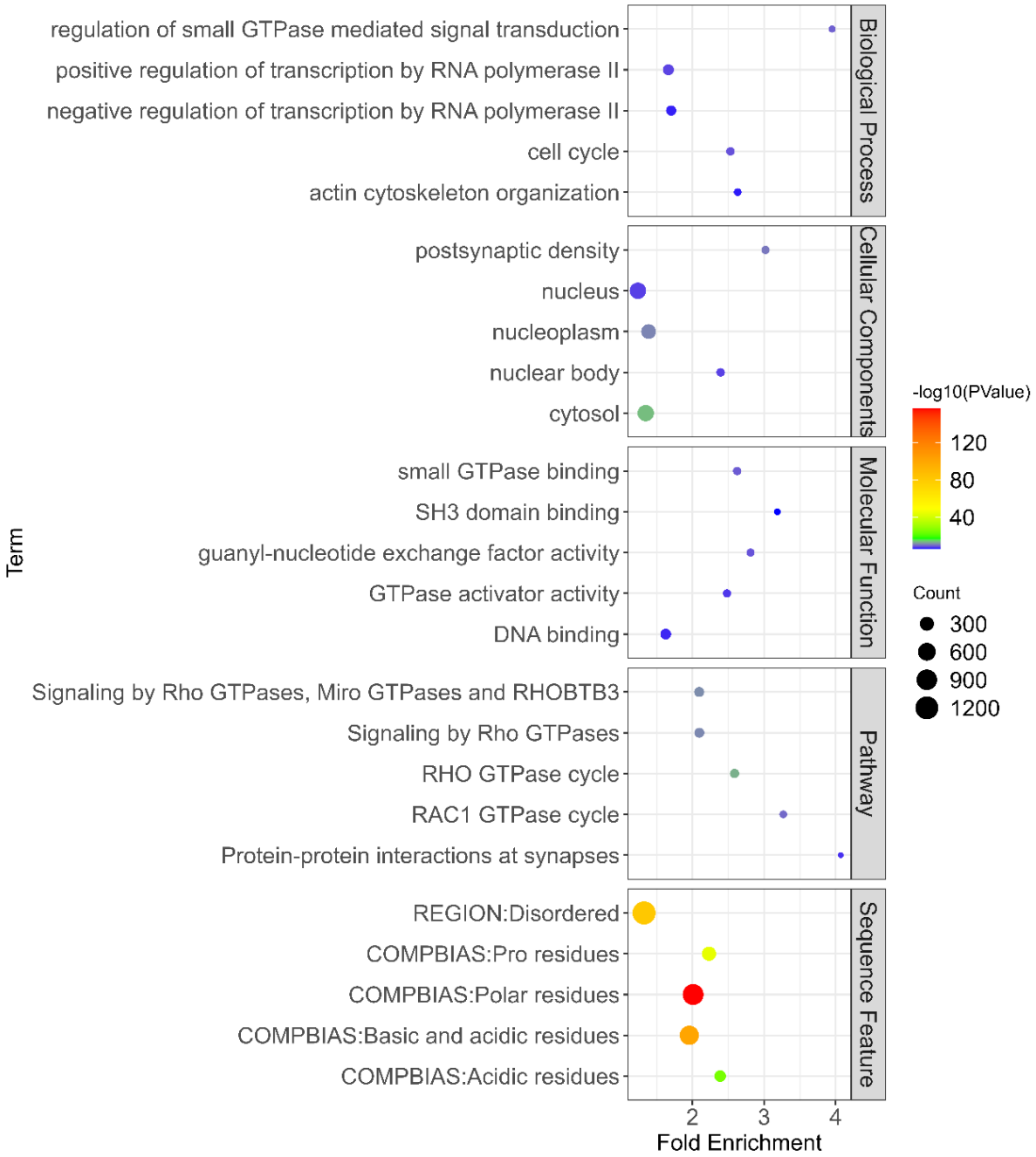

Figure 1: Bubble plot of the enrichment analysis of internal tagging preferred genes. The x- and y-axes indicate the different categories and fold enrichment, respectively. The size of the bubbles represents the number of genes enriched and the significance is shown coloured by $-\log_{10}(P)$, with red indicating the increase in significance.

Results and Discussion

Biological features of proteins that tend to tolerate tags at different sites have not been previously investigated in detail. To verify if there are any common features of the internal tagging group, genes were used to accomplish enrichment analysis separately using the DAVID online data analysis tool and these results are visualised in Figure 1.

The results show that the top biological processes attributed to the genes of the proteins that prefer internal tagging included 'regulation of small GTPase mediated signal transduction', 'positive *and* negative regulation of transcription from RNA polymerase II promoter', 'cell cycle' and 'actin cytoskeleton organisation'. Their cellular components were mostly attributed to the terms indicating localisation in the nucleus and cytoplasm, and the molecular function terms were largely attributed to the 'small GTPase binding' and 'GTPase activator activity'. The internal group was suggested to be associated with GTPase-related signalling pathways as well as biological processes.

From the REACTOME pathway analysis for the internal group, 30 pathways were significantly enriched ($P < 0.01$, FDR < 0.05). The top 5 pathways were primarily associated with the GTPase related cycle, including 'Signalling by Rho GTPase, Miro GTPases and RHOBTB3', 'Signalling by Rho GTPases', and 'RHO GTPase cycle'. Sequence feature analysis was significantly enriched in disordered and polar residue terms for the internal tagging group, which are usually exposed to the surface of the protein, generating more possibility of tag tolerance compared to buried positions. Overall, gene enrichment analysis showed that genes which prefer internal tagging are more likely to be related to GTPase associated proteins.

There are several possible properties of the proteins in the internal tagging group that make them much better targets for internal tagging, such as the longer protein length on average and a higher disordered sequence content, as predicted by pLDDT. When both protein termini are involved in important molecular functions, it may be a good approach to attempt and insert tags internally. GTPase-associated proteins should be one of the more numerous groups that arise as multi-domain proteins.

Code Availability

The code written for this project is available at:
`https://github.com/B238522-2023/property_analysis.git`.

Acknowledgements

I am deeply grateful to Prof. Joseph A Marsh for his invaluable guidance and profound impact on this project. I also extend my sincere thanks to Dr. Lukas Gerasimavicius for his generous support in data analysis and project research. Additionally, I appreciate the encouragement and support from my family and friends.

References

Aleksander, S. A. *et al.* 'The Gene Ontology Knowledgebase in 2023' Genetics **224** 1 (2023)
Bäckström, M *et al.* 'Insertion of a HIV-1-Neutralizing Epitope in a Surface-Exposed Internal Region of the Cholera Toxin B-subunit' Gene **149** 2 (1994)
Barthelmes, K. *et al.* 'Engineering Encodable Lanthanide-Binding Tags Into Loop Regions of Proteins' Journal of the American Chemical Society **133** 4 (2011)
Coyote-Maestas, W. *et al.* 'Targeted Insertional Mutagenesis Libraries for Deep Domain Insertion Profiling' Nucleic Acids Research **48** 2 (2020)
Dhar, P. *et al.* 'Genetically Engineered Protein Based Nacre-Like Nanocomposites With Superior Mechanical and Electrochemical Performance' Journal of Materials Chemistry A **8** 2 (2020)
Ford, R. C. *et al.* 'Linker Domains: Why ABC Transporters 'Live in Fragments No Longer'' Trends in Biochemical Sciences **45** 2 (2020)
George, R. A. and Heringa, J. 'An Analysis of Protein Domain Linkers: Their Classification and Role in Protein Folding' Protein Engineering, Design and Selection **15** 11 (2002)
Gillespie, M. *et al.* 'The Reactome Pathway Knowledgebase 2022' Nucleic Acids Research **50** D1 (2022)

Ki, M.-R. and Pack, S. P. 'Fusion Tags to Enhance Heterologous Protein Expression' Applied Microbiology and Biotechnology **104** 6 (2020)

Osuna, S. 'The Challenge of Predicting Distal Active Site Mutations in Computational Enzyme Design' WIREs Computational Molecular Science **11** 3 (2021)

Pamonsinlapatham, P. *et al.* 'p120-Ras GTPase Activating Protein (RasGAP): A Multi-interacting Protein in Downstream Signaling' Biochimie **91** 3 (2009)

Park, A. *et al.* 'CRISPR/Cas9 Allows Efficient and Complete Knock-In of a Destabilization Domain-Tagged Essential Protein in a Human Cell Line, Allowing Rapid Knockdown of Protein Function' PLoS ONE **9** 4 (2014)

Sherman, B. T. *et al.* 'DAVID: A Web Server for Functional Enrichment Analysis and Functional Annotation of Gene Lists (2021 Update)' Nucleic Acids Research **50** W1 (2022)

Weill, U. *et al.* 'Assessment of GFP Tag Position on Protein Localization and Growth Fitness in Yeast' Journal of Molecular Biology **431** 3 (2019)

Yofe, I. *et al.* 'One Library to Make Them All: Streamlining the Creation of Yeast Libraries via a SWAp-Tag Strategy' Nature Methods **13** 4 (2016)

Zane, L. *et al.* 'Peptide Linker Increased the Stability of Pneumococcal Fusion Protein Vaccine Candidate' Frontiers in Bioengineering and Biotechnology **11** (2023)

Zordan, R. E. *et al.* 'Avoiding the Ends: Internal Epitope Tagging of Proteins Using Transposon Tn7' Genetics **200** 1 (2015)

Van Zwam, M. C. *et al.* 'IntAct: A Nondisruptive Internal Tagging Strategy to Study the Organization and Function of Actin Isoforms' PLOS Biology **22** 3 (2024)