ProminTools: Shedding light on proteins of unknown function in biomineralization with user friendly tools illustrated using mollusc shell matrix protein sequences

Alastair W Skeffington [Corresp., 1], **Andreas Donath** [1]

[1] Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany

Corresponding Author: Alastair W Skeffington
Email address: skeffington@mpimp-golm.mpg.de

Biominerals are crucial to the fitness of many organism and studies of the mechanisms of biomineralization are driving research into novel materials. Biomineralization is generally controlled by a matrix of organic molecules including proteins, so proteomic studies of biominerals are important for understanding biomineralization mechanisms. Many such studies identify large numbers of proteins of unknown function, which are often of low sequence complexity and biased in their amino acid composition. A lack of user-friendly tools to find patterns in such sequences and robustly analyse their statistical properties relative to the background proteome means that they are often neglected in follow-up studies. Here we present ProminTools, a user-friendly package for comparison of two sets of protein sequences in terms of their global properties and motif content. Outputs include data tables, graphical summaries in an html file and an R-script as a starting point for data-set specific visualizations. We demonstrate the utility of ProminTools using a previously published shell matrix proteome of the giant limpet *Lottia gigantea*.

ProminTools: Shedding light on proteins of unknown function in biomineralization with user friendly tools illustrated using mollusc shell matrix protein sequences

Alastair William Skeffington[1], Andreas Donath[1]

[1] Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, Potsdam-Golm, 14476, Germany

Corresponding author:

Alastair Skeffington[1]

Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, Potsdam-Golm, 14476, Germany

Email address: skeffington@mpimp-golm.mpg.de

Abstract

Biominerals are crucial to the fitness of many organism and studies of the mechanisms of biomineralization are driving research into novel materials. Biomineralization is generally controlled by a matrix of organic molecules including proteins, so proteomic studies of biominerals are important for understanding biomineralization mechanisms. Many such studies identify large numbers of proteins of unknown function, which are often of low sequence complexity and biased in their amino acid composition. A lack of user-friendly tools to find patterns in such sequences and robustly analyse their statistical properties relative to the background proteome means that they are often neglected in follow-up studies. Here we present ProminTools, a user-friendly package for comparison of two sets of protein sequences in terms of their global properties and motif content. Outputs include data tables, graphical summaries in an html file and an R-script as a starting point for data-set specific visualizations. We demonstrate the utility of ProminTools using a previously published shell matrix proteome of the giant limpet *Lottia gigantea*.

Introduction

Mineralized structures are formed by many organisms across the tree of life including bacteria, metazoans, plants and algae (Skinner & Jahren 2007). These biominerals are critical for fitness, playing roles in support, defence, buoyancy, regulation of ion budgets and orientation among others. Proteins have been found to be associated with many biominerals, and are hypothesised to have a key role in mineral synthesis (Evans 2019a; Evans 2019b; Wang & Nilsen-Hamilton 2012). In some cases the roles of such proteins is relatively well understood, and some of the best studies examples come from molluscs (Song et al. 2019). For example, the proteolytic products of the Pif protein in molluscs have been shown to bind $CaCO_3$ crystals and induce formation of the aragonite polymorph of CaCO3 *in vitro* (Suzuki et al. 2009). Knock-down of the Pif gene results in disordered growth of the aragonite crystals in the nacreous layer of the shell. In other systems, well studies examples include Amelogenin from tooth enamel, Silicatein from sponge spicules and Mms6 from magnetosome synthesising bacteria (Wang & Nilsen-Hamilton 2012). However in the majority of cases the function of biomineral associated proteins remains elusive.

A common workflow in biomineralization research is to first clean a mineral preparation using detergents or oxidizing agents to remove loosely associated organic matter, and subsequently to dissolve the mineral, releasing tightly mineral-associate proteins into solution that can then be analysed using proteomic methods (Marie et al. 2013b). It is generally hypothesised that these proteins are likely to be involved in mineralization, and that proximity to the site of mineralization results in their incorporation into the mineral as it grows. Some of the proteins identified may have homology to proteins of known function or recognisable domains strongly suggestive of a certain function. For example, carbonic anhydrases have been found associated

with calcium carbonate minerals in several organisms (Le Roy et al. 2014) and may aid generation of bicarbonate as a substrate for calcification. However there are generally many proteins in such data sets which lack similarity to proteins of known function (e.g. Jackson et al. 2015; Kotzsch et al. 2016; Mann et al. 2006). Intriguingly, these proteins of unknown function often display unusual primary sequence characteristics, such as low complexity, biased composition and a high degree of predicted intrinsic disorder.

Informatic tools which allow biologists to easily investigate the global features of groups of proteins of unknown function relative to the background proteome are currently lacking. Thus many studies restrict their analysis of these proteins to noting the compositional biases or motifs which are obvious from manual inspection of the protein sequences. This method has the risk that important patterns in the data are missed and that rules are not applied consistently in identifying these patterns. Ideally the context of the proteome as a whole should also be taken into account. The more specific a feature is to the proteins of interest (POIs) the more likely it is to be involved in the specific function of those proteins. This notion is based on the well-established biological principle that the primary sequence of a protein is a strong determinant of molecular function, and that proteins with similar functions tend to share regions of sequence similarity. Thus, sequence motifs shared by a group of biomineral associated proteins are more likely to be involved in the specific function of those proteins if they are rare in the background proteome than if they are commonly found motifs. This principle is already used in various sequence analysis tools, including those seeking to identify important motifs (Wagih et al. 2016).

Although there are many tools available that allow researchers to investigate the properties of protein sequences *in silico*, including analysis of compositional bias, sequence complexity, intrinsic disorder and sequence motifs, such tools are not always easy to use. Some require

command line use, data input formats differ, some can only run on one protein at a time and most require post-processing of the output to format the data for statistical and graphical analyses in commonly used environments such as Microsoft® Excel® or R. These tools also rarely allow researchers to compare two sets of sequences.

Here we present ProminTools, a set of easy-to-use tools for the statistical comparison of two sets of protein sequences, available as apps in the CyVerse Discovery Environment (https://de.cyverse.org/) (Merchant et al. 2016) or to run locally from a Docker™ container. The inputs are simply two fasta files containing the proteins of interest (POIs) and the background proteome respectively, while the outputs include data tables and an html document containing graphical summaries of the data and interactive tables for data exploration. To demonstrate the utility of these tools, we reanalyse a published data set of shell matrix proteins (SMPs) from the giant limit *Lottia gigantea* (Mann & Edsinger 2014).

Materials and Methods

ProminTools structure

The inputs for ProminTools are two fasta formatted files: the first containing the protein set of interest (POI set), and the second the reference or background proteins (typically the predicted proteome of the organism of interest). The background proteins are used for statistical comparisons with the POI sets, allowing the user the answer the following question: 'Are the features observe in the POI common in the background sequences or are they unusual?'.

ProminTools has two component programmes: "Protein Motif Finder" and "Sequence Properties Analyzer". Both are written in Perl and R and bundled with all dependencies in Docker™ (www.docker.com) containers. They can be run from Apps within the CyVerse Discovery Environment (Merchant et al. 2016) or on a personal computer via Docker™ Desktop. The

primary outputs of the tools are data tables summarising key information from the comparison of the two sequences sets. The tools use these tables to generate an html file with a graphical summary of the information along with explanations, statistical analyses and interactive versions of certain data tables. A publication ready SVG (Scalable Vector Graphic) formatted figure is also generated by Protein Motif Finder. The R script that generates the html file from the data tables is also an output of the tools, allowing the user to reproduce the figures in the html report and to provide a starting point for further analyses specific to the data set. For licence information for all components of ProminTools the reader is referred to *Data S1*.

Analyses performed by "Protein Motif Finder"

Motif finding with motif-x

Protein Motif Finder uses the motif-x engine (Schwartz & Gygi 2005; Wagih et al. 2016) for motif finding. This engine was chosen because it breaks sequences down into their constituent motifs, by an iterative procedure that avoids oversimplification of motifs and prioritises motifs that are most enriched relative to a background sequence set. It is exhaustive for a given p-value and generates definite motifs rather than a position weight matrix, which simplifies downstream analyses and is more useful to molecular biologists. In this work it was always run with the recommended, conservative, binomial *p*-value of 10^{-6}, but this parameter is user customisable in Protein Motif Finder. The motif width is also user customisable, while the minimum occurrences parameter is hard coded at a value of five. Motif-x is run via the R module rmotif-x, centred on each amino acid sequentially and the results are combined. This procedure means that some motifs are likely to be redundant. For example, if the central residue is 'S' and the motif width 7 then the motifs '…S.S..' and '..S.S… (where dot represents any amino acid) may both be identified, but these would be collapsed to the single motif 'S.S'. Note that this procedure is

conservative with respect to the original p-value calculated by motif-x. Significant motifs are then enumerated in the POI and the background sequence sets, and motif counts and enrichments reported in the output tables. Downstream analyses do not rely on the motif-x p-value, but only on calculated enrichment values for the motifs.

Graphical representations of motif data

To provide a visual summary of the motif data, the motifs are represented in three wordclouds in the Protein Motif Finder output, which take into account two distinct measures of 'importance'. The first is the number of proteins in which that motif is found. The more proteins containing the motif, the more likely it is to have general importance in the function of the group of proteins. The second measure is the enrichment of the motif. The more enriched the motif the more unusual it is and thus is more likely to be involved in the specific function of these sets of proteins. A third measure attempts to combine the previous two by scaling them equivalently and then taking the product of the scaled values (PS-value). This measure prioritises motifs that are both highly enriched and found in a high proportion of the proteins.

In the output, proteins that are biased in sequence composition are also clustered based on their motif number and motif enrichment. The distance measure for clustering was calculated as one minus the Distance Correlation (Szekely et al. 2013) for all pairwise combinations of proteins or motifs, since this method is especially robust to outliers and produces reasonable results across a variety of datasets. Hierarchical clustering was performed using the Ward.D method. In Heatmap 1, the following filters are applied to select proteins and motifs for clustering: 1) Select proteins that contain a biased region (fLPS p-value $< 10^{-20}$; user adjustable), 2) Remove infinitely enriched motifs. 3) Remove motifs not present in at least 3 POI proteins (if >10 proteins in POI set). 4) Select the 70 overall most enriched motifs. 5) Select proteins with a good motif based

correlation to at least one other protein (dcor > 0. 65, user adjustable). The filters for Heatmap 2 are the same except that filter 3 is not applied. The same protein and motif set is displayed in Heatmap 3 as in Heatmap 2 except that motif count is displayed instead of motif enrichment.

Analyses performed by "Sequence Properties Analyzer"

Sequence Properties Analyzer performs the following analyses:

Amino acid enrichment

Compositional bias is analysed using fLPS (Harrison 2017) and the results collated to several files described in the html output of the program.

Significance of sequence bias

To estimate the probability of obtaining the observed bias in amino acid composition in the POI set by random sampling of the background proteome, the following procedure was implemented. First the degree of bias was quantified by calculating a bias index (BI):

$$BI = \sum_{Amino\ acids} \sqrt{(POI\ freq. - Proteome\ freq.)^2}$$

Where *POI freq.* is the frequency of the amino acid in the POI proteins, while *Proteome freq*. is the frequency of the amino acid in the background sequence set. The *BI* is calculated for 1000 random samples of the background sequence set, each containing the same number of sequences as the POI set. A kernel density estimate of the distribution of *BI* is calculated, and a function approximating this distribution is generated. The area under the curve greater than the *BI* value of the POI set is used as an estimate of the probability of obtaining a sequence set of this degree of bias by chance, given this particular background proteome.

The only program we are aware of that makes a similar calculation is Composition Profiler (Vacic et al. 2007). However this makes the assumption that all POI sequences come from the same underlying distribution of amino acid frequencies and tests whether this distribution is

significantly different from the background. ProminTools does not make this assumption, but accepts that the POI set my contain proteins with different types of bias, and thus analyses bias *per se* without reference to the type of bias.

Sequence complexity

The program SEG (Wootton & Federhen 1993) is used to identify low complexity regions in the datasets using default parameters, although these are customisable by the user in Sequence Properties Analyzer. For each protein, the percentage of the sequence identified as low complexity is calculated, and a Wilcoxon rank sum test with continuity correction is used to test whether there is a significant difference in the distribution of this percentage length between the POI and the background sequence set.

Intrinsic disorder

Predicted intrinsic disorder was calculated using the VSL2 predictor (Peng et al. 2006), due to its speed and good accuracy (Nielsen & Mulder 2019). This is the most time consuming step of Protein Sequence Analyzer and is thus parallelized in the implementation. For each protein, the percentage of the sequence identified as intrinsically disordered is calculated and a Wilcoxon rank sum test with continuity correction is used to test whether there is a significant difference in the distribution of this percentage length between the POI and the background sequence set.

Charged clusters

Clusters of charged amino acids are identified using the SAPS software (Brendel et al. 1992).

Data and methods for validation of ProminTools

Representative CxxC Zn finger proteins were chosen from the Wingender database (Wingender et al. 2013) and compared to the human proteome Swissprot database accessed on the 28/05/20. For the analysis of human low complexity proteins, all models were downloaded from Ensemble

version 100. Models shorter than 100 amino acids were removed, as were models with internal stop codons, resulting in 89562 proteins that were used as the background sequence in the analysis. The foreground sequence set was the 500 most biased proteins identified using fLPS (Harrison et al. 2017). These proteins were annotated using eggNOG mapper (Huerta-Cepas et al. 2017, 2019) with parameters "taxonomic scope hominidae, -target_orthologs all --seed_ortholog_evalue 0.001 --seed_ortholog_score 60 --query-cover 20".

L. gigantea shell matrix proteome data

To illustrate the utility of the ProminTools package, we used the shell matrix proteome of the giant limpet, *L. gigantea* as published by Mann and Edsinger (2014) which is a reanalysis of their original data (Mann et al. 2012). The protein identifiers were extracted from table S1 of (Mann & Edsinger 2014) and the protein sequences extracted from files Lotgi1_GeneModels_AllModels_20070424_aa.fasta and Lotgi1_GeneModels_FilteredModels1_aa.fasta which were downloaded from the JGI (https://mycocosm.jgi.doe.gov/Lotgi1/Lotgi1.home.html) on the 5/02/2020. The final set of proteins consisted of 381 sequences, and are available in *Data S2*. This is one less than the number of accepted identifications in (Mann & Edsinger 2014) since the protein Lotgi|172500 was not available in any database.

Analyses of _L. gigantea_ data using ProminTools

ProminTools was run locally using the shell matrix proteins as the foreground sequences and the 'Lotgi1_GeneModels_FilteredModels1_aa.fasta' file as the background proteome. The filtered models were chosen as they were considered likely to be a closer representation of the true proteome of *L. giga*ntea than the 'All Models' set, and thus the more appropriate set for statistical comparisons.

Additional analyses

Proteins were clustered based on motif content as an output of the Protein Motif Finder tool. To determine an optimal cluster number, manual inspection of a plot of the cindex (Hubert & Levin 1976) for cluster sizes 2 – 50 was carried out. A cluster number of 8 seemed appropriate for the present work, since it captured the major patterns in the data without becoming too granular. These clusters were the input for further runs of Protein Motif Finder.

Sequence similarity was quantified using and all vs all pairwise BLASTp analysis, reporting the percentage identity of the top scoring high scoring pair (HSP), after applying an e-value cut-off of 0.01 and a cut-off specifying that the HSP alignment length must be at least 20% of the query length.

Results

ProminTools provides a user-friendly method to analyse biomineralization proteomes

The Docker™ image containing ProminTools can be run via a GUI on the CyVerse Discovery Environment without any need for use of the command line. Runtimes on CyVerse are variable due to variable resource availability, but a typical analysis with either Protein Motif Finder or Sequence Properties Analyzer takes between 30 and 120 minutes to complete. Although ProminTools is designed to run in a Unix environment, it can also be run on a windows PC via Docker™ Desktop with simple commands in Windows® Power Shell™ (for details see https://github.com/skeffington/Promin-tools). On a Window® 10 machine with an Intel® Core™ i7-2600 3.4 GHz processer and 16 GB RAM, Protein Motif Finder completed analysing the *L. gigantea* data set in 11 min 30 s provided with 1 core and 2.5 GB RAM, while the Sequence Properties Analyser completed in 32 min 2 s provided with 5 cores and 5 GB RAM. The ProminTools workflow is summarised in *Fig. 1*.

Validation of ProminTools

Suitability for a range of data inputs

To ensure stability and good performance, we have tested ProminTools on a number of published and unpublished biomineralization datasets and used synthetic data to ensure that the program deals sensibly with unusual situations, such as small numbers of POI sequences or no motifs being found. An example analysis of a second data set, of proteins from freshwater mussel shells (Marie et al. 2017), is provided in *Data S3*.

Validation with negative control protein sets

Five sets of 100 proteins were drawn at random from the *L. gigantea* proteome and each used as the POI set to run ProminTools in five separate analyses. No enriched motifs were reported in any of the analyses. In all analysis, there were no significant differences in the degree of sequence bias, sequence complexity or intrinsic disorder between the random 'POI' set and the background proteome. Representative analyses are provided in *Data S4*.

Validating motif retrieval in Protein Motif Finder

Motif finding in ProminTools relies on the motif-x motif finding engine, which has already been well validated (Schwartz and Gygi 2005). However to ensure that there were no bugs in our post-processing of the output we spiked motifs at known frequencies into a set of protein sequences and ran Protein Motif Finder with these sequences as foreground, and the un-spiked sequences as background. The spiked motifs were recovered at the expected frequencies.

We also validated Protein Motif Finder on a groups of sequences containing motifs that have already been established as important for protein function. For example CXXC class zinc finger factors are transcription factors and histone methyltransferases that bind to CpG elements via zinc fingers. The Zn binding residues consist of cysteines arranged in CGxCxxC motifs

(Wingender et al. 2013). Using a set of CXXC zinc fingers factors as the POI set and the human proteome as the background set, Protein Motif Finder correctly identified CG.C..C as the most important motif, and found it to be 194 fold enriched in these proteins relative to the background proteome (*Data S5*).

It should be noted that not all motifs important for a protein's function will be enriched relative to the background. For example the L..LL motif is important in protein-protein interactions that regulate transcription (Plevin et al. 2005). Using the 55 Swissprot proteins annotated as possessing an L..LL motif as the POI set and the Swissprot human proteome as background, Protein Motif Finder does not recover the L..LL motif (*Data S6*). This is because L..LL is relatively common in other contexts, and so is not significantly enriched in the POI set.

Validating the biological meaning of clustering by motif enrichment

A key output of Protein Motif Finder is clustering of the POIs based on their motif enrichment. The usefulness of this clustering is based on the assumption that proteins within a cluster are likely to be involved in similar molecular processes. To test this assumption on a well annotated proteome, but focusing on biased sequences similar to those expected from biomineral associated proteins, we analysed the 500 most biased sequences from the human proteome. Of these, 303 could be annotated by eggNOG mapper (Huerta-Cepas et al. 2017) and they fell into 120 clusters when analysed with Protein Motif Finder (*Data S7*). Remarkably, proteins within a cluster all shared the same eggNOG functional annotation in 117 of the clusters, even when the proteins diverged significantly in primary sequence similarity as assessed by global alignments (*Table 1*). Of the input proteins, 53 were collagens, and 11 different types of collagen were successfully separated into separate clusters. The members of three clusters mapping to more than one annotation were clearly related within a cluster: two clusters contained two different types of collagen, while one cluster contained two types of epidermal growth factor-like domains.

Global properties of *L. gigantea* shell matrix proteome

Previous analyses of the *L. gigantea* shell matrix proteome (Mann et al. 2012; Mann & Edsinger 2014; Marie et al. 2013a) had noted a tendency for the proteins to be low complexity and disordered and that some proteins were enriched in particular residues. Here, ProminTools was used to put these observations on a more quantitative footing (*Data S8*, *Data S9*) and to discover enriched sequenced motifs in the data set from Mann *et al.* (2014), which contained 381 proteins. G and P rich motifs were found to be enriched most frequently among the proteins (*Fig. 2A*). Given that we are seeking to find the motifs that are shared within a group of proteins, Protein Motif Finder excludes motifs found in fewer than four proteins from certain plots to prevent the picture being dominated by a highly enriched motif found in very few proteins. The result can be seen in *Fig. 2B*, where Q containing motifs displayed the greatest enrichments relative to the background proteome. For example QQP was enriched 7.5 fold while Q.N.Q was enriched 6.1 fold (see data tables in *Data S8* for these numbers). In general there is often a negative correlation between the number of proteins in a set containing motifs and the enrichment of those motifs. These two measures are combined (see Materials and Methods) in *Fig. 2C*, which emphasises motifs found in a high number of proteins with high enrichment (a high PS-value). For example, GG is found in 270 proteins and is 1.8 fold enriched, G..D in 234 proteins at 1.4 fold enrichment and NG in 249 proteins at 1.53 fold enrichment (*Data S8*).

The analysis of sequence bias and amino composition bias with Sequence Properties Analyzer was concordant with the motif finding results, in that G, P, Q and A were the most enriched amino acids (*Fig. 2D*). H, I, K, L, W, E and F were found to be the most depleted relative to the background proteome. The most commonly enriched amino acids were P, G, A and C (enriched in 87, 82, 74 and 69 proteins respectively, *Fig. 2E*). Amino acid residues C and A are not found

among the most enriched motifs or the motifs with the highest PS-value, indicating that the proteins are sometimes enriched in an amino acid without that amino acid being embedded in a particular primary sequence context.

The shell matrix proteins showed a clear tendency to contain more low complexity sequence than the background proteome (Wilcoxon rank sum test, $p = 7.6 \times 10^{-6}$, *Fig. 2F*) but there was no significant tendency for the sequences to contain a greater proportion of predicted disordered sequence than the background proteome ($p = 0.1$). The shell matrix protein set contained a similar proportion of proteins with negative and positive clusters of amino acids to the background proteome (*Fig. 2H*).

Clustering of proteins based on motif content reveals relationships not found by blast searches

The Sequence Properties Analyzer carries out three hierarchical clustering analyses (*Data S8*, Materials and Methods). Eight protein clusters were identified (*Fig. 3*A), six of which contained more than two proteins. To investigate the nature of each cluster, Protein Motif Finder was re-run on each of the six main clusters (*Data S10*). Clusters 1 and 2 were riche in G containing motifs, especially NG.GG in cluster 1. Cluster 3 contained proteins rich in D containing motifs (especially D.NDD); cluster 4 in a variety of Q containing motifs; cluster 5 in C.I.P.D and C..YC..G and cluster 6 in various T and P containing motifs. By analysing the specific set of proteins in each cluster, the motifs identified are more specific to those proteins, and thus differ from the most enriched motifs in the data set as a whole displayed in *Fig. 3A*. For example D.NDD is the most prominent motif from the reanalysis of cluster 3, but is not among the most

enriched motifs in the global analysis of the entire data set, demonstrating the value of this iterative approach.

Given the results of our validation analysis with human low complexity proteins, it can be hypothesised that proteins found within the same cluster have related function. Only one of the proteins (Lotgi1|143247, cluster 5) has an annotation: a 'four disulphide core domain protein' (Pfam PF00095), suggesting that it may function as a protease inhibitor. Given the lack of annotations, it was not possible to further test the relationship between cluster membership and function using the *L. gigantea* data.

We next asked whether the motif clusters reflected larger scale sequence similarity between the proteins within a cluster. To this end, protein sequences in each cluster were subject to an all-vs-all pairwise BLASTp analysis, which is summarised in the matrices in *Fig. 3C* for six of the clusters. In general larger scale sequence similarity was low within clusters, with only three protein pairs from the five clusters displaying identity above 50% for the highest scoring HSP. This demonstrates that the clustering method can be used to find similarities that are not obvious from BLAST searches.

Discussion

Proteins are a prominent part of the organic matrices of many biominerals and are thought to have a number of roles including catalysis, templating, and control of nucleation and crystal growth. Studies of biomineral associated proteins understandably often emphasise proteins with conserved domains, which lend themselves to discussions of their possible molecular functions. However most studies also identify many proteins of unknown function, many of which appear to be low complexity in nature, with biased compositions and a high proportion of intrinsic disorder. Although authors often carefully inspect their protein sequences and note sequences

that appear particularly rich in certain residues or motifs, and note the degree of disorder, this information is rarely put in the context of the predicted proteome as a whole.

Here we introduce ProminTools, a user friendly package that allows researchers to glean more information from primary sequences of proteins of unknown function and put this in the context of the background proteome. Importantly, ProminTools allows users with minimal bioinformatic skills to run a suite of analyses and produce visualization that would otherwise require a lot of scripting. The giant limpet *L. gigantea* has a complex shell matrix proteome for which two different data sets exist. The data analysed in the present study derives from all shell layers (Mann et al. 2012, Mann & Edsinger 2014), and is thus more complex than the second data set that is derived from the aragonite shell layers only, excluding the calcitic layers (Marie et al. 2013a).

ProminTools revealed a complex array of strongly enriched motifs in the Mann et al. (2014) data set, which were not uncovered in the original study. Q, P and G rich motifs were particularly prominent and the proteins could be clustered based on their motif content even when they shared little larger scale sequence similarity. Re-running Protein Motif Finder on each of these clusters revealed unique motif profiles that could be hypothesised to be important for the molecular function of proteins in the group. For example, one group was enriched in acidic (D rich) motifs, another in Q and P rich motifs and other in G rich motifs. Interestingly, the Marie *et al.* study also identify a group of low complexity proteins rich in Q, suggesting that the functions of these proteins could be important for formation of all shell layers or just the aragonite layers, but that they are unlikely to be specific to the calcite layers.

We hypothesise that clustering protein sequences with biased composition based on their motif enrichment patterns can be used to group proteins of related function. Although this hypothesis

has yet to be confirmed on biomineral associated protein data sets, we show that this procedure can group functionally related proteins of biased composition from humans. Additional support for the idea comes from a previous study in which accurate predictors of enzyme function were built using the motif content of protein sequences (Ben-Hur et al. 2006).

We found that the shell matrix proteins as a group were significantly lower in complexity than the background proteome, providing a statistical underpinning for this observation, and supporting the conclusion of Marie *et al.* (2013a) who noted the high proportion of low complexity sequences in their data set. The Mann *et al.* studies (Mann et al. 2012; Mann & Edsinger 2014) highlight several proteins in their data which have high degrees of intrinsic disorder. Here, using the Sequence Properties Analyzer we were able to demonstrate that this is not a general feature of the data set, which is not predicted to be significantly more disordered than the background proteome. This highlights the importance of the proteome context when assigning significance to protein features, and demonstrates that the generally observed correlation between protein disorder and low complexity (Mier et al. 2019) does not hold in every data set.

The role of low complexity regions in biomineralization has only been determined in a very few cases. For example, the enamel protein Amelogenin has a central block of hydrophobic sequence rich in P, H and Q. Intramolecular hydrophobic interactions involving this regions are thought to be critical for self-assembly of Amelogenin into nanospheres and higher order structures that regulated crystal growth (Wang & Nilsen-Hamilton 2012). It is possible that the Q and P rich regions in the *L. gigantea* shell matrix proteins might have a similar role in driving self-assembly processes.

Although at present we can only speculate on the role of low complexity proteins in biomineralization, it is clear that low complexity sequences are not unique to biomineralization related proteins. Depending on the species, 22 – 36 % of residues in eukaryotic proteins fall into low complexity regions (Wootton 1994). It remains to be investigated whether the low complexity regions of biomineralization related proteins have features that set them apart from other low complexity regions in proteomes, and ProminTools could be used to investigate such questions.

ProminTools allows researchers to easily find patterns in their data, but it has limitations and judgement should be applied in interpreting the output. For example, patterns found by ProminTools can reflect technical biases as well as biological signals. Post-translational modifications of particular residues could affect peptide detectability and thus protein inference, leading to biases in the input data. It should also be remembered that ProminTools is primarily a tool for hypothesis generation. For example, proteins which share similar motifs can be hypothesised to perform similar molecular functions, but this may or may not be the case for a particular biological system, and experimental validation is required. ProminTools will be at its most useful when combined with other methods for spotting repeating patterns in sequences (e.g. HhpreID (Zimmermann et al. 2018), Meme (Bailey et al. 2009) or simply inspecting dot-plots) and when put in the context of additional information such as known domain content, post-translational modifications, phylogenetic distributions and expression patterns.

We would like to point out that ProminTools can be used for any pairwise comparison of sets of protein sequences. For examples, protein sets associated with different part of a biomineral or different developmental stages could also be compared, and if carefully carried out, cross-species comparisons could also be made. The latter could be particularly useful, since the fast evolving

nature of low complexity sequences (McDougall et al. 2013) can make it difficult to detect homology. It could also be applied to other protein sets rich in low complexity sequences, such as proteins found in pathological amyloids associated with diseases such as Alzheimer's and Parkinson's (Kumari et al. 2018).

Conclusions

ProminTools will help researchers generate new hypotheses about the important of particular motifs and protein chemistries in their system of interest and provide new directions for experimental work. Putting the patterns identified into the context of the rest of the proteome ensures that features that are genuinely overrepresented in the POIs are prioritised for further study.

Acknowledgements

We gratefully acknowledge Dr. André Scheffel for advice and critical reading of the manuscript.

Author contributions

AS conceived the study, wrote the tools, compiled the figures and wrote the manuscript. AD developed the container structure and reviewed the manuscript.

Software availability

ProminTools is available in two apps with graphical user interfaces from the Cyverse Discovery Environment (https://de.cyverse.org/). Users need to create an account, but then have access to a large number of tools and high performance computing. Docker images are most easily accessed from Docker Hub (https://hub.docker.com/repositories/biologistatsea/), but the user must have a (free) account. The images can also be downloaded from FigShare (*Data S11*, Sequence

Properties Analyzer: https://doi.org/10.6084/m9.figshare.12670817.v1 Protein Motif Finder: https://doi.org/10.6084/m9.figshare.12667070.v1) and detailed and accessible instructions on usage are available at https://github.com/skeffington/Promin-tools.

References

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren JY, Li WW, and Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research* 37:W202-W208. 10.1093/nar/gkp335

Ben-Hur A, Brutlag D. 2006. Sequence motifs: highly predictive features of protein function. *In*: Guyon I, Nikravesh M, Gunn S, Zadeh L, eds. *Feature Extraction*. 625-645 Berlin, Heidelberg: Springer.

Brendel V, Bucher P, Nourbakhsh IR, Blaisdell BE, and Karlin S. 1992. Methods and algorithms for statistical analysis of protein sequences. Proc Natl Acad Sci U S A 89:2002-2006. 10.1073/pnas.89.6.2002

Evans JS. 2019a. The biomineralization proteome: Protein complexity for a complex bioceramic assembly process. *Proteomics* 19. 10.1002/pmic.201900036

Evans JS. 2019b. Composite materials design: Biomineralization proteins and the guided assembly and organization of biomineral nanoparticles. *Materials* 12. 10.3390/ma12040581

Harrison PM. 2017. fLPS: Fast discovery of compositional biases for the protein universe. *Bmc Bioinformatics* 18. 10.1186/s12859-017-1906-3

Huerta-Cepas J, Forslund K, Pedro Coelho L, Szklarczyk D, Jensen JL, von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Molecular Biology and Evolution 34:2115-2122.

462 Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende D,
463 Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. 2019. eggNOG 5.0: a hierarchical,
464 functionally and phylogenetically annotated orthology resource based on 5090 organisms
465 and 2502 viruses. Nucleic Acids Res. 47: D309–D314. doi: 10.1093/nar/gky1085
466 Hubert LJ, and Levin JR. 1976. A general statistical framework for assessing categorical
467 clustering in free recall. *Psychological Bulletin* 83:1072-1080. 10.1037/0033-
468 2909.83.6.1072
469 Jackson DJ, Mann K, Haussermann V, Schilhabel MB, Luter C, Griesshaber E, Schmahl W, and
470 Worheide G. 2015. The *Magellania venosa* biomineralizing proteome: A window into
471 brachiopod shell evolution. *Genome Biology and Evolution* 7:1349-1362.
472 10.1093/gbe/evv074
473 Kotzsch A, Pawolski D, Milentyev A, Shevchenko A, Scheffel A, Poulsen N, Shevchenko A,
474 and Kroger N. 2016. Biochemical composition and assembly of biosilica-associated
475 insoluble organic matrices from the Diatom *Thalassiosira pseudonana*. *Journal of*
476 *Biological Chemistry* 291:4982-4997. 10.1074/jbc.M115.706440
477 Kumari B, Kumar R, Chauhan V, and Kumar M. 2018. Comparative functional analysis of
478 proteins containing low-complexity predicted amyloid regions. *PeerJ* 6.
479 10.7717/peerj.5823
480 Le Roy N, Jackson DJ, Marie B, Ramos-Silva P, and Marin F. 2014. The evolution of metazoan
481 alpha-carbonic anhydrases and their roles in calcium carbonate biomineralization.
482 *Frontiers in Zoology* 11. 10.1186/s12983-014-0075-8

Mann K, Edsinger-Gonzales E, and Mann M. 2012. In-depth proteomic analysis of a mollusc shell: acid-soluble and acid-insoluble matrix of the limpet *Lottia gigantea*. *Proteome Science* 10. 10.1186/1477-5956-10-28

Mann K, and Edsinger E. 2014. The Lottia gigantea shell matrix proteome: re-analysis including MaxQuant iBAQ quantitation and phosphoproteome analysis. *Proteome Science* 12. Artn 28

10.1186/1477-5956-12-28

Mann K, Macek B, and Olsen JV. 2006. Proteomic analysis of the acid-soluble organic matrix of the chicken calcified eggshell layer. *Proteomics* 6:3801-3810. 10.1002/pmic.200600120

Marie B, Jackson DJ, Ramos-Silva P, Zanella-Cleon I, Guichard N, and Marin F. 2013a. The shell-forming proteome of *Lottia gigantea* reveals both deep conservations and lineage-specific novelties. *Febs Journal* 280:214-232. 10.1111/febs.12062

Marie B, Ramos-Silva P, Marin F, and Marie A. 2013b. Proteomics of $CaCO_3$ biomineral-associated proteins: How to properly address their analysis. *Proteomics* 13:3109-3116. 10.1002/pmic.201300162

Marie B, Arivalagan J, Mathéron L, Bolbach G, Berland S, Marie A, Marin F. 2017. Deep conservation of bivalve nacre proteins highlighted by shell matrix proteomics of the *Unionoida Elliptio complanata* and *Villosa lienosa* . J R Soc Interface 14:pii20160846

McDougall C, Aguilera F, and Degnan BM. 2013. Rapid evolution of pearl oyster shell matrix proteins with repetitive, low-complexity domains. *Journal of the Royal Society Interface* 10. 10.1098/rsif.2013.0041

504 Merchant N, Lyons E, Goff S, Vaughn M, Ware D, Micklos D, and Antin P. 2016. The iPlant
505 Collaborative: Cyberinfrastructure for enabling data to discovery for the life sciences.
506 *Plos Biology* 14. 10.1371/journal.pbio.1002342
507 Mier P, Paladin L, Tamana S, Petrosian S, Hajdu-Soltesz B, Urbanek A, Gruca A, Plewczynski
508 D, Grynberg M, Bernado P, Gaspari Z, Ouzounis CA, Promponas VJ, Kajava AV,
509 Hancock JM, Tosatto SCE, Dosztanyi Z, and Andrade-Navarro MA. 2019. Disentangling
510 the complexity of low complexity proteins. *Brief Bioinform*. 10.1093/bib/bbz007
511 Nielsen JT, and Mulder FAA. 2019. Quality and bias of protein disorder predictors. *Scientific*
512 *Reports* 9. 10.1038/s41598-019-41644-w
513 Peng K, Radivojac P, Vucetic S, Dunker AK, and Obradovic Z. 2006. Length-dependent
514 prediction of protein intrinsic disorder. *Bmc Bioinformatics* 7. 10.1186/1471-2105-7-208
515 Plevin MJ, Mills MM, Ikura M. 2005. The LxxLL motif: a multifunctional binding sequence in
516 transcriptional regulation. *Trends Biochem Sci.* 30:66–69. doi: 10.1016/j.tibs.2004.12.001
517 Schwartz D, and Gygi SP. 2005. An iterative statistical approach to the identification of protein
518 phosphorylation motifs from large-scale data sets. *Nature Biotechnology* 23:1391-1398.
519 10.1038/nbt1146
520 Skinner HCW, and Jahren AH. 2007. Biomineralization. *Treatise on Geochemistry*, 1-69.
521 Song XR, Liu ZQ, Wang LL, and Song LS. 2019. Recent advances of shell matrix proteins and
522 cellular orchestration in marine molluscan shell biomineralization. *Frontiers in Marine*
523 *Science* 6. 10.3389/fmars.2019.00041
524 Suzuki M, Saruwatari K, Kogure T, Yamamoto Y, Nishimura Y, Kato T, Nagasawa H. 2009.
525 An Acidic matrix protein, Pif, is a key macromolecule for nacre
526 formation. *Science* 325: 1388-1390.

527 Szekely, Gabor & Rizzo, Maria. (2013). Partial distance correlation with methods for
528 dissimilarities. *The Annals of Statistics*. 42. 10.1214/14-AOS1255.
529 Wagih O, Sugiyama N, Ishihama Y, and Beltrao P. 2016. Uncovering phosphorylation-based
530 specificities through functional interaction networks. *Molecular & Cellular Proteomics*
531 15:236-245. 10.1074/mcp.M115.052357
532 Wang L, and Nilsen-Hamilton M. 2012. Biomineralization proteins: from vertebrates to bacteria.
533 *Frontiers in Biology* 8:234-246. 10.1007/s11515-012-1205-3
534 Wingender E, Schoeps T and Dönitz J. TFClass: An expandable hierarchical classification of
535 human transcription factors. *Nucleic Acids Res*. 41, D165-D170 (2013). doi:
536 10.1093/nar/gks1123
537 Wootton JC. 1994. Non-globular domains in protein sequences: automated segmentation using
538 complexity measures. *Computers & Chemistry* 18:269-285. 10.1016/0097-
539 8485(94)85023-2
540 Wootton JC, and Federhen S. 1993. Statistics of local complexity in amino-acid-sequences and
541 sequence databases. *Computers & Chemistry* 17:149-163. Doi 10.1016/0097-
542 8485(93)85006-X
543 Zimmermann L, Stephens A, Nam SZ, Rau D, Kubler J, Lozajic M, Gabler F, Soding J, Lupas
544 AN, and Alva V. 2018. A completely reimplemented MPI bioinformatics toolkit with a
545 new HHpred server at its core. *Journal of Molecular Biology* 430:2237-2243.
546 10.1016/j.jmb.2017.12.007
547

Table 1(on next page)

Annotations of the 10 largest clusters identified by ProminTools in an analysis of human proteins of biased sequence composition

Table 1:
Annotations of the 10 largest clusters identified by ProminTools in an analysis of human proteins of biased sequence composition.

Cluster ID	Cluster size	Identity (%) in all-vs-all alignments (min ; mean ; max)	eggNog annotation	Proteins in cluster carrying this annotation (%)
1	21	68.9 ; 87.7; 100	Extracellular domain of unknown function in nidogen (entactin) and hypothetical proteins	100
2	19	93.4; 96.9; 100	Collagen type XI alpha 2	100
3	14	99.8; 99.9; 100	Coiled-coil 2A	100
4	13	76.2; 88.1; 100	Collagen triple helix repeat (20 copies)	100
30	5	99.6; 99.7; 100	Nebulin repeat	100
33	5	78.0; 86.6; 99.5	Titin Z	100
39	5	57.4; 74.5; 100	Neuroblastoma breakpoint family member	100
40	5	94.6; 97.8; 100	WAS WASL interacting protein family member 1	100
41	5	67.9; 80.5; 95.9	Golgin subfamily A member	100
114	5	99.8; 99.9; 100	Neurogenic locus notch homolog protein 4	100

Figure 1

Summary of ProminTools

Proteomic datasets derived from analysis of biomineralizing organisms result in a fasta file containg the POI set which, along with the fasta file of the background proteome, make up the inputs for the ProminTools apps. The graphical interface shown is from the Cyverse Discovery Envronment (reproduced with permission). The outputs of the tools are detailed at the bottom of the figure. Attributions for the biomineral images are as follows: top left - *Vittina waigiensis* by H. Zell (licence: CC BY-SA 3.0), top right - Radiolarian skeleton by Hannes Grobe (licence: CC BY 3.0), bottom left - *Coccolithus pelagicus* by Richard Lampitt and Jeremy Young, The Natural History Museum, London (licence: CC BY 2.5), bottom right - *Lottia mesoleuca* by H. Zell (licence: CC BY-SA 3.0).

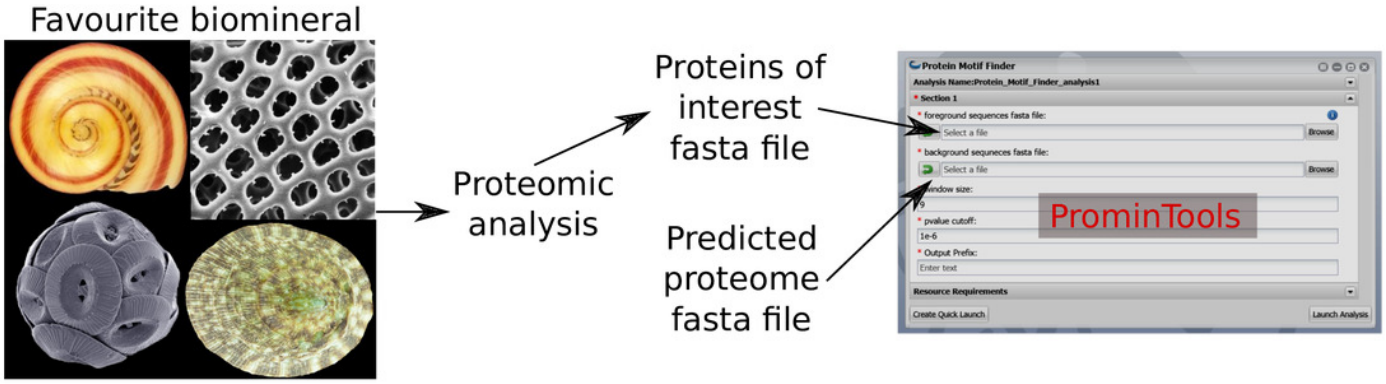


Figure 2

Properties of *L. gigantea* shell associated soluble proteome revealed by ProminTools.

Wordclouds are displayed where the height of the letter is proportional to: **A** the number of proteins in the SMP set containing the motifs, **B** the enrichment of the motif relative to the background proteome, and **C** the product of protein number and enrichment after scaling (PS-value). **D** The enrichment of amino acids in the SMP set relative to the background proteome. Values above zero indicate enrichment, and values below zero depletion. **E** The number of proteins in the SMP set enriched in each amino acid. Insert is a wordcloud summarizing the same data. **F** Density plot showing the distribution of the proportion of sequence length that is low complexity for the SMP proteins (labelled POI for Proteins Of Interest) and the background proteome. **G** Density plot showing the distribution of the proportion of sequence length that is predicted to be intrinsically disordered for the SMP proteins (POI) and the background proteome. **H** The proportion of sequences containing negatively and positively charged clusters of amino acids in the SMP proteins (POI) and the background proteome.

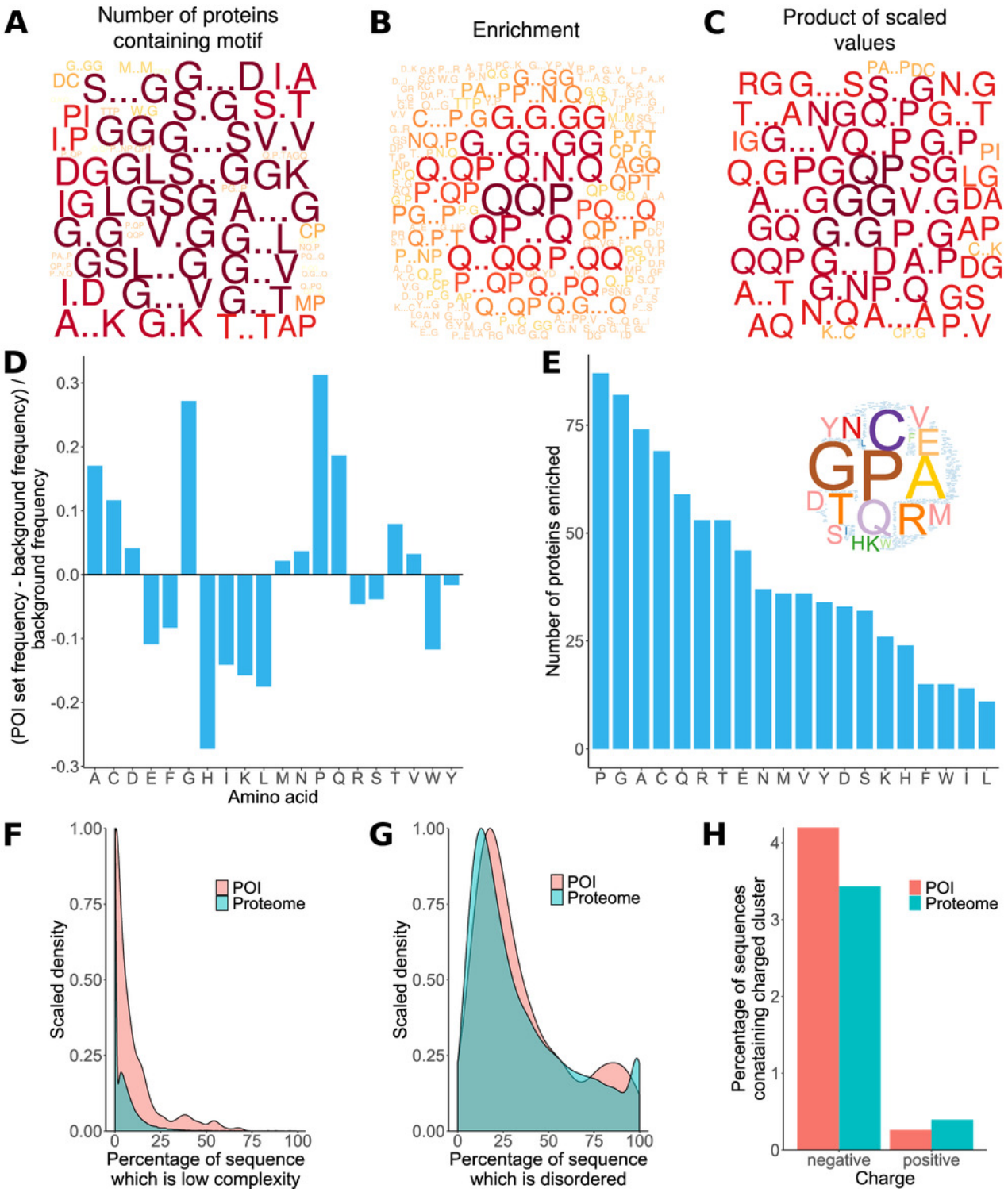
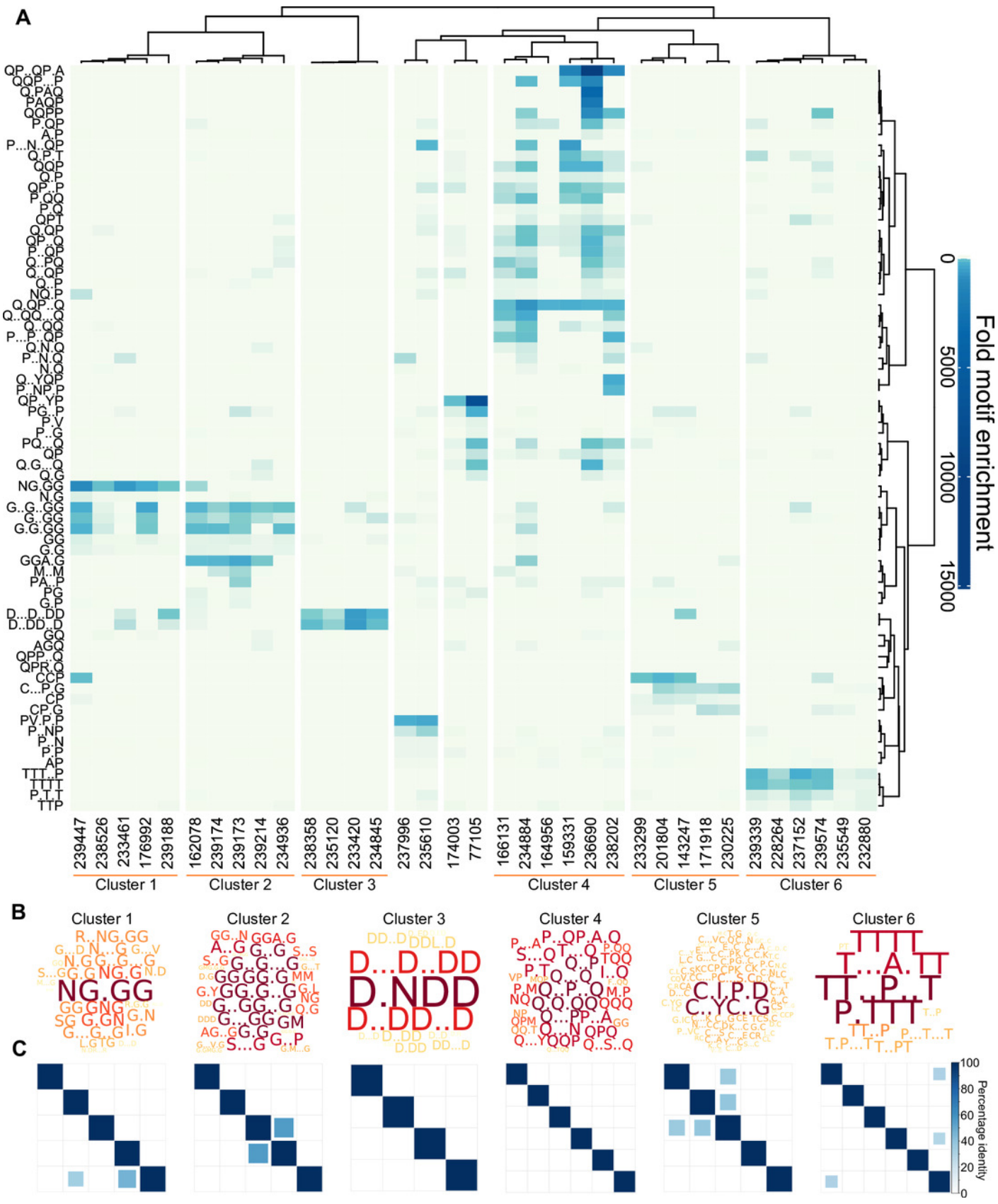

A
Number of proteins containing motif
B
Enrichment
C
Product of scaled values
D
(POI set frequency - background frequency) / background frequency
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
A C D E F G H I K L M N P Q R S T V W Y
Amino acid
E
Number of proteins enriched
75
50
25
0
P G A C Q R T E N M V Y D S K H F W I L
F
Scaled density
1.00
0.75
0.50
0.25
0.00
POI
Proteome
0 25 50 75 100
Percentage of sequence which is low complexity
G
Scaled density
1.00
0.75
0.50
0.25
0.00
POI
Proteome
0 25 50 75 100
Percentage of sequence which is disordered
H
Percentage of sequences conataining charged cluster
4
3
2
1
0
POI
Proteome
negative
positive
Charge

Figure 3

L. gigantea shell matrix proteins can be clustered based on motif content despite low sequence identity.

Figure 3: *L. gigantea* shell matrix proteins with biased composition can be clustered based on motif content despite low sequence identity. The heatmap displays (**A**) motif enrichment in the SMP set relative to the background proteome. Proteins are clustered by their motif enrichment pattern and motifs are clustered by their distribution amongst the proteins. Each motif is a row in the heatmap and each protein is a vertical column. For clusters1 - 6, a wordcloud (**B**) representing the PS-value of the enriched motifs are displayed in addition to a heatmap (**C**) representing the percentage identity between all pairs of proteins in the cluster in an all-vs-all blastp analysis (see Materials and Methods).


A
Fold motif enrichment
0
5000
10000
15000
239447 238526 233461 176992 239188
162078 239174 239173 239214 234936
238358 235120 233420 234845
237996 235610
174003
77105
166131 234884 164956 159331 236690 238202
233299 201804 143247 171918 230225
239339 228264 237152 239574 235549 232880
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
B
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
C
Percentage identity
100
80
60
40
20
0