Accurate, automatic annotation of peptidases with Hotpep-protease

Busk, Peter Kamp

Published in:
Green Chemical Engineering

DOI:
10.1016/j.gce.2020.11.008

Publication date:
2020

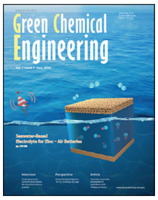
Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Busk, P. K. (2020). Accurate, automatic annotation of peptidases with Hotpep-protease. *Green Chemical Engineering*, *1*(2), 124-130. https://doi.org/10.1016/j.gce.2020.11.008

Contents lists available at ScienceDirect

Green Chemical Engineering

journal homepage: www.keaipublishing.com/en/journals/green-chemical-engineering



Accurate, automatic annotation of peptidases with hotpep-protease

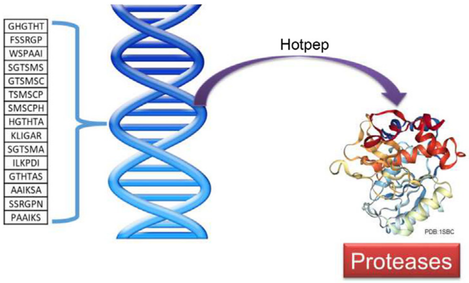
Peter Kamp Busk [*]

Department of Science and Environment, Roskilde University, Universitetsvej 1, DK-4000, Roskilde, Denmark

HIGHLIGHTS

- Examination of published annotations of proteases in fungal genomes showed that the outcome depends on the annotation method.
- Hotpep-protease is a new, k-*mer*-based method for genome-wide annotation of proteases.
- Hotpep-protease annotated human proteases with the same accuracy as the manually curated Mammalian Degradome Database.
- Hotpep-protease displayed a positive prediction rate of 0.90 compared to 0.67 for BLAST search.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Gene annotation
Proteinase
Proteolytic enzyme
Peptidase
Merops

ABSTRACT

Peptidases are essential for intracellular protein processing, signaling and homeostasis, physiological processes and for digestion of food. Moreover, peptidases are important biotechnological enzymes used in processes such as industrial food processing, leather manufacturing and the washing industry. Identification of peptidases is a crucial step in their characterization but peptidase annotation is not a trivial task due to their large sequence diversity.

In the present study short, conserved sequence profiles were generated for all peptidase families with more than four members in the comprehensive Merops peptidase database. The sequence profiles were combined with the Homology to Peptide Pattern (Hotpep) method for automatic annotation of peptidases. This method is a stand-alone software that annotates protease sequences to Merops family and subgroup and is suitable for large-scale sequence analysis. Compared to the Mammalian Degradome Database Hotpep-protease had an accuracy of 92% and a sensitivity of 96% for annotation of the human degradome. Annotation by commonly used methods (Blast and conserved domains) had an accuracy of 69% and a sensitivity of 78%. For fungal genomes, there were large differences between annotation with Hotpep-protease, Blast- and Hidden Markov Model-based annotation and the Merops annotation, which confirms the difficulty of large-scale peptidase annotation. Manual annotation indicated that Hotpep-protease had a positive prediction rate of 0.90 compared to a positive prediction rate of 0.67 for Blast search. Hence, Hotpep-protease is highly accurate method for fast and accurate annotation of peptidases.

[*] Corresponding author.
E-mail address: pbusk@ruc.dk (P. Busk).

https://doi.org/10.1016/j.gce.2020.11.008

Available online 28 November 2020

1. Introduction

Peptidases and their inhibitors are found in all organisms where they perform essential functions in protein homeostasis, intra- and extracellular signaling and digestion of food peptides [1]. Moreover, peptidases are important industrial enzymes used e.g., for food processing, biomedicine, in the washing industry and the leather industry [2–5]. Merops is a database containing 1,103,662 sequences encoding peptidases and peptidase inhibitors distributed in 465 protein families and is cross-referenced to the PANTHER database, which is the other major database for peptidases [6,7]. The database provides a description of the enzymatic and biological properties, distribution and other useful information for each family. Moreover, the sequence information in the Merops database can be used for annotation of all or a subset of proteases in a genome, e. g., by BLAST search [8] or with Hidden Markov Models (HMM) [9]. However, this option requires download of the Merops database and installation of the correct software packages or relies on online services with limited capacity for large scale annotation [6].

Annotation based on recognition of short, conserved peptides found in the members of a protein family has proven efficient for large protein classes with highly divergent sequences [10–13]. Furthermore, this approach allows for expansion of protein families to sequences that are not yet in the protein databases thereby achieving annotation of otherwise overlooked protein family members [14–17].

In the present study, short, conserved peptide profiles were generated for the 419 Merops families containing more than four different polypeptide sequences. The peptide profiles were combined with the Hotpep algorithm [18] to generate the application Hotpep-protease for identification of peptidase and peptidase inhibitor sequences and annotation to the Merops protein families. Annotation of all protease-encoding genes in the human genome and comparison to the manually curated Mammalian Degradome Database (MDD) [19,20] showed an accuracy for annotation with Hotpep-protease of 0.92. This is a clear improvement compared to annotation by Blast and conserved domain search, which had an accuracy of 0.69. Moreover, annotation of fungal protease with Hotpep-protease and verification by manual annotation suggested a positive prediction value for this approach of 0.90 ± 0.04 whereas Blast-based annotation only had a positive prediction value of 0.67 ± 0.06. These results suggested that Hotpep-protease performs better for annotation of proteases than conventional methods for large-scale annotation.

2. Methods

2.1. Sequences

The non-redundant library (pepunit.lib) of the peptidase units and inhibitor units of all the peptidases and peptidase inhibitors were downloaded from MEROPS [6].

The predicted proteins of the genomes of the fungi *Chaetomium thermophilum* (Genbank accession: GCA_000221225.1), *Penicillium chrysogenum* (Genbank accession: GCA_000710275.1), *Rhizopus delemar* (Genbank accession: GCA_000149305.1), *Talaromyces stipitatus* (Genbank accession: GCA_000003125.1) and *Thielavia terrestris* (Genbank accession: GCA_000226115.1) were downloaded from Genbank.

The *Homo sapiens* reviewed proteome (Swiss-Prot accession number up000005640) was downloaded from UniProt [21].

2.2. Identification of short, conserved peptides

The Peptide Pattern Recognition (PPR) algorithm [22] was used to identify short, conserved peptides in the Merops non-redundant library as previously described [10].

2.3. Hotpep annotation

The conserved peptide patterns were combined with the Hotpep algorithm [18] for annotation of peptidases and peptidase inhibitors from amino acid sequences as previously described [10] with minor modifications (Fig. 1). Briefly, each amino acid sequence was given a score for each peptide list that was present in the sequence by finding all the conserved peptides that were present in the amino acid sequence. A hit was considered significant if it a) included four or more conserved peptides, b) these peptides represented at least ten amino acids of the protein sequence, c) the sum of the frequency of these peptides in the protein family was higher than 1.0.

When a sequence generated a significant hit in more than one family, it was assigned to the family with the highest score.

Hotpep-protease including source code, user manual and a user guide with a detailed description of the output is available at https://sourceforge.net/projects/hotpep-protease/.

2.4. Manual annotation

Protein sequences were annotated by Blast search [23] and Conserved Domain Database search [24] followed by manually inspection of the result.

2.5. Statistical analysis

The following values were calculated for pairwise comparison of Hotpep-protease annotation to the annotation of *H. sapiens* in MDD [19,20]:

True positives = Proteins in the MDD also identified by Hotpep-protease. False positives = Proteins identified by Hotpep-protease but not in The Mammalian Degradome Database. False negatives = Proteins not found by Hotpep-protease but listed in The Mammalian Degradome Database.

Sensitivity was calculated as True positives/(True positives + False negatives); Precision (positive prediction value) was calculated as True positives/(True positives + False positives) and accuracy/F1 score (the harmonic mean of precision and sensitivity) was calculated as (2 × True positives)/(2 × True positives + False positives + False negatives).

3. Results and discussion

The library of sequences of peptidase and inhibitor units of the families in the Merops database of proteases and protease inhibitors was used as a starting point for identifying short, conserved peptides in each Merops family with the PPR algorithm [22]. The advantage of using the

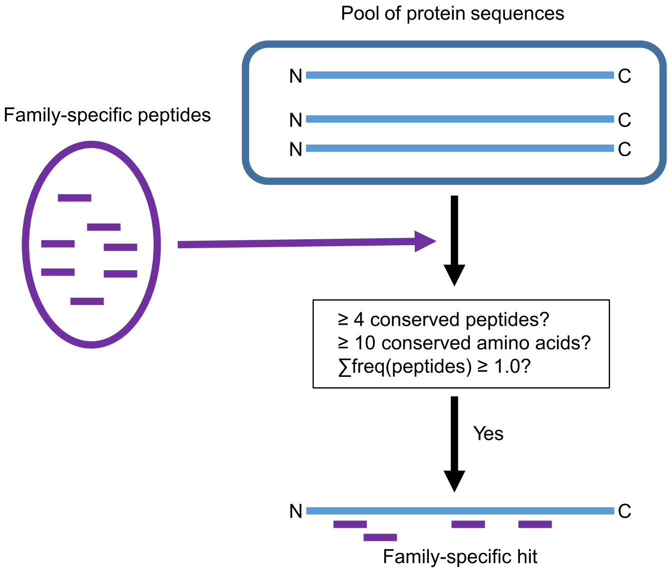


Fig. 1. Diagram of the Hotpep-protease algorithm.

protein domains carrying the Merops activity rather than the full-length sequences is that the resulting short, conserved peptides represent the peptidase- or inhibitor-encoding domains rather than sequences of associated protein domains with a non-peptidase function [10]. Some of the Merops families contain too little sequence information for annotation by HMMs [9]. Similarly, families with too few sequences to generate valid peptide patterns were excluded from the present study. PPR analysis of the rest of the families resulted in conserved sequence patterns for 419 families of peptidases and peptidase inhibitors. These conserved peptide patterns were used for the Hotpep method [18] to generate Hotpep-protease for annotation of amino acid sequences encoding peptidases and peptidase inhibitors (Fig. 1). The results for each family provides a list of hits, a link to the family description in the Merops Database and the EC number(s) of family members.

The input is a number of protein sequences in fasta format saved as a ".faa" or a ".txt" file in the same folder as "hotpep_protease.exe" and the directory "protease_patterns" including subdirectories and files.

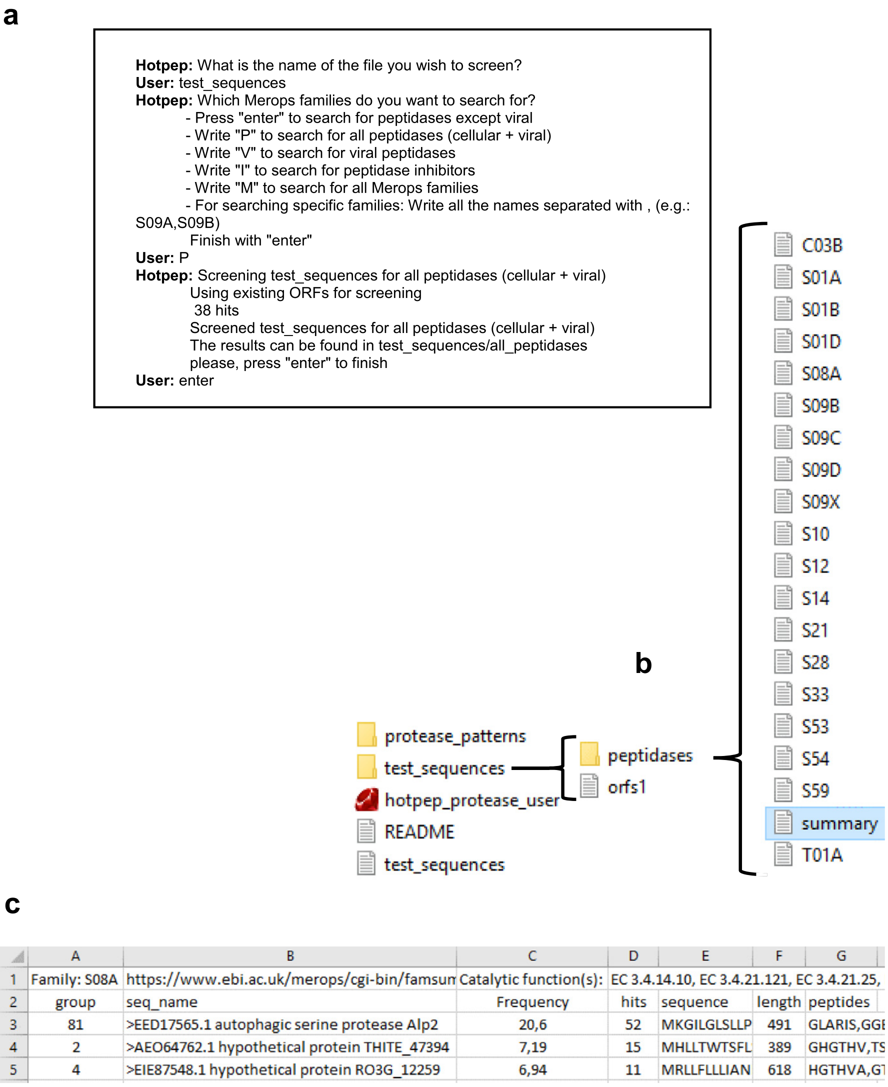
Table 1
The total list of Hotpep-protease search options.

Option	Search	Output directory
"default"	All peptidases except viral	peptidases
P	All peptidases (cellular + viral)	all_peptidases
V	Viral peptidases	viral_peptidases
I	Peptidase inhibitors	peptidase_inhibitors
M	All Merops families	Merops
Specific family names separated with comma (e.g.: S09A, S09B)	Specific families	selected_fams

Double-clicking the "hotpep_protease_user.exe" icon opens a DOS prompt, asking for the name of the input file containing the fasta-formatted protein sequences (Fig. 2a). A number of search options will be listed in the DOS window (Fig. 2a). The default option (selected by pressing "enter") is to search for all peptidases except viral (Table 1).

The results are stored in a directory with the same name as the input file (Fig. 2b). This directory contains a directory named according to the search performed (Table 1) and the search file "orf1. txt" used by Hotpep-protease.

The directory (e.g.; named "peptidases") contains a number of text files named after the Merops families where hits were found (Fig. 2b). The files are ready for import into MSExcel, LibreOffice or similar



```
Hotpep: What is the name of the file you wish to screen?
User: test_sequences
Hotpep: Which Merops families do you want to search for?
        - Press "enter" to search for peptidases except viral
        - Write "P" to search for all peptidases (cellular + viral)
        - Write "V" to search for viral peptidases
        - Write "I" to search for peptidase inhibitors
        - Write "M" to search for all Merops families
        - For searching specific families: Write all the names separated with , (e.g.:
S09A,S09B)
         Finish with "enter"
User: P
Hotpep: Screening test_sequences for all peptidases (cellular + viral)
         Using existing ORFs for screening
          38 hits
         Screened test_sequences for all peptidases (cellular + viral)
         The results can be found in test_sequences/all_peptidases
         please, press "enter" to finish
User: enter
```

	A	B	C	D	E	F	G
1	Family: S08A	https://www.ebi.ac.uk/merops/cgi-bin/famsun	Catalytic function(s):	EC 3.4.14.10, EC 3.4.21.121, EC 3.4.21.25,			
2	group	seq_name	Frequency	hits	sequence	length	peptides
3	81	>EED17565.1 autophagic serine protease Alp2	20,6	52	MKGILGLSLLP	491	GLARIS,GGE
4	2	>AEO64762.1 hypothetical protein THITE_47394	7,19	15	MHLLTWTSFL	389	GHGTHV,TS
5	4	>EIE87548.1 hypothetical protein RO3G_12259	6,94	11	MRLLFLLLIAN	618	HGTHVA,GT

Fig. 2. Hotpep-protease user interface and results. (a) User interface where the input file name is provided and the Merops families to annotate are chosen. (b) Structure of the Hotpep-protease result. Hotpep-protease generates a directory with the same name as the input sequences ("test_sequences"). A sub-directory named as shown in Supplementary Table 2 (in this case "peptidases") contains a result file for each Merops family and a file with a summary of the results (highlighted in light blue). (c) Example of Hotpep-protease output with hits for the S08A family opened in MSExcel. The first column contains information on the family. The next columns (from left to right) contain the family group of the sequence, the name of the sequence, the sum of the frequency of the conserved peptides, the number of conserved peptides, the protein sequence, length of the sequence and the sequences of the conserved peptides.

spreadsheet applications. When imported into MSExcel, the first row in each result file provides the name of the family, a link to the collected information on the family in the Merops Database and, as this is not available in the Merops database, the EC number(s) of all known family members (Fig. 2c).

Each annotated genes is listed in a single row in the spread sheet containing relevant information on the annotation: the family group where the sequence is annotated (group), the name of the sequence, the sum of the frequencies of the conserved peptides (Frequency), the number of conserved peptides (hits), the protein sequence, length of the sequence and the sequences of the conserved peptides.

In general, the higher sum of frequencies and the number of conserved hits, the more reliable is the annotation [10,18].

In addition, to the result file for each Merops family, the summary file (Fig. 2c) contains an overview of all the number of hits in each Merops families.

To test Hotpep-protease it was used to annotate all proteases in the human proteome. Hotpep found 875 genes encoding proteins belonging to 94 protease families and 578 genes encoding proteins belonging to 24 protease inhibitor families (Table S1, Supplementary data). In comparison, the MDD lists 622 human proteases and 159 protease inhibitors [19,20]. Examination of the annotations showed that the largest discrepancies between Hotpep and the MDD was for the families S33 where Hotpep-protease found 32 hits compared to five in MDD, and S71 with nine Hotpep-protease hits but no MDD hits. According to Merops, families S33 and S71 contain several human members, e.g.; the family type peptidase of S71 is *H. sapiens* self-cleaving mucin but no human S71 members are listed in the MDD. There are no available explanation as to why no S71 proteases are listed in MDD and why there are so few S33 members [19,20]. However, these two families were excluded from the analysis in order to compare Hotpep-protease exclusively to correctly annotated families in the MDD.

Moreover, Blast search [23] in GenBank [25] showed that the MDD contains a misannotated transcription factor in family C80 and a C14-encoding gene that was erroneously annotated in an earlier version of the human genome sequence. The products of these two genes were also excluded from the analysis. After curation, this resulted in 594 proteases annotated in the human degradome and 657 proteases annotated by Hotpep-protease (Table S2, Supplementary data). There was a good correlation ($R^2 = 0.968$) between the number of genes assigned to each protease family for the two annotations (Fig. 3). Assuming that the annotation of the human degradome is correct, Hotpep-protease had an accuracy of 92% and a sensitivity of 96% in predicting human proteases (Table 2). In comparison, annotation of the human degradome by automatic Blast search combined with information on domain structure [6,23,24] had an accuracy of 69% and a sensitivity of 78% in predicting human proteases (Table 1, Table S2, Supplementary data). This comparison indicated that Hotpep-protease performs a better for predicting human proteases than a commonly used standard method. For example, Hotpep-protease annotated 26% more true positive protease genes than Blast/domain search while it reported less than 1/3 the number of false positives (Table 2). The high number of false positives found by Blast/domain search can be reduced by increasing the stringency of the search. However, this will lead to more false negatives. The Blast/domain method already overlooks six times more proteases than Hotpep (Table 2) and thus, does not tolerate further increase in the number of false negatives.

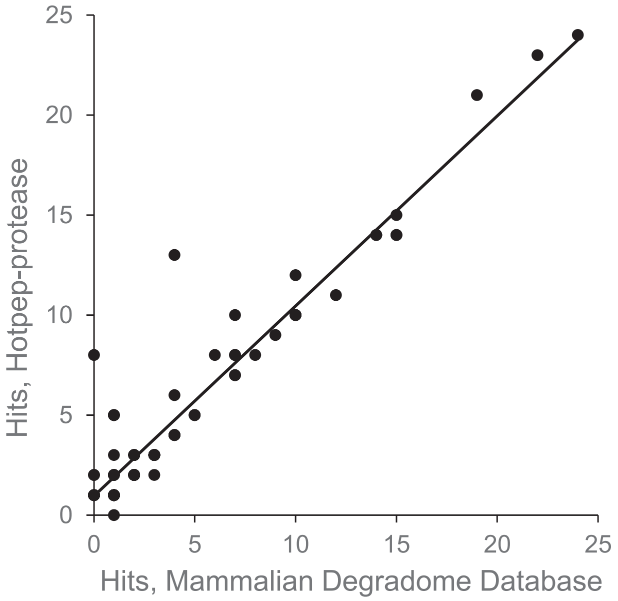


Fig. 3. Correlation between annotation of the human degradome by Hotpep-protease to the MDD. Each dot represents the number of proteases in the MDD versus the number of proteases annotated by Hotpep-protease for one Merops family. Families M12 (53 and 47 genes annotated), C19 (55 and 75 genes annotated) and S01 (133 and 123 genes annotated) were omitted for clarity. The complete data set can be found in Table S2.

Table 2
Annotation of human proteases.

	Mammalian Degradome Database	Hotpep-protease	Blast + conserved domain search
Annotated proteases	594	657	703
True positives	–	573	453
False positives	–	84	273
False negatives	–	21	131
Sensitivity	–	0.96	0.78
Precision	–	0.87	0.62
Accuracy	–	0.92	0.69

Some of the genes annotated by Hotpep-protease are true peptidase hits not annotated in the MDD but found in Merops and described as peptidases, e.g., one of the Leishmanolysin-like peptidases in family M8 (Table S2, Supplementary data). This suggests that using the MDD as reference probably overestimates the number of false positives found by Hotpep-protease. On the other hand, some Hotpep hits clearly appear to be false positives, e.g., MDD only lists four M41 family members whereas Hotpep found 13 M41 family members (Fig. 3, Table S2, Supplementary data). The nine additional sequences annotated by Hotpep contain a 15 amino acid motif shared with the M41 family but no other similarity and have been assigned other functions such as ATPase (Table S2, Supplementary data). Overall, automatic annotation of the human degradome with Hotpep yielded results similar to manual annotation and better than automatic annotation by standard methods.

Hence, Hotpep-protease is a useful tool for *de novo* annotation and reannotation of peptidases in mammalian genomes. As Hotpep-protease is based on the information in Merops it is reasonable to assume that Hotpep annotation will yield good results for other animal genomes and for plant genomes with peptidase sequences that are closely related to know Merops family members.

Hotpep-protease was generated with the sequence information from Merops. Hence, this method annotates non-peptidase members listed in Merops families but not found in MDD. An example of this is that Hotpep annotated eight sequences to Merops family M61, whereas no M61 protease family members are listed in the MDD (Fig. 3, Table S2, Supplementary data). The Hotpep hits have around 50% similarity to non-peptidase members of the Merops family M61. The Merops families are established on the basis of statistically significant sequence identity between family members [26]. There is a manual curation of the database but non-peptidase members are kept in the database if they have high sequence similarity to other family members or for the trivial fact that the mere size of Merops makes manual curation a tremendous task [6]. Thus,

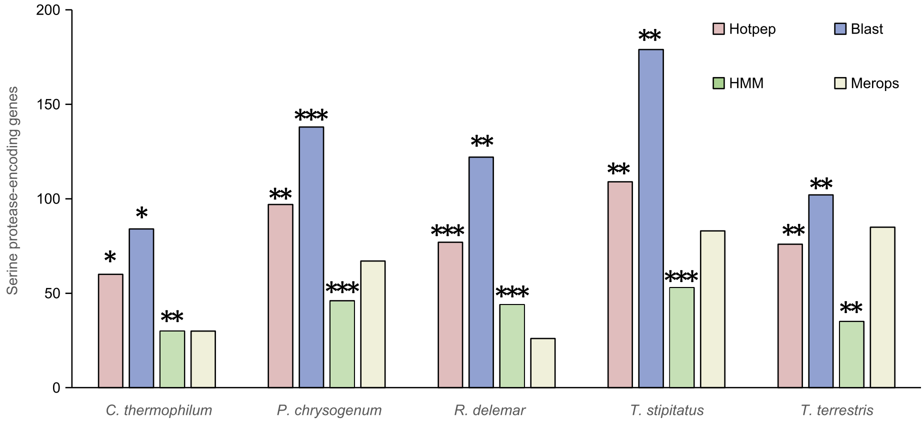


Fig. 4. Number of serine proteases annotated in five fungal genomes by Hotpep, Blast [8], Merops-based HMM models [9] and in the Merops database. Genome accession numbers are listed in Methods. The annotation for each method was compared to the Merops annotation by paired, Student's T-test of normalized data for each protein family: * indicates $P < 0.05$, ** indicates $P < 0.01$ and *** indicates $P < 0.001$.

the family M61 members found by Hotpep should be considered as true Merops hits although they are non-peptidases and therefore not included in the MDD.

There were large differences in the number of human protease inhibitor-encoding genes annotated in the MDD (159 hits), in Merops (1745 hits) and found by Hotpep-protease (578 hits) suggesting that the characterization of protease inhibitor sequences is still too limited to allow for comprehensive genome-wide annotation with sufficient accuracy (Table S1, Supplementary data).

The products of human genes have often been experimentally characterized or the gene products of similar genes in closely related organisms, other mammals, have been characterized. Hence, annotation of human genes is relatively straightforward as compared to annotation of genes of microorganisms where annotation depend on low sequence identity to characterized genes due to the large number of phylogenetically distant species and the low number of characterized genes [27]. The Hotpep method has previously proven useful for solving this kind of annotation challenge [10,12,18], hence, it was of interest to test if similar results could be obtained by Hotpep annotation of proteases in microbial degradomes. Recently, the serine proteases in the genomes of the fungi *C. thermophilum*, *P. chrysogenum*, *R. delemar*, *T. stipitatus* and *T. terrestris* were annotated independently by Blast search [8] and by HMMs [9] with Merops as reference. To compare Hotpep-protease to these methods the serine proteases in the same five genomes were annotated with the PPR-generated peptide patterns and Hotpep. The result showed a large difference in the number of serine protease-encoding genes found by the three methods and listed in the Merops database (Fig. 4, Table S3, Supplementary data). HMM models found only 208 serine proteases in the five genomes [9] whereas BLAST search found a total of 625 serine proteases [8]. The number of serine proteases in Merops (291) and found by Hotpep-protease (419) were in between. Neither the number of serine proteases found by Blast ($p = 0.008$, Student's T-test) nor by HMM ($p = 0.001$, Student's T-test) were similar to the number of serine proteases found by Hotpep (Fig. 4). The discrepancy between the methods indicate that there is a large uncertainty in genome wide protease annotation in phylogenetically distant genomes of organisms whose degradome has not been well characterized. Thus, genome wide protease annotation should be interpreted with care. The uncertainty extends to the Merops database. For example, Merops only annotates serine protease families S1, S8, S9 and S10 in *R. delemar* although this fungal genome contains genes encoding up to nine other families of serine proteases according to Hotpep, Blast and HMM (Table S3, Supplementary data). *R. delemar* belongs to the Zygomycota that are only distantly related to the more well-characterized Ascomycota [28] and have very different enzyme sequences [18] that are likely to be overlooked in automatic annotation of genomes.

It was of interest to investigate if all protease families displayed the same discrepancies in annotation as found for fungal serine proteases encoding genes (Fig. 4). To access this, all protease families in the five fungal genomes were annotated with Hotpep-protease. Hotpep found 1422 protease-encoding genes in the five genomes, 30% more than listed for the same genomes in Merops (Fig. 5a, Table S4, Supplementary data). Blast search in the five genomes found 2134 protease-encoding genes [8]. Comparison of the hits for each family to the Merops hits showed

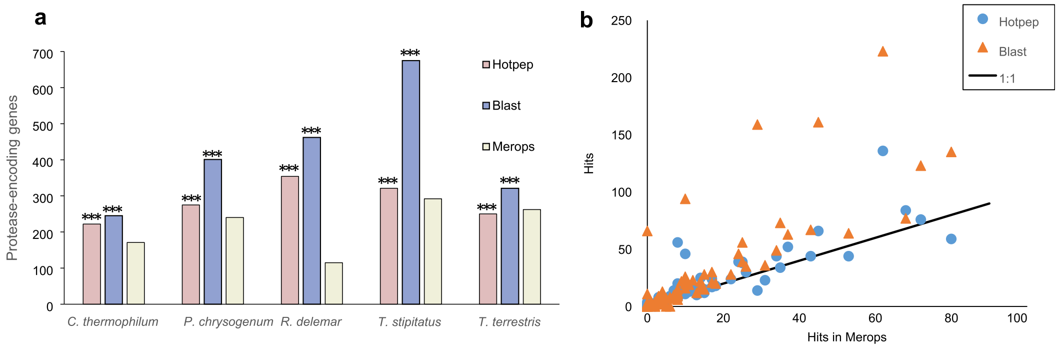


Fig. 5. (a) Number of proteases annotated in five fungal genomes by Hotpep, Blast [8] and in the Merops database. The annotation for Blast and Hotpep-protease was compared to the Merops annotation as for Fig. 4, (b) Comparison of the number of hits for each protease family. Horisontal axis: Merops, Vertical axis: Hotpep (Blue circles) or Blast (orange squares). The black line indicates a 1:1 ratio of hits on the horizontal and the vertical axis. Genome accession numbers are listed in Methods.

that the large number of Blast hits is mainly due a high number of genes annotated for a few families (Fig. 5b). For example, Blast annotated 223 genes as protease family S9 members in the five fungal genomes in contrast to only 136 members according to Hotpep-protease and 62 members according to Merops (Fig. 5b, Table S4, Supplementary data). The Blast search did not include any procedure to avoid assigning the same sequence to more than one protease family. Hence, the high number of proteases annotated may partly be due to multiple annotations of the same sequences. In support of this possibility, family S9 consists of four related subfamilies where the same sequence may easily be annotated to several of these. Hotpep-protease found 136 family S9 members. However, if correction for multiple annotations of the same gene was omitted, this increased to 235 hits in the S9 family.

Another possible cause for annotation of a high number of genes in one family is the annotation of retroviral or other viral genes; e.g. the 140 Blast annotated family A11 members in the genome of *T. stipiatus* (Table S4, Supplementary data) may be closely related viral genes. This gives rise to 159 Blast annotated family A11 members in contrast to only 29 members according to Merops (Fig. 5b, Table S4, Supplementary data).

There is no well-established reference annotation of fungal protease genes to validate the Hotpep result. Hence, to access the validity of the Hotpep annotation 60 sequences were selected at random from the 1422 genes from the five species annotated as proteases. Each of the sequences was annotated by Blast and conserved domain search [23,24] followed by manual inspection of the result. Of the 60 sequences, 54 had high sequence identity to a protease or possessed a protease domain (Table S5, Supplementary data). This result indicates that Hotpep annotation of the degradome of the five fungi had a positive prediction rate of 0.90 ± 0.04. This is higher than the positive prediction rate of 0.86 for carbohydrate-active enzymes with Hotpep [10,12,13]. The superior performance of Hotpep-protease is probably due to that the PPR patterns for proteases were based on the catalytic domains of the peptidases [6] whereas the carbohydrate-active enzyme PPR patterns were made from full-length protein sequences including other sequences than the catalytic domain [29]. According to this possibility, Hotpep-protease would only score conserved peptides that are likely to be directly involved in the catalytic function of the peptidases or in the structure of the catalytic domain. In agreement with this, it was previously shown that conserved peptides in protein families map to protein regions that are either important for catalytic function or play a crucial structural role for the protein structure [11,14–16]. For example, all functionally important amino acids except the disulfide bridge in the enzyme TaGH5 are identified by conserved peptides [11]. In a more general study of a family of α/β-hydrolase fold enzymes it was found that conserved peptides map to the catalytic domain and identify the catalytic triad Glu/Asp-His-Ser [14].

To access the positive prediction rate of conventional methods for gene annotation 60 sequences predicted by Blast search to encode proteases [8] were randomly selected and annotated manually as described above. Of the 60 sequences, 40 had high sequence identity to a protease or possessed a protease domain (Table S6, Supplementary data). This result indicated that Blast annotation of the degradome of the five fungi had a positive prediction rate of 0.67 ± 0.06 and that the high number of Blast annotated genes compared to other methods is due to a relatively high number of false positives. Comparison of the two methods indicate that Hotpep-protease should be the preferred method for fungal protease annotation due to a higher rate of true positives.

Interestingly, the positive prediction rate of Hotpep-protease indicated that 377 ± 15 of the 419 peptidase-encoding genes found by Hotpep in *C. thermophilum*, *P. chrysogenum*, *R. delemar*, *T. stipitatus* and *T. terrestris* are true positives. This suggests that the HMM and the Merops annotation underestimates the number of fungal proteases by at least 45 and 23%, respectively. When mining genomes for proteases with putative industrial applications it is important to detect proteases with sequences distinct from known industrial enzymes [2–5]. Hence, the high positive discovery rate of Hotpep-protease makes this approach especially useful for finding novel proteases of industrial relevance.

The performance of the Hotpep algorithm for protease annotation is in agreement with that k-mer based methods like Hotpep perform better for annotation than homology search or HMM [13]. A relevant future application of this method would be to generate conserved peptide patterns for the families in the ESTHER database of the α/β-hydrolase fold superfamily of proteins [30] and implement them for annotation of the many families with an α/β-hydrolase fold. An approach more focused on function than on tertiary structure would be to classify all the enzyme families in a dedicated enzyme database such as BRENDA [31]. Proteins with completely different fold and thus unrelated primary structure and conserved peptides can evolve the same enzymatic activity [32]. Molecular convergent evolution is not uncommon in biology and does not only include single enzymes but extent to complete biochemical pathways such as the remarkable case of convergent evolution of the four step biosynthesis of caffeine from xanthosine in coffee, cacao and tea [33]. However, such events occur a limited number of times and can easily be handled by the Hotpep algorithm as unrelated sets of conserved peptide patterns that identify enzymes with the same function [18]. Curated and benchmarked implementations of Hotpep with conserved peptides for ESTHER, BRENDA or similar databases would be useful tools for fast, simple and reliable annotation of the protein families and enzymes in these databases. To this end, the source code of both PPR [22] and Hotpep (this study) are available.

4. Conclusion

Hotpep-protease is a well-performing method for annotation of peptidases and for degradome-annotation, especially of genomes encoding proteases that are phylogenetically distant from consensus protease sequences.

The software can be used on a desktop computer and is available as source code for modification and improvement.

Availability and requirements

Project name: Hotpep-protease.

Project home page: https://sourceforge.net/projects/hotpep-protease/

Operating systems: Windows 7 or higher.

Programming language: Ruby 2.2.4.

License: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Any restrictions to use by non-academics: Commercial rights reserved.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

I thank Dr. Andre Rodrigues for sharing raw data on Blast annotations of fungal genomes. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gce.2020.11.008.

References

[1] J.S. Bond, Proteases: history, discovery, and roles in health and disease, J. Biol. Chem. 294 (2019) 1643–1651.

[2] L. Lange, Y. Huang, P.K. Busk, Microbial decomposition of keratin in nature-a new hypothesis of industrial relevance, Appl. Microbiol. Biotechnol. 100 (2016) 2083–2096.

[3] A. Razzaq, S. Shamsi, A. Ali, Q. Ali, M. Sajjad, A. Malik, M. Ashraf, Microbial proteases applications, Front Bioeng Biotechnol 7 (2019) 110.

[4] M.A. Hassan, D. Abol-Fotouh, A.M. Omer, T.M. Tamer, E. Abbas, Comprehensive insights into microbial keratinases and their implication in various biotechnological and industrial sectors: a review, Int. J. Biol. Macromol. 154 (2020) 567–583.

[5] S. Thapa, H. Li, J. OHair, S. Bhatti, F.-C. Chen, K.A. Nasr, T. Johnson, S. Zhou, Biochemical characteristics of microbial enzymes and their significance from industrial perspectives, Mol. Biotechnol. 61 (2019) 579–601.

[6] N.D. Rawlings, A.J. Barrett, P.D. Thomas, X. Huang, A. Bateman, R.D. Finn, The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database, Nucleic Acids Res. 46 (2018) D624–D632.

[7] H. Mi, A. Muruganujan, D. Ebert, X. Huang, P.D. Thomas, PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools, Nucleic Acids Res. 47 (2019) D419–D426.

[8] T.B. de Oliveira, C. Gostincar, N. Gunde-Cimerman, A. Rodrigues, Genome mining for peptidases in heat-tolerant and mesophilic fungi and putative adaptations for thermostability, BMC Genom. 19 (2018) 152.

[9] A. Muszewska, M.M. Stepniewska Dziubinska, K. Steczkiewicz, J. Pawlowska, A. Dziedzic, K. Ginalski, Fungal lifestyle reflected in serine protease repertoire, Sci. Rep. 7 (2017) 9147.

[10] P.K. Busk, B. Pilgaard, M.J. Lezyk, A.S. Meyer, L. Lange, Homology to peptide pattern for annotation of carbohydrate-active enzymes and prediction of function, BMC Bioinf. 18 (2017) 214.

[11] P.K. Busk, L. Lange, Function-based classification of carbohydrate-active enzymes by recognition of short, conserved peptide motifs, Appl. Environ. Microbiol. 79 (2013) 3380–3391.

[12] H. Zhang, T. Yohe, L. Huang, S. Entwistle, P. Wu, Z. Yang, P.K. Busk, Y. Xu, Y. Yin, dbCAN2: a meta server for automated carbohydrate-active enzyme annotation, Nucleic Acids Res. 46 (2018) W95–W101.

[13] J. Xu, H. Zhang, J. Zheng, P. Dovoedo, Y. Yin, eCAMI: simultaneous classification and motif identification for enzyme annotation, Bioinformatics 36 (2020) 2068–2075.

[14] J.W. Agger, P.K. Busk, B. Pilgaard, A.S. Meyer, L. Lange, A new functional classification of glucuronoyl esterases by peptide pattern recognition, Front. Microbiol. 8 (2017) 309.

[15] P.K. Busk, L. Lange, Classification of fungal and bacterial lytic polysaccharide monooxygenases, BMC Genom. 16 (2015) 368.

[16] A.S. Godoy, C.S. Pereira, M.P. Ramia, R.L. Silveira, C.M. Camilo, M.A. Kadowaki, L. Lange, P.K. Busk, A.S. Nascimento, M.S. Skaf, I. Polikarpov, Structure, computational and biochemical analysis of PcCel45A endoglucanase from Phanerochaete chrysosporium and catalytic mechanisms of GH45 subfamily C members, Sci. Rep. 8 (2018) 3678.

[17] M. Wilding, M. Nachtschatt, R. Speight, C. Scott, An improved and general streamlined phylogenetic protocol applied to the fatty acid desaturase family, Mol. Phylogenet. Evol. 115 (2017) 50–57.

[18] P.K. Busk, M. Lange, B. Pilgaard, L. Lange, Several genes encoding enzymes with the same activity are necessary for aerobic fungal degradation of cellulose in nature, PloS One 9 (2014) e114138.

[19] J.G. Pérez-Silva, Y. Español, G. Velasco, V. Quesada, The Degradome database: expanding roles of mammalian proteases in life and disease, Nucleic Acids Res. 44 (2016) D351–D355.

[20] V. Quesada, G.R. Ordóñez, L.M. Sánchez, X.S. Puente, C. López-Otín, The Degradome database: mammalian proteases and diseases of proteolysis, Nucleic Acids Res. 37 (2009) D239–D243.

[21] UniProt Consortium, UniProt: a worldwide hub of protein knowledge, Nucleic Acids Res. 47 (2019) D506–D515.

[22] P.K. Busk, Peptide Pattern Recognition for High-Throughput Protein Sequence Analysis and Clustering, BioRxiv (2018) 181917.

[23] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.

[24] S. Lu, J. Wang, F. Chitsaz, M.K. Derbyshire, R.C. Geer, N.R. Gonzales, M. Gwadz, D.I. Hurwitz, G.H. Marchler, J.S. Song, N. Thanki, R.A. Yamashita, M. Yang, D. Zhang, C. Zheng, C.J. Lanczycki, A. Marchler-Bauer, CDD/SPARCLE: the conserved domain database in 2020, Nucleic Acids Res. 48 (2020) D265–D268.

[25] D.A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, E.W. Sayers, GenBank, Nucleic Acids Res. 41 (2013) D36–D42.

[26] N.D. Rawlings, A.J. Barrett, A. Bateman, Using the MEROPS database for proteolytic enzymes and their inhibitors and substrates, Current Protocols in Bioinformatics 48 (2014) 1.25.1-1.25.33.

[27] E. McDonnell, K. Strasser, A. Tsang, Manual gene curation and functional annotation, Methods Mol. Biol. 1775 (2018) 185–208.

[28] M. Richardson, The ecology of the Zygomycetes and its impact on environmental exposure, Clin. Microbiol. Infect. 15 (2009) 2–9.

[29] V. Lombard, H. Golaconda Ramulu, E. Drula, P.M. Coutinho, B. Henrissat, The carbohydrate-active enzymes database (CAZy) in 2013, Nucleic Acids Res. 42 (2014) D490–D495.

[30] N. Lenfant, T. Hotelier, E. Velluet, Y. Bourne, P. Marchot, A. Chatonnet, ESTHER, the database of the α/β-hydrolase fold superfamily of proteins: tools to explore diversity of functions, Nucleic Acids Res. 41 (2013) D423–D429.

[31] L. Jeske, S. Placzek, I. Schomburg, A. Chang, D. Schomburg, BRENDA in 2019: a European ELIXIR core data resource, Nucleic Acids Res. 47 (2019) D542–D549.

[32] G.J. Davies, M.L. Sinnott, Sorting the diverse: the sequence-based classifications of carbohydrate-active enzymes, Biochem. J. 275 (2008) 382–392.

[33] F. Denoeud, L. Carretero-Paulet, A. Dereeper, G. Droc, R. Guyot, M. Pietrella, C. Zheng, A. Alberti, F. Anthony, G. Aprea, J.-M. Aury, P. Bento, M. Bernard, S. Bocs, C. Campa, A. Cenci, M.-C. Combes, D. Crouzillat, C. Da Silva, L. Daddiego, F. De Bellis, S. Dussert, O. Garsmeur, T. Gayraud, V. Guignon, K. Jahn, V. Jamilloux, T. Joët, K. Labadie, T. Lan, J. Leclercq, M. Lepelley, T. Leroy, L.-T. Li, P. Librado, L. Lopez, A. Muñoz, B. Noel, A. Pallavicini, G. Perrotta, V. Poncet, D. Pot, null Priyono, M. Rigoreau, M. Rouard, J. Rozas, C. Tranchant-Dubreuil, R. VanBuren, Q. Zhang, A.C. Andrade, X. Argout, B. Bertrand, A. de Kochko, G. Graziosi, R.J. Henry, null Jayarama, R. Ming, C. Nagai, S. Rounsley, D. Sankoff, G. Giuliano, V.A. Albert, P. Wincker, P. Lashermes, The coffee genome provides insight into the convergent evolution of caffeine biosynthesis, Science 345 (2014) 1181–1184.