BIOINFORMATICS ORIGINAL PAPER
Vol. 30 no. 15 2014, pages 2137–2141
doi:10.1093/bioinformatics/btu195

Sequence analysis
Advance Access publication April 11, 2014

Computational identification of natural peptides based on analysis of molecular evolution

Amir Toporik[1,2,†], Itamar Borukhov[2,†], Avihay Apatoff[2], Doron Gerber[1,*] and Yossef Kliger[2,*]
[1]The Mina and Everard Goodman Faculty of Life Sciences, Bar Ilan University, 52900 Ramat-Gan and [2]Compugen Ltd., 69512 Tel Aviv, Israel

Associate Editor: John Hancock

ABSTRACT

Motivation: Many secretory peptides are synthesized as inactive precursors that must undergo post-translational processing to become biologically active peptides. Attempts to predict natural peptides are limited by the low performance of proteolytic site predictors and by the high combinatorial complexity of pairing such sites. To overcome these limitations, we analyzed the site-wise evolutionary mutation rates of peptide hormone precursors, calculated using the Rate4Site algorithm.

Results: Our analysis revealed that within their precursors, peptide residues are significantly more conserved than the pro-peptide residues. This disparity enables the prediction of peptides with a precision of ∼60% at a recall of 40% [receiver-operating characteristic curve (ROC) AUC 0.79]. Subsequently, combining the Rate4Site score with additional features and training a Random Forest classifier enable the prediction of natural peptides hidden within secreted human proteins at a precision of ∼90% at a recall of 50% (ROC AUC 0.96). The high performance of our method allows it to be applied to full secretomes and to predict naturally occurring active peptides. Our prediction on *Homo sapiens* revealed several putative peptides in the human secretome that are currently unannotated. Furthermore, the unique expression of some of these peptides implies a potential hormone function, including peptides that are highly expressed in endocrine glands.

Availability and implementation: A pseudocode is available in the Supplementary information.

Contact: doron.gerber@biu.ac.il or kliger@cgen.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 6, 2014; revised on March 23, 2014; accepted on April 5, 2014

1 INTRODUCTION

Many secretory proteins and peptides are initially synthesized as larger precursors, usually in the form of pre-pro-proteins. (Note that the term peptide is sometimes used also to describe a protein fragment. This study, however, focuses on endogenous peptides that are synthesized by translation of messenger RNA (mRNA) followed by proteolysis to generate the mature form.) Such precursor proteins undergo post-translational proteolysis: the N-terminal pre-region, known as signal peptide, is cleaved by a well-characterized signal peptidase (Dalbey and Von Heijne, 1992; Paetzel *et al.*, 2002), whereas various proteases liberate the active peptides from the pro-proteins (Seidah and Chretien, 1999). They are often subjected to proteolysis to generate the mature form.

Examples of peptide hormones, whose proteolytic processing regulates their activities, include insulin, somatostatin, parathyroid hormone, glucagon and GLP-1. Many of these are used as therapeutic peptides for treating various disorders.

The importance of identifying mature peptides fuels both experimental and computational approaches aimed at discovering and predicting proteolytic sites. Traditionally, the main method for discovering short peptides, such as hormones or GPCR ligands, was performed through biochemical separation coupled with functional assays. This approach is challenging, as peptides often have restricted expression, exist at low abundance in complex biological matrices together with other secreted and shedded proteins and have highly diverse physiological functions. In recent years, there have been several attempts based on biochemical separation and purification coupled with various mass spectrometry techniques (Kalkum *et al.*, 2003; Ohyama *et al.*, 2008).

Today, prediction of natural peptides is possible because of advancements in computational biology algorithms enabling the analysis of the fast-growing genome and proteome databases. Nevertheless, many computational approaches are limited to the classical furin and dibasic proteolytic sites, which can be identified through the use of simple regular expressions (Shi *et al.*, 2012; Shichiri *et al.*, 2003), whereas predictors aiming to treat more complex pro-hormone convertase cleavage sites have relatively low performance (Duckert *et al.*, 2004; Hummon *et al.*, 2003; Kliger *et al.*, 2008), which leads to many falsely predicted peptides. Approaches where putative peptides are computationally identified directly, based on a hidden Markov model or machine learning, are limited to cases where sequence similarity to known peptides is significant (Mirabeau *et al.*, 2007; Shemesh *et al.*, 2008; Sonmez *et al.*, 2009). We speculate that sequence and structural divergence impairs the computational identification of many peptides using this approach. Thus, we looked for methods that can circumvent these obstacles.

In this study, we test the hypothesis that the amino acid sequence of functional peptides is more conserved (among orthologous proteins) in comparison with the rest of the precursor sequence. This is true in several well-known examples, like corticoliberin, tachykinin-3 and insulin (Fig. 1). Using a secretome-wide analysis, we demonstrate that the relative conservation of peptide sequences is high enough to allow prediction of peptides

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

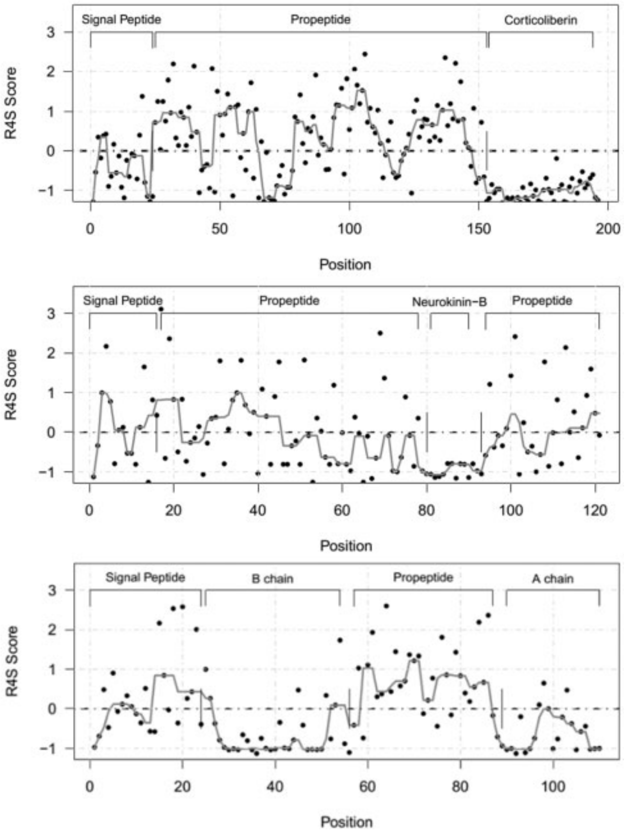

Fig. 1. Evolutionary profiles of known peptides. The Rate4Site score is plotted as a function of the residue position (dots) together with a smoothed curved (solid gray) to guide the eye based on a running median of length 3 using the '3RS3R' method in R's smooth function (Tukey, 1977). Top panel: corticoliberin (CRF_HUMAN), where the corticoliberin peptide at 154–194 is conserved. Middle panel: tachykinin-3 (TKNK_HUMAN), where neurokinin-B at 81–90 is conserved. Bottom panel: insulin (INS_HUMAN), where chains A and B are conserved, but the pro-peptide between them, which has no known biological function, is not

with reasonable performance. A machine learning approach, which combines additional features, achieves even higher performance and highlights several predicted peptides as potential novel hormones.

2 METHOD

2.1 Precursor screening and analysis

Human protein sequences were downloaded from the Uniprot knowledge base (Release 2013_05, May 1, 2013). Signal peptides were predicted using SignalP version 3.0 (Bendtsen *et al*., 2004; Nielsen *et al*., 1997), and the presence and positions of transmembrane domains were predicted using TMHMM version 2.0 (Krogh *et al*., 2001; Sonnhammer *et al*., 1998). The 'human secretome' was constructed as the list of human proteins possessing a signal peptide and no transmembrane domain.

2.2 Data preparation

For each secreted protein (from UniProt), orthologs were collected using reciprocal Blast queries. A version of the NCBI non-redundant protein sequence database (from March 18, 2012) cleaned from synthetic constructs and partial sequences was compiled. Reciprocal Blast was used to eliminate the false identification as orthologs of non-orthologous proteins that share relatively high sequence similarity and was implemented in two rounds of Blast: the first round collects the top hits to the query protein with scores $\geq$90% of the highest score (for each organism, separately). The second round selects one ortholog in each organism only if its best hit in *Homo sapiens* is the original query protein. Multiple sequence alignment (MSA) of the orthologs was performed using MAFFT version 6.240 (Katoh and Standley, 2013; Katoh *et al*., 2002), and relative amino acid substitution evolutionary rate of each position was calculated using Rate4Site version 3.0.0 (Mayrose *et al*., 2004; Pupko *et al*., 2002) or entropy. Rate4Site assigns a variability score for each position in the MSA using an empirical Bayesian inference, which considers the topology and branch lengths of the phylogenetic tree generated based only on the amino acids in this position, as well as the underlying stochastic process.

2.3 Peptide prediction

The positions of proteolytic sites in secreted human proteins were predicted using the Random Forest (RF) classifiers described in our previous work (Kliger *et al*., 2008). Sites that are not conserved in at least five orthologs were excluded. Presumed peptides were taken as segments of 3–70 residues located either between predicted two cleavage sites or between a predicted cleavage site and one of the pro-protein termini. For each precursor and potential peptide, the mean and standard deviation of the Rate4Site score were calculated and used as features for a peptide predictor. Additional features include the precursor and peptide lengths, the number of cysteine residues and its parity and two flags indicating whether the peptide starts at the signal peptide cleavage site or ends at the protein C-terminus. The peptide predictor was based on the RandomForest algorithm (Breiman, 2001) implemented in R (Team, 2005). The predictor was trained on a positive set of known peptides extracted from SwissProt. Segments of the same length distribution from other secreted proteins, which are not known to be proteolytically processed, were used as a negative set.

2.4 Tissue specificity

mRNA expression across different tissues was analyzed using MED (Erez *et al*., 2004; Helpman *et al*., 2009). Protein expression was based on ProteinAtlas (www.proteinatlas.org).

3 RESULTS

We tested the hypothesis that there are more constraints on the primary structure (i.e. sequence) of functional peptides relative to that of non-functional segments from the same precursor. To start, we analyzed the relative sequence variability per position of a few well-known examples using Rate4Site (Fig. 1). This analysis reveals that the functional peptide corticoliberin, which is proteolytically cleaved from the precursor CRH, is more conserved than the pro-peptide (Fig. 1, top panel). Similarly, the functional peptide neurokinin-B is more conserved in comparison with the rest of the TAC-3 precursor (Fig. 1, middle panel). In a more complex example, the functional segments of insulin (chains A and B) are significantly more conserved than the pro-peptide (chain C) that resides between them (Fig. 1, bottom panel). One important point to note is that irrespective of the peptide's location within the precursor, we can see that this conservation is still observed. For example, corticoliberin is at the most C-terminal part of its precursor, neurokinin-B resides between non-functional pro-peptides and insulin is

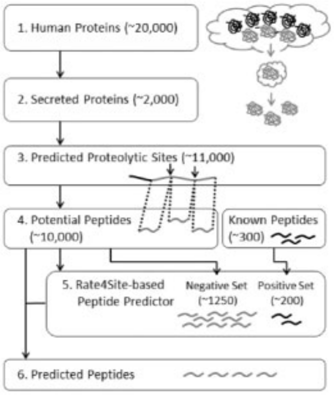
1. Human Proteins (~20,000)
2. Secreted Proteins (~2,000)
3. Predicted Proteolytic Sites (~11,000)
4. Potential Peptides (~10,000)
Known Peptides (~300)
5. Rate4Site-based Peptide Predictor
Negative Set (~1250)
Positive Set (~200)
6. Predicted Peptides

Fig. 2. Schematic view of the sequential analysis

composed of two peptides; nevertheless, the same phenomenon is observed in all these examples.

To test whether this phenomenon is strong enough to distinguish between functional peptides and other segments that reside between two predicted proteolytic sites, we followed the scheme presented in Figure 2. Analysis of the average Rate4Site scores of known peptides (R4S_in) reveals that a fraction of them are significantly more conserved (have low Rate4Site scores) when compared with the negative set (Fig. 3, left panel). Although this result indicates that it is possible to computationally discover functional peptides, the counts histogram of the same two datasets (Fig. 3, right panel) points that this is not a trivial task because the peptides are masked by many non-peptidic segments that reside between two predicted proteolytic sites and are conserved owing to other evolutionary constraints, like functional sites. Nevertheless, sorting potential peptides solely by their R4S_in scores achieves a reasonable predictive performance (Fig. 4, top panels, solid line), with a precision of ~60% at a recall of 40% [receiver-operating characteristic curve (ROC) AUC 0.79]. This naïve Rate4Site-based score performs better than an entropy-based score (Fig. 4, top panels, dotted line; a precision of ~45% at a recall of 40%). This fortifies our decision to favor Rate4site over a simpler entropy calculation. Rate4Site acts as a better predictor of local conservation compared with entropy because of the fact that its estimations of residue mutation rates take into account the global divergence between different orthologs.

Performance was significantly improved (Fig. 4, bottom panels) when an RF classifier was trained with R4S_in and additional features, achieving a precision of ~90% at a recall of 50% (ROC AUC 0.96). The added features included the standard deviation of R4S_in, the precursor and predicted peptide lengths, the number of cysteine residues and its parity in the predicted peptide, whether the predicted peptide starts immediately after the signal peptide and whether the predicted peptide ends at the end of the protein precursor.

The classifier was trained with a positive set consisting of known functional peptides that also exist in the list of putative list of peptides and a sample of the negative set that included only one randomly selected potential peptide from each precursor (Fig. 2; pseudocode is provided in the Supplementary Material). Because the classifier performance is high enough to enable computational discovery of functional peptides, we then ran the classifier over the entire putative negative set in an attempt to identify new peptides.

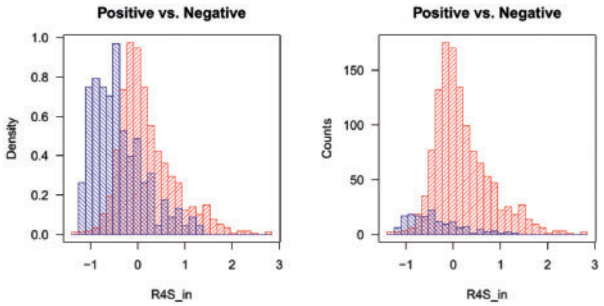

Fig. 3. Histograms of the probability density (left) and counts (right) for the average Rate4Site scores within the peptides belonging to the positive set (blue descending diagonal hatching) and negative set (red ascending diagonal hatching)

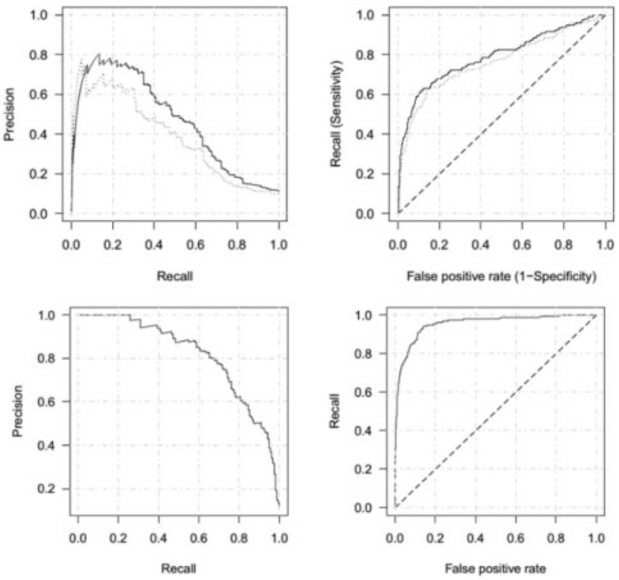

Fig. 4. Top panel: performance analysis of a naïve score based only on the average mutation rate calculated by Rate4Site (solid line) or by entropy (dotted line) within the peptide. Bottom panel: performance analysis of an RF classifier with 5-fold cross-validation. Recall-precision (left) and ROC (right) curves are presented

The performance evaluation presented in Figure 4 reflects the predictor's performance on the dataset used, and therefore, it is inaccurate because of the following reasons: (i) unknown peptides are labeled 'negative', and therefore, there is an underestimation of the real precision; (ii) approximately one-third of the peptides annotated in SwissProt are not labeled 'positive' because they involve proteolytic sites that are not predicted by the cleavage site predictor (e.g., non-R/K sites), and this causes overestimation of the recall; (iii) to remove bias from the learning set, we retained in the negative set only one case from each precursor and none from precursors with a known peptide, leading to an overestimation of the precision.

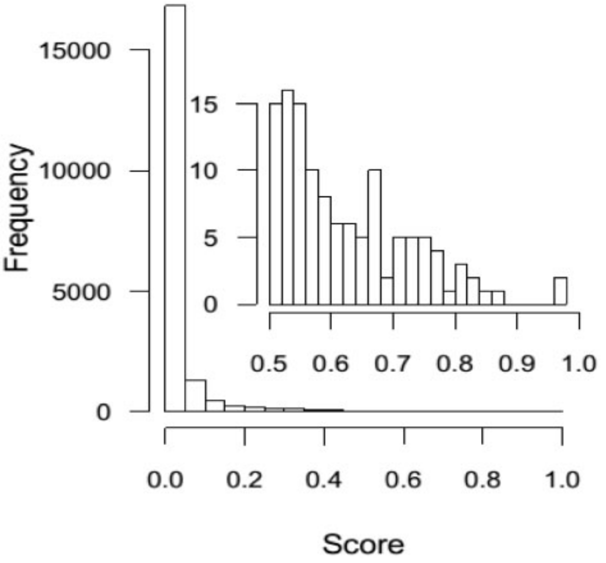

Fig. 5. Histogram of classifier scores for putative peptides. The inset focuses on the high-score cases

Dozens of predicted peptides were highly scored by the classifier (Fig. 5). Manual curation of the top scores resulted in Supplementary Table S1, which also describes the mRNA and protein expression of the precursors. The high expression in endocrine glands of some of the predicted peptides supports their authenticity.

4 DISCUSSION

To the best of our knowledge, usage of evolutionary profile to predict functional native peptides has not been studied before, although it has been implemented in the past to uncover functional residues (Lopez *et al.*, 2007, 2011) and functional patches on protein surfaces (Ashkenazy *et al.*, 2010; Nimrod *et al.*, 2005). We can speculate that others have not chosen this approach for peptide identification because of the concern that functional native peptides could not be distinguished from functional segments or domains within proteins (which reside between falsely predicted proteolytic sites) on the basis of their relative sequence conservation. Our study found that in many cases peptides can be separated from functional segments. A reasonable explanation for this success is that although there are few constraints on sequences that are removed during peptide maturation (pro-peptides), there are many more constraints applied to sequences adjacent to functional segments and domains, as these segments are part of a mature protein.

Sequence-based methods for peptide identification, whether based on HMM or other techniques, can only find novel peptides with some sequence similarity to known ones. One of the advantages of the approach described in this study is that its learning phase is sequence-independent and therefore can identify peptides that have no sequence similarity to any known peptide. For example, we predicted the existence of a few peptides in the CF057_HUMAN and F150B_HUMAN precursors (Fig. 6). In both cases, there is no sequence similarity to any known peptide. In CF057_HUMAN, we predict that there is a proteolytic site between residues 104 and 105 (RKGR104|C) followed by the removal of R104 by carboxypeptidase E (Day *et al.*, 1998; Friis-Hansen *et al.*, 2001) and amidation of K102. Note that C-terminal amidation is a well-known post-translational modification that increases the stability of many native peptides (Bradbury *et al.*, 1982). There are two options for the N-terminus of this potential peptide—G90 and Y96—which cannot be easily distinguished by this approach. According to ProteinAtlas, the precursor of this potential peptide is highly expressed in the testis' Leydig cells, the placenta's trophoblastic cells, melanocytes, some neuronal cells and glandular cells of various tissues including gall bladder, digestive tract, uterus, fallopian tube, parathyroid gland and adrenal gland. For F150B_HUMAN, we predict a peptide at the C-terminus (Fig. 6, bottom panel). It is an interesting peptide because it possesses two cysteines that might form a stabilizing intra-peptide disulphide bridge. The expression of this precursor in the adrenal gland, duodenum and small intestine glandular cells supports the functionality of this peptide as genuine. It is noteworthy that a similar predicted peptide appears in the paralogous F150A_HUMAN precursor.

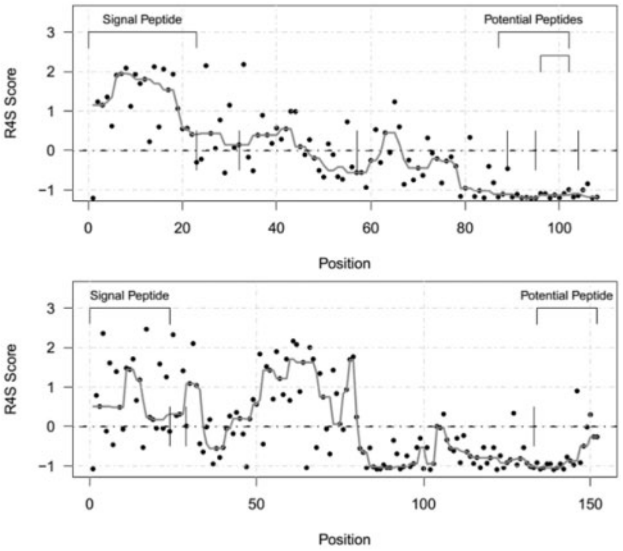

Fig. 6. Evolutionary profiles of novel peptides predicted by our approach in two precursors: CF057_HUMAN, where we identify two potential peptides between predicted proteolytic sites in a conserved region near the C-terminus of the protein (top), and F150B_HUMAN, where we suggest a novel peptide at a conserved segment between a predicted proteolytic site and the C-terminus of the protein (bottom). See Figure 1 for details

Another potential benefit of the peptide classifier described here is to determine whether currently annotated peptides possess a function, i.e. detecting false-positive findings in the scientific literature and databases. Notably, calcitonin (CALC_HUMAN, residues 85–116) is conserved relative to the upstream pro-peptide, whereas katacalcin (CALC_HUMAN, residues 121–141) is variable. Thus, we posit that katacalcin is not a functional peptide (Fig. 7, top). Similarly, we suggest that the highly variable 'potential peptide' (residues 105–134) and

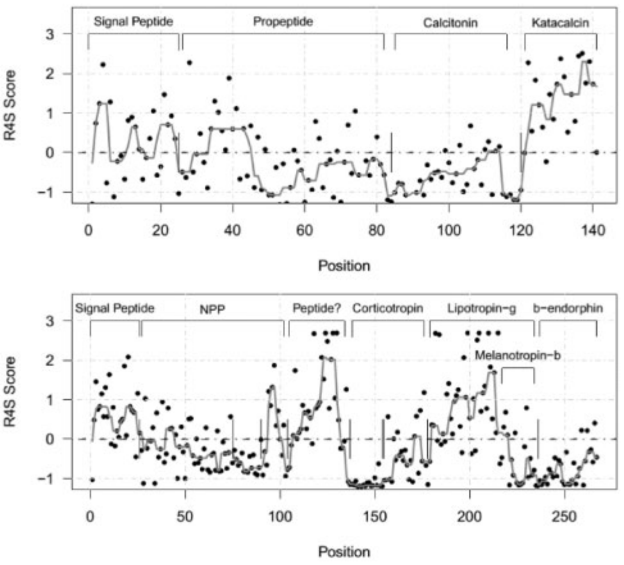

Fig. 7. Top panel: evolutionary profiles of calcitonin (CALC_HUMAN). Our analysis highlights the conservation of the known peptide (residues 85–116) but reveals high variability for katacalcin (121–141). Therefore, we predict that katacalcin is not a functional peptide. Bottom panel: evolutionary profiles of pro-opiomelanocortin (COLI_HUMAN). Our analysis highlights the conservation of some of the annotated peptides. In addition, our analysis reveals high variability for the 'potential peptide' (105–134) and for lipotropin γ (179–234). We predict that these highly variable segments are not functional peptides. For simplicity, some of the potential peptides were omitted from the figure. See Figure 1 for details

lipotropin γ (residues 179–234) of pro-opiomelanocortin (COLI_HUMAN) are not functional peptides (Fig. 7, bottom).

In this study, we preferred Rate4Site over entropy because the former takes into account the topology and branch lengths of the phylogenetic tree. Thus, Rate4Site accounts for sequence redundancy without discarding part of the data and is therefore expected to be more accurate than entropy. Figure 4 demonstrates that Rate4Site has a stronger predictive power than entropy: at a recall of 40%, Rate4Site gives a precision of ~60% compared with that of ~45% given by entropy.

In conclusion, this study highlights how evolutionary analysis can be instrumental in functional annotation.

ACKNOWLEDGEMENTS

The authors are grateful to Haim Ashkenazy, Ehud Schreiber, Arieh Bragg, Amit Novik and Emmanuel A. Weyl for useful comments and to Gady Cojocaru for kind help in expression analysis.

Funding: This work was supported by the ERC-STG grant no. 309600.

Conflict of Interest: none declared.

REFERENCES

Ashkenazy,H. *et al.* (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.*, **38**, W529–W533.

Bendtsen,J.D. *et al.* (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.

Bradbury,A.F. *et al.* (1982) Mechanism of C-terminal amide formation by pituitary enzymes. *Nature*, **298**, 686–688.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Dalbey,R.E. and Von Heijne,G. (1992) Signal peptidases in prokaryotes and eukaryotes—a new protease family. *Trends Biochem. Sci.*, **17**, 474–478.

Day,R. *et al.* (1998) Prodynorphin processing by proprotein convertase 2. Cleavage at single basic residues and enhanced processing in the presence of carboxypeptidase activity. *J. Biol. Chem.*, **273**, 829–836.

Duckert,P. *et al.* (2004) Prediction of proprotein convertase cleavage sites. *Protein Eng. Des. Sel.*, **17**, 107–112.

Erez,A. *et al.* (2004) Sil overexpression in lung cancer characterizes tumors with increased mitotic activity. *Oncogene*, **23**, 5371–5377.

Friis-Hansen,L. *et al.* (2001) Attenuated processing of proglucagon and glucagon-like peptide-1 in carboxypeptidase E-deficient mice. *J. Endocrinol.*, **169**, 595–602.

Helpman,L. *et al.* (2009) Systematic antigenic profiling of hematopoietic antigens on ovarian carcinoma cells identifies membrane proteins for targeted therapy development. *Am. J. Obstet. Gynecol.*, **201**, 196.e1–e7.

Hummon,A.B. *et al.* (2003) From precursor to final peptides: a statistical sequence-based approach to predicting prohormone processing. *J. Proteome Res.*, **2**, 650–656.

Kalkum,M. *et al.* (2003) Detection of secreted peptides by using hypothesis-driven multistage mass spectrometry. *Proc. Natl Acad. Sci. USA*, **100**, 2795–2800.

Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

Katoh,K. *et al.* (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.

Kliger,Y. *et al.* (2008) Predicting proteolytic sites in extracellular proteins: only halfway there. *Bioinformatics*, **24**, 1049–1055.

Krogh,A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Lopez,G. *et al.* (2007) firestar—prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res.*, **35**, W573–W577.

Lopez,G. *et al.* (2011) firestar—advances in the prediction of functionally important residues. *Nucleic Acids Res.*, **39**, W235–W241.

Mayrose,I. *et al.* (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.*, **21**, 1781–1791.

Mirabeau,O. *et al.* (2007) Identification of novel peptide hormones in the human proteome by hidden Markov model screening. *Genome Res.*, **17**, 320–327.

Nielsen,H. *et al.* (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6.

Nimrod,G. *et al.* (2005) In silico identification of functional regions in proteins. *Bioinformatics*, **21** (**Suppl. 1**), i328–i337.

Ohyama,K. *et al.* (2008) Identification of a biologically active, small, secreted peptide in Arabidopsis by in silico gene screening, followed by LC-MS-based structure analysis. *Plant J.*, **55**, 152–160.

Paetzel,M. *et al.* (2002) Signal peptidases. *Chem. Rev.*, **102**, 4549–4580.

Pupko,T. *et al.* (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18** (**Suppl. 1**), S71–S77.

Seidah,N.G. and Chretien,M. (1999) Proprotein and prohormone convertases: a family of subtilases generating diverse bioactive polypeptides. *Brain Res.*, **848**, 45–62.

Shemesh,R. *et al.* (2008) Discovery and validation of novel peptide agonists for G-protein-coupled receptors. *J. Biol. Chem.*, **283**, 34643–34649.

Shi,L. *et al.* (2012) Identification of Peptide lv, a novel putative neuropeptide that regulates the expression of L-type voltage-gated calcium channels in photoreceptors. *PLoS One*, **7**, e43091.

Shichiri,M. *et al.* (2003) Salusins: newly identified bioactive peptides with hemodynamic and mitogenic activities. *Nat. Med.*, **9**, 1166–1172.

Sonmez,K. *et al.* (2009) Evolutionary sequence modeling for discovery of peptide hormones. *PLoS Comput. Biol.*, **5**, e1000258.

Sonnhammer,E.L. *et al.* (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 175–182.

Team,R.D.C. (2005) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Tukey,J.W. (1977) *Exploratory Data Analysis*, Reading, MA, p. 231.