molecules

MDPI

Article

High-Throughput Identification of Mammalian Secreted Proteins Using Species-Specific Scheme and Application to Human Proteome

Jian Zhang [1,*,†], Haiting Chai [2,†], Song Guo [1], Huaping Guo [1] and Yanling Li [1]

[1] School of Computer and Information Technology, Xinyang Normal University, Xinyang 464000, China; songguo_xynu@yeah.net (S.G.); hpguoxynu@sina.com (H.G.); yanlingli639@163.com (Y.L.)

[2] College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK; h.chai.1@research.gla.ac.uk

* Correspondence: jianzhang@xynu.edu.cn; Tel.: +86-376-639-0765

† These authors contributed equally to this work.

Received: 13 April 2018; Accepted: 30 May 2018; Published: 14 June 2018

Abstract: Secreted proteins are widely spread in living organisms and cells. Since secreted proteins are easy to be detected in body fluids, urine, and saliva in clinical diagnosis, they play important roles in biomarkers for disease diagnosis and vaccine production. In this study, we propose a novel predictor for accurate high-throughput identification of mammalian secreted proteins that is based on sequence-derived features. We combine the features of amino acid composition, sequence motifs, and physicochemical properties to encode collected proteins. Detailed feature analyses prove the effectiveness of the considered features. Based on the differences across various species of secreted proteins, we introduce the species-specific scheme, which is expected to further explore the intrinsic attributes of specific secreted proteins. Experiments on benchmark datasets prove the effectiveness of our proposed method. The test on independent testing dataset also promises a good generalization capability. When compared with the traditional universal model, we experimentally demonstrate that the species-specific scheme is capable of significantly improving the prediction performance. We use our method to make predictions on unreviewed human proteome, and find 272 potential secreted proteins with probabilities that are higher than 99%. A user-friendly web server, named iMSPs (identification of Mammalian Secreted Proteins), which implements our proposed method, is designed and is available for free for academic use at: http://www.inforstation.com/webservers/iMSP/.

Keywords: secreted proteins; species-specific; high-throughput; human proteome

1. Introduction

Secreted proteins (SPs) are the proteins that are released by a cell or tissue into the extracellular space. Generally, these proteins are produced through two pathways, namely the classical Endoplasmic Reticulum and Golgi routes [1] and the unclassical secretory routes [2]. Secreted proteins play important roles in living organisms. According to their functions, they could be divided into many categories, which include hormones [3], cytokines [4], enzymes [5], toxins [6], and antibiotics [5]. In humans, the liver is the most important secretory organ. It produces a large number of plasma proteins, such as albumin, fibrinogen, and transferrin. Secreted proteins are easy to detect in body fluids, urine, and saliva in clinical trials [7], which endows them with the capability of being a rich source of biomarkers and drug targets. Since the majority of the blood diagnostic tests are directly towards secreted proteins, it is not unusual to emphasize the significance of this class of proteins.

Recent years have witnessed a number of computation-based approaches in this field. In 2011, Hong et al. used physiochemical properties and amino acid composition features to predict whether a protein can be excreted into urine [8]. Liu et al. adopted an information-retrieval (i.e., manifold ranking) technique for the identification of blood-secretory proteins [9]. Huang et al. used 531 physicochemical properties together with support vector machine to recognize secreted proteins [10]. Soon after that, Restrepo-Montoya et al. calculated a set of sequence-based features to predict secreted proteins and proposed a classifier named NClassG+ [11]. Yu et al. predicted bacterial secreted proteins by using the general concept of pseudo amino acid composition [12]. They also constructed a public web server, named SecretP. In [13], Luo et al. combined position-specific scoring matrix and used auto-covariance theory to encode secreted proteins. In 2013, Wang et al. collected a series of physicochemical properties and several sequence-based features for identifying human salivary proteins from blood circulation. They also used their method in diagnostic biomarker recognition [14]. Yu et al. built a multi-classifier to predict various types of secreted proteins [15]. Besides simply predicting secreted proteins, Sun et al. applied their method in the identification of head and neck cancer biomarkers [16]. Additionally, in secreted proteins, signal peptides are destined towards the secretory pathway [17]. Therefore, the research on signal peptides contribute to the knowledge of secreted proteins [18–20]. However, proteins with signal peptides are not necessarily secreted [21]. In eukaryotes, a protein with signal pepteides will cotranslationally translocate across the membrane. While in prokaryotes, this process takes across the cytoplasmic membrane [21].

The above-mentioned studies all contributed to the development of the research in secreted proteins. However, there still exist some shortcomings that need to be further investigated: (i) the structure-based methods, which could achieve high accuracy, are limited in real application due to the small number of known protein structures. Although sequence-based predictors are featured in their wide application, they often suffer from the unsatisfactory prediction performance; (ii) many methods used structure- or sequence-based features in order to construct feature matrix without analyzing the features in detail. That is, it is unknown whether these features could successfully differentiate secreted proteins from non-secreted proteins; (iii) some predictors simply predict general secreted proteins without considering the differences across various species of secreted proteins. Based on our investigation, the differences do exist and they help to recognize specific secreted proteins.

In view of these three issues, we aim to focus on the challenge of proposing an accurate computational method for the identification of mammalian secreted proteins based on primary sequences. Instead of using general secreted proteins, we compile several sub-datasets of prevalent species of secreted proteins. The features which have been proved to be involved in secreted proteins are collected to encode secreted proteins. We analyze the differences between secreted proteins and non-secreted proteins in detail, especially across several mammalian species. In addition, Fisher-Markov selector together with incremental feature selection scheme is introduced to remove redundant features, as well as to explore optimal feature subset. Experimental results on benchmark datasets and independent testing dataset prove the effectiveness and generalization capability of our method. Additionally, we also make predictions on unreviewed human proteome and find potential secreted proteins with high confidence.

2. Results and Discussion

2.1. The Characteristics of the Calculated Features

In this paper, we encode the proteins by using three types of features, including amino acid composition (AAC), sequence motifs (MTF), and physicochemical properties (PCP). Before constructing the prediction model, we investigate the differences of the considered features between secreted proteins and non-secreted proteins. As shown in Figure 1 (SPs-all), eight amino acids are overrepresented in secreted proteins against that in non-secreted proteins. When compared with non-secreted proteins, five out of eight show relative higher overrepresented. Lysine is less favored in secreted proteins when compared with that in non-secreted proteins. In five specific-specific datasets, the top five enriched amino acids keep consensus with that in SPs-all. However, the frequencies vary in different species. These results indicate that amino acids composition help to discriminate secreted proteins from non-secreted proteins.

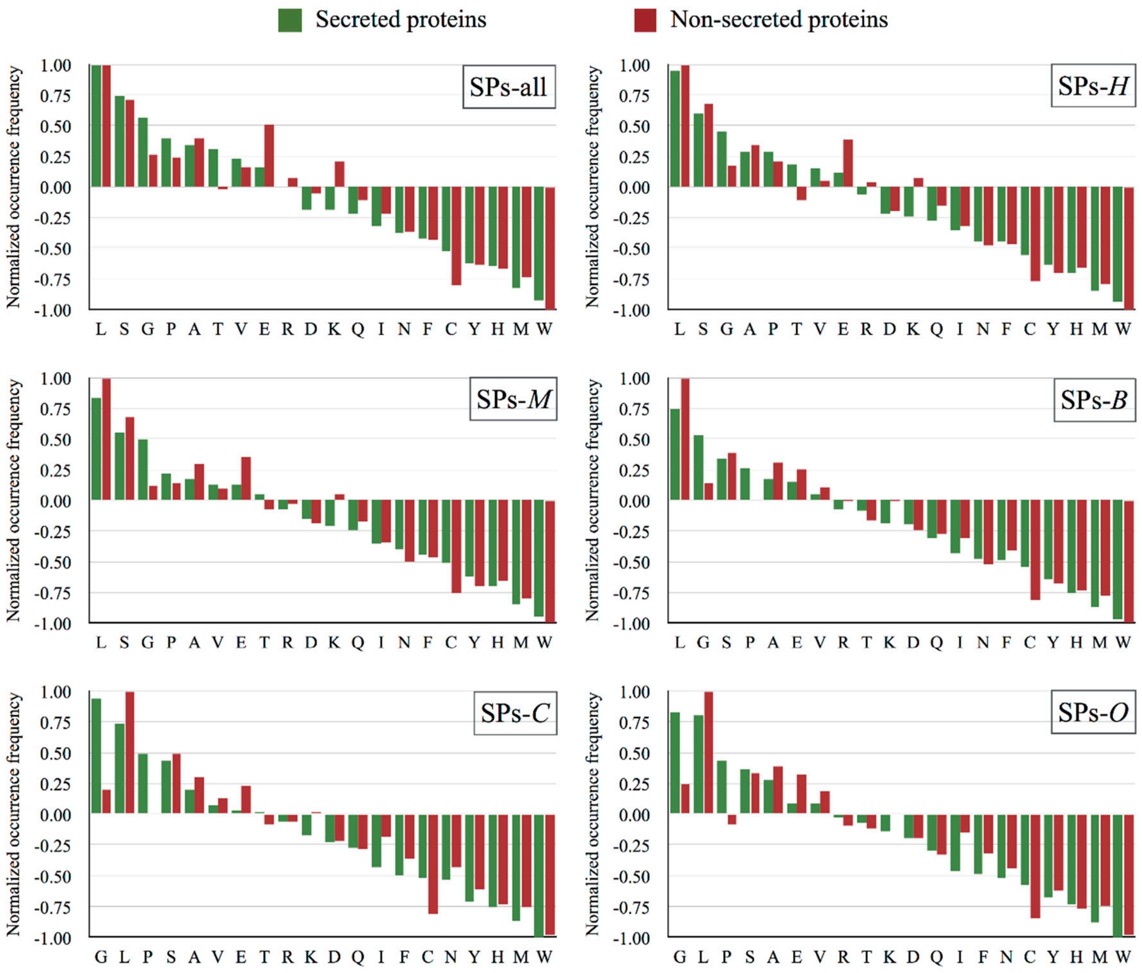

Figure 1. The relative amino acid composition of secreted proteins and non-secreted proteins in various datasets. The amino acids are sorted according to their enrichments in secreted proteins.

We further illustrate the distribution of physicochemical properties in Figure 2. Midline, box boundaries, and whiskers indicate median, quartiles, and 10th and 90th percentiles. The *x*-axis indicates the normalized values and *y*-axis stands for twelve properties. For instance, the distribution of secreted proteins against the non-secreted proteins varies obviously in hydrophobicity (Panel A). This phenomenon keeps consistent in SPs-*H*, SPs-*M*, SPs-*B*, and SPs-*C*. In SPs-*H* or SPs-*B*, a significant difference is found on the distributions of the entropy of formulation and protein kinase A. In SPs-*M*, the difference on protein kinase A is mild, but that on polarity is remarkable. In SPs-*C* and SPs-*O*, the majority of the considered physicochemical attributes show a big difference in secreted proteins against the non-secreted proteins.

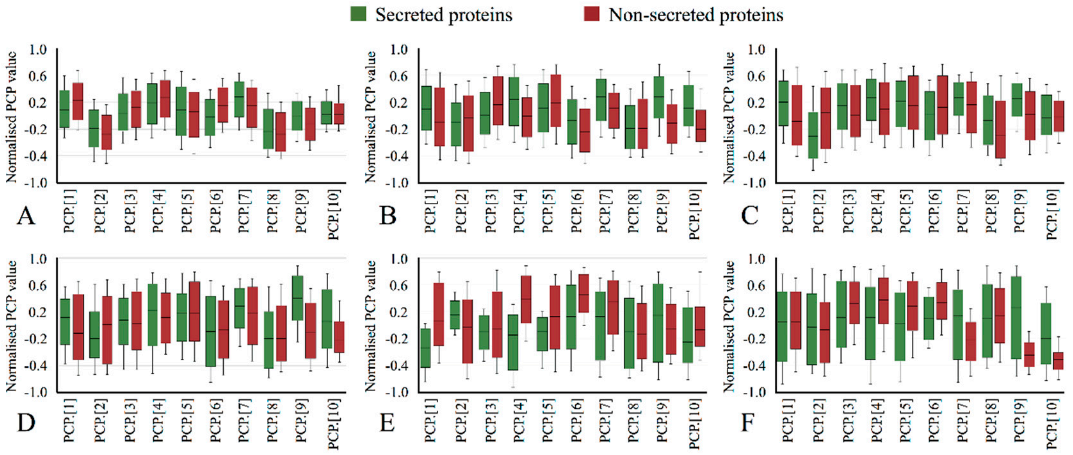

Figure 2. Physicochemical properties of secreted proteins and non-secreted proteins in (**A**) Secreted proteins (SPs)-all, (**B**) SPs-*H*, (**C**) SPs-*M*, (**D**) SPs-*B*, (**E**) SPs-*C* and (**F**) SPs-*O*. Physicochemical properties (PCP) represent hydrophobicity (PCP. [1]), polarity (PCP. [2]), solvation free energy (PCP. [3]), graph shape index (PCP. [4]), transfer free energy (PCP. [5]), correlation coefficient in regression analysis (PCP. [6]), residue accessible surface area (PCP. [7]), partition coefficient (PCP. [8]), entropy of formulation (PCP. [9]) and protein kinase A (PCP. [10]), respectively..

Listed in Table 1 are the calculated top 20 informative motifs in various datasets. We find that 'L'-rich (leucine-rich) MTFs are highly favored in SPs-all and SPs-*H* (exemplified by Figure 3). Extracellular leucine-rich pattern domains are proved to be the key organizers of connectivity among the development of neural circuits in secreted proteins [22]. It also regulates axon guidance, target selection, synapse formation, and the stabilization of connections [23]. The 'L'-rich MTFs in different secondary structures usually indicates various structure functions. As shown in Figure 3, the 'L'-rich MTFs are always located at the intrinsically disordered region (Figure 3A, 'LLLL' motif), the middle of the coil (Figure 3B, 'LAL-L' motif), and the edge of the helix (Figure 3C, 'L-LLA' motif). For instance, 'L'-rich MTFs in α-helices often shows pronounced curvature, while the β-strand usually expresses effective binding interaction [24]. Since 'L'-rich MTFs is an efficient structure, it endows them the capability of regulating intercellular communication and cell adhesion. This can explain why they are most favored in secreted proteins [24]. 'G'-rich motifs are prevalent in SPs-*M*, SPs-*C*, and SPs-*O*. These phenomena keep consistent with that in amino acid compositions. However, although 'C' is under-represented in secreted proteins, it plays important roles in the compositions of MTFs. The enriched conditions of 'L' and 'G' might be a reason for such phenomenon. Although 'C' residues are depleted in secreted proteins, we find that the 'C'-rich motifs are enriched in various species of secreted proteins. More detailed information of these MTFs is provided in Table S1. The physicochemical index data for twenty standard amino acids is listed in Table S2.

Table 1. The top 20 informative motifs in various datasets. '-' denotes arbitrary 20 amino acids.

SPs-All		SPs-*H*		SPs-*M*		SPs-*B*		SPs-*C*		SPs-*O*	
MTF	**RDI**	**MTF**	**RDI**	**MTF**	**RDI**	**MTF**	**RDI**	**MTF**	**RDI**	**MTF**	**RDI**
LLLL	0.035	LLLL	0.039	LLLL	0.037	LLLL	0.045	C-CR	0.053	G-CP	0.064
LL-LLL	0.034	LL-LLL	0.038	LL-LLL	0.035	G-CP	0.035	G-CP	0.047	C-VP	0.057
LLL-LL	0.032	LLL-LL	0.036	C-CP	0.027	CP-G	0.033	CG-C	0.047	GC-P	0.052
CP-G	0.022	LAL-L	0.027	C-QG	0.027	C-PG	0.032	C-AG	0.047	CS-C	0.051
LAL-L	0.022	L-LLA	0.024	CP-G	0.026	C-CL	0.032	KGD	0.046	SC-C	0.051
G-TC	0.021	LL-LA	0.024	C-NG	0.024	L-LLA	0.032	CP-Q	0.043	C-SC	0.049
C-PG	0.020	L-LLG	0.024	G-TC	0.024	S-SC	0.032	CC-P	0.043	C-CR	0.049
LLL-A	0.020	LL-LG	0.024	CQ-G	0.024	CS-S	0.031	GR-C	0.042	SC-P	0.047

Table 1. *Cont.*

SPs-All		SPs-*H*		SPs-*M*		SPs-*B*		SPs-*C*		SPs-*O*	
MTF	**RDI**	**MTF**	**RDI**	**MTF**	**RDI**	**MTF**	**RDI**	**MTF**	**RDI**	**MTF**	**RDI**
LL-LA	0.020	LLL-A	0.023	C-PG	0.024	AC-P	0.031	CG-R	0.041	C-SG	0.046
L-LLA	0.020	LLL-G	0.023	GG-C	0.022	LL-LA	0.030	C-CL	0.040	SG-C	0.046
GT-C	0.019	C-PG	0.022	CA-G	0.022	CA-P	0.030	SC-C	0.040	CG-C	0.046
GS-C	0.018	LLA-L	0.022	C-SC	0.022	LLL-A	0.030	CC-R	0.040	GC-G	0.045
L-LW	0.018	CP-G	0.022	C-GG	0.021	LCL	0.029	CV-P	0.039	CC-P	0.044
ALL-L	0.017	ALL-L	0.021	GE-C	0.021	G-SC	0.029	CA-G	0.039	LLLL	0.044
LLA-L	0.017	L-LAL	0.021	GK-C	0.020	SC-S	0.029	PQG	0.037	KPG	0.044
LL-AL	0.017	LL-AL	0.020	GT-C	0.019	C-SS	0.028	CS-C	0.037	C-PT	0.043
G-RC	0.017	LA-LL	0.020	G-SC	0.019	G-CS	0.028	C-SC	0.037	GDR	0.043
P-CP	0.017	AL-LL	0.020	C-PR	0.019	CG-G	0.027	RGP	0.037	CS-G	0.042
CP-P	0.017	G-TC	0.020	WL-L	0.019	AC-S	0.027	C-PT	0.036	C-GC	0.041
CA-P	0.016	L-LW	0.019	G-RC	0.019	A-CL	0.027	PGQ	0.036	S-SC	0.039

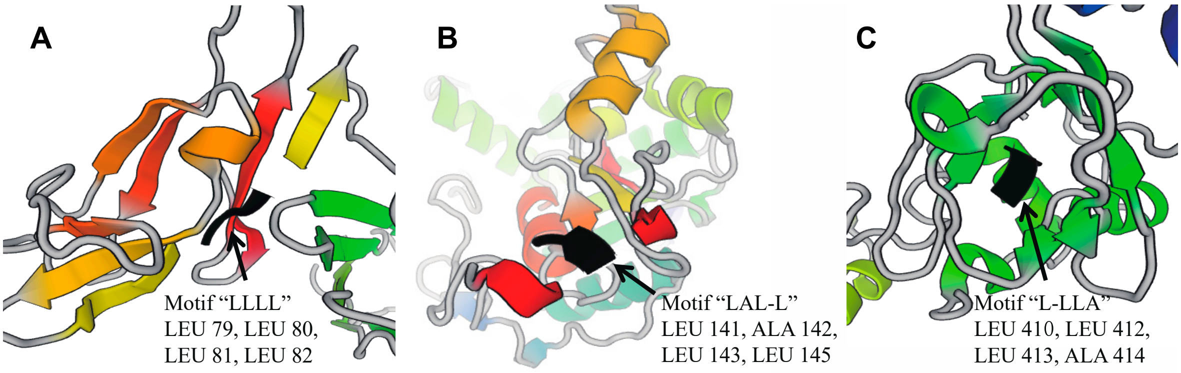

Figure 3. Example of leucine-rich motifs in mammalian secreted proteins. Panels (**A**–**C**) are captured from protein 3D structure 4GRW (human IL-23 with 3 Nanobodies), 1T8T (human 3-*O*-Sulfotransferase-3 with bound PAP), and 5NV6 (human transforming growth factor beta-induced protein), respectively.

2.2. *The Performance of the Extracted Features*

In Section 2.1, we analyze the differences across various species of secreted proteins and non-secreted proteins on considered features. However, it is still unknown whether these features can be used to distinguish secreted proteins from non-secreted proteins. Here, we test these features on general SPs-all and five species-specific datasets.

Table 2 shows the prediction performance of the considered different features on the training datasets over five-fold cross-validation. Overall, the features of AAC, MTF, and PCP produce promising results on the general mammalian secreted proteins datasets and six species-specific secreted proteins. In detail, AAC-based features perform the best among three types of features with the highest Matthews Correlation Coefficient (MCC) and AUC values on SPs-all. Although MTF-based features could not achieve the highest prediction performance, they are featured by the high capability in recognizing non-secreted proteins (Specificity > 0.84). For Mammalia, *B. taurus*, and *C. lupus familiaris* secreted proteins, the MTF-based features give out high specificity, which is above 0.9. In comparison with ACC- and MTF-based features, PCP-based features produce similar results on six training datasets.

Table 2. The prediction performance of different features on six training datasets over five-fold cross-validation.

Dataset	Feature	Sensitivity	Specificity	Accuracy	MCC	AUC
SPs-all	AAC	0.695	0.734	0.714	0.429	0.773
	MTF	0.354	0.910	0.632	0.317	0.660
	PCP	0.707	0.702	0.705	0.410	0.754
SPs-*H*	AAC	0.697	0.719	0.708	0.416	0.736
	MTF	0.469	0.846	0.657	0.340	0.677
	PCP	0.670	0.755	0.712	0.426	0.746
SPs-*M*	AAC	0.685	0.734	0.709	0.419	0.754
	MTF	0.361	0.896	0.628	0.304	0.658
	PCP	0.652	0.722	0.687	0.374	0.732
SPs-*B*	AAC	0.663	0.781	0.722	0.447	0.765
	MTF	0.247	0.988	0.618	0.350	0.682
	PCP	0.401	0.953	0.676	0.424	0.731
SPs-*C*	AAC	0.612	0.791	0.701	0.410	0.762
	MTF	0.418	0.925	0.672	0.398	0.667
	PCP	0.463	0.900	0.682	0.404	0.759
SPs-*O*	AAC	0.677	0.797	0.737	0.477	0.744
	MTF	0.563	0.870	0.716	0.454	0.725
	PCP	0.490	0.807	0.648	0.313	0.693

2.3. The Performance of Feature Selection Scheme

We empirically prove the prediction capability of proposed features in Section 2.2. In this section, we combine three types of features together to construct the feature space. When considering the existence of redundant features, we firstly use Fisher-Markov Selector [25] to calculate the coefficients between each of the features and labels. The ranked feature lists are provided in Figure S1. Next, we iteratively add features into the feature subset according to the incremental feature selection strategy.

Table 3 shows the prediction results that are based on the optimal feature subsets. The numbers for six optimum feature subsets are 30, 20, 45, 30, 30, and 35, respectively. They achieve MCC ranging from 0.490~0.644 and AUC ranging from 0.783~0.835. On SPs-*M* datasets, the MCC and AUC increase by 0.061 and 0.033 when compared with the best individual features. On SPs-*B*, the AUC slightly climbs to 0.815. More significant improvements are shown on SPs-*O* (MCC 0.644 versus 0.477) and SPs-*C* (MCC 0.546 versus 0.410). These results illustrate the effectiveness of our feature selection scheme. The prediction performances of the detailed different numbers of features on six training sets are provided in Table S3.

Table 3. The performance of the optimum feature subset general Mammalia secreted proteins and five species-specific secreted proteins over five-fold cross-validation.

Dataset	Sensitivity	Specificity	Accuracy	MCC	AUC
SPs-all	0.705	0.783	0.744	0.490	0.806
SPs-*H*	0.673	0.833	0.753	0.513	0.799
SPs-*M*	0.634	0.847	0.740	0.492	0.783
SPs-*B*	0.728	0.825	0.777	0.556	0.815
SPs-C	0.657	0.876	0.766	0.546	0.784
SPs-*O*	0.771	0.870	0.820	0.644	0.835

2.4. Comparison of Species-Specific Models with Traditional Universal Ones

Based on our previous investigation, different species of secreted proteins show various attributes in many aspects. Then, we are inspired to introduce species-specific strategy for the specific identification of various mammalian secreted proteins. When compared with universal models, species-specific ones are based on specific feature construction and optimal feature subsets. To investigate the effectiveness of this strategy, we compare these two kinds of models based on same benchmark training datasets over five-fold cross-validation. As shown in Table 4, species-specific models all achieve relatively higher (2~11%) prediction accuracy. The improvements are much more obvious on the sensitivity (3~18%) for different species expect for *M. musculus*. When considering MCC, which is capable of balancing the measurements between sensitivity and specificity, the species-specific model all produce higher values. Figure 4 displays the AUCs of species-specific and universal models. The grey bars indicate the species-specific models, while the black ones stand for the universals. For *H. sapiens*, *M. musculus*, and *B. taurus*, the improvements on AUC are about 0.018, 0.021, and 0.023. For SPs-*C* and SPs-*O*, the AUC values sharply increase from ~0.59 to ~0.78 and ~0.71 to ~0.83.

Table 4. Comparison between species-specific and universal schemes on different species of training datasets over five-fold cross-validation.

Dataset	Model	Sensitivity	Specificity	Accuracy	MCC
SPs-*H*	iMSP-*H* iMSP-*U*	0.673 0.647	0.833 0.820	0.753 0.733	0.513 0.474
SPs-*M*	iMSP-*M* iMSP-*U*	0.634 0.652	0.847 0.789	0.740 0.721	0.492 0.446
SPs-*B*	iMSP-*B* iMSP-*U*	0.728 0.695	0.825 0.811	0.777 0.753	0.556 0.509
SPs-*C*	iMSP-*C* iMSP-*U*	0.657 0.473	0.876 0.841	0.766 0.657	0.546 0.337
SPs-*O*	iMSP-*O* iMSP-*U*	0.771 0.615	0.870 0.823	0.820 0.719	0.644 0.447

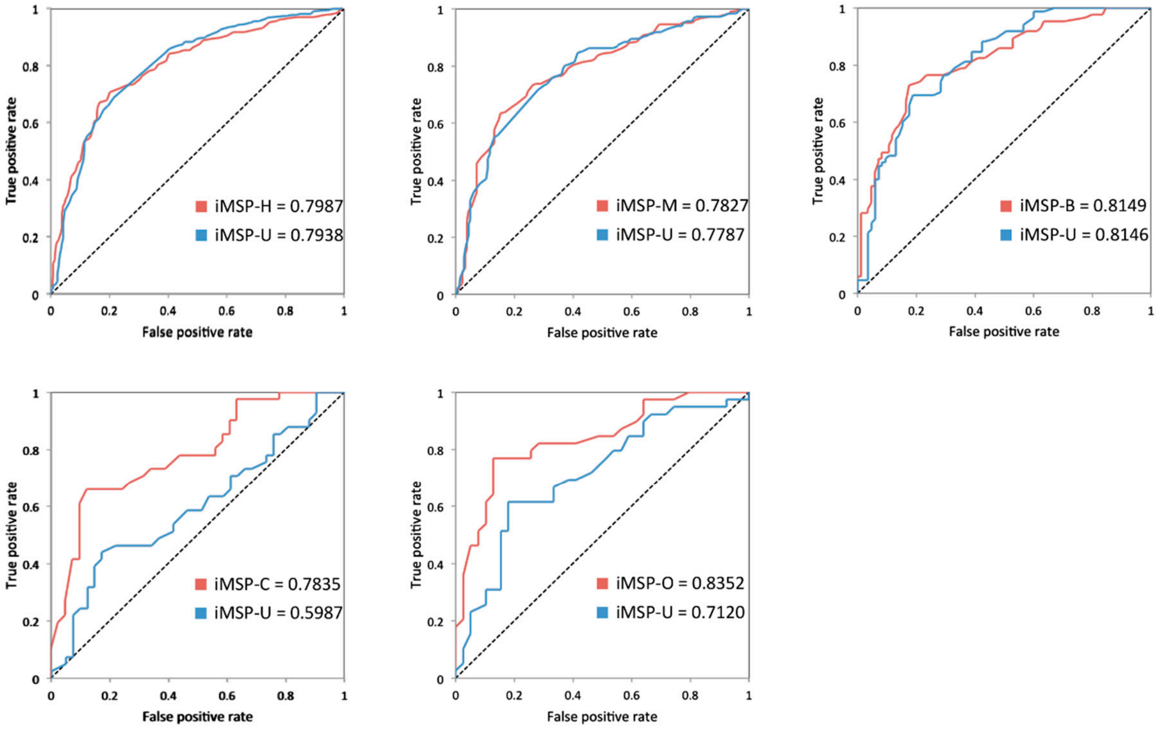

Figure 4. Comparison of predicted AUC values between species-specific and universal models.

2.5. Comparison with Other Predictors on Independent Testing Datasets

To evaluate the generalization capability of the proposed predictor as well as to compare with previous methods, we further test our method on the independent testing dataset. Recent years have witnessed several powerful predictors for identification of SPs, such as SecretomeP, NClassG+, and SRTpred. The criteria that used for selecting efficient methods include (1) the outputs of the predictors are scores and (2) the predictors can successfully predict an average length protein sequence with 200 residues within 30 min. As a result, we select two predictors, namely SecretomeP [26] and SRTpred [27], as of January 2018.

Table 5 lists the prediction results of considered predictors on various types of testing datasets. The predicted values of SecretomeP and SRTpred are directly obtained through their software. All of the predictors achieve good performance on the universal and various species-specific datasets. Our universal module (iMSP-*U*) produces the MCC of 0.427, 0.455, 0.507, 0.359, 0.324, and 0.332 on six testing datasets respectively. On the former four testing datasets, our iMSP-*U* outperforms SecretomeP and SRTpred. On SPs-*C* and SPs-*O*'s testing sets, SecretomeP and SRTpred show much better than our iMSP-*U*. When adopting species-specific models (iMSP-*H*, iMSP-*M*, iMSP-*B*, iMSP-*C*, and iMSP-*O*) on the corresponding species-specific testing datasets, the prediction performance shows obvious improvements.

Table 5. The performance of different methods on six testing datasets.

Dataset	Method	Sensitivity	Specificity	Accuracy	MCC	AUC
SPs-all	SecretomeP	0.611	0.798	0.763	0.355	0.729
	SRTpred	0.652	0.824	0.792	0.419	0.781
	iMSP-U	0.590	0.865	0.814	0.427	0.802
SPs-*H*	SecretomeP	0.632	0.787	0.762	0.340	0.764
	SRTpred	0.678	0.802	0.782	0.392	0.770
	iMSP-H	0.631	0.866	0.829	0.443	0.821
	iMSP-U	0.538	0.908	0.850	0.441	0.817
SPs-*M*	SecretomeP	0.629	0.832	0.731	0.471	0.776
	SRTpred	0.707	0.793	0.751	0.503	0.785
	iMSP-M	0.742	0.776	0.759	0.519	0.809
	iMSP-U	0.703	0.802	0.753	0.507	0.803
SPs-*B*	SecretomeP	0.575	0.861	0.824	0.367	0.768
	SRTpred	0.670	0.857	0.833	0.431	0.787
	iMSP-B	0.547	0.901	0.856	0.411	0.795
	iMSP-U	0.679	0.766	0.755	0.327	0.763
SPs-*C*	SecretomeP	0.549	0.921	0.865	0.470	0.779
	SRTpred	0.686	0.866	0.839	0.478	0.782
	iMSP-C	0.412	0.962	0.880	0.457	0.789
	iMSP-U	0.667	0.670	0.670	0.247	0.718
SPs-*O*	SecretomeP	0.729	0.782	0.775	0.390	0.747
	SRTpred	0.792	0.842	0.835	0.509	0.820
	iMSP-O	0.646	0.913	0.876	0.521	0.841
	iMSP-U	0.521	0.805	0.766	0.264	0.716

2.6. Application to Predict Secreted Proteins from Human Proteome by Using iMSP

We implement the proposed method as a public web server, named iMSP, which is deployed at http://www.inforstation.com/webservers/iMSP/. iMSP offers efficient high-throughput predictions for biologists. In this work, our new-compiled benchmark dataset was generated from UniProt (http://www.uniprot.org/, accessed on 1 January 2018). In the UniProt database, sequence similarity search programs are used to identify orthologs. Since *H. sapiens* and *M. musculus* secreted proteins occupy a large part of all secreted proteins, they would somehow influence other species of secreted proteins, such as *B. taurus*, *C. lupus* familiaris, and *O. cuniculus*. In our benchmark dataset, the number of secreted proteins in SPs-*H* and SPs-*M* is much higher than that of SPs-*B*, SPs-*C*, and SPs-*O*. As a result, the accuracy of the latter three species-specific models will be affected by that of the first two. The users are suggested to choose universal model for Bos, Canis and Oryctolagus proteins, and species-specific models for Homo and Mus proteins.

In this part, we aim to adopt iMSP to predict potential secreted proteins from human proteome. There are a total of 71,772 proteins in the human proteome. Among them, 20,303 items are reviewed records, and the rest 51,469 are unreviewed ones. Particularly, by our universal model (iMSP-*U*) and species-specific model (iMSP-*H*), we also calculate the probabilities of unreviewed human proteins to be secreted proteins. All of the proteins were ranked according to the predicted probabilities. Based on iMSP-*H*, we find that 7601 (14.77%) proteins have the probabilities higher than 0.8, while a large number of proteins are not secreted proteins (shown in Table 6). When considering the highest probabilities ($\geq$99%), we find 272 (or 0.528%) out of all 51,469 proteins to be predicted secreted proteins. Finally, we listed the predicted scores for all unreviewed human proteome (Table S4) and potential SPs with highest probabilities (Table S5).

Table 6. Predicted probabilities to be potential secreted proteins in human proteome.

Probability	**0–10%**	**10–20%**	**20–30%**	**30–40%**	**40–50%**
iMSP-*U*	1155 (2.24%)	5984 (11.63%)	7803 (15.16%)	7684 (14.93%)	7100 (13.79%)
iMSP-*H*	1904 (3.70%)	6213 (12.07%)	7333 (14.25%)	7219 (14.03%)	6769 (13.15%)
Probability	**50–60%**	**60–70%**	**70–80%**	**80–90%**	**90–100%**
iMSP-*U*	5551 (10.79%)	4745 (9.22%)	4028 (7.83%)	3768 (7.32%)	3651 (7.09%)
iMSP-*H*	5536 (10.76%)	4848 (9.42%)	4046 (7.86%)	3993 (7.76%)	3608 (7.01%)

3. Materials and Methods

3.1. Datasets Preparation

In this study, we collect 17,209 mammalian secreted proteins and 29,479 non-secreted proteins from UniProt. We take consideration of the prevalent several species, which include *Homo sapiens* (*H. sapiens*), *Mus musculus* (*M. musculus*), *Bos taurus* (*B. taurus*), *Canis lupus familiaris* (*C. lupus familiaris*), and *Oryctolagus cuniculus* (*O. cuniculus*). The species-specific datasets are used to explore the differences across the various mammalian secreted proteins. Next, Blastclust [28] is used to cluster these proteins with a threshold of 30%. We pick the longest protein from each cluster as the representative. For each dataset, we randomly pick four-fifths of secreted proteins and an equal number of non-secreted proteins to build the training/cross-validation dataset. The remaining proteins are used for independent testing. Table 7 summarizes the newly-compiled dataset. These datasets are freely available on the iMSP server.

Table 7. A breakdown of newly-compiled datasets used in this work.

Dataset	**Species**	**All Dataset**	**Training Dataset**	**Testing Daset**
		(numP, numN) *	**(numP, numN) ***	**(numP, numN) ***
SPs-all	*Mammalia*	(2560, 4299)	(2048, 2048)	(512, 2251)
SPs-*H*	*Homo sapiens*	(1986, 3714)	(1588, 1588)	(398, 2126)
SPs-*M*	*Mus musculus*	(1144, 1147)	(915, 915)	(229, 232)
SPs-*B*	*Bos taurus*	(529, 1148)	(423, 423)	(106, 725)
SPs-*C*	*Canis lupus familiaris*	(252, 492)	(201, 201)	(51, 291)
SPs-*O*	*Oryctolagus cuniculus*	(240, 490)	(192, 192)	(48, 298)

* numP and numN represent the numbers of secreted proteins and non-secreted proteins respectively.

3.2. Feature Construction

3.2.1. Amino Acid Composition-Based Features

Amino acids are the fundamental elements of proteins. The features of amino acid composition (AAC) reflect the distribution of amino acids in proteins [29,30]. AAC is widely used in predicting protein function or structures. Given a protein *P*, the features of AAC are defined, as follows:

$$f_{aa} = \{f_1, f_2, f_3, \ldots, f_{20}\} \tag{1}$$

where f_{aa} represents the calculated frequency of 20 types of amino acids in the sequence *P*. Then, these frequencies are normalized to the interval [−1, 1] by using:

$$f_{AAC} = (\frac{f_n - \min(f_{aa})}{\max(f_{aa}) - \min(f_{aa})} - \frac{1}{2}) \times 2 \tag{2}$$

where f_n, $\max(f_{aa})$, and $\min(f_{aa})$ are the original, maximum, and minimum calculated frequency of the amino acids.

3.2.2. Sequence Motif-Based Features

Proteins in the same family tend to share similar attributes. These attributes are usually located on the highly conserved parts of the proteins. These conserved parts can be recognized by sequence patterns/motifs [31]. In this study, we adopt information theory [32] in order to calculate the features of sequence motif (MTF) from protein sequences. Given a protein, the information entropy of the MTF can be formulated, as follows:

$$I(S) = logN \tag{3}$$

where *N* is the number of the considered proteins. Next, we reclassify these proteins with MTF '*M*'. The updated information entropy can be formulated as:

$$I(S|M) = P(M) \times \log(P(M) \times N) + P(\overline{M}) \times log(P(\overline{M}) \times N) \tag{4}$$

where $P(M)$ represents the percentage of proteins containing '*M*', while $P(\overline{M})$ means the opposite. The Information Gain (IG), which is produced by the introduction of MTF '*M*' can be calculated as:

$$IG(M) = I(S) - I(S|M) \tag{5}$$

In real-world cases, the imbalance on number of SPs to non-SPs would somewhat lead to potential bias on the selected motifs based on IG. Considering this, we further calculate the ratio of the difference value of IG (RDI) for target MTF '*M*', which is defined as follows:

$$RDI(M) = \frac{IG_P(M)}{I_P(S)} - \frac{IG_N(M)}{I_N(S)} \tag{6}$$

where $IG_P(M)$ and $IG_N(M)$ are IG of MTF '*M*' on SPs and non-SPs, $I_P(S)$ and $I_N(S)$ are the original information entropy of secreted proteins and non-secreted proteins. In this study, we select the top 20 informative MTFs to encode each protein. Finally, the feature of MTF is defined as:

$$f_{\mathrm{MTF}} = [M_1, M_2, \ldots, M_{20}] \tag{7}$$

where M_{n} represents the existence or not of the *n*-th motif ('1' stands for existence; '−1' refers to the opposite).

3.2.3. Physicochemical Properties-Based Features

The physiochemical properties (PCP) of residues reveal microscopic environment of proteins. These microscopic environments includes protein energy, fore, and dynamics [33]. For example, the interfaces are often associated with hydrophobic or polar residues [33]. Graph shape can somewhat determine the surface of the function regions. In view of this, we collect ten popular physicochemical properties to encode the secreted proteins. These properties include hydrophobicity [34], polarity [35], solvation free energy [36], graph shape index [37], transfer of free energy [38], correlation coefficient in regression analysis [39], residue accessible surface area [40], partition coefficient [41], entropy of formulation [42], and protein kinase A [43].

The index data for twenty standard amino acids can be formulated as:

$$\begin{bmatrix} I_{1,1} & I_{1,2} & \cdots & I_{1,20} \\ I_{2,1} & I_{2,2} & \cdots & I_{2,20} \\ \vdots & \vdots & & \vdots \\ I_{10,1} & I_{10,2} & \cdots & I_{10,20} \end{bmatrix} \tag{8}$$

where $I_{m,n}$ represented the m-th index data for the n-th type of amino acid. Detailed information of these index data are provided in Supplementary Table S1. Given a protein, its total sequence can be mathematically formulated as $SEQ = [A_1, A_2, \ldots, A_L]$, where L is length of the protein, A_n is a 20×1 submatrix representing amino acids (digital "1" for the occupation and "0" for the opposite). Then, the feature of physicochemical patterns can be formulated as: $f_{\text{PCP}} = [PCP_1, PCP_2, \ldots, PCP_{20}]$, where PCP_n was the average value of the n-th column in the matrix product of Equation (8) and SEQ. These elements are scaled between -1 and 1 using Equation (2).

3.3. Feature Selection Strategy

In information theory, the existence of 'bad' (noisy or irrelevant) features will potential destroy the classifier or will lead to overfitting [44]. Therefore, it is necessary to remove the bad features before constructing a powerful model. In this study, we introduce Fisher-Markov Selector (FMS) [25], together with incremental feature selection (IFS) strategy to search the optimal feature subset. It uses Markov random field optimization techniques to identify the most informative features in describing the native labels. Incremental feature selection strategy is adopted to build different feature subset, according to the scored feature lists. For each feature subset, a classifier is built and evaluated. The classifier that achieves the highest prediction performance will be chosen as the final prediction model. The corresponding feature subset will be the optimal feature subset.

3.4. Model Construction and Performance Evaluation

In this work, LIBSVM 3.20 [33] is utilized to empirically train and optimized the prediction model. The radial basis function is adopted as the kernel function and grid search is used to search for optimal parameters.

We assess our method using two statistical cross-validation methods, namely five-fold cross-validation and the independent test. A five-fold cross-validation is adopted for evaluating the performance of proposed predictor on the training dataset. First, we randomly divide the training dataset into five parts. In each run, four of them are used to train a classifier and test on the holdout fold. Then, we combine the predictions in all five iterations to compute the following threshold-dependent measurements: accuracy, sensitivity, specificity, and Matthews Correlation Coefficient (MCC). They are defined, as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{10}$$

$$Specificity = \frac{TN}{TN + FP} \tag{11}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}} \tag{12}$$

where *TP* is the number of correctly recognized secreted proteins, *TN* is the number of correctly recognized non-secreted proteins, *FP* is the number of incorrectly recognized secreted proteins, and *FN* is the number of incorrectly recognized non-secreted proteins. Since the abovementioned threshold-dependent measurements are sensitive to thresholds, we also adopt AUC (area under Receiver Operating Characteristic (ROC) curve), which has been proved to be a robust assessment criterion for imbalanced testing datasets [34].

4. Conclusions

Secreted proteins are widely spread in living organisms and cells. Featured by easily being detected in body fluids, urine, and saliva in clinical, they play important roles in potential biomarkers for disease diagnosis and vaccine production. In this study, we present a novel high-throughput predictor for the identification of mammalian SPs from primary protein sequences. We analyze the differences across various types of secreted proteins and non-secreted proteins by using considered features, including AAC, MTF, and PCP. When compared with the traditional universal model, the introduced species-specific scheme proves to be capable of improving the prediction performance for corresponding species of secreted proteins. Tests on independent testing dataset promise a good generalization capability of our proposed method. We also apply the proposed predictor to predict unreviewed human proteome. We list 272 potential secreted proteins, which are predicted with high confidence ($\geq$99%), for further investigation by biologists.

Supplementary Materials: The following are available online http://www.mdpi.com/1420-3049/23/6/1448/s1. Table S1: The selected 20 motifs in six datasets. Table S2: Physicochemical index data for twenty standard amino acids. Table S3: Performance of different numbers of features in six training datasets over five-fold cross-validation. Table S4: The predicted scores for all unreviewed human proteome by iMSP. Table S5: The predicted scores for potential SPs with highest probabilities by iMSP. Figure S1 Feature ranking in six training sets.

Author Contributions: J.Z. conceived the idea of this research and was in charge of the iMSP implementation. H.C. and S.G. performed the research including data collection, test and analysis. H.C. and H.G. optimized the research and participated in the development and validation of the Web server. Y.L. suggested extension and modifications to the research. All authors have read and approved the final manuscript.

Funding: This reseach was supported in part by the National Natural Science Foundation of China under the grants 61672261, 61572417, and 61502199, by Nanhu Scholars Program for Young Scholars of XYNU, and in part by the China Scholarship Council under Grant 201706620069.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gilmore, R.; Walter, P.; Blobel, G. Protein translocation across the endoplasmic reticulum. II. Isolation and characterization of the signal recognition particle receptor. *J. Cell Biol.* **1982**, *95*, 470–477. [CrossRef] [PubMed]
2. Nickel, W. The mystery of nonclassical protein secretion. *Eur. J. Biochem.* **2003**, *270*, 2109–2119. [CrossRef] [PubMed]
3. Trayhurn, P.; Drevon, C.A.; Eckel, J. Secreted proteins from adipose tissue and skeletal muscle–adipokines, myokines and adipose/muscle cross-talk. *Arch. Physiol. Biochem.* **2011**, *117*, 47–56. [CrossRef] [PubMed]
4. Abraham, C.; Medzhitov, R. Interactions between the host innate immune system and microbes in inflammatory bowel disease. *Gastroenterology* **2011**, *140*, 1729–1737. [CrossRef] [PubMed]
5. Kulp, A.; Kuehn, M.J. Biological functions and biogenesis of secreted bacterial outer membrane vesicles. *Annu. Rev. Microbiol.* **2010**, *64*, 163–184. [CrossRef] [PubMed]
6. Schrank, A.; Vainstein, M.H. Metarhizium anisopliae enzymes and toxins. *Toxicon* **2010**, *56*, 1267–1274. [CrossRef] [PubMed]

7. Mudrak, B.; Kuehn, M.J. Specificity of the type II secretion systems of enterotoxigenic Escherichia coli and Vibrio cholerae for heat-labile enterotoxin and cholera toxin. *J. Bacteriol.* **2010**, *192*, 1902–1911. [CrossRef] [PubMed]
8. Hong, C.S.; Cui, J.; Ni, Z.; Su, Y.; Puett, D.; Li, F.; Xu, Y. A computational method for prediction of excretory proteins and application to identification of gastric cancer markers in urine. *PLoS ONE* **2011**, *6*, e16875. [CrossRef] [PubMed]
9. Liu, Q.; Cui, J.; Yang, Q.; Xu, Y. In-silico prediction of blood-secretory human proteins using a ranking algorithm. *BMC Bioinform.* **2010**, *11*, 250. [CrossRef] [PubMed]
10. Hung, C.-H.; Huang, H.-L.; Hsu, K.-T.; Ho, S.-J.; Ho, S.-Y. Prediction of non-classical secreted proteins using informative physicochemical properties. *Interdisciplin. Sci.* **2010**, *2*, 263–270. [CrossRef] [PubMed]
11. Restrepo-Montoya, D.; Pino, C.; Nino, L.F.; Patarroyo, M.E.; Patarroyo, M.A. NClassG+: A classifier for non-classically secreted Gram-positive bacterial proteins. *BMC Bioinform.* **2011**, *12*, 21. [CrossRef] [PubMed]
12. Yu, L.; Guo, Y.; Li, Y.; Li, G.; Li, M.; Luo, J.; Xiong, W.; Qin, W. SecretP: Identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition. *J. Theor. Biol.* **2010**, *267*, 1–6. [CrossRef] [PubMed]
13. Luo, J.; Yu, L.; Guo, Y.; Li, M. Functional classification of secreted proteins by position specific scoring matrix and auto covariance. *Chemometr. Intell. Lab. Syst.* **2012**, *110*, 163–167. [CrossRef]
14. Wang, J.; Liang, Y.; Wang, Y.; Cui, J.; Liu, M.; Du, W.; Xu, Y. Computational prediction of human salivary proteins from blood circulation and application to diagnostic biomarker identification. *PLoS ONE* **2013**, *8*, e80211. [CrossRef] [PubMed]
15. Yu, L.; Luo, J.; Guo, Y.; Li, Y.; Pu, X.; Li, M. In silico identification of Gram-negative bacterial secreted proteins from primary sequence. *Comput. Biol. Med.* **2013**, *43*, 1177–1181. [CrossRef] [PubMed]
16. Sun, Y.; Du, W.; Zhou, C.; Zhou, Y.; Cao, Z.; Tian, Y.; Wang, Y. A Computational Method for Prediction of Saliva-Secretory Proteins and Its Application to Identification of Head and Neck Cancer Biomarkers for Salivary Diagnosis. *IEEE Trans. Nanobiosci.* **2015**, *14*, 167–174. [CrossRef] [PubMed]
17. Kapp, K.; Schrempf, S.; Lemberg, M.K.; Dobberstein, B. Post-Targeting Functions of Signal Peptides. In *Madame Curie Bioscience Database*; Landes Bioscience: Austin, TX, USA, 2013.
18. Käll, L.; Krogh, A.; Sonnhammer, E.L. Advantages of combined transmembrane topology and signal peptide prediction—The Phobius web server. *Nucleic Acids Res.* **2007**, *35* (Suppl. 2), W429–W432. [CrossRef] [PubMed]
19. Reynolds, S.M.; Käll, L.; Riffle, M.E.; Bilmes, J.A.; Noble, W.S. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput. Biol.* **2008**, *4*, e1000213. [CrossRef] [PubMed]
20. Petersen, T.N.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* **2011**, *8*, 785. [CrossRef] [PubMed]
21. Nielsen, H. Predicting secretory proteins with SignalP. *Protein Funct. Predict. Methods Protoc.* **2017**, *1611*, 59–73.
22. De Wit, J.; Hong, W.; Luo, L.; Ghosh, A. Role of leucine-rich repeat proteins in the development and function of neural circuits. *Annu. Rev. Cell Dev. Biol.* **2011**, *27*, 697–729. [CrossRef] [PubMed]
23. Kusuzawa, S.; Honda, T.; Fukata, Y.; Fukata, M.; Kanatani, S.; Tanaka, D.H.; Nakajima, K. Leucine-rich glioma inactivated 1 (Lgi1), an epilepsy-related secreted protein, has a nuclear localization signal and localizes to both the cytoplasm and the nucleus of the caudal ganglionic eminence neurons. *Eur. J. Neurosci.* **2012**, *36*, 2284–2292. [CrossRef] [PubMed]
24. Kobe, B.; Kajava, A.V. The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.* **2001**, *11*, 725–732. [CrossRef]
25. Cheng, Q.; Zhou, H.; Cheng, J. The fisher-markov selector: Fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1217–1233. [CrossRef] [PubMed]
26. Bendtsen, J.D.; Jensen, L.J.; Blom, N.; Von Heijne, G.; Brunak, S. Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng. Des. Sel.* **2004**, *17*, 349–356. [CrossRef] [PubMed]
27. Garg, A.; Raghava, G.P. A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. *In Silico Biol.* **2008**, *8*, 129–140. [PubMed]

28. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef] [PubMed]
29. Zhang, J.; Ma, Z.; Kurgan, L. Comprehensive review and empirical analysis of hallmarks of DNA-, RNA-and protein-binding residues in protein chains. *Brief. Bioinform.* **2017**, 1–19. [CrossRef] [PubMed]
30. Zhang, J.; Kurgan, L. Review and comparative assessment of sequence-based predictors of protein-binding residues. *Brief. Bioinform.* **2017**, bbx022. [CrossRef] [PubMed]
31. Chai, H.; Zhang, J. Identification of Mammalian Enzymatic Proteins Based on Sequence-Derived Features and Species-Specific Scheme. *IEEE Access* **2018**, *6*, 8452–8458. [CrossRef]
32. Chen, Z.; Chen, Y.-Z.; Wang, X.-F.; Wang, C.; Yan, R.-X.; Zhang, Z. Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. *PLoS ONE* **2011**, *6*, e22930. [CrossRef] [PubMed]
33. Zhang, J.; Gao, B.; Chai, H.; Ma, Z.; Yang, G. Identification of DNA-binding proteins using multi-features fusion and binary firefly optimization algorithm. *BMC Bioinform.* **2016**, *17*, 323. [CrossRef] [PubMed]
34. Li, C.-H.; Tu, S.-C. Active site hydrophobicity is critical to the bioluminescence activity of Vibrio harveyi luciferase. *Biochemistry* **2005**, *44*, 12970–12977. [CrossRef] [PubMed]
35. Iden, S.; Collard, J.G. Crosstalk between small GTPases and polarity proteins in cell polarization. *Nat. Rev. Mol. Cell Biol.* **2008**, *9*, 846–859. [CrossRef] [PubMed]
36. Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. Prediction of absolute solvation free energies using molecular dynamics free energy perturbation and the OPLS force field. *J. Chem. Theory Comput.* **2010**, *6*, 1509–1519. [CrossRef] [PubMed]
37. Randic, M. Novel shape descriptors for molecular graphs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 607–613. [CrossRef] [PubMed]
38. Schuler, B.; Lipman, E.A.; Eaton, W.A. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature* **2002**, *419*, 743–747. [CrossRef] [PubMed]
39. Nie, L.; Wu, G.; Zhang, W. Correlation between mRNA and protein abundance in Desulfovibrio vulgaris: A multiple regression to identify sources of variations. *Biochem. Biophys. Res. Commun.* **2006**, *339*, 603–610. [CrossRef] [PubMed]
40. Samanta, U.; Bahadur, R.P.; Chakrabarti, P. Quantifying the accessible surface area of protein residues in their local environment. *Protein Eng.* **2002**, *15*, 659–667. [CrossRef] [PubMed]
41. Skopp, G.; Pötsch, L.; Mauden, M.; Richter, B. Partition coefficient, blood to plasma ratio, protein binding and short-term stability of 11-nor-Δ 9-carboxy tetrahydrocannabinol glucuronide. *Forensic Sci. Int.* **2002**, *126*, 17–23. [CrossRef]
42. Kerwin, B.A. Polysorbates 20 and 80 used in the formulation of protein biotherapeutics: Structure and degradation pathways. *J. Pharm. Sci.* **2008**, *97*, 2924–2935. [CrossRef] [PubMed]
43. Edwards, A.S.; Scott, J.D. A-kinase anchoring proteins: Protein kinase A and beyond. *Curr. Opin. Cell Biol.* **2000**, *12*, 217–221. [CrossRef]
44. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]

Sample Availability: Samples of the compounds are not available from the authors.