Characterization of peptides in the SwePep database using physico-chemical descriptors and multivariate data analysis

UPPSALA UNIVERSITET

Mathias Norrman, Maria Falth, Karl Skold, Marcus Svensson, Anna Nilsson, Per E Andren

Laboratory for Biological and Medical Mass Spectrometry, Uppsala University, PO Box 594, SE-75124 Uppsala, Sweden

OVERVIEW

- ***Purpose*** Theoretical characterisation of the endogenous peptides in the SwePep database.
- ***Method*** The characterisation was made by using calculated physico-chemical descriptors related to size, electrostatics and hydrophobicity.
- ***Results*** The relationships between the peptides as well as their descriptors are visualized with principal component analysis.

INTRODUCTION

This work serves to provide an overview of the of the physico-chemical space that the peptides in the SwePep database (www.swepep.com) comprise. The database contains endogenous peptides and is designed to primarily target mass spectrometry users. Most of the peptides are extracted from peptide annotations in the SwissProt database but there are also peptides described in literature and several newly discovered peptides (neuropeptides). Many of the ~3000 peptides in the database act or could act as therapeutic peptides, and knowledge about the chemical properties of the amino acids composing the peptides could be used to find similar peptides and to increase the understanding of their biological role.

In the last two decades, a number of studies have used amino acid descriptors to describe properties or activities of proteins or peptides. These studies are often specific for a certain application i.e. chromatographic retention data or receptor activity. The population in our database is very diverse and the purpose of this study was to give a general picture of the physico-chemical space that the peptides in the SwePep database comprise. It was also in our interest to have easily interpretable descriptors. Therefore, the descriptors used are of a general nature, reflecting basic properties related to hydrophobicity and electrostatics. Descriptors are partly found in the literature and partly calculated by in house softwares.

METHODS

DESCRIPTOR CALCULATION

A number of different descriptors were calculated for each of the three classes; size, hydrophobicity and electrostatics. Most of the hydrophobic properties were taken from amino acid hydrophobicity indices described in the literature, like Kyte & Doolittle, Meek, Guo. The electrostatic descriptors are related to number of positive/negative charges, number of hydrogen donors/acceptors, pI, etc. Since the peptides all vary in length, and since it was in the interest of this study to get a general description of the peptide properties, all descriptors were represented both as a sum over the whole peptide, and as an average by normalising against number of amino acids in the peptide. Thus, no position specific properties were calculated. Further, no post translational modifications were taken into account nor any three-dimensional information.

MODELLING

The data set was analyzed with Principal Component Analysis (PCA) (Wold et al. 1987) to get a visual overview of the relationships between the peptides, and the descriptors. The PCA score plots where plotted in two dimensions using the two principal components describing the largest variation in the dataset (PC1 and PC2). To get an less biased result due to homologous sequences in other species, the results shown here are based on human peptides. The distribution of the human peptide data set was also compared with three separate datasets of protein sequences. The sequences were randomly selected from the SwissProt database, release 47, and the sequence length were randomly varied between 1 and 60 amino acids.

To investigate whether the descriptors actually measure relevant properties, a model to predict the peptide retention time on a reversed phase chromatography system was built using peptides with known retention time as training set. The peptide retention data was collected on a reversed phase C18 capillary column, (LC Packings) during a 60 min gradient run with 35% acetonitrile and 0.25% acetic acid with a flow rate of 150 nl/min. The RP-column was directly coupled to an ESI Q-TOF mass spectrometer for peptide identification.

A set of seven peptides were selected as training set. All three types of descriptors were used (in total 41 descriptors) to describe the peptide properties. A model was then developed to fit the descriptors to the retention time, by using partial least square (PLS). The resulting cross-validated model was then used to predict the retention time of another set of nine peptides, not included in the training set. All PCA and PLS studies were done using the Simca-P+ software.

RESULTS

PEPTIDE PROPERTIES

A PCA score plot where the human peptides in the SwePep database are combined with one of the sequence datasets randomly built from the SwissProt database is shown in Figure 1. The analysis is based on all average scaled descriptors. The SwissProt sequences are more uniformly distributed in the two dimensions, while the distribution of the endogenous human peptides in SwePep almost have a "<"-shape. The first component explains most of the variation, 49.5%, and the second 21.1%. Five components describe in total 95.3% of the variation in the dataset, with a predictive power of 90.1%.

Figure 2 gives the loading plot, i.e. the correlation pattern between the descriptor values and the peptide pattern. Most of the difference in the first dimension is explained by hydrophobicity related descriptors, while charge related descriptors dominate the second dimension. The "<"-shaped distribution of the SwePep peptides in Fig.1 suggest that these peptides are more polar and have more positive and negative charges. This is also reflected in the amino acid statistics, Table 1. There is a higher frequency of Arginine and Glutamate in the SwePep peptides compared to the SwissProt sequences used in this study and the overall protein statistics, taken from Creighton (1997). There is also a two fold higher frequency of Cysteine in the SwePep peptides.

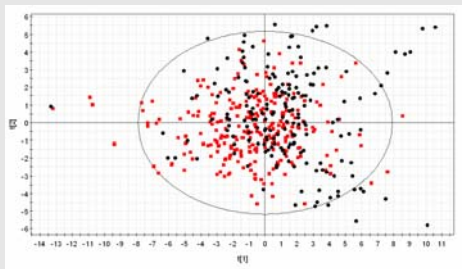

Figure 1. Score plot of human peptides in the SwePep database (black) and one of the sequence dataset randomly built from the SwissProt database (red). Five components describe in total 95.3% of the variation in the dataset, with a predictive power of 90.1%.

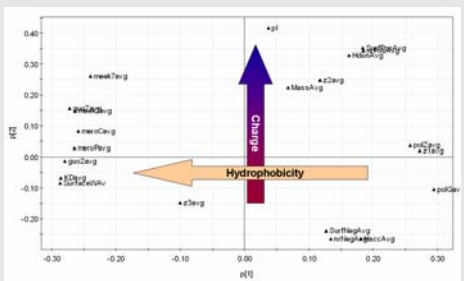

Figure 2. The correlation pattern between the descriptor values are given in a loading plot. Hydrophobicity is the dominating property in the first dimension, while charge dominates the second.

Table 1. Amino acid frequency of occurrence in the SwePep database (human peptides) compared with tabulated values in Creighton (1997) and a randomly selection of amino acid sequences from SwissProt. The frequencies are compared against the Creighton values as a ratio. Colored numbers indicate a > 10% deviation from the Creighton frequencies (red lower, blue higher)

	Creighton	SwePep human peptides non-redundant (n = 226)		Random SwissProt sequences (n = 268)	
AminoAcid	Freq.	Freq.	Ratio[1]	Freq.	Ratio[1]
Ala	8.3	6.93	0.84	8.08	0.97
Arg	5.7	6.79	1.19	4.96	0.87
Asn	4.4	3.49	0.79	4.23	0.96
Asp	5.3	5.57	1.05	5.45	1.03
Cys	1.7	3.74	2.20	1.63	0.96
Gln	4	4.96	1.24	3.82	0.95
Glu	6.2	7.34	1.18	6.40	1.03
Gly	7.2	8.17	1.13	7.65	1.06
His	2.2	3.96	1.80	2.29	1.04
Ile	5.2	2.54	0.49	6.14	1.18
Leu	9	7.77	0.86	9.47	1.05
Lys	5.7	6.02	1.06	5.39	0.95
Met	2.4	2.22	0.92	2.35	0.98
Phe	3.9	4.03	1.03	4.44	1.14
Pro	5.1	5.79	1.14	4.55	0.89
Ser	6.9	7.94	1.15	6.21	0.90
Thr	5.8	3.81	0.66	5.68	0.98
Trp	1.3	1.48	1.14	1.27	0.98
Tyr	3.2	3.21	1.00	3.14	0.98
Val	6.6	4.25	0.64	6.84	1.04

[1] Amino acid frequency divided by tabulated Creighton frequencies.

PREDICTION

The prediction model used to predict the retention time on a RPC system was based on the seven peptides shown in black in the score plot, Fig. 3A, the test peptides, not included in the model set are shown in red. (The loading plot, describing the descriptor distribution is not shown.) A cross-validated two component model was obtained with an explained variation (R2) of 87% and a predictive power Q2 of 75%. The observed versus predicted retention time are shown in Figure 3B. The variable importance is shown in Figure 4. As expected, the hydrophobicity related descriptors are the dominating.

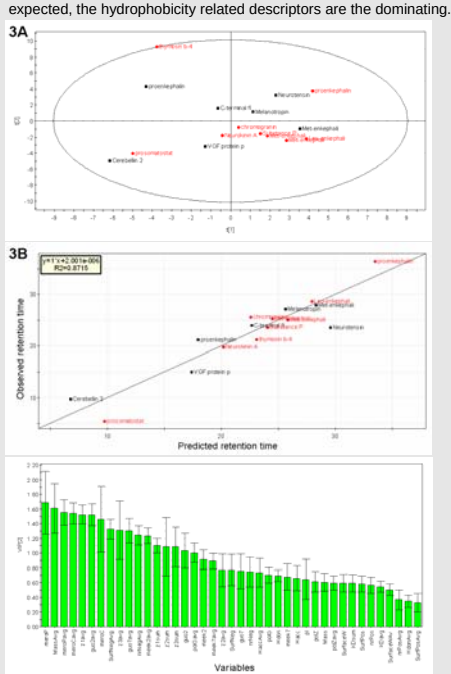

Figure 4. Variable importance in the prediction model.

CONCLUSIONS

The physico-chemical properties of the endogenous peptides in the SwePep database have been characterized and the relationship between the peptides has been visualized with principal component analysis. A predictive model for the retention time on an RPC system was built to verify the usefulness of the descriptors. The quality of the model suggest that the descriptors actually measure relevant properties.

Besides being used for retention prediction, the calculated data should be useful for finding peptides with related properties, which could be useful both during peptide identification processes as well as in biological activity studies.

ACKNOWLEDGEMENTS

This work was supported by the Swedish Research Council, Medicine #11565, Swedish Foundation for International Cooperation in Research and Higher Education (STINT), the Knowledge Foundation, Karolinska Institute Programme in Medical Bioinformatics, and K&A Wallenberg Foundation.