

2007/006 WO

TITLE

5 Method for producing peptide libraries and use thereof

FIELD OF THE INVENTION

10 The present invention relates to the field of computational biochemistry and computer aided design of bioactive peptides. It combines methods used in biological sequence analysis, bioinformatics data mining, information representation and classification algorithms using supervised learning. In addition it relates to the design of peptide libraries and the use of bioactive peptides for biomedical research.

15

BACKGROUND OF THE INVENTION

A primary goal of drug discovery today is to identify biologically active molecules that have practical clinical utility. Many, if not all, biologically active peptides (e.g. peptide hormones) have profound effects both in health and disease, either by growth stimulating roles, growth inhibitory roles, or the regulation of critical metabolic pathways.

25 Peptide hormones are produced as precursors in different cell types and organs like glands, neurons, intestine, brain, etc. Peptide hormones are initially synthesized as larger precursors, or prohormones, and may acquire a number of post-translational modifications during transportation through the ER and Golgi stacks. They are processed and transported to their final destination to act as active substances (first messengers) to trigger a cellular response by binding to a cell surface receptor.

30 Peptide hormones are the key messengers in many physiological processes including regulation of production; growth; water and salt metabolism; temperature control; cardiovascular, gastrointestinal, and respiratory control; behavior; memory; and affective states.

35 Peptide hormones play a key role in physiological processes that are relevant to many areas of biomedical research such as diabetes (Insulin), blood pressure regulation (Angiotensin), anemia (Erythropoietin- α), multiple sclerosis (Interferon- β), obesity (Leptin) and others.

Therefore, novel bioactive peptides have the potential to be used as therapeutic polypeptides, targets for drug intervention, ligands to discover relevant targets (eg. GPCR deorphaning) or biomarkers to monitor diseases.

- 5 Peptide libraries have successfully been used to identify bioactive peptides, including antimicrobial peptides, receptor agonists and antagonists, ligands for cell surface receptors, protein kinase inhibitors and substrates, T-cell epitopes, peptides binding to MHC molecules and peptide mimotopes of receptor binding sites. Peptide libraries can be categorized according to their origin in gene- and synthetic based libraries (Falciani et al., 2005).
- 10 In gene based libraries the combinatorial positions within the polypeptides are introduced at the DNA level that encodes the sequence of the target polypeptide in order to introduce diversity. In contrast to the gene based libraries, synthetic libraries achieve their diversity at the level of chemical synthesis.
- Many peptide libraries are based on one scaffold or use a random combinatorial approach to
- 15 generate different polypeptide primary structures.

- The disadvantage of both approaches is that the combination of the 20 naturally occurring amino acids allows the construction of polypeptides which are most variable and account for a very large number of different structures. To give an example on how many different
- 20 structures can be obtained, consider the 160.000 different primary structure possibilities for a peptide containing only 4 amino acids.

- There was a need to provide an accurate and high-throughput method to significantly reduce the potential number of structures in a peptide library, to enable the processing of large
- 25 amounts of data and to distinguish between peptides that have an activity *in vivo* and peptides that do not have an activity *in vivo*.

- The object of the present invention solves the problem of the prior art. The present invention relates to a method to construct novel bioactive peptide hormone libraries using a
- 30 bioinformatics strategy. A support vector machine (SVM) algorithm is used to identify bioactive peptides. This method allows to discovering potential bioactive peptide hormones *in silico* searching the human proteome by taking advantage of the conserved protein features and short motifs present in peptide hormone precursors. While these features are common to peptide hormones and are responsible for their maturation, there is, surprisingly,
- 35 very little sequence similarity between peptide hormone precursors that would allow a database search on the protein sequence level alone (e.g. BLAST, FASTA). However, combinations of co-occurring protein features and motifs for post-translational modifications

in peptide hormone precursors (e.g. short protein sequence length of precursor, signal peptide, disulfide bonds, amidation sites, sulfation sites, glycosylation sites, etc) can be used to discover novel peptide hormones with a high specificity.

SUMMARY

In a first aspect, there is provided, in a computer-based system, a method for identifying bioactive peptides derived from protein precursor sequences using a binary support vector machine (SVM) based algorithm, said method comprising the steps of:

- a) training a SVM algorithm to learn to distinguish between bioactive and non bioactive peptides, said training comprising the steps of:
 - a₁) generating vectors with 49 dimensions, each dimension resulting from the calculation of a molecular descriptor value, for a set of labelled known bioactive peptides derived from precursor sequences and labelled known non bioactive peptides derived from precursor sequences, where the labels indicate whether the peptide is, respectively, bioactive or non bioactive;
 - a₂) transferring the vector data generated in step a₁) to the SVM based algorithm, said algorithm calculating the optimal hyperplane that separates the vectors corresponding to the bioactive peptides and the non bioactive peptides, respectively;
- b) providing protein precursor sequences from a publicly available human protein database;
- c) predicting secondary structure and cleavage sites within a protein precursor sequence provided in step b) using computational methods; a set of 7 molecular descriptors is calculated based on said prediction step resulting in the generation of peptide fragments;
- d) calculating a set of 42 molecular descriptors corresponding to the physico-chemical properties of the peptide fragments generated in step c);
- e) transforming the calculated values from step c) into scaled values between 0 and 1 to generate dimensions 1 to 7 of a 49-dimension-vector for each peptide fragment and transforming the calculated values from step d) into scaled values between 0 and 1 to generate dimensions 8 to 49 of said vector for each peptide fragment;

- f) presenting the vectors generated in the step e) to the trained SVM algorithm from step a) to measure the distance of each vector to the hyperplane calculated in step a₂); and
- g) classifying each peptide fragment as bioactive peptide or non bioactive peptide, according to the distance measured in step f), wherein dimensions 1 to 7 generated in step e) are the following: Dimension 1: N-terminal ProP score; Dimension 2: N-terminal Hmcut score; Dimension 3: N-terminal fragment; Dimension 4: C-terminal ProP score; Dimension 5: C-terminal Hmcut score; Dimension 6: C-terminal Hamid score; Dimension 7: C-terminal fragment.

In a second aspect, there is provided a computational device configured to identify bioactive peptides derived from protein precursor sequences by using a binary support vector machine (SVM) based method, said method comprising the steps of:

- a) training a SVM algorithm to learn to distinguish between bioactive and non bioactive peptides, said training comprising the steps of:
 - a₁) generating vectors with 49 dimensions, each dimension resulting from the calculation of a molecular descriptor value, for a set of labelled known bioactive peptides derived from precursor sequences and labelled known non bioactive peptides derived from precursor sequences, where the labels indicate whether the peptide is, respectively, bioactive or non bioactive;
 - a₂) transferring the vector data generated in step a₁) to the SVM based algorithm, said algorithm calculating the optimal hyperplane that separates the vectors corresponding to the bioactive peptides and the non bioactive peptides, respectively;
- b) providing protein precursor sequences from a publicly available human protein database;
- c) predicting secondary structure and cleavage sites within a protein precursor sequence provided in step b) using computational methods; a set of 7 molecular descriptors is calculated based on said prediction step resulting in the generation of peptide fragments;
- d) calculating a set of 42 molecular descriptors corresponding to the physico-chemical properties of the peptide fragments generated in step c);
- e) transforming the calculated values from step c) into scaled values between 0 and 1 to generate dimensions 1 to 7 of a 49-dimension-vector for each peptide fragment and transforming the calculated values from step d) into scaled values between 0 and 1 to generate dimensions 8 to 49 of said vector for each peptide fragment;

3d

- f) presenting the vectors generated in the step e) to the trained SVM algorithm from step a) to measure the distance of each vector to the hyperplane calculated in step a₂); and
- g) classifying each peptide fragment as bioactive peptide or non bioactive peptide, according to the distance measured in step f), wherein dimensions 1 to 7 generated in step e) are the following: Dimension 1: N-terminal ProP score; Dimension 2: N-terminal Hmcut score; Dimension 3: N-terminal fragment; Dimension 4: C-terminal ProP score; Dimension 5: C-terminal Hmcut score; Dimension 6: C-terminal Hamid score; Dimension 7: C-terminal fragment.

In a third aspect, there is provided a use of the method as defined above for the identification of therapeutic polypeptides, targets for drug intervention, ligands to discover relevant targets or biomarkers to monitor diseases.

One subject-matter of the present disclosure refers to a method for identifying bioactive peptides using a binary support vector machine (SVM) based algorithm in a computer-based system, wherein:

- a) a SVM algorithm is trained to learn to distinguish between bioactive and non bioactive peptides, said training comprising the steps of:
 - a₁) generating vectors with 49 dimensions, each dimension resulting from the calculation of a molecular descriptor value, for a set of labelled known bioactive and labelled known non bioactive peptides, where the labels indicate whether the peptide is, respectively, bioactive or non bioactive;
 - a₂) transferring the vector data generated in step a₁) to the SVM based algorithm, said algorithm calculating the optimal hyperplane that separates the vectors corresponding to the bioactive peptides and the non bioactive peptides, respectively;
- b) protein sequences are provided from a publicly available human protein database;
- c) secondary structure and cleavage sites within a protein sequence provided in step b) are predicted using computational methods; a set of 7 molecular descriptors is calculated based on said prediction step resulting in the generation of peptide fragments;
- d) a set of 42 molecular descriptors corresponding to the physico-chemical properties of the peptide fragments generated in step c) is calculated;
- e) the calculated values from step c) are transformed into scaled values between 0 and 1 to generate dimensions 1 to 7 of a 49-dimension-vector for each peptide fragment

and the calculated values from step d) are transformed into scaled values between 0 and 1 to generate dimensions 8 to 49 of said vector for each peptide fragment;

f) the vectors generated in the step e) are presented to the trained SVM algorithm from step a) to measure the distance of each vector to the hyperplane calculated in step a₂); and

g) each peptide fragment is classified as bioactive peptide or non bioactive peptide, according to the distance measured in step f).

In general, dimensions 1 to 7 generated in step e) are the following: Dimension 1: N-terminal ProP score; Dimension 2: N-terminal Hmcut score; Dimension 3: N-terminal fragment; Dimension 4: C-terminal ProP score; Dimension 5: C-terminal Hmcut score; Dimension 6: C-terminal Hamid score; Dimension 7: C-terminal fragment; and dimensions 8 to 49 generated in step e) are the following: Dimension 8: Percentage of acidic amino acids (E, N, Q) per polypeptide; Dimension 9: Percentage of positively charged amino acids (R, H) per polypeptide; Dimension 10: Percentage of aromatic amino acids (F, Y, W) per polypeptide; Dimension 11: Percentage of aliphatic amino acids (G, V, A, I) per polypeptide; Dimension 12: Percentage of Proline per polypeptide; Dimension 13: Percentage of reactive amino acids (S, T) per polypeptide; Dimension 14: Percentage of Alanine per polypeptide; Dimension 15: Percentage of Cysteine per polypeptide; Dimension 16: Percentage of Glutamic acid per polypeptide; Dimension 17: Percentage of Phenylalanine per polypeptide; Dimension 18: Percentage of Glycine per polypeptide; Dimension 19: Percentage of Histidine per polypeptide; Dimension 20: Percentage of Isoleucine per polypeptide; Dimension 21: Percentage of Asparagine per polypeptide; Dimension 22: Percentage of Glutamine per polypeptide; Dimension 23: Percentage of Arginine per polypeptide; Dimension 24: Percentage of Serine per polypeptide; Dimension 25: Percentage of Threonine per polypeptide; Dimension 26: Percentage of non-canonical amino acid per polypeptide; Dimension 27: Percentage of Valine per polypeptide; Dimension 28: Percentage of Tryptophane per polypeptide; Dimension 29: Percentage of Tyrosine per polypeptide; Dimension 30: Cysteine content; Dimension 31: Percentage of coiled secondary structure per polypeptide; Dimension 32: Percentage of helical secondary structure per polypeptide; Dimension 33: Percentage of random secondary structure per polypeptide; Dimension 34: Score for structure around N-terminal cleavage site; Dimension 35: Score for structure around C-terminal cleavage site; Dimension 36: Number of helical blocks per polypeptide; Dimension 37: Isoelectric point of polypeptide; Dimension 38: Average molecular weight of polypeptide; Dimension 39: Sum of Van-der-Waals forces of each amino acid within polypeptide; Dimension 40:

Sum of hydrophobicity values of each amino acid within polypeptide; Dimension 41 – 48: Mean values calculated based on principle component score vectors of hydrophobic, steric, and electronic properties per polypeptide; Dimension 49: Length of polypeptide.

- 5 In a preferred embodiment of the method of the present disclosure, protein sequences from step b) are only naturally occurring protein sequences found in the human secretome.

In another preferred embodiment, bioactive peptides are bioactive peptide hormones derived from precursor hormones.

10

Another subject-matter of the present disclosure refers to a bioactive peptide selected from the human secretome by using the method of the present disclosure.

- 15 In a preferred embodiment, the bioactive peptide is a bioactive peptide hormone. In a more preferred embodiment, the bioactive peptide hormone derives from a precursor protein.

In another preferred embodiment, the bioactive peptide has a sequence selected from the group consisting of the amino acid sequences of SEQ ID NOS: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185.

- 25
30 The disclosure pertains further to a peptide library comprising bioactive peptides identified through the method of the present disclosure.

In a preferred embodiment, the peptide library comprises bioactive peptides having a sequence selected from the group consisting of the amino acid sequences of SEQ ID NOS 1-185 cited above.

35

In a more preferred embodiment, the peptide library comprises bioactive peptide hormones.

In another more preferred embodiment, the peptide library comprises bioactive peptide hormones derived from precursor proteins.

Another subject-matter of the present disclosure refers to a computational device configured to identify bioactive peptides by using a binary support vector machine (SVM) based method, wherein:

a) a SVM algorithm is trained to learn to distinguish between bioactive and non bioactive peptides, said training comprising the steps of:

a₁) generating vectors with 49 dimensions, each dimension resulting from the calculation of a molecular descriptor value, for a set of labelled known bioactive and labelled known non bioactive peptides, where the labels indicate whether the peptide is, respectively, bioactive or non bioactive;

a₂) transferring the vector data generated in step a₁) to the SVM based algorithm, said algorithm calculating the optimal hyperplane that separates the vectors corresponding to the bioactive peptides and the non bioactive peptides, respectively;

b) protein sequences are provided from a publicly available human protein database;

c) secondary structure and cleavage sites within a protein sequence provided in step b) are predicted using computational methods; a set of 7 molecular descriptors is calculated based on said prediction step resulting in the generation of peptide fragments;

d) a set of 42 molecular descriptors corresponding to the physico-chemical properties of the peptide fragments generated in step c) is calculated;

e) the calculated values from step c) are transformed into scaled values between 0 and 1 to generate dimensions 1 to 7 of a 49-dimension-vector for each peptide fragment and the calculated values from step d) are transformed into scaled values between 0 and 1 to generate dimensions 8 to 49 of said vector for each peptide fragment;

f) the vectors generated in the step e) are presented to the trained SVM algorithm from step a) to measure the distance of each vector to the hyperplane calculated in step a₂); and

- g) each peptide fragment is classified as bioactive peptide or non bioactive peptide, according to the distance measured in step f).

The disclosure pertains further to the use of the method of the present disclosure for the identification of therapeutic polypeptides, targets for drug intervention, ligands to discover relevant targets or biomarkers to monitor diseases.

The disclosure pertains further to the use of the peptide library of the present disclosure in a screening approach to interrogate intracellular signalling pathways; to create reagents to further the understanding of a pathway; to create novel forms of therapies and to identify pharmaceutically active compounds, targets for drug intervention, ligands to discover relevant targets or biomarkers to monitor diseases.

The disclosure is also directed to a pharmaceutical composition comprising a bioactive peptide having a sequence selected from the group consisting of the amino acid sequences of SEQ ID NOS 1-185 as bioactive agent.

DETAILED DESCRIPTION OF THE INVENTION

The present invention is directed to novel bioactive polypeptides and to an *in silico* method to identify such bioactive polypeptides.

In the present invention, a polypeptide is considered as bioactive if it has an interaction with or an effect on any cell tissue in the human body. Bioactive peptides have the potential to be used as therapeutic polypeptides, targets for drug intervention, ligands to discover relevant targets (eg. GPCR deorphaning) or biomarkers to monitor diseases. Bioactive peptides include, among others, bioactive peptide hormones. Peptide hormones are characterized by their high specificity as well as their effectiveness in very low concentrations. Peptide hormones are initially synthesized as larger precursors, or prohormones.

A precursor is a substance from which another usually more active or mature substance is formed. A protein precursor is an inactive protein (or peptide) that can be turned into an active form by post-translational modification. Several cleavage sites are involved in the modification of the precursor to produce the mature protein: signal sequence cleavage sites, protease cleavage sites, amidation sites, etc.

The name of the precursor for a protein is often prefixed by pro or pre. Precursors are often used by an organism when the subsequent protein is potentially harmful, but needs to be available on short notice and/or in large quantities.

5 The terms "polypeptide", "peptide" and "protein" are used interchangeably herein to refer to a polymer consisting of amino acid residues linked by covalent bonds. These terms include parts or fragments of full length proteins, such as, for example, peptides, oligopeptides and shorter peptide sequences consisting of at least 2 amino acids, more particularly peptide sequences consisting of 4-45 amino acids.

10 In addition, these terms include polymers of modified amino acids, including amino acids which have been post-translationally modified, for example by chemical modification including but not restricted to amidation, glycosylation, phosphorylation, acetylation and/or sulphation reactions that effectively alter the basic peptide backbone. Accordingly, a polypeptide may be derived from a naturally-occurring protein, and in particular may be
15 derived from a full-length protein by chemical or enzymatic cleavage, using reagents such as CNBr, or proteases such as trypsin or chymotrypsin, amongst others. Alternatively, such polypeptides may be derived by chemical synthesis using well known peptide synthetic methods.

20 An amino acid is any molecule that contains both amine and carboxylic acid functional groups. Amino acid residue is what is left of an amino acid once a molecule of water has been lost (an H⁺ from the nitrogenous side and an OH⁻ from the carboxylic side) in the formation of a peptide bond, the chemical bond that links the amino acid monomers in a protein chain.

25 Each protein has its own unique amino acid sequence that is known as its primary structure. Primary structure is fairly straightforward and refers to the number and sequence of amino acids in the protein or polypeptide chain. The covalent peptide bond is the only type of bonding involved at this level of protein structure. The sequence of amino acids in a protein
30 is dictated by genetic information in DNA, which is transcribed into RNA, which is then translated into protein. So protein structure is genetically determined.

The next level of protein structure generally refers to the amount of structural regularity or shape that the polypeptide chain adopts. A natural polypeptide chain will spontaneously fold into a regular and defined shape. Two main types of secondary structure have been found in
35 proteins namely α -helix, and β -pleated sheet.

The tertiary structure of a polypeptide chain is the next level of conformation or shape adopted by the α -helices or β -pleated sheets of the chain. Most proteins tend to fold

into shapes that are broadly classified as globular in arrangement, and some, particularly structural proteins form long fibres. These are the main forms of gross tertiary structure. A term often used is domain, which refers to a compact unit of globular structure in a polypeptide chain.

- 5 The unique shape of each protein determines its function in the body.

Also included within the scope of the definition of a "polypeptide" are amino acid sequence variants. These may contain one or more preferably conservative, amino acid substitutions, deletions, or insertions, in a naturally-occurring amino acid sequence which do not alter at least one essential property of said polypeptide, such as, for example, its biological activity. Such polypeptides may be synthesized by chemical polypeptide synthesis. Conservative amino acid substitutions are well-known in the art. For example, one or more amino acid residues of a native protein can be substituted conservatively with an amino acid residue of similar charge, size or polarity, with the resulting polypeptide retaining functional ability as described herein. Rules for making such substitutions are well known.

More specifically, conservative amino acid substitutions are those that generally take place within a family of amino acids that are related in their side chains.

Genetically-encoded amino acids are generally divided into four groups: (1) acidic=aspartate, glutamate; (2) basic=lysine, arginine, and histidine; (3) non-polar=alanine, valine, leucine, isoleucine, proline, phenylalanine, methionine, and tryptophan; and (4) unchargedpolar=glycine, asparagine, glutamine, cysteine, serine, threonin, and tyrosine.

Phenylalanine, tyrosine and tryptophan are also jointly classified as aromatic amino acids. One or more replacements within any particular group such as, for example, the substitution of leucine for isoleucine or valine are alternatively, the substitution of aspartate for glutamate or threonin for serine, or of any other amino acid residue with a structurally-related amino acid residue will generally have an insignificant effect on the function of the resulting polypeptide.

Included in the scope of the definition of the term "polypeptide" is a peptide whose biological activity is predictable as a result of its amino acid sequence corresponding to a functional domain. Also encompassed by the term "polypeptide" is a peptide whose biological activity could not have been predicted by the analysis of its amino acid sequence.

In the present invention, a support vector machine algorithm (SVM) is used to distinguish between polypeptides that have an activity in vivo and polypeptides that do not have an activity in vivo.

Support Vector Machine (SVM):

A Support Vector Machine (SVM) is a universal learning machine that, during a training phase, determines a decision surface or "hyperplane". The decision hyperplane is determined by a set of support vectors selected from a training population of vectors and by a set of corresponding multipliers. The decision hyperplane is also characterised by a kernel function.

The mathematical basis of a SVM is explained in the book by John Shawe Taylor & Nello Cristianini - Cambridge University Press, 2000, entitled "Support Vector Machines and other kernel-based learning methods" and in an article by Chih-Chung Chang and Chih-Jen Lin entitled "LIBSVM - A Library for Support Vector Machines", 2001.

Subsequent to the training phase, a SVM operates in a testing phase during which it is used to classify test vectors on the basis of the decision hyperplane previously determined during the training phase (Noble, 2006).

Support Vector Machines find application in many and varied fields. For example, in a paper by H. Kim and H. Park entitled "Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3d local descriptor", SVMs are applied to the problem of predicting high resolution 3D structure in order to study the docking of macro-molecules.

In the present invention, a support vector machine algorithm (SVM) is used to distinguish between polypeptides that have an activity in vivo and polypeptides that do not have an activity in vivo.

From a practical point of view, a SVM is implemented by means of a computational device such as a personal computer in the present invention.

The computational device includes one or more processors that execute a succession of different softwares, as described in the exemplification section (1.1.), containing instructions for implementing a method according to the present invention.

Training of SVM and model generation:

In order to train the SVM model, vectors with 49 dimensions were generated using the program routine described in the experimentation section (1.1.) and schematically shown in Fig 1.

- 5 For the SVM training set, information on known bioactive peptides can be extracted from any public available human protein database, such as Swissprot. Preferably bioactive peptides with a length between 4 and 55 amino acids were extracted from their precursor according to their annotation in Swissprot and labeled as positive examples used for training of the SVM algorithm. All other fragments generated with a length between 4-55 amino acids from the
- 10 same known peptide hormone precursors that have no assigned function were used as negative trainings set for SVM training. As the SVM is a binary system, bioactive peptides were labeled as +1 and non bioactive peptides were labeled as -1.
- Similarly, bioactive and non bioactive peptides with a length between 56 and 300 amino acids were used to train a second model to predict longer polypeptides. In order not to over-
- 15 represent negative examples, the final SVM training sets for short (4-55 amino acids) and long (56-300 amino acids) respectively were adjusted to an equal number of positive and negative training data by randomly selecting the same number of negatives from all negative peptides.
- 20 To transform the information hidden in the bioactive and non bioactive peptides, a set of 49 descriptors was defined and used for training of a SVM. The performance of a SVM model strongly depends on the quality of the chosen descriptors used to describe the peptides. In the present invention, the first 7 descriptors reflect the likelihood of a polypeptide to be produced by the human body. These 7 dimensions were calculated by employing a set of
- 25 protease prediction site tools to the peptide hormone precursor sequence (Fig1). The resulting scores of each program output were directly used as descriptors. The remaining 42 dimensions reflect important physico-chemical properties of each generated fragment (i.e. a bioactive or a non bioactive peptide). The 49 descriptors used in the present invention are listed in the point 3 of the exemplification section.
- 30 To each peptide corresponds a unique combination of 49 descriptors. The different peptides can be represented as points in a multidimensional space where each dimension corresponds to one of the descriptors. The SVM seeks to find a boundary that best separates the two sets of points corresponding to the bioactive and the non bioactive
- 35 peptides. This boundary is called the optimal hyperplane that best separates the two classes of objects in an n-dimensional space, namely the vectors corresponding to the bioactive peptides and the non bioactive peptides, respectively.

The resulting SVM models learn to distinguish between bioactive and non bioactive peptides. The best model is chosen which has the highest performance based on the ranking of an independent test set of bioactive and non bioactive peptides. To test the models, the performance of all generated models was tested and the two best models for short peptides (4-55 amino acids) and longer polypeptides (56-300 amino acids) were chosen, respectively.

Identification of bioactive peptides:

After training, the resulting trained SVM model is able to identify bioactive peptides for which no bioactivity had been characterized.

A schematic overview of the method disclosed in the invention is given in Fig.1 to explain the steps involved in peptide library generation. As input value a protein sequence provided from a publicly available human protein database, such as Swissprot, is used. In step 1, all potential protease cleavage sites are predicted using a set of tools to predict these events. The respective cleavage site positions are saved for each precursor sequence. In addition, the secondary structure is deduced for the entire protein precursor sequence. Based on the predicted cleavage sites within the precursor sequence, all potential fragments are generated (step 2) and are used as input for step 3.

Step 3 comprises the calculation of physico-chemical properties of each peptide fragment (list in point 3 of the exemplification section). In general, information on the amino acid frequency within each fragment, the secondary structure of each fragment, the isoelectric point of each fragment, average molecular mass of each fragment, hydrophobicity of each fragment, the sum of all van-der-Waals forces for each amino acid within the fragment, the sum of all commonly used amino acid descriptors (i.e. VHSE value for each amino acid based on Mei et al., 2005) for each amino acid within the fragment and the fragment length are taken into account to transform the biological information into numerical values.

Calculated values from step 1 and 3 are transformed in steps 4a and 4b to give scaled values between 0 and 1, respectively, to generate a 49 dimensional vector for each fragment. In step 5 the vectors are presented to the trained SVM model to measure the distance of each vector to the hyperplane. The SVM output is then used in step 6 to decide whether the peptide is likely to be bioactive or not. 49-dimensional vectors corresponding to the bioactive peptides identified through the method of the present invention are listed in figure 3.

In order to significantly reduce the potential number of structures in a peptide library, in the present invention only naturally occurring protein sequences found in the human secretome

were used as primary structures to generate peptide libraries. The human secretome is the whole information encoded in the DNA that corresponds to all human proteins that are secreted by the cells.

Potentially secreted human proteins which were used as precursor sequences to find novel bioactive peptides were extracted from the publicly available sequence databases listed in point 1.1. of the exemplification section.

Distinct parts of the primary sequences of secreted proteins, i.e. protein precursors, were used as templates to deduce novel bioactive peptides. The peptide length was restricted to 4-45 amino acids to render the peptides amenable to chemical synthesis.

Subsequent to the identification of novel bioactive peptides through the method of the present invention, antimicrobial assays were performed to test the bioactivity of the latter peptides. These assays are detailed in point 6 of the exemplification section.

The present invention further relates to a peptide library comprising bioactive peptides identified through the SVM model method described above. The amino acid sequences of the 185 bioactive peptides identified through the method of the present invention and comprised in the peptide library of the present invention are listed in figure 2.

A peptide library is a newly developed technique for protein related study. A peptide library contains a great number of peptides that have a systematic combination of amino acids. Usually, peptide libraries are synthesized on solid phase, mostly on resin, which can be made as flat surface or beads. A peptide library provides a powerful tool for drug design, protein-protein interactions, and other biochemical as well as pharmaceutical applications.

The peptide library of the present invention can be used in a screening approach to interrogate intracellular signalling pathways, to create reagents to further the understanding of a pathway, to create novel forms of therapies and to identify pharmaceutically active compounds, targets for drug intervention, ligands to discover relevant targets or biomarkers to monitor diseases.

The polypeptides of the present invention have hormonal activity. Therefore, the polypeptides of the invention are useful as drugs, for example therapeutic polypeptides, ligands to discover relevant targets (e.g. GPCRs), targets for drug intervention (e.g. targets for monoclonal antibodies, receptor fragments), biomarkers to monitor diseases (in combination with tool antibodies to detect peptide fragments in body fluids), protein kinase inhibitors and substrates, T-cell epitopes, peptide mimotopes of receptor binding sites, etc.

The DNAs coding for the peptide or precursor of the invention are useful, for example, as agents for the gene therapy, treatment or prevention of cardio-vascular diseases, hormone-producing tumours, diabetes, gastric ulcer and the like, hormone secretion inhibitors, tumour growth inhibitors, neural activity etc. Furthermore, the DNAs of the invention are useful as agents for the gene diagnosis of diseases such as cardio vascular disease, hormone-producing tumours, diabetes, gastric ulcer and the like.

EXEMPLIFICATION

The invention now being generally described will be more readily understood in reference to the following examples, which are included merely for purposes of illustration of certain aspects and embodiments of the present invention and are not intended to limit the invention.

1. Databases and computer programs

1.1. Databases

The following publicly available sequence databases were used to extract potentially secreted human proteins which were used as precursor sequences to find novel bio-active peptides:

Human genome (NCBI 33 assembly, 1st of July 2003) translated into protein, subset; International Protein Index, Swissprot (Release 50.3 of 11-July-2006) and TrEMBL (Releases: August 2003 – March 2006);

For training of SVM based algorithms, information on known bioactive peptides was extracted from Swissprot.

1.2. Computer programs

1.1 Signal P version 2.0 (Nielsen et al., 1997)

Objective: This program was used to detect potential signal sequences and determine the potential human secretome. It was used with a cut off score of 0.98. Signal P version 2.0 predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: The method incorporates a prediction of cleavage sites and a

signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks and hidden Markov models.

1.2 ProP version 1.0 (Duckert et al., 2004)

Objective: This program was used to detect potential cleavage sites in protein sequences. The cut off score used was set to 0.11. This program predicts arginine and lysine propeptide cleavage sites in eukaryotic protein sequences using an ensemble of neural networks. Furin-specific prediction is the default. It is also possible to perform a general proprotein convertase (PC) prediction.

1.3. Amidation site prediction and prediction of protease cleavage sites (Rohrer, 2004)

Objective: The program Hamid predicts amidation sites in protein sequences. The program Hmcut predicts protease cleavage sites in protein sequences that take place before a basic amino acid residue (Lys, Arg). Both programs are based on Hidden Markov Models and utilize the software version Hmmer 2.3.2 (Durbin et al. 1998).

1.4 Support Vector Machine (Chang and Lin, 2001)

LIBSVM is an integrated software for support vector classification, (C-SVC, nu-SVC), regression (epsilon-SVR, nu-SVR) and distribution estimation (one-class SVM).

The following SVM specifications were used: SVM_type: nu-SVC; Kernel_type: radial basis function.

1.5. PsiPred version 2.45 (Jones, 1999)

Method for protein secondary structure prediction. The method was used as described in Jones, 1999.

1.6. Calculation of isoelectric points

Objective: Calculation of isoelectric points of polypeptides. This was done according to Gasteiger et al. 2005.

1.7. Perl – Practical extraction and report language

Objective: Perl is a dynamic programming language created by Larry Wall and first released in 1987.

2. Training of SVM

For the supervised learning process, known bioactive polypeptide precursors were extracted from commonly used public databases such as Swissprot using the following SRS (Sequence Retrieval System on www.expasy.org) query statement: Organism = vertebrata; Sequence_length = 30:300; Feature_key = signal; Keywords = cytokine or hormone or bombesin or bradykinin or glucagon or growth factor or insulin or neuropeptide or opioid peptide or tachykinin or thyroid hormone or vasoconstrictor or vasodilator. This query yields a set of known peptide hormone precursors in which their bioactive peptides are readily available by the annotation of the Swissprot database. Therefore these sequences can be used to deduce a set of bioactive and non bioactive peptides for training of an SVM based model.

3. Molecular descriptors used to build the vectors

The performance of a SVM model strongly depends on the quality of the chosen descriptors used to describe the peptides.

In the present invention, the following descriptors were chosen:

Dimension 1-7 represent the likelihood of a polypeptide to be produced in the human body and was calculated by a combination of different protease cleavage site prediction tools. The results of these tools represent in the first 7 dimensions of the vector.

Dimension 1: N-terminal ProP score;

Dimension 2 : N-terminal Hmcut score ;

Dimension 3 : N-terminal fragment (fixed value of 0.2)

Dimension 4 : C-terminal ProP score ;

Dimension 5 : C-terminal Hmcut score ;

Dimension 6 : C-terminal Hamid score ;

Dimension 7 : C-terminal fragment (fixed value of 0.2)

Physico-chemical properties of the polypeptides were calculated and represent the following 42 dimensions of the vector.

Dimension 8: Percentage of acidic amino acids (E, N, Q) per polypeptide

Dimension 9: Percentage of positively charged amino acids (R, H) per polypeptide

Dimension 10: Percentage of aromatic amino acids (F, Y, W) per polypeptide

Dimension 11: Percentage of aliphatic amino acids (G, V, A, I) per polypeptide

- Dimension 12: Percentage of Proline per polypeptide
- Dimension 13: Percentage of reactive amino acids (S, T) per polypeptide
- Dimension 14: Percentage of Alanine per polypeptide
- Dimension 15: Percentage of Cysteine per polypeptide
- 5 Dimension 16: Percentage of Glutamic acid per polypeptide
- Dimension 17: Percentage of Phenylalanine per polypeptide
- Dimension 18: Percentage of Glycine per polypeptide
- Dimension 19: Percentage of Histidine per polypeptide
- Dimension 20: Percentage of Isoleucine per polypeptide
- 10 Dimension 21: Percentage of Asparagine per polypeptide
- Dimension 22: Percentage of Glutamine per polypeptide
- Dimension 23: Percentage of Arginine per polypeptide
- Dimension 24: Percentage of Serine per polypeptide
- Dimension 25: Percentage of Threonine per polypeptide
- 15 Dimension 26: Percentage of non-canonical amino acid (undefined) per polypeptide
(Please not that this dimension does not contain any value other than 0 as input)
- Dimension 27: Percentage of Valine per polypeptide
- Dimension 28: Percentage of Tryptophane per polypeptide
- Dimension 29: Percentage of Tyrosine per polypeptide
- 20 Dimension 30: Cysteine content (zero, even or odd number set to 0.5, 1 or 0, respectively)
- Dimension 31: Percentage of coiled secondary structure per polypeptide
- Dimension 32: Percentage of helical secondary structure per polypeptide
- Dimension 33: Percentage of random secondary structure per polypeptide
- Dimension 34: Score for structure around N-terminal cleavage site
- 25 Dimension 35: Score for structure around C-terminal cleavage site
- Dimension 36: Number of helical blocks per polypeptide
- Dimension 37: Isoelectric point of polypeptide
- Dimension 38: Average molecular weight of polypeptide
- Dimension 39: Sum of Van-der-Waals forces of each amino acid within polypeptide
- 30 Dimension 40: Sum of hydrophobicity values of each amino acid within polypeptide
- Dimension 41 – 48: Mean values calculated based on principle component score vectors of hydrophobic, steric, and electronic properties per polypeptide (Mei et al. 2005)
- Dimension 49: Length of polypeptide
- Wherever applicable, the values for dimension 1-49 were scaled to be in the range between
- 35 0 and 1.

The input vectors for training and prediction contain 49 dimensions, however in the current format only 48 are utilized since dimension 26 is set to zero for all fragments (percentage of

non-canonical amino acids per fragment). This is due to the lack of appropriate training data containing non-canonical amino acids, but can be included in future models.

5 4. **Testing of the models**

The best model is chosen which has the highest performance based on the ranking of an independent test set of bioactive and non bioactive peptides. To test the models, the performance of all generated models was tested and the two best models for short peptides (4-55 amino acids) and longer polypeptides (56-300 amino acids) were chosen, respectively. As a result, an overall prediction accuracy of 90.7 % for short peptides and 94 % for longer peptides was achieved. Using an independent test set the disclosed method correctly identifies around 93 % of bio-active peptides and around 91 % of the non-active peptides.

15 5. **Identification of bioactive peptides**

During the ranking step (Step 6, figure 1), the highest scoring peptides per precursor are chosen that have a length shorter than 46 amino acids. In this ranking process, all fragments that have after SVM classification a distance greater than $|0,65|$ and are localized with the negative training data set (i.e. a score of -0,65 or lower) are readily discarded even if they represent the highest scoring peptides per protein precursor.

25 6. **Antimicrobial assays to test the bioactivity of the peptides identified through the method of the present invention**

6.1. **Assay technology**

30 The micro dilution test represents a homogenous method for determining the number of viable bacterial or yeast cells in culture. It relies on the fact that living bacteria or yeast are turbid in culture. Turbidity can be measured as light absorbance with a photometer and is correlated with the number of cells in the sample.

35 6.2. **Materials and methods**

Bacterial and yeast strains

The strains used in the course of the experiments are *Escherichia coli* (*E. coli* ATCC 25922), *Staphylococcus aureus* (*S. aureus* ATCC 29213) and *Candida albicans* (*C. albicans* FH 2173).

Pre-cultivation of all test strains

The cultivation of the strain starts with building up a cryostock that can be used for multiple inoculations of pre-cultures.

1. Streak the bacteria onto the surface of a Mueller Hilton (MH)-agar plate by using an inoculation loop, and incubate the agar plate for 3 days at 37 °C. For yeast use the same procedure but with Sabouraud dextrose (SD)-agar.
2. Inoculate a 100 ml shaking flask containing 30 ml MH broth with one loop of bacteria and incubate the flask for 1 day at 37 °C and 180 rpm. For yeast apply the same conditions in SD broth.
3. Remove the hypertonic cryo-preservative solution from the Cryobank (CRYO/G) plastic vials, each containing 25 green glass beads, by using a sterile pipette.
4. Fill each vial with 2 ml of the bacterial/yeast suspension, close the vial, and mix carefully.
5. Remove as much of the bacterial/yeast culture supernatant from the vial as possible. The surface of the beads is now covered with bacteria/yeast. The amount of liquid remaining in the vial should be as low as possible to prevent clumping of the beads. One bead is used for the inoculation of one pre-culture (30 ml MH/SD broth in a 100 ml shaking flask).
6. Store the Cryobank (CRYO/G) vials at –80 °C.
7. Quality/sterility check: Remove a Cryobank (CRYO/G) vial from the freezer and place it into a Cryoblock (CRYO/Z). Open the vial, remove one bead and immediately streak the bead over the surface of a MH/SBD agar plate. Incubate the plate for 3 days at 37 °C. Verify that only the test strain has grown by examining the colony morphology.

Preparation of test culture using MH broth

The test strain vial is removed from the Cryobank. One bead is removed with a sterile pipette and inoculated in a 100 ml Erlenmeyer with 30 ml MH and SD broth for bacteria and yeast, respectively. Grow the culture for 18 h at 37°C and 180 rpm. The optical density is adjusted with MH broth to a cell density corresponding to 10^8 cells/ml for all test strains. The standard inoculum culture for the assay is diluted 1:100 to the final concentration of 10^6 CFU/ml (colony forming units/ml).

Peptide Dilutions

The compounds are diluted serially (10 dilution steps) from the standard initial concentration of 125 μ M to a final concentration of 0,24 μ M. The initial DMSO concentration is 1,4 % in all samples and controls.

Standard Antibiotic Dilutions for Dose Response Curves

For dose response experiments dilute the compounds serially (16 dilution steps) with MH broth. Final compound concentrations range between 64 μ g/ml and 0.002 μ g/ml. The initial DMSO concentration is 1,4 % in all samples and controls.

	Supplier	Cat No	Function
Mueller Hinton (MH) broth	Becton Dickinson	275730	Culture medium
Sabouraud dextrose (SD) broth	Becton Dickinson	238230	Culture medium
DMSO	Merck	102 931	Solvent
Nystatin	Calbiochem	475914	Antibiotics
Cyprobay 100	Bayer		
Greiner, 384	Greiner	781182	Assay Plates
SPECTRAFluor Plus	Tecan	-	Reader Absorbance

Assay Protocol

- * Pre-culture the bacteria in 30 ml MH broth at 37°C for 18 h (100 ml Erlenmeyer flask)
- * Pre-culture the yeast in 30 ml SD broth at 37°C for 18 h (100 ml Erlenmeyer flask)
- * Adjust the cell suspension with MH broth to 10^6 CFU/ml (test culture)

Assay :

- * Add 10 μ l compound in DMSO and 30 μ l MH broth to the first vial

- * Transfer 20 µl from the first vial in the second that contains 20 µl MH broth
 - * The last step is repeated 8 times (peptides, 10 dilution steps) or 14 times (antibiotics, 16 dilution steps)
 - * Add 10 µl test culture suspension to each vial (10 vials for the peptides and 16 vials for the antibiotics)
- 5 ⇒ start cell inoculum : 5 x 10⁵ CFU
- ⇒ start DMSO concentration : 12,5 %
- ⇒ start/final compound concentration : 125 µM - 0,24 µM
- ⇒ start/final antibiotic concentration : 64 µg/ml - 0,002 µg/ml
- 10 * Incubate at 37°C for 18 h by 5% relative humidity and 5% CO₂
- * Read absorbance at 590 nm with 5 flashes

Controls :

- * High controls : MH broth with bacteria (growth control, high signal)
- 15 * Low controls : MH broth without bacteria (sterile control, low signal)

6.3. Sensitivity testing with antibiotics

20 In order to evaluate the suitability of the assay for the identification of potential drugs, the dose dependent effects of a number of antibiotics were tested using the conditions described under 'Materials and Methods'. Cyprofloxacin expected to be active against *E. coli* and *S. aureus* and Nystatin against *C. albicans*. The calculated IC₅₀ values for these antibiotics are given in figure 4 in µg/ml.

25 6.4. Assay results

The peptides were tested against the test strains *E. coli* (ATCC 25922), *S. aureus* (ATCC 29213) and *C. albicans* (FH 2173). The peptides A003500589 and A003500548 showed IC₅₀ values of 7,25 µg/ml and 6,79 µg/ml, respectively, against *E. coli*. No activities were found against *S. aureus* and *C. albicans*.

30

REFERENCES

- Chih-Chung Chang and Chih-Jen Lin; "LIBSVM: a library for support vector machines"; 2001
- 5 Peter Duckert, Søren Brunak and Nikolaj Blom; "Prediction of proprotein convertase cleavage sites"; *Protein Engineering, Design and Selection*, 17:107-112, 2004
- Durbin R, Eddy S, Krogh A and Mitchison G; "The theory behind profile HMMs: Biological sequence analysis: probabilistic models of proteins and nucleic acids"; Cambridge University
 10 Press, 1998.
- C. Falciani, L. Lozzi, A. Pini, L. Bracci; "Bioactive Peptides from Libraries"; *Chemistry & Biology*, Volume 12, Issue 4, Pages 417-426, 2005
- 15 Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A.; "Protein Identification and Analysis Tools on the ExPASy Server"; (In) John M. Walker (ed): *The Proteomics Protocols Handbook*, Humana Press, 2005
- Jones, D.T.; "Protein secondary structure prediction based on position-specific scoring
 20 matrices"; *J. Mol. Biol.* 292:195-202, 1999
- H. Kim and H. Park; "Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3d local descriptor"; *Proteins*, 54(3): 557-62, 2004
- 25 Mei, H., Liao, T.H., Zhou, Y., and Li, S.Z.; "A new set of amino acid descriptors and its application in peptide QSARs"; *Biopolymers Vol. 80*, 775-786, 2005
- Henrik Nielsen, Jacob Engelbrecht, Søren Brunak and Gunnar von Heijne; "Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites"; *Protein
 30 Engineering*, 10:1-6, 1997
- Noble W.S.; "What is a support vector machine?"; *Nat. Biotechnol.* 24(12):1565-7, 2006
- Rohrer, S.; "Prediction of post-translational processing sites in peptide hormone precursors";
 35 Diplomarbeit, Universität Würzburg, 2004

John Shawe Taylor & Nello Cristianini; "Support Vector Machines and other kernel-based learning methods"; Cambridge University Press, 2000

5

DESCRIPTION OF THE FIGURES

Figure 1:

10 A schematic overview of the method disclosed in the invention is given in Fig 1 to explain the steps involved in peptide library generation.

Figure 2:

15 Figure 2 shows the amino acid sequences of the 185 bioactive peptides selected based on shared physico-chemical properties.

Figure 3:

Figure 3 shows the input vectors of the 185 peptides identified as bioactive by the trained SVM algorithm.

20 Figure 4:

Figure 4 shows the calculated IC₅₀ values for antibiotics in µg/ml.

CLAIMS

1. In a computer-based system, a method for identifying bioactive peptides derived from protein precursor sequences using a binary support vector machine (SVM) based algorithm, said method comprising the steps of:

a) training a SVM algorithm to learn to distinguish between bioactive and non bioactive peptides, said training comprising the steps of:

a₁) generating vectors with 49 dimensions, each dimension resulting from the calculation of a molecular descriptor value, for a set of labelled known bioactive peptides derived from precursor sequences and labelled known non bioactive peptides derived from precursor sequences, where the labels indicate whether the peptide is, respectively, bioactive or non bioactive;

a₂) transferring the vector data generated in step a₁) to the SVM based algorithm, said algorithm calculating the optimal hyperplane that separates the vectors corresponding to the bioactive peptides and the non bioactive peptides, respectively;

b) providing protein precursor sequences from a publicly available human protein database;

c) predicting secondary structure and cleavage sites within a protein precursor sequence provided in step b) using computational methods; a set of 7 molecular descriptors is calculated based on said prediction step resulting in the generation of peptide fragments;

d) calculating a set of 42 molecular descriptors corresponding to the physico-chemical properties of the peptide fragments generated in step c);

e) transforming the calculated values from step c) into scaled values between 0 and 1 to generate dimensions 1 to 7 of a 49-dimension-vector for each peptide fragment and transforming the calculated values from step d) into scaled values between 0 and 1 to generate dimensions 8 to 49 of said vector for each peptide fragment;

f) presenting the vectors generated in the step e) to the trained SVM algorithm from step a) to measure the distance of each vector to the hyperplane calculated in step a₂); and

g) classifying each peptide fragment as bioactive peptide or non bioactive peptide, according to the distance measured in step f), wherein dimensions 1 to 7 generated in step e) are the following: Dimension 1: N-terminal ProP score; Dimension 2: N-terminal Hmcut score; Dimension 3: N-terminal fragment; Dimension 4: C-terminal ProP score; Dimension 5: C-terminal Hmcut score; Dimension 6: C-terminal Hamid score; Dimension 7: C-terminal fragment.

2. A method according to claim 1, wherein dimensions 8 to 49 generated in step e) are the following: Dimension 8: Percentage of acidic amino acids (E, N, Q) per polypeptide; Dimension 9: Percentage of positively charged amino acids (R, H) per polypeptide; Dimension 10: Percentage of aromatic amino acids (F, Y, W) per polypeptide; Dimension 11: Percentage of aliphatic amino acids (G, V, A, I) per polypeptide; Dimension 12: Percentage of Proline per polypeptide; Dimension 13: Percentage of reactive amino acids (S, T) per polypeptide; Dimension 14: Percentage of Alanine per polypeptide; Dimension 15: Percentage of Cysteine per polypeptide; Dimension 16: Percentage of Glutamic acid per polypeptide; Dimension 17: Percentage of Phenylalanine per polypeptide; Dimension 18: Percentage of Glycine per polypeptide; Dimension 19: Percentage of Histidine per polypeptide; Dimension 20: Percentage of Isoleucine per polypeptide; Dimension 21: Percentage of Asparagine per polypeptide; Dimension 22: Percentage of Glutamine per polypeptide; Dimension 23: Percentage of Arginine per polypeptide; Dimension 24: Percentage of Serine per polypeptide; Dimension 25: Percentage of Threonine per polypeptide; Dimension 26: Percentage of non-canonical amino acid per polypeptide; Dimension 27: Percentage of Valine per polypeptide; Dimension 28: Percentage of Tryptophane per polypeptide; Dimension 29: Percentage of Tyrosine per polypeptide; Dimension 30: Cysteine content; Dimension 31: Percentage of coiled secondary structure per polypeptide; Dimension 32: Percentage of helical secondary structure per polypeptide; Dimension 33: Percentage of random secondary structure per polypeptide; Dimension 34: Score for structure around N-terminal cleavage site; Dimension 35: Score for structure around C-terminal cleavage site; Dimension 36: Number of helical blocks per polypeptide; Dimension 37: Isoelectric point of polypeptide; Dimension 38: Average molecular weight of polypeptide; Dimension 39: Sum of Van-der-Waals forces of each amino acid within polypeptide; Dimension 40: Sum of hydrophobicity values of each amino acid within polypeptide; Dimension 41 – 48: Mean

values calculated based on principle component score vectors of hydrophobic, steric, and electronic properties per polypeptide; Dimension 49: Length of polypeptide.

3. A method according to claim 1 or 2, wherein protein sequences from step b) are only naturally occurring protein sequences found in the human secretome.

4. A method according to any of claims 1 to 3, wherein said bioactive peptides, classified in step g), are bioactive peptide hormones derived from precursor hormones.

5. A method according to any of claims 1 to 4, wherein the bioactive peptide, classified in step g), has a sequence selected from the group consisting of the amino acid sequences of SEQ ID NOS: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185.

6. A method according to claim 5, wherein the bioactive peptide is present in a peptide library.

7. A computational device configured to identify bioactive peptides derived from protein precursor sequences by using a binary support vector machine (SVM) based method, said method comprising the steps of:

5 a) training a SVM algorithm to learn to distinguish between bioactive and non bioactive peptides, said training comprising the steps of:

10 a₁) generating vectors with 49 dimensions, each dimension resulting from the calculation of a molecular descriptor value, for a set of labelled known bioactive peptides derived from precursor sequences and labelled known non bioactive peptides derived from precursor sequences, where the labels indicate whether the peptide is, respectively, bioactive or non bioactive;

15 a₂) transferring the vector data generated in step a₁) to the SVM based algorithm, said algorithm calculating the optimal hyperplane that separates the vectors corresponding to the bioactive peptides and the non bioactive peptides, respectively;

20 b) providing protein precursor sequences from a publicly available human protein database;

25 c) predicting secondary structure and cleavage sites within a protein precursor sequence provided in step b) using computational methods; a set of 7 molecular descriptors is calculated based on said prediction step resulting in the generation of peptide fragments;

d) calculating a set of 42 molecular descriptors corresponding to the physico-chemical properties of the peptide fragments generated in step c);

30 e) transforming the calculated values from step c) into scaled values between 0 and 1 to generate dimensions 1 to 7 of a 49-dimension-vector for each peptide fragment and transforming the calculated values from step d) into scaled values between 0 and 1 to generate dimensions 8 to 49 of said vector for each peptide fragment;

35

- f) presenting the vectors generated in the step e) to the trained SVM algorithm from step a) to measure the distance of each vector to the hyperplane calculated in step a₂); and
- 5 g) classifying each peptide fragment as bioactive peptide or non bioactive peptide, according to the distance measured in step f), wherein dimensions 1 to 7 generated in step e) are the following: Dimension 1: N-terminal ProP score; Dimension 2: N-terminal Hmcut score; Dimension 3: N-terminal fragment; Dimension 4: C-terminal ProP score; Dimension 5: C-terminal Hmcut score; Dimension 6: C-terminal Hamid score; Dimension 7: C-terminal fragment.
- 10
8. Use of the method according to any of claims 1 to 6 for the identification of therapeutic polypeptides, targets for drug intervention, ligands to discover relevant targets or biomarkers to monitor diseases.
- 15