

Review

Annotation of Peptide Structures Using SMILES and Other Chemical Codes—Practical Solutions

Piotr Minkiewicz *, Anna Iwaniak and Małgorzata Darewicz

Chair of Food Biochemistry, Faculty of Food Science, University of Warmia and Mazury in Olsztyn, Plac Cieszyński 1, 10-726 Olsztyn-Kortowo, Poland; ami@uwm.edu.pl (A.I.); darewicz@uwm.edu.pl (M.D.)

* Correspondence: minkiew@uwm.edu.pl; Tel.: +48-89-523-3715

Received: 16 October 2017; Accepted: 25 November 2017; Published: 27 November 2017

Abstract: Contemporary peptide science exploits methods and tools of bioinformatics, and cheminformatics. These approaches use different languages to describe peptide structures—amino acid sequences and chemical codes (especially SMILES), respectively. The latter may be applied, e.g., in comparative studies involving structures and properties of peptides and peptidomimetics. Progress in peptide science “in silico” may be achieved via better communication between biologists and chemists, involving the translation of peptide representation from amino acid sequence into SMILES code. Recent recommendations concerning good practice in chemical information include careful verification of data and their annotation. This publication discusses the generation of SMILES representations of peptides using existing software. Construction of peptide structures containing unnatural and modified amino acids (with special attention paid on glycosylated peptides) is also included. Special attention is paid to the detection and correction of typical errors occurring in SMILES representations of peptides and their correction using molecular editors. Brief recommendations for training of staff working on peptide annotations, are discussed as well.

Keywords: peptides; glycopeptides; chemical modifications; bioinformatics; cheminformatics; chemical information; good practice; molecular editors; SMILES code

1. Introduction

Contemporary peptide science encompasses the biological and chemical approach. Medical sciences, pharmacology, biotechnology, and last, but not least, food and nutrition sciences need both biology and chemistry. Peptides are in the focus of interest of all the above-mentioned areas.

Tools for in silico peptide research, such as databases and programs, utilize both of the approaches that are classified as bioinformatics and cheminformatics, respectively, although most of the specialized databases and programs that are dedicated for peptides may be classified as bioinformatic tools [1]. Both of the approaches use different languages to describe the structures of biomolecules (e.g., peptides) [2,3]; bioinformatics operates based on amino acid sequences, whereas cheminformatics, on universal chemical codes. Communication between these two areas requires translation from the biological into the chemical language [2,3]. Dataset concerning biological activities of taste-affecting peptides, published in our review [4], may serve as an example of benefits from merging the biological and chemical approaches. It could not be completed without screening databases using peptide structures, annotated in the SMILES code [5,6], as a query. Another example of in silico research, utilizing chemical approach, has been recently published by Ortiz-Martinez and others [7]. They used the SwissTargetPrediction program [8,9], as provided by the Swiss Institute of Bioinformatics, Lausanne, Switzerland, to predict interactions between small peptides from maize and proteins of human organism. Chemical modifications of peptides, aimed to alter their biological activity, may recently be considered as a “hot topic” [10–14]. Processing of the peptide sequences, including non-protein or modified amino acids, is possible, e.g., using PepstrMod program [15,16] (provider Institute of

Microbial Technology, Chandigarh, India). The above program utilizes hundreds of non-protein or modified amino acid residues. Possible space of non-natural or modified amino acids and other possible constituents of peptides, contains however even billions of possible molecules or molecule fragments [17]. SMILES and other chemical codes and formats enable the description of any artificially inserted substituents for in silico study of properties of modified peptides. Another approach is the search for peptidomimetics, which are potentially useful as drugs [18], on the basis of known peptide structures.

Many programs, utilizing the SMILES code, are recently available, including, e.g., BioTriangle program provided by the Central South University, Changsha, China [19,20], which serves to calculate, e.g., physicochemical and topological parameters of small molecules. Some programs that are utilizing SMILES are available via the website of the Swiss Institute of Bioinformatics, Lausanne, Switzerland [21]. This website offers access to, e.g., to SwissADME program [22,23] which allows predicting properties that affect substance applicability as a drug. Another example of program utilizing the chemical code is WebMolCS [24,25] and other programs developed at the University of Bern, Bern, Switzerland [26] within the Chemical Space Project [17]. Apart from the above, free accessible programs, there are also commercial tools, such as JChem or MadFast [27], both are provided by ChemAxon, Budapest, Hungary, utilizing chemical codes for database screening or calculations. There are also specialized peptide databases using the SMILES code, such as Brainpeps [28,29] or Quorumpeps [30,31], provided by the University of Ghent, Belgium; AHTPDB [32,33], CancerPPD [34,35], Hemolytik [36,37], ParaPep [38,39] or PepLife [40,41], provided by the Institute of Microbial Technology, Chandigarh, India. The above resources are integrated via the SATPdbmetabase [42,43]. Another example is BIOPEP database of sensory peptides and amino acids [44,45], provided by the University of Warmia and Mazury in Olsztyn, Poland. The number of programs and databases utilizing chemical codes successively increases. More links to such tools are available via metabases and metaservers [46–50].

The bioinformatics approach concerning peptides involves, e.g., modeling structures and predicting interactions with biomacromolecules on the basis of amino acid sequences [1,51]. Structure modeling, involving amino acid sequences, may be performed using programs such as PepstrMod [15,16], Pep-Fold [52,53] (provider: University of Paris, Diderot, Paris, France), or (PS)².v3 (provider: National Chiao Tung University, Hsinchu, Taiwan) [54,55]. For instance, the Quantitative Structure-Activity Relationship (QSAR) approach involves a set of parameters that are describing the structure and physicochemical properties of particular amino acid residues [56,57]. The sequence-based approach is expanded using pseudo-amino acid composition [58–60]. The application of chemical information for annotation of peptides and for processing their structures may enlarge the array of tools available for peptide research in silico.

Cheminformatics tools cannot, however, be considered and used uncritically as “black boxes”. Many published datasets contain errors. Users or curators of databases and programs using chemical information should be prepared to recognize and correct possible errors [61–64]. Validation of representations, identification, and correction of mislabeled compounds is recommended as one of the crucial steps of compound dataset preparation and curation [2,63]. The preparation of peptide datasets, involving translation from amino acid sequences into chemical codes, is not an exception. Peptide data, annotated using chemical information codes, requires careful inspection before use.

The aim of this review article is to present practical solutions concerning translation of peptide annotation from biological into chemical language and correction of possible errors using contemporary software (with special attention of non-commercial programs) without extensive historical background. Proposed recommendations are based on our experience with completion and curation of the BIOPEP database of sensory peptides and amino acids [44,45] and MetaComBio website (University of Warmia and Mazury in Olsztyn, Poland) [48,49].

2. Codes for Annotation of Peptide Sequences and Structures

2.1. Annotation of Peptides Using Biological Codes

The most common biological codes for the annotation of peptide sequences are: single letter code and multi-letter code. The first is applied for annotation of protein sequences and peptides consisting of amino acids present in proteins. Annotation of the peptides containing non-protein or non-natural amino acid residues requires a multi-letter code (usually three or four characters). The most comprehensive list of abbreviations of amino acid names may be found in the SwissSidechain (Swiss Institute of Bioinformatics, Lausanne, Switzerland) [65,66] and Norine (University of Sciences and Technologies of Lille, Villeneuve d'Ascq, France) [67,68] databases. L-amino acids are annotated using capital letters, whereas D-amino acids—using small letters. Such a layout is used in, e.g., CycloPS program (University College Dublin, Ireland) [69,70], and the SATPdb database (Institute of Microbial Technology, Chandigarh, India) [42,43]. The last database annotates protein amino acids and their D-enantiomers using a single letter code, whereas non-protein amino acids—using a multi-letter code. Peptides containing amino acids from both groups are annotated using a “mixed code”, utilizing both single- and multi-letter amino acid abbreviations. The multi-letter symbols are divided by dashes. Amino acid symbols that are used in the SwissSidechain database do not include dashes within amino acid abbreviations. Such abbreviations may be sufficient to construct machine-readable sequences with the help of multi-letter and mixed codes. The Norine database utilizes symbols, including dashes. Amino acid sequences utilizing symbols containing dashes may be annotated using the LINUCS code, designed originally for the description of oligosaccharides [71]. LINUCS is also used to annotate peptide sequences in the PubChem database (National Center for Biotechnology Information, Bethesda, MD, USA) [72,73]. LINUCS code may offer machine-readable representation of glycopeptides. Another code applicable for this purpose is HELM [74,75], utilizing single letter symbols of amino acid residues and describing modifications at atomic level. HELM representations of peptides may be found e.g., in PubChem and ChEMBL (European Bioinformatics Institute, Hinxton, UK) [76,77] databases. Macrocyclic peptides may be annotated using amino acid sequences with SMILES rules for description of multiple rings [78].

2.2. Representation of Peptides Using SMILES Code

The most known chemical codes, which are used to annotate compound structures, are: SMILES [5,6], InChI [79], and InChIKey [79]. SMILES is the most popular chemical code used in databases of low molecular-weight compounds. It is applied in such databases as PubChem [72,73], ChemSpider (Royal Society of Chemistry, London, UK) [80,81], ChEMBL [76,77], or ZINC 15 (University of California San Francisco, CA, USA) [82,83]. The so-called isomeric SMILES, taking into account stereoisomers (e.g., configuration around asymmetric carbon atoms), is usually applied as a representation of chiral compounds, including amino acids and peptides. Databases and programs supporting mass spectrometric analysis may use the canonical SMILES (without the discrimination of configurations around asymmetric carbon atoms) due to the fact that mass spectrometry is unable to discriminate between stereoisomers. The canonical SMILES is applied to annotate peptides in, e.g., HMDB database (University of Alberta, Edmonton, AB, Canada) [84,85].

Multiple SMILES representations for the same compound are also possible. They may differ by, e.g., order of symbols indicating particular atoms [3,6]. The order of symbols, as proposed by Siani et al. [2], and presented in Figure 1, i.e., α -amine group; asymmetric carbon atom C2; side chain, and carboxyl group, is sufficient for the construction of SMILES representations of peptides. Such SMILES codes are presented in, e.g., source codes to the Cyclops program [69,70] and in SwissSidechain database [65,66]. SMILES strings presented in the PubChem and ChemSpider databases do not maintain the above order.

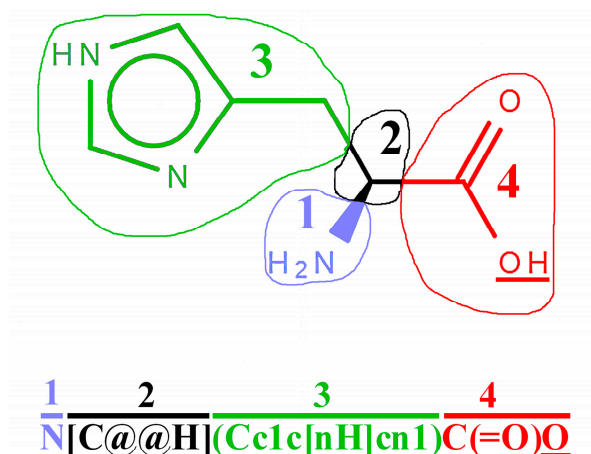


Figure 1. Construction of SMILES string of exemplary amino acid (L-histidine), sufficient for peptide annotation. (1) α -amine group; (2) carbon atom C2; (3) side chain (in this case methylene group and imidazole ring); and, (4) α -carboxyl group. Underlined hydroxyl group in the structure or oxygen atom in SMILES string is replaced by the next amino acid residue during construction of peptide representation. Figure prepared with the help of Ketcher 2.0 molecular editor (EPAM Systems, St Petersburg, Russia) (Demo version) [86,87].

Manual construction of SMILES representation of the exemplary tripeptide, based on recommendation of Siani and co-workers [2] is illustrated in Figure 2. The SMILES strings of an individual amino acids follow the order presented in Figure 1. The last oxygen atom in the carboxyl group (underlined in Figures 1 and 2) may be replaced by the fragment that is corresponding to the next residue.

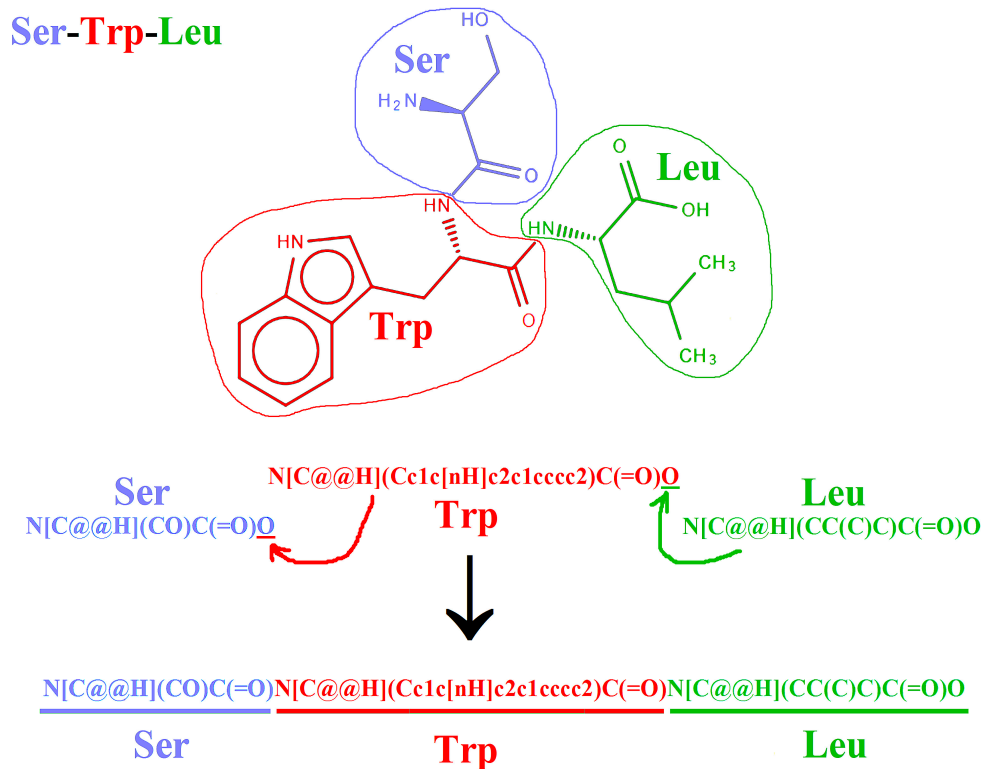


Figure 2. Scheme illustrating manual construction of SMILES representation of peptide SWL (Ser-Trp-Leu): Underlined oxygen atom in SMILES strings of individual amino acid molecules is replaced by the next amino acid residues during construction of peptide representation. Picture of peptide structure prepared with the help of Ketcher 2.0 molecular editor (Demo version).

There are two freely accessible programs that are able to convert amino acid sequences of peptides into SMILES representations—downloadable program Open Babel [88,89] and CycloPs [69,70], available at its own server.

The Open Babel program (current version: 2.4.1) is able to translate peptide sequences annotated in a single-letter code into SMILES or other chemical codes and formats (recently 110 formats). The program utilizes FASTA format [90,91] as an input. The FASTA format is widely used for protein sequence annotation in, e.g., the UniProt database (European Bioinformatics Institute, Hinxton, UK) [92,93]. Conversion of peptide sequences from the FASTA format into SMILES is not described in the program manual, and requires a special correction procedure, as described in the supplement to our previous article [44]. Open Babel program utilizes sequences consisting of 20 common protein amino acids. The program is able also to convert peptide structures, annotated as pdb files [94] e.g., created by (PS)². v3 or PepstrMod program. The second opportunity allows taking into account amino acid modifications, as accepted by the program. The order of symbols in the SMILES strings created by Open Babel differs from that created manually, as presented in Figure 2.

The CycloPs program utilizes common protein amino acids, as well as their D-enantiomers. The program applies an algorithm, as presented in Figure 2, to create SMILES representations of linear peptides. The creation of SMILES representations of cyclic peptides using CycloPs is also feasible. Another advantage of CycloPs as compared with Open Babel is its speed, especially in the case of processing sequences with length exceeding 10 amino acid residues.

Biomolecule toolkit (Provider: ChemAxon, Budapest, Hungary) is a commercially available resource for creation of biomolecule (e.g., peptide) annotations. It utilizes both peptide sequences (including non-natural and modified amino acids) and SMILES. This toolkit accepts peptides annotated using HELM notation [74,75] as an input.

2.3. Construction of SMILES Representations of Peptide Containing Modified Amino Acid Residues Using Glycosylated Amino Acids As an Example

Non-protein, non-natural, or modified amino acid residues may be inserted into SMILES representations of peptides via two ways. The first one includes the manual insertion of amino acid representations, which are taken from such databases as SwissSidechain. SMILES codes of amino acids, taken from PubChem, ChemSpider or ZINC 15 may need rearrangement to obtain the order of atom symbols presented in Figure 1. The second option is to apply molecular editors. They serve to display and modify structures of chemical compounds [95]. They may serve for implementation of simplified version of “forward translation flowchart” proposed by Siani and co-workers [2]. The editors sufficient to modify SMILES representations of peptides should provide the following opportunities: input and output of the molecular structure as a SMILES string, as well as display of the absolute configuration of substituents around asymmetric carbon atoms (Rectus or Sinister). Ketcher version 2.0 [86,87] and Marvin editor versions (Marvin Sketch, Marvin JS) (ChemAxon, Budapest, Hungary) [27] are examples of molecular editors that are fulfilling these requirements. The Marvin JS is widely used in chemical databases, e.g., ChEMBL [76,77] or HMDB [84,85].

The scheme of inserting unnatural or modified amino acids into a peptide structure, annotated via the SMILES code with the help of molecular editor, may include the following steps: finding a protein amino acid or its enantiomer most similar to the desired one; construction of peptide SMILES representation using a selected amino acid residue; display and modification of a peptide structure using molecular editor; and, the conversion of the resulting, modified structure into SMILES. The above scheme is presented in Figure 3, with peptide containing a glycosylated threonine residue as an example. We use sugar moiety as an example of peptide modification for two reasons. The first one is fact that glycan residues are commonly present in peptides and proteins, and play very significant role in their biological activity [96–98]. The second reason is that sugar moieties are relatively complex as compared with other non-amino acid residues that are present in peptides. Someone who can correctly introduce sugar moiety may do that with other, less complex residue. The initial sequence of peptide includes an unmodified threonine residue as the most similar to the glycosylated threonine (Figure 3a). The second step (Figure 3b) involves the construction of SMILES

representation of peptide (built manually or using appropriate software). The third step includes the import and display of the peptide structure in the molecular editor (Figure 3c). This step enables the verification of structure and the correction of errors in the peptide backbone structure (For details see Section 3.). The fourth step (Figure 3d) includes building additional groups from the so-called “basic primitives” [99]—the simplest fragments of molecule structure that is used by molecular editors. If the additional group contains chirality centers, absolute configuration around asymmetric carbon atoms should be checked. The additional group (e.g., glycosidic, as shown in Figure 3d) usually mimics individual compound, which can be used as a reference molecule to check details of structure and chirality. In the example discussed herein, it is *N*-Acetyl- α -D-galactosamine (IUPAC name: *N*-[(2*S*,3*R*,4*R*,5*R*,6*R*)-2,4,5-trihydroxy-6-(hydroxymethyl)oxan-3-yl]acetamide; PubChem CID 84265).

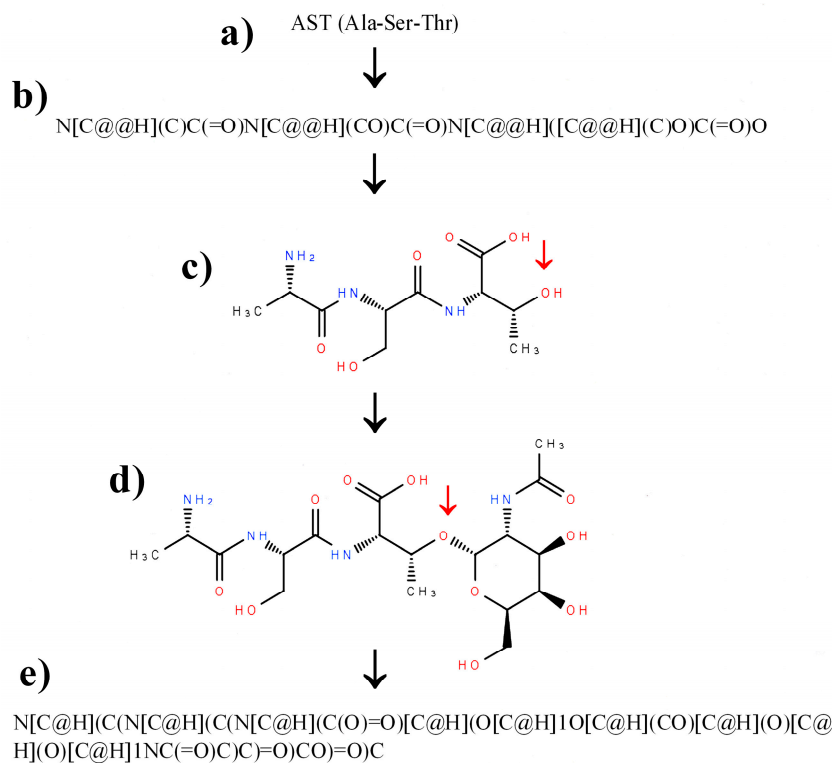


Figure 3. Scheme illustrating construction of SMILES representation of a modified peptide AST (Ala-Ser-Thr): (a) peptide sequence; (b) SMILES representation of unmodified peptide, constructed in agreement with scheme presented in Figure 2; (c) structure of unmodified peptide, displayed using program Ketcher (v. 2.0); and, (d) structure of peptide with a glycosylated threonine residue in position 3 (added residue of *N*-Acetyl- α -D-galactosamine—PubChem CID 84265); (e) SMILES representation of glycosylated peptide. Red arrows indicate modification site in structures of peptide and glycopeptide.

A comparison of absolute configurations around asymmetric carbon atoms in *N*-acetylgalactosamine and glycopeptide containing *N*-acetylgalactosamine residue is displayed in Figure 4. In the case of the compound presented above, the addition of the peptide chain does not affect configuration around asymmetric carbon atom being anomeric in Figure 4a and is involved in glycosidic bond in Figure 4b, in the sugar residue, as judged using Cahn-Ingold-Prelog priority rules [100,101]. The addition of the peptide chain does not change the priority of substituents around carbon atom from *N*-acetylgalactosamine residue, involved in the formation of *O*-glycosidic bond. Configuration “Sinister” is thus retained for this atom. Configuration (Rectus) remains the same also for carbon atom C3 in a threonine residue (See Figures 3c,d and 4b). Change of the absolute configuration of asymmetric carbon atoms is, however, possible for other groups that are added to the peptide chain.

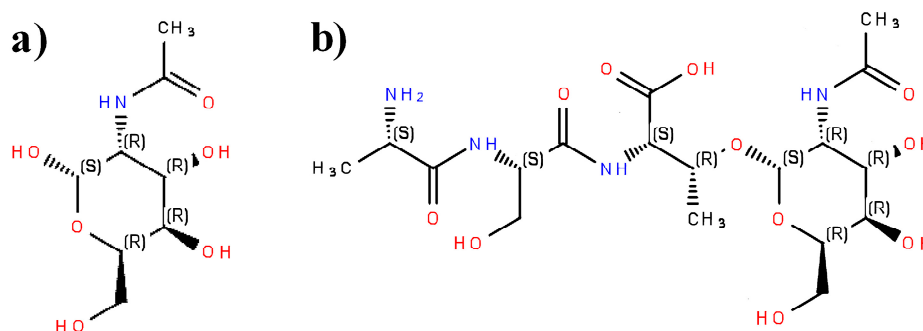


Figure 4. Structures of (a) *N*-Acetyl- α -D-galactosamine; (b) exemplary glycopeptide containing *N*-Acetyl- α -D-galactosamine residue with displayed absolute configuration of substituents around asymmetric carbon atoms. R—configuration Rectus; S—configuration Sinister. Structures displayed using Ketcher 2.0 program.

We have chosen the addition of a sugar residue as an example of peptide modification due to difficulties that are posed by the presence of several asymmetric carbon atoms in a single residue or to a discrepancy in numbering the carbon atoms. According to the rules that are accepted in sugar chemistry and glycobiology, the anomeric carbon atom in aldoses and their derivatives possesses number 1. Computer programs such Chemical Identifier Resolver (National Institutes of Health, Bethesda, MD, USA) [102,103] or Chemical Translation Service (University of California Davis, Davis, CA, USA) [104,105] use IUPAC rules for heterocyclic compounds for numbering carbon atoms in sugar rings. According to these rules, the anomeric carbon atom possesses number 2. IUPAC name of *N*-Acetyl- α -D-galactosamine, present in PubChem database (CID 84265), follows rules that are proposed for heterocyclic compounds. The introduction of many other modifications, such as C-terminal amidation, esterification, or charge of ionizable groups, is much easier.

Translation from the SMILES code into amino acid sequences of peptides is possible using the Smiles2Monomers program (University of Sciences and Technologies of Lille, Villeneuve d'Ascq, France) [106,107], which is associated with the Norine database [67,68]. This program utilizes SMILES strings of amino acid residues that are present in peptides annotated in the Norine database. Another, commercially available option for this purpose is Biomolecule Toolkit, provided by ChemAxon. This program utilizes chemical codes as an input via Marvin editor [27].

2.4. Other Chemical Codes

InChI and InChIKey [79] are another common formats applied to describe chemical compounds including peptides. In contrast to SMILES, inChI and InChIKey provide unique representation of a molecule structure. InChI code describes the structure of a molecule, whereas InChIKey, always containing 27 characters, does not reflect the molecule structure, but may serve as a query for search using popular search engines, such as Google™ [108,109]. In addition, InChIKeys are utilized by specialized programs for finding compounds in multiple databases, such as Chemical Translation Service.

SMILES or InChI may be converted into many other formats using, e.g., OpenBabel program. Chemical formats are utilized, for instance, by programs designed to model interactions between proteins and small molecules. SwissDock (Swiss Institute of Bioinformatics, Lausanne, Switzerland) [110,111], AMMOS2 (Université Paris Diderot, Paris, France) [112,113], and ProteinsPlus (University of Hamburg, Germany) [114,115] are examples of such programs.

3. Verification and Correction of Errors in Peptide Structures Annotated Using SMILES Code

3.1. Typical Errors in Peptide Representations and Their Correction

Characteristic errors in structures of peptides, as annotated using the SMILES code, are summarized in Table 1 and Figure 5. Inappropriate structure or configuration of a molecule fragment may lead to false negative results of database searching, or generate errors in prediction of physicochemical properties or interactions with biomacromolecules. Some of the errors may be detected and displayed automatically (Figure 5). The Marvin JS molecular editor is sufficient for this purpose. It displays inappropriate valence of atoms or missed chirality centers.

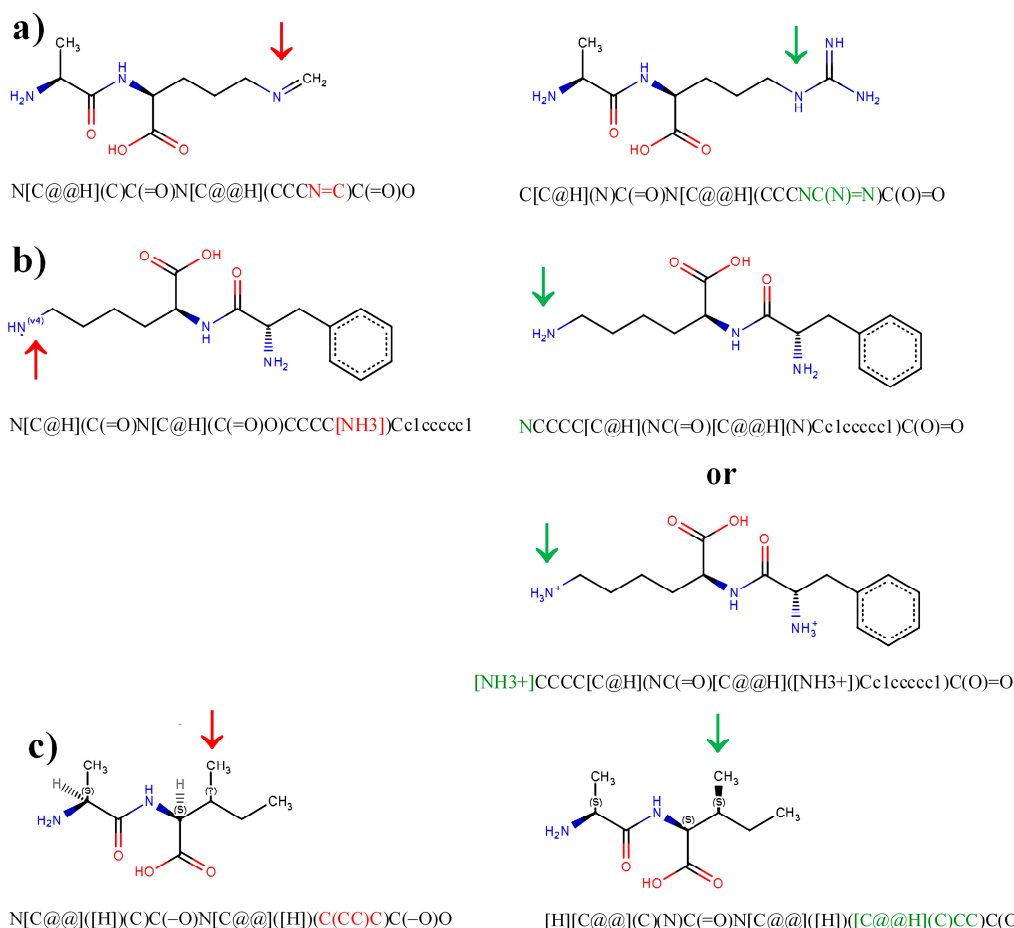


Figure 5. Structures of peptides displayed by molecular editor with errors (left column) and after correction (right column). (a) Peptide AR (Ala-Arg) with missed nitrogen atoms in guanidine group; (b) peptide FK (Phe-Lys) with inappropriate valence of nitrogen atom in side chain of lysine; (c) peptide AI (Ala-Ile) with missed chirality center in Ile residue. Red arrows and red fonts in SMILES representations indicate molecule fragments with errors. Green arrows and green fonts in SMILES representations indicate the same fragments after correction. R—configuration *Rectus*; S—configuration *Sinister*; ?—undefined configuration of substituents around asymmetric carbon atom. Figure prepared with the help of Marvin JS editor [27].

Missed atoms or inappropriate structures of functional groups (Figure 5a) are not detected automatically. In this case, the structures that are displayed reflect existing or at least stable molecules, but differ from the desired ones. The group indicated using red arrow in Figure 5a contains a nitrogen atom possessing valence 3, and a carbon atom with valence 4 (typical valences of both elements), but differs from true guanidine group in arginine residue (indicated using green arrow). Structures not possessing inappropriate valence or undefined chirality centers are accepted by the molecular editor and can be verified and corrected only manually—i.e., by deletion of the

inappropriate fragment of a molecule and construction of an appropriate one from basic primitives. Guanidine group in Figure 5a has been corrected in this way.

Table 1. Possible errors in SMILES strings of peptides and their consequences.

Error	Consequences
Missed nitrogen atoms in guanidine groups	Possible errors in results of modeling interactions with biomacromolecules, inappropriate InChIKey, insufficient to be a query in database searching
Inappropriate valence of nitrogen atoms, e.g., in amine or guanidine groups	Possible errors in results of modeling interactions with biomacromolecules, unavailable for database search engines (e.g., for search engine of ZINC15 database), inappropriate InChIKey, insufficient to be a query in database searching
Undefined or inappropriate configuration of substituents around asymmetric carbon atom, e.g., C3 atom in isoleucine or threonine	Possible errors in results of modeling interactions with biomacromolecules, inappropriate InChIKey, insufficient to be a query in database searching
Spaces in SMILES strings	Disabled processing of SMILES strings by database search engines and other programs

Figure 5b presents an example of a structure containing atom with inappropriate valence. The example presents the ϵ -amine group of a lysine residue, containing nitrogen atom with valence 4 ($-\text{NH}_3$ group without formal charge). This error may be important in the light of recommendations for possible molecular docking studies. Peptide structure input should include hydrogen atoms [88]. The error may be corrected by deletion of the inappropriate $-\text{NH}_3$ group and insertion of an amine group $-\text{NH}_2$, or a protonated amine group $-\text{NH}_3^+$ if necessary. Inappropriate valence is clearly shown in structures displayed using Marvin JS editor (Figure 5b, left column).

Missed chirality center is another kind of common error in compound structures. Configuration around asymmetric carbon atoms in molecules strongly affects interactions with biomacromolecules and hence biological activity. In the case of peptides, configuration around asymmetric carbon atom C3 in isoleucine or threonine residues should be taken into account. Peptide containing the isoleucine residue with missed chirality center at C3 carbon atom is presented in Figure 5c. Missed chirality center is displayed in InChI strings, which is converted from SMILES by the Open Babel or Marvin JS program, using “?” character. The same character is shown in the structure displayed using the second program (Figure 5c, left column). The structure should be corrected by insertion of an appropriate basic primitive [99] to achieve appropriate configuration around the chiral carbon atom. In L-isoleucine, the C3 carbon atom has *Sinister* (S) configuration, whereas in L-threonine—*Rectus* (R) configuration. In D enantiomers of these amino acids, the configuration around carbon atom C3 is reverse to the above. Another opportunity for error detection is Structure Checker application (provider. ChemAxon, Budapest, Hungary) [27].

3.2. Verification and Correction of Representations of Non-Peptidic Moieties with Special Attention on Sugar Residues

The following additional procedure of verification of non-peptidic groups is possible if their structure corresponds to the known compounds that are present in general databases, such as PubChem. The SMILES code may be imported to the molecule editor. A peptide chain may be removed from the displayed structure. For instance, the removal of a peptide chain from the compound presented in Figure 4b provides the compound presented in Figure 4a. The resulting structure of non-peptidic moiety may be converted into SMILES and InChIKey, and may be used as a query in database searching (directly or via Chemical Translation Service). Finding compound in databases implies correctness of its structure. Apart from general databases such as PubChem or ChemSpider, Chemical Translation Service covers also LIPID MAPS® (University of California San Diego, San Diego, CA, USA) [116,117]—a database designed for lipid annotation. It may serve for verification of lipid moieties in lipopeptides.

Verification of structures of sugar residues such as this presented in Figure 4, is possible also via specific tools designed for annotation and processing of carbohydrate structures. Databases of sugars

and tools applied in glycoinformatics were reviewed by Campbell and co-workers [118]. WURCS program (WURCS Working Group, Japan) [119,120] and GlyTouCan database (Soka University, Tokyo, Japan) [121,122] may be helpful in the verification of correctness of simple carbohydrate moieties. The protocol presented below may be used to this end.

Structures of sugar moiety and entire glycopeptide (See Figure 4a,b, respectively) should be converted into MDL molfile format [123]. Molecule editors, such as Ketcher 2.0 or Marvin JS, enable output of compound structure in MDL molfile format. OpenBabel program is also able to translate a molecule structure into this format. The structure converted into the above format may be transferred to the molecular editor at WURCS program website ("Chemical structure to WURCS" tab). This tab serves to translate sugar structure from MDL molfile format into WURCS code. The last one is a machine-readable code designed for description of carbohydrates [124]. The above-mentioned option results in the same WURCS representation of glycopeptide and its sugar moiety. For instance, *N*-acetyl- α -D-galactosamine (Figure 4a) and glycopeptide containing this sugar moiety (Figure 4b) possesses the following WURCS representation: WURCS=2.0/1,1,0/[a2112h-1a_1-5_2*NCC/3=O]/1/. This representation is identical also for *N*-acetyl- α -D-galactosamine structure taken from PubChem (CID 84265) and ChemSpider (ID 76020). The use more databases is recommended to avoid potential errors that are associated with, e.g., missed chirality centers in carbohydrate structures. WURCS representation may also be used as a query in GlyTouCan database screening using text search option. GlyTouCan accession number of *N*-acetyl- α -D-galactosamine is G57321FI. Finding sugar structure in GlyTouCan is a way to prove its correctness. The construction of glycan moiety of glycopeptide may be facilitated by the use of Marvin JS editor that is installed at the website of JCGGDB database [125,126]. The editor installed there offers additional option for construction of glycan moiety from building blocks (e.g., hexose ring, pentose ring etc.). Building blocks do not contain information concerning the configuration around particular asymmetric carbon atoms, which should be added manually.

Although time-consuming, the application of systematic procedures for verification and correction of molecular structures of compounds of interest (e.g., peptides) may allow avoiding errors in research conducted with the help of cheminformatic and bioinformatic methods. Alves and co-workers [127] have also pointed out the significance of transparency of cheminformatic procedures. Details of a procedure for verification and correction of compound structures that are used in a database or a dataset should be included in publications or their supplements. Readers would be thus able to apply and/or improve the workflow. Correction of peptide structures annotated in the BIOPEP database of sensory peptides and amino acids [45] may serve as an example of a procedure fulfilling this recommendation.

4. Brief Recommendations for Training

The importance of training in cheminformatics has recently been emphasized by Tetko and co-workers [128]. They pointed out that cheminformatics needs experience in two areas; chemistry and informatics. In the case of peptide science, the users of existing databases and/or software (such as authors of this article who are not informaticians) need skills in two areas: chemistry and biochemistry with molecular biology. Apart from skills in informatics, development of new software requires sufficient experience in the two above-mentioned areas. Recommendations below concern information from the area of chemistry, which is sufficient in the approach that is mentioned in this publication. Information from this area may be important for biologists using amino acid sequences and abbreviation-based codes in describing other biomolecules (sugars, nucleotides, and acylglycerols).

The checklist of skills includes training in use of a molecular editor, taking into account molecule drawing, input and output options, as well as strong and weak points. The second step involves knowledge of all details of peptide structures, including protein, non-protein and non-natural amino acids as well as modifications (e.g., sugar or lipid moieties). Special attention should be paid to the configuration of substituents around asymmetric carbon atoms. Some amino acids contain more than one chirality center (e.g., C3 atoms in isoleucine and threonine). Additional moieties may also contain asymmetric carbon atoms. Configuration around some asymmetric carbon atoms in biologically-

active molecules, such as acylglycerols or sugars, may be described in different languages. IUPAC recommends the use of absolute configuration (R or S) in systematic names of chiral compounds. Attention should thus be paid to the translation of stereospecific numbering of acylglycerols (*sn*-1 and *sn*-3) or nomenclature of anomeric carbon atoms in sugar residues (α or β) into absolute configuration of asymmetric carbon atoms (atom *sn*-2 in acylglycerols or anomeric atoms in sugars). User of this software should recognize its strong and weak points, as well as the possible errors (e.g., generation of structures containing atoms with inappropriate valence, missed chirality centers or reversed configuration around asymmetric carbon atoms). In the case of using few programs, it is important to check co-operation between them.

The training scheme involves options and protocols that are described in the previous sections. The initial training dataset should contain simple compounds, mainly di- and tripeptides built from typical protein amino acids. They are usually annotated in general, chemical databases, such as PubChem, ChemSpider, and ChEMBL, and specialized peptide databases, such as BIOPEP or AHTPDB. Chemical codes that are generated during training may thus be easily verified. We confirm our previous recommendation [3] to try more databases. Simple di- and tripeptides are usually well described, but representations of more complex moieties (e.g., sugars) may contain errors. They may be recognized and removed via the confrontation of data from few databases. Sometimes, the creation of an appropriate structure representation requires a few attempts. For instance, any asymmetric carbon atom may occur in two possible configurations: R or S. The first attempt to obtain the correct one with the help of the molecular editor is successful with 50% likelihood (for one asymmetric carbon atom). The configuration of asymmetric carbon atoms is usually displayed in PubChem, or in names that are generated by Chemical Identifier Resolver, although for sugar moieties the numbering of carbon atoms, fulfilling IUPAC recommendations designed for heterocyclic compounds, may be confusing.

5. Final Remarks

A language barrier between biology and chemistry, or more precisely, between bioinformatics and cheminformatics, is a fact. Peptide science involves both approaches. Breaking this barrier would enhance progress in all research areas and disciplines interested in the structure and properties of peptides (e.g., medical, pharmaceutical, and food and nutritional sciences). The goal understood as correct and fluent translation from biological (amino acid sequences) into chemical (SMILES, InChI etc.) is not achieved to date. Remarks and recommendations, presented herein, may help the users of databases and software working with sequences and structures of peptides to omit the language barrier. Delivery of new generation of translating software may provide opportunity to break this barrier in the area of peptides. The recommendations presented herein may be applied to develop datasets serving for testing such software. The applicability of some proposals presented here may also be extended to other classes of small biomolecules, such as sugars and lipids.

Acknowledgments: This work was supported from the funds of the University of Warmia and Mazury in Olsztyn (Project No. 17.610.014-300).

Author Contributions: A.I., M.D. and P.M. performed data mining. Options serving for verification and correction of peptide representations in chemical codes were tested by A.I., M.D. and P.M. P.M. prepared the manuscript. A.I., M.D. and P.M. are curators of the BIOPEP database and MetaComBio website.

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

Abbreviations

ADME	Absorption, Distribution, Metabolism and Excretion
AHTPDB	Antihypertensive peptide database
CID	Compound Identifier (in PubChem database)
EMBL	European Molecular Biology Laboratory
α -GalNAc	N-Acetyl- α -D-galactosamine-PubChem CID 84265
HELM	Hierarchical Editing Language for Macromolecules
HMDB	Human Metabolome Database
InChI	International Chemical Identifier

InChIKey	Key of International Chemical Identifier
IUPAC	International Union of Pure and Applied Chemistry
LINUCS	Linear Notation for Unique Description of Carbohydrate Sequences
LIPID MAPS	Lipid Metabolites and Pathways Strategy
MDL	Molecular Design Limited Inc. (company name)
QSAR	Quantitative Structure-Activity Relationship
R	One-letter symbol of amino acid arginine or absolute configuration of substituents around asymmetric carbon atom- <i>Rectus</i> (depending on context)
S	One-letter symbol of amino acid serine or absolute configuration of substituents around asymmetric carbon atom- <i>Sinister</i> (depending on context)
SATPdb	Structurally Annotated Therapeutic Peptide database
SMILES	Simplified Molecular Input Line Entry System or Simplified Molecular Input Line Entry Specification
WURCS	Web3 Unique Representation of Carbohydrate Structures

References

1. Senthilkumar, B.; Rajasekaran, R. Computational resources for designing peptide based drugs preferred in the field of nanomedicine. *J. Bionanosci.* **2016**, *10*, 1–14, doi:10.1166/jbns.2016.1354.
2. Siani, M.A.; Weininger, D.; Blaney, J.M. CHUCKLES: A method for representing and searching peptide and peptoid sequences on both monomer and atomic levels. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 588–593.
3. Minkiewicz, P.; Darewicz, M.; Iwaniak, A.; Bucholska, J.; Starowicz, P.; Czyrko, E. Internet databases of the properties, enzymatic reactions, and metabolism of small molecules-search options and applications in food science. *Int. J. Mol. Sci.* **2016**, *17*, doi:10.3390/ijms17122039.
4. Iwaniak, A.; Minkiewicz, P.; Darewicz, M.; Hryniewicz, M. Food protein-originating peptides as tastants—Physiological, technological, sensory, and bioinformatic approaches. *Food Res. Int.* **2016**, *89*, 27–38, doi:10.1016/j.foodres.2016.08.010.
5. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
6. OpenSMILES. Available online: <http://opensmiles.org/> (accessed on 29 September 2017).
7. Ortiz-Martinez, M.; Gonzalez de Mejia, E.; García-Lara, S.; Aguilar, O.; Lopez-Castillo, L.M.; Otero-Pappatheodorou, J.T. Antiproliferative effect of peptide fractions isolated from a quality protein maize, a white hybrid maize, and their derived peptides on hepatocarcinoma human HepG2 cells. *J. Funct. Foods* **2017**, *34*, 36–48, doi:10.1016/j.jff.2017.04.015.
8. SwissTargetPrediction Program. Available online: <http://www.swisstargetprediction.ch/> (accessed on 29 September 2017).
9. Gfeller, D.; Grosdidier, A.; Wirth, M.; Daina, A.; Michielin, O.; Zoete, V. SwissTargetPrediction: A web server for target prediction of bioactive small molecules. *Nucleic Acids Res.* **2014**, *42*, W32–W38, doi:10.1093/nar/gku293.
10. Deming, T.J. Functional modification of thioether groups in peptides, polypeptides, and proteins. *Bioconj. Chem.* **2017**, *28*, 691–700, doi:10.1021/acs.bioconjchem.6b00696.
11. Obexer, R.; Walport, L.J.; Suga, H. Exploring sequence space: Harnessing chemical and biological diversity towards new peptide leads. *Curr. Opin. Chem. Biol.* **2017**, *38*, 52–61, doi:10.1016/j.cbpa.2017.02.020.
12. Sengupta, S.; Mehta, G. Late stage modification of peptides via CAH activation reactions. *Tetrahedron Lett.* **2017**, *58*, 1357–1372, doi:10.1016/j.tetlet.2017.02.069.
13. Stone, T.A.; Deber, C.M. Therapeutic design of peptide modulators of protein-protein interactions in membranes. *Biochim. Biophys. Acta Biomembr.* **2017**, *1859*, 577–585, doi:10.1016/j.bbamem.2016.08.013.
14. Chingle, R.; Proulx, C.; Lubell, W.D. Azapeptide synthesis methods for expanding side-chain diversity for biomedical applications. *Acc. Chem. Res.* **2017**, *50*, 1541–1556, doi:10.1021/acs.accounts.7b00114.
15. PepstrMod. Available online: <http://osddlinux.osdd.net/raghava/pepstrmod/> (accessed on 14 November 2017).
16. Singh, S.; Singh, H.; Tuknait, A.; Chaudhary, K.; Singh, B.; Kumaran, S.; Raghava, G.P.S. PEPstrMOD: Structure prediction of peptides containing natural, non-natural and modified residues. *Biol. Direct* **2015**, *10*, doi:10.1186/s13062-015-0103-4.
17. Raymond, J.-L. The chemical space project. *Acc. Chem. Res.* **2015**, *48*, 722–730, doi:10.1021/ar500432k.
18. Floris, M.; Moro, S. Mimicking peptides... In silico. *Mol. Inf.* **2012**, *31*, 12–20, doi:10.1002/minf.201100093.
19. BioTriangle Program. Available online: <http://biotriangle.scbdd.com/> (accessed on 29 September 2017).

20. Dong, J.; Yao, Z.-J.; Wen, M.; Zhu, M.-F.; Wang, N.-N.; Miao, H.-Y.; Lu, A.-P.; Zeng, W.-B.; Cao, D.-S. BioTriangle: A web-accessible platform for generating various molecular representations for chemicals, proteins, DNAs/RNAs and their interactions. *J. Cheminform.* **2016**, *8*, doi:10.1186/s13321-016-0146-2.
21. Swiss Institute of Bioinformatics. Available online: <http://www.sib.swiss/services-resources/databases-tools> (accessed on 29 September 2017).
22. SwissADME Program. Available online: <http://www.swissadme.ch/> (accessed on 29 September 2017).
23. Daina, A.; Michielin, O.; Zoete, V. SwissADME: A free web tool to evaluate pharmacokinetics, druglikeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* **2017**, *7*, doi:10.1038/srep42717.
24. WebMolCS Program. Available online: <http://www.gdbtools.unibe.ch:8080/webMolCS/> (accessed on 29 September 2017).
25. Awale, M.; Probst, D.; Reymond, J.-L. WebMolCS: A web-based interface for visualizing molecules in three-dimensional chemical spaces. *J. Chem. Inf. Model.* **2017**, *57*, 643–649, doi:10.1021/acs.jcim.6b00690.
26. University of Bern. Available online: <http://www.gdb.unibe.ch/> (accessed on 29 September 2017).
27. ChemAxon. Available online: <https://www.chemaxon.com/> (accessed on 14 November 2017).
28. Brainpeps Database. Available online: <http://brainpeps.ugent.be/> (accessed on 29 September 2017).
29. Van Dorpe, S.; Bronselaer, A.; Nielandt, J.; Stalmans, S.; Wynendaele, E.; Audenaert, K.; Van De Wiele, C.; Burvenich, C.; Peremans, K.; Hsueh, H.; et al. Brainpeps: The blood-brain barrier peptide database. *Brain Struct. Funct.* **2012**, *217*, 687–718, doi 10.1007/s00429-011-0375-0.
30. Qorumpeps Database. Available online: <http://quorumpeps.ugent.be/> (accessed on 9 November 2017).
31. Wynendaele, E.; Bronselaer, A.; Nielandt, J.; D'Hondt, M.; Stalmans, S.; Bracke, N.; Verbeke, F.; Van De Wiele, C.; De Tré, G.; De Spiegeleer, B. Quorumpeps database: Chemical space, microbial origin and functionality of quorum sensing peptides. *Nucleic Acids Res.* **2013**, *41*, D655–D659, doi:10.1093/nar/gks11.
32. AHTPDB Database. Available online: <http://crdd.osdd.net/raghava/ahtpdb/> (accessed on 15 November 2017).
33. Kumar, R.; Chaudhary, K.; Sharma, M.; Nagpal, G.; Chauhan, J.S.; Singh, S.; Gautam, A.; Raghava, G.P.S. AHTPDB: A comprehensive platform for analysis and presentation of antihypertensive peptides. *Nucleic Acids Res.* **2015**, *43*, D956–D962, doi:10.1093/nar/gku1141.
34. CancerPPD Database. Available online: <http://crdd.osdd.net/raghava/cancerppd/index.php> (accessed on 15 November 2017).
35. Tyagi, A.; Tuknait, A.; Anand, P.; Gupta, S.; Sharma, M.; Mathur, D.; Joshi, A.; Singh, S.; Gautam, A.; Raghava G.P.S. CancerPPD: A database of anticancer peptides and proteins. *Nucleic Acids Res.* **2015**, *43*, D837–D843, doi:10.1093/nar/gku892.
36. Hemolytik Database. Available online: <http://crdd.osdd.net/raghava/hemolytik/> (accessed on 15 November 2017).
37. Gautam, A.; Chaudhary, K.; Singh, S.; Joshi, A.; Anand, P.; Tuknait, A.; Mathur, D.; Varshney, G.C.; Raghava, G.P.S. Hemolytik: A database of experimentally determined hemolytic and non-hemolytic peptides. *Nucleic Acids Res.* **2014**, *42*, D444–D449, doi:10.1093/nar/gkt1008.
38. ParaPep Database. Available online: <http://crdd.osdd.net/raghava/parapep/> (accessed on 15 November 2017).
39. Mehta, D.; Anand, P.; Kumar, V.; Joshi, A.; Mathur, D.; Singh, S.; Tuknait, A.; Chaudhary, K.; Gautam, S.K.; Gautam, A.; et al. ParaPep: A web resource for experimentally validated antiparasitic peptide sequences and their structures. *Database* **2014**, doi:10.1093/database/bau051.
40. PepLife Database. Available online: <http://crdd.osdd.net/raghava/peplife/> (accessed on 15 November 2017).
41. Mathur, D.; Prakash, S.; Anand, P.; Kaur, H.; Agrawal, P.; Mehta, A.; Kumar, R.; Singh, S.; Raghava, G.P.S. PEPlife: A repository of the half-life of peptides. *Sci. Rep.* **2016**, *6*, doi:10.1038/srep36617.
42. SATPdb Database. Available online: <http://crdd.osdd.net/raghava/satpdb/links.php> (accessed on 29 September 2017).
43. Singh, S.; Chaudhary, K.; Dhanda, S.K.; Bhalla, S.; Usmani, S.S.; Gautam, A.; Tuknait, A.; Agrawal, P.; Mathur, D.; Raghava, G.P.S. SATPdb: A database of structurally annotated therapeutic peptides. *Nucleic Acids Res.* **2016**, *44*, D1119–D1126, doi:10.1093/nar/gkv1114.
44. BIOPEP Database. Available online: <http://www.uwm.edu.pl/biochemia/index.php/pl/biopep> (accessed on 29 September 2017).
45. Iwaniak, A.; Minkiewicz, P.; Darewicz, M.; Sieniawski, K.; Starowicz, P. BIOPEP database of sensory peptides and amino acids. *Food Res. Int.* **2016**, *85*, 155–161, doi:10.1016/j.foodres.2016.04.031.
46. OmicTools. Available online: <https://omictools.com/> (accessed on 29 September 2017).

47. Henry, V.J.; Bandrowski, A.E.; Pepin, A.-S.; Gonzalez, B.J.; Desfeux, A. OMICtools: An informative directory for multi-omic data analysis. *Database* **2014**, doi:10.1093/database/bau069.
48. MetaComBio. Available online: <http://www.uwm.edu.pl/metachemibio/index.php/about-metacombio> (accessed on 29 September 2017).
49. Minkiewicz, P.; Iwaniak, A.; Darewicz, M. Using internet databases for food science organic chemistry students to discover chemical compound information. *J. Chem. Educ.* **2015**, *92*, 874–876, doi:10.1021/ed5006739.
50. LabWorm. Available online: <https://labworm.com/> (accessed on 29 September 2017).
51. Udenigwe, C.C. Bioinformatics approaches, prospects and challenges of food bioactive peptide research. *Trends Food Sci. Technol.* **2014**, *36*, 137–143, doi:10.1016/j.tifs.2014.02.004.
52. PEP-FOLD Program. Available online: <http://bioserv.rpbs.univ-paris-diderot.fr/services/PEP-FOLD/> (accessed on 10 November 2017).
53. Shen, Y.; Maupetit, J.; Derreumaux, P.; Tufféry, P. Improved PEP-FOLD approach for peptide and miniprotein structure prediction. *J. Chem. Theory Comput.* **2014**, *10*, 4745–4758, doi:10.1021/ct500592m.
54. (PS)² v3. Available online: <http://ps2v3.life.nctu.edu.tw/> (accessed on 10 November 2017).
55. Huang, T.-T.; Hwang, J.-K.; Chen, C.-H.; Chu, C.-S.; Lee, C.-W.; Chen, C.-C. (PS)²: Protein structure prediction server version 3.0. *Nucleic Acids Res.* **2015**, *43*, W338–W342, doi:10.1093/nar/gkv454.
56. Iwaniak, A.; Minkiewicz, P.; Darewicz, M.; Protasiewicz, M.; Mogut, D. Chemometrics and cheminformatics in the analysis of biologically active peptides from food sources. *J. Funct. Foods* **2015**, *16*, 334–351, doi:10.1016/j.jff.2015.04.038.
57. Nongonierma, A.B.; FitzGerald, R.J. Learnings from quantitative structure-activity relationship (QSAR) studies with respect to food protein-derived bioactive peptides: A review. *RSC Adv.* **2016**, *6*, 75400–75413, doi:10.1039/C6RA12738J.
58. Chou, K.-C. Prediction of protein cellular attributes using pseudo-amino-acid-composition. *Proteins* **2001**, *43*, 246–255, doi:10.1002/prot.1035.
59. Chou, K.-C. Pseudo amino acid composition and its application in bioinformatics, proteomics and system biology. *Curr. Proteom.* **2009**, *6*, 262–274, doi:10.2174/157016409789973707.
60. Chou, K.-C.; Shen, H.-B. Recent progress in protein subcellular localization. *Anal. Biochem.* **2007**, *370*, 1–16, doi:10.1016/j.ab.2007.07.006.
61. Williams, A.J.; Ekins, S. A quality alert and call for improved curation of public chemistry databases. *Drug Discov. Today* **2011**, *16*, 747–750, doi:10.1016/j.drudis.2011.07.007.
62. Fourches, D.; Muratov, E.; Tropsha, A. Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* **2010**, *50*, 1189–1204, doi:10.1021/ci100176x.
63. Fourches, D.; Muratov, E.; Tropsha, A. Curation of chemogenomics data. *Nat. Chem. Biol.* **2015**, *11*, 535, doi:10.1038/nchembio.1881.
64. Fourches, D.; Muratov, E.; Tropsha, A. Trust, but verify II: A practical guide to chemogenomics data curation. *J. Chem. Inf. Model.* **2016**, *56*, 1243–1252, doi:10.1021/acs.jcim.6b00129.
65. SwissSidechain Database. Available online: <http://swissidechain.ch/> (accessed on 29 September 2017).
66. Gfeller, D.; Michielin, O.; Zoete, V. SwissSidechain: A molecular and structural database of non-natural sidechains. *Nucleic Acids Res.* **2013**, *41*, D327–D332, doi:10.1093/nar/gks991.
67. Norine Database. Available online: <http://bioinfo.lifl.fr/NRP/> (accessed on 29 September 2017).
68. Flissi, A.; Dufresne, Y.; Michalik, J.; Tonon, L.; Janot, S.; Noé, L.; Jacques, P.; Leclère, V.; Pupin, M. Norine, the knowledgebase dedicated to non-ribosomal peptides, is now open to crowdsourcing. *Nucleic Acids Res.* **2016**, *44*, D1113–D1118, doi:10.1093/nar/gkv1143.
69. CycloPS Program. Available online: <http://bioware.ucd.ie/~cyclops/cgi-bin/webpep.cgi> (accessed on 29 September 2017).
70. Duffy, F.J.; Verniere, M.; Devocelle, M.; Bernard, E.; Shields, D.C.; Chubb, A.J. CycloPs: Generating virtual libraries of cyclized and constrained peptides including nonnatural amino acids. *J. Chem. Inf. Model.* **2011**, *51*, 829–836, doi:10.1021/ci100431r.
71. Bohne-Lang, A.; Lang, E.; Förster, T.; von der Lieth, C.-W. LINUCS: Linear notation for unique description of carbohydrate sequences. *Carbohydr. Res.* **2000**, *336*, 1–11, doi:10.1016/S0008-6215(01)00230-0.
72. PubChem Database. Available online: <https://pubchem.ncbi.nlm.nih.gov/> (accessed on 29 September 2017).
73. Kim, S.; Thiessen, P.A.; Bolton, E.E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B.A.; et al. PubChem substance and compound databases. *Nucleic Acids Res.* **2016**, *44*, D1202–D1213, doi:10.1093/nar/gkv951.

74. Zhang, T.; Li, H.; Xi, H.; Stanton, R.V.; Rotstein, S.H. HELM: A hierarchical notation language for complex biomolecule structure representation. *J. Chem. Inf. Model.* **2012**, *52*, 2796–2806, doi:10.1021/ci3001925.
75. Milton, J.; Zhang, T.; Bellamy, C.; Swayze, E.; Hart, C.; Weissner, M.; Hecht S.; Rotstein, S. HELM software for biopolymers. *J. Chem. Inf. Model.* **2017**, *57*, 1233–1239, doi:10.1021/acs.jcim.6b00442.
76. ChEMBL Database. Available online: <https://www.ebi.ac.uk/chembl/db/> (accessed on 29 September 2017).
77. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A.P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L.J.; Cibrián-Uhalte, E.; et al. The ChEMBL database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954, doi:10.1093/nar/gkw1074/
78. Bartoloni, M.; Jin, X.; Marcaida, M.J.; Banha, J.; Dibonaventura, I.; Bongoni, S.; Bartho, K.; Gräbner, O.; Sefkow, M.; Darbre, T.; et al. Bridged bicyclic peptides as potential drug scaffolds: Synthesis, structure, protein binding and stability. *Chem. Sci.* **2015**, *6*, 5473–5490, doi:10.1039/C5SC01699A.
79. Heller, S.R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.* **2015**, *7*, doi:10.1186/s13321-015-0068-4.
80. ChemSpider Database. Available online: <http://www.chemspider.com/Default.aspx> (accessed on 29 September 2017).
81. Williams, A.; Tkachenko, V. The royal society of chemistry and the delivery of chemistry data repositories for the community. *J. Comput. Aided Mol. Des.* **2014**, *28*, 1023–1030, doi:10.1007/s10822-014-9784-5.
82. ZINC 15 Database. Available online: <http://zinc15.docking.org/> (accessed on 29 September 2017).
83. Sterling, T.; Irwin, J.J. ZINC 15—Ligand discovery for everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337, doi:10.1021/acs.jcim.5b00559.
84. HMDB Database. Available online: <http://www.hmdb.ca/> (accessed on 29 September 2017).
85. Wishart, D.S.; Jewison, T.; Guo, A.C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; et al. HMDB 3.0—The human metabolome database in 2013. *Nucleic Acids Res.* **2013**, *41*, D801–D807, doi:10.1093/nar/gks1065.
86. Ketcher Program. Available online: <http://lifescience.opensource.epam.com/ketcher/> (accessed on 29 September 2017).
87. Karulin, B.; Kozhevnikov, M. Ketcher: Web-based chemical structure editor. *J. Cheminform.* **2011**, *3* (Suppl. 1), doi:10.1186/1758-2946-3-S1-P3.
88. Open Babel Program. Available online: http://openbabel.org/wiki/Main_Page (accessed on 29 September 2017).
89. O’Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, doi:10.1186/1758-2946-3-33.
90. Pearson, W.R. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* **2000**, *132*, 185–219.
91. Wikipedia FASTA Format. Available online: https://en.wikipedia.org/wiki/FASTA_format (accessed on 29 September 2017).
92. UniProt Database. Available online: <http://www.uniprot.org/> (accessed on 29 September 2017).
93. The UniProt Consortium. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D158–D169, doi:10.1093/nar/gkw10.
94. Wkipedia Protein Data Bank File Format. Available online: [https://en.wikipedia.org/wiki/Protein_Data_Bank_\(file_format\)](https://en.wikipedia.org/wiki/Protein_Data_Bank_(file_format)) (accessed on 9 November 2017).
95. Ertl, P. Molecular structure input on the web. *J. Cheminform.* **2010**, *2*, doi:10.1186/1758-2946-2-1.
96. Varki, A. Biological roles of glycans. *Glycobiology* **2017**, *27*, 3–49, doi:10.1093/glycob/cww086.
97. Okerblom, J.; Varki, A. Biochemical, cellular, physiological, and pathological consequences of human loss of N-glycolylneuraminic acid. *ChemBioChem* **2017**, *18*, 1155–1171, doi:10.1002/cbic.201700077.
98. Palaniappan, K.K.; Bertozzi, C.R. Chemical glycoproteomics. *Chem. Rev.* **2016**, *116*, 14277–14306, doi:10.1021/acs.chemrev.6b00023.
99. Clark, A.M. Basic primitives for molecular diagram sketching. *J. Cheminform.* **2010**, *2*, doi:10.1186/1758-2946-2-8.
100. Cahn, R.S.; Ingold, C.K.; Prelog, V. Specification of molecular chirality. *Angew. Chem. Int. Ed.* **1966**, *5*, 385–415, doi:10.1002/anie.196603851.
101. Wikipedia Cahn–Ingold–Prelog Priority Rules. Available online: https://en.wikipedia.org/wiki/Cahn%E2%80%93Ingold%E2%80%93Prelog_priority_rules (accessed on 29 September 2017).
102. Chemical Identifier Resolver Program. Available online: <https://cactus.nci.nih.gov/chemical/structure> (accessed on 29 September 2017).

103. Muresan, S.; Sitzmann, M.; Southan, C. Mapping between databases of compounds and protein targets. *Meth. Mol. Biol.* **2012**, *910*, 145–164, doi:10.1007/978-1-61779-965-5_8.
104. Chemical Translation Service Program. Available online: <http://cts.fiehnlab.ucdavis.edu/> (accessed on 29 September 2017).
105. Wohlgemuth, G.; Haldiya, P.K.; Willighagen, E.; Kind, T.; Fiehn, O. The chemical translation service—A web-based tool to improve standardization of metabolomic reports. *Bioinformatics* **2010**, *26*, 2647–2648, doi:10.1093/bioinformatics/btq476.
106. Smiles2Monomers Program. Available online: <http://bioinfo.lifl.fr/norine/smiles2monomers.jsp> (accessed on 29 September 2017).
107. Dufresne, Y.; Noé, L.; Leclère, V.; Pupin, M. Smiles2Monomers: A link between chemical and biological structures for polymers. *J. Cheminform.* **2015**, *7*, doi:10.1186/s13321-015-0111-5.
108. Southan, C. InChI in the wild: An assessment of InChIKey searching in Google. *J. Cheminform.* **2013**, *5*, doi:10.1186/1758-2946-5-10.
109. Warr, W.A. Many InChIs and quite some feat. *J. Comput. Aided Mol. Des.* **2015**, *29*, 681–694, doi:10.1007/s10822-015-9854-3.
110. SwissDock Program. Available online: <http://www.swissdock.ch/> (accessed on 29 September 2017).
111. Grosdidier, A.; Zoete, V.; Michielin, O. SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res.* **2011**, *39*, W270–W277, doi:10.1093/nar/gkr366.
112. AMMOS2 Program. Available online: <http://drugmod.rpbs.univ-paris-diderot.fr/ammosHome.php> (accessed on 29 September 2017).
113. Labbé, C.; Pencheva, T.; Jereva, D.; Desvillechabrol, D.; Becot, J.; Villoutreix, B.O.; Pajeva, I.; Miteva, M.A. AMMOS2: A web server for protein-ligand-water complexes refinement via molecular mechanics. *Nucleic Acids Res.* **2017**, *45*, W350–W355, doi:10.1093/nar/gkx397.
114. ProteinsPlus Program. Available online: <http://proteinsplus.zbh.uni-hamburg.de/> (accessed on 29 September 2017).
115. Fährrolfes, R.; Bietz, S.; Flachsenberg, F.; Meyder, A.; Nittinger, E.; Otto, T.; Volkamer, A.; Rarey, M. ProteinsPlus: A web portal for structure analysis of macromolecules. *Nucleic Acids Res.* **2017**, *45*, W337–W343, doi:10.1093/nar/gkx333.
116. LIPID MAPS® Database. Available online: <http://www.lipidmaps.org/> (accessed on 29 September 2017).
117. Sud, M.; Fahy, E.; Cotter, D.; Brown, A.; Dennis, E.A.; Glass, C.K.; Merrill, A.H.; Murphy, R.C.; Raetz, C.R.H.; Russell, D.W.; et al. LMSD: LIPID MAPS structure database. *Nucleic Acids Res.* **2007**, *35*, D527–D532, doi:10.1093/nar/gkl838.
118. Campbell, M.P.; Ranzinger, R.; Lütke, T.; Mariethoz, J.; Hayes, C.A.; Zhang, J.; Akune, Y.; Aoki-Kinoshita, K.F.; Damerell, D.; Carta, G.; et al. Toolboxes for a standardised and systematic study of glycans. *BMC Bioinform.* **2014**, *15* (Suppl. 1), doi:10.1186/1471-2105-15-S1-S9.
119. WURCS Program. Available online: <http://www.wurcs-wg.org/software.php> (accessed 29 September 2017).
120. Matsubara, M.; Aoki-Kinoshita, K.F.; Aoki, N.P.; Yamada, I.; Narimatsu, H. WURCS 2.0 update to encapsulate ambiguous carbohydrate structures. *J. Chem. Inf. Model.* **2017**, *57*, 632–637, doi:10.1021/acs.jcim.6b00650.
121. GlyTouCan Database. Available online: <https://glytoucan.org/> (accessed on 29 September 2017).
122. Aoki-Kinoshita, K.; Agravat, S.; Aoki, N.P.; Arpinar, S.; Cummings, R.D.; Fujita, A.; Fujita, N.; Hart, G.M.; Haslam, S.M.; Kawasaki, T.; et al. GlyTouCan 1.0—The international glycan structure repository. *Nucleic Acids Res.* **2016**, *44*, D1237–D1242, doi:10.1093/nar/gkv1041.
123. Wikipedia Chemical Table File. Available online: https://en.wikipedia.org/wiki/Chemical_table_file#Molfile (accessed on 29 September 2017).
124. Tanaka, K.; Aoki-Kinoshita, K.F.; Kotera, M.; Sawaki, H.; Tsuchiya, S.; Fujita, N.; Shikanai, T.; Kato, M.; Kawano, S.; Yamada, I.; et al. WURCS: The Web3 unique representation of carbohydrate structures. *J. Chem. Inf. Model.* **2014**, *54*, 1558–1566, doi:10.1021/ci400571e.
125. JCGGBD Database. Available online: http://jcgddb.jp/database_en.html (accessed on 29 September 2017).
126. Maeda, M.; Fujita, N.; Suzuki, Y.; Sawaki, H.; Shikanai, T.; Narimatsu, H. JCGGDB: Japan consortium for glycobiology and glycotecnology database. *Methods Mol. Biol.* **2015**, *1273*, 161–179, doi:10.1007/978-1-4939-2343-4_12.

127. Alves, V.M.; Muratov, E.N.; Capuzzi, S.J.; Politi, R.; Low, Y.; Braga, R.C.; Zakharov, A.V.; Sedykh, A.; Mokshyna, E.; Farag, S.; et al. Alarms about structural alerts. *Green Chem.* **2016**, *18*, 4348–4360, doi:10.1039/C6GC01492E.
128. Tetko, I.V.; Engkvist, O.; Koch, U.; Reymond, J.-L.; Chen, H. BIGCHEM: Challenges and opportunities for big data analysis in chemistry. *Mol. Inf.* **2016**, *35*, 615–621, doi:10.1002/minf.201600073.



© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).