Computational and Structural Biotechnology Journal 19 (2021) 3628–3639

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Extended many-item similarity indices for sets of nucleotide and protein sequences

Dávid Bajusz [a,1,2], Ramón Alain Miranda-Quintana [b,*,1,3], Anita Rácz [c,4], Károly Héberger [c,*,5]

[a] Medicinal Chemistry Research Group, Research Centre for Natural Sciences, Magyar tudósok krt. 2, 1117 Budapest, Hungary
[b] Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, FL 32611, USA
[c] Plasma Chemistry Research Group, Research Centre for Natural Sciences, Magyar tudósok krt. 2, 1117 Budapest, Hungary

ARTICLE INFO

Article history:
Received 11 March 2021
Received in revised form 7 June 2021
Accepted 14 June 2021
Available online 16 June 2021

Keywords:
Multiple comparisons
DNA sequences
Protein sequences
Diversity analysis
Similarity indices
Consistency
ANOVA
Human protein kinases
Human SH2 domains
Cytochrome P450

ABSTRACT

Quantification of similarities between protein sequences or DNA/RNA strands is a (sub-)task that is ubiquitously present in bioinformatics workflows, and is usually accomplished by pairwise comparisons of sequences, utilizing simple (*e.g.* percent identity) or more intricate concepts (*e.g.* substitution scoring matrices). Complex tasks (such as clustering) rely on a large number of pairwise comparisons under the hood, instead of a direct quantification of set similarities. Based on our recently introduced framework that enables multiple comparisons of binary molecular fingerprints (*i.e.*, direct calculation of the similarity of fingerprint sets), here we introduce novel symmetric similarity indices for analogous calculations on sets of character sequences with more than two (t) possible items (*e.g.* DNA/RNA sequences with $t = 4$, or protein sequences with $t = 20$). The features of these new indices are studied in detail with analysis of variance (ANOVA), and demonstrated with three case studies of protein/DNA sequences with varying degrees of similarity (or evolutionary proximity). The Python code for the extended many-item similarity indices is publicly available at: https://github.com/ramirandaq/tn_Comparisons.

1. Introduction

Much like molecular similarity is a key concept of cheminformatics [1,2], the comparison of amino acid and nucleotide sequences is a cornerstone of bioinformatics. Both are based on the similarity principle, *i.e.* structurally similar molecules are presumed to exhibit similar properties and similar biological activities, and analogously, similar nucleotide or amino acid sequences most often encode proteins with similar biological function. Despite the common philosophical roots, the core methodologies of cheminformatics and bioinformatics are quite different, partly due to the different data representations of molecules *vs.* macromolecular sequences. Since small molecular structures were primarily conceived as drawings on paper, their computational representations were developed from scratch and refined over the past decades, yielding a rich selection of file formats and binary molecular fingerprints [3]. Molecular fingerprints offer–among other advantages [4,5]–a direct way to quantify the similarity of molecules, with the application of binary similarity metrics (yielding pairwise similarity values usually in the [0;1] range, with a value of 1 corresponding to identical objects/fingerprints). While many such metrics exist [6], the past decades of practice have cemented the Tanimoto coefficient as the most popular similarity coefficient [7,8], despite its known shortcomings [9,10].

In contrast, the representation of macromolecular sequences was quite straightforward from the start, as sequences of one-letter monomer (amino acid or nucleotide) codes. Here, an additional task is finding the optimal alignment of two (or more) sequences prior to evaluating their similarities. While the basics of (global) sequence alignments have been established already in the 1970's [11], decades of refinement have yielded local alignment algorithms, particularly BLAST (Basic Local Alignment Search Tool) as today's standard sequence alignment tool [12], and new generation alignment algorithms are still being developed [13]. For quantifying the similarity between two aligned sequences, percent identity values (for protein sequences, also percent similarity

* Corresponding authors.
E-mail addresses: quintana@chem.ufl.edu (R.A. Miranda-Quintana), heberger.karoly@ttk.mta.hu (K. Héberger).
[1] These authors have contributed equally.
[2] https://orcid.org/0000-0003-4277-9481
[3] https://orcid.org/0000-0003-2121-4449
[4] https://orcid.org/0000-0001-8271-9841
[5] https://orcid.org/0000-0003-0965-939X

https://doi.org/10.1016/j.csbj.2021.06.021

values) are used, together with the expectation value (E) of finding an equivalent alignment by chance. Protein sequence alignments can also be assessed by substitution scoring matrices, which contain additive score contributions for each possible exchange of amino acid A to amino acid B. (With Point Accepted Mutation (PAM) matrices [14] and Blocks Substitution Matrices (BLOSUM) being the most popular such tools [15].)

In our recent works with binary (molecular and other) fingerprints, we have provided statistical findings that support the use of the Tanimoto coefficient [8], but we could also identify some coefficients, which are more advantageous in some circumstances, *e.g.* for metabolomic [16] or protein–ligand interaction fingerprints [17]. We have also introduced *differential consistency analysis* (DCA), a rigorous mathematical framework to reveal consistencies between any pair of similarity metrics [18]. Most importantly, in the direct predecessors of this work, we introduced the idea of comparing more than two molecules (*i.e.* groups/sets of molecules) at a time, defined a series of extended similarity indices based on this idea, and selected the best indices for further usage [19]. In a companion paper, we have proven the computational advantage of these new indices in assessing the similarity of large sets of molecules, and provided illustrative examples for their usage in: i) the selection of diverse compound sets, ii) clustering applications, and iii) assessing the compactness of clusters corresponding to ligands of different pharmaceutical targets [20]. Much of this is possible due to the unprecedented computational efficiency of our indices in quantifying the similarities of sets with an arbitrary number of objects.

In this work, we generalize this formalism even further, introducing extended many-item–or (*t*,*n*)–similarity indices to compare any number of objects *n*, containing any finite number *t* of categorical variables. Realizing the prospective applications in bioinformatics, we showcase the usage of the new similarity indices on protein families and subfamilies relevant for current medicinal chemistry, using their DNA (t = 4), and amino acid (t = 20) sequences. We also introduce a simple amino acid categorization scheme to account for sidechains of similar character (t = 8). A thorough literature search reveals that related approaches are scarce: "integer coding" is sporadically used and definitely not for uncovering (macro)molecular similarities. Terms such as "non-binary similarity coefficients" usually refer to the use of (pairwise) similarity metrics for ordinal (integer) or continuous data [21,22]. Further uses appear in the distantly related fields of process control [23,24], feature selection [25] and multicriteria decision making [26]. Therefore, our work presents the first approach to directly compare arbitrarily large sets of DNA and amino acid sequences. Hence, we suggest a terminology of (*t*,*n*)-comparisons, *i.e.* the comparison of *n* objects (sequences) containing *t* possible characters, as an extension of: i) (2,2)-comparisons, the "traditional approach" for the pairwise comparison of binary (molecular) fingerprints, and ii) (2,*n*)-comparisons, our recent generalization to compare an arbitrary number *n* of such binary fingerprints [19,20]. More specifically for DNA and amino acids, these are (4, *n*)- and (20, *n*)-comparisons, as demonstrated in the case studies that are included in the present work

2. Theory

First, let us introduce some elements of notation. As explained before, we will use the term (*t*,*n*)-comparison for a comparison of *n* sequences, each containing *m* characters from a set of *t* items. An alternative term that we use here (also in the title) are "extended many-item comparisons", with "extended" meaning that we are comparing more than two objects (sequences) simultaneously (in contrast to pairwise comparisons) and "many-item" meaning that there are more than two possible characters in each position of the sequences

As a reference, the pairwise comparison of binary sequences, *i.e.* (2,2)-comparisons are used ubiquitously in cheminformatics to define the similarity of molecules by comparing their binary fingerprints (sequences of zeros and ones). In this case, each bit position can contribute to the occurrence of four events: (1,1), (1,0), (0,1) and (0,0), which are summed in the counters *a*, *b*, *c* and *d*, respectively. Binary similarity metrics are then defined with the use of these counters (*e.g.* the popular Tanimoto coefficient is given as $a / (a + b + c)$). Notice that the counters *a* and *d* express similarity of the two sequences, while *b* and *c* express dissimilarity in the given positions. Our core idea for the generalization of similarity metrics for the comparison of more sequences (*i.e.* (2,*n*)-comparisons), is that even for arbitrarily large sets of compared objects, we can classify each bit position as a similarity or dissimilarity counter. For example, if we compare ten sequences and there are eight co-occurring 1 bits (and two 0 bits) in a given position, that will contribute to the similarity, while five co-occurring 1 bits (and five 0 bits) will contribute to the dissimilarity of the ten objects. In our recent work, we provide a systematic approach to the classification of positions into similarity and dissimilarity counters, using an indicator we have termed $\Delta_{n(k)}=|2k-n|$ and a coincidence threshold γ [19]. In this work, we take this generalization one step further by allowing an arbitrary number of *t* different characters, instead of two.

In the more general (*t*,*n*)-comparisons we have sequences formed by $X_1, X_2, \cdots, X_t$ distinct characters (it is arbitrary how we choose to represent these characters, they can be numbers, letters, *etc.*). In this work we will only discuss the "democratic matching" case, meaning that matching *k* characters of type X_p is equivalent to matching *k* characters of type X_q. This means that we can directly study similarity indices, which yield the following form for (2,*n*)-comparisons:

$$s = \sum\nolimits_i G_i \left(\frac{g_{i1}(a+d, b+c)}{g_{i2}(a+d, b+c)} \right) \tag{1}$$

where the terms $a + d$ and $b + c$ are the counts over the similarity and dissimilarity counters, respectively, as introduced in our recent work for (2,*n*)-comparisons [19,20]. (Briefly, $a + d$ and $b + c$ are the numbers of sequence positions where the frequency of either ones or zeros is above/below a predefined confidence threshold γ, respectively, see also below.) In particular, formula (1) holds for the SM (simple matching), RT (Rogers-Tanimoto), SS2 (Sokal-Sneath), CT1, CT2 (Consonni-Todeschini), and AC (Austin-Colwell) indices.

The first step is counting, for each position of the sequences, the number of matches for each type of character. To fix ideas, let us consider the following case of five short DNA-sequences:

$$\begin{aligned} F_1 &= (CGCTACAA) \\ F_2 &= (AACAGCAC) \\ F_3 &= (CGCTCAAC) \\ F_4 &= (AACCACAA) \\ F_5 &= (AGCTTCAT) \end{aligned} \tag{2}$$

We can write up a general coincidence matrix as follows:

	1	2	3	4	5	6	7	8
A	3	2	0	1	2	1	5	2
C	2	0	5	1	1	4	0	2
G	0	3	0	0	1	0	0	0
T	0	0	0	3	1	0	0	1

where the columns label the position (b), and the rows label the type of character (j). Each entry in this table corresponds to k_{j-b}, that is, the number of times that character j appears in the position b.

The next step is to assign a coincidence value to each bit position. As we follow the philosophy to maximize the final similarity, we assign to each position the column maximum. That is, from the previous table we can extract the reduced coincidence vector:

	1	2	3	4	5	6	7	8
Coincidence	3	3	5	3	2	4	5	2

In other words, denoting the coincidence over bit b by k_b:

$$k_b = max_j\{k_{j-b}\} \tag{3}$$

Notice that, by definition (and if there are no gaps in the compared sequences):

$$\left\lceil \frac{n}{t} \right\rceil \leqslant k_b \leqslant n \tag{4}$$

where $\lceil x \rceil$ is the ceiling function.

Now, for each k_b we calculate the indicator $\Delta_{t-n(k_b)}$ that will enable us to classify the various possible values of k_b as either *similarity* or *dissimilarity* counters:

$$\Delta_{t-n(k_b)} = tk_b - n \tag{5}$$

According to Eq. (4), possible values of this indicator will be in the following range:

$$(t - n\mathrm{mod}t)\mathrm{mod}t \leq \Delta_{t-n(k_b)} \leq n(t-1) \tag{6}$$

This means that the coincidence threshold to classify the various possible values of k_b as either *similarity* or *dissimilarity* counters, γ, will have to be in the range:

$$(t - n\mathrm{mod}t)\mathrm{mod}t \leq \gamma < n(t-1) \tag{7}$$

Thus, if we want to maximize the similarity in the end, we must choose the following coincidence threshold:

$$\gamma = (t - n\mathrm{mod}t)\mathrm{mod}t \tag{8}$$

We note that $(2 - n\mathrm{mod}2)\mathrm{mod}2 = n\mathrm{mod}2$, which was the expression we used in the $t = 2$ (binary molecular fingerprints) case [19]. In the present case, $\gamma = (4 - 5\mathrm{mod}4)\mathrm{mod}4 = 3$.

In general, a k_b value will indicate similarity if:

$$\Delta_{t-n(k_b)} > \gamma \tag{9}$$

and dissimilarity if:

$$\Delta_{t-n(k_b)} \leq \gamma \tag{10}$$

Now we define the counters. A counter $C_{t-n(k)}$ will count the number of times a given k appears in the set of all k_b. In the example, we have been following so far, we have:

$$C_{4-5(5)} = 2, C_{4-5(4)} = 1, C_{4-5(3)} = 3, C_{4-5(2)} = 2 \tag{11}$$

Notice that with this new definition we will have $n - \lfloor \frac{n}{t} \rfloor$ counters. Also, as expected, the sum of all the counters is equal to the length of the sequences (m).

We classify each counter as similarity or dissimilarity counters according to Eqs. (9) and (10), so:

$$\begin{aligned} &C_{4-5(5)} \text{ is a similarity counter} \\ &C_{4-5(4)} \text{ is a similarity counter} \\ &C_{4-5(3)} \text{ is a similarity counter} \\ &C_{4-5(2)} \text{ is a dissimilarity counter} \end{aligned} \tag{12}$$

Analogously to the $t = 2$ case, certain counters might indicate a stronger measure of similarity/dissimilarity, *e.g.* in our example, $C_{4-5(5)}$ (the same character repeating in all 5 sequences at the same position) is stronger than $C_{4-5(3)}$ (the same character repeating in 3 out of 5 sequences at the same position). We can use suitable weighting functions to reflect this, using the natural generalization of the weight functions used in the $t = 2$ case [19], namely:

$$f_s(\Delta_{t-n(k)}) = \frac{\Delta_{t-n(k)}}{n(t-1)} \tag{13}$$

$$f_d(\Delta_{t-n(k)}) = 1 - \frac{\Delta_{t-n(k)} - (t - n\mathrm{mod}t)\mathrm{mod}t}{n(t-1)} \tag{14}$$

Now we have all the necessary ingredients to calculate the similarity indices. In the case of the weighted simple matching (or Sokal-Michener) index, for instance:

$$SM_{4-5\_w} = \frac{\sum_s f_s(\Delta_{n(k)})C_{n(k)}}{\sum_s f_s(\Delta_{n(k)})C_{n(k)} + \sum_d f_d(\Delta_{n(k)})C_{n(k)}} \tag{15}$$

$$SM_{4-5\_w} = \frac{f_s(\Delta_{4-5(5)})C_{4-5(5)} + f_s(\Delta_{4-5(4)})C_{4-5(4)} + f_s(\Delta_{4-5(3)})C_{4-5(3)}}{f_s(\Delta_{4-5(5)})C_{4-5(5)} + f_s(\Delta_{4-5(4)})C_{4-5(4)} + f_s(\Delta_{4-5(3)})C_{4-5(3)} + f_d(\Delta_{4-5(2)})C_{4-5(2)}} \tag{16}$$

$$\begin{aligned} SM_{4-5\_w} &= \frac{1.0 \times 2 + 0.733 \times 1 + 0.467 \times 3}{1.0 \times 2 + 0.733 \times 1 + 0.467 \times 3 + 1.0 \times 2} \\ &= \frac{4.133}{6.133} = 0.674 \end{aligned} \tag{17}$$

Analogously as for the (2,n) similarity metrics, we can choose to omit the weighting schemes from the denominator and define the non-weighted analogs of the similarity indices, *e.g.* for the *SM* index:

$$SM_{4-5\_nw} = \frac{\sum_s f_s(\Delta_{n(k)})C_{n(k)}}{\sum_s C_{n(k)} + \sum_d C_{n(k)}} \tag{18}$$

$$SM_{4-5\_nw} = \frac{f_s(\Delta_{4-5(5)})C_{4-5(5)} + f_s(\Delta_{4-5(4)})C_{4-5(4)} + f_s(\Delta_{4-5(3)})C_{4-5(3)}}{C_{4-5(5)} + C_{4-5(4)} + C_{4-5(3)} + C_{4-5(2)}} \tag{19}$$

$$\begin{aligned} SM_{4-5\_nw} &= \frac{1.0 \times 2 + 0.733 \times 1 + 0.467 \times 3}{2 + 1 + 3 + 2} = \frac{4.133}{8} \\ &= 0.517 \end{aligned} \tag{20}$$

The weighted and non-weighted formulas for each of the available similarity metrics are collected in Appendix 1.

An additional part of the formalism that is especially important for DNA and protein sequences is the question of how we handle gaps in a set of aligned sequences. In our approach, positions with a majority of gap (-) characters are omitted from further analysis (they are regarded neither as similarity, nor as dissimilarity counters), but a position with a smaller number of gaps can still be considered as a similarity or dissimilarity counter. More precisely: whenever we have a bit position for which the maximum number of identical "physical" (non-gap) characters is less than $\frac{n}{t}$, then we ignore that position, as it does not contain enough information to say if it conveys similarity or dissimilarity.

3. Results

3.1. Individual index variations

In order to explore how the extended many-item similarity metrics behave for different input data, we have generated random

four-item (t = 4) character sequences of various lengths (m = 10, 100, 1000 and 100 000) and calculated the extended similarity values for various numbers of compared objects (n), according to both the weighted (w) and non-weighted (nw) formulas. In each case, we randomly generated 16 sequences. First, let us study how the average (of the absolute value) $|s|$ of the comparisons with an individual index s changes when we change n (Fig. 1).

An interesting pattern emerges here, similarly to the case of molecular fingerprints (bitvectors, corresponding to t = 2) [19]: notice the alternating "zigzag" pattern of maxima and minima with a period of four. This behavior can be rationalized on the basis of the potential number of dissimilarity counters when changing the number of compared objects (sequences). Generally, both our results for molecular fingerprints, as well as Fig. 1 reflect an interesting characteristic of the extended comparisons: the average values will tend to show an oscillating behavior with a period of t when we change the value of n. In addition, applying a weighting scheme clearly amplifies these differences (Fig. 1).

3.2. Analysis of mean similarity indices and ranking behavior

Next, we compared the range of similarity values returned by the various similarity metrics for the same dataset (Fig. 2). The indices cover different ranges from almost zero to one.

We can plausibly assume that all quaternary similarity indices express the similarity of the sequence sets with some error. As such, analysis of variance (ANOVA) is a suitable technique to decompose the effects of different factors. Here, the following factors were considered: F2–number (n) of objects (sequences) compared, 14 levels with n = 2, 3, … 15; F3–weighting, two levels: weighted and non-weighted versions; F4–the similarity coefficients themselves: six levels (see Appendix 1); F5–length of the sequences, four levels: m = 10, 100, 1000, 100 000 (F1 is reserved for k-fold cross-validation iterations, see later). Altogether 14*2*6*4 = 6672 items (averages of similarity indices) have been decomposed into the above factors. The averages of the quaternary indices show a characteristic zigzag pattern (Fig. 3). The role of weighting is illustrated in Fig. 3 as a function of the number of n (number of objects compared).

Here, the same zigzag pattern is observed as in Fig. 1, with the "amplitude" being damped for non-weighted coefficients. Weighted and non-weighted averages cover different ranges, the non-weighted indices are dispersed between ~ 0.2 and 0.5, whereas the weighted ones range from ~ 0.2 to 0.9. The variances do not change much as the multiplicity of comparisons (n) increases. Weighting has no effect on pairwise (n = 2) comparisons; and a similarly small effect can be seen for quintic (n = 5) comparisons. The difference of weighted and non-weighted metrics is the largest for n = 4, 8 and 12, *i.e.* where n is an exact multiple of t. The coupling of various factors reflects the behavior of the individual indices nicely (Fig. 4).

The two quaternary Consonni-Todeschini indices change reversely, with the gap between weighted and unweighted versions diminishing as the length of the sequences increases. The main reason for the fringe behavior of the CT indices is the presence of terms like "1 + p" and "1 + b + c" in their expression, which mean that the values of these indices will depend heavily on the length of the compared sequences. Compare this to an index like SM, which is calculated via $(a + d)/p$, it is clear that increasing the length of the sequences, the density of a given type of character is going to remain sensibly constant, so the overall effect is going to cancel between the numerator and denominator. This is in contrast with CT1 and CT2, in which the value of the indices will tend to monotonically increase or decrease, respectively, when we increase the length of the sequences, since the "1+" terms break the proportionality. This means that the CT indices cannot retrieve similarity information from a set in a robust way, because of its marked length-dependence (note that this effect is more marked in the case of CT2). The other indices have values in a relatively narrower range and exhibit a slight increase as a function of sequence length.

To prioritize between the extended indices for further usage, a property that is even more important than the actual similarity values is the way the indices rank different groups of objects (sequences), as compared to an ideal reference method (benchmark). For this purpose, we have used Sum of Ranking Differences (SRD) [27], a robust multicriteria decision making tool that was extensively applied in our earlier works [8,17,19]. Briefly, SRD expresses the closeness of the individual methods (extended similarity metrics) to the reference method by calculating a normalized Manhattan distance (termed the SRD value) between them, after rank-transformation. The smaller the SRD value, the closer the given metric is to the reference (benchmark). Here, the reference was defined as the average of the six similarity metrics, based on the consideration that the average will cancel out the individual errors at least partially. An illustrative example of SRD calculations is shown in Supplementary Figure S1.

After completing all SRD calculations for the groups detailed above (with varying settings of n, m and weighting) with two types of sevenfold cross-validation, variance analysis (ANOVA) was completed on the resulting SRD values. The following factors were considered: F1–number (n) of compared objects (fingerprints or other representations), 14 levels: n = 2, 3, … 15; F2–the similarity coefficients themselves: 6 levels. The role of weighting (F3) was evaluated separately, because of the different scales of the weighted and non-weighted versions within the limits of 0 and 100. The ANOVA of the average similarity values (above) and the one on SRD scores exhibits an essential difference, as the smaller SRD values are better. Fig. 5 shows the comparison of the six different similarity measures in the t = 4 case.

First we can observe that weighted similarity metrics give almost identical rankings: SRD values are close to 0 for each metric, even for the two "outliers" CT1 and CT2. There is greater distinction for non-weighted metrics: here, Rogers-Tanimoto (RT) emerges as the best option (closest to the reference), while Sokal-Sneath 2 (SS2) is the least favorable one. Taken together with the marked size dependence of CT1 and CT2 (see above), we can recommend the Rogers-Tanimoto (RT), and also the Austin-Colwell (AC) and Sokal-Michener (SM) metrics for further usage. Additionally, we have compared the average SRD values in terms of the number of compared objects n (Supplementary Figure S2): for non-weighted metrics, a similar zigzag pattern (with maxima at multiples of four) emerges as for the similarity values themselves, while for weighted metrics, there is again almost no distinction between the metrics (SRD values close to 0).

3.3. Case studies of selected datasets

To evaluate our method in real-life scenarios, we have collected protein and DNA sequences for three families of proteins, each of which have great importance in medicinal chemistry. The three datasets also correspond to distinct cases in terms of the number and population of subfamilies, as well as their sequence diversities. We have used the n-ary similarity measures introduced here to quantify the similarity of these protein families and their subfamilies, based on their protein (t = 20), simplified protein (t = 8) and DNA (t = 4) sequences. For the simplified protein sequences, we have re-coded the protein sequences by classifying the amino acids into eight groups based on the chemical/pharmacophoric character of their side chains (see Table 1).

The first case study involves the sequence comparison and multiple (n-ary) similarity calculations of human protein kinases. Protein kinases are regulatory and signaling proteins that account for

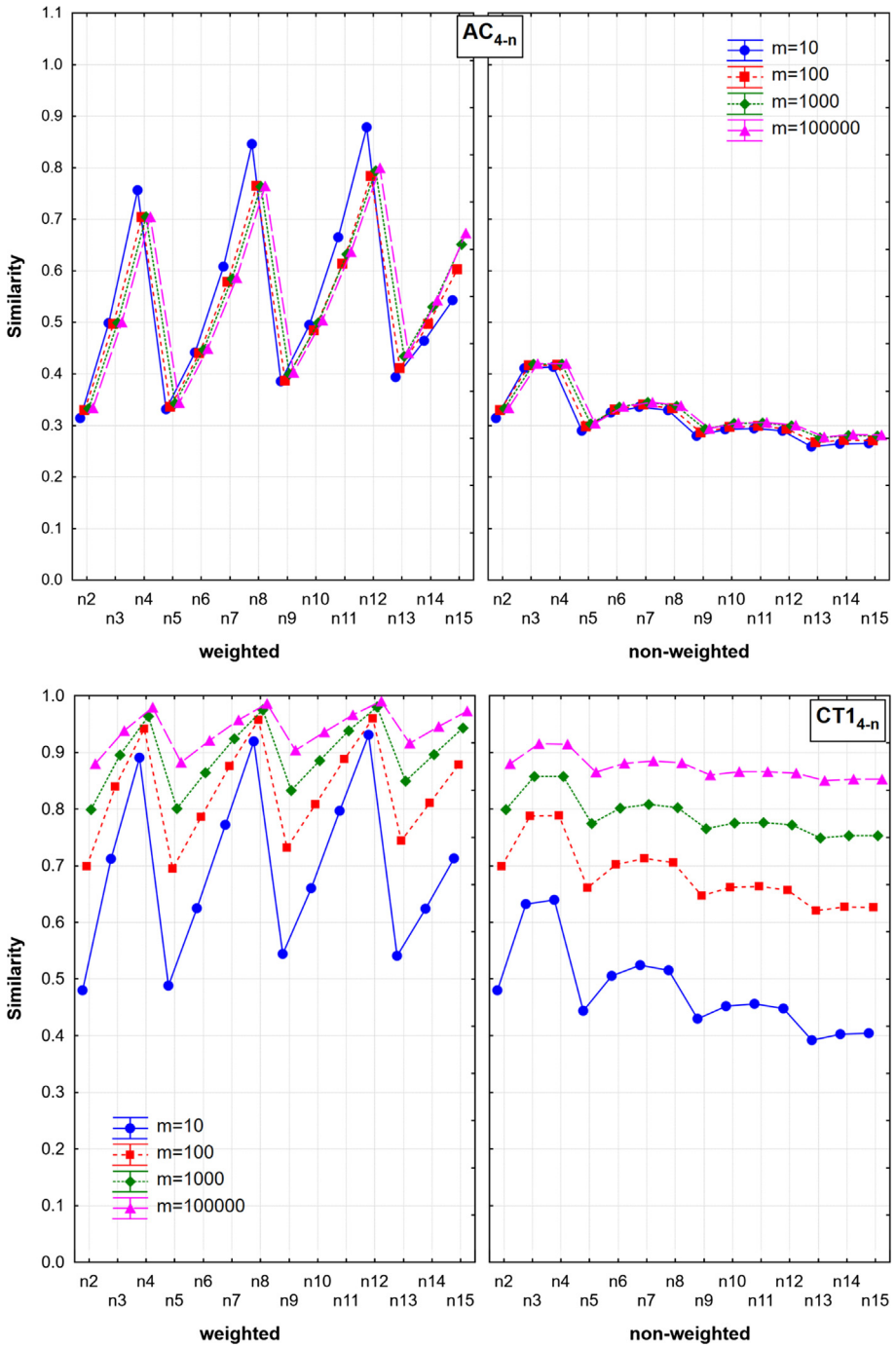


Fig. 1. Variation of the average (of the absolute value) of all possible *n*-ary comparisons over 16 sequences of length $m = 10$ to $m = 100\,000$ for different values of n for the quaternary ($t = 4$) weighted (w,left) and non-weighted (nw,right) AC (Austin-Colwell, top) and CT1 (Consonni-Todeschini 1, bottom) indices.

roughly 2% of the total human proteome [28], many of them are important pharmaceutical targets in mostly oncological indications [29]. There are close to 500 protein kinases, having relatively diverse sequences, but a conserved structure (corresponding to their identical enzymatic function of transferring a phosphoryl group to a sidechain of a downstream signaling protein), classically grouped into eight subfamilies based on the sequence similarities of their catalytic sites (Fig. 6) [30,31].

The major (and minor) subfamilies display relatively close levels of similarity, except for the more diffuse "Other" category, consisting of kinases that are not grouped into the major subfamilies based on evolutionary relations/similarity. The DNA- and protein-based similarities ($t = 4$ and $t = 20$, respectively) are roughly equal for most of the subfamilies, but the similarities based on simplified amino acid sequences ($t = 8$) are notably larger (this is particularly nicely illustrated by the otherwise diffuse group of all kinases). This supports the notion that due to the shared function of kinases (protein phosphorylation), amino acids tend to be replaced with sidechains of similar character (which is reflected in our sidechain categorization scheme, see Table 1).

Set similarities calculated individually with the six (non-weighted and weighted) indices are included in Supplementary Figures S3 and S4. We can observe that non-weighted metrics usually offer a greater level of distinction than weighted ones. The

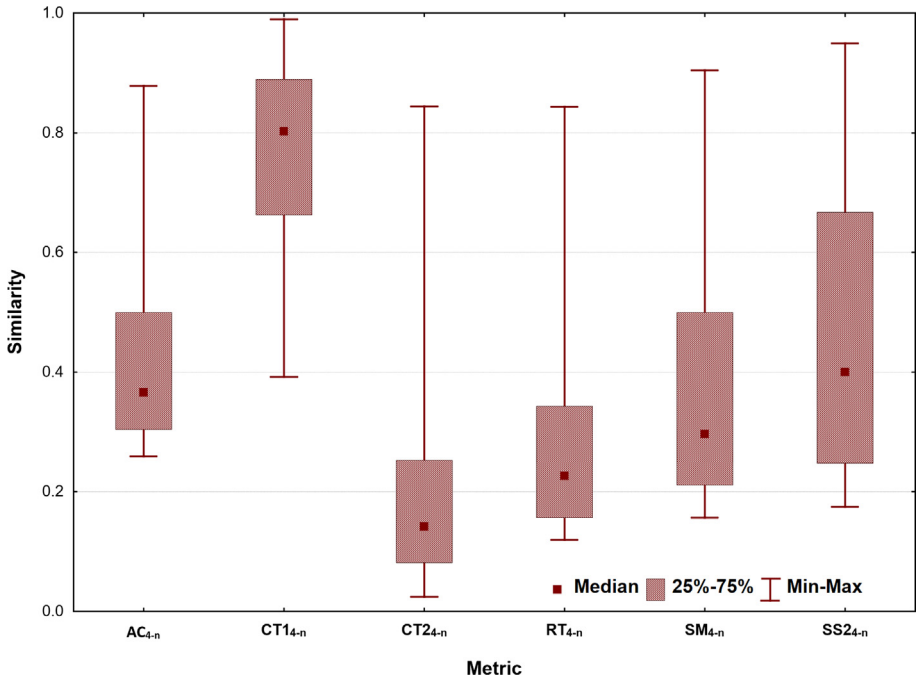


Fig. 2. Box plots with the median, interquartile range and minimum–maximum of individual quaternary (t = 4) indices.

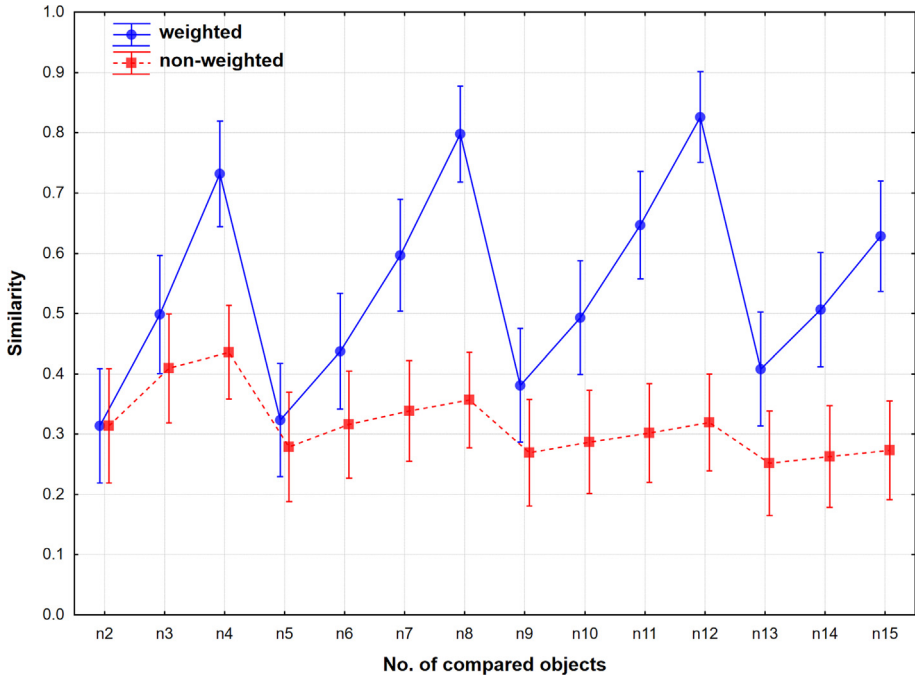


Fig. 3. The effect of weighting on the means of quaternary (t = 4) similarity coefficients as a function of the number of compared objects n (w: weighted, nw: non-weighted).

weighted formulas display a peculiar behavior, returning higher similarity values when more objects (n) are compared (this is also true for the non-weighted CT2 metric).

In the second case study, sequences of 120 human SH2 domains are compared. SH2 domains are ancient modular protein units that arose within multicellular life and are key regulators of cellular signal transduction [32]. They recognize phosphotyrosine-containing peptide motifs in a highly selective manner, depending on the contextual sequence [33]. SH2 domains have a relatively conserved fold and are found in proteins with diverse functions, including kinases, transcription factors, scaffold proteins, *etc.*, with many of them being involved in oncogenic processes [34]. Compared to kinases, we have more, relatively smaller groups of SH2-containing proteins, and interestingly, their functional grouping does not directly correspond to the phylogenetic distance of the SH2 domains themselves, as illustrated by their phylogenetic tree (Fig. 7).

Here, differences between the DNA/protein-based and simplified protein-based similarities are much smaller than in the case of kinases, suggesting a higher degree of freedom in terms of amino

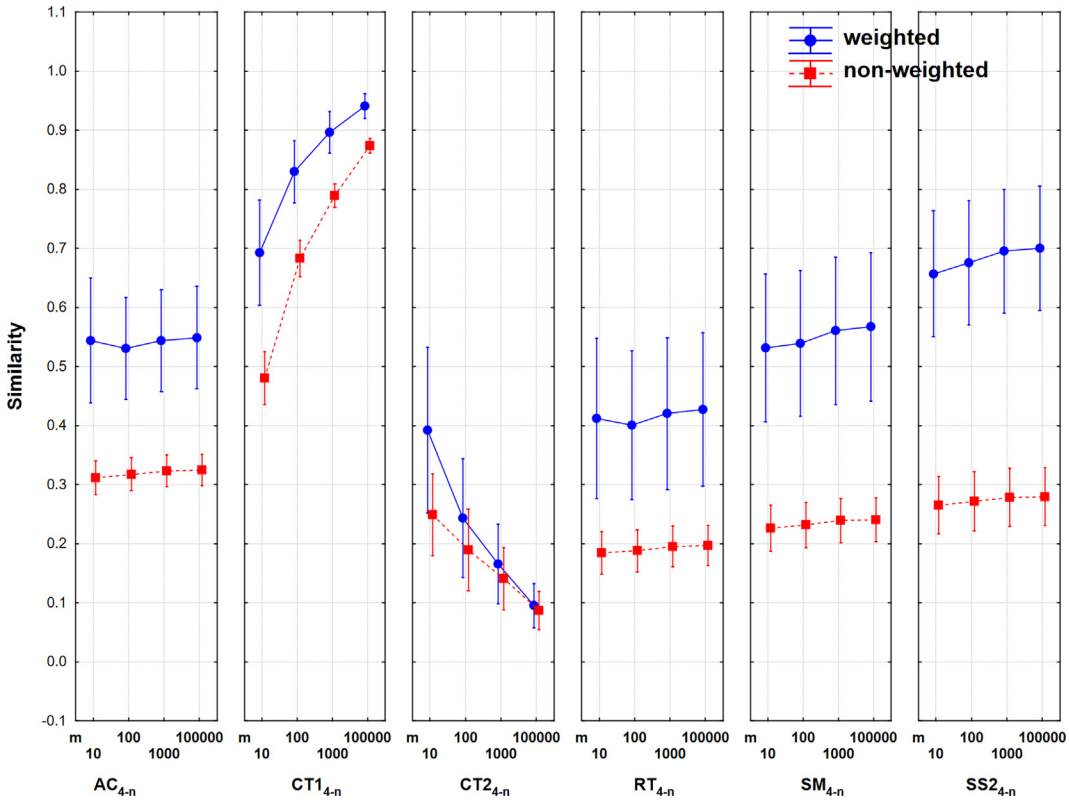


Fig. 4. Averages of extended similarity coefficients. Line plots correspond to weighted and non-weighted options. The lengths of the sequences (*m*, F5) are plotted on the upperx axis, whereas the lower x axis contains the individual extended indices.

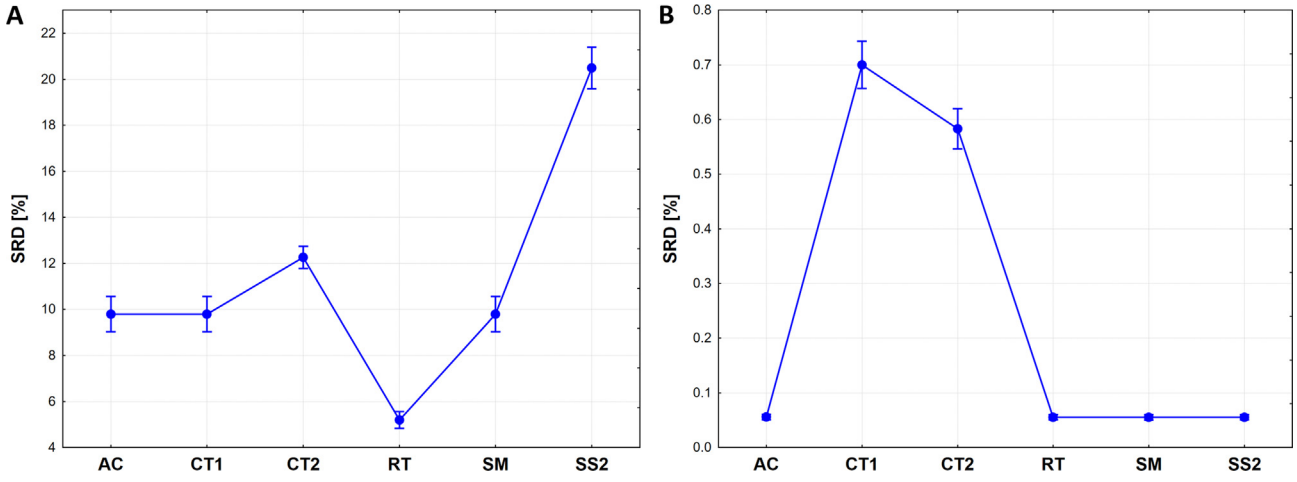


Fig. 5. Comparison of the individual non-weighted (A) and weighted (B) similarity coefficients with the ANOVA analysis of their normalized SRD values (0–100). The abbreviations can be found in Appendix 1.

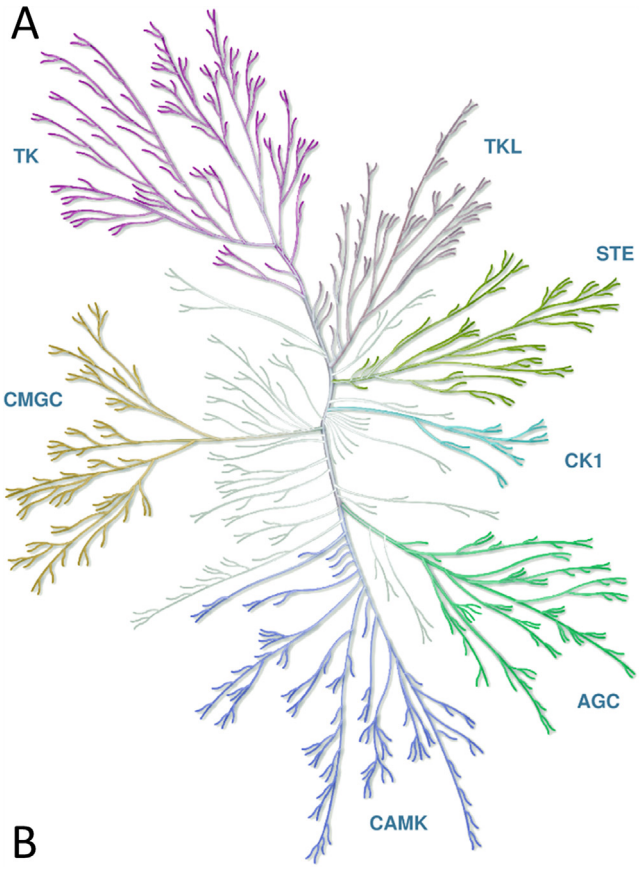
acid substitutions. This can be explained by individual SH2 domains selectively recognizing phosphotyrosine-containing peptide segments with diverse contextual sequences, requiring more specific (and equally diverse) binding motifs on the SH2 domains themselves. Also, in this case, there is a greater level of distinction between the set similarities of smaller and more compact groups (*e.g.* cytoskeletal regulators) and larger, more diffuse groups (*e.g.* small GTPase signaling proteins).

Set similarities calculated individually with the six (non-weighted and weighted) indices are included in Supplementary Figures S5 and S6. The non-weighted results are mostly in agreement with Fig. 7B, although the CT1 and CT2 metrics seem to offer a lower level of distinction (operating in a narrower range). A peculiar result in the case of the weighted metrics is that the group of all SH2 domains was assessed to be more similar than any of the subfamilies, but only based on the DNA sequences. Together with the results on the kinases, this suggests the wider applicability of the non-weighted formulas.

Finally, the last case study involves the sequence similarity calculations of a large family of cytochrome P450 (CYP) enzymes. CYP enzymes are heme-thiolate proteins that are found in virtually all organisms [35]. Commonly, they use electrons from NAD(P)H to catalyze the activation of molecular oxygen, but their reactions can be surprisingly diverse. They are involved in vital processes

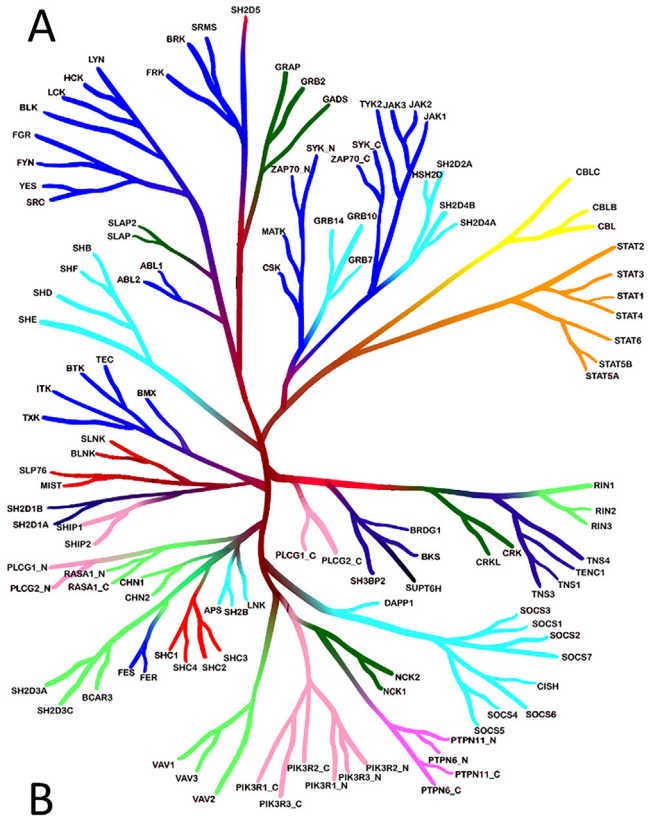
Table 1
Re-coding scheme for producing the simplified protein sequences based on the chemical/pharmacophoric character of amino acid sidechains.

Residue	Residue group
Glycine (**G**)	Glycine (**G**)
Alanine (**A**), valine (**V**), leucine (**L**), isoleucine (**I**), methionine (**M**)	Hydrophobic (**H**)
Phenylalanine (**F**), tyrosine (**Y**), tryptophan (**W**), histidine (**H**)	Aromatic (**A**)
Aspartate (**D**), glutamate (**E**)	Negative (**N**)
Lysine (**K**), arginine (**R**)	Positive (**P**)
Asparagine (**N**), glutamine (**Q**)	Amide (**D**)
Serine (**S**), threonine (**T**), cysteine (**C**)	OH/SH-based H-bond donor/acceptor (**B**)
Proline (**P**)	Proline (**R**)



	n (DNA)	*n* (AA)	DNA	Simplified AA	AA
AGC	61	63	0.531	0.646	0.504
CAMK	77	82	0.549	0.618	0.506
CK1	11	12	0.546	0.614	0.544
CMGC	57	61	0.416	0.573	0.441
RGC	5	5	0.493	0.550	0.498
STE	46	48	0.511	0.624	0.497
TK	71	93	0.471	0.605	0.516
TKL	39	43	0.425	0.524	0.424
Other	70	83	0.303	0.492	0.350
All kinases	437	491	0.313	0.568	0.415

Fig. 6. A) Phylogenetic tree of human protein kinases, with the seven larger subfamilies (illustration reproduced courtesy of Cell Signaling Technology, Inc.–www.cellsignal.com). B) *n*-ary similarities of the kinase subfamilies, calculated for DNA, amino acid (AA) and simplified amino acid sequences (average of all six, non-weighted similarity metrics), along with the numbers of compared objects (*n*).

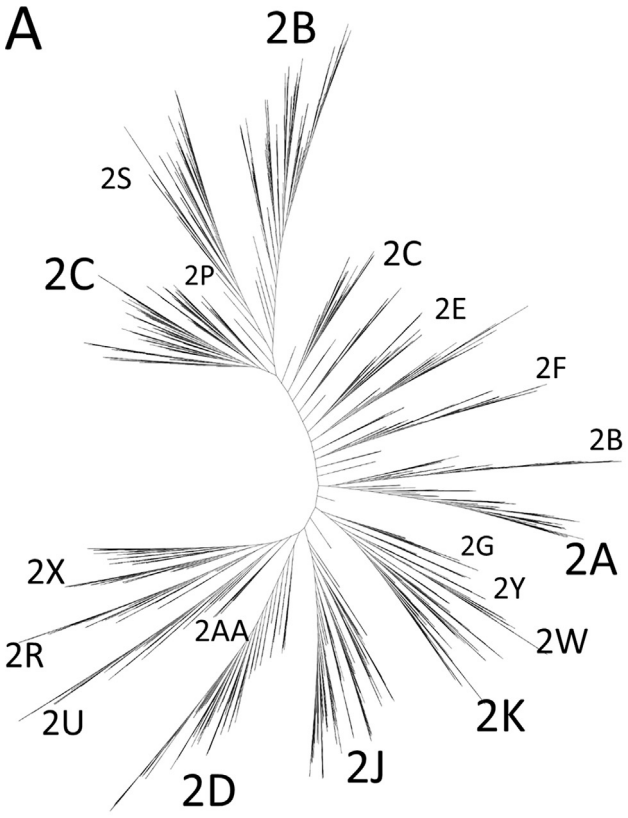


	n (DNA)	n (AA)	DNA	Simplified AA	AA
Adaptor	9	9	0.463	0.462	0.396
Cytoskeletal Regulation	4	4	0.690	0.759	0.666
Kinase	28	30	0.474	0.508	0.404
Phosphatase	4	4	0.502	0.604	0.526
Phospholipid 2nd Messenger Sign.	9	12	0.461	0.491	0.428
Scaffold	8	9	0.424	0.468	0.394
Signal Regulation	28	28	0.406	0.432	0.344
Small GTPase Signaling	10	13	0.392	0.391	0.332
Transcription	5	7	0.474	0.591	0.539
Ubiquitination	3	3	0.718	0.830	0.788
All SH2 domains	109	120	0.400	0.445	0.329

Fig. 7. A) Phylogenetic tree of human SH2 domains (adapted from the work of Liu *et al.* [32] with permission from Elsevier). While smaller protein groups generally comprise compact clusters (*e.g.* transcription factors, orange), larger groups are more diffuse (*e.g.* kinases, blue). B) *n*-ary similarities of the SH2 subfamilies (with matching colors to panel A), calculated for DNA, amino acid (AA) and simplified amino acid sequences (average of all six, non-weighted similarity metrics), along with the numbers of compared objects (*n*). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

such as chemical defense in plants or degradation of xenobiotics in animals. The latter signifies their main importance in medicinal chemistry and drug design: hepatic CYP enzymes are the main drivers of drug metabolism [36]. The number of currently known CYP enzymes is over 40 000 [37], their classification was introduced and refined by a nomenclature committee [38,39] as follows: the root symbol CYP is followed by a number for families (groups of proteins with more than 40% amino-acid sequence identity, currently there are more than 300), a letter for subfamilies (with greater than 55% identity) and a number for the protein, for example CYP2C9. Here, we quantify the similarities of the CYP2 family and its subfamilies (Fig. 8): in addition to being the

A



B

	n (DNA)	n (AA)	DNA	Simplified AA	AA
CYP2AA	26	26	0.723	0.827	0.732
CYP2A	255	259	0.643	0.820	0.779
CYP2B	245	252	0.569	0.800	0.743
CYP2C	656	665	0.605	0.692	0.672
CYP2D	304	313	0.625	0.803	0.733
CYP2E	175	178	0.654	0.772	0.684
CYP2F	95	96	0.590	0.806	0.706
CYP2G	22	22	0.722	0.758	0.757
CYP2J	414	414	0.576	0.598	0.508
CYP2K	226	227	0.597	0.733	0.662
CYP2P	20	20	0.693	0.749	0.581
CYP2R	108	108	0.765	0.827	0.751
CYP2S	60	60	0.790	0.819	0.699
CYP2U	112	112	0.650	0.765	0.713
CYP2W	85	85	0.635	0.740	0.644
CYP2X	73	73	0.754	0.809	0.724
CYP2Y	38	38	0.709	0.834	0.746
All CYP2	3261	3296	0.288	0.474	0.383
Human CYP2	54	59	0.527	0.355	0.301

Fig. 8. A) Phylogenetic tree of CYP2 enzymes. B) *n*-ary similarities of the CYP subfamilies, calculated for DNA, amino acid (AA) and simplified amino acid sequences (average of all six, non-weighted similarity metrics), along with the numbers of compared objects (*n*).

largest known family (3296 proteins), the CYP2 family contains many of the key human metabolic CYP enzymes (such as CYP2C9) with high relevance for the ADME prediction of drug candidates [40].

Here, since the large number of CYP enzymes are classified into subfamilies based on sequence homology, it is no surprise that we can observe higher levels of similarity in virtually all subfamilies. Differences between the original and simplified protein-based similarities are mostly moderate, but the DNA sequences are considerably less similar for some groups (*e.g.* 2B, 2D or 2F), suggesting that genetic codes of diverse species can yield CYP2 enzymes of similar protein sequence. The other face of the same coin is represented by the human CYP2 proteins: a group of 59 enzymes with representatives from 12 of the 18 CYP2 subfamilies in Fig. 8 (hence, constituting a more diffuse group than any individual subfamily). Here, the genetic code is more similar than the protein sequences, suggesting that within the same species, relatively fewer/smaller genetic mutations can yield a more diverse panel of CYP2 enzymes.

Set similarities calculated individually with the six (non-weighted and weighted) indices are included in Supplementary Figures S7 and S8. From the non-weighted metrics, CT2 seems to be an outlier based on its assessment of the "All CYP2" and "Human CYP2" sets, while the rest of the metrics agree with the results in Fig. 8B. At the generally much higher level of similarity of the CYP2 subfamilies, the weighted formulas offer little distinctive power, consistently returning values close to 1 (CT2 is interestingly an outlier again, this time by working in a wider operating range).

We briefly note that besides the $t = 8$ case presented here, other amino acid re-coding schemes can be introduced as well. A few of these possibilities (with varying values of t) are presented in Supplementary Figure S9, with the similarity values calculated for each protein class summarized in Figure S10. While there seem to be slightly higher similarity values for several protein classes at lower values of t, the trend is not consistent and the differences between the re-coding schemes are always much smaller than the differences across the protein classes. Nonetheless, we cannot recommend any further generalization than the one presented in the main text, since a smaller number of residue classes inevitably forces together amino acids of different character.

3.4. Diversity picking

Having a set of extended many-item (t,n) similarity metrics opens the doors to potentially many applications in bioinformatics. Here, we explore diversity picking as an illustrative example. In cheminformatics, diversity picking is a key concept for selecting a smaller number of molecules that represents the variability of a much larger chemical space (this was addressed in detail in our recent work, where we introduced (2,n) similarity metrics [20]). Analogously, selecting diverse macromolecular sequences can be important in certain situations (*e.g.* the selectivity of kinase inhibitors is often evaluated against a small, but diverse panel of kinases to cover all major branches of the phylogenetic tree [41,42]).

After implementing a diversity picker algorithm based on the (t,n) similarity metrics (as well as the MaxMin and MaxSum algorithms as two well-known diversity pickers for benchmarking [20]), we have tested it by selecting small sets of varying sizes (10, 20…, 100) from the CYP2 family of altogether 3296 enzymes (Fig. 9). Each set was selected three times to establish error ranges.

Most importantly, our *n*-ary diversity picker surpassed the benchmark methods by providing more diverse sets in every case, corroborating our earlier results for the (2,n) comparisons [20]. Interestingly, we also observe local maxima in most cases for n values that are multiples of t (*e.g.* 40 and 80 for $t = 8$). We can conclude that the new metrics present an ideal choice for diversity picking;

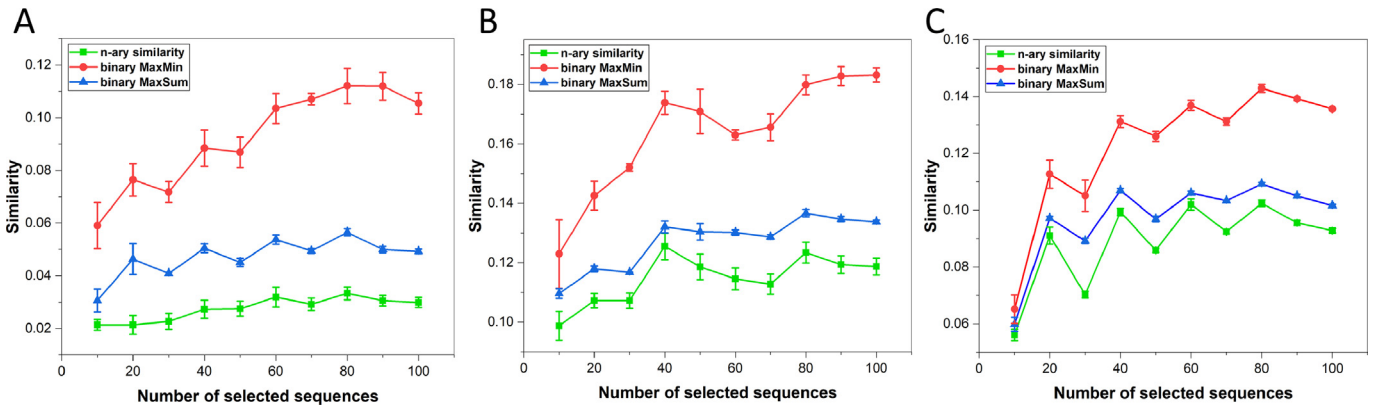


Fig. 9. Results of diversity picking from the CYP2 family of enzymes with the *n*-ary (green, bottom), and the binary MaxMin (red, top) and MaxSum (blue, middle) diversity pickers for DNA (A), simplified amino acid (B) and amino acid (C) sequences. Diversities are expressed as extended many-item similarities (average of the six metrics) for the complete sets of 10, 20, . . . 100 selected sequences. Data are shown as averages +/- standard deviations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the algorithm is available at https://github.com/ramirandaq/tn_Comparisons.

4. Conclusions

We have recently introduced a framework to extend the concept of similarity calculations from binary comparisons (similarity of two objects) to *n*-ary or multiple comparisons (similarity of sets of objects), chiefly for molecular fingerprint similarities in cheminformatics [19,20]. Expanding upon our results, here we have introduced extended many-item–or(*t*,*n*)–similarity metrics, moving from the domain of binary fingerprints (bit-strings containing two possible characters, such as 0 and 1) to character sequences (strings with an arbitrary number *t* of possible characters), such as DNA (*t* = 4) and protein sequences (*t* = 20). In our "democratic matching" approach (where matches are quantified in the same way, independently of the characters being matched), six existing similarity metrics can be extended for this purpose: SM (simple matching), RT (Rogers-Tanimoto), SS2 (Sokal-Sneath), CT1, CT2 (Consonni-Todeschini), and AC (Austin-Colwell).

In addition to a full theoretical description, we have provided a detailed study on the characteristics of the new similarity indices, including the typical ranges of similarity values returned, and how certain factors influence these results. For this purpose, we have applied analysis of variance (ANOVA).

Additionally, we have demonstrated the usage of the extended many-item similarity indices on three case studies of DNA, protein, and simplified protein (eight categories of similar amino acids) sequences of three existing protein families, and briefly explored diversity picking as one of the possible applications. The metrics present a new option for a quick calculation of the similarity of sets of sequences and have been demonstrated to provide good levels of distinction between protein groups with varying degrees of similarity. Nonetheless, we can narrow down our choice from the wide selection of new similarity metrics, accounting for some of the observations in this study, *e.g.* non-weighted formulas usually offer more distinctive power, and the CT2 metric often acts as an outlier. A multicriteria decision tool (sum of ranking differences) allowed to select the most advantageous similarity coefficents. Considering these results, along with the marked size dependency of the CT1 and CT2 metrics, we recommend the Rogers-Tanimoto (RT) coefficent as an optimum choice. The Python code for the extended many-item similarity indices is publicly available at: https://github.com/ramirandaq/tn_Comparisons

5. Methods

5.1. Statistical analysis

In section 3.2, the means of the extended similarity coefficents were analyzed using factorial analysis of variance (ANOVA) [43]. Factorial ANOVA was applied on the raw data, considering the following factors: F2–number (*n*) of objects (sequences) compare, 14 levels (*n* = 2, 3, . . .15); F3–weighted or unweighted version of the extended many-item similarity indices, two levels (w, nw); F4–the extended many item similarity coefficients themselves, six levels (AC, CT1, CT2, RT, SM, SS2); F5–sequence lengths, four levels (*m* = 10, 100, 1000, 100 000). Here, sequences were generated randomly, using four characters: A, C, G, T. The factors yield a total of 14*2*6*4 = 672 combinations and their effects were examined separately and in certain combinations (section 3.2).

Additionally, ANOVA was also performed on the normalized Sum of Ranking Differences (SRD) values obtained for the similarity measures (with the average of the six measures implemented as the reference method). Briefly, Sum of Ranking Differences is a robust multicriteria decision making tool [27] that is widely applied for method comparison in diverse fields [44–46]. SRD yields a normalized Manhattan distance (the SRD value) for each alternative method (here, similarity metric) as a measure of closeness to the reference method, which can be an independent gold standard or can be defined by a suitable data fusion method (in most cases, the average) from the compared methods. To adjust for the possibly different scales of values returned by the methods, rank transformation is applied as a data preprocessing step. SRD implements several validation steps, and is maintained for several platforms, including MS Excel (http://aki.ttk.hu/srd/), Python (https://github.com/davidbajusz/srdpy) and R Shiny (https://attilagere.shinyapps.io/srdonline/)

5.2. Collection of protein and DNA sequences

Pre-aligned kinase sequences were downloaded from http://kinase.com [28]. The SH2 domain sequence alignment was adapted from the work of Liu *et al.* [32]. Aligned sequences of the CYP2 family and its larger subfamilies (with ≥ 20 proteins) were downloaded from the Cytochrome P450 Engineering Database (current website: https://cyped.biocatnet.de/) [37]. The phylogenetic tree for the CYP2 family was drawn with Hypertree [47].

By default, our extended similarity metrics involve the detection of identical characters as similarity counters (with the exception of the "-" character for gaps). In contrast, during the comparison of protein sequences, amino acids of similar character (hydrophobic, aromatic, *etc.*) are also considered as a feature of similarity (as implemented in popular similarity scoring matrices, such as BLOSUM [15]). To reflect this, we introduce here a simple idea to "re-code" the protein sequences by grouping the natural amino acids into 8 groups, based on their sidechain character (Table 1). We have to note that this is a rather crude approach that does not account for the possible overlaps between these groups (for example, we classify histidine as aromatic, but it can assume a positively charged character upon protonation; similarly, cysteine is often deprotonated, *etc.*). The introduction of overlapping features would require serious modifications to the methodology and would constitute the basis of a separate work.

Finally, DNA sequences were downloaded from corresponding NCBI databases (chiefly, Entrez [48]) by either a look-up of the gene identifiers of the proteins (for the CYP dataset), or a BLAST search of the corresponding protein sequences (with the *tblastn* algorithm [49], for the kinase and SH2 datasets). In the latter case, at most one mismatch was allowed between the query protein sequence and the protein sequence resulting from the translation of the gene hit (to allow for transcript variants if the exact sequence was not found), retaining a sequence identity of more than 99% in every case (100% in most cases). The lookup was successful for 437 kinases (out of 491) and 109 SH2 domains (out of 120). The DNA sequences were aligned with Clustal Omega [50].

6. Availability of data and materials

Python code for calculating the extended many-item similarity metrics is freely available at: https://github.com/ramirandaq/tn_Comparisons.

Funding

- National Research, Development and Innovation Office of Hungary (OTKA), contracts K_20 134,260 (KH), K_20 135,150 (DB) and PD_20 134,416 (AR)
- University of Florida: startup grant: RAMQ
- Hungarian Academy of Sciences: János Bolyai Research Scholarship: DB
- Ministry for Innovation and Technology of Hungary: ÚNKP-20–5 New National Excellence Program: DB.

CRediT authorship contribution statement

Dávid Bajusz: Conceptualization, Software, Writing - review & editing, Data curation, Investigation, Methodology, Project administration. **Ramón Alain Miranda-Quintana:** Conceptualization, Software, Writing - review & editing, Formal analysis, Project administration, Resources. **Anita Rácz:** Conceptualization, Methodology, Writing - review & editing, Project administration. **Károly Héberger:** Conceptualization, Validation, Funding acquisition, Writing - review & editing, Project administration, Resources.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix 1:

Extended many-item similarity indices.

Metric	Label	Name	Equation
AC	$AC_{t\text{-}n\_w}$	extended many-itemAustin-Colwell	$AC_{t-n\_w} = \frac{2}{\pi}\arcsin\sqrt{\frac{\sum_s f_s(\Delta_{n(k)})C_{n(k)}}{\sum_s f_s(\Delta_{n(k)})C_{n(k)}+\sum_d f_d(\Delta_{n(k)})C_{n(k)}}}$
	$AC_{t\text{-}n\_nw}$		$AC_{t-n\_nw} = \frac{2}{\pi}\arcsin\sqrt{\frac{\sum_s f_s(\Delta_{n(k)})C_{n(k)}}{\sum_s C_{n(k)}+\sum_d C_{n(k)}}}$
CT1	$CT1_{t\text{-}n\_w}$	extended many-itemConsoni-Todeschini (1)	$CT1_{t-n\_w} = \frac{\ln\left(1+\sum_s f_s(\Delta_{n(k)})C_{n(k)}\right)}{\ln\left(1+\sum_s f_s(\Delta_{n(k)})C_{n(k)}+\sum_d f_d(\Delta_{n(k)})C_{n(k)}\right)}$
	$CT1_{t\text{-}n\_nw}$		$CT1_{t-n\_nw} = \frac{\ln\left(1+\sum_s f_s(\Delta_{n(k)})C_{n(k)}\right)}{\ln\left(1+\sum_s C_{n(k)}+\sum_d C_{n(k)}\right)}$
CT2	$CT2_{t\text{-}n\_w}$	extended many-itemConsoni-Todeschini (2)	$CT2_{t-n\_w} = \frac{\ln\left(1+\sum_s f_s(\Delta_{n(k)})C_{n(k)}+\sum_d f_d(\Delta_{n(k)})C_{n(k)}\right)-\ln\left(1+\sum_d f_d(\Delta_{n(k)})C_{n(k)}\right)}{\ln\left(1+\sum_s f_s(\Delta_{n(k)})C_{n(k)}+\sum_d f_d(\Delta_{n(k)})C_{n(k)}\right)}$
	$CT2_{t\text{-}n\_nw}$		$CT2_{t-n\_nw} = \frac{\ln\left(1+\sum_s f_s(\Delta_{n(k)})C_{n(k)}+\sum_d f_d(\Delta_{n(k)})C_{n(k)}\right)-\ln\left(1+\sum_d f_d(\Delta_{n(k)})C_{n(k)}\right)}{\ln\left(1+\sum_s C_{n(k)}+\sum_d C_{n(k)}\right)}$
RT	$RT_{t\text{-}n\_w}$	extended many-itemRogers-Tanimoto	$RT_{t-n\_w} = \frac{\sum_s f_s(\Delta_{n(k)})C_{n(k)}}{\sum_s f_s(\Delta_{n(k)})C_{n(k)}+2\sum_d f_d(\Delta_{n(k)})C_{n(k)}}$
	$RT_{t\text{-}n\_nw}$		$RT_{t-n\_nw} = \frac{\sum_s f_s(\Delta_{n(k)})C_{n(k)}}{\sum_s C_{n(k)}+2\sum_d C_{n(k)}}$
SM	$SM_{t\text{-}n\_w}$	extended many-itemSimple matching (Sokal-Michener)	$SM_{t-n\_w} = \frac{\sum_s f_s(\Delta_{n(k)})C_{n(k)}}{\sum_s f_s(\Delta_{n(k)})C_{n(k)}+\sum_d f_d(\Delta_{n(k)})C_{n(k)}}$
	$SM_{t\text{-}n\_nw}$		$SM_{t-n\_nw} = \frac{\sum_s f_s(\Delta_{n(k)})C_{n(k)}}{\sum_s C_{n(k)}+\sum_d C_{n(k)}}$
SS2	$SS2_{t\text{-}n\_w}$	extended many-itemSokal-Sneath (2)	$SS2_{t-n\_w} = \frac{2\sum_s f_s(\Delta_{n(k)})C_{n(k)}}{2\sum_s f_s(\Delta_{n(k)})C_{n(k)}+\sum_d f_d(\Delta_{n(k)})C_{n(k)}}$
	$SS2_{t\text{-}n\_nw}$		$SS2_{t-n\_nw} = \frac{2\sum_s f_s(\Delta_{n(k)})C_{n(k)}}{2\sum_s C_{n(k)}+\sum_d C_{n(k)}}$

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.06.021.

References

[1] Martin YC, Kofron JL, Traphagen LM. Do structurally similar molecules have similar biological activity?. J Med Chem 2002;45:4350–8. https://doi.org/10.1021/jm020155c.

[2] Bender A, Glen RC. Molecular similarity: a key technique in molecular informatics. Org Biomol Chem 2004;2:3204–18. https://doi.org/10.1039/B409813G.

[3] Bajusz D, Rácz A, Héberger K. Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching. In: Chackalamannil S, Rotella DP, Ward SE, editors. Compr. Med. Chem. III, Oxford: Elsevier; 2017, p. 329–78. https://doi.org/10.1016/B978-0-12-409547-2.12345-5.

[4] Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods 2015;71:58–63. https://doi.org/10.1016/j.ymeth.2014.08.005.

[5] Bender A, Jenkins JL, Scheiber J, Sukuru SCK, Glick M, Davies JW. How similar are similarity searching methods?: A principal component analysis of molecular descriptor space. J Chem Inf Model 2009;49:108–19. https://doi.org/10.1021/ci800249s.

[6] Todeschini R, Consonni V, Xiang H, Holliday J, Buscema M, Willett P. Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. J Chem Inf Model 2012;52:2884–901. https://doi.org/10.1021/ci300261r.

[7] Willett P. Similarity-based virtual screening using 2D fingerprints. Drug Discov Today 2006;11:1046–53. https://doi.org/10.1016/j.drudis.2006.10.005.

[8] Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?. J Cheminform 2015;7. https://doi.org/10.1186/s13321-015-0069-3.

[9] Flower DR. On the Properties of Bit String-Based Measures of Chemical Similarity. J Chem Inf Comput Sci 1998;38:379–86. https://doi.org/10.1021/ci970437z.

[10] Fligner MA, Verducci JS, Blower PE. A modification of the Jaccard-Tanimoto similarity index for diverse selection of chemical compounds using binary strings. Technometrics 2002;44:110–9. https://doi.org/10.1198/004017002317375064.

[11] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–53. https://doi.org/10.1016/0022-2836(70)90057-4.

[12] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

[13] Chowdhury B, Garai G. A bi-objective function optimization approach for multiple sequence alignment using genetic algorithm. Soft Comput 2020;24:15871–88. https://doi.org/10.1007/s00500-020-04917-5.

[14] Dayhoff MO, Schwartz R, Orcutt BC. A model of evolutionary change in proteins. Atlas Protein Seq. Struct., Nat. Biomed. Res. Found. 1978:345–58.

[15] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Biochemistry 1992;89:10915–9.

[16] Rácz A, Andrić F, Bajusz D, Héberger K. Binary similarity measures for fingerprint analysis of qualitative metabolomic profiles. Metabolomics 2018;14:29. https://doi.org/10.1007/s11306-018-1327-y.

[17] Rácz A, Bajusz D, Héberger K. Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints. J Cheminform 2018;10:48. https://doi.org/10.1186/s13321-018-0302-y.

[18] Miranda-Quintana RA, Bajusz D, Rácz A, Héberger K. Differential Consistency Analysis: Which Similarity Measures can be Applied in Drug Discovery?. Mol Inform 2021;40:2060017. https://doi.org/10.1002/minf.202060017.

[19] Miranda-Quintana RA, Bajusz D, Rácz A, Héberger K. Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 1: Theory and characteristics. J Cheminform 2021;13:32.. https://doi.org/10.1186/s13321-021-00505-3.

[20] Miranda-Quintana RA, Rácz A, Bajusz D, Héberger K. Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 2: speed, consistency, diversity selection. J Cheminform 2021;13:33.. https://doi.org/10.1186/s13321-021-00504-4.

[21] Al KA, Haranczyk M, Holliday J. Comparison of nonbinary similarity coefficients for similarity searching, clustering and compound selection. J Chem Inf Model 2009;49:1193–201. https://doi.org/10.1021/ci8004644.

[22] Avram SI, Crisan L, Bora A, Pacureanu LM, Avram S, Kurunczi L. Retrospective group fusion similarity search based on eROCE evaluation metric. Bioorganic Med Chem 2013;21:1268–78. https://doi.org/10.1016/j.bmc.2012.12.041.

[23] Boulif M, Atif K. A new branch-&-bound-enhanced genetic algorithm for the manufacturing cell formation problem. Comput Oper Res 2006;33:2219–45. https://doi.org/10.1016/j.cor.2005.02.005.

[24] Won Y, Lee KC. Group technology cell formation considering operation sequences and production volumes. Int J Prod Res 2001;39:2755–68. https://doi.org/10.1080/00207540010005060.

[25] Yazdani S, Shanbehzadeh J, Aminian E. Feature subset selection using constrained binary/integer biogeography-based optimization. ISA Trans 2013;52:383–90. https://doi.org/10.1016/j.isatra.2012.12.005.

[26] Farhadinia B, Effati S, Chiclana F. A family of similarity measures for q-rung orthopair fuzzy sets and their applications to multiple criteria decision making. Int J Intell Syst 2021:int.22351. https://doi.org/10.1002/int.22351.

[27] Héberger K. Sum of ranking differences compares methods or models fairly. TrAC Trends Anal Chem 2010;29:101–9. https://doi.org/10.1016/j.trac.2009.09.009.

[28] Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science 2002;298:1912–34. https://doi.org/10.1126/science.1075762.

[29] Fedorov O, Müller S, Knapp S. The (un)targeted cancer kinome. Nat Chem Biol 2010;6:166–9. https://doi.org/10.1038/nchembio.297.

[30] Bajusz D, Ferenczy GG, Keserű GM. Structure-Based Virtual Screening Approaches in Kinase-Directed Drug Discovery. Curr Top Med Chem 2017;17:2235–59. https://doi.org/10.2174/1568026617666170224121313.

[31] Roskoski R. Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes. Pharmacol Res 2016;103:26–48. https://doi.org/10.1016/j.phrs.2015.10.021.

[32] Liu BA, Jablonowski K, Raina M, Arcé M, Pawson T, Nash PD. The Human and Mouse Complement of SH2 Domain Proteins—Establishing the Boundaries of Phosphotyrosine Signaling. Mol Cell 2006;22:851–68. https://doi.org/10.1016/j.molcel.2006.06.001.

[33] Liu BA, Engelmann BW, Nash PD. The language of SH2 domain interactions defines phosphotyrosine-mediated signal transduction. FEBS Lett 2012. https://doi.org/10.1016/j.febslet.2012.04.054.

[34] de Araujo ED, Orlova A, Neubauer HA, Bajusz D, Seo H-S, Dhe-Paganon S, et al. Structural Implications of STAT3 and STAT5 SH2 Domain Mutations. Cancers (Basel) 2019;11:1757. https://doi.org/https://doi.org/10.3390/cancers11111757.

[35] Werck-Reichhart, Daniele Feyereisen R. Cytochromes P450: a success story. Genome Biol 2000;1:reviews3003.1.

[36] Zanger UM, Schwab M. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacol Ther 2013;138:103–41. https://doi.org/10.1016/j.pharmthera.2012.12.007.

[37] Fischer M, Knoll M, Sirim D, Wagner F, Funke S, Pleiss J. The Cytochrome P450 Engineering Database: a navigation and prediction tool for the cytochrome P450 protein family. Bioinformatics 2007;23:2015–7. https://doi.org/10.1093/bioinformatics/btm268.

[38] Nelson DR, Kamataki T, Waxman DJ, Guengerich FP, Estabrook RW, Feyereisen R, et al. The P450 Superfamily: Update on New Sequences, Gene Mapping, Accession Numbers, Early Trivial Names of Enzymes, and Nomenclature. DNA Cell Biol 1993;12:1–51. https://doi.org/10.1089/dna.1993.12.1.

[39] Nelson DR. The Cytochrome P450 Homepage. Hum Genomics 2009;4:59. https://doi.org/10.1186/1479-7364-4-1-59.

[40] Rácz A, Keserű GM. Large-scale evaluation of cytochrome P450 2C9 mediated drug interaction potential with machine learning-based consensus modeling. J Comput Aided Mol Des 2020;34:831–9. https://doi.org/10.1007/s10822-020-00308-y.

[41] Rachman M, Bajusz D, Hetényi A, Scarpino A, Merő B, Egyed A, et al. Discovery of a Novel Kinase Hinge Binder Fragment by Dynamic Undocking. RSC Med Chem 2020;11:552–8.

[42] Petri L, Egyed A, Bajusz D, Imre T, Hetényi A, Martinek T, et al. An electrophilic warhead library for mapping the reactivity and accessibility of tractable cysteines in protein kinases. Eur J Med Chem 2020;207:. https://doi.org/10.1016/j.ejmech.2020.112836112836.

[43] Lindman HR. Analysis of Variance in Experimental Design. New York: Springer-Verlag; 1991.

[44] Gere A, Rácz A, Bajusz D, Héberger K. Multicriteria decision making for evergreen problems in food science by sum of ranking differences. Food Chem 2020;128617. https://doi.org/10.1016/j.foodchem.2020.128617.

[45] Rácz A, Gere A, Bajusz D, Héberger K. Is soft independent modeling of class analogies a reasonable choice for supervised pattern recognition?. RSC Adv 2018;8:10–21. https://doi.org/10.1039/C7RA08901E.

[46] Bajusz D, Rácz A, Héberger K. Comparison of Data Fusion Methods as Consensus Scores for Ensemble Docking. Molecules 2019;24:2690. https://doi.org/10.3390/MOLECULES24152690.

[47] Bingham J, Sudarsanam S. Visualizing large hierarchical clusters in hyperbolic space. Bioinformatics 2000;16:660–1. https://doi.org/10.1093/bioinformatics/16.7.660.

[48] Maglott D. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 2004;33:D54–8. https://doi.org/10.1093/nar/gki031.

[49] Gertz EM, Yu Y-K, Agarwala R, Schäffer AA, Altschul SF. Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. BMC Biol 2006;4:41. https://doi.org/10.1186/1741-7007-4-41.

[50] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011;7:539. https://doi.org/10.1038/msb.2011.75.