Chemoinformatics profiling of ionic liquids – Uncovering structure-cytotoxicity relationships with network-like similarity graphs

Maykel Cruz-Monteagudo*†‡§, M. Natália D. S. Cordeiro*†

† REQUIMTE, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal.

‡ Centro de Estudios de Química Aplicada (CEQA), Facultad de Química y Farmacia, Universidad Central “Marta Abreu” de Las Villas, Santa Clara, 54830, Cuba.

§ Molecular Simulation and Drug Design Group, Centro de Bioactivos Químicos (CBQ), Universidad Central “Marta Abreu” de Las Villas, Santa Clara, 54830, Cuba.

Corresponding Author

* Departamento de Química e Bioquímica, FCUP, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal. Phone +351 220402502. Fax +351 220402659. E-mail: CRUZ-MONTEAGUDO, M.: maikelcm@uclv.edu.cu or gmailkelcm@yahoo.es; CORDEIRO, M.N.D.S.: ncordeir@fc.up.pt.

RUNNING TITLE: NSG-based mining of ionic liquids structure-cytotoxicity relationships
ABSTRACT: Ionic liquids (ILs) constitute one of the hottest areas in chemistry since they have become increasingly popular as reaction and extraction media. Their almost limitless structural possibilities, as opposed to limited structural variations within molecular solvents, make ILs “designer solvents”. They also have been widely promoted as “green solvents” although their claimed relative non-toxicity has been frequently questioned. The Thinking in Structure-Activity Relationships (T-SAR) approach has proved to be an efficient method to gather relevant toxicological information of analogue series of ILs. However, when datasets significantly grow in size and structural diversity the use of computational models becomes essential. We provided such a computational solution in a previous work by introducing a reliable, predictive, simple and chemically interpretable Classification and Regression Tree (CART) classifier enabling the prioritisation of ILs with a favourable cytotoxicity profile. Even so, an efficient and exhaustive mining of structure-activity relationships information goes beyond analogue compound series and the applicability domain of quantitative structure-activity relationships modelling. So, we decided to complement our previous findings based on the use of the CART classifier by applying the network-like similarity graphs (NSG) approach to the mining of relevant structure-cytotoxicity relationships (SCR) trends. Finally, the SCR information concurrently gathered by both, quantitative (CART classifier) and qualitative (NSG) approaches was used to design a focused combinatorial library enriched with potentially safe ILs.

KEYWORDS: Cytotoxicity, Focused combinatorial library design, Ionic liquids, Network Similarity Graphs, SAR mining, SAR Pathways, SAR Trees, Structure-Cytotoxicity Relationships.
INTRODUCTION

Ionic liquids (ILs) constitute one of the hottest areas in chemistry since they have become increasingly popular as reaction and extraction media (Ranke et al., 2007b). Their almost limitless structural possibilities, as opposed to limited structural variations within molecular solvents, make ILs “designer solvents” (Sheldon, 2005). They also have been widely promoted as “green solvents” (Rogers and Seddon, 2003; Wasserscheid and Welton, 2003; Welton, 1999). The rationale for calling ILs "green" generally is based on three arguments: i) an extremely low vapour pressure providing a reduced inhalative exposure of workers as compared to conventional molecular solvents; ii) non-flammability, which strongly reduces the risk of fast, exothermic oxidations in the case of an accident; and iii) they are claimed to be relatively nontoxic. While the first two arguments certainly contribute to their “greenness”, their claimed relative non toxicity has been frequently questioned (Ranke, et al., 2007b).

Such a concern on the toxicological profile of ILs comes from a growing effort of chemical industry, governments, academia, and nongovernmental organisations to address the principles of green chemistry (Anastas and Warner, 1998). Specifically, the green chemistry principle No. 3 states that “wherever practicable, synthetic methodologies should be designed to use and generate substances that possess little or no toxicity to human health and the environment”. So, if in the past, the synthesis of ILs was focused on obtaining unique physicochemical properties to achieve a specific behaviour considering the potential final industrial application; nowadays the main goal is to produce ILs with the desired biological features for the final application and also to facilitate the REACH registration processes (Hough et al., 2007; Pham et al., 2010).

It is now known from a growing number of studies on the hazard potential of ILs (Frade and Afonso, 2010; Pham, et al., 2010; Zhao et al., 2007), that its hazard potential and consequently
their greenness strongly depend on the structure. In this sense, Ranke and co-workers at the UFT Centre for Environmental Research and Sustainable Technology pioneered the field by introducing and successfully applying the *Thinking in Structure-Activity Relationships* (T-SAR) approach (Stolte *et al.*, 2007) to the molecular design of ILs (Arning *et al.*, 2008; Ranke, *et al.*, 2007b; Stock *et al.*, 2004; Stolte *et al.*, 2006; Stolte, *et al.*, 2007). Ranke *et al.* were also the first to propose the use of rat cell lines, namely leukaemia IPC-81 to evaluate the cytotoxicity of ILs and the use of this model as a preliminary screening (Ranke *et al.*, 2006; Ranke *et al.*, 2004).

T-SAR has proved to be an efficient approach to gather relevant toxicological information of ILs. Actually, the current structure-activity relationships (SAR) on the cytotoxicity of ILs come from the application of this approach to related series of ILs (Ranke, *et al.*, 2004; Ranke *et al.*, 2007a; Ranke, *et al.*, 2007b; Stolte, *et al.*, 2006; Stolte, *et al.*, 2007). In these studies the SAR information is extracted from rationally designed test kits of compounds and so, restricted to those (usually analogue) compound series.

For the study of the cation’s side chain length effect over cytotoxicity the test kit compounds were restricted to ILs with imidazolium head groups with varying alkyl side chain lengths and only five different anions (Cl\(^-\), Br\(^-\), BF\(_4^-\), PF\(_6^-\) and p-toluenesulphonate). No functionalization of the alkyl side chain was included in the design of this test kit (Ranke, *et al.*, 2004). Instead, for the study of the effect of the cation head group and side chain functionalization, the test kit was designed to include a set of ILs more diverse in terms of cation head groups (pyridinium, dimethylaminopyridinium, morpholinium, pyrrolidinium, piperidinium, ammonium, and phosphonium; in addition to imidazolium) and functionalizations of the alkyl side chain (alkoxy, nitrile, hydroxyl) but still restricted in terms of anions (Cl\(^-\), Br\(^-\), I\(^-\), and bistriflamide) (Stolte, *et al.*, 2007). The opposite case was the study of the anion effect were the test kit certainly included
a numerous and diverse set of 27 anions but it was restricted to imidazolium head groups with ethyl, butyl and hexyl side chains with no functionalization (Stolte, et al., 2006). As can be noted, the goal of the study determines and usually restricts the composition of the test kit to analogue series.

The most established approach to explore analogue series is the so-called R-group table. Since analogue series are relatively small, as well as the number of alternative substitution sites, SAR trends can often be deduced from R-group tables. However, this essentially subjective evaluation becomes quickly infeasible once datasets significantly grow in size, demanding the use of computational models (Stumpfe and Bajorath, 2012).

We provided such a computational solution in a previous work (Cruz-Monteagudo et al., 2013) by introducing a reliable, predictive, simple and chemically interpretable CART classifier enabling the prioritisation of ILs with a favourable cytotoxicity profile. This CART classifier faithfully reproduced the essentials of the current structure-cytotoxicity relationships of ILs, supporting its high biophysical relevance. So, the analysis of the structure of the corresponding decision tree allowed us to identify several moieties that could be regarded as “cytotoxicophores”. Even so, an efficient and exhaustive mining of SAR information goes beyond analogue compound series and the applicability domain of quantitative structure-activity relationships (QSAR) modelling (Stumpfe and Bajorath, 2012; Wawer et al., 2010).

To study not only large but structurally diverse datasets it is essential to focus on those compound subsets that might carry more and most interesting SAR information. To identify such regions, the search for SAR discontinuity becomes a major goal (Wassermann et al., 2010; Wawer, et al., 2010). Compound series of high SAR discontinuity are characterized by small structural modifications leading to substantial potency variations. Activity cliffs (Maggiora,
2006; Wassermann, et al., 2010), i.e. pairs or groups of structurally very similar compounds with dramatic potency differences, are the extreme form of SAR discontinuity. Such compound subsets are generally associated with high SAR information content and thus a major target for SAR investigation.

In this sense, the network-like similarity graphs (NSG) (Wawer et al., 2008) introduced by Bajorath and co-workers constitute one of the most efficient and practical data structures for the extraction of SAR information from large and diverse compound sets. These particular characteristics, makes the NSG approach a suitable platform for the integration and visualisation of the SAR information derived from the different analogue series designed for the application of the T-SAR approach to the study of the cytotoxicity of ILs. Or even to explore the network in search for new SAR trends guiding the design of new and safe ILs. On the other hand, combining qualitative approaches such as T-SAR or NSG with quantitative methods like the Classification and Regression Tree (CART) classifier derived in the previous work (Cruz-Monteagudo, et al., 2013) can increase the level of understanding of key SAR trends. Additionally, the presence of outliers or activity cliffs escaping to the applicability domain of QSAR methods but carrying critical SAR information provide ample reasons to continue qualitative considerations (Ranke, et al., 2007b; Stolte, et al., 2007).

Therefore, we have decided here to complement our previous findings (Cruz-Monteagudo, et al., 2013) by applying the NSG approach to the mining of SAR trends relevant for the cytotoxicity of ILs, namely, structure-cytotoxicity relationships (SCR) trends which can be used as useful tips guiding the molecular design of new and safe ILs. Finally, the SCR information concurrently gathered by both, quantitative (CART classifier) and qualitative (NSG) approaches was used to design a focused combinatorial library enriched with potentially safe ILs.
MATERIAL AND METHODS

Network-like Similarity Graphs. NSGs represent a compound dataset by showing all molecules and their similarity relationships. NSGs are formal graphs in which nodes correspond to molecules. Pair-wise similarity relationships are represented by edges that connect individual nodes. Only molecule pairs that exceed a predefined similarity threshold are connected by an edge. To visualise the potency distribution, nodes are colour coded by potency, applying a continuous spectrum from green (lowest) to red (highest potency). The compound discontinuity score reflects SAR characteristics of individual molecules and is represented by node scaling reflecting the potency deviation of a compound from its structural neighbours. Large nodes represent compounds inducing a high discontinuity and vice versa (Stumpfe and Bajorath, 2012; Wawer, et al., 2010). Thus, it identifies molecules that introduce SAR discontinuity and activity cliffs. In NSGs, combinations of large red and green nodes connected by an edge are activity cliff markers that can be easily identified. For this task we resort to SARANEA (Lounkine et al., 2009), a freely available program that implements a graphical user interface to NSGs and NSG-based data mining techniques.

In SARANEA, as a criterion for edges between nodes in NSGs, connected ILs needed to exceed a predefined Tanimoto similarity threshold value. In this work we applied two different thresholds. To search for highly discontinuous regions in the network containing “cytotoxicity cliffs” pairs encoding critical structure variations for cytotoxicity we used a Tanimoto similarity threshold of 0.95. To explore alternative routes to ILs of low cytotoxicity covering a wider range of structural moieties we selected a threshold of 0.70.

It is important to note that SARANEA also provides a set of functionalities to quantify global, local and compound-specific SAR features relying on numerical functions capturing pair-wise
compound similarity and potency comparisons including the continuity and discontinuity scores, as well as the Cliff, and SAR (SARI) Indexes (Lounkine, et al., 2009; Peltason and Bajorath, 2007; Wawer, et al., 2008). These quantitative features can aid the process of SAR mining. However, since the computation of this type of functions involves standardisation and normalisation process based on a normalisation set of molecules not including ionic liquids, we decided to prioritise the use of the different functionalities of SARANEA based on the graphic exploration of structure-similarity relationships.

The input file supplied to SARANEA for the NSG analysis consists of a list of 281 ILs including cytotoxic potency values and the corresponding structure description (see details below). The single input file used to be opened with SARANEA is a zip folder containing a subfolder including the above mentioned information, and a second subfolder including several files with the settings used for each applicable similarity threshold. This file is provided as a supplementary material in order to ensure reproducibility of results reported in this work or to facilitate further NSG analyses.

**SAR Data Collection and Structure Codification of ILs.** The NSG-based SCR mapping was conducted over a curated set of 281 ILs and related salts reporting the values of effective concentration required for 50% cytotoxicity (EC$_{50}$) expressed in micromolar units towards the rat leukaemia cell line IPC-81 (Cruz-Monteagudo, et al., 2013), extracted from the UFT/Merck Ionic Liquids Biological Effects Database.

Like most computational medicinal chemistry and QSAR tools, SARANEA was originally designed to deal with small-to medium size organic molecules represented as fully connected molecular graphs. However, ILs are represented as disconnected molecular graphs (the respective cation and anion species) and consequently, this program cannot be directly applied to
datasets of ILs without applying some adaptations to the design of the molecular fingerprints used by the program as input.

Since SARANEA accepts customized molecular fingerprint representations as input, the publicly available version of Molecular ACCess System (MACCS) structural keys (Durant et al., 2002) implemented in the CDK software (Steinbeck et al., 2003; Steinbeck et al., 2006) was modified to codify the molecular structure of ILs. MACCS structural keys are substructure-based fingerprints representing a dictionary of predefined structural fragments of fixed format and length. Each bit position of MACCS fingerprints is associated with an individual fragment and accounts for its presence (the bit is set to 1) or absence (the bit is set to 0) in a test compound. The publicly available version of MACCS keys consists of 166 structural fragments (Durant, et al., 2002). These features of MACCS keys allow their customisation for ILs based on a simple modification. This modification consists on the computation of the original 166 bits MACCS key for each IL constituent species, the differentiation of anion’s and cation’s MACCS keys by summing to each bit position on the MACCS key of one of the constituent ionic species a constant value equal to the length of the bits string (166 in this case), and further concatenation of both (cation + anion) bits strings. The result is a concatenated fingerprint of 332 bits codifying the molecular structure of ILs. The first 166 bits account for the presence or absence of one of the 166 predefined structural fragments on the cationic species while the subsequent 166 bits account for the same information for the anionic species, or vice versa. This process is graphically illustrated in Figure 1.

<Insert Figure 1 about here>

SARANEA also provides immediate interactive access to the molecular structures represented by nodes in the graphs (Lounkine, et al., 2009). Unfortunately, this feature only works for
canonical SMILES codes of fully connected molecular graphs, but not for those of disconnected molecular graphs like the case of ILs. In any case, this limitation can be partially overtaken with the aid of external molecular structure visualisation tools such as MarvinView (ChemAxon, 2012b) which do handle this kind of molecular graphs.

**Combinatorial Library Generation.** The assembling of the focused combinatorial library was based on three sets of 15 cationic head groups, 20 cationic side chains and 31 anions, which are depicted in Figure 2. A combinatorial library of 22508 unique cations was generated with the aid of the SmiLib software (Schüller *et al.*, 2003) by using as inputs the corresponding SMILES notation of the two sets of head groups and side chains. The SmiLib software generated an SDF file comprising 22508 unique cations. The SDF file comprising the 31 anions was generated by using the ChemAxon’s JChem for Excel software (ChemAxon, 2012a). Both, cation’s and anion’s SDF files were submitted to the ISIDA Fragmentor software (Varnek *et al.*, 2005; Varnek *et al.*, 2008) to compute the corresponding 371/2136 SMFs used to establish the structural reference space for the similarity assessment of the initial set of 281 ILs. Finally, the corresponding SVM output files provided by the ISIDA Fragmentor were converted to a fixed format/length vector file and concatenated into a unique vector file of size 2507 (including the corresponding vector files of 371/2136 anion/cation SMFs) for each one of the 697748 ILs of the combinatorial library. This process of format conversion and combinatorial concatenation was performed with a python implementation specifically developed in our group to process IL’s data generated with the ISIDA Fragmentor software, which is freely available for academic laboratories upon request from http://molweb.cbq.uclv.edu.cu/Downloads. The similarity assessment and the corresponding basic statistical analysis of the combinatorial library were conducted by using a MatLab implementation developed in our group. The corresponding
MatLab code is also freely available from the same address.

<Insert Figure 2 about here>

RESULTS AND DISCUSSION

Overview on the Cytotoxicity Profiling of ILs Based on the CART Classifier. In the first part of this work it was derived a reliable, predictive, simple and chemically interpretable CART classifier enabling the prioritisation of ILs with a favourable cytotoxicity profile (Cruz-Monteagudo, et al., 2013). Based on the classification performance on an external evaluation set we can expect that at about 75% of a set of new ILs will be correctly classified by the CART classifier. The other salient feature of this classifier is their simplicity and transparent chemical interpretation based on structural molecular fragments. The analysis of the structure of the corresponding decision tree allowed us to identify several moieties that can be regarded as “cytotoxicophores”, which were also used to establish a set of SAR trends specifically aimed to prioritise low cytotoxicity ILs, namely:

- A cationic linear alkyl side chain of length > 5
- Anions with a fluorocarbonated side chain of length ≥ 2
- Cationic aromatic N-heterocycles with linear alkyl side chain of length ≥ 4
- Six membered aromatic rings with a methyl substituent.

Only one moiety was found to reduce the cytotoxicity, i.e.:

- Short alkyl side chains functionalized with polar nitrile groups on aliphatic cation head groups containing nitrogen atoms.

Visual Network-like Similarity Graph Exploration. The first analysis conducted was directed to visually detect in the ILs NSG highly discontinuous regions (clusters of ILs) encoding minimal structural variations leading to significant cytotoxicity changes, with a special interest
on those containing cytotoxicity cliffs. ILs connected in this network possess a structural similarity of 95% or higher. So, any structural variation between two connected nodes (ILs) leading to a modification of the cytotoxic potency of one order of magnitude or higher comes from a cytotoxicity cliff pair. Consequently, such structural variations can be regarded as relevant for establishing the SCRs characterizing this diverse set of ILs. Figure 3 shows the NSG obtained at this similarity threshold.

<Insert Figure 3 about here>

This network is characterised by the coexistence of regions with continuous and discontinuous SARs. Regions of continuous SAR are characterized by clusters of small green nodes whereas discontinuous regions involve clusters composed of large green and red nodes (highlighted with a circle around). As deduced from the graph, this network contains similar subsets of ILs inducing SAR continuity and discontinuity, respectively. So, we can assert that at this similarity threshold, the ILs under study are characterised by a heterogeneous SCR. Similar and high values of the continuity (0.960) and discontinuity (0.963) scores estimated by SARANEA for this ILs network determine a SARI index of 0.498. Even considering the previous warning on the use of the quantitative features of SARANEA for ILs, the values estimated support the hypothesis of a heterogeneous relaxed SCR. So, the coexistence of continuous and discontinuous SARs confirms the validity of our choice of combining qualitative with quantitative approaches for increasing the level of understanding of key SCR trends.

Clusters of ILs combining large red and green nodes connected by an edge are cytotoxicity cliff markers that can be easily identified. These types of cluster were visually inspected in order to identify the key SCR trends dominating this ILs network.

The most significant cytotoxicity cliff pair in this network (see Figure 3) it is constituted by $N$-
Methyl-N,N-dioctyl-1-octanaminium bis(trifluoromethylsulfonyl)imide and N-Ethyl-N,N-dimethyl-1-butanaminium bis(trifluoromethylsulfonyl)imide (nodes 1 and 2 in the network) which is a clear example of the influence of the alkyl side chain length over the cytotoxicity of ILs. The only difference between these two ILs sharing a Tanimoto similarity value of 1.00 it is precisely a significantly longer alkyl side chains of IL 1 compared to IL 2. This slight structural modification induces an increase in cytotoxic potency of about three orders of magnitude. This can be noted by a clear difference in the spectrum of colours of the connected nodes (red to green) which reflects the significant cytotoxic influence of this kind of structural variations (Lounkine, et al., 2009; Stumpfe and Bajorath, 2012; Wawer and Bajorath, 2011; Wawer, et al., 2010; Wawer, et al., 2008).

The cluster including the ILs represented by nodes 3, 4 and 5 clearly suggest the effect of highly fluorinated anions over cytotoxicity. A quite explicit correlation between the degree of fluorination of the anion and cytotoxicity it is observed, as reported in previous studies (Stolte, et al., 2006). This kind of structural variation induces an increase in cytotoxic potency up to two orders of magnitude. The lower influence on the anion (increase the cytotoxicity up to two orders of magnitude) compared with the side chain length effect (about three orders of magnitude) has been also confirmed in previous studies and is also suggested by a smaller difference in the spectrum of colours of the this nodes (red to orange to green).

The last cluster analysed (nodes 6, 7, 8 and 9) suggest a weak influence of the cation head group over cytotoxicity. Note that the difference in the spectrum of colours in this cluster is quite low (the colour of the connected nodes only differ on the tones of green) which suggest a significantly lower influence of structural variations in the cation head group. However, another previous finding can be confirmed in this cluster: the relatively higher cytotoxicity of aromatic
cation head groups (Ranke, et al., 2007a; Stolte, et al., 2007). Note that the cytotoxic potency of an IL containing an aromatic head group it is always higher than the structurally closer aliphatic analogue, increasing the cytotoxic potency in no more than one order of magnitude.

The previous analyses demonstrates that at this level of inspection the major SCR trends identified confirms the previous SAR trends dominating the SCR of ILs (Ranke, et al., 2004; Ranke, et al., 2007a; Ranke, et al., 2007b; Stolte, et al., 2006; Stolte, et al., 2007) which supports the validity of the NSG approach for ILs SCR mining. However, NSGs are generally too complex for the extraction of detailed SAR rules and thus, no new SCR was uncovered just relying on this type of visual analysis. Consequently, for further analysis we resort to more advanced data structures implemented on SARANEA that mine the information contained in NSGs: SAR pathways and SAR trees.

**Structure-Cytotoxicity Relationships Pathways.** SAR pathways (Lounkine, et al., 2009; Wawer and Bajorath, 2009; Wawer et al., 2009) capture potency effects accompanying stepwise structural changes and consist on sequences of pairwise similar compounds (nodes). In essence, SAR pathways reflect SAR continuity. Activity cliffs are by design not covered by SAR pathways; however, all pathways leading to a particular activity cliff can be selected and analysed.

SAR pathways (or SCR pathways) were computed to assess small structural changes in ILs inducing gradual changes in cytotoxic potency. Specifically, those SCR pathways between the pair of ILs of lowest and highest cytotoxicity (*i.e.* a cytotoxicity cliff) that conform each highly discontinuous cluster in the NSG depicted in Figure 3, using a Tanimoto similarity threshold of 0.95. In this way should be possible to identify sequences of pair-wise similar ILs forming a cytotoxic potency gradient and so, determine a relevant and hopefully new SCR trend.
After exploring the different highly discontinuous clusters, it was observed that most of the SCR pathways characterising each cluster encoded the major and already known IL’s SCR trends while just one new SCR trend could be uncovered. Figure 4 shows the SCR pathway established between the most cytotoxic IL in the dataset (1-Methyl-3-tetradecyl-1H-imidazolium chloride) and one of the less cytotoxic IL in their cluster (1-(Cyanomethyl)-3-methyl-1H-imidazolium chloride). The cytotoxic potency gradient determined by this SCR pathway clearly shows the omnipresent alkyl side chain length effect and the good correlation between cation hydrophobicity and cytotoxicity. The significant influence of the introduction of polar nitrile functions on the alkyl side chain over the reduction of the cytotoxicity is also detected.

<Insert Figure 4 about here>

Figure 5 shows the potency gradient determined by a SCR pathway including a cliff pair of a significant difference of cytotoxic potency near to three orders of magnitude. The SCR trend observed here suggests that the anion species (at least sulphates and sulfonates) in ILs are affected too by hydrophobicity, the alkyl chain length, and polar functionalization effects in a similar mode that these do affect the cation cytotoxicity. This trend is more noticeable on imidazolium cations of short alkyl side chains (ethyl imidazoliums) although the migration of the spectrum of colours from orange (higher cytotoxicity) to increasingly greener tones (lower cytotoxicity) equally takes place on the butyl imidazolium cations. This trend also suggests a major role of the anion species for those ILs with short alkyl side chains. The anion effect of anionic moieties with lipophilic and hydrolysable structural elements comparable to the side chain effect established for cations was already proposed by Ranke et al. (Stolte, et al., 2006). However, this particular trend was not covered by the rules derived from our CART classifier which supports the convenience of the combination of QSAR and qualitative NSG approaches.
for an efficient SCR exploration.

<Insert Figure 5 about here>

The most significant new SAR finding discovered using the SAR Pathway functionality was related to the influence over cytotoxicity of methyl substituents on the pyridinium head group and of its relative position to alkyl side chains of different length. The pattern observed consists on a significantly higher cytotoxic potency of those pyridinium ILs with methyl substituents in position 4 compared to analogues with methyl substituent in a different position or unsubstituted. This pattern was consistently observed in three different clusters of high discontinuity including three different anion moieties, respectively (Cl\(^{-}\), BF\(_{4}^{-}\) and PF\(_{6}^{-}\)). The corresponding SCR pathways are shown in Figure 6.

<Insert Figure 6 about here>

The first cluster analysed (Figure 6A) comprises ten pyrimidinium ILs of different alkyl side chain length (ethyl, butyl, hexyl and octyl) with a common chloride anion. In this cluster the pattern found is observed for ILs with butyl alkyl side chains (nodes 7 to 9), but not for longer alkyl side chains (nodes 1 to 6). So, it seems to be a break point in the length of the alkyl side chain for the cytotoxic influence of methyl substituents. This break point interestingly matches the cut-off of 5 for the length of the alkyl side chain determining the cytotoxicity of ILs previously established using the CART classifier. A probable reason could be a dominant influence of long alkyl side chains over the hydrophobicity of the cation, compared to the influence of the methyl substituent. Thus, the hydrophobic influence of the methyl substituent over cytotoxicity does not seem to be significant beyond butyl or maybe pentyl alkyl side chains.

For these specific ILs, the cytotoxic potency of methyl substituents in position 4 trebles those of methyl substituents in position 3, implying a change of cytotoxic class from moderate (100 <
EC$_{50} \leq 5000$ µM) to low (EC$_{50} > 5000$ µM) cytotoxicity. Since 1-butylpyridinium chloride was not included on the dataset, the experimental EC$_{50}$ value (8020 µM) of the closer analogue (1-butylpyridinium bromide) was used to estimate the cytotoxic influence of methyl substituents in position 4 with respect to unsubstituted pyridinium cations. In this case, a similar cytotoxic influence of methyl substituents in position 4 is estimated. The pattern could be confirmed too for PF$_6^-$ and BF$_4^-$ anion moieties. See the details on the respective SCR pathways depicted in Figure 6B and 6C. The consistency of the pattern along different anion moieties supports its significance. It is also interesting to note that this pattern parallels a cytotoxicophore feature (six membered aromatic rings with a methyl substituent) previously identified via the CART classifier and codified by (C+):C-C*C*C-H, one of the five structural molecular fragments (SMFs) conforming the corresponding decision tree.

**Structure-Cytotoxicity Relationships Trees.** To integrate and compare the information provided by individual SAR pathways, all pathways or a subset of pathways that lead to or originate from a compound of interest can also be organised in a treelike structure termed an SAR tree (Lounkine, et al., 2009; Wawer and Bajorath, 2009). So, to explore alternative routes (sequences of gradual structure modifications) to ILs of low cytotoxicity covering a wider range of structural moieties the corresponding NSG was built at a similarity threshold of 0.70. The NSG obtained is shown in Figure 7. Note that, as expected, the use of less stringent similarity threshold produced a network conformed by a minimal number of three clusters which are expected to exhibit a wider structural variability. Like in the previous NSG (using a different similarity threshold of 0.95), high and similar continuity (0.960) and discontinuity (0.981) scores determining a SARI index of 0.489 similarly suggest a heterogeneous relaxed SCR for this NSG. Each cluster was analysed with the aid of the SAR tree utility in order to detect, if possible, some
new SCR trend implying a higher structural variability, not detected at the previous similarity threshold of 0.95.

<Insert Figure 7 about here>

To explore alternative pathways encoding different structure transitions from highly cytotoxic to safe ILs the respective SCR trees were constructed to originate from the IL of lowest cytotoxicity in each cluster. So, each SCR tree computed is rooted at a green node representing the IL of interest (the IL of lowest cytotoxic potency in the cluster), and the leaves of the tree correspond to the start (green nodes representing ILs of low cytotoxicity) and end points (red nodes representing highly cytotoxic ILs) of all pathways.

The SCR tree obtained for cluster C1 (see Figure 8) evidences that the main SCR encoded in this cluster accounts for the cytotoxic influence of a high degree of fluorination of the anion species across different linear, and cyclic (aromatic and aliphatic) cation moieties. The best example of this SCR trend it is represented by nodes 1 (1-butylpyridinium hexafluorophosphate) and 13 (1-butyl-4-methylpyridinium trifluorotris(pentafluoroethyl)phosphate), those being the start and end points of the main SCR pathway of the tree, respectively. This pair of ILs shares more than 95% of structural similarity but its respective cytotoxic potencies differ in almost four orders of magnitude. Highly fluorinated anion moieties have previously found to be a key structural feature leading to the anion cytotoxicity. So, we can say that the results of the SAR Tree analysis of this cluster essentially enclose and confirm the results of the application of the T-SAR approach to the study of the anion effect of ILs over cytotoxicity. \(^1, 15, 16, 18, 19\) The cytotoxic influence of methyl substituents (mainly on position 4) on butylpyridinium ILs previously found using SAR pathways can be confirmed with this tree. The lower similarity threshold applied in this case allowed placing PF\(_6^-\) and BF\(_4^-\) anion moieties (nodes 1 to 6) in a
common SCR pathway, unifying the SCR trend previously detected in two separate SCR pathways.

<Insert Figure 8 about here>

The SCR tree rooted at the less cytotoxic IL in the cluster C2 (and also of the whole dataset) encodes the main previous SCR findings pertaining to the effect of varying alkyl side chains (Ranke, et al., 2004), different head groups and functionalized side chains (Stolte, et al., 2007). This trend is observed in the SCR tree depicted in Figure 9 by a structure transition from cation species conformed by aromatic head groups with long alkyl side chains (highly cytotoxic ILs placed at the leaves) to aliphatic head groups with shorter and/or polar functionalised alkyl side chains (ILs of low cytotoxicity placed to the root). The effect of polar functionalizations on the reduction of the cytotoxicity was found to be more significant for terminal methoxy or hydroxyl (ILs 1 and 2) than ethoxy (ILs 3, 4 and 10) groups. The clearest example of this behaviour it is comprised by the cytotoxicity cliff pair conformed by ILs 1 and 10. This pair of ILs has a structural similarity of 88% while exhibiting significantly different degrees of cytotoxicity (> 40-fold).

<Insert Figure 9 about here>

Finally, as can be deduced from the SCR tree displayed in Figure 10, the cluster C3 condenses the main SCR trends currently established for ILs (Ranke, et al., 2004; Ranke, et al., 2007a; Ranke, et al., 2007b; Stolte, et al., 2007). The highly cytotoxic effect of long alkyl side chains it is constantly observed along the three branches of the tree, especially those ending at nodes 14 and 15. This observation confirms the omnipresence of the side chain length effect over the cytotoxic behaviour of the ILs evaluated up to date by Ranke et al. via the T-SAR approach. The branch ending at node 13 particularly shows the significant cytotoxic effect of highly fluorinated
and/or hydrophobic anions (Stolte, et al., 2006). The three branches meet at the less cytotoxic IL in C3 (1-(3-Methoxypropyl)-3-methyl-1H-imidazolium chloride), which is a good example of polar functionalization of alkyl side chains on the reduction of the cytotoxicity as well as the negligible effect of halide anions, which in turn was also observed in cluster C2.

<Insert Figure 10 about here>

Design and Assembling of a Focused Combinatorial Library Enriched with Potentially Safe ILs. Finally, the SCR trends identified by the interpretation of the rules derived from the CART classifier based on ISIDA SMFs and the application of the NSG approach were used to assemble a focused combinatorial library enriched with potentially safe ILs. This information was used to establish a general structural pattern of ILs with low cytotoxic potency. For this, the cation species used to assemble the combinatorial library should fulfil the following structural requirements:

- Aromatic head groups (i.e. imidazolium or pyridinium) with non-functionalized alkyl side chains up to length 4.
- Aliphatic head groups (i.e. pyrrolidinium, piperidinium, phosphonium or ammonium) with non-functionalized alkyl side chains up to length 5.
- Any head group with alkyl side chains up to length 5 with the following polar functionalizations: –O–, –OH, and –CN.
- Avoid substitutions in position 4 of the pyridinium head groups.

On the other hand, the anion species should be restricted to:

- Anions with no more than one –C(F)2– or –CF3 group. In these subset of anions the length of the alkyl side chain cannot be > 5 and can be functionalized (with –O–, –OH or –CN) or not.
The final result is a focused combinatorial library of 697748 ILs. Since it is not possible to actually check the quality of the library assembled we decided to estimate it based on the use of a combined scoring metric (Π) derived from the posterior probabilities of an IL to exhibit a favourable cytotoxic profile ($PP_{Class_1}$) assigned by a rigorously validated CART classifier and a fused similarity metric (ε) with demonstrated enrichment capability (Cruz-Monteagudo, et al., 2013). The result is a new scoring metric that quantifies the likelihood of a IL to exhibit a favourable cytotoxicity profile based on probabilistic ($PP_{Class_1}$) and structural similarity (ε) criteria also with demonstrated enrichment capability (Cruz-Monteagudo, et al., 2013). The values of Π near to 1 will be obtained for ILs with a high probability of exhibiting a favourable cytotoxicity profile determined by a high degree of structural similarity to the structural patterns determining a favourable cytotoxicity profile. In contrast, the value of Π will approach to 0 as the balance of $PP_{Class_1}$ and ε deteriorates, being 0 if at least one criteria takes a value of 0.

The analysis of the combinatorial library revealed that 75.57% of the ILs in the library exhibited values of Π ≥ 0.8, while just 17.72% exhibited values of Π < 0.5. The mean value of Π obtained for the library was of 0.67. However, several ILs in the combinatorial library exhibit values of Π = 0 and it is well known that such extreme values limits the usefulness of the arithmetic mean as estimator of central tendency of a population. Consequently, we resort to the median to estimate the overall quality of the library, which unlike the arithmetic mean it is not affected by extreme values. The median value of Π obtained for the library was of 0.82.

Considering both parameters (mean and median Π values) one can expect that an IL randomly selected from the library assembled will have a probability of exhibiting a favourable cytotoxicity profile between 67 and 82%. Details on the above analysis can be graphically assessed from the corresponding histogram of frequencies for the Π values of the 697748 ILs,
which is provided in Figure 11.

A file including the identification as well as the respective Π, ε, and $PP_{\text{Class}_1}$ values of the 697748 ILs comprised in the focused combinatorial library (decreasingly sorted by Π), is freely available for academic laboratories upon request from http://molweb.cbq.uclv.edu.cu/Downloads. Two SDF files are also provided, comprising the molecular structure and identification of the respective 22508 cations and 31 anions used to assemble the combinatorial library of ILs. The information comprised in these files represents a valuable decision making element to be considered by the different research groups involved on the development and REACH registration of ILs for various technical applications and committed with the principles of green chemistry.

CONCLUSIONS

The results described in this work show that only two specific SCR findings could be detected with the aid of the NSG approach that can be considered rather new. First, a significantly higher cytotoxic potency of pyridinium ILs with methyl substituents in position 4 compared to analogues with methyl substituent in a different position or unsubstituted was detected with the aid of the SAR pathway functionality and confirmed with SAR Trees. Afterward, a further SAR tree analysis revealed that the effect of polar functionalizations on the reduction of the cytotoxicity is more significant for terminal methoxy or hydroxyl than ethoxy groups. Yet, these findings were probably perceived although not explicitly reported in previous works of Prof. Ranke’s group (Stolte, et al., 2007). The SCR of ILs has been extensively studied by this group and consequently, it is really difficult to uncover completely new SCR findings. So, for this particular case we cannot go beyond than confirming and/or integrating their findings across the
different cationic and anionic moieties reported in previous and separate studies (Ranke, et al., 2004; Ranke, et al., 2007a; Ranke, et al., 2007b; Stolte, et al., 2006; Stolte, et al., 2007). Even so, the NSG approach and NSG-based data mining techniques implemented on SARANEA have proved to be an efficient tool to mine relevant SCR information guiding the design of potentially safe ILs. Of outstanding value is the suitability of NSGs to identify information rich regions by focusing on regions or clusters of high SCR discontinuity, specially, those containing cytotoxicity cliff pairs. The adaptation of the NSG approach proposed here to the particular and special case of disconnected molecular structures such as ILs also contributes to the integration of approaches like the traditional T-SAR analysis and the computational mining and visualisation of relevant SCRs of this interesting family of chemicals.

Finally, the SCR information gathered from both quantitative (CART classifier) (Cruz-Monteagudo, et al., 2013) and qualitative (NSG) approaches guided the design of a focused combinatorial library of about 700000 ILs with a likelihood to exhibit a favourable cytotoxicity profile of about 80%. Such a virtual library represents a valuable decision making element for the development of ILs for various technical applications that fulfil the principles of green chemistry.

SUPPLEMENTARY DATA

The input file supplied to SARANEA for the NSG analysis.

FUNDING

This work is supported by projects financed by FEDER through COMPETE – Programa Operacional Factores de Competitividade and by FCT – Fundação para a Ciência e a Tecnologia [PTDC/EQU-FTT/104195/2008, Pest-C/EQB/LA0006/2011, Pest-
C/EQB/LA0020/2011]. M.C-M acknowledges also FCT for the grant [SFRH/BPD/90673/2012] co-financed by the European Social Fund.

REFERENCES


Varnek, A., Fourches, D., Horvath, D., Klimchuk, O., Gaudin, C., Vayer, P., Solov’ev, V.,
198.


Wawer, M., Lounkine, E., Wassermann, A. M. and Bajorath, J. (2010). Data structures and
computational tools for the extraction of SAR information from large compound sets.


activity relationship anatomy by network-like similarity graphs and local structure-


FIGURE LEGEND

Figure 1. Graphic representation of the generation of a 166 bits MACCS key customized for the codification of the molecular structure of ILs.

Figure 2. Cationic head groups and side chains as well as anions used to generate a focused combinatorial library enriched with ILs with favourable cytotoxicity profile.

Figure 3. NSG constructed with the software SARANEA for a set of 281 ionic liquids using a Tanimoto similarity threshold of 0.95. This NSG consists of several clusters encoding different local SCRs. Nodes represent ILs connected by edges if they share a 2D similarity above a predefined threshold. The colour and size of a node reflect the cytotoxic potency and contribution to the local SCR discontinuity of the corresponding IL, respectively. The molecular structure and the respective EC$_{50}$ values for IPC981 leukaemia rat cell lines of the nine ionic liquids conforming the three clusters analysed are highlighted in three respective square boxes. These three clusters are highlighted in the NSG by red circles while the rest of clusters including ILs inducing a high discontinuity (connected large red and green nodes) are highlighted with black circles and further subjected to SAR pathway analysis.

Figure 4. SCR pathway computed for a representative cluster of ILs of high discontinuity describing the cation alkyl side chain length and polar functionalisation effect over cytotoxicity. Green nodes represent ILs of low cytotoxicity and yellow to red nodes ILs of moderate-to-very high cytotoxicity. Nodes are colour-coded and positioned in the graph according to IL cytotoxic potency. The molecular structure and cytotoxic potencies of ILs corresponding to numbered nodes are shown.

Figure 5. SCR pathway computed for a cluster of ILs of high discontinuity describing the anion alkyl side chain length and polar functionalization effect over cytotoxicity. A table is provided
that shows the molecular structure and cytotoxic potencies of ILs corresponding to numbered nodes. In this table, each cell containing the value of cytotoxic potency is colour coded similarly to the corresponding NSG and SCR pathway.

**Figure 6.** SCR pathways for three clusters of ILs with different anion moieties describing the influence over cytotoxicity of methyl substituents on the pyridinium head group and of its relative position to alkyl side chains of different length. A), B) and C) show an SCR pathway for pyridinium ILs with chloride, hexafluorophosphate and tetrafluoroborate anion moieties, respectively.

**Figure 7.** Network-like similarity graph constructed with the software SARANEA for a set of 281 ionic liquids using a Tanimoto similarity threshold of 0.70. The main three clusters of the network (denoted as \(C_1\), \(C_2\) and \(C_3\), respectively) that were subjected to SAR tree analysis are highlighted by dashed circles.

**Figure 8.** SCR tree generated for cluster \(C_1\) in the NSG obtained at a Tanimoto similarity threshold of 0.70. Green / red nodes represent ILs of low / very high cytotoxicity. The molecular structures and cytotoxic potencies of thirteen ILs represented by individual nodes in the SCR tree are shown.

**Figure 9.** SCR tree generated for cluster \(C_2\) in the NSG obtained at a Tanimoto similarity threshold of 0.70. The molecular structures and cytotoxic potencies of sixteen ILs represented by individual nodes in the SCR tree are shown.

**Figure 10.** SCR tree generated for cluster \(C_3\) in the NSG obtained at a Tanimoto similarity threshold of 0.70. The molecular structures and cytotoxic potencies of fifteen ILs represented by individual nodes in the SCR tree are shown.
Figure 11. Histogram of frequencies for the Π values of the ILs of the focused combinatorial library assembled.
15x1mm (600 x 600 DPI)
104x62mm (300 x 300 DPI)
167x156mm (300 x 300 DPI)
69x56mm (300 x 300 DPI)