Contents lists available at ScienceDirect

Journal of Hazardous Materials

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

Using cheminformatics to find simulants for chemical warfare agents

J. Lavoie, Sree Srinivasan, R. Nagarajan*

Molecular Sciences and Engineering Team, U.S. Army Natick Soldier Research, Development & Engineering Center, 15 Kansas Street, Natick, MA 01760, USA

ARTICLE INFO

Article history:
Received 28 February 2011
Received in revised form 16 July 2011
Accepted 21 July 2011
Available online 5 August 2011

Keywords:
Chemical warfare agents
Molecular descriptors
Simulants
Similarity search
Tanimoto coefficient
Euclidean distance

ABSTRACT

Direct experimentation with chemical warfare agents (CWA) to study important problems such as their permeation across protective barrier materials, decontamination of equipment and facilities, or the environmental transport and fate of CWAs is not feasible because of the obvious toxicity of the CWAs and associated restrictions on their laboratory use. The common practice is to use "simulants," namely, analogous chemicals that closely resemble the CWAs but are less toxic, with the expectation that the results attained for simulants can be correlated to how the CWAs would perform. Simulants have been traditionally chosen by experts, by means of intuition, using similarity in one or more physical properties (such as vapor pressure or aqueous solubility) or in the molecular structural features (such as functional groups) between the stimulant and the CWA. This work is designed to automate the simulant identification process backed by quantitative metrics, by means of chemical similarity search software routinely used in pharmaceutical drug discovery. The question addressed here is: By the metrics of such software, how similar are traditional simulants to CWAs? That is, what is the numerical "distance" between each CWA and its customary simulants in the quantitative space of molecular descriptors? The answers show promise for finding close but less toxic simulants for the ever-increasing numbers of CWAs objectively and fast.

Published by Elsevier B.V.

1. Introduction

Civilized nations have foresworn not to make or deploy chemical warfare agents but rogue states and terrorists have not, and CWA stockpiles presumably exist. Hence warfighters and civilians need to be prepared for CWA attacks [1,2]. At the US Army NSRDEC and similar organizations an ongoing focus is on developing protective masks, barrier clothing, and decontamination techniques to avoid exposure to CWAs that are toxic even at low dosages [3]. Also of interest are CWA disposal and environmental issues [4]. In this high-stakes context, it is desirable, even critical, to develop a quantitative understanding of the gamut of CWA phenomena besides toxicology – termed here as "handling characteristics": permeation (through barriers and human or livestock skin) as well as reactivity, and sorption. Measuring the necessary data in laboratories, however, is not practical because CWAs are so harmful that their handling is subject to severe restrictions. Direct theoretical prediction of the phenomena is also not a panacea, due to the paucity of estimates for the parameters required by the theories [5]. While neither measurement nor theory is an easy option for mapping the behavior of the toxic chemical agents directly, data can be obtained for alternative, less toxic molecules that can act as "simulants," surrogates, or analogs of the agents. The expectation is that, if a simulant resembles an agent in molecular structure or physicochemical properties then the simulant's performance may offer guidance for handling the CWA. Of course, no one simulant will be congruent in all aspects with the agent it is supposed to mimic; if so, it will probably be equally toxic. Accordingly, for a given CWA one can expect to find different simulants depending on the handling-phenomenon of interest – each simulant comparable with the agent in different molecular features or properties that correlate well with the phenomenon; e.g., similarity in hydrophobicity or polarizability indicating similarity in sorption.

Chemical similarity search is routine in medicinal drug discovery by "virtual screening" to find synthesizable chemicals that may mimic the pharmacologic activity of a synthetic or natural "lead drug" or "known active" [6–12]. It is also current in toxicology to assess chemicals of unknown toxicity, by searching for analogs with known toxicity profiles [13]. It would be of much interest to apply such powerful computational chemistry and data mining techniques and find CWA simulants by searching through databases of agent and simulant molecular structures and physicochemical properties.

Abbreviations: BCUT, Burden Chemical Abstracts Service and the University of Texas; CAS, Chemical Abstracts Service; CWAs, Chemical warfare agents; ED, Euclidean distance; HTS, high throughput screening; MACCS, molecular access system; NSRDEC, Natick Soldier Research Development and Engineering Center; QSAR, quantitative structure activity relations; QSPR, quantitative structure property relations; TC, Tanimoto coefficient; WENDI, Web Engine for Nonobvious Drug Information.

* Corresponding author. Tel.: +1 508 233 6445; fax: +1 508 233 4469.
E-mail address: Ramanathan.Nagarajan@us.army.mil (R. Nagarajan).

0304-3894/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.jhazmat.2011.07.077

Table 1
Types of chemical warfare agents.

Type of agent	Examples (and their military symbols)	Persistence	Mode of exposure
Asphyxiates	Chlorine (CL), Phosgene (CG)	Low	Inhalation
Blister Agents	Distilled mustard (HD)	High	Various
Blood Agents	Hydrogen Cyanide (AC)	Low	Inhalation
Nerve Agents, G	Tabun (GA), Sarin (GB), Soman (GD)	Low	Inhalation
Nerve Agents, V	VX	High	Skin contact

The CWA similarity search is akin to a subset of drug design – one that addresses bio-*availability* (a measure of intestinal or cutaneous absorption of the drug), which is continuously correlated with physical properties – than the hit-or-miss search for the more elusive bio-*activity*, which requires precise lock-and-key fits between receptors and ligands. Bioactivity is manifested when a chemical has the right three dimensional structure, orientation, and energetics to interact effectively with specific receptors in a living organism. Bioavailability dictates whether the chemical will reach the receptors or will fail to do so because of an inability to permeate or penetrate the various barrier membranes, storage in inert tissues, degradation by metabolic or other processes, or excretion [14]. Chemical warfare agents – which can be viewed as lethal drugs – must have not only bioactivity or lethality but also bioavailability or deliverability. While agent toxicity is a result of bioactivity, its bioavailability is the predominant concern in the development of protective barrier materials or in determining the environmental transport and fate of CWAs. The fundamental physicochemical processes that determine bioavailability – sorption, dissolution, or partition at interfaces in series with diffusion, apply also to CWA permeation through barriers or skin. Arguably, the ideal simulant will be a chemical that has all the bioavailability of the agent but none of the bioactivity.

Once the simulants have been identified, the handling behavior of CWAs can be addressed by computational methods of a different kind – based in experimental data on simulants and reference materials, and offering predictions via QSAR or QSPR statistical correlations or pattern recognition algorithms, instead of theory. This paper reports results from an exploration of the use of cheminformatics – fast, objective, and computational chemical-similarity searches – to find CWA simulants.

2. Prior art on CWA simulants

Any household or industrial chemical, if improperly used, can be harmful. Nevertheless, certain especially virulent and easily deployable chemicals have attained dubious distinction as CWAs (Table 1).

Employing simulants as a means of understanding CWAs is routine, as described in a recent review paper [4] which lists several classic CWA simulants. Previous simulant selections, however, appear to have been manual, i.e., not systematic computerized searches through chemical databases. Typically, the simulants seem to have been chosen by experienced scientists intuitively comparing a dozen or so molecules in terms of functional groups and a few physical properties. Of course, such commentary is not meant to disparage the significant progress that has been made using the simple methods. Our aim instead is to bring out the opportunity for greatly broadening the scope of CWA simulant searches by tapping into the advances in drug discovery and data mining.

3. Cheminformatics

There is an abundant and growing literature on “similarity searches” which can quickly rank a plethora of chemicals by their similarity to a specified query chemical, especially in medicinal chemistry and pharmacology, for drug design by “high throughput screening”. HTS techniques are quite sophisticated, conjugating chemistry, computer science, mathematics, and statistics [6–13] to mine “similarity spaces”.

Similarity searches start with a compilation of quantified structural features and physicochemical properties of query chemicals and prospective simulants. These “molecular descriptors” are preferably calculated as opposed to measured, in order to let the search encompass molecules that are as yet nonexistent; i.e., have been only proposed *in silico* but not yet synthesized *in actu*, or exist only in amounts too minuscule for some measurements. From this chemical database which, in drug design may include hundreds of descriptors each for thousands of molecules, the method then has to use statistical tools to weed out irrelevant or redundant descriptors and molecules [13]. Alternatively, the descriptors may be filtered through a weighting scheme [7]. Finally, the search method has to rank the molecules on their similarity or dissimilarity to the query molecule by comparing the descriptors – via graphical, statistical, pattern recognition or machine-learning means. The ranking usually involves numerical metrics of the “distance” or “association” between query and simulant.

The desired features of similarity search methodology and software are:

(1) Access to a large database of molecular descriptors for chemicals.
(2) Provision to augment the database with user calculations and data.
(3) Computational chemistry tools for calculating quantum chemical descriptors.
(4) Options to select molecules, descriptors and similarity metrics, in order to detect correlations among descriptors and compare the significance of the metrics.
(5) Capabilities for calculating the metrics of similarity between any two chemicals.

After a careful search, we settled on the commercial[1] Sarchitect® software, which, in combination with public or proprietary databases of chemicals, would meet most of these criteria.[2]

4. Scope

In addition to software and methodology, drug design also provides insights that can guide the work on CWA simulants. For instance, it has been recognized [6] that virtual screening is more successful for effects that are continuously variable with some or other physicochemical property – a change in the latter generating a non-abrupt, proportionate (but not necessarily linear) change in the former. Example: intestinal absorption or bioavailability correlating positively with hydrophobicity and negatively with molecular size. In contrast, virtual screening has struggled [6,15] in predicting effects (such as alignment-specific binding of

[1] Sarchitect Designer version 2.5, Strand Life Sciences Pvt. Ltd., Bangalore, India, 2008.

[2] The authors’ choice is not to be construed as an endorsement by the U.S. Army.

ligands to receptor sites) which are discontinuously or highly nonlinearly dependent on the causes. It is fortunate that, as noted in the Introduction, the CWA similarity search would not seek to mimic the query chemical's specific bio-activity (lethality) but instead aim at its general bio-availability (handling).

5. Molecular descriptors

It is understood that no molecule will be amenable to a single all-encompassing absolute definition, but any molecule can be characterized by a number of descriptors for each phenomenon of interest. When all these finite sets for the enormous variety of phenomena are put together, even after allowing for overlaps, the compilation of molecular descriptors can begin to approach an infinite set. Despite their large number, however, molecular descriptors lend themselves to compact classifications [16]. For instance, in terms of the level of abstraction, descriptors fall in three classes:

- Macroscopic properties such as molecular weight or the octanol/water partition coefficient, refractive index, molar refractivity, parachor, density, solubility, partition coefficient, dipole moment, chemical shift, chromatographic retention time, spectroscopic signal (or even complete spectra for each molecule), rate constant, equilibrium constant, vapor pressure, boiling/freezing point, and acid dissociation constant.
- Derived properties such as the surface distribution of electrostatic potential, the empirical absorbability index (a group-contribution index of carbon adsorption from aqueous solutions), or the various theoretical charge descriptors (calculated using quantum chemistry methods).
- More abstract measures such as BCUT, topological indices, substructural fingerprints and feature counts [7–13].

Another pertinent classification is based on molecular dimensionality:

- One-dimensional or 1D descriptors depend only on the formula (e.g., molecular weight).
- 2D descriptors depend on topology – the connectivity of bonds between the atoms (e.g., the Balaban connectivity index)
- 3D descriptors depend on stereochemistry and geometry (e.g., dipole moment).
- 4D descriptors take into account the conformational variability (e.g., the global flexibility index).

The cited literature on molecular descriptors is highly evolved and amply details descriptor types, invariance, and degeneracy, as well as the attributes of similarity metrics. Hence, these details need not be repeated here. It is instructive, however, to list a few considerations that pertain to CWA chemical similarity searches:

1. Descriptors need not be complex to be useful; e.g., simple heavy-atom counts can be telling [6].
2. As noted previously [6], descriptors are "good" if small tweaks of the descriptors cause small changes in the targeted behavior, and "poor" if the elicited responses are large or abrupt.
3. Along with the numerical values for the descriptors, it is desirable to include error bars, indicating the experimental error for measured properties and round-off error for calculated descriptors.
4. The ideal set of descriptors would be minimalist (offering sufficient representation with the fewest descriptors), fundamental instead of derivative (with little correlation among descriptors), and entirely theoretical or computational in order to include virtual or untested molecules.
5. Computed descriptors – purely theoretical constructs such as topographic indices or physical properties calculated using theoretical or empirical equations – should preferably be calculable instantaneously during the similarity search. That way, the entire database of chemicals can be probed in a single run. On the other hand, measured properties or computationally intensive descriptors calculated offline and stored – e.g., quantum chemical measures – would limit the search to those chemicals with such stored entries.
6. While a comprehensive database may be set up and continually updated with theoretical indices as well as measured properties, the similarity search per se need not use only that database or a single similarity metric. Instead, several separate forays may be made, into the large database or several smaller field-specific databases [8], each search restricted to a certain cluster of chemicals or class of descriptors and appropriate similarity metrics.

6. Method

The first phase in this ongoing work had the scope of answering a key question: By the metrics of computational chemical similarity search, how similar are traditional CWA simulants [3,4] to CWAs? That is, what is the "distance" between each CWA and its customary simulants in the quantitative space of molecular descriptors?

Based on information on a few CWAs and their traditional simulants (listed in Ref. [4]), a small database was constructed by entering the CAS Registry numbers of the CWAs as well as the known simulants in the CAS Scifinder®, entering the resulting structures in standard MOL notation into Sarchitect®, and calculating energy-minimized conformers from the structures using Sarchitect®.

Sarchitect® then calculates two types of descriptors for all the molecules in the database: "Digital" (topological and other) descriptors that can be stored as binary (0 or 1) strings, and "analog" physical properties and other (including topological) descriptors which are not confined to 0 or 1. Specifically, the binary strings are the 166 2D MACCS "finger-print" keys [17]; i.e., each molecule is characterized by a string of 166 zeroes or ones, each digit indicating simply the absence or presence of a particular topological or compositional feature, such as hetero-atoms in a four-member ring, various atomic elements, etc. Similarly, each molecule is characterized by about a thousand constitutional, topological, and conformational integer- or real-valued descriptors. The latter are pruned to remove any descriptor for which there is zero variance across all molecules. The list of all the descriptors used can be found in the supplementary data.

The next step is the determination of how similar (or dissimilar) the target chemicals in the database (simulants) are to each query chemical (CWA). The binary strings of digital descriptors are amenable to sorting by the "Tanimoto coefficient (TC)" which is a measure of association or similarity; i.e., TC = 1 if the target and query are identical, and TC = 0, if the target has nothing in common with the query. The integer- and real-valued sets of analog descriptors are assessed by the "Euclidean distance (ED)" which is a measure of dissimilarity; i.e., ED = 0 if the target and query are identical. It is pertinent to emphasize that, while Sarchitect® uses the MACCS *fingerprints* as binary records of the presence or absence of various features when computing Tanimoto coefficients, Sarchitect® also provides the option of including the integer number of counts of the same features as MACCS *descriptors* when computing Euclidean distances. The TC and ED calculations are outlined next. Ref. [7] provides a fuller description of the theory behind the TC and ED indices.*Tanimoto coefficient*: Consider two molecules A and B for each of which the string of 166 MACCS *fingerprints* can

Table 2
Tanimoto coefficients (TC) between chemical warfare agents and simulants (TC = 1.0 for perfect similarity).

Rank	Simulant	Compound	CAS Number	TC
0	GB	Sarin	107-44-8	1.000
1	DFP	Di-isopropyl fluorophosphate	55-91-4	0.875
2	DIMP	Di-isopropyl methyl phophonate	1445-75-6	0.750
3	DMMP	Dimethyl methyl phosphonate	756-79-6	0.667
4	TMP	Thimethyl phospate	512-56-1	0.640
5	DPCP	Diphenyl chloro phosphate	2524-64-3	0.500
6	TEP	Triethyl phosphate	78-40-0	0.455
7	DEEP	Diethyl ethyl phosphonate	78-38-6	0.429
8	DEHP	Diethyl ester phosphonic acid	762-04-9	0.382
9	Paraoxon	Diethyl 4-nitrophenyl phosphate	311-45-5	0.300
10	ECA	Ethyl chloro acetate	105-39-5	0.194
11	DEM	Diethyl malonate	105-53-3	0.189
12	DOP	Bis(2-ethylhexyl) phthalate	117-81-7	0.178
13	DPGME	Dipropylene glycol monomethyl ether	34590-94-8	0.146
14	Ethanol	Ethanol	64-17-5	0.103
15	BUSH	1-Butanethiol	109-79-5	0.030
0	GD	Soman	96-64-0	1.000
1	DFP	Di-isopropyl fluorophosphate	55-91-4	0.840
2	DIMP	Di-isopropyl methyl phophonate	1445-75-6	0.720
3	DMMP	Dimethyl methyl phosphonate	756-79-6	0.640
4	TMP	Thimethyl phospate	512-56-1	0.615
5	DPCP	Diphenyl chloro phosphate	2524-64-3	0.485
6	TEP	Triethyl phosphate	78-40-0	0.441
7	DEEP	Diethyl ethyl phosphonate	78-38-6	0.417
8	DEHP	Diethyl ester phosphonic acid	762-04-9	0.371
9	Paraoxon	Diethyl 4-nitrophenyl phosphate	311-45-5	0.294
10	ECA	Ethyl chloro acetate	105-39-5	0.189
11	DEM	Diethyl malonate	105-53-3	0.184
12	DOP	Bis(2-ethylhexyl) phthalate	117-81-7	0.174
13	DPGME	Dipropylene glycol monomethyl ether	34590-94-8	0.143
14	Ethanol	Ethanol	64-17-5	0.100
15	BUSH	1-Butanethiol	109-79-5	0.029
0	GA	Tabun	77-81-6	1.000
1	DEEP	Diethyl ethyl phosphonate	78-38-6	0.537
2	TEP	Triethyl phosphate	78-40-0	0.525
3	Paraoxon	Diethyl 4-nitrophenyl phosphate	311-45-5	0.472
4	DEHP	Diethyl ester phosphonic acid	762-04-9	0.463
5	DIMP	Di-isopropyl methyl phophonate	1445-75-6	0.447
6	DMMP	Dimethyl methyl phosphonate	756-79-6	0.432
7	TMP	Thimethyl phospate	512-56-1	0.421
8	DFP	Di-isopropyl fluorophosphate	55-91-4	0.381
9	ECA	Ethyl chloro acetate	105-39-5	0.333
10	DEM	Diethyl malonate	105-53-3	0.326
11	DOP	Bis(2-ethylhexyl) phthalate	117-81-7	0.294
12	DPGME	Dipropylene glycol monomethyl ether	34590-94-8	0.250
13	DPCP	Diphenyl chloro phosphate	2524-64-3	0.245
14	Ethanol	Ethanol	64-17-5	0.184
15	BUSH	1-Butanethiol	109-79-5	0.146
0	HD	Distilled mustard	505-60-2	1.000
1	CEES	2-Chloroethyl ethyl sulfide	693-07-2	0.647
2	CEMS	2-Chloroethyl methyl sulfide	542-81-4	0.529
3	CEPS	Chloroethyl phenyl sulfide	5535-49-9	0.421
4	DEA	Diethyl adipate	141-28-6	0.310
5	DEP	Diethyl pimelate	2050-20-6	0.310
6	DMA	Dimethyl adipate	627-93-0	0.207
7	DEM	Diethyl malonate	105-53-3	0.167
8	MS	Methyl salicylate	119-36-8	0.000
0	L	Lewisite	541-25-3	1.000
1	LO	Lewisite oxide	3088-37-7	0.417
2	PAO	Phenylarsine oxide	637-03-6	0.133
0	VX	O-ethyl *S*-[2-(diisopropylamino)ethyl] methylphosphonothiolate	50782-69-9	1.000
1	Amiton	O,O-diethyl-S-[2-(diethylamino)ethyl] phosphorothiolate	78-53-5	0.837
2	Malathion	S-[1,2-bis(ethoxycarbonyl) ethyl] O,O-dimethyl phosphorodithiolate	121-75-5	0.630
3	BisEHEHP	Bis(2-ethyl 1-hexyl) 2-ethyl 1-hexyl phosphonate	126-63-6	0.600
4	DEPPT	O,S-diethyl phenyl phosphonothiolate	57557-80-9	0.595
5	BisEHP	Bis(2-ethylhexyl) phosphonate	3658-48-8	0.565
6	Parathion	O,O-diethyl-O-p nitrophenyl thiophosphate	56-38-2	0.509
7	DEP	Diethyl pimelate	2050-20-6	0.391
8	DES	Diethyl sebacate	110-40-7	0.391
9	DEM	Diethyl malonate	105-53-3	0.326
10	DEPh	Diethyl phthalate	84-66-2	0.271

Table 3
Euclidean distances (ED) between chemical warfare agents and simulants (ED = 0 for perfect similarity).

Rank	Simulant	Compound	CAS Number	ED
0	GB	Sarin	107-44-8	0.000
1	DMMP	Dimethyl methyl phosphonate	756-79-6	0.193
2	DIMP	Di-isopropyl methyl phosphonate	1445-75-6	0.226
3	DEHP	Diethyl ester phosphonic acid	762-04-9	0.230
4	BUSH	1-Butanethiol	109-79-5	0.232
5	DEEP	Diethyl ethyl phosphonate	78-38-6	0.242
6	TMP	Trimethyl phospate	512-56-1	0.255
7	DFP	Di-isopropyl fluorophosphate	55-91-4	0.260
8	ECA	Ethyl chloro acetate	105-39-5	0.268
9	TEP	Triethyl phosphate	78-40-0	0.308
10	DPGME	Dipropylene glycol monomethyl ether	34590-94-8	0.320
11	Ethanol	Ethanol	64-17-5	0.321
12	DEM	Diethyl malonate	105-53-3	0.343
13	Paraoxon	Diethyl 4-nitrophenyl phosphate	311-45-5	0.456
14	DPCP	Diphenyl chloro phosphate	2524-64-3	0.541
15	DOP	Bis(2-ethylhexyl) phthalate	117-81-7	0.659
0	GD	Soman	96-64-0	0.000
1	DIMP	Di-isopropyl methyl phosphonate	1445-75-6	0.228
2	DEEP	Diethyl ethyl phosphonate	78-38-6	0.239
3	DFP	Di-isopropyl fluorophosphate	55-91-4	0.261
4	DEHP	Diethyl ester phosphonic acid	762-04-9	0.275
5	TEP	Triethyl Phosphate	78-40-0	0.302
6	DMMP	Dimethyl methyl phosphonate	756-79-6	0.303
7	BUSH	1-Butanethiol	109-79-5	0.320
8	TMP	Trimethyl phospate	512-56-1	0.328
9	ECA	Ethyl chloro acetate	105-39-5	0.353
10	DPGME	Dipropylene glycol monomethyl ether	34590-94-8	0.364
11	DEM	Diethyl malonate	105-53-3	0.380
12	Ethanol	Ethanol	64-17-5	0.404
13	Paraoxon	Diethyl 4-nitrophenyl phosphate	311-45-5	0.440
14	DPCP	Diphenyl chloro phosphate	2524-64-3	0.541
15	DOP	Bis(2-ethylhexyl) phthalate	117-81-7	0.610
0	GA	Tabun	77-81-6	0.000
1	DEHP	Diethyl ester phosphonic acid	762-04-9	0.238
2	DEEP	Diethyl ethyl phosphonate	78-38-6	0.245
3	DMMP	Dimethyl methyl phosphonate	756-79-6	0.265
4	DIMP	Di-isopropyl methyl phosphonate	1445-75-6	0.269
5	TMP	Trimethyl phospate	512-56-1	0.275
6	TEP	Triethyl phosphate	78-40-0	0.286
7	DFP	Di-isopropyl fluorophosphate	55-91-4	0.295
8	BUSH	1-Butanethiol	109-79-5	0.311
9	ECA	Ethyl chloro acetate	105-39-5	0.316
10	DPGME	Dipropylene glycol monomethyl ether	34590-94-8	0.341
11	DEM	Diethyl malonate	105-53-3	0.363
12	Ethanol	Ethanol	64-17-5	0.382
13	Paraoxon	Diethyl 4-nitrophenyl phosphate	311-45-5	0.426
14	DPCP	Diphenyl chloro phosphate	2524-64-3	0.511
15	DOP	Bis(2-ethylhexyl) phthalate	117-81-7	0.617
0	HD	Distilled mustard	505-60-2	0.000
1	CEES	2-Chloroethyl ethyl sulfide	693-07-2	0.180
2	CEMS	2-Chloroethyl methyl sulfide	542-81-4	0.256
3	CEPS	Chloroethyl phenyl sulfide	5535-49-9	0.457
4	DEM	Diethyl malonate	105-53-3	0.480
5	DMA	Dimethyl adipate	627-93-0	0.486
6	DEA	Diethyl adipate	141-28-6	0.560
7	MS	Methyl salicylate	119-36-8	0.587
8	DEP	Diethyl pimelate	2050-20-6	0.637
0	L	Lewisite	541-25-3	0.000
1	LO	Lewisite oxide	3088-37-7	0.459
2	PAO	Phenylarsine oxide	637-03-6	0.770
0	VX	O-ethyl *S*-[2-(diisopropylamino)ethyl] methylphosphonothiolate	50782-69-9	0.000
1	Amiton	O,O-diethyl-S-[2- (diethylamino)ethyl] phosphorothiolate	78-53-5	0.195
2	DEP	Diethyl pimelate	2050-20-6	0.291
3	BisEHP	Bis(2-ethylhexyl) phosphonate	3658-48-8	0.305
4	Malathion	S-[1,2-bis(ethoxycarbonyl) ethyl] O,O-dimethyl phosphorodithiolate	121-75-5	0.342
5	DEM	Diethyl malonate	105-53-3	0.349
6	DEPPT	O,S-diethyl phenyl phosphonothiolate	57557-80-9	0.367
7	DES	Diethyl sebacate	110-40-7	0.376
8	Parathion	O,O-diethyl-O-p nitrophenyl thiophosphate	56-38-2	0.395
9	DEPh	Diethyl phthalate	84-66-2	0.421
10	BisEHEHP	Bis(2-ethyl 1-hexyl) 2-ethyl 1-hexyl phosphonate	126-63-6	0.466

Table 4
Top CWA simulants found in this study in decreasing order of similarity.

CWA	Simulants based on	
	Tanimoto coefficient ≥ 0.6	Euclidean distance ≤ 0.25
GB	DFP, DIMP, DMMP, TMP	DMMP, DIMP, DEHP, BUSH, DEEP
GD	DFP, DIMP, DMMP, TMP	DIMP, DEEP
GA	None	DEHP, DEEP
HD	CEES	CEES
VX	Amiton, Malathion, BisEHEHP	Amiton
Lewisite	None	LO

be written as binary records. Between two binary strings A and B (each comprising 166 bits) the Tanimoto coefficient (TC) is calculated using the definition: $\mathrm{TC} = N_\mathrm{C}/(N_\mathrm{A} + N_\mathrm{B} - N_\mathrm{C})$, where N_A and N_B are the number of ones in strings A and B, respectively, and N_C is the number of bits which are ones in both A and B. Two examples (employing strings much shorter than the canonical 166 bits) are instructive:

Example 1. String A: 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0
SenString B: 1 0 0 0 0 0 0 1 0 0 0 1 0 0 1

It can be discerned that $N_\mathrm{A} = 3$, $N_\mathrm{B} = 4$, and $N_\mathrm{C} = 2$; TC = 0.4.

Example 2. String A: 1 1 1 1 1 1 1 1 1 1 1 1
String B: 1 1 1 1 1 1 1 1 1 1 1 1

Here, the two strings are identical: $N_\mathrm{A} = N_\mathrm{B} = N_\mathrm{C} = 12$; TC = 1.

Euclidean distance: For each pair of molecules, the Euclid distance is calculated based on up to 947 non-zero variance constitutional, topological, and conformation descriptors (computed within Sarchitect®). Each column of descriptors is "center-and-scale" normalized to have zero mean and unit variance. The Euclidean distance between two chemicals in rows p and q based on n columns of normalized descriptors d_norm is calculated as: $\mathrm{ED}_{pq} = \sqrt{(1/n)\sum_{i=1}^{i=n}(d_{\mathrm{norm},ip} - d_{\mathrm{norm},iq})^2}$.

7. Results and discussion

The Tanimoto coefficients and Euclidean distances were calculated for a number of chemicals commonly used as simulants for the chemical warfare agents GD, GA, GB, HD, VX and Lewisite. The results are listed in Tables 2 and 3, respectively. The yields for the nearest simulants based on this molecular similarity search from the two distance metrics are compared in Table 4.

These results meet the scope of this paper, namely, locating each CWA and its customary simulants in the quantitative space of molecular descriptors. It is gratifying to note that many of the simulants previously identified by experts based on intuitive comparisons of a few structural features and physical properties are also short-listed as close simulants by the present similarity search based on numerous calculated molecular descriptors, although it must be admitted that the list will be shorter, if the similarity criteria were stricter; e.g., TC > 0.9 instead of 0.6, or ED < 0.1 instead of 0.25. Searches using larger databases may point to other chemicals as closer simulants, but the simulants addressed here will retain the present Tanimoto coefficients with respect to the CWAs; the Euclidean distances may change (via the descriptor-normalization step) if the mean and the variance of descriptors change with database size and makeup. The rankings may also change, if the similarity metrics are calculated using alternative sets of fingerprints and descriptors [18,19], but it may turn out, as noted in Ref. [8], that "for any particular query chemical the set of neighbors (selected using physicochemical and topological spaces) is essentially the same with some minor variation, though the order of neighbor selection differs."

Upon closer scrutiny of Table 4, it becomes apparent that – in the similarity space of CWAs and customary simulants – the Euclidean distance is somewhat more inclusive than the Tanimoto coefficient. A possible cause may have to do with the way in which Sarchitect® calculates the two similarity metrics: Tanimoto coefficients are solely based on the binary MACCS fingerprint keys, which record the presence or absence of various features as ones or zeroes. As a result, when a query molecule (CWA) has one phosphorous atom, the Sarchitect® TC based on binary MACCS fingerprints would not discriminate between a target simulant that has one phosphorous atom and another that has multiple phosphorous atoms. In contrast, the Euclidean distance may be a more robust measure of (dis)similarity, since it is calculated using a variety of descriptors, optionally including the integer MACCS *descriptors* which record the number of counts of the same features the mere presence or absence of which is coded by the MACCS *fingerprints*. The two measures, Euclidean distance (ED) and Tanimoto Coefficient (TC), are fundamentally different. TC depends only on the presence or absence (1 or 0; yes or no) of molecular features, while ED depends on the extent (big or small, long or short) of the features. Further, in the present work, each is based on a different set of features of the molecules being compared. Accordingly, the two are not expected to be "monotonic" with each other; i.e. yield identical similarity rankings. By the same token, the two are not expected to yield completely contradictory rankings either, since the underlying feature sets may be correlated at a deeper level. The Tanimoto coefficients are bound by 0 and 1, by definition; Euclidean distances need not be bound by 0 and 1, but in this work the distances have been normalized to remain within those bounds.

Broader investigations may reveal where CWAs reside in the larger chemical similarity space, not restricted to customary simulants as in the present work. At this stage, this study can be taken as a ratification of the computational similarity search method as an efficient technique to find simulants for CWAs, which is a welcome advance considering that the number of potential CWAs for which simulants are unknown may far exceed [2] the small set with known simulants [4]. There is a danger, however, that HTS methods may be used for *finding* more CWAs. That is, by emphasizing only chemical similarity, searches may trawl in more and more toxic chemicals as candidates for new CWAs.

It is appropriate to close with a brief outline of future research avenues. Aside from perfecting the current search procedure – by augmenting the database of CWAs and potential simulants, and altering the choice of descriptors, similarity metrics, and software – alternative search techniques such as Random Forests and Artificial Neural Networks [20] may be explored:

- With random forests, the entire data base comprising both CWAs and potential simulants would get sorted into classes based on multiple randomized classification passes through numerous descriptors. Irrelevant or redundant descriptors would be eliminated as a concomitant of the sorting. Simulants can be identified from the "proximity matrix."
- Similarly, a neural network can be trained by giving it the query CWA as well as a list of traditional simulants; then the trained network can be employed to find new simulants from chemical databases.

Also, practical considerations can be accommodated by boosting the list of descriptors to include, besides purely physicochemical characteristics, practical measures such as cost, and "endpoints" such as toxicity (recognizing that quantifying toxicity is a research topic by itself [21–23]), ranking the database of potential simulants in terms of toxicity and cost, and restricting the search for simulants only to the moiety of chemicals below an acceptable toxicity level and reasonable cost.

Acknowledgements

The work was supported by an Early Applied Research (EAR) award from the U.S. Army Natick Soldier Research, Development & Engineering Center and by the Defense Threat Reduction Agency Project # BA10PHM050. The authors thank Strand Life Sciences personnel (chiefly K. Subramanian, K. Sumathy, and A. Das) for custom scripts and numerous discussions on Sarchitect® usage.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jhazmat.2011.07.077.

References

[1] J.A. Romano Jr., B.J. Lukey, H. Salem (Eds.), Chemical Warfare Agents: Chemistry, Pharmacology, Toxicology, and Therapeutics, 2nd ed., CRC Press, Cleveland, 2008.

[2] D.H. Ellison, Handbook of Chemical and Biological Warfare Agents, 2nd ed., CRC Press, Cleveland, 2008.

[3] D. Rivin, R.S. Lindsay, W.J. Shuely, A. Rodriguez, Liquid permeation through nonporous barrier materials, J. Membr. Sci. 246 (2005) 39–47.

[4] S.L. Bartelt-Hunt, D.R.U. Knappe, M.A. Barlaz, A review of chemical warfare agent simulants for the study of environmental behavior, Crit. Rev. Environ. Sci. Technol. 38 (2008) 112–136.

[5] D.N. Theodorou, Molecular simulations of sorption and diffusion in amorphous polymers, in: P. Neogi (Ed.), Diffusion in Polymers, Marcel Dekker, New York, 1996, pp. 67–142.

[6] R.C. Glen, S.E. Adams, Similarity metrics and descriptor spaces—which combinations to choose? QSAR Comb. Sci. 25 (2006) 1133–1142.

[7] P. Willett, J.M. Barnard, G.M. Downs, Chemical similarity searching, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996.

[8] S.C. Basak, B. Gute, D. Mills, Similarity methods in analog selection, property estimation and clustering of diverse chemicals, ARKIVOC 9 (2006) 157–210.

[9] S.C. Basak, Role of mathematical chemodescriptors and proteomics-based biodescriptors in drug discovery, Drug Dev. Res. 72 (2010) 1–9.

[10] J. Xu, A. Hagler, Chemoinformatics and drug discovery, Molecules 7 (2002) 566–600.

[11] L. Franke, O. Schwarz, L. Müller-Kuhrt, C. Hoernig, L. Fischer, S. George, Y. Tanrikulu, P. Schneider, O. Werz, D. Steinhilber, G. Schneider, Identification of natural-product-derived inhibitors of 5-lipoxygenase activity by ligand-based virtual screening, J. Med. Chem. 50 (2007) 2640–2646.

[12] Q. Zhu, M.S. Lajiness, Y. Ding, D.J. Wild, WENDI. A tool for finding non-obvious relationships between compounds and biological properties, genes, diseases, and scholarly publications, J. Cheminform. 2 (6) (2010) 1–9.

[13] B.D. Gute, S.C. Basak, D. Mills, D.M. Hawkins, Tailored similarity spaces for the prediction of physicochemical properties, IEJMD 1 (2002) 374–387.

[14] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Deliv. Rev. 23 (1997) 3–25.

[15] E.O. Cannon, F. Nigsch, J.B.O. Mitchell, A novel hybrid ultrafast shape descriptor method for use in virtual screening, Chem. Cent. J. 2 (3) (2008) 1–9.

[16] R. Todeschini, V. Consonni, Molecular descriptors for chemoinformatics, 2nd ed., Wiley-VCH, Weinheim, Federal Republic of Germany, 2009.

[17] J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, Re-optimization of MDL keys for use in drug discovery, J. Chem. Inf. Comp. Sci. 42 (2002) 1273–1280.

[18] J.X. Duan, S.L. Dixon, J.F. Lowrie, W. Sherman, Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods, J. Mol. Graph. Model. 29 (2010) 157–170.

[19] M. Sastry, J.F. Lowrie, S.L. Dixon, W. Sherman, Large scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments, J. Chem. Inf. Model. 50 (2010) 771–784.

[20] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009.

[21] D. Mackay, J.A. Arnot, E.P. Petkova, K.B. Wallace, D.J. Call, L.T. Brooke, G.D. Veith, The physicochemical basis of QSARs for baseline toxicity, SAR QSAR Environ. Res. 20 (2009) 393–414.

[22] E. Benfenati, Predicting toxicity through computers: a changing world, Chem. Cent. J. 1 (32) (2007) 1–7.

[23] M.T.D. Cronin, D.J. Livingstone (Eds.), Predicting Chemical Toxicity and Fate, CRC Press, Cleveland, 2004.