MayaChemTools: An open source package for computational discovery

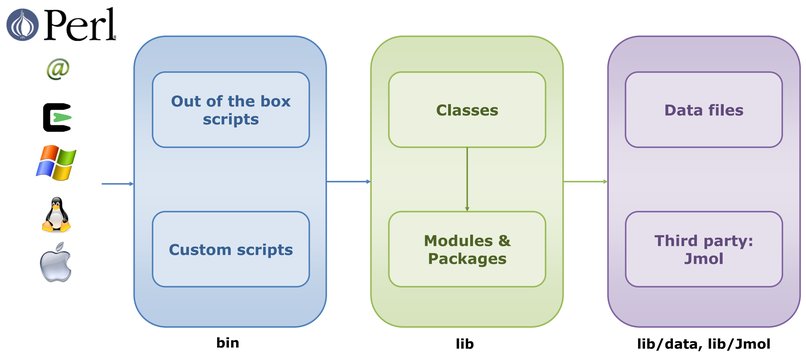
Manish Sud
COMP #306, 243rd ACS National Meeting & Exposition, March 25-29 2012, San Diego, CA

MayaChemTools.org

Introduction

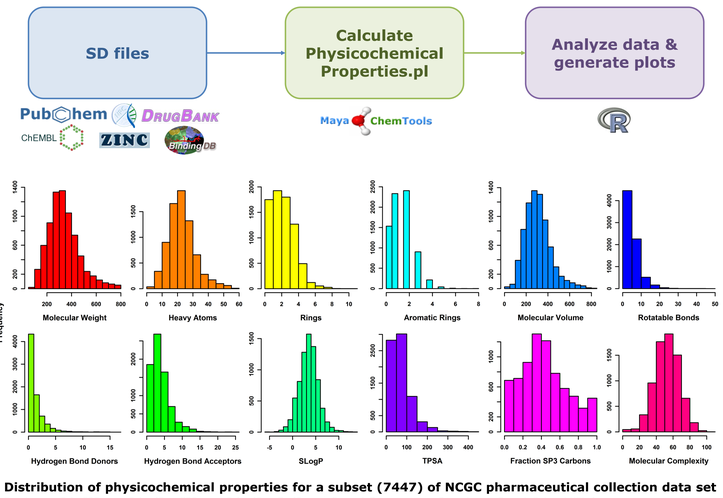
MayaChemTools is a growing collection of Perl scripts, modules and classes to support day-to-day computational drug discovery needs. It provides the following out of the box functionality: Manipulation and analysis of data in SD, CSV/TSV, sequence/alignments, PDB and fingerprints files; Properties of periodic table elements, amino acids and nucleic acids; Calculation of physicochemical properties such as hydrogen bond donors and acceptors, SLogP and topological polar surface area; Generation of fingerprints corresponding to atom neighborhoods, atom types, E-state indicies, extended connectivity, MACCS keys, path lengths, topological atom pairs/triplets/torsions and topological pharmacophore atom pairs/triplets; Similarity searching and calculation of similarity matrices. An extensive set of modules and classes are also available for custom development. MayaChemTools is freely available under the terms of the LGPL license at www.MayaChemTools.org.

Software architecture



Physicochemical properties profiling

Name	Description
Molecular Weight	Sum of atomic weights
Heavy Atoms	Number of non-hydrogen atoms
Rings, Aromatic Rings	Number of rings and aromatic rings (aromaticity detection using Hückel's rule)
Rotatable bonds	Number of non-ring single bonds involving only non-hydrogen atoms with the option to exclude: terminal bonds; attached to triple bonds; amide, thioamide and sulfonamide bonds
van der Waals Molecular Volume	Sum of atomic volumes corresponding to van der Waals atomic radii with adjustments for number of bonds, aromatic and non-aromatic rings
Hydrogen Bond Donors & Acceptors	Type1 - Donor: Any N and O with implicit/explicit H; Acceptor: Any N without implicit/explicit H and any O Type2 - Donor: Any N and O with implicit/explicit H; Acceptor: Any N and O
LogP & Molar Refractivity (SLogP & SMR)	Sum of atomic contributions from pre-defined atom types corresponding to specific structure fragments
Topological Polar Surface Area (TPSA)	Sum of atomic contributions from pre-defined N and O atom types corresponding to specific structure fragments with option to include P and N atom types
Fraction of SP3 Carbons (FSP3Carbons)	Number of SP3Carbons divided by the total number of carbons
Molecular Complexity	Number of bits-set or unique keys in 2D fingerprints. Supported fingerprints: atom types, extended connectivity, MACCS keys, path lengths, topological atom pairs/triplets/torsions and topological pharmacophore atom pairs/triplets



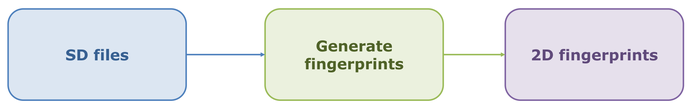
Distribution of physicochemical properties for a subset (7447) of NCGC pharmaceutical collection data set

Scripts used: FilterSDFiles.pl, ExtractFromSDFiles.pl, ExtractFromTextFiles.pl, CalculatePhysicochemicalProperties.pl, Rscript; Data set URL: tripod.nih.gov/npc

2D Fingerprints

Type	Values Type	Key Default Parameters/Description
Atom Neighborhoods	Vector	Values: Alphanumerical vector; MinNeighborhoodRadius: 0; MaxNeighborhoodRadius: 2; AtomIdentifierType: AtomicInvariants (AS, X, BO, H,F C)
Atom Types	Bit-vector or vector	Values: Numerical vector; AtomIdentifierType: AtomicInvariants (AS, X, BO, H, FC)
E-state Indicies	Vector	Values: Numerical vector; EStatAtomTypesSetSize: Arbitrary
Extended Connectivity	Bit-vector or vector	Values: Alphanumerical vector; NeighborhoodRadius: 2; AtomIdentifierType: AtomicInvariants (AS, X, BO, H, FC, MN)
MACCS Keys	Bit-vector or vector	Values: Bit-vector; Size: 166; Available sizes: 166 and 322; Keys count available
Path Lengths	Bit-vector or vector	Values: Bit-vector; Size: 1024; AtomIdentifierType: AtomicInvariants (AS); MinPathLength: 1; MaxPathLength: 8; Paths count available
Topological Atom Pairs	Vector	Values: Numerical vector; AtomIdentifierType: AtomicInvariants (AS, X, BO, H, FC); MinDistance: 1; MaxDistance: 10
Topological Atom Triplets	Vector	Values: Numerical vector; AtomIdentifierType: AtomicInvariants (AS,X,BO,H,FC); MinDistance: 1; MaxDistance: 10; TriangleInequality: No
Topological Atom Torsions	Vector	Values: Numerical vector; AtomIdentifierType: AtomicInvariants (AS, X, BO, H, FC)
Topological Pharmacophore Atom Pairs	Vector	Values: Numerical vector; AtomTypes: HBD, HBA, PI, NI, H; MinDistance: 1; MaxDistance: 10; AtomTypesWeight: None; Normalization: None; FuzzifyAtomPairsCount: No
Topological Pharmacophore Atom Triplets	Vector	Values: Numerical vector; AtomTypes: HBD, HBA, PI, NI, H, Ar; MinDistance: 1; MaxDistance: 10; DistanceBinSize: 2; TriangleInequality: Yes

Atom identifier atom types: Atomic invariants, Functional class, DREIDING, EState, MMFF94, SLogP, SYBYL, TPSA and UFF
Atomic invariants: AS(Atom symbol), X(Num of heavy atom neighbors), BO(Sum of bond orders to heavy atoms), LBO(Largest bond order to heavy atoms), SB(Num of single bonds to heavy atoms), DB(Num of double bonds to heavy atoms), TB(Num of Triple bonds to heavy atoms), H(Num of implicit and explicit hydrogens), Ar (Aromatic), RA(Ring atom), FC(Formal charge), MN(Mass number), SM(Spin multiplicity)
Functional class: HBD(Hydrogen bond donor), HBA(Hydrogen bond acceptor), PI(Positively ionizable), NI(Negatively ionizable), Ar(Aromatic), Hal(Halogen), H(Hydrophobic), RA(RingAtom), CA(ChainAtom)



Fingerprints comparisons

Fingerprints bit-vectors:

Name	Formula
Baroni Urbani & Buser	(SQRT(Nc*Nd) + Nc)/(SQRT(Nc*Nd) + Nc + (Na –Nc) + (Nb -Nc))
Cosine & Ochiai	Nc/SQRT(Na*Nb)
Dice	2*Nc/(Na + Nb)
Dennis	(Nc*Nd -((Na - Nc)*(Nb - Nc)))/SQRT(Nt*Na*Nb)
Forbes	Nt*Nc/Na*Nb
Fossum	(Nt*((Nc – 0.5)**2)/(Na*Nb)
Hamann	((Nc + Nd) - (Na - Nc) - (Nb - Nc))/Nt
Jaccard & Tanimoto	Nc/((Na - Nc) + (Nb –Nc) + Nc)) = Nc/(Na + Nb - Nc)
Kulczynski	1: Nc/(Na + Nb -2Nc) 2: 0.5*(Nc/Na + Nc/Nb)
Matching	(Nc + Nd)/Nt
McConnaughey	(Nc**2 - (Na - Nc)*(Nb - Nc))/(Na*Nb)
Pearson	((Nc*Nd) - ((Na - Nc)*(Nb - Nc))/SQRT(Na*Nb*(Na – Nc + Nd)*(Nb – Nc + Nd))
Rogers Tanimoto	(Nc + Nd)/(Na + Nb - 2Nc + Nt)
Russell Rao	Nc/Nt
Simpson	Nc/MIN(Na, Nb)
Skoal Sneath	1: Nc/(2*Na + 2*Nb -3*Nc) 2: (2*Nc + 2*Nd)/(Nc + Nd +Nt) 3: (Nc + Nd)/(Na + Nb -2*Nc)
Tversky	Nc/(alpha*(Na - Nb) + Nb)
Yule	((Nc*Nd) - ((Na - Nc)*(Nb - Nc)))/((Nc*Nd) + ((Na - Nc)*(Nb - Nc)))

Na = Num of bits set to "1" in A
Nb = Num of bits set to "1" in B
Nc = Num of bits set to "1" in both A and B
Nd = Num of bits set to "0" in both A and B
Nt = Num of bits set to "1" or "0" in A and B
Nt = Na + Nb - Nc + Nd
Na -Nc = Num of bits set to "1" in A not in B
Nb - Nc = Num of bits set to "1" in B not in A

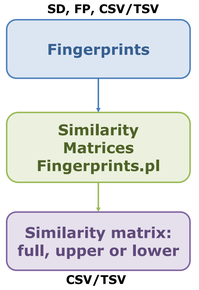
Fingerprints vectors containing ordered numerical, numerical or alphanumerical values:

Name	Albgebric Form	Binary Form	Set Theoretic Form
City Block, Hamming & Manhattan Distance	SUM(ABS (Xai –Xbi))	Na + Nb – 2*Nc	SUM(Xai) + SUM (Xbi) - 2*(SUM(MIN(Xai, Xbi)))
Cosine & Ochiai Similarity	SUM(Xai*Xbi) / SQRT(SUM (Xai**2) * SUM(Xbi**2))	Nc/SQRT(Na*Nb)	SUM(MIN(Xai, Xbi)) / SQRT(SUM(Xai) * SUM(Xbi))
Czekanowski , Dice & Sorenson Similarity	(2*(SUM (Xai*Xbi))) / (SUM (Xai**2) + SUM (Xbi**2))	2*Nc/(Na + Nb)	2*(SUM(MIN (Xai, Xbi))) / (SU (Xai) + SUM (Xbi))
Euclidean Distance	SQRT(SUM((Xai – Xbi)**2))	SQRT(Na + Nb – 2*Nc)	SQRT(SUM (Xai) + SUM (Xbi) – 2*(SUM(MIN(Xai, Xbi)))
Jaccard & Tanimoto Similarity	SUM(Xai *Xbi) / (SUM (Xai**2) + SUM (Xbi**2) – SUM (Xai*Xbi))	Nc/(Na + Nb –Nc)	SUM(MIN(Xai, Xbi)) / (SUM(Xai) + SUM (Xbi) – SUM(MIN(Xai, Xbi)))
Soergel Distance	SUM(ABS(Xai - Xbi)) / SUM(MAX(Xai, Xbi))	(Na + Nb – 2*Nc)/(Na + Nb - Nc)	(SUM(Xai) + SUM(Xbi) - 2*(SUM(MIN(Xai, Xbi)))) / (SUM(Xai) + SUM(Xbi) - SUM(MIN(Xai, Xbi)))

N = Num of values
SUM = Sum over values
Xa = Values of vector A
Xai = Value of ith element in A
Xb = Values of vector B
Xbi = Value of ith element in B
SetIntersectionXaXb = SUM(MIN(Xai, Xbi))
SetDifferenceXaXb = SUM(Xa)+ SUM(Xb) - SUM(MIN(Xai, Xbi))

Na = Num of bits set to "1" in A = SUM(Xai)
Nb = Num of bits set to "1" in B = SUM(Xbi)
Nc = Num of bits set to "1" in both A and B = SUM(Xai*Xbi)
Nd = Num of bits set to "0" in both A and B = SUM(1 - Xai - Xbi + Xai*Xbi)

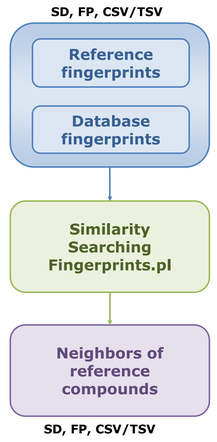
Similarity matrices



	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	1	0.29	0.41	0.43	0.43	0.38	0.39	0.33	0.3	0.5	0.52	0.36	0.4	0.33	0.29	0.45	0.58	0.24	0.33	0.61
Arg	0.29	1	0.36	0.33	0.33	0.42	0.39	0.23	0.29	0.27	0.31	0.45	0.36	0.28	0.36	0.34	0.26	0.25	0.28	0.26
Asn	0.41	0.36	1	0.59	0.46	0.52	0.42	0.32	0.42	0.36	0.42	0.39	0.38	0.46	0.29	0.48	0.35	0.34	0.46	0.36
Asp	0.43	0.33	0.59	1	0.48	0.42	0.5	0.4	0.43	0.37	0.44	0.41	0.39	0.48	0.3	0.5	0.36	0.35	0.48	0.38
Cys	0.43	0.33	0.46	0.48	1	0.42	0.44	0.33	0.34	0.37	0.44	0.41	0.39	0.38	0.3	0.5	0.36	0.28	0.38	0.38
Gln	0.38	0.42	0.52	0.42	0.42	1	0.62	0.29	0.31	0.33	0.39	0.46	0.45	0.34	0.31	0.44	0.32	0.26	0.34	0.33
Glu	0.39	0.39	0.42	0.5	0.44	0.62	1	0.36	0.32	0.34	0.41	0.48	0.46	0.35	0.32	0.46	0.33	0.27	0.35	0.35
Gly	0.33	0.23	0.32	0.4	0.33	0.29	0.36	1	0.23	0.29	0.3	0.33	0.27	0.26	0.26	0.35	0.27	0.19	0.26	0.29
His	0.3	0.29	0.42	0.43	0.34	0.31	0.32	0.23	1	0.28	0.32	0.31	0.3	0.4	0.26	0.35	0.26	0.37	0.4	0.27
Ile	0.5	0.27	0.36	0.37	0.37	0.33	0.34	0.29	0.28	1	0.44	0.32	0.35	0.3	0.27	0.38	0.48	0.23	0.3	0.5
Leu	0.52	0.31	0.42	0.44	0.44	0.39	0.41	0.3	0.32	0.44	1	0.38	0.41	0.35	0.28	0.46	0.44	0.27	0.35	0.46
Lys	0.36	0.45	0.39	0.41	0.41	0.46	0.48	0.33	0.31	0.32	0.38	1	0.43	0.33	0.34	0.42	0.31	0.26	0.33	0.32
Met	0.4	0.36	0.38	0.39	0.39	0.45	0.46	0.27	0.3	0.35	0.41	0.43	1	0.32	0.29	0.41	0.34	0.25	0.32	0.36
Phe	0.33	0.28	0.46	0.48	0.38	0.34	0.35	0.26	0.4	0.3	0.35	0.33	0.32	1	0.24	0.39	0.29	0.4	0.64	0.3
Pro	0.29	0.36	0.29	0.3	0.3	0.31	0.32	0.26	0.26	0.27	0.28	0.34	0.29	0.24	1	0.31	0.25	0.21	0.24	0.26
Ser	0.45	0.34	0.48	0.5	0.5	0.44	0.46	0.35	0.35	0.38	0.46	0.42	0.41	0.39	0.31	1	0.38	0.29	0.39	0.39
Thr	0.58	0.26	0.35	0.36	0.36	0.32	0.33	0.27	0.26	0.48	0.44	0.31	0.34	0.29	0.25	0.38	1	0.22	0.29	0.57
Trp	0.24	0.25	0.34	0.35	0.28	0.26	0.27	0.19	0.37	0.23	0.27	0.26	0.25	0.4	0.21	0.29	0.22	1	0.37	0.23
Tyr	0.33	0.28	0.46	0.48	0.38	0.34	0.35	0.26	0.4	0.3	0.35	0.33	0.32	0.64	0.24	0.39	0.29	0.37	1	0.3
Val	0.61	0.26	0.36	0.38	0.38	0.33	0.35	0.29	0.27	0.5	0.46	0.32	0.36	0.3	0.26	0.39	0.57	0.23	0.3	1

Scripts used: ExtendedConnectivityFingerprints.pl, SimilarityMatricesFingerprints.pl, TextFilesToHTML.pl

Similarity searching

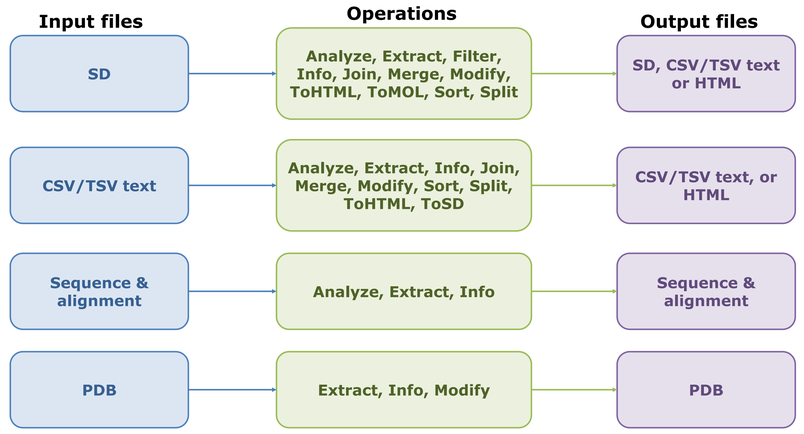


Showing table 1 of 7 Tables: 1 2 3 4 5 6 7 Next Showing compounds 1 to 15 of 100

Structure	DatabaseCmpdID	ComparisonValue	MolecularWeight	HeavyAtoms	AromaticRings	RotatableBonds	Hydr
	Elinogrel	1.00	523.95	34	3	6	4
	Ezetimibe	0.30	409.43	30	3	6	2
	Glimepiride	0.29	490.62	34	1	10	3
	Tosifen	0.28	332.42	23	2	6	2

Scripts used: PathLengthFingerprints.pl, SimilaritySearchingFingerprints.pl, SDFilesToHTML.pl

File data info, manipulation & analysis



Data retrieval from databases

Elements, amino acids & nucleic acids