VoroContacts: a tool for the analysis of interatomic contacts in macromolecular structures

Supplementary information

Kliment Olechnovič and Česlovas Venclovas

Institute of Biotechnology, Life Sciences Center, Vilnius University, Saulėtekio 7, Vilnius LT-10257, Lithuania

Contents

1 On units of length and area

All length values (e.g. radii values) in this document are in angstroms (Å). All area values — in square angstroms (Å^2). 1 Å= 10^{-10} meters.

2 Default van der Waals radii used by VoroContacts

For protein atoms, we use the detailed set published by Li and Nussinov [1], the set is shown in Table S1. For non-protein atoms, we use van der Waals radii radii values derived from multiple sources. We obtained the radii values for the most common elements in biological macromolecules (Table S2) by averaging and rounding the radii values from the Li and Nussinov set. For elements commonly appearing as ions (Table S3) we used the set of common ionic radii from the CRC Handbook of Chemistry and Physics [2]. For unrecognized atoms the radius value of 1.8 angstroms is assigned.

Table S1: Van der Waals radii used for protein atoms

ALA	C	1.75	CYS	CB	1.91	HIS	NE2	1.6	MET	SD	1.94	TRP	CA	1.9
ALA	CA	1.9	CYS	N	1.7	HIS	O	1.49				TRP	CB	1.91
ALA	CB	1.92	CYS	O	1.49				PHE	C	1.75	TRP	CD1	1.82
ALA	N	1.7	CYS	SG	1.88	ILE	C	1.75	PHE	CA	1.9	TRP	CD2	1.82
ALA	O	1.49				ILE	CA	1.9	PHE	CB	1.91	TRP	CE2	1.74
			GLN	C	1.75	ILE	CB	2.01	PHE	CD1	1.82	TRP	CE3	1.82
ARG	C	1.75	GLN	CA	1.9	ILE	CD1	1.92	PHE	CD2	1.82	TRP	CG	1.74
ARG	CA	1.9	GLN	CB	1.91	ILE	CG1	1.92	PHE	CE1	1.82	TRP	CH2	1.82
ARG	CB	1.91	GLN	CD	1.81	ILE	CG2	1.92	PHE	CE2	1.82	TRP	CZ2	1.82
ARG	CD	1.88	GLN	CG	1.8	ILE	N	1.7	PHE	CG	1.74	TRP	CZ3	1.82
ARG	CG	1.92	GLN	N	1.7	ILE	O	1.49	PHE	CZ	1.82	TRP	N	1.7
ARG	CZ	1.8	GLN	NE2	1.62				PHE	N	1.7	TRP	NE1	1.66
ARG	N	1.7	GLN	O	1.49	LEU	C	1.75	PHE	O	1.49	TRP	O	1.49
ARG	NE	1.62	GLN	OE1	1.52	LEU	CA	1.9						
ARG	NH1	1.62				LEU	CB	1.91	PRO	C	1.75	TYR	C	1.75
ARG	NH2	1.67	GLU	C	1.75	LEU	CD1	1.92	PRO	CA	1.9	TYR	CA	1.9
ARG	O	1.49	GLU	CA	1.9	LEU	CD2	1.92	PRO	CB	1.91	TYR	CB	1.91
			GLU	CB	1.91	LEU	CG	2.01	PRO	CD	1.92	TYR	CD1	1.82
ASN	C	1.75	GLU	CD	1.76	LEU	N	1.7	PRO	CG	1.92	TYR	CD2	1.82
ASN	CA	1.9	GLU	CG	1.88	LEU	O	1.49	PRO	N	1.7	TYR	CE1	1.82
ASN	CB	1.91	GLU	N	1.7				PRO	O	1.49	TYR	CE2	1.82
ASN	CG	1.81	GLU	O	1.49	LYS	C	1.75				TYR	CG	1.74
ASN	N	1.7	GLU	OE1	1.49	LYS	CA	1.9	SER	C	1.75	TYR	CZ	1.8
ASN	ND2	1.62	GLU	OE2	1.49	LYS	CB	1.91	SER	CA	1.9	TYR	N	1.7
ASN	O	1.49				LYS	CD	1.92	SER	CB	1.91	TYR	O	1.49
ASN	OD1	1.52	GLY	C	1.75	LYS	CE	1.88	SER	N	1.7	TYR	OH	1.54
			GLY	CA	1.9	LYS	CG	1.92	SER	O	1.49			
ASP	C	1.75	GLY	N	1.7	LYS	N	1.7	SER	OG	1.54	VAL	C	1.75
ASP	CA	1.9	GLY	O	1.49	LYS	NZ	1.67				VAL	CA	1.9
ASP	CB	1.91				LYS	O	1.49	THR	C	1.75	VAL	CB	2.01
ASP	CG	1.76	HIS	C	1.75				THR	CA	1.9	VAL	CG1	1.92
ASP	N	1.7	HIS	CA	1.9	MET	C	1.75	THR	CB	2.01	VAL	CG2	1.92
ASP	O	1.49	HIS	CB	1.91	MET	CA	1.9	THR	CG2	1.92	VAL	N	1.7
ASP	OD1	1.49	HIS	CD2	1.74	MET	CB	1.91	THR	N	1.7	VAL	O	1.49
ASP	OD2	1.49	HIS	CE1	1.74	MET	CE	1.8	THR	O	1.49			
			HIS	CG	1.8	MET	CG	1.92	THR	OG1	1.54			
CYS	C	1.75	HIS	N	1.7	MET	N	1.7						
CYS	CA	1.9	HIS	ND1	1.6	MET	O	1.49	TRP	C	1.75			

Table S2: Van der Waals radii used for the most common elements in biological macromolecules

C	1.8	N	1.6	O	1.5	S	1.9	P	1.9

Table S3: Van der Waals radii used for ions

Al^{+3}	0.54	As^{+3}	0.58	Au^{+1}	1.37	Ba^{+2}	1.35	Be^{+2}	0.45	Bi^{+3}	1.03
Ca^{+2}	1.00	Cd^{+2}	0.95	Co^{+2}	0.65	Cr^{+2}	0.73	Cs^{+}	1.67	Cu^{+2}	0.73
Fe^{+2}	0.61	Ga^{+3}	0.62	Ge^{+2}	0.73	Hg^{+2}	1.02	K^{+}	1.38	Li^{+}	0.76
Mg^{+2}	0.72	Mn^{+2}	0.83	Mo^{+3}	0.69	Na^{+}	1.02	Ni^{+2}	0.69	Pb^{+2}	1.19
Pd^{+2}	0.86	Pt^{+2}	0.80	Rb^{+}	1.52	Sb^{+3}	0.76	Sc^{+3}	0.75	Sn^{+4}	0.69
Sr^{+2}	1.18	Tc^{+4}	0.65	Ti^{+2}	0.86	V^{+2}	0.79	Zn^{+2}	0.74	Zr^{+4}	0.72
F^{-}	1.33	CL^{-}	1.81	Br^{-}	1.96	I^{-}	2.20				

3 Comparing computation of contacts using VoroContacts and dr_sasa software tools

3.1 Dataset used for the analysis

We selected 313 protein chains from PDB using the PISCES server [3]. by applying the following constraints:

- sequence identity between any protein pair is 20% or less;
- X-ray resolution is 1.4 angstroms or better;
- free R-factor is 0.14 or less;
- sequence length is between 50 and 500 residues.

The resulting set of PDB identifiers and chain names is listed in Table S4.

Table S4: Protein structures used for the analysis.

1C5E_A, 1EB6_A, 1G6X_A, 1GCI_A, 1I1W_A, 1IQZ_A, 1IX9_A, 1J0P_A, 1JFB_A, 1K5C_A,
1K7C_A, 1KQP_A, 1KWF_A, 1LWB_A, 1M15_A, 1M4L_A, 1MC2_A, 1MJ5_A, 1MUW_A, 1NKD_A,
1NWZ_A, 1O7J_A, 1OK0_A, 1PSR_A, 1R6J_A, 1SAU_A, 1SFS_A, 1UFY_A, 1US0_A, 1UWC_B,
1VBW_A, 1VYR_A, 1W0N_A, 1W23_A, 1X8Q_A, 1XG0_A, 1XG0_C, 1ZL0_A, 1ZUU_A, 1ZZK_A,
2A3N_A, 2AKZ_A, 2BK9_A, 2BT9_C, 2C71_A, 2CI1_A, 2CNQ_A, 2CS7_A, 2CWS_A, 2FKK_A,
2FVY_A, 2GGC_A, 2GKG_A, 2JFR_A, 2NLR_A, 2NWF_A, 2O7A_A, 2O9S_A, 2OB3_A, 2OFC_A,
2OIZ_A, 2OIZ_D, 2OV0_A, 2PND_A, 2PVB_A, 2R31_A, 2RH2_A, 2UU8_A, 2V3G_A, 2V3I_A,
2V8T_B, 2V9L_A, 2V9V_A, 2VB1_A, 2VXN_A, 2W15_A, 2WFI_A, 2XJP_A, 2XOM_A, 2XTS_A,
2YKZ_A, 2ZK9_X, 3A5F_A, 3A72_A, 3AGN_A, 3D1P_A, 3DLC_A, 3EA6_A, 3EO6_B, 3FSA_A,
3FYM_A, 3G0K_A, 3G21_A, 3G5T_A, 3HWU_A, 3HYN_A, 3IP0_A, 3JQ0_A, 3KWE_A, 3L8W_A,
3LAA_A, 3LWX_A, 3M5Q_A, 3MD7_A, 3MQD_A, 3NVS_A, 3O4P_A, 3PD7_A, 3PPL_A, 3PUC_A,
3Q46_A, 3QHB_A, 3QL9_A, 3QPA_A, 3QR7_A, 3RWN_A, 3S6E_A, 3T2C_A, 3TC8_A, 3TEU_A,
3TG2_A, 3ULJ_A, 3VII_A, 3VOR_A, 3W07_A, 3WDN_A, 3X2M_A, 3X34_A, 3ZOJ_A, 3ZSJ_A,
3ZUC_A, 3ZZP_A, 4A02_A, 4A29_A, 4ACJ_A, 4AFF_A, 4AL0_A, 4AYO_A, 4AZ6_A, 4BJ0_A,
4CE8_C, 4CHI_A, 4CJ0_B, 4CO8_A, 4E3Y_A, 4EIC_A, 4ERC_A, 4F06_A, 4F1V_A, 4FK9_A,
4G9S_B, 4H3U_A, 4HCJ_A, 4HNO_A, 4HS1_A, 4IAU_A, 4JN7_A, 4JXR_A, 4KEF_A, 4KQP_A,
4LF0_A, 4M1X_A, 4M51_A, 4M9V_F, 4MNC_A, 4MTM_A, 4MZC_A, 4N1I_A, 4NDS_A, 4NPD_A,
4O6U_A, 4QEK_A, 4QHW_A, 4QLP_A, 4QLP_B, 4REK_A, 4RXV_A, 4S39_A, 4U9H_S, 4UA6_A,
4UU3_B, 4UYR_A, 4UZG_A, 4W9Z_A, 4WBJ_A, 4WJT_A, 4WKA_A, 4WN5_A, 4WUI_A, 4WWF_A,
4X1Z_B, 4X2R_A, 4X5P_A, 4X9X_A, 4XDX_A, 4Y9W_A, 4YAA_A, 4YI8_A, 4YMY_A, 4ZGF_A,
4ZJU_A, 5A0Y_B, 5A0Y_C, 5A1I_A, 5A71_A, 5A8C_A, 5AGD_B, 5AIG_A, 5AOZ_A, 5CKL_A,
5COF_A, 5CTM_A, 5D8V_A, 5DP2_A, 5DZE_A, 5E9P_A, 5EL9_A, 5EQ7_A, 5EWO_A, 5FBF_A,
5FEW_A, 5GJI_A, 5GTQ_A, 5GV8_A, 5IDB_A, 5IG6_A, 5II6_A, 5IMA_A, 5J4L_A, 5JBX_A,
5JRY_A, 5JUG_A, 5JVI_E, 5L87_A, 5LP9_A, 5LS7_B, 5LS7_D, 5M0W_A, 5M17_A, 5MK9_A,
5MX9_A, 5NFM_A, 5NQO_A, 5NWP_A, 5O2X_A, 5O45_A, 5OHQ_A, 5OL4_A, 5OL4_B, 5OPZ_A,
5SY4_B, 5T5L_A, 5TAB_A, 5TDA_A, 5TIF_A, 5U3A_A, 5W8Q_A, 5WGI_A, 5X9L_A, 5Y0M_A,
5YDE_A, 5Z3E_A, 5ZW7_A, 6A9S_A, 6B8F_A, 6C4Q_A, 6CNW_A, 6DTV_A, 6EF7_A, 6EIO_A,
6EQE_A, 6ER4_B, 6ETL_A, 6EY1_A, 6FM7_A, 6FMC_A, 6GX2_A, 6H10_A, 6H40_A, 6HAV_A,
6HFQ_A, 6I6M_A, 6KFN_A, 6KLZ_A, 6L27_A, 6LL8_A, 6NIB_A, 6NNR_B, 6NP3_A, 6Q00_A,
6QHJ_A, 6RI6_A, 6RK0_A, 6RRV_A, 6RY0_A, 6RYG_A, 6SSD_A, 6TWT_B, 6WQY_A, 6Y4E_A,
6ZEG_B, 6ZEG_C, 7A3H_A, 7A5M_A, 7ADR_A, 7ADR_E, 7ADR_F, 7CN7_C, 7COF_A, 7JJA_A,
7KR0_A, 7JJA_A, 7KR0_A

3.2 Comparing contact areas

We processed all the 313 structures with VoroContacts (as a script running Voronota [4]) and dr_sasa [5] using the same set of van der Waals radii (taken from the dr_sasa configuration) and the same rolling probe radius of 1.4. As dr_sasa produces asymmetric contact matrices, we took an average of two dr_sasa areas for every contact. We compared three categories of contact areas:

- per-atom solvent accessible areas (Figure S1)
- residue-residue contact areas (Figure S2)
- atom-atom contact areas (Figure S3)

The results show that solvent accessible surface areas produced by VoroContacts and dr_sasa are nearly identical, the residue-residue contact areas are highly correlated (Pearson c.c. = 0.97), and atom-atom contact areas are moderately correlated (Pearson c.c. = 0.83). Analysis of the results showed that oftentimes a contact identified by dr_sasa was not identified as such by VoroContacts. The reason for such discrepancies is that, to identify contacts, VoroContacts uses the Voronoi tessellation, which takes into account not only the proximity between atoms, but also the structural context. Thus, in the structure tessellation, atoms close by distance may not always be direct neighbors, because of the shielding effect by other neighbors.

Figure S1: Comparison of per-atom solvent accessible areas.

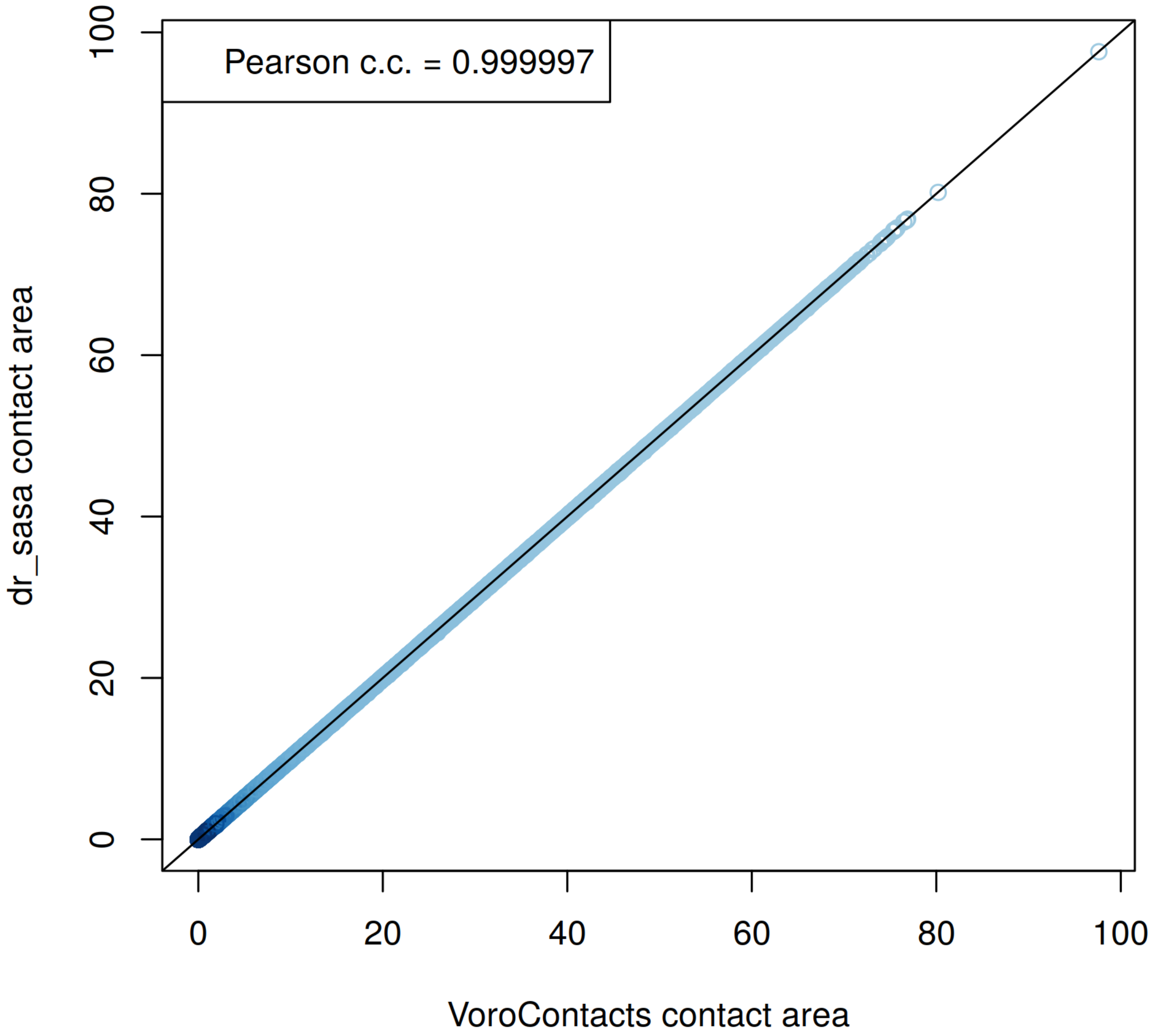


Figure S2: Comparison of residue-residue contact areas.

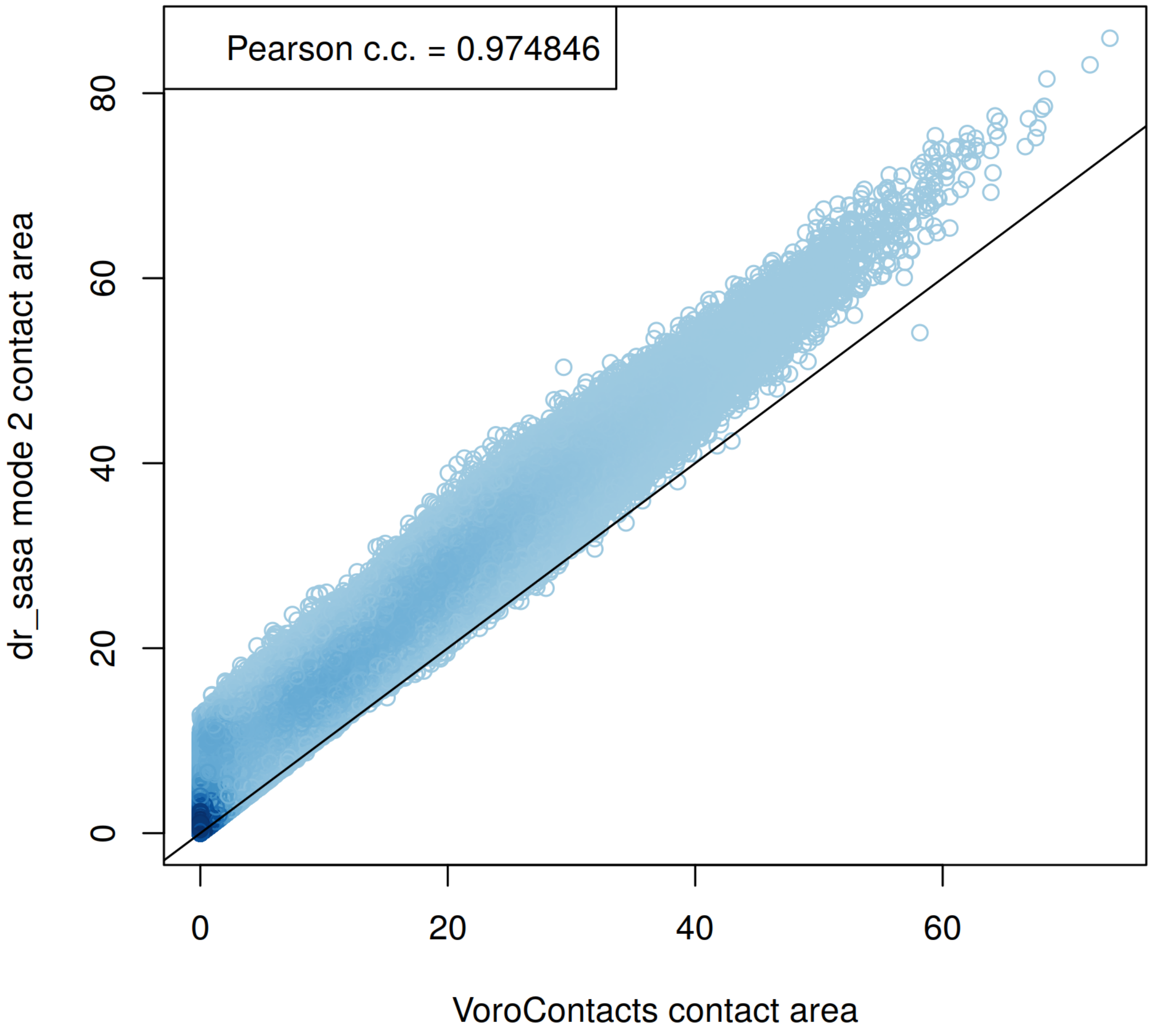
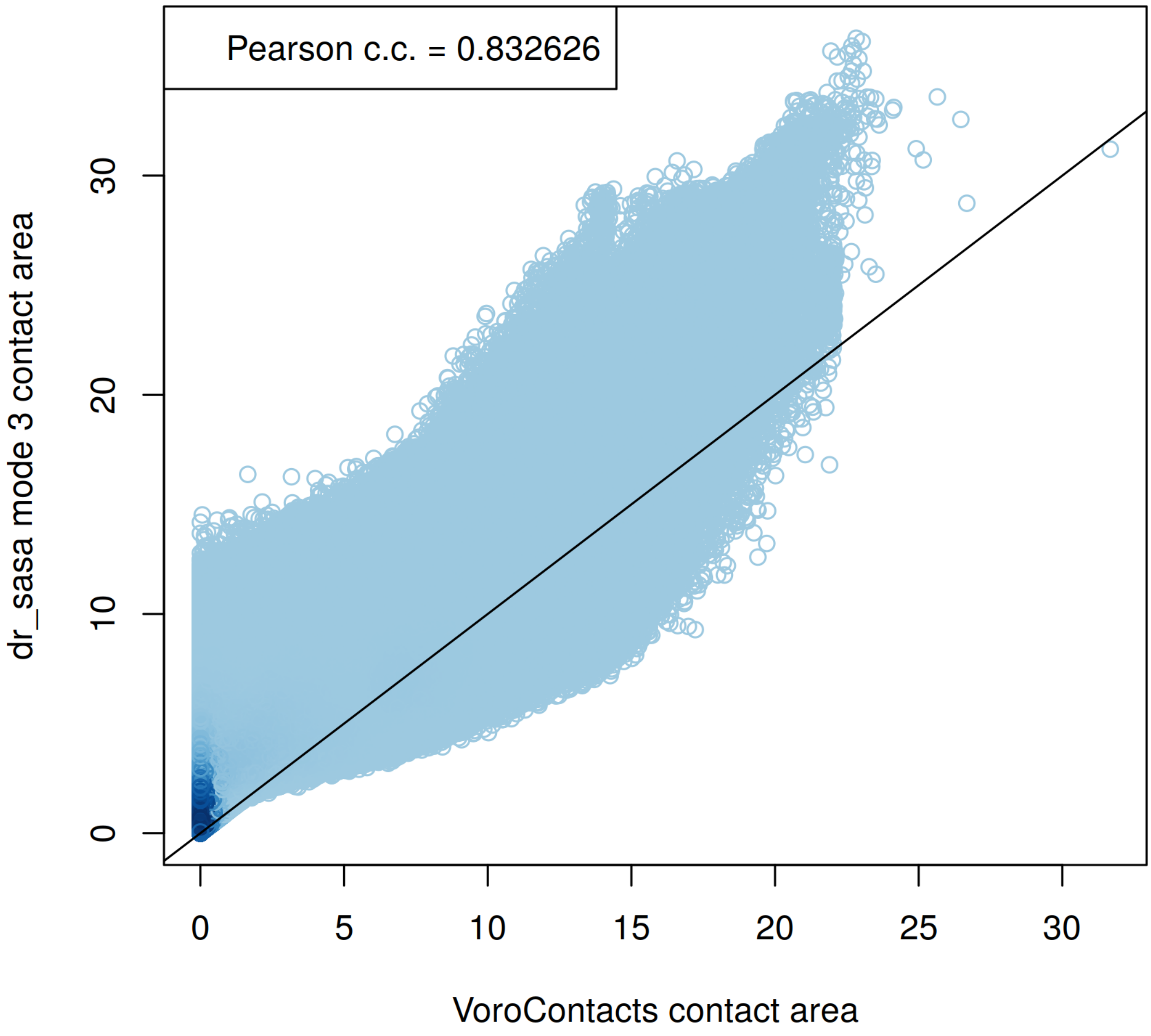


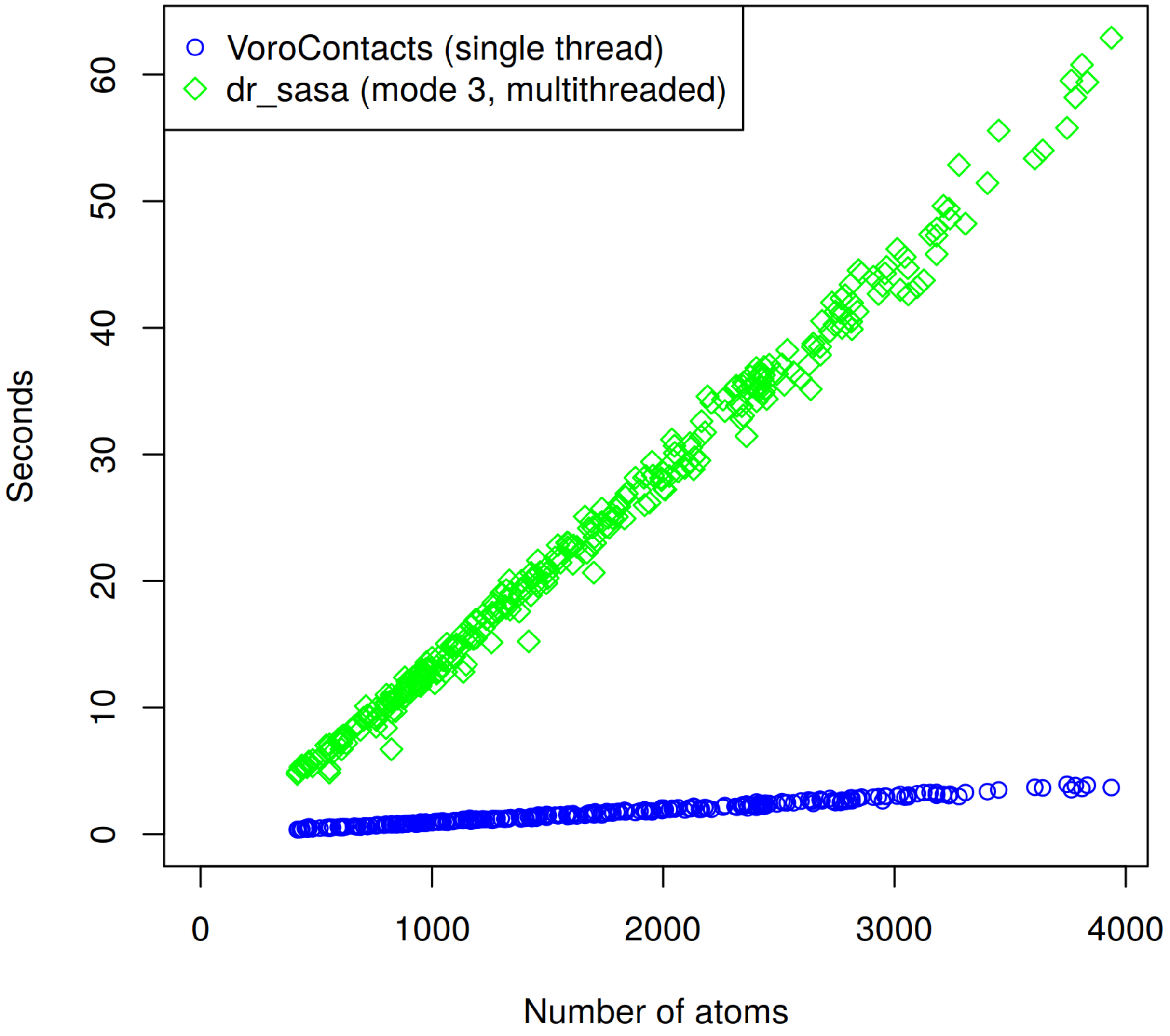
Figure S3: Comparison of atom-atom contact areas.



3.3 Comparing running times

We recorded running times on the same machine. Figure S4 shows that VoroContacts was several times faster, despite using only a single thread while dr_sasa used several.

Figure S4: Comparison of execution times.



3.4 Comparing dependency on the rolling probe radius

We investigated how the results of VoroContacts and dr_sasa depend on the rolling probe radius. We run the software tools on three non-similar protein structures (1C5E_A, 1EB6_A, 1G6X_A) using probe radius parameters ranging from 0.1 to 3.0 angstroms. For every probe radius we counted the number of non-zero-area contacts of the following three categories:

- per-atom solvent accessible surfaces (Figure S5)
- residue-residue contacts (Figure S6)
- atom-atom contacts (Figure S7)

As expected, the numbers of solvent accessible atoms were similar for every probe radius. However, dr_sasa overestimated the number of residue-residue and atom-atom contacts when executed with larger probe radius parameters. On the other hand, VoroContacts took advantage of the Voronoi tessellation and was much less sensitive to the probe radius increase because it did not consider atoms to be in contact when there are atoms in between them. The performed analysis also suggests that VoroContacts should produce more consistent results when executed using different van der Waals radii.

Figure S5: Number of solvent accessible atoms for various rolling probe radii.

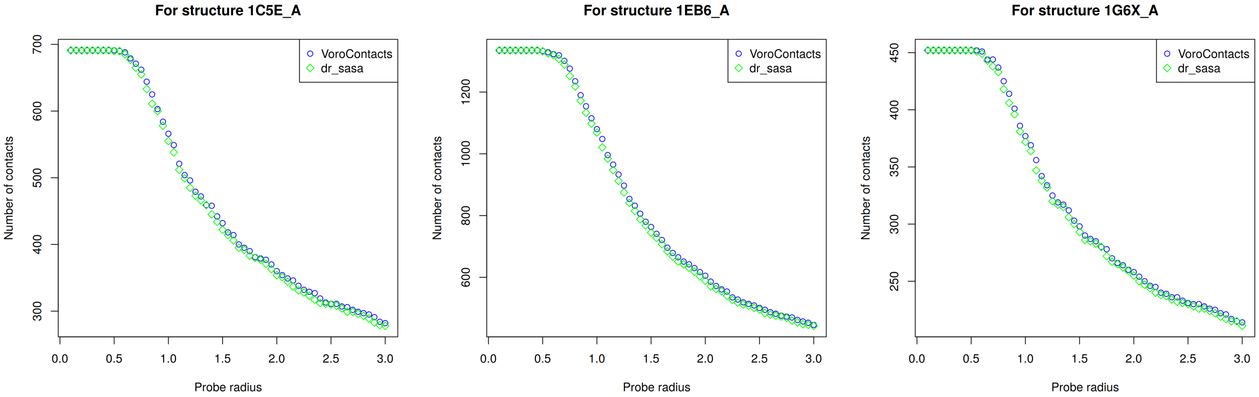


Figure S6: Number of residue-residue contacts for various rolling probe radii.

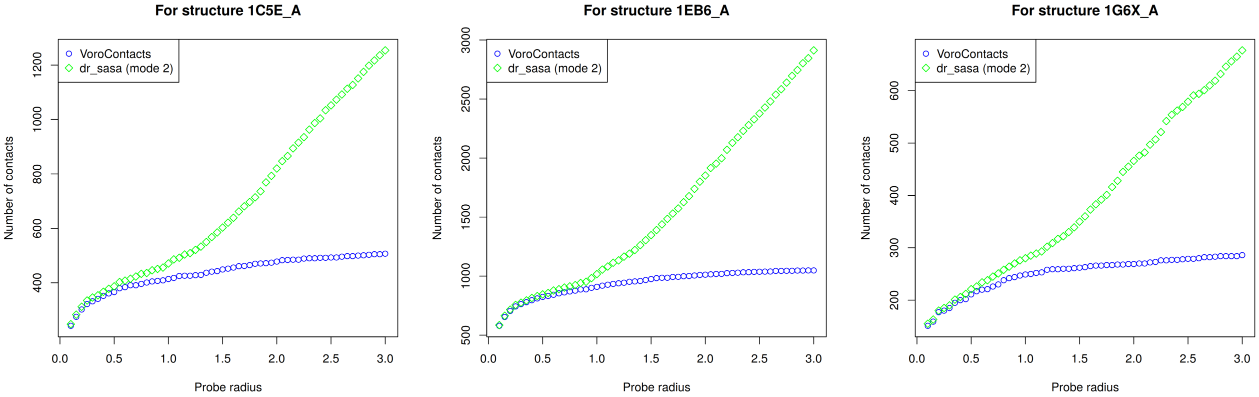
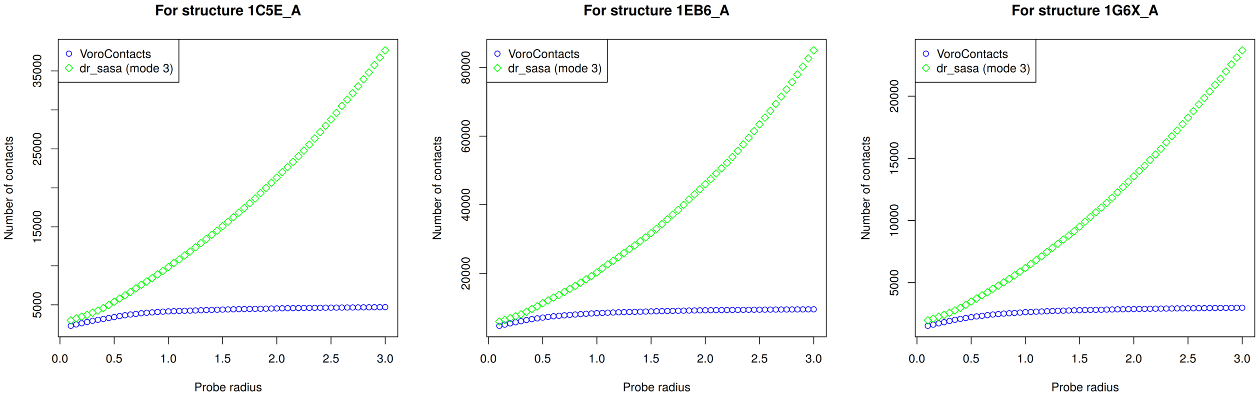


Figure S7: Number of atom-atom contacts for various rolling probe radii.



References

[1] A. J. Li and R. Nussinov, "A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking," *Proteins*, vol. 32, pp. 111–127, July 1998.

[2] D. R. Lide, ed., *CRC handbook of chemistry and physics: a ready-reference book of chemical and physical data*. Boca Raton: CRC Press, 82. ed., 2001-2002 ed., 2001. OCLC: 248346773.

[3] G. Wang and R. L. Dunbrack, "PISCES: a protein sequence culling server," *Bioinformatics*, vol. 19, pp. 1589–1591, Aug. 2003.

[4] K. Olechnovič and C. Venclovas, "Voronota: A fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls," *J Comput Chem*, vol. 35, pp. 672–681, Mar. 2014.

[5] J. Ribeiro, C. Ríos-Vera, F. Melo, and A. Schüller, "Calculation of accurate interatomic contact surface areas for the quantitative analysis of non-bonded molecular interactions," *Bioinformatics*, vol. 35, pp. 3499–3501, Sept. 2019.