On the bond graphs in the Delaunay-tetrahedra of the simplicial decomposition of spatial protein structures

Rafael Ördög[1,2], Vince Grolmusz[1,2]

[1] Protein Information Technology Group
Department of Computer Science
Eötvös University, H-1117 Budapest Hungary
[2] Uratim Ltd.
H-4400 Nyíregyháza, Hungary
E-mail: {devill, grolmusz}@cs.elte.hu

Abstract. The examination of straightforwardly definable discrete structures in nucleic acids and proteins turned out to be perhaps the most important development in our present knowledge and understanding the their form and function. These discrete structures are sequences of nucleotides and amino acid residues, respectively. Bioinformatics was born as the science of analyzing these sequences. The discretization of the biological information into easy-to-handle sequences of 4 or 20 symbols made possible the application of deep mathematical, combinatorial and statistical tools with enormous success. The tools, resulting from this process, changed our perception of genetics, molecular biology, and life itself.

Straightforward discrete structures can also be defined in the spatial descriptions of proteins and nucleic acids. The definition and examination of discrete objects, using the spatial structure of proteins instead of amino acid sequences would intercept spatial characteristics, that are more conservative evolutionary than the polypeptide sequences.

In the present work we analyze the Delaunay tessellations of more than 5700 protein structures from the Protein Data Bank. The Delaunay tessellations of the heavy atoms of these protein structures give certainly a more complex structure than the polymer sequences themselves, but these tessellations are still easily manageable mathematically and statistically, and they also well describe the topological simplicial complex of the protein.

Our main result is Table 1, describing the relation between van der Waals and covalent bonds in the edges of the Delaunay tessellation. Among other findings, we show that there is only a single one Delaunay tetrahedron in the analyzed 5757 PDB entries with more than 81 million tetrahedra, where all six edges of the tetrahedron correspond to atom-pairs in van der Waals distance, but none of them to atom-pairs in covalent distance.

1 Introduction

Recognizing the importance and decoding the rich information of polypeptide sequences of proteins and nucleotide sequences of nucleic acids were the bases of the exponential growth of the biological knowledge in the 20th century.

Beside these sequential information, numerous other discretized or discretizeable biological data sources wait to be exploited. One of these is the large, rich and reliable Protein Data Bank [1], storing the mainly crystallographical information of more than 50,000 entries (proteins and nucleic acids) today.

In our earlier work [2] we defined a certain simplicial decomposition on the heavy atoms of the protein structures in the PDB, and analyzed geometrical properties of the tetrahedra of different atomic composition.

1.1 Delaunay-Decompositions

Definition 1. *Given a finite set of points $A \subseteq R^3$, and a $H \subseteq A$ such that the points of H are on the surface of a sphere and the sphere does not contain any further points of A, then the convex hull of H is called a Delaunay region.*

Theorem 1. *Delaunay regions define a partition of the convex hull of A. If the points of A are in general position, (i.e., no five of the points are on the surface of a sphere), then all regions are tetrahedra (cf. Figure 1).*

□

We are interested in the Delaunay tessellation of the point-sets, since it is well-defined, it can be computed easily [3], and the resulting tetrahedra are as close to the regular tetrahedra as possible, in the sense that circumspheres do not contain further points from the point-set.

Figure 1 shows an example for the Delaunay tessellation on the plane.

Singh, Tropsha and Vaisman [4] applied Delaunay decomposition to protein-structures as follows: they selected A to be the set of C_α atoms of the protein, and analyzed the relationship between Delaunay regions volume and “tetrahedrality” and amino acid order in order to predict secondary protein structure. They gave the following definition:

Definition 2 ([4]). *The tetrahedrality of the tetrahedron with edge-lengths $l_1, l_2, l_3, l_4, l_5, l_6$ is defined*

$$4\left(\sum_k l_k\right)^2 \sum_{i<j} \frac{(l_i - l_j)^2}{15}$$

where l_i is the length of the i^{th} edge.

In a more recent work, Masso, Hijazi, Parvez and Vaisman [5] applied Delaunay tessellation combined with AI tools to predict residue-structure compatibility in case of point mutations of the *E. coli* lac repressor.

In an earlier work of us [2], we computed the Delaunay tessellations of all the heavy atoms of more than 5700 ”perfect” PDB entries from the Protein Data

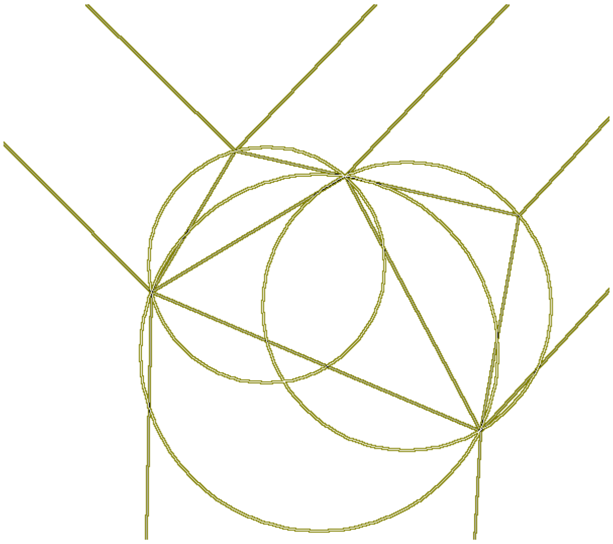


Fig. 1. *The Delaunay decomposition of 5 points on the plane. Note the empty circumcircles. The figure was created by using a Java applet, available at http://www.cs.cornell.edu/home/chew/Delaunay.html*

Bank, where "perfect" means that they contain no missing atoms in their spatial coordinate section (c.f. Figures 2 and 3 for illustrations.) Next, we examined the protein-ligand complexes by reviewing all ligand atoms in protein-ligand complexes, and characterizing the protein-atoms in the four vertices of the tetrahedra containing the ligand-atoms ([2], Tables 2 and 3). We found intricate geometrical relations in the distribution of tetrahedral vertices. We also found different volume-tetrahedrality characteristics of tetrahedra with different atomic vertex-sets ([2], Figure 3).

2 Materials and methods

In what follows $A \subseteq R^3$ is always a subset of the *heavy* atoms (i.e., non-hydrogen atoms) of a protein.

To find the Delaunay decomposition of a set, we have used the algorithm *qhull*. Its implementation source is available at: http://www.qhull.org/ [3].

For visualizing the Delaunay tessellations of protein structures, we applied the PyMol software together with the publicly available PyDet plug-in [6].

In an earlier work of us, we applied a rigorous cleaning and re-structuring procedure for the entries in the Protein Data Bank [7], and created the RS-PDB database. We made use of non-trivial mathematical, mainly graph-algorithms: Computing the InChI$^{\mathrm{TM}}$ code [8, 9] applied a graph-isomorphism testing,

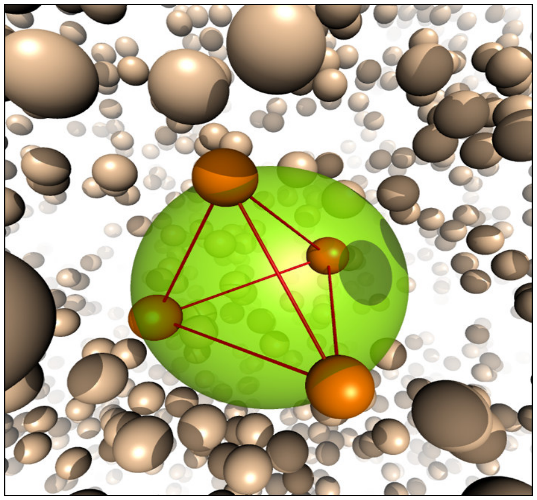


Fig. 2. *A Delaunay-tetrahedron in a spatial atom set.*

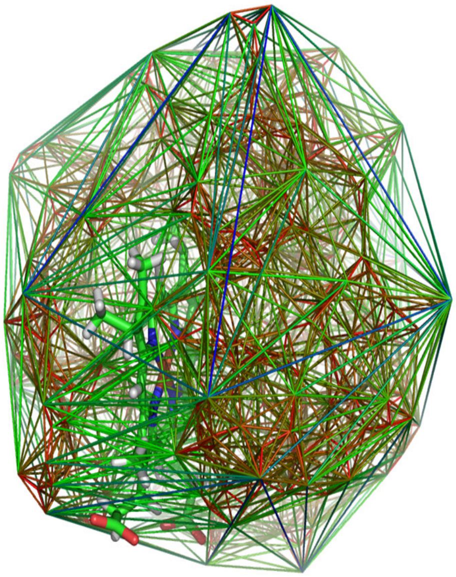


Fig. 3. *The Delaunay decomposition of the PDB entry 1n9c.*

transforming aromatic notation to Kekule-notation used a non-bipartite graph-matching algorithm [10], breadth-first-search graph traversals [11] were used throughout the work [7], depth-first search [11] was used in building the ligand molecules and identifying ring structures, kd-trees [12] were applied for computing covalent bonds, and hashing [11] were utilized for the fast generation of protein-sequence ID's.

We applied the qhull algorithm for those PDB [1] entries, that contained

- at least one protein molecule,
- with no missing atoms,
- the resolution of the structure is at least 2.2 Å.

We have found 5757 such entries in the RS-PDB database. Note, that the requirement for the missing atoms is perhaps too strict, but in this study we do not intend to deal with stability questions, i.e., the effects of the missing atoms for the whole tessellation.

In contrast with the article [4], we have taken A to be the set of heavy atoms of the protein, in the recent work as well as the in the [2]. Note that in that case we can not suppose that points are in general position, as for example in a (perfect) benzene ring at least 6 carbon atoms lie on a sphere. However, we have found that - probably due to imprecision of data in the PDB, all regions turn out to be tetrahedra.

Our present work also use this set of filtered data for characterizing the bond graphs in spatial protein structures.

3 Results: Van der Waals edges vs. covalent edges in Delaunay tetrahedra

We suggest that the Delaunay tessellation could be a discrete structure catching some deep properties of the 3D protein data. In [2], we analyzed protein-ligand complexes with the help of Delaunay tessellations.

Here we characterize the edges of Delaunay tetrahedra into three classes:

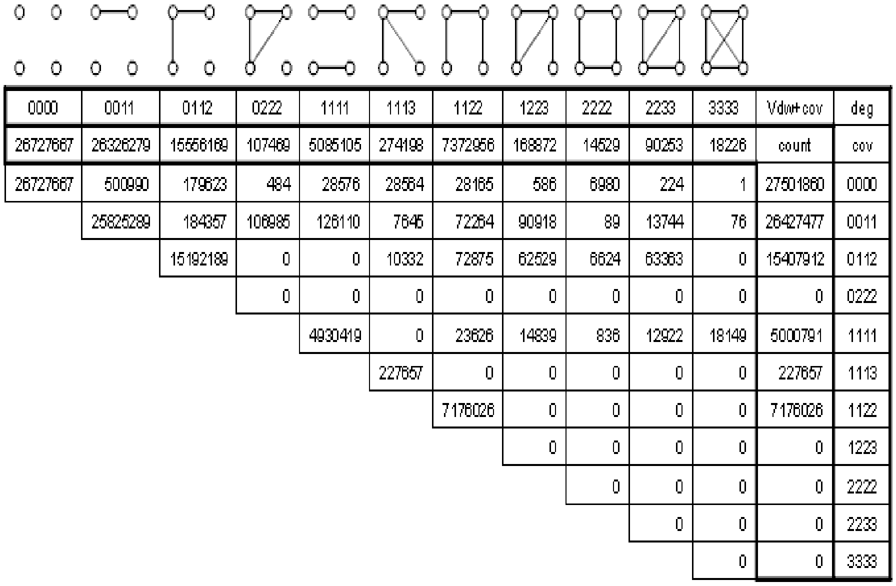
i Edges, longer than the van der Waals bond distance of the two atoms in the vertices;
ii Edges, shorter than or equal to the van der Waals bond distance of the two atoms in the vertices;
iii Edges, shorter than or equal to the covalent bond distance of the two atoms in the vertices.

Clearly, type (iii) edges are also type (ii) edges, but the reverse is not true.

Heuristically, the Delaunay tetrahedra will contain lots of atom pairs in bond distance. Our main result in the present work is Table 1.

The columns of Table 1 correspond to the graphs, with vertices correspond to the vertices of the Delaunay tetrahedra, and edges of type (ii). The rows of Table 1 correspond to the same 4-vertex graphs in the same order as in the columns, but his time with type (iii) (i.e., covalent) edges. The very first row and the last column describes the degrees of the vertices in the 4-vertex graphs.

For example, on Table 1, in the intersection of column 1111 and row 0011, the number 126,110 means that from the more, than 5 million tetrahedra with two



0000	0011	0112	0222	1111	1113	1122	1223	2222	2233	3333	Vdw+cov	deg
26727667	26326279	15556169	107469	5085105	274198	7372956	168872	14529	90253	18226	count	cov
26727667	500990	179623	484	28576	28564	28165	586	6980	224	1	27501860	0000
	25825289	184357	106985	126110	7645	72264	90918	89	13744	76	26427477	0011
		15192189	0	0	10332	72875	62529	6624	63363	0	15407912	0112
			0	0	0	0	0	0	0	0	0	0222
				4930419	0	23626	14839	836	12922	18149	5000791	1111
					227657	0	0	0	0	0	227657	1113
						7176026	0	0	0	0	7176026	1122
							0	0	0	0	0	1223
								0	0	0	0	2222
									0	0	0	2233
										0	0	3333

Table 1. *The columns correspond to the Delaunay tetrahedra where the connected vertices are in van der Waals distance, while the non-connected ones are longer than the atom-specific van der Waals distance. The rows correspond to the same graphs in the same order, but there the edges correspond to vertices in covalent distance, and the non-edges vertex-pairs in non-covalent distance. The first row and the last column contains the degree sequences of the bond-edges in the tetrahedra. The item in the intersection of a row and in a column contains the number of tetrahedra satisfying the definition both of its row and its column. Since van der Waals distances are larger than covalent distances, the lower left half of the table is empty.*

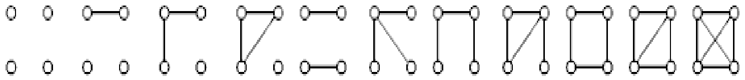
van der Waals edges with degree-sequence 1111, only 126,110 contain exactly one covalent edge.

It is worth to mention, that while there are - albeit very few - length-3 and length-4 cycles in the van der Waals graphs, there is no a single one in the covalent graphs (c.f., the highlighting in Table 2), this observation correlates well with facts from basic biochemistry.

Rows with degree-sequences 1113 and 1122 show that tetrahedra with at least 3 covalent bond edges do not admit further van der Waals edges.

Row with degree sequence 0112 shows that its intersection with column 0222 is 0, that is, a length-2 path of covalent edges prohibits a third van der Waals edge closing the path to a triangle.

On the other hand, the shaded part of Table 3 shows that almost all complete 4-vertex van der Waals graph has two non-adjacent covalent edges. There is only a single complete 4-vertex van der Waals graph without covalent bonds, depicted on Figure 4.



0000	0011	0112	0222	1111	1113	1122	1223	2222	2233	3333	vdw	deg
26727667	26326279	15556169	107469	5085105	274198	7372956	168872	14629	90253	18226	count	cov
26727667	500990	179623	484	28576	28564	28165	586	6980	224	1	27501860	0000
	25825289	184357	106985	126110	7645	72264	90918	89	13744	76	26427477	0011
		15192189	0	0	10332	72875	62529	6624	63363	0	15407912	0112
			0	0	0	0	0	0	0	0	0	0222
				4930419	0	23626	14839	836	12922	18149	5000791	1111
					227657	0	0	0	0	0	227657	1113
						7176026	0	0	0	0	7176026	1122
							0	0	0	0	0	1223
								0	0	0	0	2222
									0	0	0	2233
										0	0	3333

Table 2. *The highlights show that there are no 3- or 4-cycles in the covalent graph (it is not surprising), but there are numerous such cycles in the van der Waals graph.*

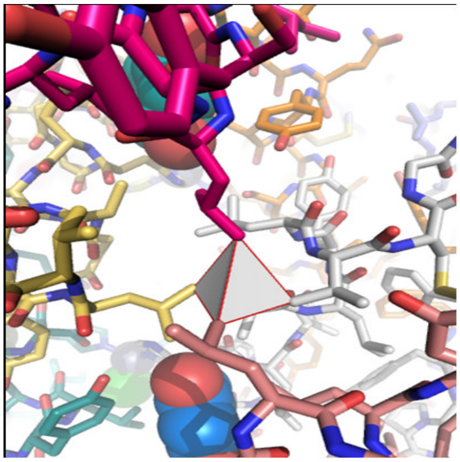
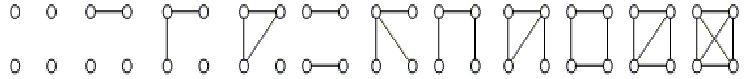


Fig. 4. *The single complete 4-vertex van der Waals graph without covalent bonds in PDB entry 1qiz. The vertices are carbon atoms, each from different polypeptide chains; more exactly, the van der Waals tetrahedron is formed from the $C_{\delta 2}$ atom of the LEU^{13} of chain E, the $C_{\delta 2}$ atom of the LEU^{13} of chain K, the $C_{\delta 2}$ of the LEU^{17} of chain L, and $C_{\gamma 2}$ atom of the VAL^{18} of chain F.*



0000	0011	0112	0222	1111	1113	1122	1223	2222	2233	3333	vdw	deg
26727667	26326279	15556169	107469	5085105	274198	7372956	168872	14629	90253	18226	count	cov
26727667	500990	179623	484	28576	28564	28165	586	6980	224	1	27501860	0000
	25825289	184357	106985	126110	7645	72264	90918	89	13744	76	26427477	0011
		15192189	0	0	10332	72875	62529	6624	63363	0	15407912	0112
			0	0	0	0	0	0	0	0	0	0222
				4930419	0	23626	14839	836	12922	18149	5000791	1111
					227657	0	0	0	0	0	227657	1113
						7176026	0	0	0	0	7176026	1122
							0	0	0	0	0	1223
								0	0	0	0	2222
									0	0	0	2233
										0	0	3333

Table 3. *The highlighted column shows that almost all complete 4-vertex van der Waals graph has two non-adjacent covalent edges. There is only a single complete 4-vertex van der Waals graph without covalent bonds.*

This single exception can be found in an interesting configuration in the PDB entry 1qiz, in a human insulin hexamer structure. As we depicted on Figure 4, the four vertices of the tetrahedron consists of four carbon atoms, each in different polypeptide chains. More exactly, from the $C_{\delta 2}$ atom of the LEU^{13} of chain E, the $C_{\delta 2}$ atom of the LEU^{13} of chain K, the $C_{\delta 2}$ of the LEU^{17} of chain L, and $C_{\gamma 2}$ atom of the VAL^{18} of chain F.

In [2], we examined the relation between the tetrahedrality and the volume of the tetrahedra in the Delaunay tessellations of the protein structures. We give here the triple logarithmic plot of the complete data set in Figure 5.

We have found that the volume-tetrahedrality relation is strongly dependent on the bond-graph of the Delaunay tetrahedra examined. Figure 6 gives the highly different triple-logarithmic color-coded plots, that is a decomposition of Figure 5, according to bond-graphs.

Acknowledgments. This research was partially supported by the NKTH grants OMFB-01295/2006 and OM-00219/2007 and OTKA grant NK-67867.

References

[1] Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., Bourne, P.: The Protein Data Bank. Nucleic Acids Research **28** (2000) 235–242

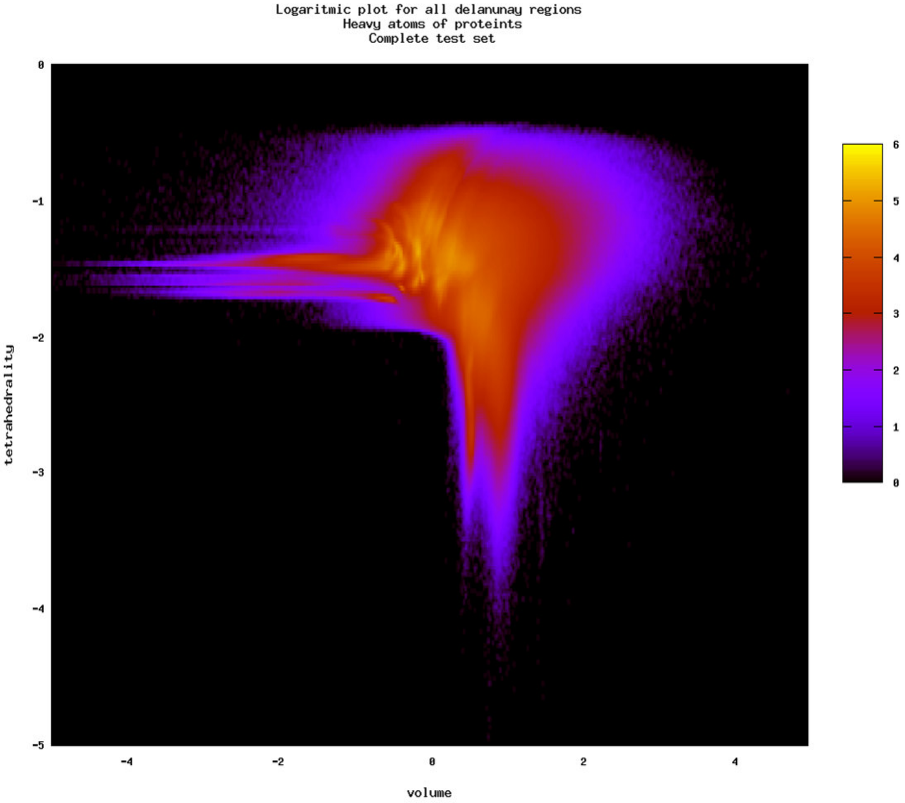


Fig. 5. *The triple logarithmic plot of the density of Delaunay regions. That is a point with coordinates (x, y) on the plot corresponds to all Delaunay regions whose volume is $10^{x\pm0.01}$ and tetrahedrality is $10^{y\pm0.01}$ and the color of the point corresponds to* $\log(z+1)$ *where z is the number of such regions.*

[2] Ördög, R., Szabadka, Z., Grolmusz, V.: Analyzing the simplicial decomposition of spatial protein structures. BMC Bioinformatics **9**(S11) (2008)

[3] Barber, C.B., Dobkin, D.P., Huhdanpaa, H.: The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software **22**(4) (1996) 469–483

[4] Singh, R.K., Tropsha, A., Vaisman, I.I.: Delaunay tessellation of proteins: Four body nearest-neighbor propensities of amino acid residues. Journal of Computational Biology **3**(2) (1996) 213–222

[5] Masso, M., Hijazi, K., Parvez, N., Vaisman, I.I.: Computational mutagenesis of E. coli lac repressor: Insight into structure-function relationships and accurate

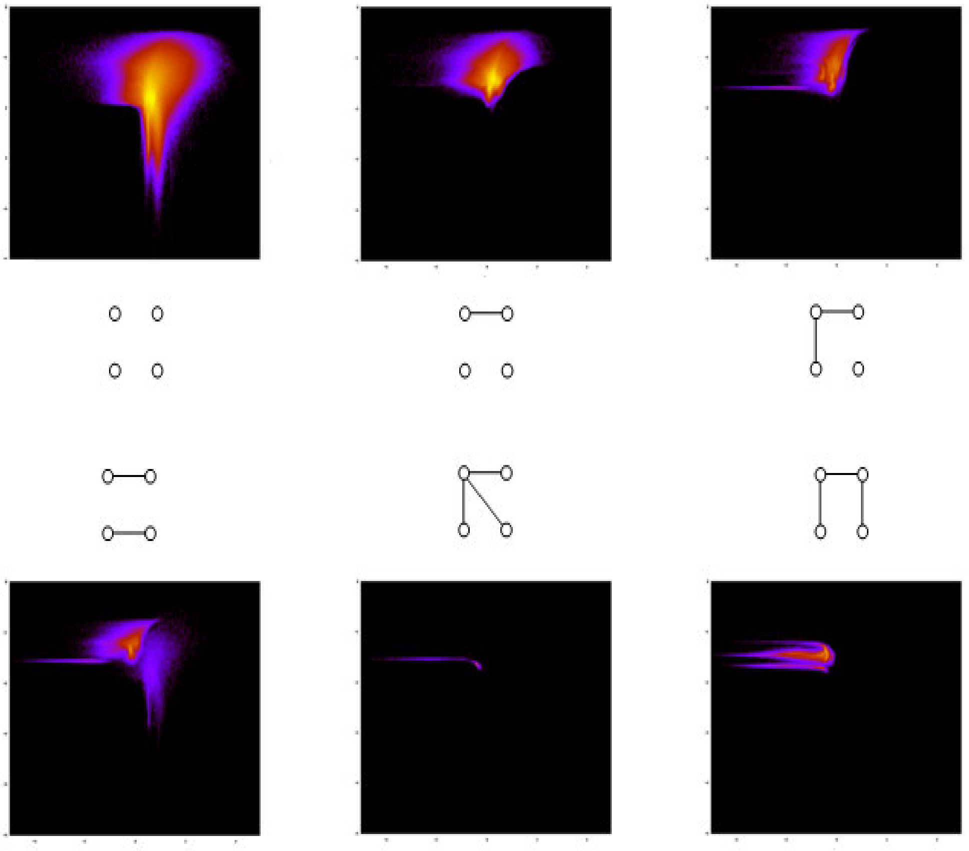


Fig. 6. *Decomposition of Figure 5 by the covalent bond-graphs. The color codes and axes labels are the same as in Figure 4.*

prediction of mutant activity. In: Proceedings of the 4-th International Symposium on Bioinformatics Research and Applications, May 6-9, 2008, Atlanta, Georgia, Springer Verlag Lecture Notes in Bioinformatics LNBI 4983. (2008) 390–401

[6] Ördög, R.: Pydet, a pymol plug-in for visualizing geometric concepts around proteins. Bioinformation **2**(8) (2008) 346–347

[7] Szabadka, Z., Grolmusz, V.: Building a structured PDB: The RS-PDB database. In: Proceedings of the 28th IEEE EMBS Annual International Conference, New York, NY, Aug. 30-Sept 3, 2006. (2006) 5755–5758

[8] Rovner, S.L.: Chemical 'naming' method unveiled. Chem. & Eng. News **83** (2005) 39–40

[9] Adam, D.: Chemists synthesize a single naming system. Nature **417**(369) (2002)

[10] Lovász, L., Plummer, M.D.: Matching theory. Volume 121 of North-Holland Mathematics Studies. North-Holland Publishing Co., Amsterdam (1986) Annals of Discrete Mathematics, 29.

[11] Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to algorithms. Second edn. MIT Press, Cambridge, MA (2001)

[12] Bentley, J.L.: Multidimensional binary search trees used for associative searching. Communications of the ACM **18**(9) (1975) 509–517