

Comparing Voronoi and Laguerre tessellations in the protein-protein docking context

Thomas Bourquard¹, Julie Bernauer², Jérôme Azé¹, Anne Poupon^{3*}

1: Bioinformatics group – Laboratoire de Recherche en Informatique, Université Paris-Sud, 91405 Orsay, France; 2: Algorithms Biology Structure, INRIA Sophia-Antipolis, Y209, 2004 route des Lucioles, BP 93, 06902 Sophia-Antipolis, France; 3: BIOS group, INRA, UMR85, Unité Physiologie de la Reproduction et des Comportements, F-37380 Nouzilly, France; CNRS, UMR6175, F-37380 Nouzilly, France ; Université François Rabelais, 37041 Tours, France.

*: to whom correspondence should be addressed anne.poupon@tours.inra.fr

Abstract

Most proteins fulfill their functions through the interaction with other proteins. Because most of these interactions are transitory, they are difficult to detect experimentally, and obtaining the structure of the complex is generally not possible. Consequently, prediction of the existence of these interactions and of the structure of the resulting complex has received a lot of attention in the last decade. However, proteins are very complex objects, and classical computing methods have led to computer-time consuming methods, whose accuracy is not sufficient for large-scale exploration of the so-called “interactome”, the ensemble of protein-protein complexes in the cell. In order to design an accurate and high-throughput prediction method for protein-protein docking, the first step was to model a protein structure using a formalism allowing fast computation, without losing the intrinsic properties of the object. In our work, we have tested two different, but related, formalisms: the Voronoi and Laguerre tessellations. We present here a comparison of these two models in the context of protein-protein docking.

1. Introduction

Recent scientific and technological progress in molecular biology and biochemistry has led to the techniques gathered under the term “omics”. These experiments tell us about the genome, the proteome, the metabolome, the lipidome, etc. of the cell. Although the techniques allowing the observation of objects, such as genes, proteins or small molecules, are very accurate, “omics” also covers a series of techniques designed to observe dynamic events, such as the complexation of two proteins.

Experimental detection of protein-protein complexes is a difficult task. For example in Yeast, which is certainly the most studied species, different techniques have been used to establish a comprehensive list of binary interactions: Yeast two hybrid (Y2H) [1-4], high-throughput co-affinity purification followed by mass spectrometry (AM/MS) [5, 6]. These lead to the detection of 18.000 to 25.000 binary interactions. Estimation of false-positive and false-negative rates of these methods leads to an estimation of 40.000 to 100.000 binary interactions in Yeast, among which 10.000 to 15.000 detected experimentally [7]. Because of the difficulties and time involved in such experimental techniques, their low-coverage (only about 20% of the interactions can be detected), and the high rate of false positives (20 to 50%), in silico prediction methods have received a lot of attention. Many different prediction methods, based on different principles have been designed: gene neighbour and gene cluster methods, phylogenetic profile methods, Rosetta stone, etc. (see [8, 9] for reviews). However, these methods also suffer from a low rate of detection, and a high rate of false positives.

Another challenge is, when the three-dimensional structures of both partners are available, the prediction of their assembly, called “docking”. The experimental determination of protein-protein complex three-dimensional structures is a very difficult and uncertain task. However, some examples have been well studied, leading to a better knowledge of the physico-chemical properties involved. Most interfaces are constituted of a buried core of mostly hydrophobic residues surrounded by a partially accessible rim of more polar residues with the overall size of about 1600 Å² [10-12]. The docking problem has also received a lot of

attention in the computational biology community. A community-wide blind experiment has been created to evaluate periodically the available methods: the CAPRI experiment [13]. This experiment has shown that some methods are now able to generate very accurate predictions in a limited number of cases. However, none of these is suitable for large-scale exploration of protein-protein complexes. Indeed, most are computationally expensive; others require the availability of biological data. More importantly, none of these was found to be accurate on all examples. A docking method usually involves two successive steps: the generation of a large number of candidate conformations, and the ranking of these candidates using a scoring function. The generation of a candidate conformation is accomplished by sampling all the possible arrangements of the two proteins, each of them constituted by a few thousand atoms. The scoring function aims at estimating the conformity of each interface relatively to known interfaces. Executing these two steps in short times, which is a requirement for large-scale studies, requires the modelling of both protein partners by simplified and computer-friendly objects. However, this model should be precise enough to retain the properties that distinguish specific from non-specific interfaces. Moreover, if the scoring function measures accurately the adequacy between a given interface and the standards, then we should be able to tell, from the score only, if a given interface does exist in vivo.

The work we present here is based on this idea. We initially used a Voronoi tessellation for modelling the protein structure, replacing each amino acid by a cell. This led to very efficient candidate generation method and scoring function [14, 15]. Each amino acid is represented by one single Voronoi cell. However, amino acids used in proteins have very different number of atoms, and consequently very different volumes, which was not accurately reflected by the volumes of the cells in the Voronoi construct. To evaluate if this discrepancy had an impact on the accuracy of the scoring function, we repeated the work using the Laguerre tessellation (also known as power diagram). In this paper we present a comparison of the results obtained with both models.

2. Material and Methods

2.1. Tessellations

Preparatory to the tessellation, the atoms of each amino acid are replaced by a node, which is the geometrical centre of the side chain plus $C\alpha$ atoms. The reasons for this choice can be found in [14]. Voronoi and Laguerre tessellations were derived from their duals, the Delaunay and regular triangulations respectively, built using the CGAL [16]. For the

Laguerre tessellation, each amino acid type was assigned a power chosen so that the volume of the Laguerre cell is as close as possible to the volume computed from an atomic Voronoi tessellation [17] (see Results).

Solvent accessibility of the residues was computed using a 1.4 Å probe.

2.2. Conformation generation

The generation of candidate conformations for a given pair of protein structures is obtained through the following steps:

- (i) The Delaunay triangulation of each partner is built;
- (ii) For each node exposed to the solvent, a normal vector is built by adding the vectors linking this node to each of its neighbours, inverting this vector in convex regions, and setting its length to 6.5 Å;
- (iii) For each pair of nodes, the two partners are placed so that the extremities of the two normal vectors are in congruence and the orientations opposed. Then a rotation along the axis of the two vectors is applied with steps of 5°, thus generating 72 candidate assemblies for each pair of nodes.
- (iv) Only candidates with an interface larger than 400 Å² are retained.

This method leads to the generation of more than 100,000 candidate conformations on average for each pair of protein structures. This number of course much depends on the size of each partner.

2.3. Learning and test sets

The learning sets were built based on a list of 187 complexes of known three-dimensional structure. This list results from a merge between the protein-protein benchmark [18] and the list of structures we used in our previous study [15], eliminating sequence redundancy using a 30% sequence identity threshold. These 187 complexes constitute the “native” set (N).

For each complex in the N set, non-native assemblies (F set) were generated using our conformation generation method. For each complex in the N set, about 10 non-native conformations (with IRMSD > 10Å, IRMSD being the root mean square deviation computed between the atoms of the amino acids belonging to the interface, as defined for the CAPRI evaluation [19]), and an interaction area larger than 600 Å² were retained for the non-native, or false, set (F).

2.4. Scoring functions

Scoring functions were learned using a genetic algorithm optimizing the area under the ROC curve, as done in previous work [14, 15]. The non-linear scoring function learned has the form:

$$S = \sum_i w_i |X_i - c_i|$$

Where for each parameter X_i , w_i and c_i are the weight and centring value respectively, optimized by the learning procedure. All functions were learned in 10 fold cross-validation, each run being repeated 9 times. Consequently, each conformation is evaluated using 9 different scoring functions, and for final ranking, the sum of the ranks obtained with each function is used.

We used 96 parameters. Among these, 84 are those we used in our previous studies [14, 15]: number of residues and size of the interface, for each type of residue frequency at the interface and mean volume of the cell, for pairs of residues at the interface (residues binned in six categories) frequencies and distance. We added 12 new parameters: the frequencies and mean volumes for the six different categories of amino acids (small, hydrophobic, aromatic, polar, positively charged and negatively charged).

We learned two different scoring functions, using both Voronoi or Laguerre constructions, using each time the relevant N and F sets.

2.5. Mean contribution of individual parameters to the scoring function

In order to compare the importance of individual parameters in the different scoring functions, we compute the mean contribution:

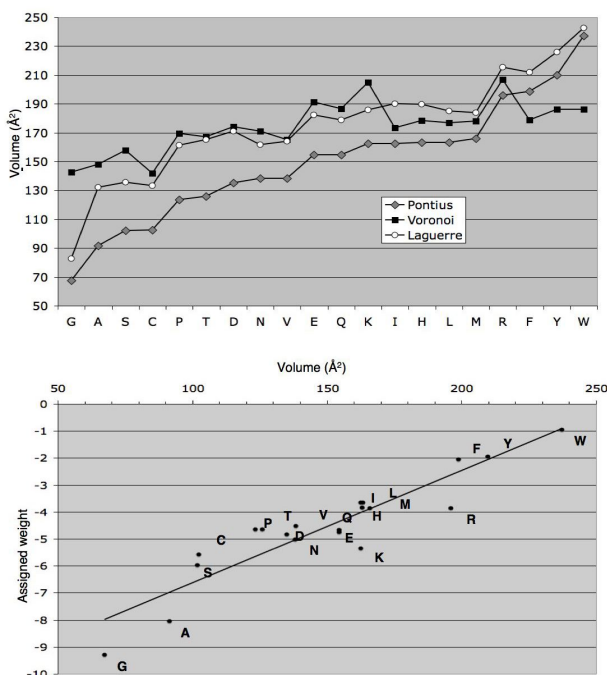


Figure 2: Normalized differences of pair frequencies (top) and pair distances (bottom) in native conformations between Laguerre and Voronoi constructs.

$$\text{cont}(P) = \sum_f w_f(P) | \langle P \rangle - c_f(P) |$$

Where $w_f(P)$ is the weight of parameter P in function f , $\langle P \rangle$ its mean value in native conformations; and $c_f(P)$ its centring value.

3. Results and Discussion

3.1. From Voronoi to Laguerre

Our initial work had been conducted using a residue-Voronoi tessellation. In this construction, all the sites, in our case all the amino acids, are considered equivalent. This modeling lead us to an accurate scoring function for protein-protein docking [14, 15], and allowed us to efficiently distinguish biological and crystallographic dimers [20]. However, the mean volumes of Voronoi cells computed using this construction are very different from the volumes that can be computed from an atomic-Voronoi construction, and considered as reference volumes (Figure 1 top). One way of getting these volumes closer to reality is to use a Laguerre tessellation.

To build the Laguerre tessellation, the first task was to assign weights to the different amino acid types. To obtain a first approximation of these weights, we first explored a range of weights from 0 to -10 for each amino acid, keeping the other 19 weights at 0. The obtained weights were then refined iteratively. The

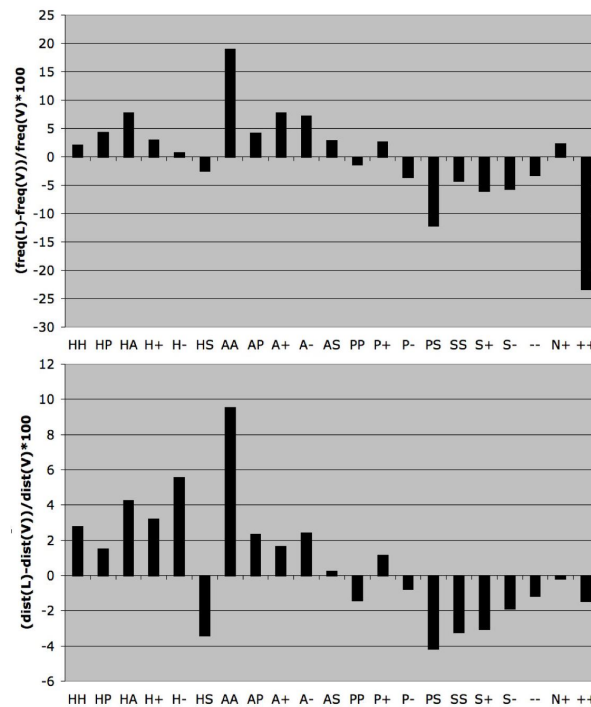


Figure 1: Amino acid weights. Left: comparison of the standard volumes with Voronoi and Laguerre cell volumes. Right: weight used in the Laguerre tessellation, the equation of the linear regression is $y=0.0417x - 10.796$.

obtained weights show a very high correlation (0.92) with the standard volumes of the amino acids (Figure 1 bottom).

The volume of the Laguerre cell is closer to the standard volume than the volume of the Voronoi for most residues (average differences are 25.27 \AA^2 and 31.02 \AA^2 respectively), and much better correlated (correlation coefficients 0.98 and 0.79 respectively). As can be seen from the right panel in Figure 1, weights might still be adjusted further. However, because of the computation time involved in each trial

(several hundred hours of CPU-time), it was decided to make further tests with these values, since, even if not perfect, they would enable us to estimate the benefits from using the Laguerre tessellation as a model.

3.2. Values of the Parameters

Having discriminating parameters is key for efficient machine learning. Consequently, the careful comparison of parameters values obtained in the different data sets (N and F, using either Voronoi or Laguerre tessellation) is an essential step.

Comparison of parameter values obtained for native

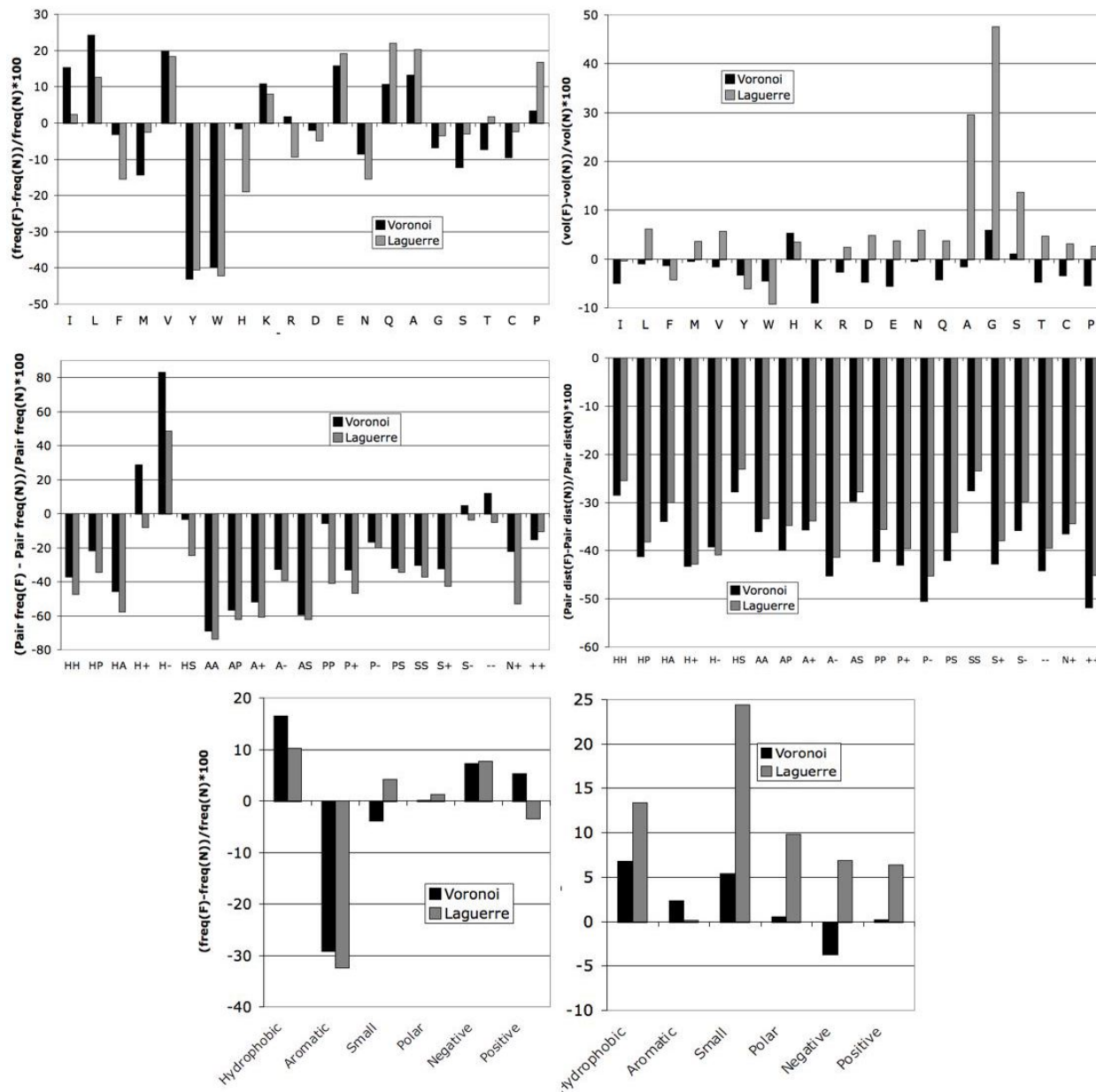


Figure 3: Normalized differences of parameter values between non-native (F) and native (N) conformations in Voronoi and Laguerre constructs.

conformations using both tessellations shows that, as expected, mean volumes for amino acids and for amino acid groups at the interface are affected. Volumes are smaller for small residues and larger for large residues in Laguerre than in Voronoi, which is a direct consequence of the chosen weights. Frequencies of individual residues, and frequencies of groups at the interface are almost not affected. Pairs involving hydrophobic or aromatic residues are more frequent, and the partner residues are closer to one another in Laguerre, whereas pairs involving polar or charged residues are less frequent, and the distance between the partner residues is larger (Figure 2).

Next, we compared the differences observed for parameters between non-native and native conformations for Voronoi and Laguerre tessellations. The differences have the same sign in Voronoi and Laguerre for all parameters that exhibit a significant difference between native and non-native. For most parameters, the differences between non-native and native in Voronoi and Laguerre are of comparable amplitude (Figure 3). However, some interesting exceptions appear. In particular, the differences in mean volumes, computed individually or for groups, are dramatically increased when the Laguerre tessellation is used. For example, the volume for glycine is found to be 50% smaller at the interface of native conformation when using Laguerre, whereas the ratio is only 5% when using Voronoi. This is probably due to the fact that the over-estimation of the volume for this residue in the Voronoi tessellation is hiding this difference. Variations can also be seen in the differences between pair frequencies at the interface, especially for hydrophobic-charged pairs.

A general comment would be that, for those parameters which behavior is different between Voronoi and Laguerre constructs, the differences between the natives and non-natives are generally larger in the Laguerre. This would tend to give a higher discriminating power to these parameters in the Laguerre construct than in the Voronoi. However, we also observed that the standard deviation for these parameters is higher in the Laguerre construct, which goes in the direction of less discriminating power. Consequently, one of the goal in this work was to determine if this balance between higher difference in values between the two categories to be discriminated (natives and non-natives) would win over larger standard deviations or not.

Another important difference we have observed concerns missing values. Indeed, for a given conformation (native or non-native), not all the parameters are filled. For example, tryptophane (W) is an amino acid present in small quantities in proteins, and many conformations will present interfaces

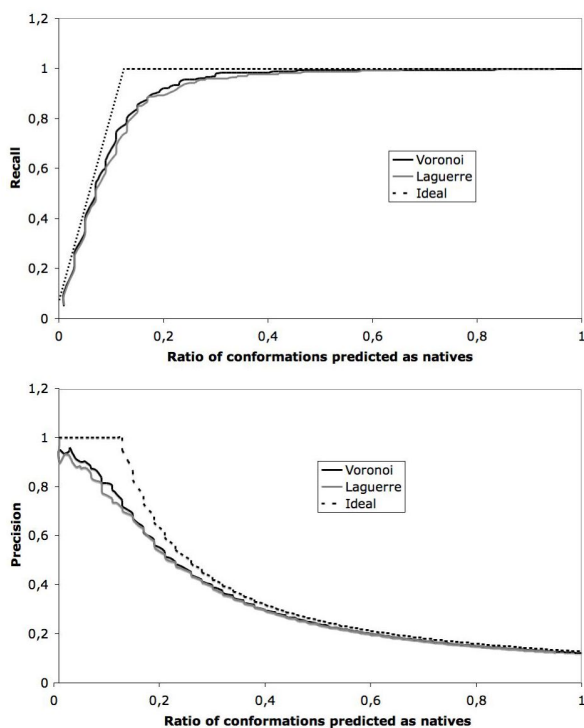


Figure 4: Recall (top) and precision (bottom) as a function of the number of conformations, ranked by value of the score, predicted as native, for SV, SL and ideal case.

containing no W residue. Consequently, the frequency of W for this interface will be 0, which is an acceptable value for this parameter. However, the mean volume at the interface for W residues won't be filled either. In our procedure, this missing value is replaced by the median value of this parameter. We have noted that, if the native conformations in Voronoi and Laguerre present comparable mean number of missing values (13.18 and 13.39 respectively), the non-native ones exhibit important differences (17.24 and 20.83 respectively).

3.3. Accuracy of the Scoring Function

Two non-linear scoring functions were optimized as described in Material and Methods:

- SV was learned on natives against non-natives using the Voronoi construct, in 10-fold cross-validation;
- SL was learned on natives against non-natives using the Laguerre construct in 10-fold cross-validation.

The precision and recall of the two functions are shown on figure 4. On this figure, we have also materialized the ideal case (green line). Since there are 18.8 native conformations in each subset (there is a total of 188 natives, and we use 10-fold cross-validation), the recall should increase from 0 to 1 on the first 19 predictions. Similarly, the precision should be 1 on the first 19 predictions, then decrease gradually, since each

conformation encountered after the 19 first ones should be non-native conformations.

As can be seen for figure 4, the precision and recall obtained in both cases (Voronoi and Laguerre) are close to the ideal case. At the 19th prediction, SV gives a precision of 0.74 and a recall of 0.78; SL gives a precision of 0.71 and a recall of 0.75. The performance of SL is in general slightly lower than that of SV. However, the ranked conformation using the SL function was a native 10 times, whereas it was a native 9 times out of 10 using the SV function. It has also to be noted that SL reaches a recall of 1 (meaning that all

the natives have been detected) sooner (at the 98th prediction on average) than SV (at the 125th prediction on average).

3.4. Contributions of parameters in the Scoring Functions

SV and SL functions exhibit very close performances. We have next compared the mean contribution of each parameter to these two functions.

Although the weight and centering values are highly variable from one function to the other, computed on the same data sets and using the same constructs, the mean contribution is very stable. Consequently, this

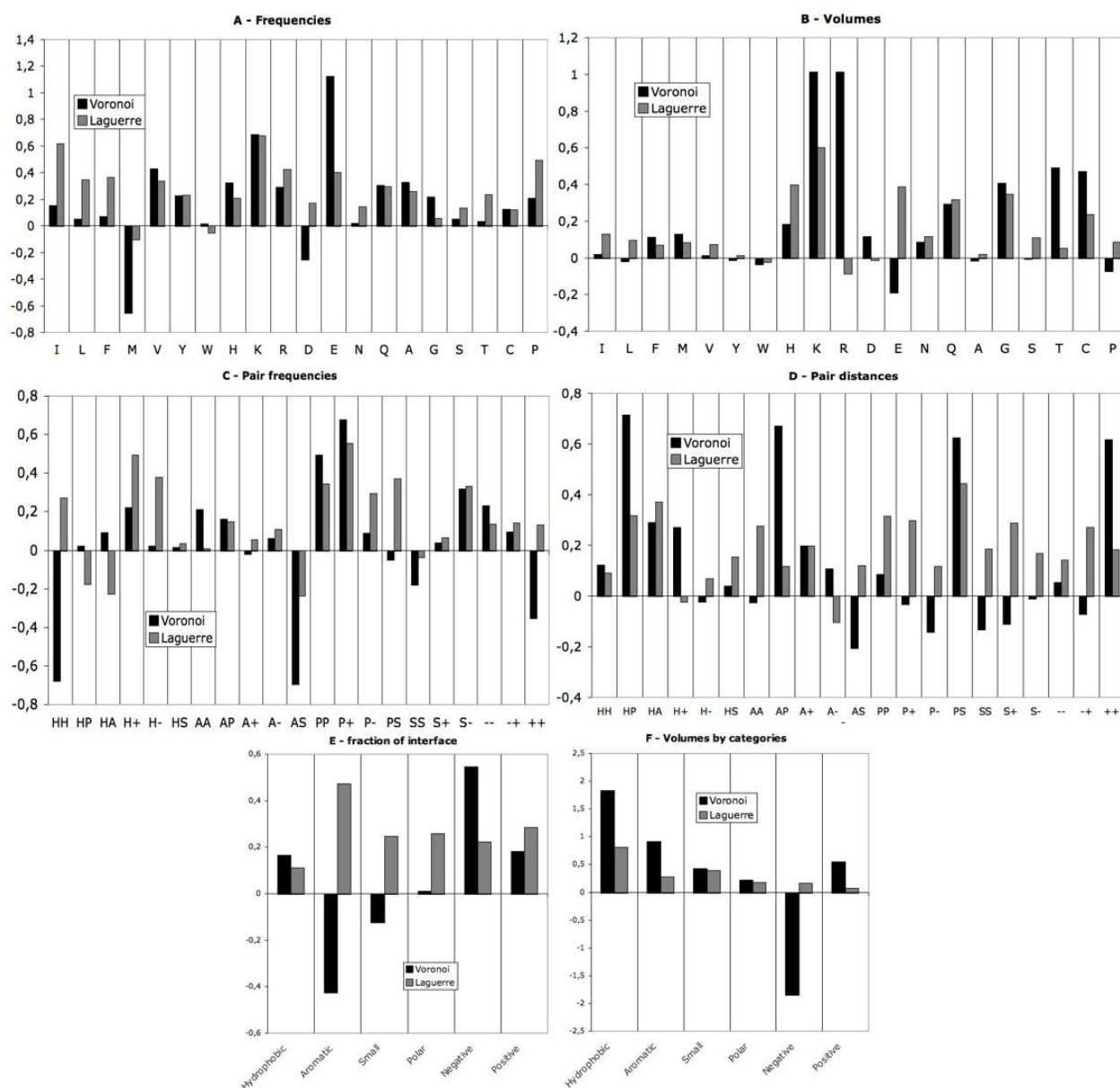


Figure 5: Mean contribution of individual parameters to the scoring functions SV and SL.

mean contribution is a good indicator of the importance of each parameter, and a good way to compare the SV and SL functions.

About half of the parameters, encountered in all categories, have comparable contributions in the two functions. The other half exhibits very different contributions. Although amino acids frequencies at the interface are found to be very close in Voronoi and Laguerre constructs, their contributions to the scoring functions can be different. For example, SL and not SV favors the presence of hydrophobic amino acids I, L and F at the interface; whereas SV but not SL disfavors the presence of hydrophobic amino acids M. The largest difference is observed for the contribution of the volume of amino acid R: the contribution of this parameter is almost 1 in SV, which makes it one of the three most contributing parameters to the scoring function, whereas its contribution is -0.8 in SL. Concerning the volumes, the amino acids which volume contribution are most different between SV and SL are not those which volume was most affected by the Voronoi-Laguerre conversion.

From Figure 5E, we can deduce that the types of interfaces favored by both functions are quite different: SV favors interfaces with hydrophobic and charged residues, but few aromatics; SL favors interfaces with more aromatics and less hydrophobic residues.

4. Conclusion

Modeling of protein structure using Voronoi tessellation had already been shown in previous studies to be a very efficient tool for the optimization of accurate scoring function able to distinguish native and non-native protein-protein complexes conformations. Because of its non-weighted nature, Voronoi tessellation does not faithfully model volumes occupied by amino acids in the protein structure. For this reason we tested the Laguerre tessellation, adjusting the individual weights such as the volumes would get closer to their standard values.

Many differences are observed with the two constructions: parameters values in native conformations; differences in the values of the parameters between native and non-native conformations. However, the two optimized scoring functions SV (Voronoi construct) and SL (Laguerre construct) exhibit very close precision and recall. Finally, the contributions of individual parameters are very different in each scoring function.

Conclusion is that the use of Laguerre tessellation instead of Voronoi tessellation does not change the performances of the scoring function. However, since these performances rely on very different elements in both cases, combining the two approaches could

certainly lead to a better recognition of native and non-native conformations.

References

- [1] Fromont-Racine, M., Rain, J., and Legrain, P., "Toward a functional analysis of the Yeast genome through exhaustive two-hybrid screens," *Nat. Genetics*, vol. 16, pp. 277-282, 1997.
- [2] Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., and Sakaki, Y., "Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins," *Proc Natl Acad Sci U S A*, vol. 97, pp. 1143-7, 2000.
- [3] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M., "A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*," *Nature*, vol. 403, pp. 623-7, 2000.
- [4] Yu, H., Braun, P., Yildirim, M., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebread, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R., and Vidal, M., "High-quality binary protein interaction map of the Yeast interactome network," *Science*, vol. 322, pp. 104-110, 2008.
- [5] Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreaux, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D., and Tyers, M., "Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry," *Nature*, vol. 415, pp. 180-3, 2002.
- [6] Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edlmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G., "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, pp. 141-7, 2002.
- [7] Aloy, P. and Russel, R., "Structural systems biology: modelling protein interactions," *Nat. Mol. Cell. Biol.*, vol. 7, pp. 188-197, 2008.
- [8] Shoemaker, B. and Pachenko, A., "Deciphering protein-protein interactions. Part I. Experimental techniques and databases," *PLOS Comput. Biol.*, vol. 3, pp. e42, 2007.

- [9] Shoemaker, B. and Pachenko, A., "Deciphering protein-protein Interactions. Part II. Computational methods to predict protein and domain interaction partners," *PLoS Comput. Biol.*, vol. 3, pp. e43, 2007.
- [10] Rodier, F., Bahadur, R. P., Chakrabarti, P., and Janin, J., "Hydration of protein-protein interfaces," *Proteins*, vol. 60, pp. 36-45, 2005.
- [11] Chakrabarti, P. and Janin, J., "Dissecting protein-protein recognition sites," *Proteins*, vol. 47, pp. 334-43, 2002.
- [12] Janin, J. and Wodak, S. J., "Protein modules and protein-protein interaction. Introduction," *Adv Protein Chem*, vol. 61, pp. 1-8, 2002.
- [13] Janin, J., "Assessing predictions of protein-protein interaction: the CAPRI experiment," *Protein Sci*, vol. 14, pp. 278-83, 2005.
- [14] Bernauer, J., Poupon, A., Aze, J., and Janin, J., "A docking analysis of the statistical physics of protein-protein recognition," *Phys Biol*, vol. 2, pp. S17-23, 2005.
- [15] Bernauer, J., Azé, J., Janin, J., and Poupon, A., "A new protein-protein docking scoring function based on interface residues properties," *Bioinformatics*, vvil 23, pp. 557-62, 2007.
- [16] Boissonnat, J.-D., Devillers, O., Pion, S., Teillaud, M., and Yvinec, M., "Triangulations in CGAL," *Comput. Geom. Theory Appl.*, vol. 22, pp. 5-19, 2002.
- [17] Pontius, J., Richelle, J., and Wodak, S. J., "Deviations from standard atomic volumes as a quality measure for protein crystal structures," *J Mol Biol*, vol. 264, pp. 121-36, 1996.
- [18] Hwang, H., Pierce, B., Mintseris, J., Janin, J., and Weng, Z., "Protein-protein docking benchmark version 3.0," *Proteins*, vol. 73, pp. 705-709, 2008.
- [19] Mendez, R., Leplae, R., Lensink, M. F., and Wodak, S. J., "Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures," *Proteins*, vol. 60, pp. 150-69, 2005.
- [20] Bernauer, J., Bahadur, R., Rodier, F., Janin, J., and A, P., "DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions.," *Bioinformatics*, vol. 24, pp. 652-658, 2008.