VILNIUS UNIVERSITY

KLIMENT OLECHNOVIČ

METHODS FOR THE ANALYSIS AND ASSESSMENT OF THE THREE-DIMENSIONAL STRUCTURES OF PROTEINS AND NUCLEIC ACIDS: DEVELOPMENT AND APPLICATIONS

Doctoral dissertation
Physical sciences, informatics (09P)

Vilnius, 2017

The dissertation work was carried out at the Department of Bioinformatics, Institute of Biotechnology, Vilnius University from 2012 to 2016.

Scientific supervisor:
Dr. Česlovas Venclovas (Vilnius University, Biomedical Sciences, Biology — 01B).

Scientific consultant:
Prof. Dr. Habil. Feliksas Ivanauskas (Vilnius University, Physical Sciences, Informatics — 09P).

VILNIAUS UNIVERSITETAS

KLIMENT OLECHNOVIČ

BALTYMŲ IR NUKLEORŪGŠČIŲ ERDVINIŲ STRUKTŪRŲ ANALIZĖS IR VERTINIMO METODAI: KŪRIMAS IR TAIKYMAS

Daktaro disertacija
Fiziniai mokslai, informatika (09P)

Vilnius, 2017 metai

Disertacija rengta 2012–2016 metais Vilniaus universiteto Biotechnologijos instituto Bioinformatikos skyriuje.

Mokslinis vadovas:
dr. Česlovas Venclovas (Vilniaus universitetas, biomedicinos mokslai, biologija — 01B).

Mokslinis konsultantas:
prof. habil. dr. Feliksas Ivanauskas (Vilniaus universitetas, fiziniai mokslai, informatika — 09P).

Acknowledgements

I am grateful to my supervisor Česlovas Venclovas for his guidance and invaluable insights into bioinformatics and science in general.

I would like to thank present and former colleagues from VU Institute of Biotechnology Department of Bioinformatics Justas Dapkūnas, Mindaugas Margelevičius, Kęstutis Timinskas, Rytis Dičiūnas, Albertas Timinskas, Darius Kazlauskas, Visvaldas Kairys, Eleonora Kulberkytė, Nerijus Verseckas and Mantas Marcinkus for their help and stimulating discussions.

My thanks also go to Andriy Kryshtafovych, Jürgen Haas, Alessandro Barbato, Rimvydas Krasauskas, Kęstutis Karčiauskas, Severinas Zubė, Janusz Bujnicki and Andrius Merkys for providing unique insights into their respective fields. I am also grateful to the scientific consultant, Feliksas Ivanauskas, and to the dissertation reviewers, Romas Baronas and Saulius Gražulis.

Finally, I would like to thank my parents for their understanding and support.

Contents

Introduction

Research area

Science builds and organizes knowledge about the universe on different scales: from galaxy clusters to subatomic particles. Being part of the universe, life also needs to be studied on different levels: ecosystems are comprised of populations of organisms, organisms are systems of organs, organs are comprised of cells, cells and cellular organelles are built from and driven by biological macromolecules.[1] There are two main classes of such molecules: proteins and nucleic acids. The latter class encompasses RNA (ribonucleic acids) and DNA (deoxyribonucleic acids). Proteins and nucleic acids are polymers; they are made up of smaller molecular units that are sequentially attached to one another in long chains. The chain sequences are encoded in genomes of organisms. Protein molecules vary greatly in size and structure and these differences allow them to perform vastly different tasks. The spatial arrangement of protein atoms is what ultimately determines how a protein functions. Nucleic acids, especially RNA, also exhibit a variety of structures and functions, although, historically, their structural aspects were less studied than those of proteins.

As of 2017, there are more than 126,000 structures of proteins, nucleic acids and their complexes that were experimentally determined and published in the Protein Data Bank (PDB),[2] the only global archive of macromolecular structural data. Figure 1 provides a glimpse of the structural variety and complexity of the PDB data, but even if it showed all the PDB structures, the picture would still be incomplete. The number of known protein structures (about 123,000) is vastly behind the number of protein sequences that are encoded in the genomes of organisms: due to the recent advances in genome sequencing, several million pro-

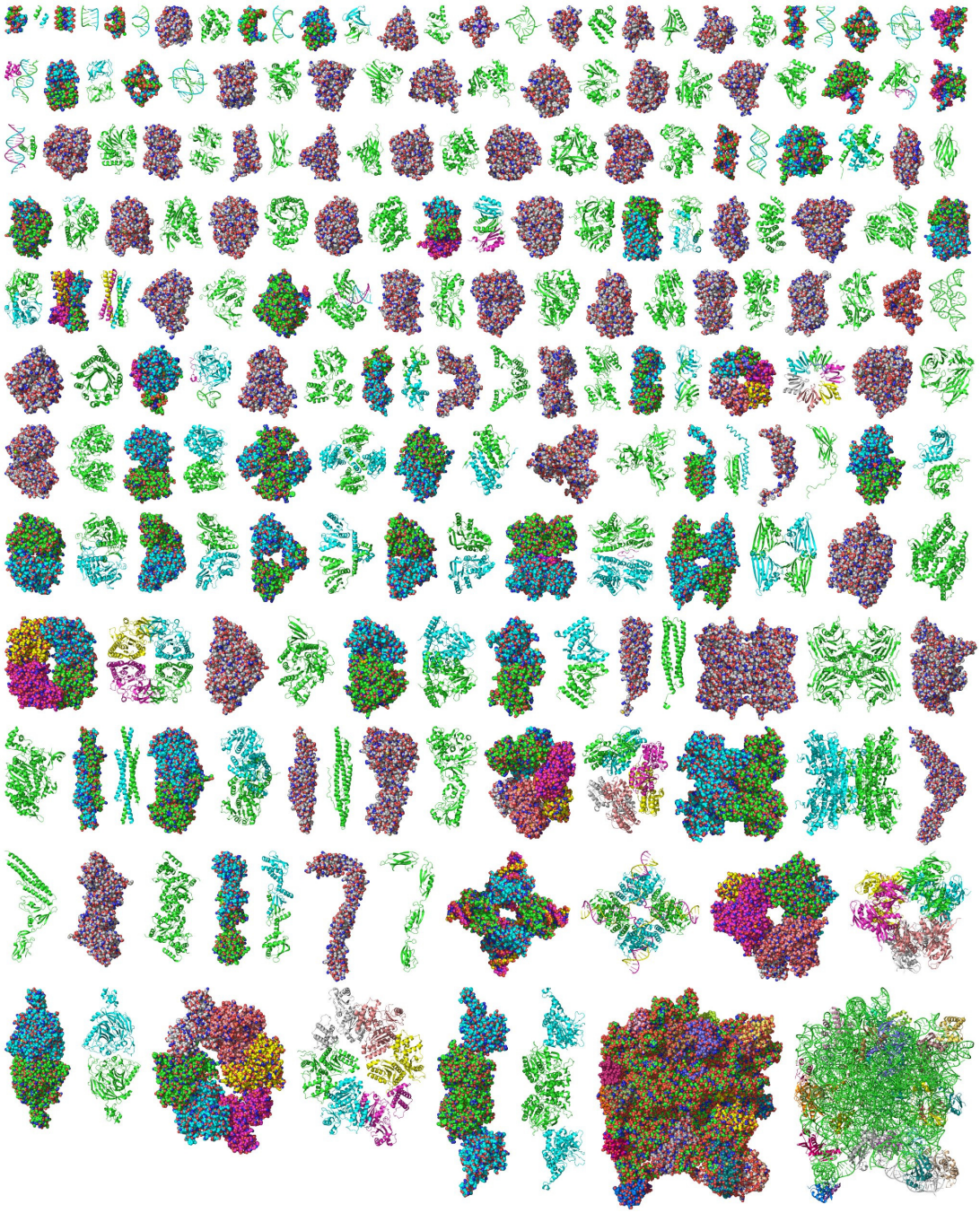


Figure 1: Random selection of 100 structures from PDB, each structure is shown in two representations (rendered with PyMol software): atomic balls and schematic cartoon. All the structures are shown in the same scale to emphasize the differences in their sizes.

tein sequences are already known and publicly available in specialized databases such as the Universal Protein Resource (UniProt).[3] As for the number of experimentally determined nucleic acid structures, it barely exceeds 3,000 (9,000 if also counting those from protein-nucleic acid complexes). In general, experimental determination of both protein and nucleic acid structures is expensive, slow and not always successful. Therefore, PDB misses a great part of knowledge about three-dimensional structures of biomolecules that is crucial for understanding and manipulating processes in living cells. Moreover, even the available structural data does not provide automatic answers, PDB structures are just large sets of three-dimensional atomic coordinates that need to be properly interpreted before being useful.

Life sciences take advantage of computer science for solving some of its main problems related to biomolecular structures: analyzing the known structures and modeling the unknown ones.[4] The problem of predicting spatial structures of biopolymers from their sequences is far from being solved for either proteins or nucleic acids, but some approaches, especially homology-based modeling, are already exceedingly useful in practice.[5] Most current structure prediction methods work in two stages:

1. Generating a set of candidate models, i.e. predicted structures.
2. Selecting the best model, i.e. the model most similar to the native (real) structure.

This dissertation is focused on the development of computational methods for testing and improving the second stage. More specifically, it focuses on the analysis and evaluation of structural models.

Research goals and tasks

The goal of this work is to develop novel better methods for solving the following interrelated problems:

1. Analysis of geometric features of biological macromolecular structures to provide a foundation for their evaluation.
2. Evaluation of model structure quality by comparing it to the reference (native) structure, i.e. reference-based assessment.
3. Evaluation of model structure quality when the native structure is not known, i.e. referenceless assessment.

The development of each method is to be accomplished by carrying out the following common set of tasks: define and describe the method; implement the method as a standalone software tool; use the developed software to perform large-scale tests using both experimentally determined and computationally predicted biological macromolecular structures; discuss the testing results; if appropriate, develop an easy-to-use web server for the method software.

Research results

The presented dissertation is a sequence of six studies carried out in 2012–2016 and published in well-reputed international journals.[6–11] The central study[6] of this work focuses on the evaluation of protein models against the native structure, which is essential for the development and benchmarking of protein structure prediction methods. Although a number of evaluation scores have been proposed before, many aspects of model assessment still lacked desired robustness. To improve the assessment we developed CAD-score, a new evaluation function quantifying differences between physical contacts in a model and the reference structure. The new score uses the concept of residue-residue contact area difference (CAD) introduced by Abagyan and Totrov.[12] Contact areas, the underlying basis of the score, are derived using the Voronoi tessellation of protein structure. The newly introduced CAD-score is a continuous function, confined within fixed limits, free of any arbitrary thresholds

or parameters. The built-in logic for treatment of missing residues allows consistent ranking of models of any degree of completeness. We have tested CAD-score on a large set of diverse models and compared it to GDT-TS,[13] a widely accepted measure of model accuracy. Similarly to GDT-TS, CAD-score showed a robust performance on single-domain proteins, but displayed a stronger preference for physically more realistic models. Unlike GDT-TS, the new score revealed a balanced assessment of domain rearrangement, removing the necessity for different treatment of single-domain, multi-domain, and multi-subunit structures. Moreover, CAD-score makes it possible to assess the accuracy of inter-domain or inter-subunit interfaces directly.

The CAD-score method uses interatomic contacts derived from the Voronoi diagram[14] of protein structure. There are several different types of the Voronoi tessellation.[15] The Voronoi diagram of balls, corresponding to atoms of van der Waals radii, is particularly well-suited for the analysis of three-dimensional structures of biological macromolecules. However, due to the shortage of practical algorithms and the corresponding software, simpler approaches are often used instead. We dedicated a special study[7] to develop a simple and robust algorithm for computing the vertices of the Voronoi diagram of balls. The vertices correspond to the centers of the empty tangent spheres defined by quadruples of atomic balls; they can be used in unequivocally defining atomic neighborhoods. The algorithm is implemented as an open-source software tool, Voronota. Large-scale tests showed that Voronota is a fast and reliable tool for processing both experimentally determined and computationally modeled macromolecular structures.

The subsequent development of the CAD-score method was its adaptation for quantifying discrepancies between RNA 3D models and reference structures.[8] A growing interest in computational prediction of ribonucleic acid (RNA) three-dimensional structure[16] has highlighted the need for reliable and meaningful methods for comparing models and ex-

perimental structures. To meet this need, we explored a possibility of using contact area-based assessment for the RNA 3D structure. Despite significant differences between proteins and nucleic acids, it turned out that in the case of RNA this approach is as efficient as it is in the case of proteins. In the same way as for proteins, CAD-score for RNA closely reflects physical interactions, has a simple definition, a fixed range of values and no arbitrary parameters. It is based on the correspondence of respective contact areas between nucleotides or their components (base or backbone). The better the agreement between respective contact areas in a model and the reference structure is, the more accurate the model is considered to be. Since RNA bases account for the largest contact areas, we further distinguish stacking and non-stacking contacts. We have extensively tested the contact area-based evaluation method and found it effective in both revealing local discrepancies and ranking models by their overall quality. Compared to other reference-based RNA model evaluation methods, the new method shows a stronger emphasis on stereochemical quality of models. In addition, it takes into account model completeness, enabling a meaningful evaluation of full models and those missing some residues.

The CAD-score method was made more accessible by developing the web server[9] that provides a universal framework to compute and analyze discrepancies between different 3D structures of the same biological macromolecule or complex. The server accepts both single-subunit and multi-subunit structures and can handle all the major types of macromolecules (proteins, RNA, DNA and their complexes). In addition to entire structures and interfaces, the server can assess structural subsets defined by contact-based queries. The CAD-score server performs both global and local numerical evaluations of structural differences, it also provides a rich set of means for interactive exploration and visualization of the results.

The CAD-score method was further developed to serve as a foundation

for the PPI3D web server[11] that is focused on searching and analyzing the structural data on protein-protein interactions. Reducing the data redundancy by sequence similarity-based and CAD-score-based clustering and analyzing the properties of interaction interfaces using Voronoi tessellation made PPI3D a highly effective tool for addressing different questions related to protein interactions.

The final study of this dissertation is focused on the referenceless estimation of the quality of predicted protein structures, which is important not just for selecting the best model in the second stage of structure prediction, but also for estimating the utility of a computational model for addressing biological questions. One of the approaches to this problem is the use of knowledge-based statistical potentials. Such methods typically rely on the statistics of distances and angles of residue-residue or atom-atom interactions collected from experimentally determined structures. In this work, a new method for the estimation of protein structure quality is presented. The method, called VoroMQA (Voronoi tessellation-based Model Quality Assessment),[10] combines the idea of statistical potentials with the use of interatomic contact areas instead of distances. Thus, VoroMQA is in large part based on the groundwork established by the development of the Voronota and CAD-score methods. Contact areas, derived using Voronoi tessellation of protein structure, are used to describe and seamlessly integrate both explicit interactions between protein atoms and implicit interactions of protein atoms with solvent. VoroMQA produces scores at atomic, residue and global levels, all in the fixed range from 0 to 1. The method was tested on the CASP data and compared to several other single-model quality assessment methods. VoroMQA showed strong performance in the recognition of the native structure and in the structural model selection tests, thus demonstrating the efficacy of interatomic contact areas in estimating protein structure quality.

Scientific novelty

The most prominent novel aspect of this work is the construction and application of interactomic contact areas, derived from the Voronoi tessellation of atomic balls, for the analysis and evaluation of biological macromolecular structures. CAD-score is the first method to use tessellation-derived contact areas for the reference-based evaluation of structural models. VoroMQA is the first method for the referenceless assessment of protein structural models that uses inter-atom contact areas, derived directly from the cells of the Voronoi tessellation of atomic balls, to provide model quality estimates on an absolute scale. The key novel aspect of the Voronota method is utilizing some common geometric properties of macromolecular structures for the efficient construction of the additively-weighted Voronoi diagram of atoms.

Practical value

The software implementations of the Voronota, CAD-score and VoroMQA methods are freely available as standalone open-source applications (and, in the case of CAD-score and VoroMQA, as web servers) from the following addresses:

- `http://bioinformatics.lt/software/voronota`
- `http://bioinformatics.lt/software/cad-score`
- `http://bioinformatics.lt/software/voromqa`

CAD-score already became one of the standard assessment methods in CASP (Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction)[17,18] and CAMEO (Continuous Automated Model EvaluatiOn)[19] experiments that periodically monitor the state of the art in the field of protein structure prediction. CAD-score

is also applicable for the analysis of structures of nucleic acids, the CAD-score web server features special modes for it. CAD-score-based PPI3D web server (`http://bioinformatics.lt/software/ppi3d`) serves predictors of protein-protein complexes by helping them find and analyze suitable templates for homology-based modeling.

VoroMQA software can be useful for anyone who creates or uses protein structural models and wishes to assess the realism of a single model or select the best model out of several.

The Voronota software package allows other scientists to create different structure analysis tools that utilize the Voronoi tessellation of balls and the related contact areas.

Propositions to be defended

- The developed method for computing the vertices of the Voronoi tessellation of balls is capable of processing macromolecular structures efficiently by exploiting common patterns of atomic spatial arrangements. The method serves as an effective tool for defining interatomic interactions, it is also easily parallelizable.

- The developed method for reference-based evaluation of macromolecular structural models avoids common problems of traditionally employed reference-based assessment methods by using Voronoi tessellation-derived contact areas. The method is universally applicable for the efficient comparison of structures of all the major types of macromolecules (proteins, nucleic acids and their complexes).

- The developed method for referenceless evaluation of protein structural models efficiently combines the idea of knowledge-based statistical potential with the concept of interatomic contact areas derived from the Voronoi tessellation of atomic balls. The method con-

sistently outperforms other statistical potential-based protein structure quality assessment methods.

Structure of the dissertation

Chapter 1 provides basic information about biomolecular structures and concise reviews of previously published works on analysis and assessment of macromolecular structural data. Chapters 2, 3 and 4 are dedicated to the detailed description of the developed methods (Voronota, CAD-score and VoroMQA) and their performance results. The subsequent chapters contain conclusions followed by the list of bibliographic references and other supporting information.

1 Literature overview

1.1 Structures of proteins

Proteins are biochemical compounds which are essential parts of all organisms and participate in every biological process.[20] Proteins have a great variety of functions which are determined by protein structure. A brief introduction to protein structure is provided below: it is stripped of many details that may be important when looking from biochemical or biophysical viewpoints, but still includes a bare minimum of information necessary for the understanding of this dissertation (the same considerations were also used when introducing nucleic acids in the subsequent section).

Proteins are polymers, each protein consists of one or more single linear chains of amino acids.[20] Amino acids are small molecules that share a common structural pattern and can be bonded together in a sequence (Figure 1.1 A). Once linked in the protein chain, an amino acid is called a residue, and the linked series of carbon, nitrogen and oxygen atoms are known as the main chain. The group of residue atoms that are not in the main chain is called the side chain. The main chain carbon atom to which the side chain connects is known as the alpha-carbon or Cα. Amino acids differ in side chains attached to Cα atoms. There are 20 standard amino acids naturally occurring in proteins. Side chains differ in chemical structure and physical properties. Each protein folds into three-dimensional structure (Figure 1.1 B-C) that is ultimately determined by the combined effect of all the interactions involving the amino acid side chains in the protein.[20,21] Therefore, protein structure is explicitly defined by amino acid sequence,[20,22] usually called protein sequence. Protein sequences are encoded in the genomes of biological organisms. Naturally occurring protein sequences can have a length from approximately 30 to thousands

of amino acids[23] (chains shorter than ≈ 30 residues also exist and are called peptides, but they usually don't have a stable structure).

An all-atom representation of a protein structure (Figure 1.1 C) is detailed, but not visually comprehensible. One of the simplified representations of a protein chain structure is a Cα-trace: a sequence of line segments connecting consecutive Cα atom centers (Figure 1.1 D). Looking at a Cα-trace, there are commonly noticeable local structural patterns, most prominently helices (called alpha-helices) and extended regions (called beta-strands): these structural elements are emphasized in one of the most popular simplified protein structure visual representations, a cartoon representation (Figure 1.1 E).

The structure of a single protein chain is commonly described as a multi-level hierarchy:[20] the primary structure (chain amino acid sequence); the secondary structure (alpha-helices, beta-strands and remaining less structured parts called loops); the tertiary structure (fully folded chain). Single-chain structures of the same or different sequences often interact and form complexes (Figure 1.1 F). Such protein-protein complexes represent the fourth level of the protein structure hierarchy — the quaternary structure.

1.2 Structures of nucleic acids

Another important class of biological macromolecules is the class of nucleic acids, which contains two major subclasses: ribonucleic acids (RNA) and deoxyribonucleic acids (DNA).[24,25] Like proteins, RNA and DNA are polymers, they are comprised of linear chains of small molecules — nucleotides. Like amino acids, nucleotides share a common structural pattern that can be sectioned into the main chain and the side chain (Figure 1.2 A). The nucleotide main chain is composed of a five-carbon sugar (ribose in RNA or deoxyribose, i.e. ribose without one oxygen atom, in DNA) and at least one phosphate group. The nucleotide side chain is

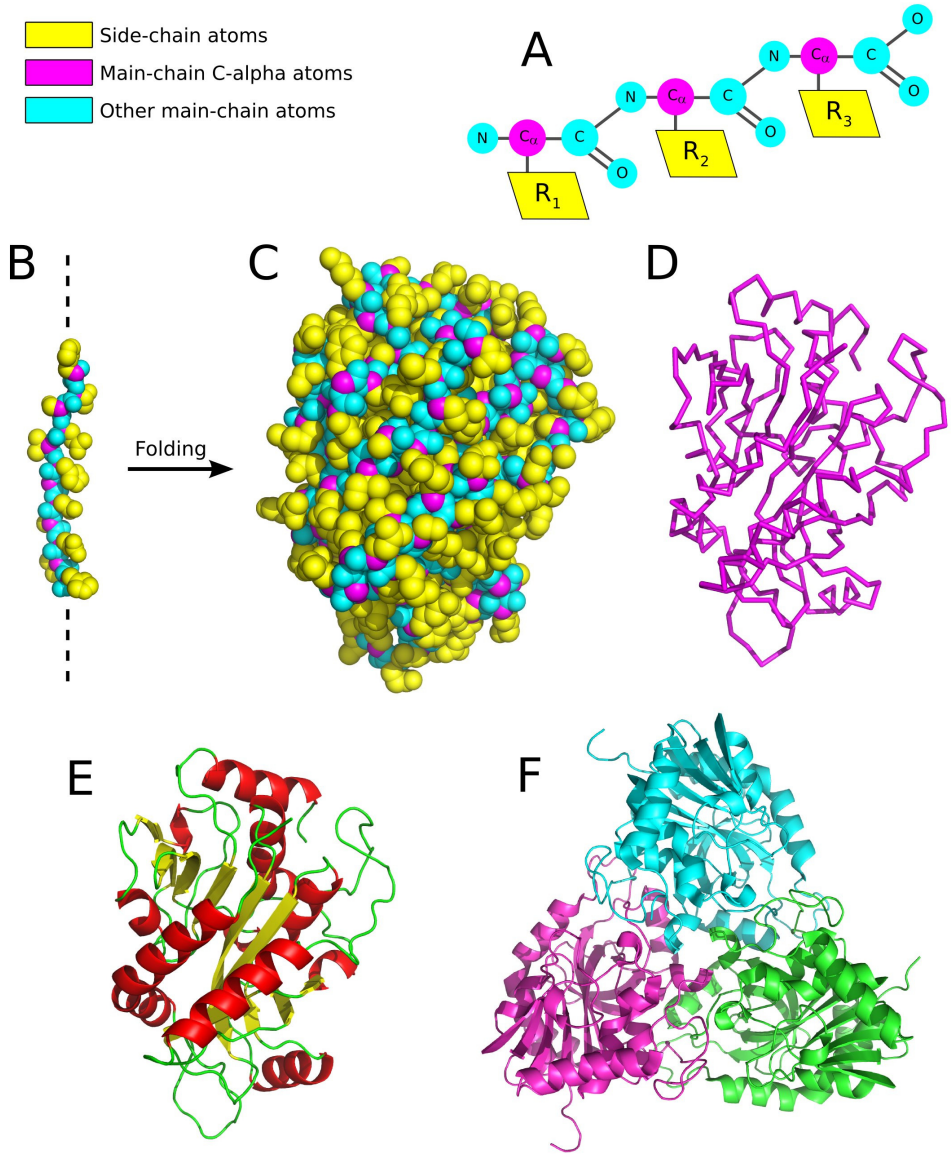


Figure 1.1: Brief introduction to protein structure, using entry 4EAR from the Protein Data Bank as an example. (A) Schematic display of amino acids forming a linear chain. Different side chains are denoted as $R_{1,2,3}$. Hydrogen atoms are omitted. (B) Part of an unfolded protein amino acid chain structure. Coloring of atoms is the same as in (A). (C) Folded amino acid chain, i.e. a final single-chain protein three-dimensional structure. (D) Cα-trace representing the folded protein amino acid chain from (C). (E) Cartoon representation of the protein structure from (C) with differently colored secondary structure elements: alpha-helices in red, beta-strands in yellow, loops in green. (F) Complex formed by three interacting chains.

called a nucleobase or, simply, a base. Different nucleotides differ by their nucleobases. There are 5 standard nucleobases: adenine (abbreviated as A), guanine (G), cytosine (C), thymine (T) and uracil (U). The first three occur naturally in both DNA and RNA, thymine is specific to DNA, uracil — to RNA.

A nucleotide sequence heavily influences the spatial structure of the nucleic acid.[24,25] Interestingly, RNA structures are exceedingly more variable than DNA structures. DNA mostly exist in a form of stable double-helical compounds, which conforms with its main function — storing genetic information. RNA can form much more intricate 3D structures because ribose has one more oxygen atom attached to it than deoxyribose. Additional oxygen atoms in RNA facilitate additional interatomic interactions that support more sophisticated structures[25] (one such structure is shown in Figure 1.2 B-C). Accordingly, the functions of RNA are of wider variety than those of DNA. As a consequence, the field of structural bioinformatics of nucleic acids is more focused on analyzing and modeling RNA structures. Another aspect of nucleic acids is that both DNA and RNA commonly interact and form complexes with proteins[26] (one such complex is shown in Figure 1.3).

1.3 Methods for the construction of the Voronoi tessellation of atomic balls

As mentioned above, proteins and RNA typically function as complex three-dimensional (3D) shapes. These shapes are determined by the combined effect of interatomic interactions both within the macromolecule itself and with the environment (e.g. water or lipid bilayer). For comprehensive understanding of these interactions it is essential to unambiguously identify all the neighbors of a given atom, to determine whether it is in contact with any of the neighboring atoms or with the environment, and how extensive these contacts are. The atomic neighborhood analysis

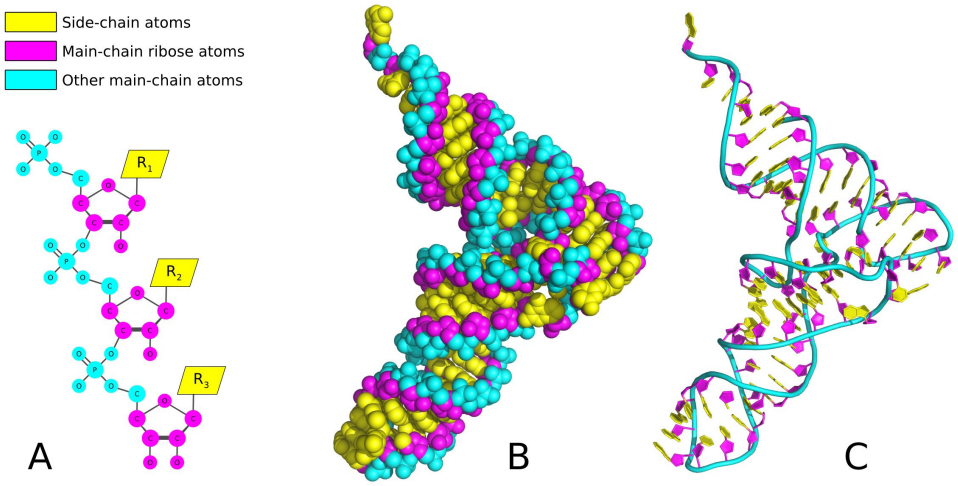


Figure 1.2: Brief introduction to RNA structure, using entry 4TNA from the Protein Data Bank as an example. (A) Schematic display of nucleotides forming a linear chain. Different nucleobases (side chains) are denoted as $R_{1,2,3}$. Hydrogen atoms are omitted. (B) Folded ribonucleotide chain, i.e. a final single-chain RNA three-dimensional structure. Coloring of atoms is the same as in (A). (C) Cartoon representation of the RNA structure from (B).

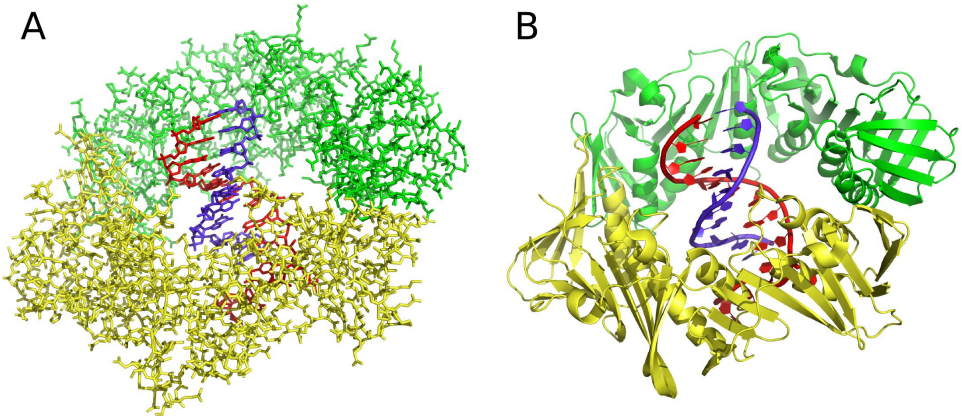


Figure 1.3: Protein-DNA complex (entry 3BEP from the Protein Data Bank), colored by chains (DNA double helix chains are colored in red and blue). (A) Atom bonds displayed as sticks, hydrogen atoms are omitted. (B) Cartoon representation of the same structure.

can also be used for studying various geometric features of 3D structure including voids, pockets and channels, for deriving molecular and solvent accessible surfaces and other geometric parameters. For these types of analyses, the Voronoi tessellation seems to be among the most suitable approaches.[27]

Voronoi diagram is named after Georgy Voronoi, who defined it back in 1908.[14] Given a set of points (centroids) in space, Voronoi diagram partitions the space into so-called Voronoi cells. The Voronoi cell may be considered as the volume "owned" by the centroid, because every point within the cell is closer to the centroid of the cell than to any other centroid. The Voronoi cell can be constructed as follows. Every line connecting a given centroid with other centroids is bisected by the plane perpendicular to that line. The smallest polyhedron formed around the centroid by such planes is termed the Voronoi cell (also known as the Voronoi region). Collectively, Voronoi cells corresponding to the set of points define the Voronoi tessellation, partitioning the space without any voids or overlaps. An important property of the Voronoi diagram is that every Voronoi cell has unambiguously defined neighbors without using any distance cutoffs.

However, the representation of protein or nucleic acids atoms as discrete points in many cases is an unacceptable oversimplification as it fails to reflect that different atoms have measurable volumes of different sizes. A more physically relevant representation of atoms is balls/spheres of van der Waals (VDW) radii and of molecules as unions of such balls. In such case the Voronoi procedure for points (or balls of the same radii) has to be modified. Richards, who was the first to apply the Voronoi method to protein structures,[28] accounted for atomic diversity by introducing VDW radius-dependent weights for positioning the separating planes. Although this method became widely used, it has a serious drawback. Namely, the separating planes no longer intersect at common points resulting in some unallocated volume between the cells. One of

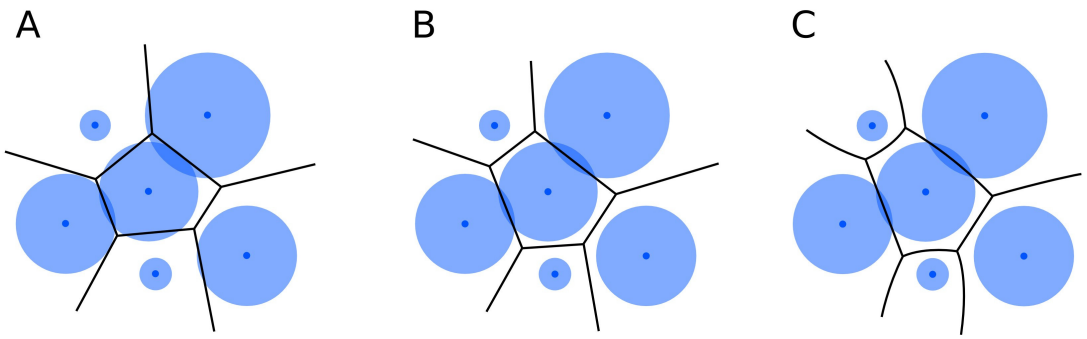


Figure 1.4: (A) Voronoi diagram of ball centroids on a plane. (B) Laguerre-Voronoi diagram of 2D balls. (C) Additively weighted Voronoi diagram of 2D balls.

the proposed solutions to this problem was to use radical plane as a separating plane between atomic balls.[29] This solution represents another weighted Voronoi scheme producing so-called Laguerre or power diagram. The advantage of the Laguerre diagram is that the cells all have flat faces making computations simpler. In addition, there is no unallocated volume in the resulting tessellation. The downside is that the weights assigned to two atoms are not directly proportional to the distance from each atom to the separating plane. This makes physical interpretation of the Laguerre tessellation problematic. Goede et al.[30] proposed a weighted Voronoi procedure resulting in a straightforward physical interpretation. In this procedure the weights assigned to atoms are linearly related to their respective distance to the dividing surface. The dividing surface is no longer a plane but a quadric surface (hyperboloid) producing Voronoi cells with faces that in general are not flat. This type of diagram is known as the Voronoi diagram of balls/spheres,[31] the additively weighted Voronoi diagram[15] or the Apollonius diagram.[32] 2D examples of ordinary, Laguerre, and additively weighted Voronoi diagrams are shown in figure 1.4.

Although the Voronoi diagram of balls is particularly well-suited for the analysis of 3D structures of biological macromolecules, so far this approach has not been utilized as widely as it might be expected. The main

reason of its limited use appears to be the shortage of efficient algorithms and the associated software tools. Therefore, in most applications, in which the Voronoi diagram of balls would be the most appropriate approach, simpler methods such as the ordinary Voronoi diagram of points or the Laguerre (power) diagram are adopted instead.

To our knowledge, there are only few algorithms available for computing Voronoi diagram of balls with the focus on structures of biological macromolecules. One of the practical algorithms applied to protein structures was proposed by Kim et. al.[31] The algorithm sequentially discovers the vertices of the Voronoi cells by tracing the edges of the cells. This algorithm was later improved by applying geometric filters for spatial search.[33,34] Medvedev et al.[35] published a similar algorithm, but it was reported[36] that the software implementing their algorithm is not suitable for typical proteins. Kim et. al[37,38] introduced an algorithm for constructing the quasi-triangulation, which is a data structure dual to the Voronoi diagram of balls. Thus, the quasi-triangulation is analogous to the Delaunay triangulation,[39] the dual of the Voronoi diagram of points. Previously, we used the Voronoi diagram of balls in Voroprot, an interactive tool for the analysis of complex geometric features of protein structure.[40] However, Voroprot was developed mainly as a visual analysis tool, not intended for batch processing or analysis of extremely large biomolecular structures.

1.4 Methods for the reference-based evaluation of protein structural models

Effective assessment of protein structural models against the experimentally determined protein structure (the reference) is at the heart of development and objective comparison of protein structure prediction methods. It may seem that one-to-one correspondence of amino acids in a model and the reference structure should make such a task trivial. How-

ever, this impression is misleading. The task is complex and, despite the fact that many evaluation scores have been devised over the years, it continues to be an active area of research.

One of the earliest and best known scores is Root Mean Square Deviation (RMSD).[41] RMSD indicates the mean distance between the corresponding atoms in the two protein structures after their optimal rigid-body superposition (Figure fig:literatureoverview-figure5). It is typically calculated for Cα atoms, but it can be applied to any subset of residue atoms. Although RMSD is a popular score, it is informative only if the differences are reasonably small and fairly equally distributed. The main disadvantage of RMSD is its sensitivity to large local deviations. Even few poorly modeled residues, which may be of little structural and/or biological importance (e.g. poorly structured protein termini or a flexible loop), may have a large impact on the resulting RMSD score. If different models include different number of residues, corresponding RMSD values may be entirely misleading as to the true accuracy of models. In particular, the inadequacy of RMSD for evaluation and ranking of very different and often incomplete protein models became apparent during early CASP experiments,[42,43] established for monitoring the state-of-the-art in protein structure prediction.

Thus, CASP experiments revealed a need for scores that would be robust in a wide range of model accuracy and completeness. Global Distance Test (GDT)[13] was one of the scores developed to overcome shortcomings of RMSD. GDT identifies the largest subset of model residues (represented by their Cα atoms) that can be superimposed with the corresponding residues in the reference structure under specific distance threshold. The overall model accuracy is summarized by GDT Total Score (GDT-TS), a single value derived by averaging the fractions of residues obtained in the four independent superpositions under 1, 2, 4 and 8 Å distance thresholds.[44] Due to multiple superpositions of different stringency, GDT-TS is able to rank models quite effectively in a wide range

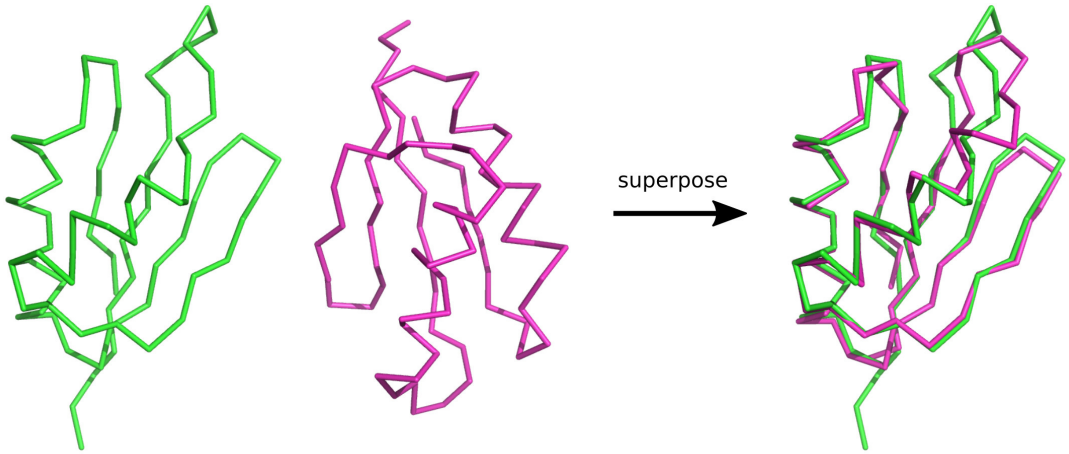


Figure 1.5: Example of a rigid-body superposition of two proteins structures, shown with Cα-traces.

of accuracy. Unlike RMSD, GDT-TS rewards the good bits of the model without adding a penalty for the inaccurately modeled regions. As a result, GDT-based benchmarking promotes methods that attempt to construct not only the most accurate, but also the most complete structural models. Other scores, similar to GDT-TS, include MaxSub[45] and TM-score.[46] MaxSub, just like GDT-TS, aims at identifying the largest subset of residues that can be superimposed under specific distance threshold. However, in contrast to GDT-TS, MaxSub uses only a single 3.5 Å distance threshold. This makes MaxSub somewhat less robust in ranking models, in particular those of lower accuracy.[46] TM-score considers all the corresponding residue pairs. It uses the distance-dependent weighting scheme, which reduces the contribution from significantly deviating residue pairs. In addition, the distance-dependent down-weighting varies with the protein size, making the score less size-dependent in comparison with either GDT-TS or MaxSub. Yet, similarly to MaxSub, TM-score is derived from a single superposition. When size-dependence is not an issue (e.g. evaluating models against the same reference) multiple superpositions as implemented in GDT-TS offer an obvious advantage.

Not surprisingly, GDT-TS has *de facto* become the central score in the au-

tomated reference-based model evaluation during CASP experiments.[47] However, despite its common use, GDT-based scoring is not without weaknesses. Since GDT-TS is based on the rigid-body superposition, it performs poorly on multi-domain proteins. A slight change in the mutual domain orientation may be biologically irrelevant, yet it may strongly affect the GDT-TS score. Another GDT weakness is that it uses only Cα atoms and therefore lacks information about the correctness of residue side chain modeling. However, this is an important component in benchmarking high accuracy comparative modeling or protein structure refinement methods. One additional and perhaps the most disconcerting issue is the lack of direct relationship between the GDT-TS score and the physicochemical characteristics of a protein model. A model having unrealistic features such as extensive interatomic clashes or systematic structural distortions may still receive a favorable GDT-TS score.[48–50] The same limitations are characteristic of similar scores, MaxSub and TM-score.

In attempt to address some of these issues, a number of modifications to the GDT-TS score have been proposed. Some of them were directed at a better resolution of higher accuracy models. Thus, GDT-HA,[51] a more stringent version of GDT-TS, uses distance thresholds half the size of those for GDT-TS. GDC, another modified score, is capable of including different thresholds and different subsets of residue atoms.[52,53] To make the score mindful of steric clashes, the inclusion of repulsion term into GDT-TS was proposed.[48] However, each modification addresses only one of several limitations of the GDT-TS score.

Therefore, there is a clear need to have the best features of GDT-TS (robustness over the wide model accuracy range and the ability to compare models of different degree of completeness) combined with a more physically meaningful representation of protein structure. Globular proteins fold into specific 3D structures that are defined by residue-residue interactions, which are reflected by physical contacts. Therefore, it seems that

contacts might be well-suited for quantifying deviations in a model with respect to the reference structure. Besides, the comparison of contacts does not require structure superposition with all the associated caveats. Indeed, a number of scores that use the concept of residue-residue contacts have been proposed.[50,53–56] However, typically, "contacts" in these scores are represented by distances between Cα, Cβ or all atoms within the arbitrarily specified threshold. Obviously, the physical meaning of contacts in such scores is lost. If only a single atom per residue (e.g. Cα) is used, important structural details are lost as well.

An interesting idea of using the explicit description of physical residue-residue contacts for model evaluation was introduced by Abagyan & Totrov.[12] They proposed to use the residue-residue contact area as the basis for comparing a model and the reference structure. Furthermore, they introduced a single-number score, contact area difference (CAD), as a measure of the overall model accuracy. CAD, as defined by Abagyan & Totrov, has a number of appealing features. It is continuous and threshold-free, works in a wide range of model accuracies, adequately penalizes domain, fragment and side-chain rearrangements and captures essential geometrical characteristics of protein structure.[12]

However, the original CAD has some properties that make its use for evaluation of methods on a large scale (e.g. CASP experiments) problematic. First, CAD considers only residues common for both the model and the reference structure. It means that a complete model would be evaluated against the complete reference structure, while a modeled short fragment would be evaluated against the corresponding reference fragment. In other words, the exact choice of the reference depends on the completeness of the model. This can hardly be considered an objective mode for benchmarking different methods. Second, the normalizing CAD term includes inter-residue contact areas not only of the reference structure but also of the model. Although this is not expected to have a large impact on the total score, it nevertheless makes the CAD normal-

ization model-specific.

1.5 Methods for the reference-based evaluation of RNA structural models

In recent years the repertoire of known biological functions that RNA performs in the cell has greatly expanded.[57] Many of these different functions are performed by RNA molecules or their regions adopting complex three-dimensional (3D) structures. Not surprisingly, the interest in RNA 3D structure has also increased considerably. However, the determination of RNA 3D structure using experimental approaches such as X-ray crystallography or NMR remains a formidable challenge. Therefore, computational RNA structure prediction methods are rapidly gaining importance.[16] A critical component in both the development and comparison of such methods is the ability to evaluate computational models against the experimentally determined reference structure. Only through the effective reference-based model evaluation, one can hope to obtain useful comparison of the performance by different methods. Moreover, the quantitative data regarding discrepancies between models and corresponding reference RNA structures can provide much-needed guidance to methods developers. Therefore, the progress in RNA 3D structure prediction is tightly coupled with the availability of both informative and objective scores that quantify discrepancies between modeled and experimental structures.

The best-known score for measuring the differences between two 3D structures is RMSD,[41] which has several major drawbacks discussed in the previous section 1.4. For protein structure analysis, the recognition of RMSD shortcomings has recently led to introduction of several alternative scores. Global Distance Test (GDT)[13] (described in section 1.4) is one of the scores in the protein field adopted for RNA.[58–60] However, the representation of a residue by a single atom (Cα for proteins and C3'

for RNA), while appropriate for proteins, seems to be too coarse-grained for RNA. Moreover, GDT-TS distance cutoffs selected to be meaningful for protein models may not be optimal for RNA. Several new scores, including Interaction Network Fidelity,[61] Deformation Index,[61] Deformation Profile[61] and RNAlyzer,[62] have been developed specifically for RNA 3D structure. Interaction Network Fidelity (INF) compares how closely base pairing and base stacking interactions within the reference RNA 3D structure are reproduced in a model.[61] Deformation Index (DI) is RMSD adjusted by INF and has been introduced as an attempt to improve RMSD properties on RNA models. The other two new scores, Deformation Profile (DP) and RNAlyzer are also based on RMSD. DP highlights dissimilarities between a model and the reference structure at the nucleotide resolution. RNAlyzer works by comparing how well corresponding local neighborhoods in the reference structure and a model agree with each other.[62] These new RNA-specific scores significantly expand the list of available model evaluation methods. However, it should be noted that they all, except INF, are based on RMSD and, therefore, inherit at least some of its drawbacks.

1.6 Methods for the referenceless assessment of protein structure model quality

The ability to predict protein three-dimensional (3D) structure from sequence is one of the most important and challenging problems in computational biology. Protein structure prediction methods tackling this problem are being developed continuously and in many cases they can produce models that are close to the native structure. The performance of such methods is systematically assessed during community-wide CASP experiments[17,18] that not only reveal successes, but also point out the bottlenecks in the field of protein structure prediction. One of the most prominent bottlenecks is the model quality assessment (QA). Current

structure prediction methods typically produce multiple models for a given protein, and then QA methods are used to identify the best model and to estimate how realistic the model is. However, according to the results of recent CASP experiments,[63,64] model quality assessment remains a difficult task and there is a clear need for better QA methods.

There are two major classes of QA methods: multi-model and single-model.[63] A multi-model method evaluates a model by quantifying how well do the structural features of the model correspond to the consensus of the structural features of a diverse ensemble of other models. A single-model method does not rely on additional models for the assessment of a single structure. Therefore, single-model methods are particularly well-suited for the practical use outside of CASP-like settings. Some of the most successful single-model QA methods, e.g. ProQ2[65] and QMEAN,[66] are meta-methods that combine several sources of information about an input structure. Such meta-methods often employ machine learning techniques to produce a single generalized quality score out of several lower-level scores such as the estimates of free energy and agreement scores that tell how well some of the observed structural features, such as secondary structure and residue solvent-accessibility, correspond to the sequence-based predictions. A viable approach for creating a better QA method is designing better techniques to combine available scores, another approach is to design better independent scores that perform well on their own or become useful components of meta-methods.

Prominent examples of independent QA methods are knowledge-based statistical potentials. Over the last twenty years or so a number of different statistical potentials have been developed. Most of them rely on statistics of pairwise interaction distances,[67–72] some also utilize information about interaction angles.[73–75] However, distance-based metrics may not necessarily be best-suited for the description and analysis of physical properties of protein structure. A possible alternative approach is to use interatomic contact areas. The first attempt to employ contact areas

as a foundation for knowledge-based potentials was made over a decade ago by McConkey et al.[76] Contact areas in the McConkey method are derived from the Voronoi tessellation of atomic centers. Voronoi and related tessellation methods proved to be an effective means in the analysis of various structural features,[27,28,40,77,78] including the identification of physical contacts that could be utilized in deriving distance-based statistical potentials.[79–81] However, to the best of our knowledge, the study by McConkey et al. so far has been the only QA method based on tessellation-derived contact areas. Their method achieved respectable results in discriminating native protein structures from decoys; however, perhaps mainly due to the lack of publicly available software implementations, the prospects of applying contact areas for the assessment of protein structural models remained largely unexplored.

2 Voronota: a method for computing the vertices of the Voronoi diagram of atomic balls

Voronota is a simple yet efficient algorithm and the corresponding open-source software for computing the vertices of the Voronoi diagram of 3D balls. The algorithm can be applied to 3D structures of various biological macromolecules including proteins, nucleic acids, protein-protein and protein-nucleic acids complexes. The computed Voronoi vertices can be used in unequivocally defining atomic neighborhoods, describing internal cavities in molecular structures or constructing edges and faces of Voronoi cells of atoms. Here, we provide a detailed description of the algorithm, then we describe the software implementation and provide large-scale tests illustrating its speed and robustness. In addition, we compare the performance of our software with the performance of QTFier (`voronoi.hanyang.ac.kr/software.htm`) and awVoronoi (`sourceforge.net/projects/awvoronoi`) that, to the best of our knowledge, are the only other publicly available tools that include similar functionality.

2.1 Method description

2.1.1 The Voronoi diagram of 3D balls and the corresponding Voronoi vertices

Let $B = \{b_1, b_2, \ldots, b_n\}$ be a set of balls, where $b_i = \langle c_i, r_i \rangle$ is a ball with a center $c_i \in \mathbb{R}^3$ and a radius $r_i \in \mathbb{R}_0^+$. A signed distance $d(p, b_i)$ from a point

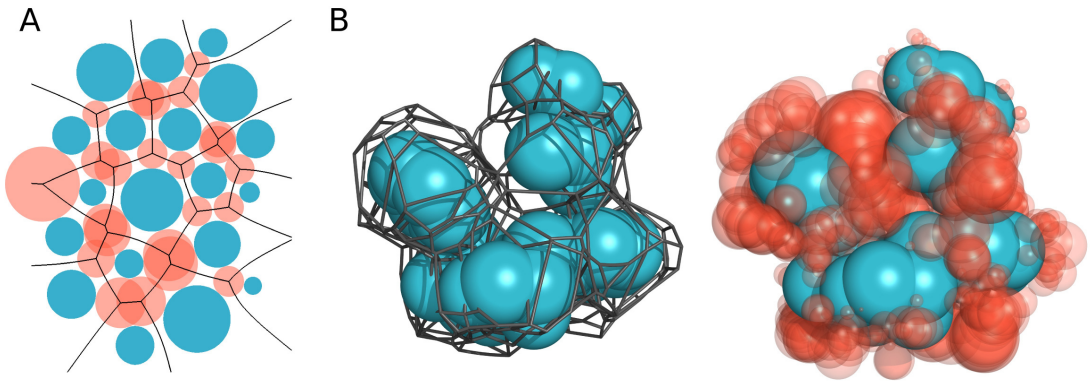


Figure 2.1: (A) Voronoi cells of 2D balls (blue) and the empty tangent spheres (red) corresponding to the Voronoi vertices. (B) Edges of the Voronoi cells of 3D balls (left) and the empty tangent spheres corresponding to the Voronoi vertices (right).

$p \in \mathbb{R}^3$ to a ball b_i is defined as follows:

$$d(p, b_i) = \|p - c_i\| - r_i \tag{2.1}$$

The Voronoi cell V_i for a ball b_i is a region containing all points closest to b_i:

$$V_i = \{p \in \mathbb{R}^3 | d(p, b_i) \leq d(p, b_j), \forall b_j \in B \setminus b_i\} \tag{2.2}$$

A set $\{V_1, V_2, \ldots, V_n\}$ is the Voronoi diagram for B. Figure 2.1 contains examples of the Voronoi cells of balls. Two balls are considered to be neighbors if their Voronoi cells intersect. The intersection of four Voronoi cells defines a point termed the Voronoi vertex. It is the center of an empty sphere tangent to the four neighboring balls (Figure 2.2 A). Notably, some Voronoi cells of balls may have no vertices – such situations are analyzed separately.

2.1.2 Outline of the algorithm for finding the Voronoi vertices

Given an input set of balls B, our goal is to find the quadruples of balls that define all the vertices of the Voronoi diagram for B. In other words,

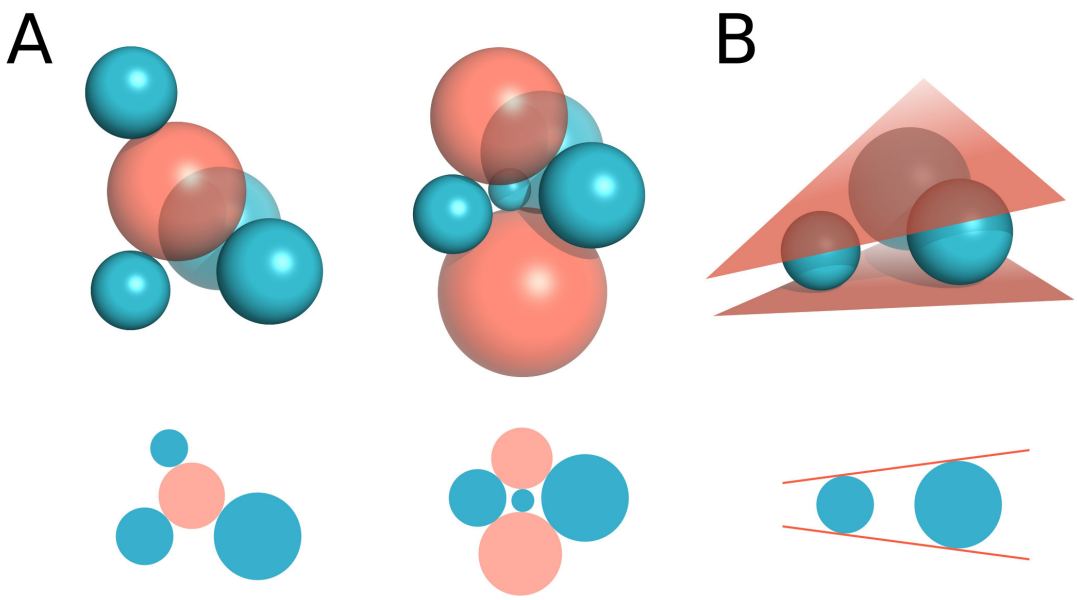


Figure 2.2: (A) Quadruples of 3D balls having either one (left) or two (right) tangent spheres. (B) A triple of 3D balls having two tangent planes. Diagrams below each 3D example show corresponding similar cases in 2D space.

we search for quadruples of balls that have at least one tangent sphere, which does not intersect with any ball from the input set. We term such quadruples valid. Any quadruple is a union of exactly four different triples, e.g. $\{a, b, c, d\} = \{a, b, c\} \cup \{a, b, d\} \cup \{a, c, d\} \cup \{b, c, d\}$. We call a triple valid if it is a subset of a valid quadruple. Starting with a single valid triple, we can discover valid quadruples by finding valid neighbors for previously detected valid triples. This principle, commonly known as "gift wrapping",[82] is used in algorithms for both the construction of the Delaunay triangulation of points[83,84] and for the construction of the quasi-triangulation of balls.[31,35] We exploit the same principle, but use a different take on searching for valid triples and their neighbors.

In Procedure 1 we implement the "gift wrapping" strategy with an important modification: we take into account that a network of valid quadruples may be disconnected.[85] This is achieved by having two "while" cycles. The inner cycle (starting at line 8) finds as many quadruples as possible starting from a valid triple. The outer cycle (starting at

line 6) runs while there still are valid triples containing balls that are not part of any of the already found quadruples.

In the next sections we explain the algorithm in detail. To begin, we briefly describe the technique we use for computing tangent spheres. We then define the two complex subprocedures incorporated into Procedure 1: finding the first valid triple (lines 5 and 22) and finding all neighbors for a valid triple (line 11). Both subprocedures utilize the same technique for efficient searching in a large set of balls, which is also described later in the text.

Procedure 1 Find valid quadruples

input: $B =$ (a set of balls)
output: $Q =$ (a set of valid quadruples for B)

```
 1: Q ← (an empty set for found quadruples)
 2: T ← (an empty set for processed triples)
 3: M ← (an empty map to associate triples with sets of their neighbors)
 4: stack ← (an empty stack for triples)
 5: t_f ← (for B, find a first valid triple)
 6: while t_f ≠ ∅ do
 7:    push(stack, t_f)
 8:    while stack is not empty do
 9:       t ← pop(stack)
10:       T ← T ∪ t
11:       X ← (for B, find a set of all neighbors of t, excluding M[t])
12:       for all x ∈ X do
13:          q ← (a quadruple from t and x)
14:          Q ← Q ∪ q
15:          T_q ← (a set of all triples from q)
16:          for all t_q ∈ T_q do
17:             x_q ← (a neighbor of t_q in q)
18:             M[t_q] ← M[t_q] ∪ x_q
19:             if t_q ∉ T then
20:                push(stack, t_q)
21:    U ← (detect balls not included in any q ∈ Q)
22:    t_f ← (for B, find a first valid triple containing any u ∈ U)
23: return Q
```

2.1.3 Computing tangent spheres

To compute a tangent sphere for four balls $\{b_1, b_2, b_3, b_4\}$ (examples shown in Figure 2.2 A) we use the method proposed by Gavrilova and Rokne.[86] Let us assume without the loss of generality that b_4 has the smallest radius. We reduce the radii of all the four balls by the radius of b_4. We then move the balls such that b_4 coincides with the origin. This allows us to define an easily solvable system of equations for finding the coordinates and the radius of the tangent sphere. This system can have none, one or two solutions. After solving the system, we restore the original positions and radii of $\{b_1, b_2, b_3, b_4\}$ and transform each computed tangent sphere accordingly. Note that if the centers of tangent spheres are located inside the intersection of all four balls, tangent spheres will have negative radii.

There also may be tangent spheres of infinite radius. Their surfaces can be regarded as planes. To compute tangent planes for three 3D balls (an example shown in Figure 2.2 B), we use an approach similar to the one used when computing tangent spheres for four balls.

2.1.4 Finding the first valid triple

We make an assumption that balls close to each other are likely to form valid quadruples. Thus, we take the first ball $b_0 \in B$ and select a set of its nearest neighbors $B_0 \subset B$. We then start enumerating quadruples for B_0. If a quadruple has a tangent sphere that does not intersect any ball from B, then a quadruple is valid and one of its triples is returned. If no valid quadruples are found, B_0 is expanded and more quadruples are enumerated and tested.

In line 22 of Procedure 1 we need to find a first valid triple that should contain a ball that was previously not included in any of the already found valid quadruples. It may not always be possible, therefore for such constrained search we limit the maximum size of B_0 to avoid enumera-

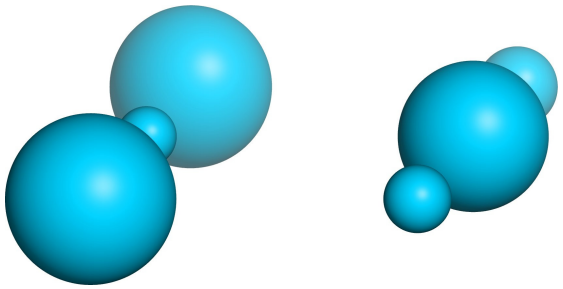


Figure 2.3: Two examples of loose triples.

tion of all quadruples for B.

2.1.5 Finding all neighbors for a valid triple

Constricted and loose triples

Consider a valid triple of balls $t = \{a, b, c\} \subset B$. Generally, t can have infinitely many possible tangent spheres, and a ball d can have a tangent sphere with t if and only if d intersects or touches the volume defined by the union of all the possible tangent spheres of t. If t has infinitely large tangent spheres, then these spheres can be regarded as tangent planes. If t has exactly two tangent planes, we call it a *constricted* triple. Otherwise we call it a *loose* triple (see figure 2.3). We define separate algorithms of finding neighbors for constricted and loose triples, i.e. triples that do not have two tangent planes.

Search space for a constricted triple

A constricted triple t has two tangent planes, therefore the union of all the possible tangent spheres of t is a union of the following three regions:

- the half-space h_1 defined by the first tangent plane;
- the half-space h_2 defined by the second tangent plane, h_1 and h_2 may intersect;

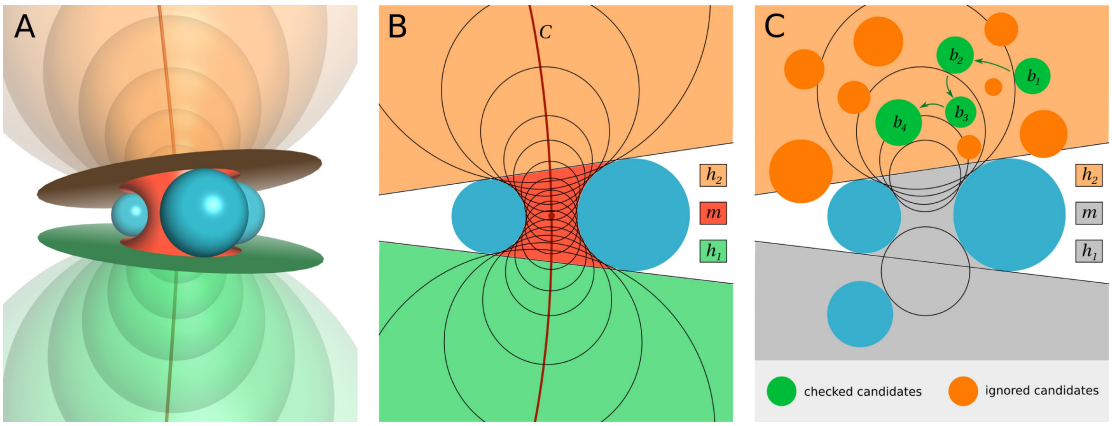


Figure 2.4: (A) Regions defined by two tangent planes in 3D. (B) Regions defined by two tangent planes in 2D. (C) Illustration of Procedure 2 for halfspace h_2: starting with b_1 the procedure runs until encountering b_4, which produces an empty h_2-related tangent sphere.

- the region m located between the two tangent planes.

Figure 2.4 (A, B) provides an illustration of such a subdivision. The centers of all the possible tangent spheres of t belong to a continuous curve.[31] Let us denote this curve as C. C intersects the plane defined by the centers of the three balls in t at a single point p, which corresponds to the center of the smallest possible tangent sphere of t. When moving away from p along the curve C, the radius of the corresponding tangent sphere always grows.

Finding neighbors in the halfspaces defined by a constricted triple

Let us assume that for a constricted triple t there is a ball d_i such that d_i intersects halfspace $h_x \in \{h_1, h_2\}$. If t and d_i have a single tangent sphere s_i, let us call s_i a h_x-related tangent sphere for t and d_i (if t and d_i have two tangent spheres, then one of them closer to h_x is called h_x-related). Another tangent sphere s_j of t can be produced by moving the center of s_i along the curve C (with the radius of s_j changing so that s_j remains tangent to t). If the movement is directed towards h_x, then s_j intersects d_i, therefore s_j is not empty. Otherwise the movement is directed away

from h_x and s_j does not even touch d_i. In this case, if s_i is empty, then s_j does not touch any ball in h_x. Therefore, if s_i is empty, then it is the only empty h_x-related tangent sphere for t.

The properties of h_x-related tangent spheres allow us to define Procedure 2 for finding a valid neighbor of t in halfspace h_x. Along with the pseudocode, we provide a simplified description of the procedure:

1. The procedure starts with any ball that intersects h_x and produces a h_x-related tangent sphere with t;
2. The procedure selects any ball that intersects both h_x and the previously produced tangent sphere and produces another h_x-related tangent sphere;
3. If step 2 has produced a tangent sphere, then step 2 is repeated;
4. If the last produced tangent sphere is empty, then the last selected ball is a valid neighbor.

Procedure 2 is greedy, it does not check all the balls that intersect h_x. See Figure 2.4 (C) for an illustration of the procedure run. Procedure 2 does not need to be called if a valid neighbor of t from h_x is already known from the previously found valid quadruple that contains t. Therefore, for most valid triples the procedure is performed only once. Also, the running time of the procedure can be reduced if in line 2 a ball is selected from a close neighborhood of t.

Finding neighbors in the middle region defined by a constricted triple

After valid neighbors of t from both h_1 and h_2 are determined, there may be remaining valid neighbors that do not intersect h_1 or h_2 but intersect middle region m and have empty tangent spheres with t. For a fast intersection checking we need a simple approximation of m. Let us consider the surface of m. It is known to be a part of the Dupin cyclide de-

Procedure 2 Find a valid neighbor of a triple in a halfspace

input: B = (a set of balls), t = (a triple of balls, $t \subset B$), h_x = (a halfspace, $h_x \in \{h_1, h_2\}$)

output: d = (a valid neighbor of t) and s = (an empty h_x-related tangent sphere of t and d)

1: $\langle d, s \rangle \leftarrow \langle \emptyset, \emptyset \rangle$
2: $d_0 \leftarrow$ (select a ball $d_0 \in B$ such that there exists a h_x-related tangent sphere s_0 for t and d_0)
3: **while** $d_0 \neq \emptyset$ **do**
4: $\quad intersection \leftarrow false$
5: $\quad replacement \leftarrow false$
6: $\quad$ **repeat**
7: $\quad\quad d_1 \leftarrow$ (select another ball $d_1 \in B$ such that d_1 intersects s_0)
8: $\quad\quad$ **if** $d_1 \neq \emptyset$ **then**
9: $\quad\quad\quad intersection \leftarrow true$
10: $\quad\quad\quad$ **if** there exists a h_x-related tangent sphere s_1 for t and d_1 **then**
11: $\quad\quad\quad\quad replacement \leftarrow true$
12: $\quad$ **until** $replacement = true$ or $d_1 = \emptyset$
13: $\quad$ **if** $replacement = true$ **then**
14: $\quad\quad \langle d_0, s_0 \rangle \leftarrow \langle d_1, s_1 \rangle$
15: $\quad$ **else if** $intersection = true$ **then**
16: $\quad\quad \langle d_0, s_0 \rangle \leftarrow \langle \emptyset, \emptyset \rangle$
17: $\quad$ **else**
18: $\quad\quad \langle d, s \rangle \leftarrow \langle d_0, s_0 \rangle$
19: $\quad\quad \langle d_0, s_0 \rangle \leftarrow \langle \emptyset, \emptyset \rangle$
20: **return** $\langle d, s \rangle$

fined by balls in t (Figure 2.5 A).[31,87] A Dupin cyclide is an envelope surface of spheres tangent (both externally and internally) to the three fixed spheres. The surface of m is part of the Dupin cyclide that corresponds only to externally tangent spheres, i.e. tangent spheres that do not overlap the three fixed spheres. Each externally tangent sphere of t has three points touching balls in t. The circumcircles of such triples of touching points lie on the surface of m (Figure 2.5 B).[88] There are two circumcircles that also lie on the touching planes of t because they correspond to the largest possible tangent spheres. We use a bounding cylinder of these two circumcircles as an initial approximation of m (Figure 2.5 C). We can reduce the size of this bounding cylinder by considering circumcircles that correspond not to the largest possible tangent spheres of t, but to the largest empty tangent spheres of t, i.e. empty h_1-related and h_2-related tangent spheres, if such exist. If a ball intersects the defined bounding cylinder and is located between the two tangent planes of t, then this ball is checked for having at least one empty tangent sphere with t.

Notably, circumcircles defined by tangent spheres of negative radii (see Figure 2.5 (B3) for an example) may lie outside of the cylinder approximating region m. However, this does not present a problem, because if a valid neighbor of t corresponds to a tangent sphere of negative radius, then this neighbor overlaps the center of that tangent sphere and therefore intersects the cylinder approximating m.

Finding all neighbors for a loose triple

Let us now consider a case, in which a valid triple t does not have exactly two tangent planes. In this case the search for valid neighbors is performed in a brute-force manner: each ball $b \in B \setminus t$ is checked for having at least one empty tangent sphere with t. It should be possible to define a faster but a more complex procedure for handling loose triples. However, this would not significantly improve the overall performance of the algorithm, because our tests (described later in the text) show that

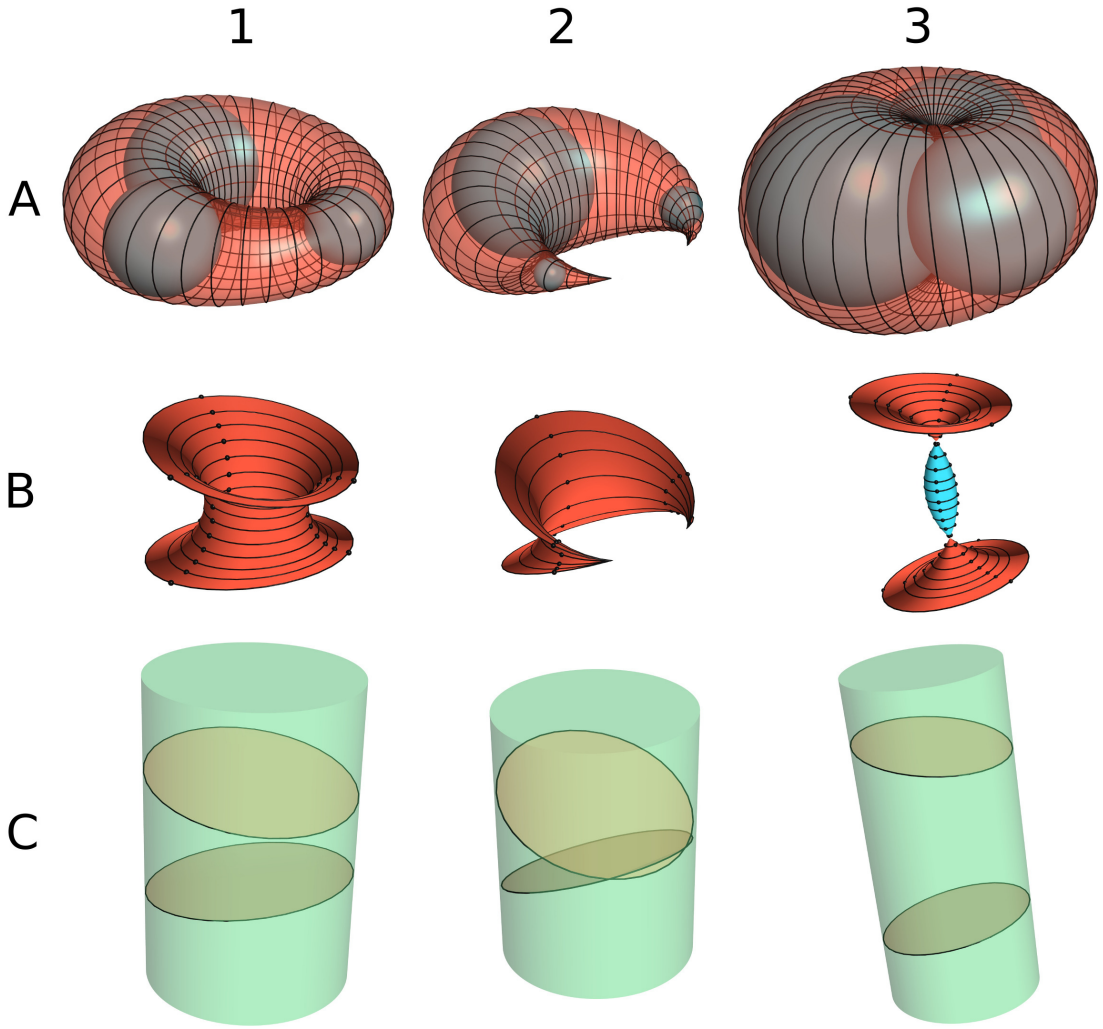


Figure 2.5: (A) For constricted triples of balls it is possible to define Dupin cyclides of three types: ring-like (1), horn-like (2) and spindle-like (3). Horn-like cyclides are generally two-part, but we only show the part relative to the middle region m. (B) 3D surfaces of middle regions are parts of Dupin cyclides displayed above. Black circles indicate circumcircles of the points, at which externally tangent spheres touch the balls of a triple. The circles lie on the surface of a middle region. The topmost and bottommost circles also lie on touching planes and are used for approximating a middle region. In the last case (3) the middle part of the surface corresponds to tangent spheres that have negative radii because they lie inside the intersection of all the balls of a triple. (C) Bounding cylinders of the topmost and bottommost circles shown in (B).

in macromolecular structures loose triples occur very rarely.

2.1.6 Efficient searching in a large set of balls

Let us summarize geometric search operations that we need to implement: search for balls that intersect a halfspace; search for balls that intersect a sphere; search for balls that intersect a cylinder. To implement them efficiently we need a search data structure. We chose to use a bounding spheres hierarchy (BSH)[89] because it does not add any additional complexity when implementing geometric queries that we need: checking any bounding sphere for an intersection with some object is no different from checking an input ball for the same thing. Our approach to the construction of BSH for a set of input balls B can be summarized as follows:

1. The elements of B form the leaf nodes of the tree;
2. Nodes created in the previous step are grouped and enclosed within bounding spheres which form the higher level of nodes;
3. Step 2 is performed in a recursive fashion eventually resulting in a tree structure with a single bounding sphere at the top of the tree.

In step 2 we can use the following algorithm:

1. Select group centers from the input spheres using the greedy Procedure 3;
2. Assign each input sphere to the group with the nearest center;
3. Construct a bounding sphere for each group.

This algorithm is practical only for a relatively small number (less than 10^5) of input spheres, because Procedure 3 has quadratic time complexity. To overcome this problem we provide input in smaller portions. The portions are determined by recursively subdividing the input set of spheres using k-d tree subdivision algorithm.[90]

Procedure 3 Select group centers in BSH construction

input: $S =$ (a list of spheres), $l_{min} =$(minimal distance between group centers)
output: $S_{selected} =$ (a set of selected group centers)

```
1: S ← (order S by the distance to S[0])
2: S_selected ← ∅
3: S_locked ← ∅
4: for all a ∈ S do
5:    if a ∉ S_locked then
6:       S_selected ← S_selected ∪ a
7:       for all b ∈ S do
8:          if distance(a, b) < l_min then
9:             S_locked ← S_locked ∪ b
10: return S_selected
```

Two examples of bounding spheres hierarchies are shown in Figure 2.6. Searching in BSH is performed as in any other tree structure – children are not examined if their parent does not satisfy the predefined condition. In the case of BSH, a node is not examined if the bounding sphere of the parent node does not satisfy the predefined constraint. A search starts from the root node and can be performed in either depth-first or breadth-first manner. If we need to find out whether at least one ball from B satisfies some condition (for example, if any ball intersects a tangent sphere), then the depth-first search method is more beneficial because it reaches the leaves level faster.

2.1.7 Handling special situations

The Voronoi diagram of balls may exhibit various special cases and anomalies.[38,91] Since our algorithm searches for Voronoi vertices, we focus on handling two special situations that relate to the existence of Voronoi vertices and valid quadruples.

Firstly, let us consider a ball that has the Voronoi cell without vertices and, therefore, is not part of any valid quadruple. As noted by Medvedev et al.,[35] such orphaned balls can be identified and handled separately after the search for the Voronoi vertices. We choose to simply report the

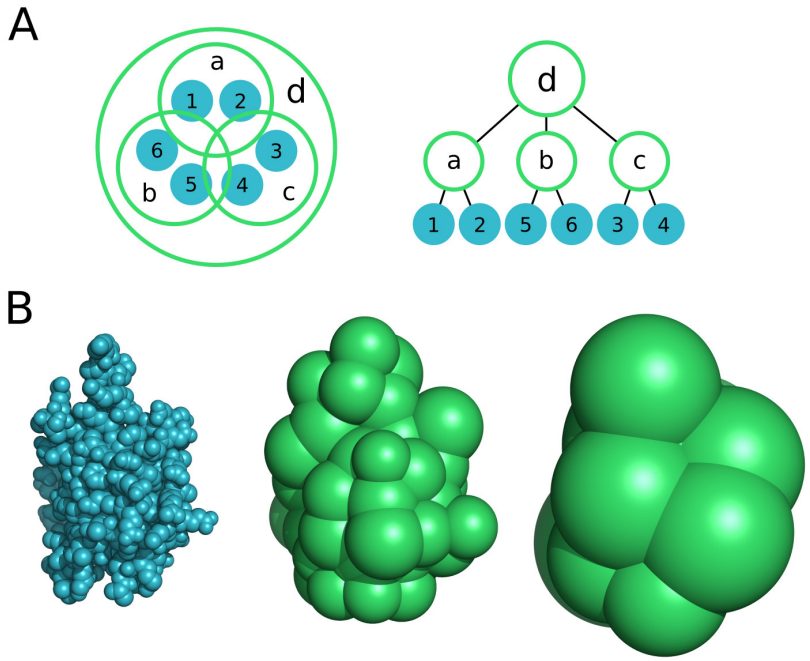


Figure 2.6: (A) 2D example of a bounding spheres hierarchy. (B) Illustration of a bounding spheres hierarchy applied to a protein structure: protein atoms (left), the first layer (middle) and the second layer (right) of bounding spheres.

orphaned balls. Our tests, described later in the text, show that occurrences of orphaned balls in macromolecular structures are extremely rare and that they represent physically non-realistic stereochemistry.

Secondly, let us consider a situation where more than four balls share the same empty tangent sphere. If n is the number of these spheres, then it is possible to select up to $\binom{n}{4}$ quadruples defining the same Voronoi vertex. Our algorithm selects a smaller set of quadruples because it considers halfspaces defined by triples of balls. For example, if the algorithm is applied to a set of points where more than four points share the same circumsphere, it produces a valid triangulation where simplices meet edge-to-edge or vertex-to-vertex and do not overlap. However, we provide an optional procedure for finding all possible valid quadruples. After computing all the Voronoi vertices we use a bounding spheres hierarchy to search for all the touching balls for each of the constructed tangent spheres. We then report surplus quadruples for each tangent sphere that

has more than four touching balls.

2.1.8 Parallelization of the algorithm

We parallelize the algorithm by implementing the following strategy:

1. Subdivide the set of input balls B into k smaller sets $B_1, B_2, \ldots B_k$.
2. In parallel: for each $B_i \in \{B_1, B_2, \ldots B_k\}$ find a set Q_i of all valid quadruples that contain at least one ball from B_i.
3. Return the full set of quadruples $Q = Q_1 \cup Q_2 \cup \ldots \cup Q_k$.

In step 1 we recursively subdivide the input set of balls using k-d tree subdivision algorithm,[90] so that during each subdivision step the input is divided into parts that are as similar in size as possible. In step 2 we run the algorithm for B as defined in Procedure 1, but maintain the following constraint: every triple pushed into the stack should contain at least one element of B_i. This requires a simple modification in the procedure for finding a first valid triple (called in lines 5 and 22 of Procedure 1): a returned first valid triple should contain at least one element of B_i.

2.1.9 Convergence of the algorithm

Let us show that both sequential and parallel versions of our algorithm find all the Voronoi vertices. Missing a Voronoi vertex implies missing a valid quadruple (otherwise it would imply a wrongful rejection of an empty tangent sphere, which is not possible because the check for the sphere emptiness is performed explicitly). Let us assume that there is a valid quadruple q_m that was missed. Let us consider a ball $b_m \in q_m$. If b_m is not a part of some found valid quadruple, then the procedure for finding the first valid triple containing b_m would be called, which would find either q_m or some other valid quadruple containing b_m. Therefore b_m

must be a part of some found valid quadruple $q_f \neq q_m$. Let us now consider the Voronoi cell V_m of b_m. For any two vertices of V_m there is a path of Voronoi edges between them (none of the special cases[38] of Voronoi cells of balls have disjoint sets of vertices). Each Voronoi edge of V_m corresponds to a valid triple that contains b_m. Therefore, there is a path of valid triples containing b_m between any two valid quadruples containing b_m. The sequential version of the algorithm would follow such a path because it would search for neighbors of every valid triple (including the triples that are subsets of q_f). The parallel version would follow such a path because it would search for neighbors of every valid triple that contains b_m when processing the subset of the input balls that contains b_m. Thus, if q_f was found, then q_m would be found too, which contradicts the initial assumption that q_m was missed.

2.1.10 Implementation

Our algorithm for computing the vertices of the Voronoi diagram of 3D balls is implemented as an open-source C++ program, named Voronota. It has no external dependencies, and only a C++ compiler is needed to build it. In addition, we developed parallel implementations of the algorithm using OpenMP and MPI technologies.

All the geometric calculations are implemented using double precision floating point numbers (C++ "double" data type). To reduce the effects of numerical errors we apply several techniques. We use Kahan summation algorithm[92] in the code for solving equations when computing tangent spheres and tangent planes. For quadratic equations we use a numerically safer solving algorithm.[93] After computing each tangent sphere or plane we calculate to what extent the computed object is really tangent: the obtained tangency error estimate is used when performing intersection queries.

As an input, Voronota accepts a list of balls in the plain text format. The

software provides a way to create such lists of balls from files in PDB and mmCIF formats. By default, all heteroatoms and all hydrogen atoms are ignored, but this behavior can be altered using command-line options. Voronota also offers the possibility to customize VDW radii of atoms. The output of Voronota is an easily parseable list of valid quadruples and the corresponding empty tangent spheres.

2.2 Testing results

2.2.1 Testing on Protein Data Bank structures

The software was tested on Intel Core i7-2600 3.40GHz processor. Firstly, we compared the performance of our software and of two other tools: QTFier (`voronoi.hanyang.ac.kr/software.htm`, version 1.0) and awVoronoi (`sourceforge.net/projects/awvoronoi`, version 1.0.0). These tools output both the Voronoi vertices and the topological links between them, while Voronota outputs only Voronoi vertices. Therefore we provide a running time comparison only for an informational purpose.

The test set consisted of all asymmetric units available from the Protein Data Bank (PDB) database[2] as of 2013.05.15 (90365 structures, each having at least 4 non-hydrogen "ATOM" records). Hydrogen atoms and heteroatoms were removed from the input structures. All the three tools were given the same set of coordinates and used the same set of VDW radii[94] (it was the only set of radii available in QTFier). For speed analysis we measured CPU-time needed to process every input structure (Figure 2.7 (A, B)). Voronota processed the test set in about 34.8 hours, QTFier and awVoronoi in 138.2 and 172.9 hours, respectively. Importantly, Voronota did not fail on any of the input PDB structures whereas QTFier failed on 259 and awVoronoi failed on 104.

Results obtained on the PDB test set enabled us to make some general-

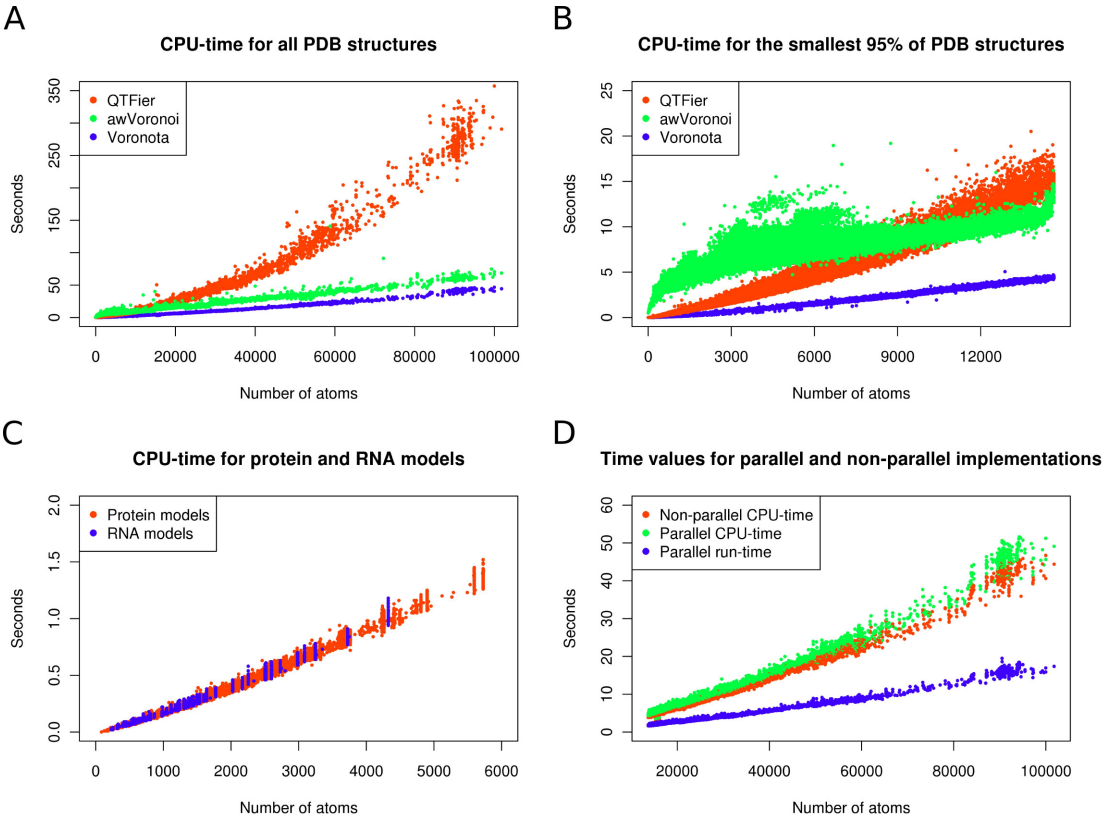


Figure 2.7: (A) CPU-time values for the macromolecular structures available from the PDB database. (B) CPU-time values for for 95% of smallest structures from PDB. (C) Voronota CPU-time values for protein and RNA structural models. (D) Voronota run-time and CPU-time values for non-parallel and parallel implementations. Parallel implementation was executed on 4 computational units.

izations. For example, it turned out that the number of valid quadruples linearly correlates with the number of atoms (Pearson's correlation coefficient being greater than 0.99). On average, the number of valid quadruples was about 6.6 time greater than the number of atoms and about 2 times smaller than the number of valid triples. Only less than 0.005% of all the valid triples were not constricted, and only 18 of about $4.5 \cdot 10^8$ atomic balls did not have any Voronoi vertices. On average, about 11 quadruples had to be examined to find the first valid triple.

We also asked if there is any quadruple q that meets both of the following two conditions: 1) q is found by QTFier or awVoronoi, but not by Voronota; 2) q is valid with respect to the 10^{-10} angstroms threshold used for checking if any of the tangent spheres defined by q is empty. There were some quadruples meeting the first condition, but none of them met the second condition. One of the reasons for slight differences in the output is that the three programs handle floating point arithmetic errors differently. The tools may also be using different approaches to handle degenerate situations.

To check if Voronota is capable of properly handling molecules with hydrogen atoms, we performed a similar test routine with all the NMR entries from the initial PDB set (9883 structures, only first structural model from each entry was used). This time hydrogen atoms were retained. Voronota successfully processed all the input structures. In comparison with the hydrogen-free testing results, there were more non-constricted valid triples (approximately 0.5%) and atomic balls that did not have any Voronoi vertices (117 atomic balls from 39 input structures). Voronota processed the input set in about 1.3 hours, QTFier and awVoronoi in 3.3 and 15.4 hours, respectively.

We analyzed all the situations where either a hydrogen or non-hydrogen atom did not have any corresponding Voronoi vertices. We found that these cases represent either unrealistically short covalent bonds or severe steric clashes of non-bonded atoms. Therefore, such situations may

be considered to be indicators of dubious low-quality macromolecular structures.

2.2.2 Testing on protein and RNA structural models

Computational structural models are becoming widely used for various applications. However, models, especially of lower accuracy, may have a number of physically unfeasible features. Therefore, we decided to test whether Voronota is sufficiently robust to be used for computational models. To this end we used 28806 protein models ranging widely in their quality submitted by modeling servers to CASP9[95] and CASP10[17] experiments. Voronota successfully processed all the structural models (QTFier failed on 31 and awVoronoi failed on 875 input structures). In addition, Voronota was successful in processing all 42585 RNA models from the "randstr" decoys set.[60] For the protein and RNA models the relation between structure size and CPU-time (Figure 2.7 (C)) was consistent with the results for the PDB structures (Figure 2.7 (A, B)).

2.2.3 Testing parallel implementations

We tested the performance of our OpenMP-based parallel implementation on the 5000 largest structures from PDB. The execution was performed on the same machine as before, 4 computational units were used for each input structure. Figure 2.7 (D) shows the recorded run-time (real time) and CPU-time (total amount of time spent by all the used computational units) values for both non-parallel and parallel implementations.

The MPI-based parallel implementation is likely most suitable for processing very large structures on a computing cluster. For example, we processed the HIV virus capsid structure (PDB ID 3J3Q, 2440800 atoms) on a cluster of Intel Xeon X5650 2.66GHz processors. When 9 CPU cores were used, run-time and CPU-time values were 511 and 4176 seconds, respectively. For 17 cores the values were 293 and 4330 seconds, for 33

cores – 189 and 5018 seconds.

2.3 Discussion

We presented a simple and robust algorithm for computing the vertices of the Voronoi diagram of balls. The algorithm is particularly well-suited for processing 3D structures of biological macromolecules. It takes advantage of the observation that in the case of macromolecular structures the overwhelming majority of valid ball (atom) triples are constricted (have two tangent planes). When processing constricted triples, our algorithm efficiently combines the knowledge of the search space with the use of hierarchical spatial indexing. The algorithm uses a bounding spheres hierarchy to iteratively search for neighbors so that the search space is reduced after each iteration. Importantly, we introduce a simple approximation for the middle region of the search space defined by the constricted triple. When processing rare loose triples (triples without two tangent planes) our algorithm does not attempt to reduce the search space, but still uses a bounding spheres hierarchy to speed up the search for neighbors. This strategy works well in terms of speed and simplicity of the algorithm implementation. Another important feature of the algorithm is the simplicity and generality of the procedure for finding the first valid triple, which enabled us to parallelize the algorithm in a straightforward manner. We implemented the algorithm as an open-source console application, Voronota, which can be run on either single or multiple processors. Large-scale tests showed that Voronota is a fast and reliable tool for processing both experimentally determined and computationally modeled macromolecular structures.

3 CAD-score: a method for contact area-based comparison of structures and interfaces of proteins, nucleic acids and their complexes

CAD-score (Contact Area Difference Score) is a universal method to quantify both local and global similarity of macromolecular structures. The method employs the Voronoi diagram of atomic balls for deriving interatomic contact areas. Here, we provide a description of CAD-score and present the results of extensive testing procedures performed on both protein and RNA structural models. We then present the CAD-score web server and describe the application of CAD-score for large-scale clustering of protein-protein interaction interfaces.

3.1 Method description

3.1.1 Construction of inter-atom contacts

Atom-atom contact areas are derived using the Voronoi tessellation of 3D balls (also known as the additively weighted Voronoi diagram or the Apollonius diagram), where balls correspond to the heavy atoms of van der Waals (VDW) radii.[7] Here we used van der Waals radii for heavy atoms derived by Li & Nussinov.[96] For each atom we can define the Voronoi cell, a set of all points closer to this particular atom than to any other atom. Two atoms are said to be Voronoi neighbors if their Voronoi cells share a common subset of points.

Interatomic contacts are derived from the Voronoi diagram of atoms based on the idea proposed by McConkey et al.[97] Neighboring atoms are

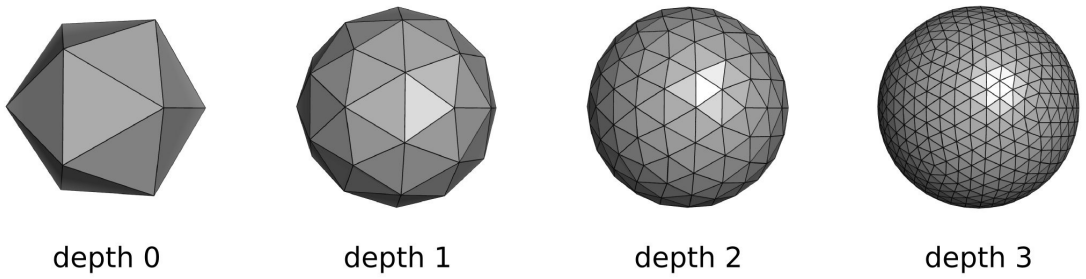


Figure 3.1: Constructing a triangulated representation of a sphere using the recursive icosahedron subdivision technique.

defined as contacting each other if a water molecule cannot fit between them. Thus, the complete contact surface of an atom is represented by the sphere of the radius equal to the sum of van der Waals radius of the atom and the standard radius (1.4 Å) of a water molecule. We term it a contact sphere. Point p on the contact sphere of atom i belongs to the contact surface with atom j if the following two conditions are satisfied: 1) i and j are Voronoi neighbors; 2) p is closer to j than to i or any other neighbor of i. If p is closer to i than to any neighbor of i, then it belongs to the solvent-accessible surface.

For a given atom we use the recursive icosahedron subdivision technique[98] to produce a triangulated representation of its contact sphere (Figure 3.1). We then construct inter-atom contact surfaces by intersecting the triangulated surface of the atom contact sphere with hyperboloids that correspond to the junctures of neighboring Voronoi cells, i.e. Voronoi faces (Figure 3.2). For an analytic representation of such hyperboloids we use the method proposed by Kim et al.[31]

The method of cutting triangles with hyperboloids can also be applied to construct contact surfaces that do not lie on the contact sphere of an atom, but correspond to the portions of the atomic Voronoi cell faces that are inside the contact sphere: this idea is illustrated in Figure 3.3. Such contact surfaces, which can be called constrained Voronoi faces, are suitable for both visualizing and quantifying interatomic contacts, but we initially

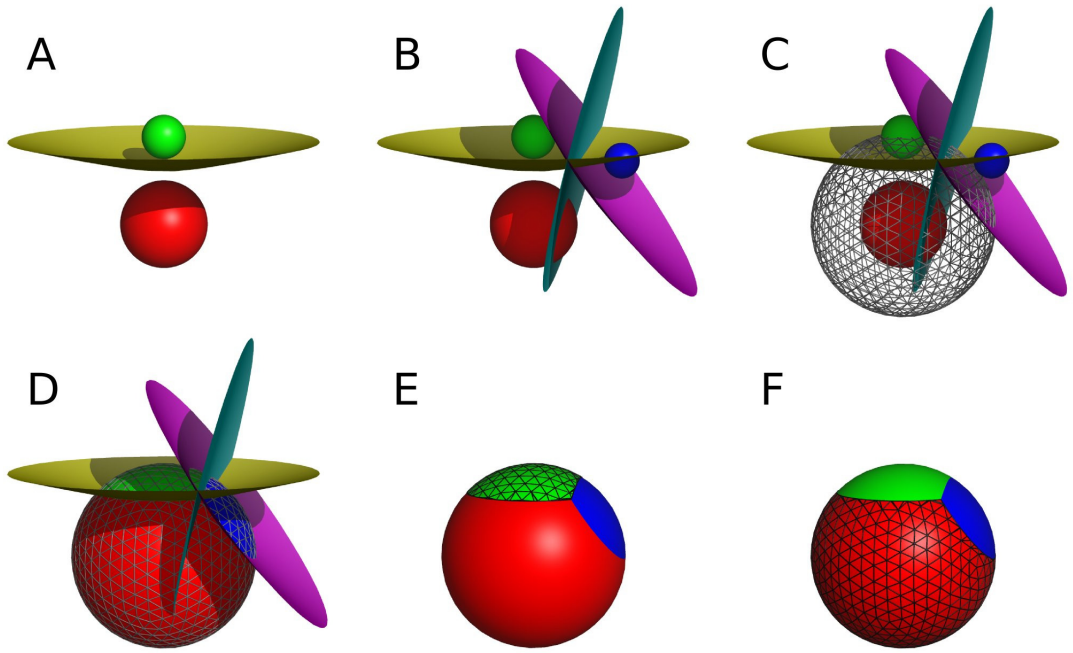


Figure 3.2: (A) A hyperboloid defined for a pair of balls. (B) Pairwise hyperboloids defined for three balls. (C) Hyperboloids intersecting the triangulated representation of the contact sphere of the red ball. (D) Hyperboloids cutting the contact sphere into three contact surfaces. (E) The contacts defined by the intersections in (D), triangulation of the contact with the green ball is overlaid on top. (F) Similar to (E), but with the overlaid triangulation belonging to the solvent-accessible surface of the red ball.

used on-sphere contacts because their implementation was simpler and worked faster.

3.1.2 Construction of inter-residue contacts

Residue-residue contacts are constructed by simply grouping contacts between atoms of corresponding residues. The inter-atom contacts corresponding to the covalent bonds that connect residues adjacent in sequence are not considered. Since contacts are resolved at the level of atoms, we can define contacts not only for the entire residue but also for various subsets of its atoms (e.g. main chain and side chain). Figure 3.4, created using our Voroprot[40] software, illustrates how the combination of Voronoi cells and contact spheres is used to construct contact surfaces for an atom (Figure 3.4 A) and for a residue (Figure 3.4 B).

3.1.3 Partitioning of nucleobase-nucleobase contacts into stacking and non-stacking contacts

For nucleic acids, we additionally characterize base-base contacts by partitioning them into stacking and non-stacking ones, Let us consider base i, which is in contact with base j, $i \neq j$. If all atoms (represented as spheres of VDW radii) of base j are entirely on one side of the plane of base i, the base-base contact is defined as the stacking contact. Conversely, if one or more atoms (or part of their VDW spheres) of base j appear on the other side of the plane of base i than the remaining atoms of base j, the contact is defined as non-stacking. The illustration of this simple definition is provided in Figure 3.5.

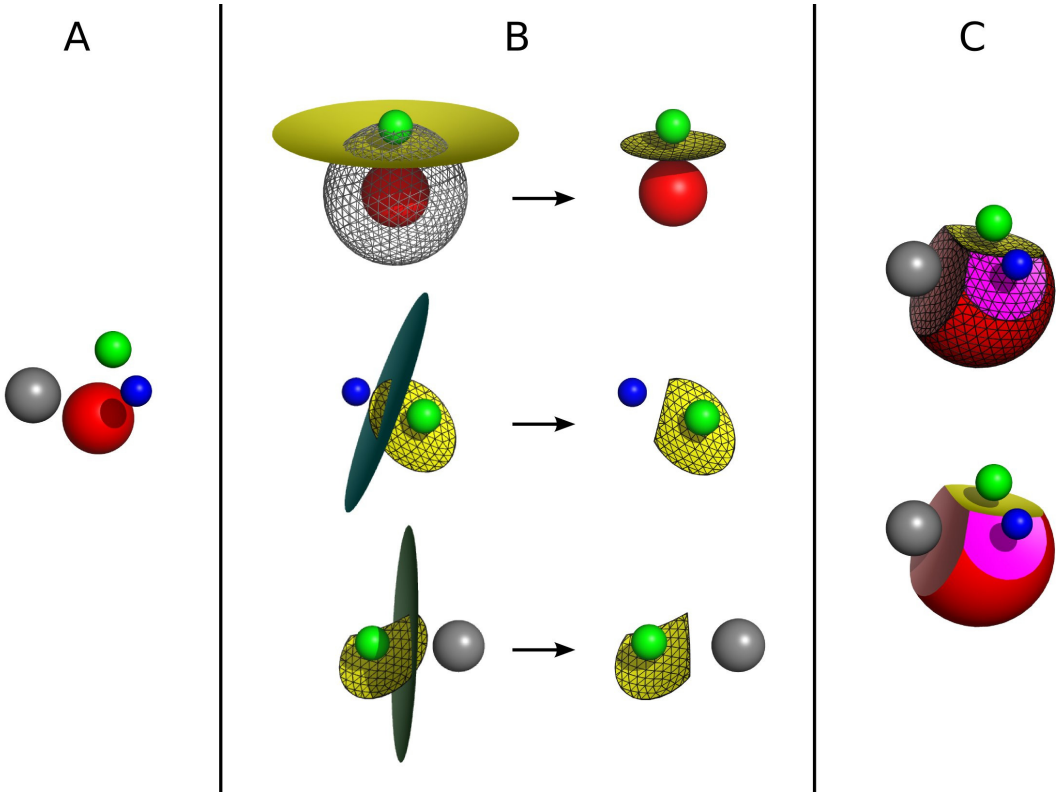


Figure 3.3: Constructing constrained Voronoi faces. (A) Four neighboring balls. (B) Constructing a contact between the red and the green balls in three steps: selecting a portion of the inter-ball hyperboloid that is inside the contact sphere of the red ball, triangulating it can be done by projecting the cut-out part of the contact sphere triangulation on the hyperboloid; cutting the previously initialized triangulated patch with the next hyperboloid that correspond to the contact between the green and the blue balls; cutting the previously modified triangulated patch with the last hyperboloid that correspond to the contact between the green and the gray balls. (C) Constrained Voronoi faces obtained using the approach demonstrated in (B), plus the corresponding solvent-accessible surface of the red ball.

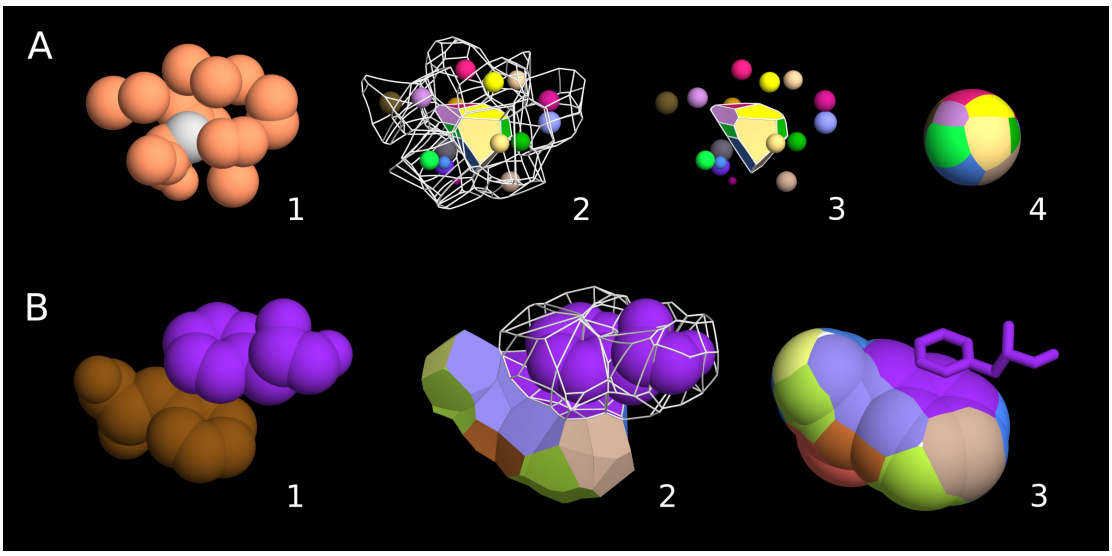


Figure 3.4: Illustration of the procedure for deriving contact surfaces for atoms (A) and residues (B). (A) Interatomic contacts: 1 — the considered atom (grey) surrounded by neighboring atoms; 2 — the Voronoi cell of the considered atom (solid) and neighboring Voronoi cells (wireframe); small colored spheres correspond to the same neighboring atoms shown as large spheres in 1; 3 — the Voronoi cell with its faces colored according to the color of neighboring atoms; 4 — interatomic contact surfaces mapped onto the contact sphere of the atom. (B) Inter-residue contacts: 1 — two interacting phenylalanine residues in the space-filling representation; 2 — Voronoi cells of the same residues; faces of one of the residues are colored according to the color of neighboring residues; 3 — the map of inter-residue contact surfaces for one of the interacting residues.

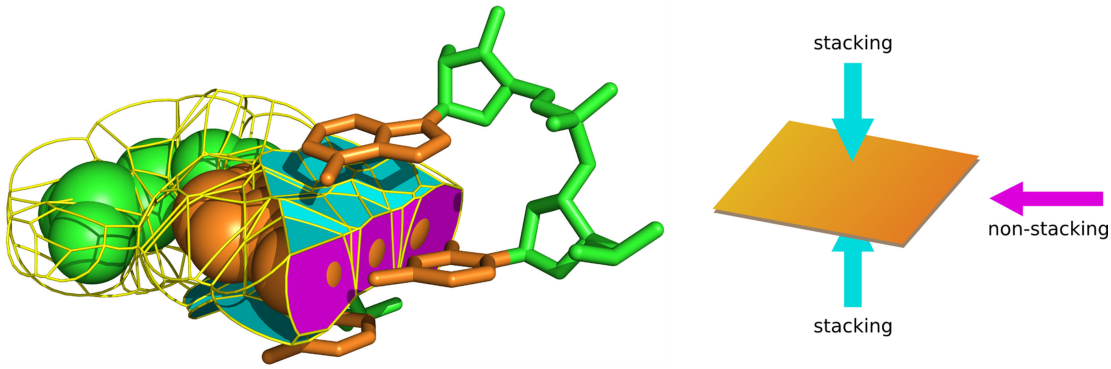


Figure 3.5: Illustration of the definition of stacking/non-stacking base-base contacts. On the left, a nucleotide in the space-filling representation is shown in contact with the three neighbors. Contacts are represented as faces of the Voronoi cells constrained by the contact spheres. Cyan and magenta indicate stacking and non-stacking contacts respectively. On the right, the same contacts are shown in schematic representation.

3.1.4 CAD-score definition

Global score

We defined CAD-score based on the three main considerations: 1) contacts in the model should be evaluated according to the contacts in the reference structure (target); 2) any missing residues in the model should be treated in the same way as if none of their contacts were correctly predicted; 3) strong over-prediction (non-physical overlap) of a particular contact should be equivalent to missing that contact entirely. The mathematical definition of CAD-score is presented below.

Let G denote the set of all the pairs of residues (i, j) that have a non-zero contact area $T_{(i,j)}$ in the target structure. Then for every residue pair $(i, j) \in G$ we calculate the contact area $M_{(i,j)}$ in the model. If the model has additional residues not present in the target, these residues are excluded from the calculation of contact areas. If some residue is present in the target, but is missing from the model, all the contact areas for that residue in the model are assigned zeroes.

For every residue pair $(i, j) \in G$ we can then define contact area difference as the absolute difference of contact areas between residues i and j in target T and in model M:

$$\mathrm{CAD}_{(i,j)} = |T_{(i,j)} - M_{(i,j)}| \tag{3.1}$$

To impose symmetrical treatment of over-prediction and under-prediction of the contact area, instead of the raw $\mathrm{CAD}_{(i,j)}$ value, we use bounded $\mathrm{CAD}_{(i,j)}$ defined as follows:

$$\mathrm{CAD}_{(i,j)}^{\mathrm{bounded}} = \min(\mathrm{CAD}_{(i,j)}, T_{(i,j)}) \tag{3.2}$$

CAD-score for the whole model is then defined as:

$$\text{CAD-score} = 1 - \frac{\sum_{(i,j)\in G} \text{CAD}^{\text{bounded}}_{(i,j)}}{\sum_{(i,j)\in G} T_{(i,j)}} \tag{3.3}$$

The sum in the numerator of equation (3.3) never exceeds the sum of all contact areas $T_{(i,j)}$ in the target structure. In other words, CAD-score defined by equation (3.3) is always within the [0,1] range. If model and target structures are identical, CAD-score=1. At the other extreme, if not a single contact is reproduced with sufficient accuracy (there are no cases satisfying the condition: $\text{CAD}_{(i,j)} < T_{(i,j)}$), CAD-score=0.

Local scores

For the analysis and visualization of local differences between two structures some additional scores need to be defined. Two types of local error values (raw and normalized) can be derived for every residue. A raw local error for residue i is defined as follows:

$$\delta(i) = \sum_{(i,j)\in G} \min(|T_{(i,j)} - M_{(i,j)}|, T_{(i,j)}) \tag{3.4}$$

A normalized local error, which is referred later in the text simply as "local error", is a raw local error divided by the sum of the corresponding target contact areas:

$$\varepsilon(i) = \frac{\delta(i)}{\sum_{(i,j)\in G} T_{(i,j)}} \tag{3.5}$$

Local error values for individual residues may show large variation. To make the signal less noisy, both raw and normalized local errors can be smoothed along the residue sequence using a window of w residues to the left and to the right:

$$\delta_w(i) = \sum_{k \in [-w,w]} \frac{\delta(i+k)}{2w+1} \tag{3.6}$$

$$\varepsilon_w(i) = \frac{\sum_{k \in [-w,w]} \delta(i+k)}{\sum_{k \in [-w,w]} \sum_{(i+k,j) \in G} T_{(i+k,j)}} \tag{3.7}$$

3.1.5 CAD-score variants

Our algorithm computes inter-residue contact areas at the resolution of individual atoms. Therefore, we can define contact area as well as contact area difference not only for the entire residue, but also for any subset of its atoms. In all cases contact areas are calculated with all atoms present, but if a subset of residue atoms is considered, only contact areas corresponding to this subset are retained. We consider two standard subsets: main chain and side chain for proteins and correspondingly backbone and base for nucleic acids. This results in nine CAD-score variants shown in Figure 3.6. For nucleic acids, further partitioning of base-base ("S-S") contacts into stacking and non-stacking ones results in two additional CAD-score variants.

Three pairs of CAD-score variants (A-S and S-A, A-M and M-A, and S-M and M-S) are not entirely symmetric. For example, glycine does not have a side chain and therefore cannot form any S-A contacts, but it can form A-S contacts. Nevertheless, for practical purposes these three pairs of CAD-score variants may be considered to be redundant. As a result, for standard subsets of residue atoms there are six non-redundant CAD-score variants that can be used to address different questions in evaluating models against the reference structure: A-A, A-S, A-M, S-M, M-M, S-S. For nucleic acids there are two additional variants: "S-S stacking" and "S-S non-stacking".

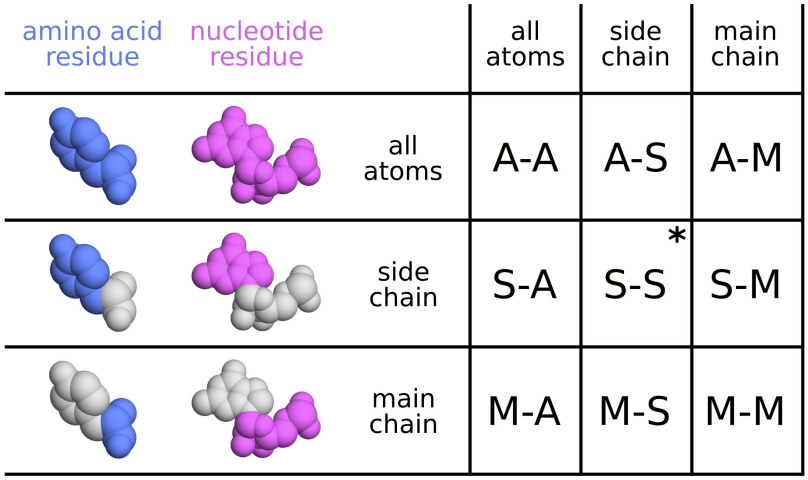


Figure 3.6: CAD-score variants based on standard subsets of residue (amino acid or nucleotide) atoms. "A", "S" and "M" denote all atoms, side-chain (base) and main chain (backbone), respectively.

3.1.6 Additional global scores for interfaces

A straightforward way to compare the inter-chain interfaces of two structures of the same sequence is to use Equation 3.3 with G limited to just inter-chain contacts. Below is an alternative representation of the same computation:

$$\text{CAD-score}^{\text{iface}} = 1 - \frac{\sum_{(i,j)\in I\times J} \min(|T_{(i,j)} - M_{(i,j)}|, T_{(i,j)})}{\sum_{(i,j)\in I\times J} T_{(i,j)}} \tag{3.8}$$

Here, I and J are the sets of interface residues of the first and the second subunits (chains), respectively, in the target (reference) protein complex. $T_{(i,j)}$ is the area of the contact between residues i and j in the target protein complex, $M_{(i,j)}$ is the corresponding area in the model protein complex. If i and j are not in contact, then the corresponding contact area equals zero.

Using the same notation we can also quantify how each interface residue is exposed to the other chain by summing the relative contact areas, like

in the example below:

$$T_i = \sum_{(i,j)\in i\times J} T_{(i,j)} \tag{3.9}$$

A set of T_i values with all $i \in I$ describes the binding site of the first chain in the target structure. The corresponding binding site in the model structure is defined in the same way. We then can compute a similarity score of the target and the model binding sites:

$$\text{CAD-score}^{\text{bsite}} = 1 - \frac{\sum_{i\in I} \min(|T_i - M_i|, T_i)}{\sum_{i\in I} T_i} \tag{3.10}$$

$\text{CAD-score}^{\text{bsite}}$ is more forgiving than $\text{CAD-score}^{\text{iface}}$ because it uses less detailed information. We can define even less detailed (and, therefore, less stringent) similarity measures using total interface areas. The interaction interface area similarity is calculated as follows:

$$\text{CAD-score}^{\text{iface-area}} = \min\left(1, \frac{\sum_{k\in I\cup J} M_k}{\sum_{k\in I\cup J} T_k}\right) \tag{3.11}$$

The binding site area similarity is defined as follows:

$$\text{CAD-score}^{\text{bsite-area}} = \min\left(1, \frac{\sum_{i\in I} M_i}{\sum_{i\in I} T_i}\right) \tag{3.12}$$

3.2 Testing results for protein structures

3.2.1 Testing data set

To test the properties of CAD-score and its effectiveness in evaluating and ranking models we applied it to models obtained during CASP9, the ninth community-wide assessment of protein structure prediction methods.[95] CASP models are generated by a large array of different meth-

ods and, therefore, represent a wide range of accuracies. In addition, the set contains models of different degree of completeness including complete models, those missing a few residues and only short structural fragments. Moreover, models differ greatly by their physical plausibility. Some of them feature structural characteristics reminiscent of high resolution experimental structures, while some others have a number of unrealistic features such as steric clashes and strongly deviating covalent bond geometries. All these aspects of CASP models make them an excellent test set for an automatic reference-based model evaluation score as the set presents a serious challenge for objective and fair model ranking.[99] To have a representative and least redundant set, we only considered CASP9 models generated by automatic methods (servers) taking a single most confident (first) model per method for a given prediction target. Since CAD-score is an all-atom measure, we excluded from our analysis models produced by methods representing amino acid residues in a simplified or incomplete form. Models for one of the targets (T0629; the long tail fiber protein gp37 of the T4 bacteriophage) were also excluded. T0629 forms the needle-shaped parallel homo-trimer, and considering the isolated single chain is both structurally and biologically meaningless.[100]

3.2.2 CAD-score is a robust measure for evaluating and ranking single-domain models

As a first step, we decided to compare CAD-score with GDT-TS, a standard CASP score that withstood the test of time and is generally recognized as the single most effective reference-based score.[56] To make an overall comparison of CAD-score and GDT-TS, we selected CASP9 models for individual domains ("assessment units" to be more precise) of prediction targets as defined by the assessors.[101] For the resulting diverse set of 8429 models we compiled both GDT-TS and CAD scores. GDT-TS values were taken from the data archive of the Prediction Center (`www.predictioncenter.org`) while different variants of

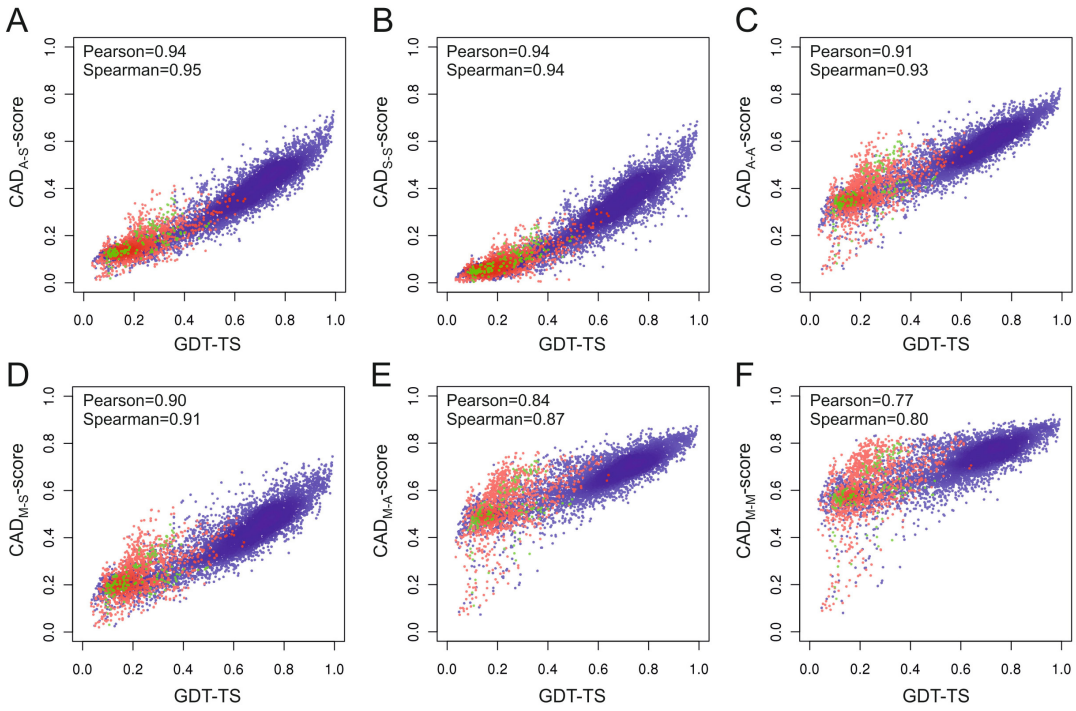


Figure 3.7: Relationship between GDT-TS (horizontal axis) and different variants of CAD-score (vertical axis) for CASP9 models. CAD-score variants (A-F) are arranged in the order of their decreasing correlation with GDT-TS. Blue, red and green colors represent models assessed in template-based (TBM), free modeling (FM) and unresolved (TBM/FM) categories respectively. Higher color intensity reflects higher density of models. Pearson's correlation coefficients and Spearman's rank correlation coefficients are indicated for each plot.

CAD-score were calculated as described in the method definition section. The plots displaying the relationship between GDT-TS and six non-redundant CAD-score variants are shown in Figure 3.7.

It is evident that there is a strong correlation between GDT-TS and CAD-score values, which is surprising considering the different nature of scores. Notably, this is true not only for Pearson's correlation coefficient, which depends on the linear relationship between the two scores. Even better values in all cases are obtained for Spearman's rank correlation, which indicates the extent to which ranking by GDT-TS agrees with ranking by CAD-score without the assumption of the linear rela-

tionship between the two scores. In particular, three types of CAD-score ("all atoms – side chain" (A-S), "side chain – side chain" (S-S) and "all atoms – all atoms" (A-A)) show the strongest correlation (Figure 3.7 A-C). For these three CAD-score variants Pearson's correlation coefficients are in the (0.91-0.94) range, and Spearman's rank correlation values are in the (0.93-0.95) range. The other three types of CAD-score, in particular the variant based on "main chain – main chain" (M-M) contacts, correlate somewhat weaker. We reasoned that the lower correlation to a large degree might be determined by the abundance of local M-M contacts that are not linked to the global topology of the structure. If this is true, the type of secondary structure should be a major factor. Indeed, when analyzed separately, the correlation for proteins rich in β-strands (many non-local M-M contacts) improved, while for α-helical proteins (mostly local M-M contacts) it decreased further (Supplementary Figure S1[6]).

We also looked at the correlation between CAD-score and GDT-HA,[51] a more stringent variant of GDT-TS. GDT-HA is similarly derived from four independent superpositions, but their threshold distances (0.5, 1, 2 and 4 Å) are half the size of those used for standard GDT-TS. Therefore, GDT-HA can provide a better resolution for models of higher accuracy. The best correlating CAD-score variants are the same (A-S, S-S and A-A) and their correlation values remain very similar. Namely, the ranges for Pearson's and Spearman's correlation coefficients are (0.91-0.95) and (0.92-0.95) respectively (Supplementary Figure S2[6]).

The only adjustable parameter used in CAD-score is the values of van der Waals (VDW) radii of protein atoms. Since different VDW radii sets have been reported in the literature we asked whether the results are sensitive to the choice of a particular set. To this end, in addition to the assessment of CASP9 models using standard VDW radii reported by Li & Nussinov,[96] we repeated the analysis using the set of minimal VDW radii derived by the same authors.[96] Although differences between the two VDW sets are variable and some are fairly significant (up to 0.45Å),

we observed only negligible differences in CAD-score values and their correlation with either GDT-TS or GDT-HA (Supplementary Table S1[6]). This finding should not be too surprising after all, since CAD-score is based on contact area differences rather than the absolute contact area sizes.

Taken together, these analyses revealed a robust performance of CAD-score on single-domain proteins. In particular, the three CAD-score types (A-S, S-S and A-A) stand out. They provide some of the highest resolution and the best correlation with GDT-TS/GDT-HA. Therefore, we will further focus mostly on the properties of these three CAD-score variants.

3.2.3 CAD-score promotes the physical realism of structural models

It is generally assumed that the better model score indicates a more accurate representation of the reference structure. However, it has been noticed that some model evaluation scores including GDT-TS are fairly insensitive to unrealistic structural features such as steric clashes or deviations in residue geometries.[48] Therefore, an improvement according to a particular score may come at the expense of physical realism of structural models. In other words, some protein structure prediction methods, especially if they are optimized against a particular score, may seemingly "improve" their performance according to that score without real improvement in model accuracy.

What about CAD-score? How the improvement of models according to CAD-score relates to their physical realism? Since CAD-score is highly correlated with GDT-TS (Figure 3.7), how does it fare in comparison to GDT-TS in this regard? To answer these questions we analyzed pairs of models for which CAD-score and GDT-TS rankings were in conflict, namely, CAD-score and GDT-TS assigned better values to different models within the considered pair. We asked which score in those cases is

more consistent with the physical realism of models. We chose the MolProbity score[102] as a measure of physical realism. MolProbity is one of the widely used structure quality evaluation suites. The MolProbity score is a single number that represents the central MolProbity protein statistics collected from a large number of high quality protein crystal structures. The score takes into account clashes between non-bonded atoms, backbone Ramachandran conformations outside the favored regions and side chain rotamer outliers.[102] Unlike GDT-TS and CAD-score, the MolProbity score is not a reference-centric measure. It does not tell how close the model is to the native structure. Instead, it reports how "protein-like" the model is. Therefore, MolProbity may be considered as an independent "judge" for resolving ranking conflicts between the two reference-based scores.

We limited our analysis to reasonably accurate models of single domains (assessment units) as it would be meaningless to consider the physical realism of grossly incorrect models. Thus, we selected models above the GDT-TS threshold of 0.6 (60%) and compiled pairs of models with the conflicting rankings between GDT-TS and each of the three CAD-score variants. We then looked at how the MolProbity score would rank models within the same pairs. The results of this analysis show that in conflicting rankings, CAD-score is supported by the MolProbity score much stronger than GDT-TS (Figure 3.8 A). Among the three CAD-score variants, $CAD_{A\text{-}A}$ received the greatest MolProbity support, followed by $CAD_{A\text{-}S}$ and then by the most stringent variant, $CAD_{S\text{-}S}$. However, model pairs with small differences of MolProbity, GDT-TS or CAD-score values might be expected to contribute a certain level of noise to the results. Therefore, we performed two additional tests aimed at the progressive elimination of the impact of noise. First, we looked only at those conflicting rankings, for which the absolute MolProbity score difference is greater than the standard deviation of the MolProbity score distribution on all considered models (Supplementary Figure S3[6]). As a result, the

CAD-score agreement with MolProbity increased dramatically (Figure 3.8 B). For the second test, in addition to the constraint on the MolProbity score difference, we asked that either GDT-TS or CAD-score values would also differ more than the corresponding standard deviation (Figure S3B-E). The second test has further emphasized the overwhelming MolProbity support for CAD-score (Figure 3.8 C). For example, the ranking by $CAD_{A\text{-}A}$ agreed with the MolProbity score in 24 out of 25 cases, and only in 1 case this was true for GDT-TS. Collectively, these analyses indicate that if there is a disagreement about the relative ranking of models, CAD-score assigns a better score to the physically more realistic model much more often than does GDT-TS. This CAD-score property might be especially relevant for tasks such as ranking models of higher accuracy and assessing model refinement, because the better performance according to CAD-score would strongly imply the improvement in physical realism as well.

CAD-score removes the necessity to split multi-domain proteins into domains for model evaluation purposes

Many proteins are composed of multiple structural domains. However, GDT-TS and other scores based on the rigid-body superposition (e.g. TM-score, RMSD) are sensitive to even small differences in domain orientation. As a result, the score of the model for the entire structure may often be disconnected from the scores of the models for individual domains. This problem can be alleviated by splitting the target structure into domains and performing domain-based evaluation. However, as there are no universal criteria for domain definition, it is often impossible to unequivocally define both the number of domains and their exact boundaries. Moreover, it is not always clear whether it is necessary to split the multi-domain target structure into domains for evaluation purposes. A simple method for helping to decide whether or not the splitting into domains is required was recently introduced by Grishin and

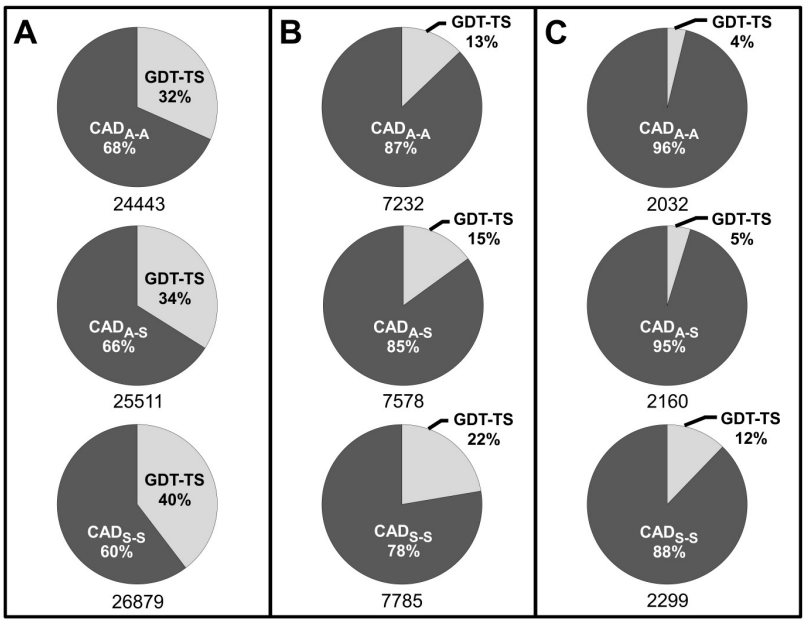


Figure 3.8: Pairs of CASP9 models with conflicting ranking by GDT-TS and CAD-score. Only models with GDT-TS over 0.6 (60%) were considered. Pie charts represent the MolProbity score agreement with rankings by GDT-TS and each of the three variants of CAD-score. Numbers of analyzed model pairs are indicated below each chart. (A) Complete MolProbity score data. (B) Data for model pairs with the absolute MolProbity score difference greater than the standard deviation (0.9). (C) Data for model pairs derived as in (B) with the additional requirement that the absolute difference of either GDT-TS or CAD-score difference would be greater than the corresponding standard deviation, i.e. 0.06 (6%) for GDT-TS, 0.05 for $CAD_{A\text{-}A}$, 0.06 for $CAD_{A\text{-}S}$, and 0.07 for $CAD_{S\text{-}S}$).

colleagues.[56] The method, used in the "official" CASP9 evaluation,[101] is based on the analysis of correlation between GDT-TS scores of the whole-chain models and the weighted sum of GDT-TS for individual domains. The weighted sum is defined as follows: GDT-TS scores for each individual domain, multiplied by its length, are summed up and divided by the sum of the domain lengths.[56] The main idea is that if the scores for the whole-chain models are systematically lower (or higher) than the weighted sum of domain scores, then the splitting into domains should be considered. Since this idea is quite general, we decided to perform a similar analysis based on CAD-score and to compare the results with those obtained for GDT-TS. However, some CASP9 whole-chain target structures have additional residues compared to the sum of individual domains. To make the analysis entirely objective, we removed these additional residues from multi-domain whole-chain target structures, so that the whole-chain structure and the sum of domains would have exactly the same residues. We then assessed models against these whole-chain targets by both CAD-score and GDT-TS. The latter data was recalculated using the LGA (Local-Global Alignment) method software.[103] The resulting analysis of 1287 models for 24 multi-domain targets is presented in Figure 3.9. There is a stark difference between the GDT-TS plot (3.9 A) and those based on CAD-score (Figure 3.9 B-D). In the case of GDT-TS, essentially for all the models the weighted sum of domain scores is higher than the score for the entire structure. This reaffirms the choice made by the assessors to parse these CASP9 targets into domains (assessment units) for performing robust model evaluation using GDT-TS. In contrast to GDT-TS, all three CAD-score variants (Figure 3.9 B-D) show at most only small differences between scores of the whole-chain structure and the combined scores of domains. In other words, the evaluation based on CAD-score allows the objective comparison of models for multi-domain proteins even without parsing the structures into domains.

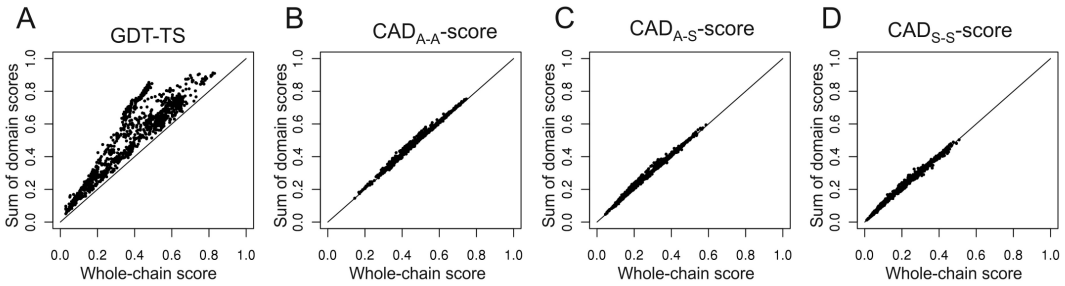


Figure 3.9: Correlation between the model scores for the whole-chain (horizontal axis) and the weighted sum of domain scores (vertical axis) for CASP9 multi-domain targets. Different plots represent the analysis of the same models using different scores: (A) GDT-TS, (B) CAD_{A-A}, (C) CAD_{A-S}, and (D) CAD_{S-S}.

CAD-score provides a balanced assessment of the inter-domain arrangement accuracy in models for multi-domain proteins

Although CAD-score shows little or no difference between domain-based and whole-chain evaluation (Figure 3.9), the important question is whether or not this reflects an adequate scoring of domain rearrangement. In our view, the accuracy of predicting mutual domain arrangement should not be judged by simple error in the directional orientation between the domains. If domains are kept together only by a connecting linker, any fixed mutual orientation might be structurally and/or biologically irrelevant (especially if the linker is flexible). In such case, the penalty for not predicting a particular orientation observed in the crystal structure would be unfair. In contrast, if domains share extensive interface, their specific arrangement suggests structural and/or biological importance and therefore should contribute to the evaluation score more significantly. In other words, the larger is the fraction of protein surface area buried at the domain interface, the larger potential impact (positive or negative) it should be able to exert on the total score of the model. Following this logic, we analyzed the expected and the observed contributions of the domain arrangement to the total model score. We defined the expected contribution as the fraction of solvent accessible sur-

face (SAS) buried at the domain-domain interface(s) of a target corrected for the accuracy of a given whole-chain model. The correction was performed by simply multiplying the SAS fraction buried at the interface by the whole-chain score. Buried SAS was determined by subtracting SAS of the whole-chain structure from the sum of SAS for individual domains and dividing by two. Of course, the definition of expected contribution of the domain arrangement is simplistic, as we consider the accuracy of the interface prediction to be the same as the average accuracy of all domains. Nevertheless, this concept is useful for exploring the relationship between the expected and the observed contributions. The observed contribution was defined as the difference between the whole-chain scores and the weighted sum of domain scores (as shown in Figure 3.9).

We analyzed the relationship between the expected and the observed contributions of the inter-domain interface prediction component to the total score of the model for both CAD-score and GDT-TS. The results are presented in Figure 3.10. We only included data for those multi-domain protein models, for which all individual domains had GDT-TS values over 0.4 (40%) and therefore were expected to represent at least a correct structural fold. Despite some data noisiness, the figure reveals a strikingly different behavior of GDT-TS and CAD-score.

Based on the data for GDT-TS (Figure 3.10 A), two important observations can be made. Firstly, the largest observed contributions to the total score are several times that of the largest expected contributions. This is the result of the GDT-TS property to strongly exaggerate the domain rearrangement making the domain-based evaluation a necessity. Secondly, this exaggeration is most strongly pronounced for models with some of the smallest expected values. In other words, given similar average quality of individual domains, models for targets having the smallest inter-domain interface are more likely to produce poor scores for the whole-chain structure.

In contrast, for CAD-score (Figure 3.10 B-D) the observed contribution

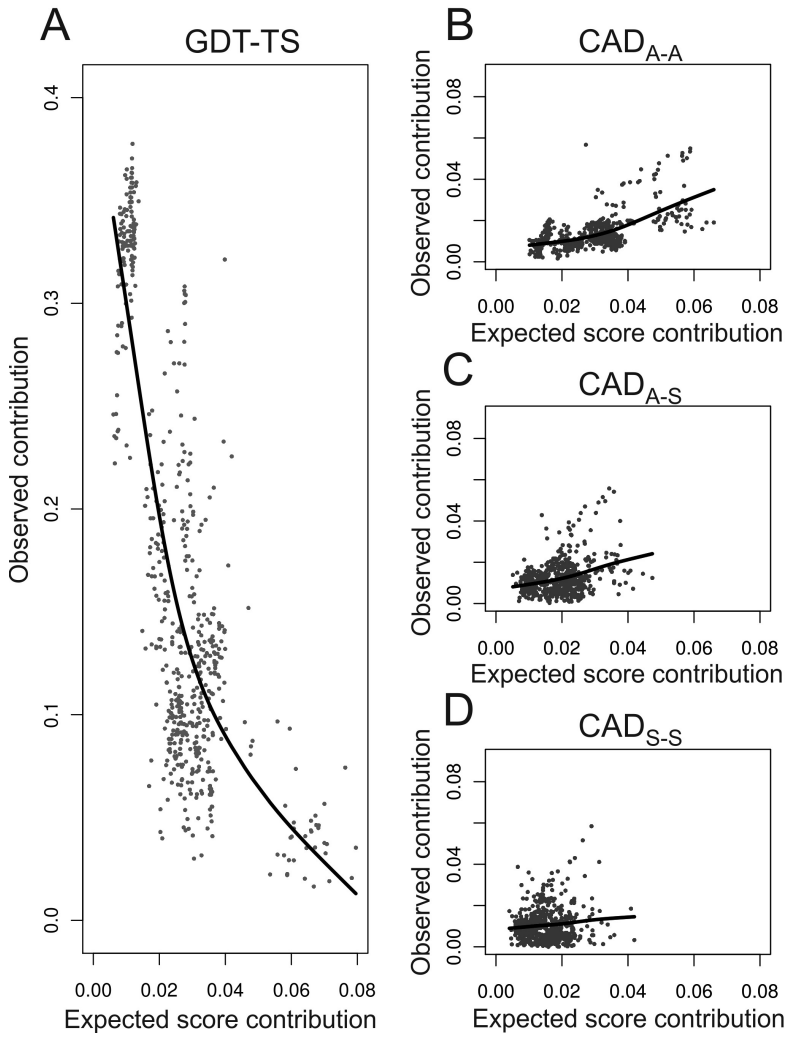


Figure 3.10: Figure 5. Relationship between the absolute values of expected (horizontal axis) and observed (vertical axis) contributions of the domain rearrangement to the total model score. For definitions of expected and observed contributions see the main text. Only data for models with GDT-TS > 0.4 (40%) for any individual domain are included. General trends for each plot are indicated by a cubic spline applied to the data (solid line). (A) GDT-TS, (B) $CAD_{A\text{-}A}$, (C) $CAD_{A\text{-}S}$ and (D) $CAD_{S\text{-}S}$ data.

of the domain arrangement score to the total score tends to increase as the expected contribution increases. The best agreement is displayed by $CAD_{A\text{-}A}$-score followed by $CAD_{A\text{-}S}$ and $CAD_{S\text{-}S}$ scores. Although the relationship is somewhat noisy, the observed contributions almost never exceed the expected ones, indicating the balanced impact of domain arrangement errors to the total score.

An illustrative example of GDT-TS problems upon evaluation of models for multi-domain targets that disappear with the application of CAD-score is provided in Figure 3.11. GDT-TS scores for both domains of CASP9 model TS453 (Figure 3.11 B) are better than those for TS245 (Figure 3.11 C). However, despite the visually very similar mutual domain arrangement in both models (Figure 3.11 A), TS453 is assigned a worse full-chain GDT-TS value. Obviously, this cannot be considered a fair assessment. In contrast, CAD-score assigns better scores not only for individual domains of TS453, but, as might be expected, also for the full-chain model. The tendency of GDT-TS to overestimate tiny differences in mutual domain arrangement is apparent even within the same model. It would be reasonable to expect the accuracy for a full-chain model to be in between the worst-scoring and the best-scoring domains. However, according to GDT-TS, both full-chain models in Figure 3.11 are worse than their least accurate domain. Again, this problem is non-existent for CAD-score.

Since the mutual arrangement of domains is not conceptually different from the arrangement of protein chains, CAD-score can also be used to evaluate the accuracy of models for protein complexes. The larger is the inter-subunit interface, the bigger impact of its prediction accuracy on the total CAD-score of the protein complex may be expected.

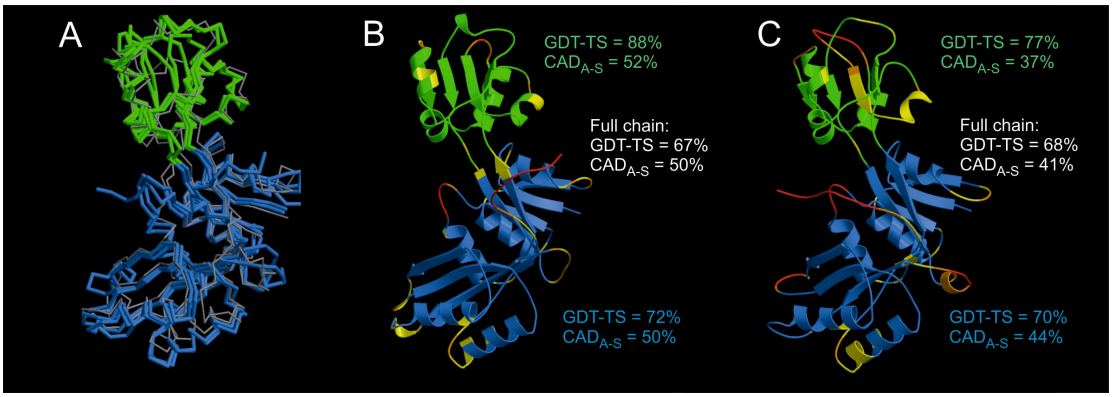


Figure 3.11: An example of multi-domain structure evaluation by GDT-TS and CAD-score. (A) Two models, TS453 and TS245, colored by domains (blue and green) are superimposed with the target T0533 structure (grey). Cartoon representations show models TS453 (B) and TS245 (C). Increasingly larger deviations of Cα-atoms are indicated by yellow, orange and red colors respectively. GDT-TS and CAD$_{A-S}$-score values in blue and green are for the corresponding domains, white — for the entire model.

3.2.4 CAD-score can directly evaluate the accuracy of inter-domain or inter-subunit interfaces

In addition to scoring models for entire multi-domain or multi-subunit structures, CAD-score provides a direct way for assessing the accuracy of the interface prediction. The only difference is the reference against which the model is evaluated. In this case the reference would be defined as contact areas between residues originating from either different protein domains (inter-domain interface) or different protein subunits (inter-subunit interface). Figure 3.12 provides specific examples of inter-domain and inter-subunit interfaces of different accuracy. First example (Figure 3.12 A) illustrates the accuracy of the inter-domain interface for two models of target T0533 that have been analyzed in detail above. It confirms once again that model TS453 has a more accurate inter-domain interface than TS245. Another example (Figure 3.12 B) features inter-subunit interfaces of different accuracy within two oligomeric predictions for target T0576. One of the two models, TS458, was identified

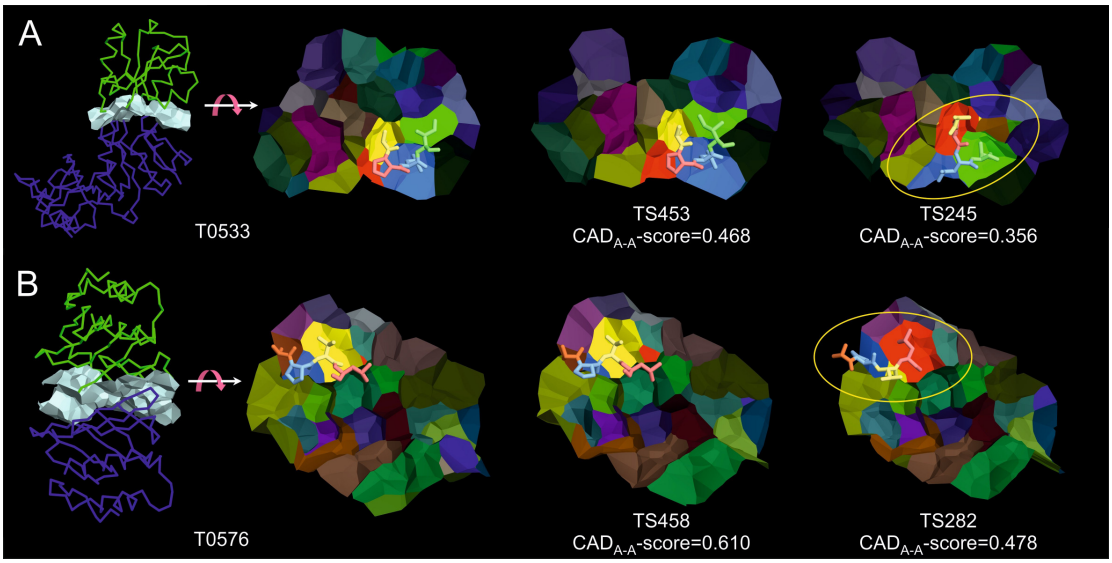


Figure 3.12: Examples of direct evaluation of the interface between domains (A) and subunits (B). (A) The inter-domain interface within the two-domain target T0533 (left) is compared with interfaces in two models, TS453 and TS245. For ease of comparison, interfaces are represented as sets of colored faces of Voronoi cells in the same orientation. Different colors correspond to different residues at the interface. Interface CAD$_{A-A}$-score values are indicated for each model. Major errors within the less accurate interface are indicated with ellipse. The corresponding protein chain fragment is shown as sticks in the target and both models. (B) The inter-subunit interface within the dimeric structure of target T0576 (left) is compared with interfaces in two multi-chain models, TS458 and TS282. Notations are the same as in (A).

by CAD-score as having the most accurate interface for this target. This CAD-score assignment completely agrees with the CASP9 assessment of oligomeric predictions.[53]

3.2.5 Discussion

The development of protein structure prediction methods and scores used for their benchmarking are interdependent. Robust and effective scores promote improvements in protein structure prediction methods. On the other hand, the overall improvement in model accuracy necessitates a more sensitive and more comprehensive evaluation. At present, due to both the improvement of structure prediction methods and the

dominance of template-based models, the focus is shifting towards the accuracy of structural features beyond the backbone. More emphasis is put on the physical plausibility of computational models. The ability to evaluate the accuracy of mutual domain arrangement in models for multi-domain proteins and the arrangement of subunits within protein complexes is also becoming increasingly important.

In this study we present CAD-score, a new model scoring function for comprehensive evaluation of structural models. CAD-score builds upon the concept of contact area difference (CAD) originally introduced by Abagyan & Totrov.[12] However, the new score differs significantly in its design and algorithmic implementation.

One of the key differences is the treatment of missing residues in the model. The original CAD only takes into account the subset of residues that are common for both target and model. In this regard it is reminiscent of RMSD, which can be calculated only on a common set of residues. In the newly defined CAD-score both the failure to include the residue into the model and the failure to predict all of its contacts are treated identically. To put it differently, CAD-score encourages the construction of the complete model. Incorrectly modeled regions can make at most only negligible improvements to the score; however, even grossly incorrect regions of the model cannot make the score worse compared to the situation when they are not modeled at all. In this respect, the design of CAD-score is similar to that of GDT-TS, which does not reward, but at the same time does not penalize grossly inaccurate regions. We believe that this is a very positive feature of a reference-based model evaluation score, as it allows testing of new bold ideas in protein structure prediction without being penalized for large local errors.

The second difference is the normalization procedure. The normalizing factor in the CAD number as proposed by Abagyan & Totrov is different for different models of the same reference structure (target). This makes the ranking of models for a given target problematic. In our case, the

normalizing term is constant for a given target, no matter how unusual or how different evaluated models are.

Yet another difference is the range of values. The originally proposed CAD number is not always guaranteed to fall within the range from 0 to 1 (0%-100%). In contrast, the newly defined CAD-score can never be outside of the [0,1] range. This is assured by "symmetric" boundaries of a maximal contact area difference for a given residue pair. We treat the failure to predict an existing contact in the same way as its "strong" over-prediction. The "strong" over-prediction is defined as the case when the absolute contact area difference is larger than the reference contact area itself. In both extremes we consider the prediction to be equally wrong, and therefore the contact area difference is bounded by the reference contact area. As a result, the sum of bounded contact area differences for the model can never exceed the sum of contact areas of the target.

Algorithms for deriving contact areas in our case and the original CAD study are also substantially different. We derive contact areas using a protein structure tessellation approach. It allows us to take into account the influence of other residues surrounding the considered residue pair. In the original CAD study, the contact area for a pair of residues is calculated in isolation, thereby tending to overestimate the size of contact area. In addition, the resolution of contact areas is different in the two methods. In contrast to Abagyan & Totrov, we calculate contact areas at the level of heavy atoms, and that allows us to derive contact areas not only for entire residues, but also for subsets of residue atoms such as main chain and side chain. In turn, this allows us to define a number of CAD-score variants, addressing different aspects of model accuracy and providing different degrees of sensitivity.

In this study we explored properties of the newly introduced CAD-score and compared it primarily with GDT-TS, a widely accepted score for reference-based model evaluation. We found that for single structural domains CAD-score shows a strong correlation with GDT-TS (Figure 3.7)

and GDT-HA (Supplementary Figure S2[6]). In both cases the strongest correlation is obtained for those CAD-score variants that include either all residue atoms or side chain in any combination. It may seem somewhat surprising that contacts between all atoms and side chains (A-S) and even those between side chains (S-S) correlate with GDT-TS better than all atom to all atom (A-A) contacts. However, side chains make up about two thirds of the protein structure and apparently their packing is what gives rise to a specific folding pattern. CAD-score variants that include only main chain atoms on at least one side of the contact show somewhat weaker correlation, with the main chain to main chain (M-M) variant occupying the lower end. The character of main chain to main chain contacts differs significantly depending on the secondary structure type. While in β-sheets these contacts are defined by the global topology, for α-helices they are local and are mostly defined by the accuracy of secondary structure assignment. Apparently, the lack of non-local contacts within α-helical structures is a major factor in making the M-M variant least correlated with GDT-TS (Supplementary Figure S1[6]).

One of the important advantages of CAD-score compared to GDT-TS and other structure superposition-based methods is the robust evaluation of models for multi-domain proteins and protein complexes. Our analysis showed that in contrast to GDT-TS, CAD-scores of individual domains and the whole-chain structure are tightly connected (Figure 3.9). Moreover, the accuracy of the inter-domain or inter-subunit interface is an integral part of the total score. The more extensive is the interface, the more potential improvement or deterioration to the total score it may contribute (Figure 3.10). Although the domain-based model evaluation is perfectly possible, CAD-score removes the necessity to chop the structure into domains to get meaningful results. Moreover, even if the structure is split into domains, the performance of CAD-score cannot be strongly affected by imprecise or even outright wrong domain boundary definition, which would have a large impact in the GDT-TS-based evaluation.

According to CAD-score, the accuracy of the model depends only on how closely the contact areas between residues (or subsets of residue atoms) correspond to those in the reference structure. However, what may seem a simplistic definition of model accuracy in fact incorporates many structural features such as interatomic distances, dihedral angles, hydrogen bonds and bond lengths. Protein structure prediction methods trained using a particular model evaluation score, in some cases may "improve" their performance by optimizing some of the model structural parameters at the expense of others. Here, we showed that CAD-score is associated with physical realism of models much stronger than GDT-TS (Figure 3.8). In particular, this property of CAD-score may be relevant for assessing model refinement, which turns out to be a surprisingly hard problem.[104]

Although we developed the new CAD-score with the reference-based model evaluation in mind, the approach may be a valuable tool for other tasks such as clustering of structural models. Model clustering is one of the steps employed by many current protein structure prediction approaches, especially if there are no suitable structural templates. The clustering step is used for the identification of near-native structures from a large set of candidate structures (decoys). Since contact areas between residues directly reflect the strength of physical interactions, CAD-score values may be more suitable for grouping models with similar energies compared to Cartesian distance-based approaches such as RMSD or GDT. As clustering typically involves large numbers of models, the clustering method needs to be fast. In CAD-based clustering, the slowest step is the computation of contact areas between residues in individual models. However, once it is done, subsequent calculation of pairwise CAD-scores is very fast. An example of model clustering results using CAD-score is presented in Supplementary Figure S4.[6]

CAD-score is based on interatomic contacts and as such it is not exclusively restricted to protein structures. Similar approach could be ap-

plied for evaluation of models of other biomolecules forming complex 3D structures such as RNA. Similarly, evaluation of the protein-protein interface (inter-domain or inter-subunit) accuracy could be easily extended to the more general case of protein-ligand interfaces. Obviously, the CAD-score based evaluation would be most appropriate for large interfaces such as those in protein-nucleic acids complexes, but perhaps it may be sufficiently informative even for interfaces between proteins and small molecules.

In summary, the newly introduced CAD-score has a number of attractive properties. It is based on physical contacts between residues, thereby directly reflecting interactions within the protein structure. It is a continuous, threshold-free function that returns quantitative accuracy scores within the strictly defined boundaries. The definition of CAD-score does not contain any arbitrary parameters. CAD-score provides a single uniform framework for assessing single-domain, multi-domain and even multi-subunit protein structural models of varying degree of accuracy and completeness. While being highly correlated with GDT-TS on single-domain structures, CAD-score displays a stronger emphasis on the physical realism of models. We believe that all these attractive properties make CAD-score a valuable tool for the development and assessment of protein structure prediction and refinement methods as well as for clustering models based on their mutual similarity.

3.3 Testing results for RNA structures

3.3.1 Testing data sets

PDB structure set

For comparative analysis of contacts in RNA and proteins we used experimentally determined 3D structures that were selected from PDB (as

of 2013.06.01). The selection included only x-ray structures solved at the resolution of 3.0 Å or better. In addition, 30% sequence identity cutoff was applied to the initial selection to make the set non-redundant.

RNA model test set

We used RNA models and corresponding experimental structures available as part of RNA-puzzles,[105] a collective experiment for blind RNA structure prediction. We used the data of all the challenges completed as of 2013, namely 1, 2, 3, 4, and 6. Prior to the analysis, residue numbering and chain identities of raw models were set to match the naming of corresponding nucleotides in experimental structures. No coordinates of any model were modified.

3.3.2 Base-base contacts dominate RNA 3D structures

Physical basis of interatomic contacts is the same in both protein and RNA 3D structures. We considered that therefore the contact area-based model evaluation score as defined in its general form should be also feasible for RNA. On the other hand, considerably different roles of main chain (backbone) and side chain (base) atoms in defining secondary and tertiary structures in proteins and RNA compelled us to perform a more thorough investigation of corresponding contacts.

To investigate the contribution of different types of contacts in proteins and RNA we performed the following analysis. We selected well-resolved non-redundant protein and RNA structures from PDB (see Methods for details). For every structure we computed the total area of all contacts as well as fractions of contact area contributed by three types of contacts: 1) main chain-main chain (backbone-backbone), 2) side chain-side chain (base-base) and 3) the remaining contacts that consist of side chain-main chain (base-backbone) and main chain-side chain (backbone-base) contacts. The results (Figure 3.13) show that, except for

the smallest structures, the individual contributions to the total contact area by the three contact types are largely independent of the structure size in both proteins and RNA. However, these contributions differ significantly in proteins and RNA. In the case of proteins the contributions by the three types of contacts are well-balanced. Although the share of the main chain-main chain contacts is the largest (36% on average), the fractions of both side chain-side chain and the remaining contacts are comparable (correspondingly 30% and 34% on average). In the case of RNA the picture is dramatically different. Base-base contacts strongly dominate, on average making up about half (49%) of all contact areas. These results indicate that base-base interactions in RNA make a significantly larger impact than side chain-side chain interactions in proteins and therefore merit a more detailed analysis.

3.3.3 Contact area is an effective means for describing base-base interactions

To perform a more detailed analysis of RNA base-base contacts, we divided them into bins according to the size of contact area. The area size corresponds to the physical impact of a contact; therefore, we also looked at the cumulative impact of contacts (frequency multiplied by the area size) for each bin. To compare our results with established approaches, for the same set of RNA structures we identified base-base interactions using MC-Annotate, a widely used RNA annotation method.[106] MC-Annotate detects and annotates base-base interactions using a procedure involving both geometric and probabilistic considerations.[106,107] Figure 3.14 shows the comparison of base-base contact data derived using our approach and MC-Annotate. Since MC-Annotate does not compute contact areas, its contact data was generated by our approach according to the MC-Annotate annotations. If the contact frequency is considered (Figure 3.14 A, left), the two approaches show a reasonably close agreement, except for the contacts characterized by small area sizes. Appar-

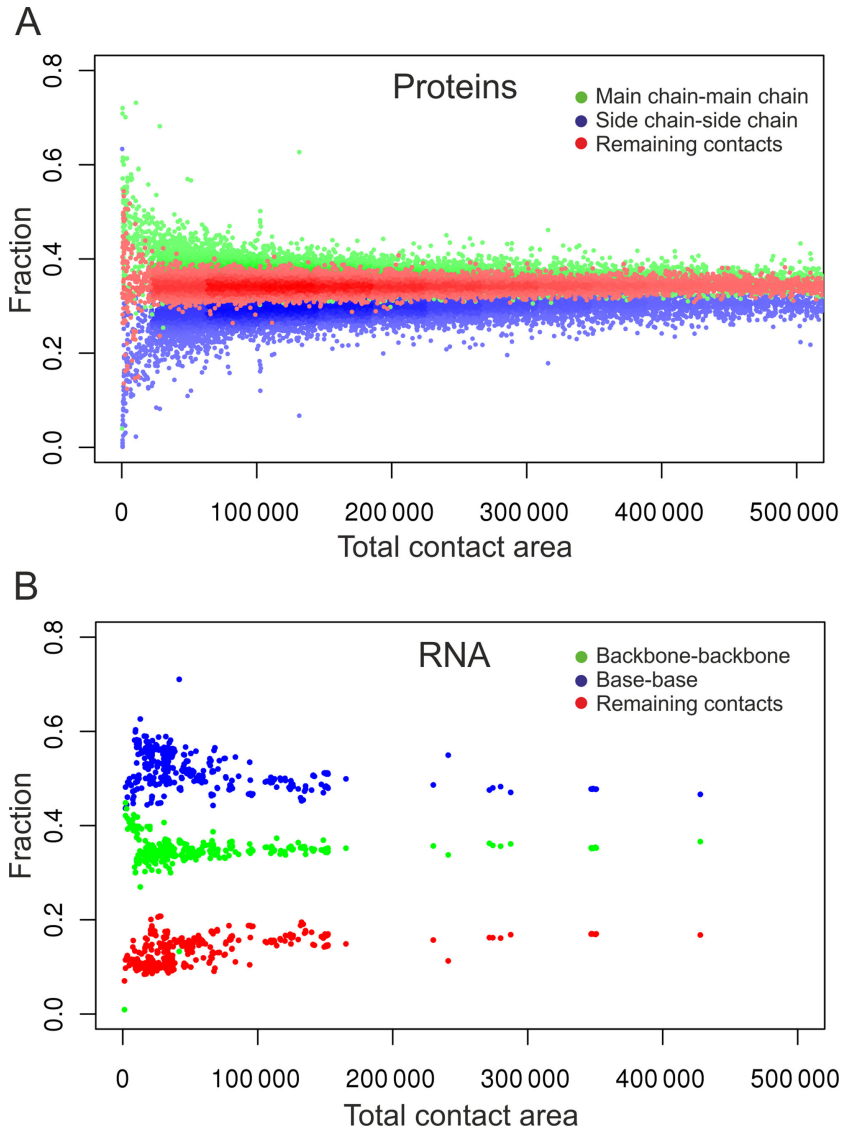


Figure 3.13: Contribution of the three components of all atom-all atom contacts to the total contact areas in 13336 protein (A) and 445 RNA (B) structures from PDB.

ently, MC-Annotate does not annotate bases as interacting if they barely contact each other. If the cumulative area size is considered (Figure 3.14 A, right), the agreement is significantly better, since contacts with the negligible area size, despite their abundance, contribute almost nothing to the cumulative impact. One of the conclusions that can be made from this comparison is that the definition of contacts only as binary information (present/absent) may be misleading. A more appropriate way is to also consider contact strength, expressed here as the contact area size.

3.3.4 Simple contact-based definition provides a useful approximation of base stacking and base pairing

There are two major types of base-base interactions: base stacking and base pairing. Therefore, it would be desirable to assign at least approximately base-base contacts to one of these two interaction types. We devised an extremely simple definition to partition base-base contacts into the two types (as described in the method definition section and illustrated in Figure 3.5) and applied it to the base-base contact data (Figure 3.14 A). If we consider undivided base-base contacts, there are three peaks common to both the frequency plot (Figure 3.14 A, left) and cumulative area plot (Figure 3.14 A, right). According to our definition, the two rightmost peaks correspond to base stacking (Figure 3.14 B) while the leftmost of the three peaks corresponds to non-stacking contacts (Figure 3.14 C). To see how well this partitioning works, we compared it with the classification provided by MC-Annotate. Again, the agreement with MC-Annotate improves if the total cumulative contact area instead of the contact frequency is considered. In particular, stacking interactions characterized by the largest contact areas agree almost ideally. At the same time even for relatively large contact areas there is a visible gap between cumulative values of base stacking curves (Figure 3.14 B). According to our visual analysis at least some of these cases can be assigned to either adjacent or non-adjacent base stacking interactions (examples are pro-

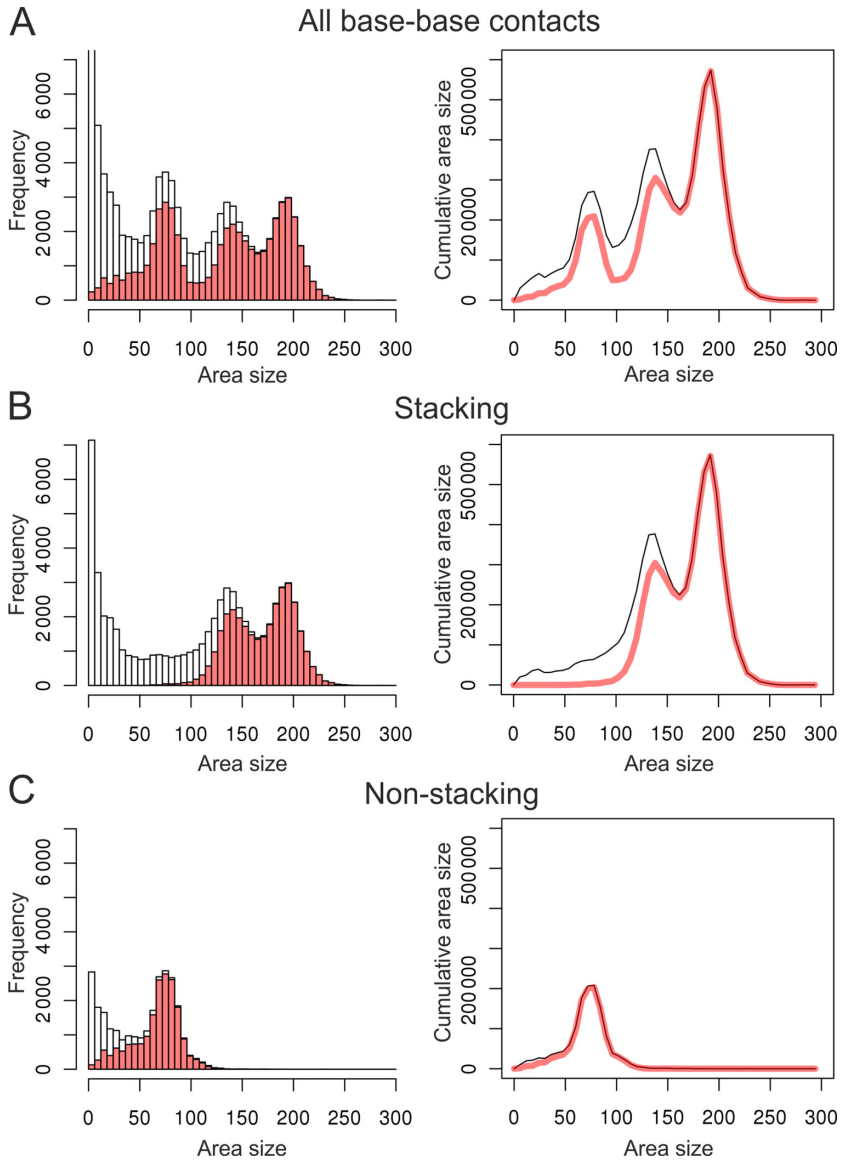


Figure 3.14: Dependence of the base-base contact frequency (left) and cumulative contact areas (right) on the contact area size. The data on all base-base contacts (A), base stacking (B) and non-stacking (pairing) (C) are shown. Gray bars and lines correspond to contacts determined by the approach reported here; red bars and lines correspond to definitions by MC-Annotate.

vided in Supplementary Figure S1[8]). Many other differences represent inter-strand base-base overlaps. Although these overlaps are not identified as base stacking by MC-Annotate, many of them feature fairly large contact areas indicating important contribution to the interaction network. Quite unexpectedly, although non-stacking contacts (Figure 3.14 C) do not involve special considerations for hydrogen bonding, they very closely recapitulate base pairing interactions defined by MC-Annotate.

In the case of unambiguously classified contacts (there was an agreement between our approach and MC-Annotate) we also looked into the nature of stacked bases and the number of hydrogen bonds in base pairs (Supplementary Figure S2[8]). As might be expected, purine-purine stacking dominates the largest contact areas, while pyrimidine-pyrimidine stacking is at the lower end of stacking contact area size. Purine-pyrimidine stacking shows bimodal distribution. As for base pairs, most of them have two or three hydrogen bonds. Only a small fraction of contacts, both in numbers and in the cumulative area size, correspond to other base pairings.

Since our approach considers all base-base contacts, their division into stacking and non-stacking is, of course, oversimplification. However, even this extremely simple classification is able to provide a useful distinction between most stacking and pairing interactions and thus to reveal model errors specific to each interaction type.

3.3.5 CAD-score provides a direct link between local discrepancies in an RNA model and the global score

By its nature CAD-score is a local score, as it analyzes discrepancies only within the immediate 3D neighborhood. The most inclusive CAD-score variant quantifies discrepancies that involve entire nucleotides by taking into account all atom-all atom contacts. Additionally defined partial CAD-scores measure other types of discrepancies by considering

contacts between various sets of nucleotide atoms (all atoms, backbone or base) or even different types of base-base contacts (stacking, non-stacking). A simple combination of the local discrepancies of each kind produces a global score that summarizes the overall accuracy of a model with respect to the reference structure.

Different CAD-score variants allow addressing different questions. However, for practical applications the variants that consider either all atom-all atom or base-base contacts appear to be the most useful. The usefulness of all atom-all atom CAD-score is understandable, since contacts between all atoms represent the most complete description of the structure. On the other hand, as we have shown, base-base contacts represent the dominant contact fraction in the RNA and are largely responsible for its specific 3D shape. Therefore, the base-base CAD-score and its partial (stacking and non-stacking) scores can be particularly useful in figuring out the cause of discrepancies between the two structures.

Figure 3.15 shows an example of the evaluation of both local and global accuracy of two RNA-puzzles models (Challenge 3) using major variants of CAD-score. The two models are of different accuracy, appropriately reflected by the "summarizing" CAD-score values. Furthermore, both the local discrepancies and the global accuracy values reveal that one of the major reasons of the second model being inferior to the first one is poorly modeled non-stacking (base pairing) interactions. Often, base stacking and non-stacking CAD-score values alone may reveal the source of error and indicate whether the errors are confined to specific regions or dispersed throughout the modeled structure (Supplementary Figure S3[8]).

3.3.6 CAD-score is an effective RNA model ranking index

CAD-score efficiently accounts for all the local discrepancies between a model and the reference structure. The question is whether the global

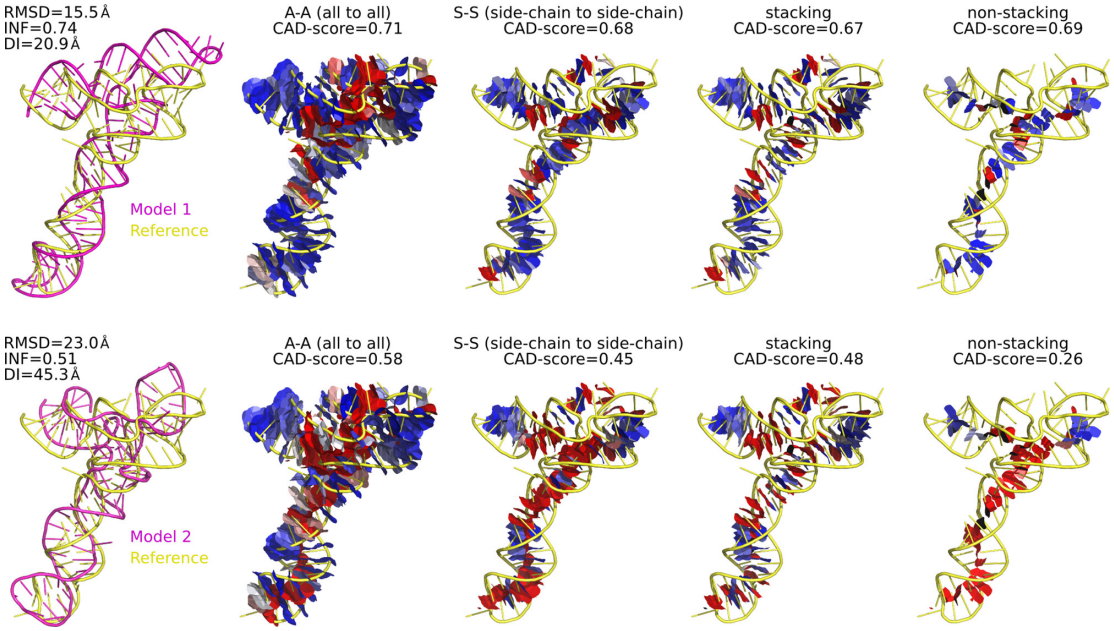


Figure 3.15: Example of CAD-score evaluation of two models of different accuracy at both global and local levels. Different panels show the contact areas considered by the indicated CAD-score variants. Contacts are represented as faces of the Voronoi cells constrained by the contact spheres. Blue-white-red color gradient represents the accuracy of reproduced contacts (blue — accurate, red — inaccurate).

score, expressed as a simple combination of local errors, is also effective in ranking models by their overall accuracy. Model ranking is inherently subjective, because of the multiple features that have to be assessed simultaneously. On the other hand, to be considered effective, a new evaluation score should at least roughly agree with the currently used scores. To analyze model ranking by CAD-score, we compared it with the scores used in the RNA-puzzles experiment,[105] namely, Interaction Network Fidelity (INF), Deformation Index (DI) and RMSD. We took all the models generated as part of the RNA-puzzles experiment, scored them against corresponding reference structures and analyzed how well CAD-score correlates with each of the other three scores. It turned out that CAD-score correlates best with INF, less well with DI and least with RMSD. This order does not depend on whether we use Pearson's correlation coefficient, which assumes the linear relationship between scores, or Spearman's ranking correlation coefficient, which makes no such assumption. Figure 3.16 shows the relationship between two representative CAD-score variants (all atom-all atom and base-base) and INF, DI and RMSD. The correlation between CAD-score and INF reaches as high as 0.95 indicating a good agreement between the two scores. The agreement with DI and RMSD is worse, but correlation values are still fairly high. Diverse models available as part of the RNA-puzzles experiment represent an excellent test set, but their number is relatively small (104 models for 5 reference structures). To make the test more rigorous, we performed the same analysis using over 30 000 models (for 67 reference structures) of the randstr decoy set.[60] Although correlation coefficients calculated using the randstr decoy set are slightly smaller, we obtained the same correlation trend: INF > DI > RMSD (Supplementary Table S1,[8] Supplementary Figure S4[8]). Thus, overall results of the correlation analysis indicate that CAD-score model ranking properties are closest to those of INF, reflecting their common focus on the similarity of interactions.

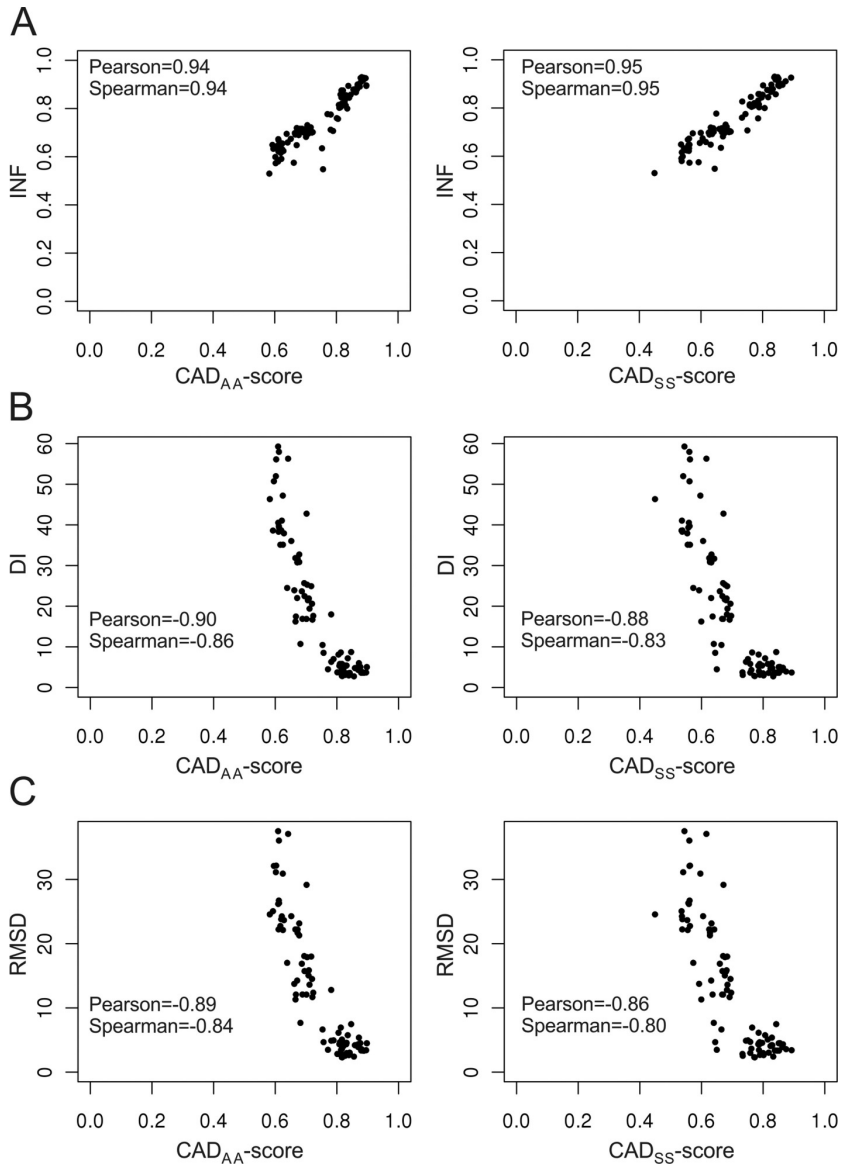


Figure 3.16: Relationship between CAD-score and INF (A), DI (B), and RMSD (C). Data is shown for CAD_{AA}-score (left) and CAD_{SS}-score (right). For each plot Pearson's correlation coefficients and Spearman's ranking correlation coefficients are indicated.

3.3.7 CAD-score favors physical realism of RNA structural models

Correlation analysis revealed that CAD-score shows a fairly close agreement with other scores, INF in particular. However, inevitably there are cases when the scores disagree. For example, it may be that according to CAD-score, model A is more accurate than model B, but according to another score it is the opposite. An important question is which score to trust in such cases. One way to address this question is to consider physical realism of models, the feature that does not depend on how closely a model agrees with the reference structure.[6] The idea is that if we take two scores, the score that shows stronger tendency to select physically more realistic models as the more accurate ones is likely to be more objective.

We asked how CAD-score compares to the other three scores (INF, DI and RMSD) in the light of physical realism of models. For the assessment of physical realism we used the *clash score*, *bad angles* and *bad bonds* as reported by MolProbity,[102] a well-known structure validation software suite. We considered one of the two models to be more physically realistic if the model was better according to at least one of the three MolProbity scores and the other two scores did not contradict that (for example, the *clash score* was lower, while the values of *bad bonds* and *bad angles* were identical). To perform this analysis, we used the RNA-puzzles data set. For every reference structure we compiled all the possible model pairs and identified those in which the relative ranking of models was in conflict according to CAD-score and either INF, DI or RMSD. We then analyzed the same model pairs with MolProbity. It turned out that CAD-score agreed with MolProbity more often than did any of the other three scores (Table 3.1). As could be expected by the highest correlation values, the smallest number of conflicting rankings was between CAD-score and INF. Nevertheless, the support of CAD-score by MolProbity was stronger than that of INF. In particular, the CAD_{SS}-score (evaluating base-base

Model pairs with conflicting ranking	Supported by Molprobity			
	First score	Model pairs	Second score	Model pairs
	CAD_{AA}-score (all atom-all atom contacts)			
94	CAD_{AA}-score	54 (57%)	INF	40 (43%)
133	CAD_{AA}-score	109 (82%)	DI	24 (18%)
145	CAD_{AA}-score	122 (84%)	RMSD	23 (16%)
	CAD_{SS}-score (base-base contacts)			
80	CAD_{SS}-score	54 (67.5%)	INF	26 (32.5%)
143	CAD_{SS}-score	121 (85%)	DI	22 (15%)
167	CAD_{SS}-score	140 (84%)	RMSD	27 (16%)

Table 3.1: MolProbity's "judgment" on model pairs with the conflicting assignment of accuracy by CAD-score and either INF, DI or RMSD.

contacts) most closely corresponding to INF was supported in about two out of three cases. These results show that CAD-score favors physical realism of models more strongly than either INF, DI or RMSD.

3.3.8 CAD-score accounts for RNA model completeness

Structural models may not necessarily include all the residues. Most often, difficult-to-predict structural regions are omitted. A reference-based model evaluation score should be able to take this into account properly in order to make a fair comparison. We asked how well CAD-score, INF, DI and RMSD cope with structural models that are heterogeneous as to their completeness. To this end we performed the following analysis using RNA-puzzles models. We iteratively truncated each model by 20% (removing equal number of residues from both 5′ and 3′ ends) and recalculated the scores at every step. We monitored the number of models for which the score has improved after each truncation step. The idea behind this test was that if the removed fragment had at least some correct features, its removal should make the score worse. Even if the removed fragment was completely incorrect, the score of the truncated model should be the same at best.

The results of this test are presented in Table 3.2. CAD-score (both all atom-all atom and base-base contacts) did not improve even once upon

Model completeness	80%	60%	40%	20%	
	Number of models with the increased score				**Total**
CAD_{AA}-score	0	0	0	0	**0**
CAD_{SS}-score	0	0	0	0	**0**
INF	0	5	1	1	**7**
DI	18	19	17	36	**90**
RMSD	56	88	69	102	**315**

Table 3.2: The effect of model truncation on evaluation scores.

iterative truncation of models. INF has improved seven times, DI - 90 times and RMSD - 315 times. Thus it may be concluded that CAD-score is suitable for evaluation of a mixture of complete/incomplete models. INF is not as good as CAD-score, while DI and RMSD could be applied only to models consisting of exactly the same residues.

3.3.9 Discussion

Our results show that contact area-based approach can be highly effective in quantifying discrepancies between modeled and reference structures not only for proteins but also for RNA. The same general definition of CAD-score can be applied to the both types of macromolecules despite their significant differences.

A number of features make CAD-score attractive as a similarity measure. First of all, since CAD-score is based on comparing contact areas, it does not require structure superposition. Moreover, contact areas not only define physical contacts in the structure, but also indicate their relative strength. Therefore, CAD-score reflects physical interactions that are relevant to the formation and stability of 3D structure. The global CAD-score is constructed by accounting for all the local discrepancies, thereby providing a transparent relationship between local errors and the overall model accuracy. Unlike some other scores such as RMSD or DI, CAD-score has a fixed value range, simplifying the comparison of different

models. One other attractive feature of CAD-score is that its definition does not involve any arbitrary parameters. In fact, the only adjustable parameter used in computing CAD-score is VDW radii of heavy atoms.

In addition to CAD-score based on all atoms, a number of partial CAD-score variants can be defined based on subsets of residue atoms. Since RNA and proteins differ considerably, we explored the relative impact of contacts contributed by either main chain (sugar-phosphate backbone) or side chain (base). Given the importance of base-base interactions in the formation and maintenance of both the secondary and the tertiary RNA structures it came as no surprise that base-base contact areas represent by far the largest fraction of all contact areas. Typically, base-base interactions are classified into only two types: base stacking and base pairing. Therefore, we reasoned that it would be useful for CAD-score also to have the ability to consider these two types of interactions individually. We devised an extremely simple partitioning of all base-base contacts into two types (stacking and non-stacking). Our intention was not to substitute RNA annotation algorithms but rather to provide a useful approximation of the two interaction modes. RNA annotation algorithms are selective in defining base stacking and base pairing, while our approach takes into consideration all physical contacts. Thus, it was surprising to see that our approach and the annotation by MC-Annotate show fairly close agreement. It should be emphasized, however, that the agreement is good only when the cumulative contact area and not the contact count is considered. Perhaps most surprising observation was that non-stacking contact areas very closely correspond to base pairings defined by MC-Annotate. Since our definition of stacking/non-stacking contacts does not involve any special treatment of hydrogen bonds, such close agreement suggests that the absolute majority of significant non-stacking contacts originate from base pairs. Disagreement between the contact area approach and MC-Annotate largely coincides with smaller areas of stacking contacts. Many of these cases represent either tiny over-

laps of base planes or bases contacting at an angle and therefore do not represent canonical base stacking. However, some large base overlaps ignored by MC-Annotate appear to represent typical base stacking, suggesting that the definition of stacking in current annotation algorithms could be improved. The contact area approach may help to increase the sensitivity of detecting candidate stacking interactions that subsequently could be refined using additional criteria. Overall, the analysis of base-base interactions suggested that our contact-based definition is specific enough to enable CAD-score to focus onto discrepancies related to base stacking and base pairing separately.

No matter how a score is defined, its usefulness depends entirely on the performance. To make a thorough analysis of CAD-score performance, we compared it with the three other scores, INF, DI and RMSD, used for the model assessment during the first round of the RNA-puzzles experiment.[105] We made a comparison of scores according to their model ranking properties, the preference of physical realism and the ability to take into account model completeness. These tests revealed that, according to the overall behavior, CAD-score is most similar to INF, less so to DI and least similar to RMSD. Taking into account that DI was designed as an attempt to improve RMSD properties,[61] the trend of CAD-score agreement with other scores is exactly what should be expected from an effective score. The similar behavior of CAD-score and INF should not be surprising, since both are assessing local interactions. However, despite the strong correlation between these two scores, CAD-score appears to be superior.

Firstly, CAD-score shows a stronger preference towards more physically realistic models than INF. We believe that this is an important property since the improvement according some reference-dependent score should not come at the expense of stereochemical quality, which is the reference-independent property. The stronger emphasis on physical reality by CAD-score might be due to the fact that CAD-score takes into

account all physical contacts, while INF uses only selected set of interactions defined by the structure annotation. Furthermore, CAD-score takes into account contact strength. The penalty for missed contact depends on its area size. Missing important contacts (large contact area) is penalized strongly, while missing contacts with negligible contact area have almost no effect on the score. In contrast, INF considers only the presence or absence of interactions, without taking into account how important they are.

Secondly, CAD-score is able to properly account for the absence of nucleotides or their parts in a model. Although INF, unlike DI or RMSD, shows similar trend, it is not entirely consistent. When all the evaluated models are complete this feature has no bearing on model comparison. However, if models generated by different methods are compared, some heterogeneity of model completeness might be expected. In such cases the ability to account for missing regions would be important.

In summary, we believe that the attractive properties of CAD-score relevant to the RNA 3D structure make CAD-score an important addition to the reference-based RNA structure evaluation methods. Moreover, taking into account the applicability of the method to both nucleic acids and protein 3D structures, CAD-score offers new capabilities for the assessment of 3D structural models of protein-nucleic acid complexes.

3.4 CAD-score web server

Previously we introduced Contact Area Difference Score (CAD-score), a method to quantify both local and global similarity of structures and interfaces.[6] CAD-score was initially developed for proteins; however, we extended its application to RNA 3D structure.[8] In general, the universal nature of the method makes it applicable to any major type of macromolecular structures. Here, we describe a web-based interface for the CAD-score computation and interactive analysis of the results for pro-

teins, nucleic acids and their complexes.

3.4.1 Input

Input data

The inputs to the web server are macromolecular structure files in PDB format: one file for the reference structure (target) and one or more files of structures (models) to be compared with the target. A user is also asked to specify the type of the input structures: proteins, nucleic acids or protein-nucleic acid complexes.

By default, the web server verifies that the residue sequence, residue numbering and chain naming in each model are consistent with the target structure. If inconsistencies are detected, the web server stops and reports an error. The user can alter this behavior by selecting an option to allow mismatches between model and target sequences. In such case only the consistency of residue numbering and chain naming is verified. This option might be useful for comparison of structures with one or few mismatches such as native structures and point mutants.

Evaluation modes

The CAD-score web server provides a flexible way to choose which residue-residue contacts to analyze. The most straightforward choice is to analyze contacts within the entire structure.

Another option is to evaluate inter-chain interfaces. In this case by default only contacts between residues belonging to different chains are analyzed. The user may choose to extend the reference set of contacts by additionally including contacts between the interface residues from the same chain. In both cases the reference set of interface residues is the same, only the contact reference sets differ.

Finally, the most flexible option is to instruct the CAD-score web server

String	Meaning	Examples of evaluation target
(A)(B)	Contacts between chains A and B	Interface between two subunits in a multisubunit structure
(A,B)(C)	Contacts of chains A and B with C	Interface between a protein dimer and the RNA in a protein-RNA complex
(A)(A)	Contacts between residues within chain A	Contacts within a single subunit in a multisubunit structure
(A1-A9,A21-A90)(B1-B90)	Contacts between two explicitly specified groups of residues	Interface between two domains in a multisubunit structure

Table 3.3: Examples of custom selection strings that define subsets of residue-residue contacts.

to analyze only contacts between custom selections of residue groups. A selection can be specified by writing chain names or residue identifiers (or ranges of residue identifiers) in a simple notation. Examples of custom selections are given in Table 3.3.

3.4.2 Output

Representation of global scores

The default view of the results generated upon the completion of a user-submitted job is a summary, presented as a sortable table of global score values. Independently of the molecule type the table has columns for "A-A", "A-S" and "S-S" CAD-scores. Other table columns are specific for the macromolecule type. In the case of proteins, TM-score, GDT-TS and GDT-HA scores as computed by the TM-score software[46] are included. In the case of nucleic acids, "S-S" CAD-score evaluating base-base contacts is further subdivided into S-S stacking and S-S non-stacking scores. A table of global scores for protein-nucleic acid complexes has only the columns that are available for both proteins and nucleic acids. A table of global scores including all CAD- score variants is also available for downloading in the flat text format. In addition to the summary table, there are sortable tables of global scores for specific CAD-score variants

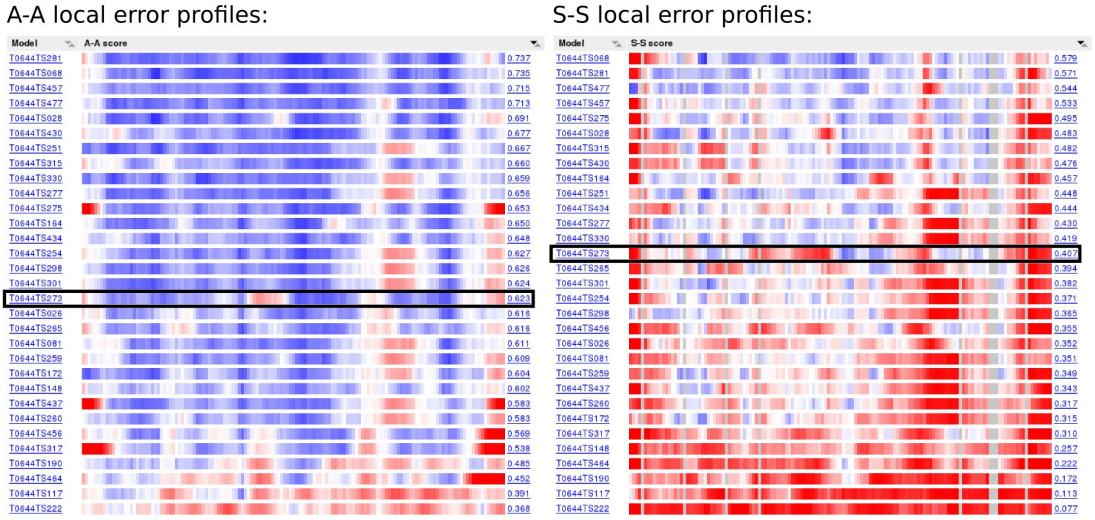


Figure 3.17: Example of a global view of A-A and S-S local error profiles generated for some models of the CASP10 prediction target T0644. Black frames indicate profiles for the same model, TS273.

for all processed models. A global score for each model in these tables is accompanied by the color-coded profile of local errors (described in detail in the following section). This view is particularly useful for the simultaneous analysis of multiple models as it enables to contrast and compare local discrepancies of individual models in the overall context (Figure 3.17).

Visualization of local errors

Each global CAD-score value in either sortable table is linked to a detailed report of local errors for the corresponding model. Local errors are primarily displayed as profile images where the value of the local error for each analyzed residue is color-coded using blue-white-red gradient. Blue and red colors represent good and poor agreement, respectively. For each CAD-score variant four profile versions are generated using smoothing windows of 0 (no smoothing), 1, 2 and 3 residues on both sides of each analyzed residue. Additionally, four profiles of raw local errors with the same smoothing windows are generated for compar-

ison. The actual values of both raw and normalized local errors without smoothing can be viewed in plain text format. The combined contacts file detailing corresponding contact areas for each residue in the target and the model is also available as a text file for the off-line analysis.

Figure 3.18 provides examples of local error profiles and their relation to the superimposed contact maps for the model and the target. Such maps, generated by the server, show contacts represented as colored points on the black background. Residue contacts in the target are red, in the model are green, and the color of coinciding contacts consists of red and green components mixed with a ratio proportional to the corresponding contact areas in the target and in the model. Therefore, yellow color indicates that the areas of corresponding contacts are of approximately same size. Images of both local error profiles and contact maps are interactive: a user can click on them to see the corresponding residue numbers. Another way to analyze local errors is to visualize them in the context of 3D structures with Jmol, an interactive molecular viewer. Local errors are converted into the B-factor values of PDB files for both target and model and are represented by the same color gradient as in the corresponding linear profiles. Local discrepancies mapped onto 3D structures are exemplified with the predicted and experimental protein structures (Figure 3.19) and with the two x-ray structures of a protein crystallized in a free state and with the bound DNA (Figure 3.21).

Local error profiles can also be used when analyzing subsets of residue-residue contacts. An example in Figure 3.20 features comparison of a decoy and the native structure of a protein-RNA complex used in a study aimed at scoring protein-RNA docking solutions.[108] The complete local error profile shows that in the decoy structure the contacts both inside the protein chain and inside the RNA chain are reproduced relatively accurately. However, the comparative contact map of the protein-RNA complex shows that contacts between protein and RNA differ significantly. Therefore, it is also useful to analyze the local error profile produced only

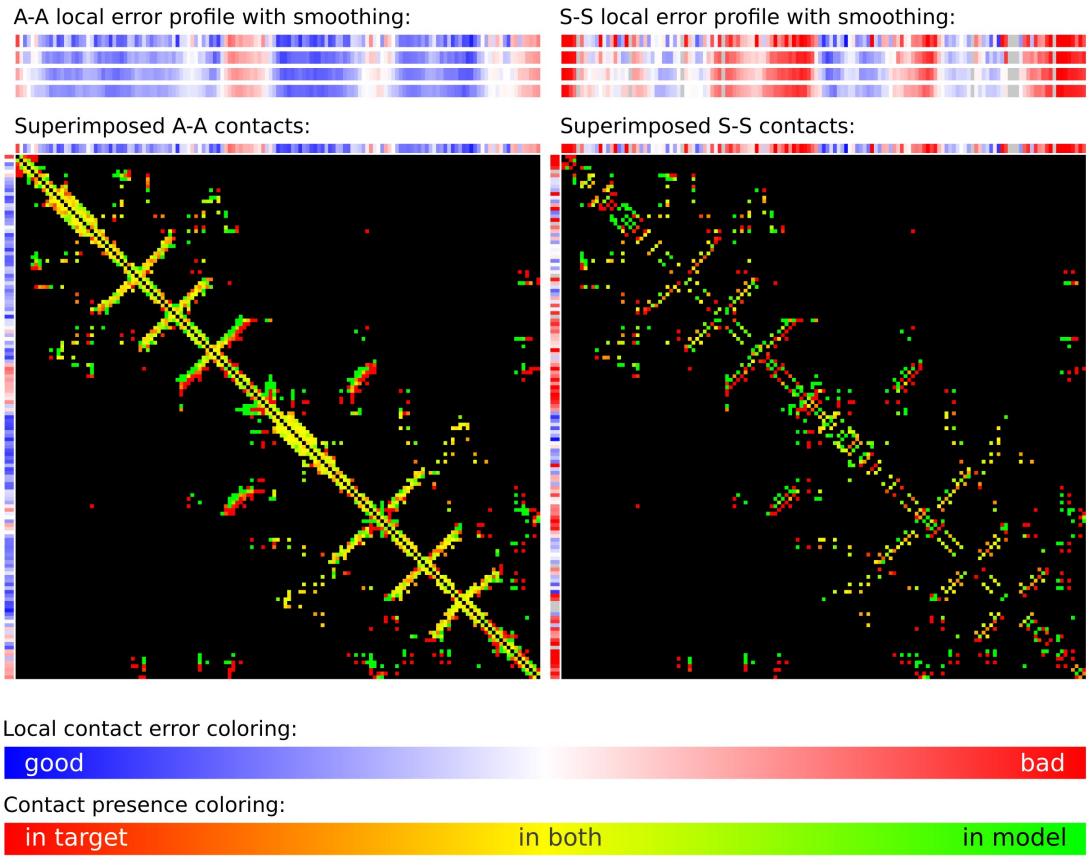


Figure 3.18: Local error profiles and superimposed contact maps of A-A and S-S contacts for the model highlighted in Figure 3.17.

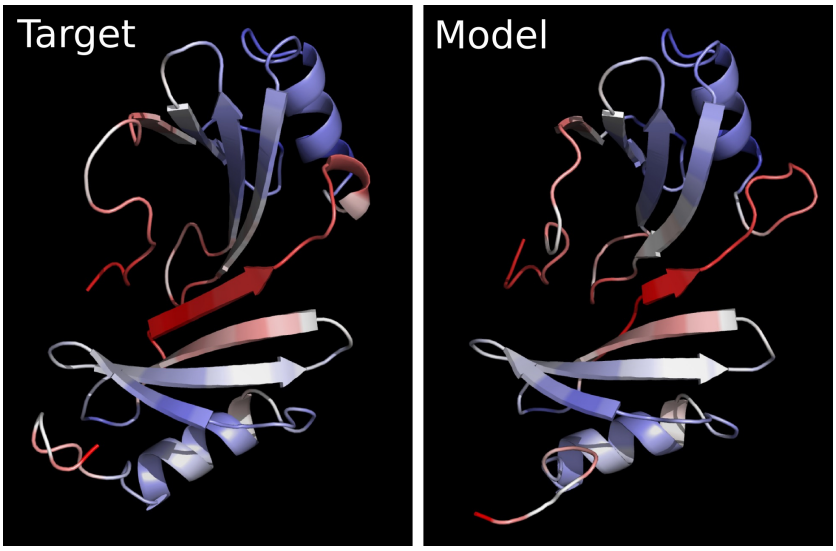


Figure 3.19: Experimentally solved (target) and predicted (model) structures colored according to the local errors in the model highlighted in Figure 3.17.

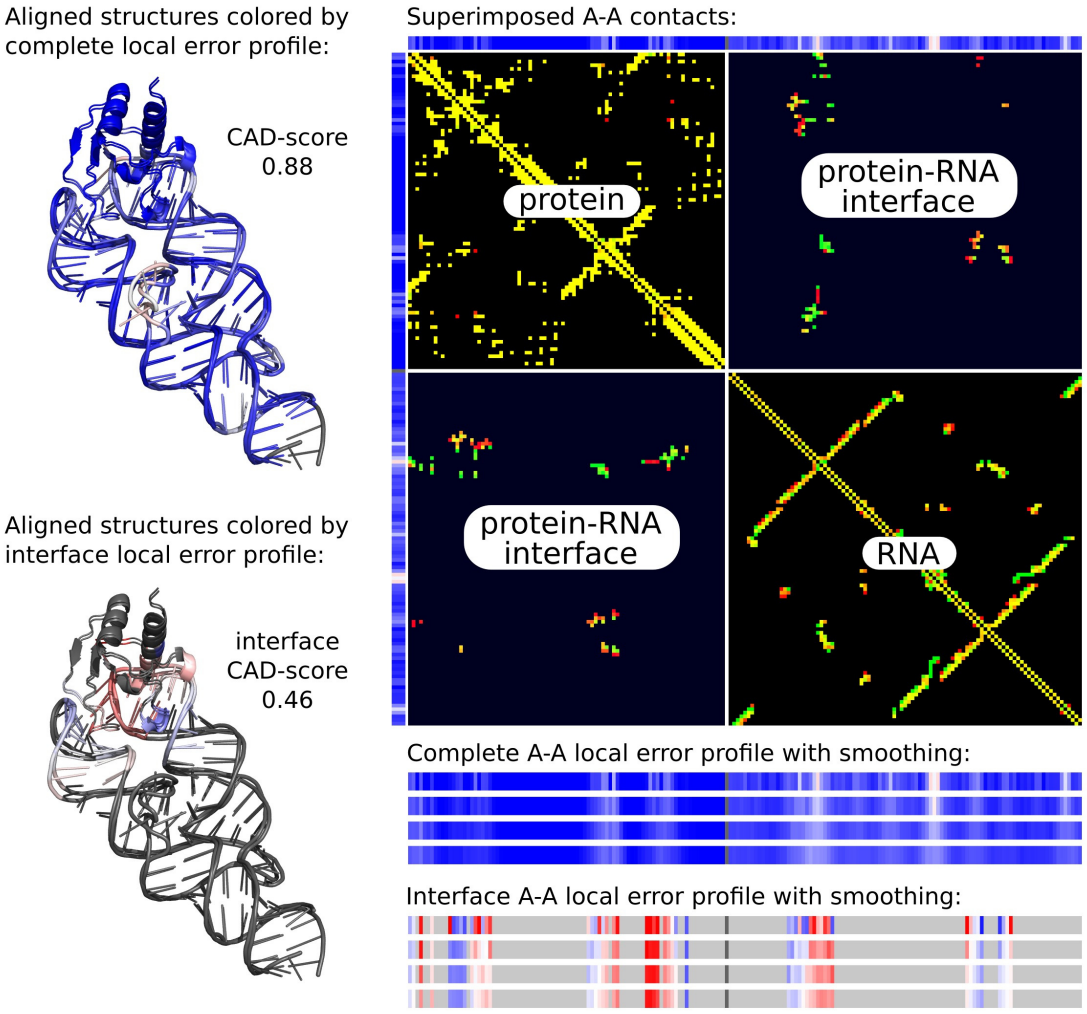


Figure 3.20: Example of the evaluation of a protein-RNA complex model against the reference structure. The model corresponds to the structure 1364 from the decoy set used in the assessment of protein-RNA docking solutions.

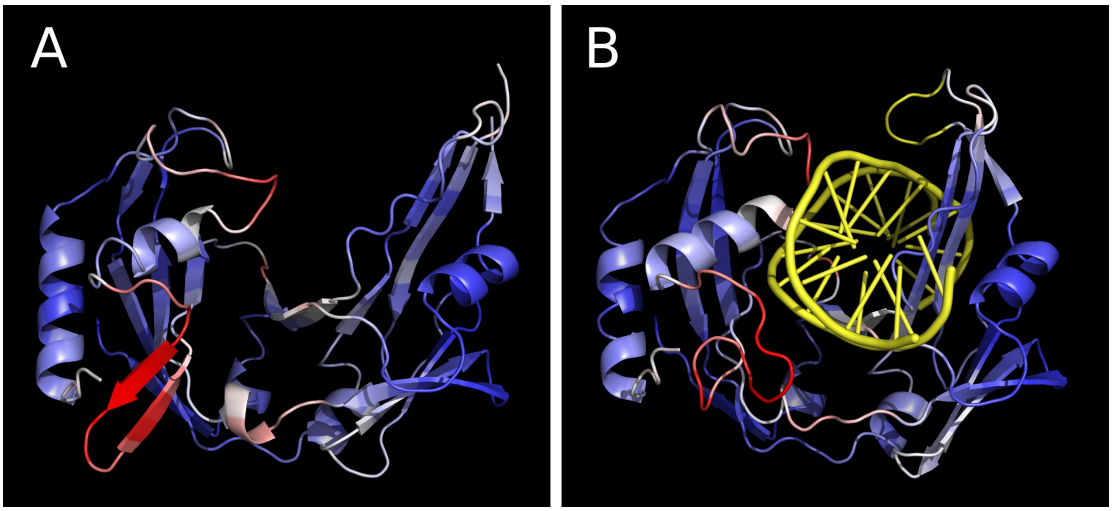


Figure 3.21: Local differences between the structures of the restriction endonuclease BcnI crystallized (A) in the apo form (PDB id: 2ODH) and (B) in complex with the cognitive DNA (PDB id: 2Q10).[109] The DNA and the loop unresolved in the apo form are shown in yellow.

for the interface residues.

3.4.3 Discussion

The web server provides a simple and intuitive interface for the use of the CAD-score method in the interactive manner. In particular, the server features highly interactive visualization options of local contact differences. The server is universal in several ways. It accepts both single-chain and multi-chain structures, works with all the major types of macromolecules (proteins, RNA, DNA and various complexes), allows flexible designation of substructures for the analysis and performs both global and local evaluation of structural differences. Thus, the CAD-score server provides a single framework for addressing a variety of questions related to structural similarity for all the major types of biological macromolecules.

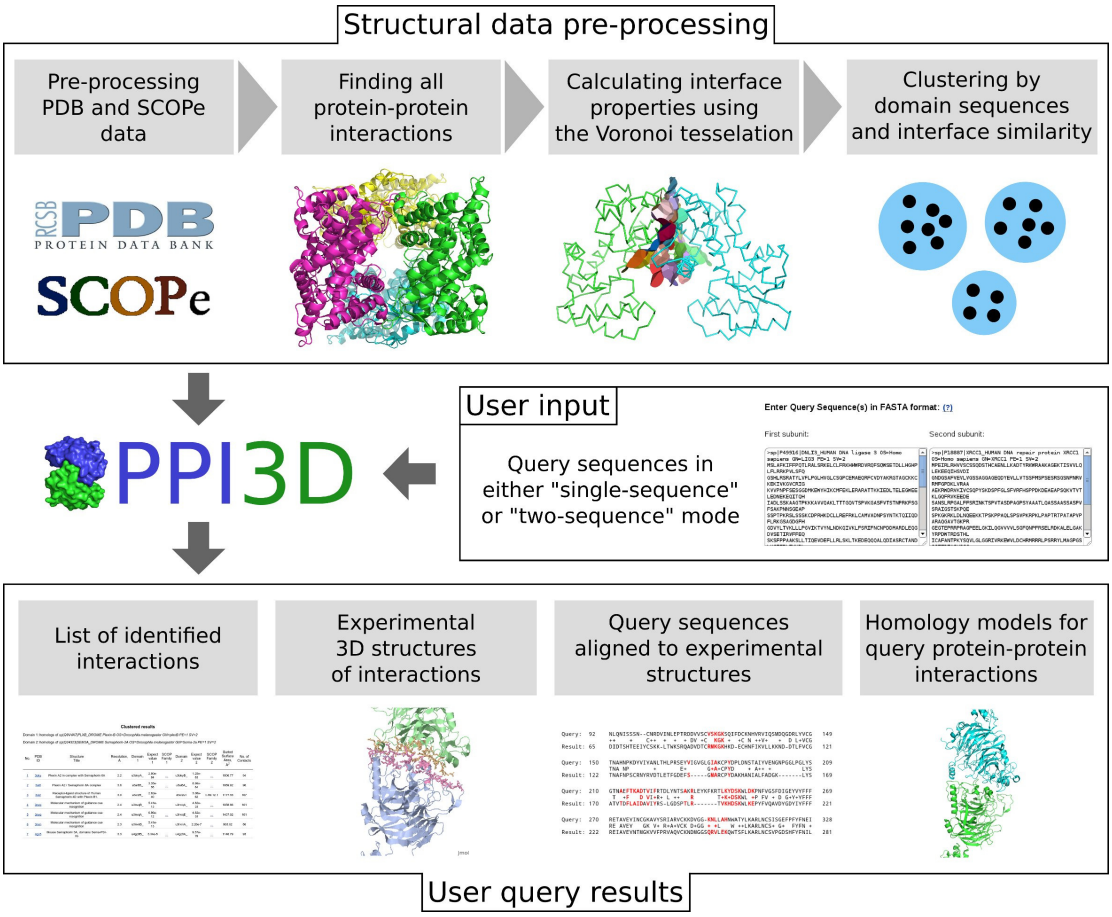


Figure 3.22: Schematic workflow of the PPI3D web server.

3.5 CAD-score application in PPI3D system

The CAD-score method is employed as one of the core components of the PPI3D (Protein-Protein Interactions in 3D) web server.[11] PPI3D is focused on searching and analyzing the structural data on protein-protein interactions. CAD-score enables one of the main features of PPI3D: reducing the data redundancy by clustering and analyzing the properties of interaction interfaces using the Voronoi tessellation. The overall workflow of the PPI3D system is shown in Figure 3.22, CAD-score is utilized in the structural data pre-processing stage of the workflow.

3.5.1 Adapting CAD-score for structure-based clustering of protein-protein interactions

For the clustering of interfaces and binding sites, we use interface-focused CAD-score variations defined in Section 3.1.6. All described CAD-score values are asymmetric and may be different for the same pair of proteins depending on which protein is chosen as reference. Therefore for each pair it is calculated twice by taking the first and the second interaction interface as the reference. Minimal CAD-score value is then selected as the similarity between these interfaces. The same procedure is used for asymmetry correction when comparing protein binding sites.

The original CAD-score definition assumes that the comparison involves structures of proteins having identical sequences. In PPI3D we use CAD-score modification that allows comparison of proteins with different sequences provided the corresponding residues in related structures are known. The residue correspondence is obtained from multiple sequence alignments of clustered protein sequences, generated with MAFFT.[110]

Also, in PPI3D we use constrained Voronoi faces (Figure 3.3) instead of on-sphere contacts. This decision was made due to two reasons. Firstly, for an amino acid residue, the sum of face-based contact areas is less dependent on the number of residue atoms than the sum of on-sphere contact areas. Thus, using areas of constrained Voronoi faces allowed us to make CAD-score less dependent on the differences in amino acid contents when comparing structures of proteins that have non-identical sequences. Secondly, we believe that constrained Voronoi faces provide a more natural and intuitive representation of physical contacts: we report face-based areas of protein-protein interfaces in PPI3D server, therefore, the areas used for CAD-score in PPI3D also need to be face-based.

3.5.2 Clustering of protein interaction interfaces and binding sites

The clustering of protein interaction interfaces and binding sites according to structural similarity is performed using the Taylor-Butina algorithm,[111] which makes later updating of the clusters with new data very straightforward. This algorithm takes as input the similarity matrix and has one tunable parameter: the threshold that defines which elements are similar enough to be joined into one cluster.

The interface similarity criteria for creating the clustered datasets were chosen after careful examination of the data. First, all protein-protein interaction interfaces were clustered by the similarity of protein sequences > 95%. Then CAD-score similarity matrices were calculated for each of resulting clusters. Each matrix was used for further clustering using the Taylor-Butina algorithm and varying the similarity thresholds that define which interaction interfaces may belong to the same cluster. The results are represented in Figure 3.23. It can be seen that if we lower the interface similarity threshold for clustering protein complexes that are nearly identical at the sequence level, the number of clusters stays about the same when the threshold is below 50-60 % (CAD-score 0.5-0.6). It means that the pairs of interaction interfaces usually have either very high or very low similarity values. In other words, the same protein pairs usually interact either using the same or completely different binding sites.

Therefore, to identify protein complexes having nearly identical structure, we used sequence identity > 95% and the similarity of interface contacts > 50% (interface CAD-score > 0.5, calculated using Equation 3.8). This is the default clustering mode in the PPI3D server and it filters out the largest part of the PDB data redundancy such as multiple complexes of the same proteins or their point mutants.

Highly similar protein interaction interfaces of homologous protein complexes (interologs) were clustered together if their sequence similarity

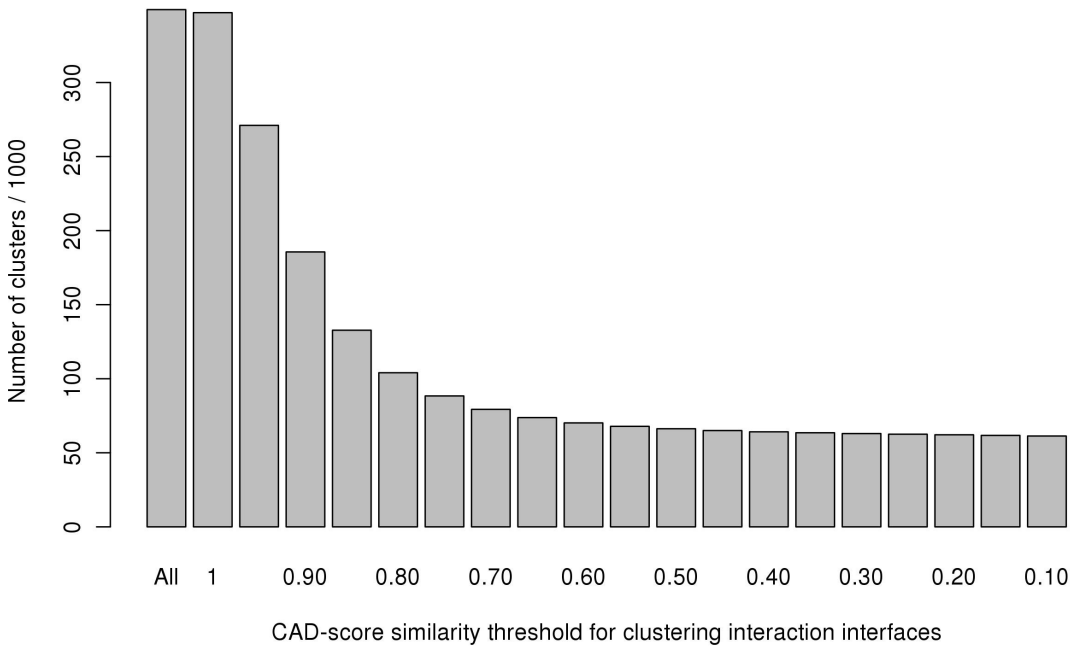


Figure 3.23: Relationship between the interface similarity threshold used for clustering and the number of clusters for protein-protein interfaces having sequences > 95% similar. PDB data released before December 4, 2015 were used for clustering.

was above 40% and the similarity of interface contacts >50% (interface CAD-score > 0.5).

Homologous protein complexes having lower sequence similarity may retain the same binding sites, but may have a significant rearrangement of pairwise residue contacts across the interface. To capture such cases, we used a less stringent similarity measure, the similarity of the interface areas (Equation 3.11).

In case of protein/peptide binding sites, the overall clustering results were the same as for clustering protein-protein interaction interfaces. Consequently, the same thresholds were used for defining the similarity of binding site residue areas (Equation 3.10) and binding site areas (Equation 3.12).

3.5.3 Discussion

The clustering of the protein interactions allows reduction of the data redundancy more than sevenfold for interaction interfaces and more than tenfold for protein-protein binding sites. The summary of results is given in the Table 3.4. In all cases the number of clusters derived by accounting for the structural similarity is higher than the number of just sequence similarity-based clusters. This difference indicates that some related protein complexes have significantly different or even alternative interaction interfaces. Consequently, clustering protein complexes only by sequence similarity would result in a loss of structural data for these proteins.

	Number of clusters
Protein-protein interaction interfaces:	
All protein-protein interaction interfaces	349,208
Clustered at sequence similarity > 95%:	
By sequence similarity only	41,602
By sequence similarity and interface contacts similarity > 50%	66,234
Clustered at sequence similarity > 40%:	
By sequence similarity only	26,951
By sequence similarity and interface contacts similarity > 50%	51,984
By sequence similarity and interface area similarity > 50%	43,845
Protein and peptide binding sites:	
All protein and peptide binding sites	723,546
Clustered at sequence similarity > 95%:	
By sequence similarity only	35,749
By sequence similarity and binding site residue areas similarity > 50%	93,058
Clustered at sequence similarity > 40%:	
By sequence similarity only	21,635
By sequence similarity and binding site residue areas similarity > 50%	68,040
By sequence similarity and binding site area similarity > 50%	58,505

Table 3.4: The summary of results of clustering structural data on protein interactions by sequence and structure similarity in PPI3D (based on the PDB data released before December 4, 2015).

4 VoroMQA: a method for referenceless assessment of protein structure quality using interatomic contact areas

VoroMQA (Voronoi tessellation-based Model Quality Assessment) is an all-atom statistical potential-based method for protein structure quality assessment. The method considers protein structure as a set of balls corresponding to heavy atoms and characterize interactions through interatomic contact areas derived from the Voronoi tessellation of atomic balls.[7] Here, we present a description of the method and compare its performance with both statistical potentials and composite model quality assessment scores.

4.1 Method description

4.1.1 Construction of contacts

Given a protein structure, it can be represented as a set of atomic balls, each ball having a van der Waals radius depending on the atom type. A ball can be assigned a region of space that contains all the points that are closer (or equally close) to that ball than to any other. Such a region is called a Voronoi cell and the partitioning of space into Voronoi cells is called a Voronoi tessellation. Two adjacent Voronoi cells share a set of points that form a surface called a Voronoi face. A Voronoi face can be viewed as a geometric representation of a contact between two atoms. However, if a pair of contacting atoms is near the surface of a protein structure, the corresponding Voronoi face may extend far away from the atoms. Here, this problem is solved by constraining the Voronoi cells

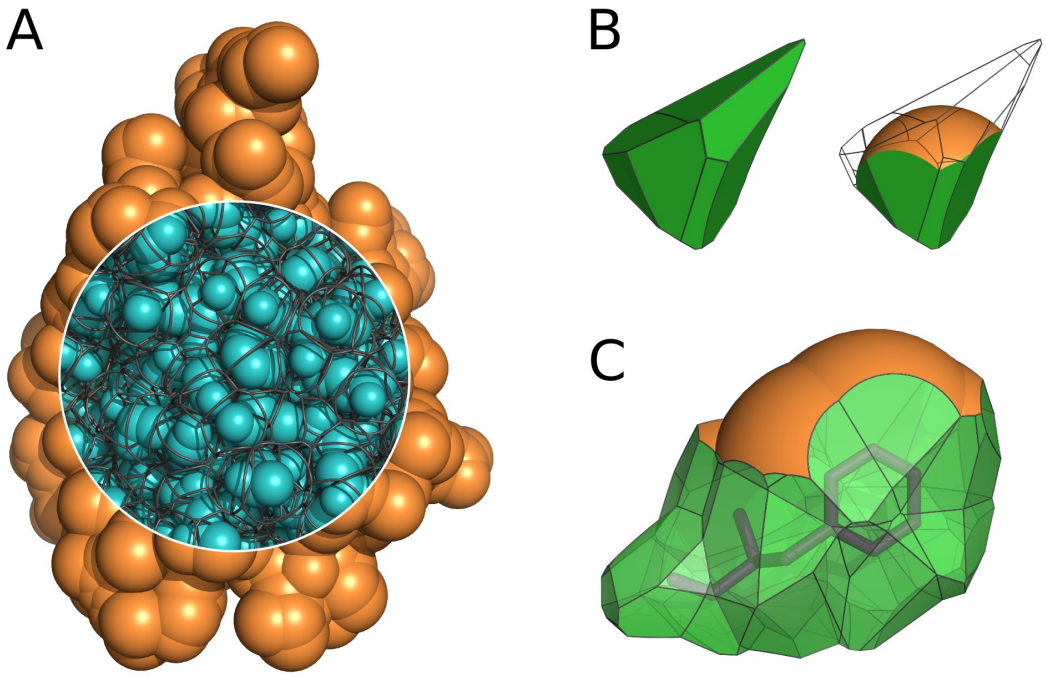


Figure 4.1: (A) Edges of the Voronoi cells constrained inside the solvent accessible surface of a protein structure. (B) Cutting a Voronoi cell with a sphere corresponding to the rolling probe surface results in constrained Voronoi faces and SAS patches. (C) An integral surface of a phenylalanine residue constructed by combining atomic contact surfaces.

of atomic balls inside the boundaries defined by the solvent accessible surface (SAS) of the same balls, as illustrated in Figure 4.1 (A, B). The resulting constrained Voronoi faces and SAS patches can be combined into integral surfaces of larger components of protein structure, e.g. amino acids (Figure 4.1 C). Construction of interatomic contact surfaces is implemented as part of the Voronota software.[7] The construction procedure uses triangulated representations of Voronoi faces and spherical surfaces. Contact areas are calculated as the areas of the corresponding triangulations.

In this study the Voronoi tessellation-based analysis is also used to describe the centrality of contacts. Given a pair of contacting atoms, the contact between them is called *central* if the line segment connecting the

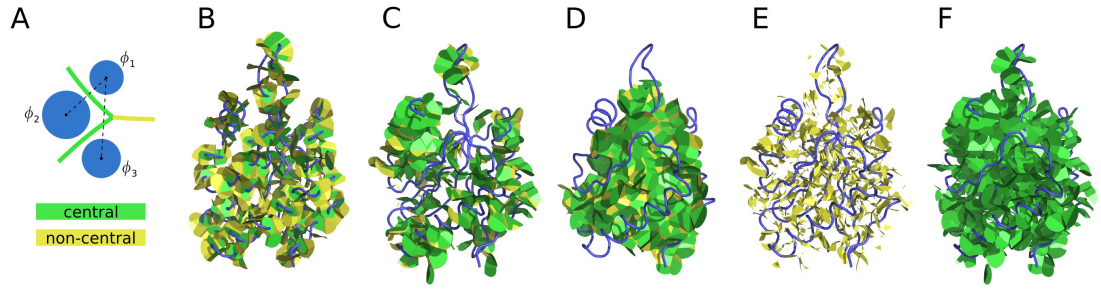


Figure 4.2: (A) 2D illustration of central and non-central contacts: the contact between balls ϕ_1 and ϕ_3 is non-central, the other contacts are central. (B) Central (green) and non-central (yellow) contacts for sequence separation 1. (C) Central and non-central contacts for sequence separation from 2 to 6. (D) Central and non-central contacts for sequence separation greater than 6. (E) Only non-central contacts for sequence separation greater than 1. (F) Only central contacts for sequence separation greater than 1. The PDB ID of the protein structure used in this figure is 1T3Y.

centers of the atoms intersects the corresponding constrained Voronoi face. Otherwise, the contact is called *non-central*. The definition of central and non-central contacts is illustrated in Figure 4.2 (A). Another categorization of contacts used in this work is based on the sequence separation between the residues of the contacting atoms. It is illustrated in Figure 4.2 (B–F) in combination with the centrality-based categorization.

4.1.2 Definition of the quality scoring method

Interatomic and solvent contact areas may be used to evaluate quality of protein structural models by employing the idea of a knowledge-based statistical potential as was first shown by McConkey et al.[76] Our method is aimed to employ the same principle using more elaborate contact descriptions and to be able to produce both local (atom-level) and global (structure-level) scores in a fixed range of values from 0 to 1.

In order to formulate our method, the first step is to define a set of possible contact types. Let $A = \{a_0, a_1, ..., a_n\}$ be a set of atom types and $C = \{c_0, c_1, ..., c_m\}$ be a set of contact categories. A contact type is described by

a tuple $(a_i, a_j, c_k) \in A \times A \times C$, which is equivalent to (a_j, a_i, c_k) because contacts are undirected. The atom type a_0 represents solvent and the contact category c_0 represents solvent-accessible areas, therefore a_0 and c_0 always come together and the set of all possible contact types can be narrowed down to $T = ([A \setminus a_0] \times [A \setminus a_0] \times [C \setminus c_0]) \cup ([A \setminus a_0] \times \{a_0\} \times \{c_0\})$.

A contact type can be assigned a pseudo-energy value $\mathrm{E}(a_i, a_j, c_k)$ calculated from the corresponding expected and observed probabilities:

$$\mathrm{E}(a_i, a_j, c_k) = \log \frac{\mathrm{P}_{\mathrm{exp}}(a_i, a_j, c_k)}{\mathrm{P}_{\mathrm{obs}}(a_i, a_j, c_k)} \tag{4.1}$$

The probability values can be estimated empirically using the contact area values calculated for a learning set of high-quality experimentally determined protein structures. Let $\mathrm{S}(a_i, a_j, c_k)$ be a sum of all the areas of the contacts of type (a_i, a_j, c_k) observed in the learning set. Also, let us define that if $(a_i, a_j, c_k) \notin T$, then $\mathrm{S}(a_i, a_j, c_k) = 0$. Let $\mathrm{S}_{\mathrm{sol}}$ and $\mathrm{S}_{\mathrm{int}}$ be sums of solvent and interatomic contact areas, respectively:

$$\mathrm{S}_{\mathrm{sol}} = \sum_{1 \le i \le n} \mathrm{S}(a_i, a_0, c_0) \tag{4.2}$$

$$\mathrm{S}_{\mathrm{int}} = \sum_{1 \le i \le n} \sum_{1 \le j \le i} \sum_{1 \le k \le m} \mathrm{S}(a_i, a_j, c_k) \tag{4.3}$$

Then the observed probability of the contact type (a_i, a_j, c_k) is defined as the following ratio of areas:

$$\mathrm{P}_{\mathrm{obs}}(a_i, a_j, c_k) = \frac{\mathrm{S}(a_i, a_j, c_k)}{\mathrm{S}_{\mathrm{int}} + \mathrm{S}_{\mathrm{sol}}} \tag{4.4}$$

The corresponding expected probability should represent how often the contacts of the same type would occur in a set of randomly folded structures of the same sequences as in the learning set. It is estimated using the observed probabilities of the isolated components of the contact type

(a_i, a_j, c_k):

$$\mathrm{P}_{\mathrm{exp}}(a_i, a_j, c_k) = \begin{cases} \mathrm{P}_{\mathrm{obs}}(a_i) \cdot \mathrm{P}_{\mathrm{obs}}(c_0) & \text{if } j = 0 \\ \mathrm{P}_{\mathrm{obs}}(a_i) \cdot \mathrm{P}_{\mathrm{obs}}(a_j) \cdot \mathrm{P}_{\mathrm{obs}}(c_k) & \text{if } j \geq 1, i = j \\ \mathrm{P}_{\mathrm{obs}}(a_i) \cdot \mathrm{P}_{\mathrm{obs}}(a_j) \cdot 2 \cdot \mathrm{P}_{\mathrm{obs}}(c_k) & \text{if } j \geq 1, i \neq j \end{cases} \tag{4.5}$$

$$\mathrm{P}_{\mathrm{obs}}(a_i) = \frac{\sum_{0 \leq j \leq n} \sum_{0 \leq k \leq m} \mathrm{S}(a_i, a_j, c_k)}{2\,\mathrm{S}_{\mathrm{int}} + \mathrm{S}_{\mathrm{sol}}} \tag{4.6}$$

$$\mathrm{P}_{\mathrm{obs}}(c_k) = \frac{\sum_{0 \leq i \leq n} \sum_{0 \leq j \leq i} \mathrm{S}(a_i, a_j, c_k)}{\mathrm{S}_{\mathrm{int}} + \mathrm{S}_{\mathrm{sol}}} \tag{4.7}$$

Having the derivation of pseudo-energy values defined using equations (4.1-4.7), let us describe how the derived values are used for scoring protein structures. In order to assign a quality score to a single atom ϕ, a set of related contacts Ω_ϕ is selected. Atom-related contacts are defined as not only the immediate contacts of the considered atom, but also all the contacts of the neighboring atoms. A normalized pseudo-energy value $\mathrm{E}_{\mathrm{n}}(\Omega_\phi)$ is computed using the information known about each contact $\omega \in \Omega_\phi$, namely the contact area (area_ω) and the contact type ($\mathrm{type}_\omega \in T$):

$$\mathrm{E}_{\mathrm{n}}(\Omega_\phi) = \frac{\sum_{\omega \in \Omega_\phi} \mathrm{E}(\mathrm{type}_\omega) \cdot \mathrm{area}_\omega}{\sum_{\omega \in \Omega_\phi} \mathrm{area}_\omega} \tag{4.8}$$

An atom quality score $\mathrm{Q}_{\mathrm{a}}(\Omega_\phi) \in [0, 1]$ is defined using the Gauss error function:

$$\mathrm{Q}_{\mathrm{a}}(\Omega_\phi) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\mathrm{E}_{\mathrm{n}}(\Omega_\phi) - \mu_{\mathrm{type}_\phi}}{\sigma_{\mathrm{type}_\phi}\sqrt{2}}\right)\right) \tag{4.9}$$

The values of μ (mean) and σ (standard deviation) are estimated for each atom type from the normalized pseudo-energy values calculated for the atoms in the learning set of protein structures.

Given a set Φ of all atoms in a protein structure, a global structure quality score $\mathrm{Q_g}(\Phi)$ is defined as a weighted arithmetic mean of the atoms quality scores:

$$\mathrm{Q_g}(\Phi) = \frac{\sum_{\phi \in \Phi} \mathrm{Q_a}(\Omega_\phi) \cdot \mathrm{weight}_\phi}{\sum_{\phi \in \Phi} \mathrm{weight}_\phi} \tag{4.10}$$

The weights here indicate how deep each atom is buried inside a structure: solvent-accessible atoms have weight 1, their direct contacting neighbors have weight 2, the neighbors of the direct neighbors have weight 3, and so on.

The quality score of a residue is defined as an average of quality scores of its atoms. A sliding window with four residues on both sides is used to smooth residue scores along the sequence. Let us denote an unsmoothed residue score at position n as $\mathrm{Q_r}(n)$, then the corresponding smoothed value $\mathrm{W_r}(n)$ is computed as a normalized weighted sum of the scores of the neighboring residues:

$$\mathrm{W_r}(n) = \frac{\sum_{-5<m<5} \mathrm{Q_r}(n+m) \cdot (5 - |m|)}{\sum_{-5<m<5} (5 - |m|)} \tag{4.11}$$

4.1.3 Implementation of the quality scoring method

Implementation of the method requires a protein structure dataset (learning dataset) for collecting data on interatomic contacts, the set of atom types and the set of contact categories. Protein structures for the learning set were obtained from the Protein Data Bank[2] (`www.rcsb.org`). Only protein structures solved by X-ray at better than 2.5Å resolution were considered. The set was limited to monomeric or oligomeric (up to 12 subunits) proteins with each chain longer than 99 residues. Proteins solved in complex with nucleic acids, membrane proteins, proteins with modified polymeric residues were excluded. From the remaining structures only representatives at 50% sequence identity were retained. For

each of the resulting PDB entries (totaling 12825 as of 2015.06.11), the structure of the first biological assembly was used for deriving contact areas. Only non-bonded contacts between atoms of different residues were considered.

In the case of multi-chain biological assemblies the set of derived contacts is redundant. To remove this redundancy, the contact areas are multiplied by the ratio $\frac{N_u}{N}$, where N is the total number of chains and N_u is the number of unique protein chains.

Twenty standard amino acids have 167 different heavy atom names, however, seven atom pairs are interchangeable because of the molecular symmetry: Arg NH1 and NH2, Asp OD1 and OD2, Glu OE1 and OE2, Phe CD1 and CD2, Phe CE1 and CE2, Tyr CD1 and CD2, Tyr CE1 and CE2. Therefore, the final set contains 160 distinct atom types, plus one special type representing solvent.

As for the set of contact categories, a hybrid scheme is used: solvent contacts are treated separately; each non-solvent contact is categorized as either near or far depending on the sequence separation between the residues of the contacting atoms; each non-solvent contact is categorized as either central or non-central as illustrated in Figure 4.2 (A). This results in 5 distinct categories: "solvent", "near and central", "near and non-central", "far and central", "far and non-central". During the method learning stage, when the empirical probabilities are computed, a contact is considered far if the corresponding sequence separation is greater than 6: this is done to separate the contacts that may be largely induced by the close sequence proximity of the contacting residues from the contacts that are more likely to occur because they are favorable. During the method application stage, when calculating normalized pseudo-energies of atoms using equation (4.8), only far or solvent contacts are considered, but the sequence separation threshold for contacts considered as far is lowered so that only contacts between the atoms of residues adjacent in sequence are categorized as near. This allows to take into account the vast

Category	$\mathbf{P_{obs}^{high}}$	$\mathbf{P_{obs}^{low}}$
near and central	0.159	0.147
near and non-central	0.168	0.165
far and central	0.225	0.166
far and non-central	0.056	0.052
solvent	0.392	0.470

Table 4.1: Observed probabilities of the contact categories estimated for the learning set of high quality structures (P_{obs}^{high}) and the set of lower quality structures comprised of CASP models (P_{obs}^{low}).

majority of contacts while excluding the ones that are likely to appear in a structural model regardless of its correctness.

When estimating the probabilities of the contact categories using equation (4.7), we tried two datasets for input: the learning set of high quality structures and a set of lower quality structures that was comprised of the models of the monomeric targets from CASP8,[112] CASP9[95] and CASP10.[17] Table 4.1 contains the two resulting sets of probability values, the most prominent difference between them being the solvent contact probabilities, meaning that the lower quality structures are not as well packed as the high quality ones. We reasoned that random protein-like structures should also be packed worse than the native protein structures, therefore for equation (4.5) we employed the probabilities of the contact categories that were estimated from the set of lower quality structures.

The last required information is the mean and standard deviation values used in equation (4.9). These values were calculated for each atom type after applying equation (4.8) to every atom in the learning set of protein structures.

The VoroMQA software is available both as a standalone application and as a web-server at `bioinformatics.lt/software/voromqa`. Our standalone software does not require any third-party programs or libraries to work. However, in some cases it may be beneficial to employ

an external tool to rebuild the side chains in input protein structures before evaluation as this may reduce chances of overly penalizing structural models that have good backbone but poor side-chain packing.

4.1.4 Expected scores for native protein structures

The implemented method was used to compute the global quality scores of the protein structures in the original learning set to estimate what scores to expect from realistic structural models. Each structure was evaluated twice: first time, using the default method configuration, and second time, after rerunning the learning stage with the structure of interest removed from the learning set. The mean difference between the first and the second global scores was less than 0.00028, for 99% of the structures the difference was less than 0.0006, the maximum observed difference was 0.0038. This allows us to conclude that the performance of the method is largely insensitive to the presence or absence of any single structure in the learning set.

The summary of global quality scores calculated by the second procedure is presented in Figure 4.3. Plot (A) shows the empirical distribution of global scores leading to the following observations: 1) it is unlikely for a realistic protein structure to have a global score lower than 0.3 or greater than 0.7, and 2) a global quality score is not heavily dependent on the prevailing type of secondary structure. Plot (B) shows that, on average, smaller protein structures receive slightly lower global quality scores than larger structures, and the variance is greater for smaller structures. Another aspect of the method is that the scoring time scales linearly with the structure size, as illustrated in Figure 4.3 (C).

4.1.5 A note on the older version of the method

The initial simplified variant of the VoroMQA method was tested in the QA category of the CASP11 experiment.[64] Compared to the current ver-

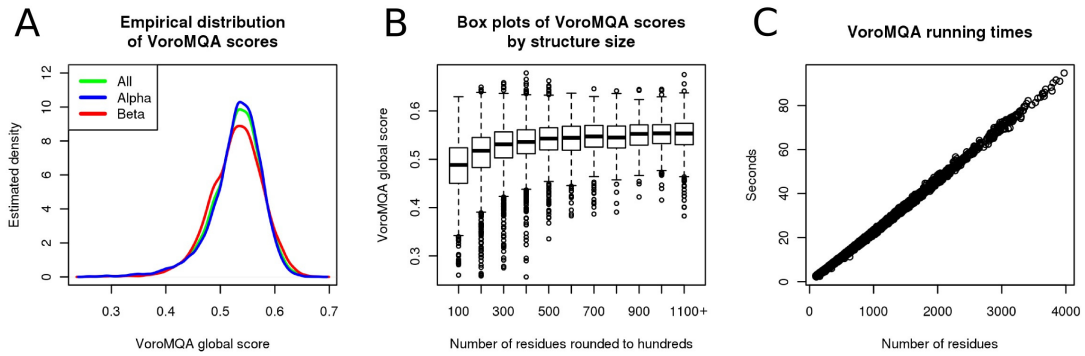


Figure 4.3: Recap of the global quality scores calculations performed for the protein structures in the learning set. (A) Estimated empirical density functions of the scores for all the structures (green), the structures with prevailing alpha helices (blue) and the structures with prevailing beta sheets (red). (B) Box plots of global scores for different thresholds of structure sizes (the rightmost box plot also covers all the structures from the learning set that have more than 1100 residues). (C) Software running times plotted against the corresponding structure sizes (the test was performed using CPU Intel® Xeon® E5-2670 v3 @ 2.30GHz).

sion, the older one did not utilize contact categories, did not distinguish between different atom types when converting from pseudo-energy values to atomic quality scores and did not assign weights to the atomic scores when calculating global quality scores. Also, only single-chain protein structures from the PISCES[113] database were used in the learning stage of the older version. To assess the effect of the differences between the older and the newer versions of VoroMQA, we recorded the results achieved by both versions in the tests described later in this paper. When describing the test results, the older version of VoroMQA is denoted as "VoroMQA-old" and the current version is denoted as "VoroMQA-new" or simply "VoroMQA".

4.2 Testing arrangements

4.2.1 Datasets used to assess the performance of the method

In order to analyze the ability of VoroMQA to select a native structure from a set of its models of varying quality, we downloaded the target and model structures from the last four CASP experiments (CASP8-11). We did not consider targets that correspond to individual subunits of obligatory protein complexes representing biologically unrealistic oligomeric state or those solved with poor resolution. Consequently, we used 140 CASP targets that conform to the following criteria: a target must correspond to a PDB entry that has at least one single-chain biological assembly and the experimental method used to determine the structure must be X-ray crystallography with resolution better than 2.5 Å. For every target, all the available complete models were downloaded, excessive regions in models were trimmed to exactly match the target structure. Scores for target and model structures were then computed using the old and the new variations of VoroMQA as well as DOOP,[72] GOAP,[75] and dDFIRE.[73]

In order to analyze the ability of VoroMQA to evaluate protein structural models, we used the CASP11 Quality Assessment (QA) data. We considered all the 88 targets that were used in the official assessment of CASP11 QA results,[64] but, after manual inspection, excluded four of them (T0775, T0787, T0799 and T0813) because their structures are single chains pulled out from obligatory oligomers. For the remaining 84 targets, the following data was downloaded from the CASP11 website: the structures of server models (best150 and sel20 sets[63]); the corresponding reference-based quality scores (GDT-TS,[13,103] LDDT,[114] SphGr (SphereGrinder)[115] and CAD_AA[6]); the available global scores calculated by the single-model quality assessment methods (MULTICOM-CLUSTER,[116] MULTICOM-NOVEL,[117] ProQ2,[65]

ProQ2-refine,[118] Wang_SVM,[119] Wang_deep_{1,2,3}[119] and the old version of VoroMQA). The latest version of our method was applied to produce two more scores for each model: VoroMQA-new (calculated for the unmodified input structure) and VoroMQA-new-sr (calculated after rebuilding the side-chains of the input structure using SCWRL4[120]). Additionally, we computed the following statistical potential-based scores: DOOP, GOAP, and dDFIRE. We compiled the retrieved and the computed scores together and removed duplicate model entries and entries with at least one score absent. The final combined tables of scores characterize 11627 models from the best150 sets and 1583 models from the sel20 sets (the tables are available for download from the VoroMQA web page).

The best150 and sel20 sets, composed at the time by CASP11 organizers, differ in both their size and nature. For each target, the best150 set contains the best 150 models selected using a consensus-based QA algorithm, while the sel20 set contains 20 diverse models selected based on a clustering of all the available models for the target. We reasoned that it may also be interesting to perform tests on sets that are small (like sel20) but contain better models (like best150). Therefore, in addition to the best150 and sel20 sets for each target, we used sets of models produced by the three well-performing prediction servers: BAKER-ROSETTASERVER,[121] Zhang-Server,[122] and QUARK.[123] We dubbed these sets "BZQ15" because only up to 15 models of the three servers are available for each CASP11 target. Each BZQ15 set simulates the real-life scenario, when a researcher needs to choose the best model from a few generated by several well-known servers.

4.2.2 Reference-based scores used to assess the model selection capabilities

For assessing the VoroMQA performance when selecting the best model from a set of models of the same target, we chose to employ the same four reference-based scores (GDT-TS, LDDT, SphGr and CAD_AA) used in the official CASP11 QA assessment. However, each of the four scores focuses on somewhat different structural properties and they often disagree in deciding which model is closer to the native structure. Moreover, each score has a degree of uncertainty so that the score difference for close models may not always be significant.[124] To take care of these issues, we additionally introduced a simple tournament-based methodology described below.

Let us take two models a and b of the same target. Let us say that a "wins" against b if all the four reference-based scores are higher for a than for b. If there is a disagreement, for example, if $\mathrm{GDT-TS}_a > \mathrm{GDT-TS}_b$ but $\mathrm{LDDT}_a < \mathrm{LDDT}_b$, then the outcome of the duel between a and b is a draw. Using the defined rules, all the possible duels are executed for the models in the input data set. For each model the numbers of wins, draws and losses are recorded. The results of the performed "tournament" are used as a basis for our ensuing analysis.

A straightforward way to utilize the tournament results is to assess how well a QA method is able to select the best model out of two. Let us consider a set of models M, let N be the total number of non-draw duels among the elements of M and N_p be the number of non-draw duels that the QA method correctly predicts the winner for. Then the QA method performance can be quantified using the agreement percentage score:

$$\text{Agreement-score}(M) = N_p/N \cdot 100\% \tag{4.12}$$

The next step is to assess the ability of a QA method to select the best

model out of many. In our tournament-based framework we define the true best model of a target as the model with the highest number of won duels. If two models have the same number of wins, the one with more draws, i.e. less losses, is considered better. We combine the numbers of wins and the numbers of draws into a single score called Wins-score. Let us consider a target t that has N_t models and its best model b has w_b wins with d_b draws. Given a model m of t that has w_m wins with d_m draws, let us define Wins-score score for m. To ensure that even a single win has more weight than $N_t - 1$ draws, numbers of wins are multiplied by N_t in the following formula:

$$\text{Wins-score}(m) = \frac{w_m \cdot N_t + d_m}{w_b \cdot N_t + d_b} \tag{4.13}$$

$\text{Wins-score}(m)$ can range from 0 (when m has no wins and draws) to 1 (when m is the best model). The score can be interpreted as a measure of success achieved by model m compared to the remaining $(N_t - 1)$ models of t. We use the Win-scores of the models selected by a QA method to quantify the ability of the method to select the best possible models.

Summarizing Agreement-scores (or Wins-scores of selections) for multiple different targets can be done by calculating their mean value. However, when comparing the performances of two different QA methods, a simple comparison of the corresponding mean values is not sufficient as it lacks the information about the significance of the difference. We use the Wilcoxon signed-rank test[125] to assess whether two sets of per-target scores come from two populations with different means. We chose this particular test and not the paired Student's t-test because we cannot assume that a population of Agreement-scores (or a population of selection Wins-scores) is distributed normally. We first run the two-sided Wilcoxon test: if the computed p-value is sufficiently small, i.e. the two population means differ significantly, then the one-sided version of the test is used to check if the first population mean is likely larger.

4.3 Testing results

4.3.1 Overview of testing procedures

We tested the performance of VoroMQA in several ways which are outlined in this section and presented in detail in the subsequent sections.

Firstly, we focused on global VoroMQA scores and assessed if the method is able to distinguish a native structure from its decoys. We used data from several CASP experiments to form sets of decoys: such sets are comprised of models of various quality generated by a variety of structure prediction methods. We compared the performance of VoroMQA with the performance of three other methods that are based solely on analyzing geometric features and applying knowledge-based statistical potentials, namely DOOP,[72] GOAP,[75] and dDFIRE.[73]

Next, we analyzed how VoroMQA global scores computed for models relate to the observed differences between models and the native structures (targets). To this end, we used CASP11 structural models and the corresponding reference-based quality scores. As the official assessment of CASP11 QA results[64] was done using primarily GDT-TS,[13,103] LDDT,[114] SphGr (SphereGrinder)[115] and CAD_AA (CAD-score$_{\mathrm{AA}}$),[6] the same four scores were also employed in our study.

During another test we analyzed the ability of VoroMQA to select the best model out of several or many. For this test we applied the four reference-based scores and the newly introduced tournament-based methodology (see "Testing arrangements"), which allows multiple reference-based scores to be considered simultaneously. In addition to DOOP, GOAP and dDFIRE, we compared VoroMQA with single-model quality assessment methods that participated in CASP11, namely MULTICOM-CLUSTER,[116] MULTICOM-NOVEL,[117] ProQ2,[65] ProQ2-refine,[118] Wang_SVM[119] and Wang_deep_{1,2,3}.[119] Unlike VoroMQA, these methods employ additional data such as secondary structure and

solvent accessibility predictions, in effect incorporating evolutionary information derived from homologous sequences. Therefore, matching or surpassing their performance may be considered a serious challenge for VoroMQA.

The last testing procedure was dedicated to the local scoring. We used data from the CAMEO project (www.cameo3d.org)[19] to investigate some properties of VoroMQA local scores in relation with reference-based local scores.

4.3.2 Selecting native structures from sets of decoys

We tested VoroMQA alongside DOOP, GOAP and dDFIRE according to the ability to distinguish a native structure amidst a variety of its models (decoys) using the data corresponding to the 140 monomeric targets from CASP8-11 experiments (see "Testing arrangements" for details).

The performance of each structure evaluation method was assessed by counting how many times a native (target) structure was missed. Also, differences between the target scores and the corresponding model mean scores were computed and converted to z-scores. According to the number of missed native structures VoroMQA performed on par with DOOP and surpassed the others. The summary of the results is presented in Table 4.2, the per-target results are shown in Supplementary Table S1.[10]

4.3.3 Relationship between VoroMQA global scores and model quality

As we have shown above (Figure 4.3), VoroMQA global scores do not significantly depend on either prevalent secondary structure content or protein size. Thus in principle, it should be possible to decide if a computational model is close to the native structure solely on the basis of the VoroMQA global score. Figure 4.3 (A) shows that a vast majority

Method	Missed targets	Mean z-score
VoroMQA-new	8	3.19
DOOP	8	3.00
GOAP	16	2.87
VoroMQA-old	27	2.67
dDFIRE	46	2.09

Table 4.2: Results of the target selection ability analysis performed for the set of 140 monomeric targets from CASP8, CASP9, CASP10 and CASP11. The "Missed targets" column values show how many times each QA method failed to distinguish a target structure among its models. The "Mean z-score" column values show the average z-scores of the QA scores of the target structures.

of high quality experimentally determined structures have VoroMQA scores greater than 0.4. Also, almost none of the native structures have VoroMQA scores less than 0.3. Following these observations, we computed empirical distribution densities of the four reference-based quality scores (GDT-TS, LDDT, SphGr and CAD_AA) of the CASP11 models that have VoroMQA-new scores in the intervals $(0, 0.3)$, $[0.3, 0.4]$ and $(0.4, 1)$. The results, shown in Figure 4.4, allow us to formulate the following simple rule for interpreting a VoroMQA-new value v of a protein structural model: if $v < 0.3$, then the model is likely bad; if $v > 0.4$, then the model is likely good; if $v \in [0.3, 0.4]$, then the model quality cannot be reliably classified as bad or good using VoroMQA alone. This rule is most useful when just a single model is available. Results for CASP11 models also showed that VoroMQA-new and VoroMQA-new-sr global scores are highly correlated (Pearson correlation coefficient is about 0.98). Therefore, the same rule can be applied for both scores.

4.3.4 Results of the per target analysis of CASP11 data

In model selection tests, we first analyzed how different QA scores perform on best150 sets using the tournament-based methodology. For each available QA method, we calculated agreement percentage scores for all the targets combined and for every target separately. When considering

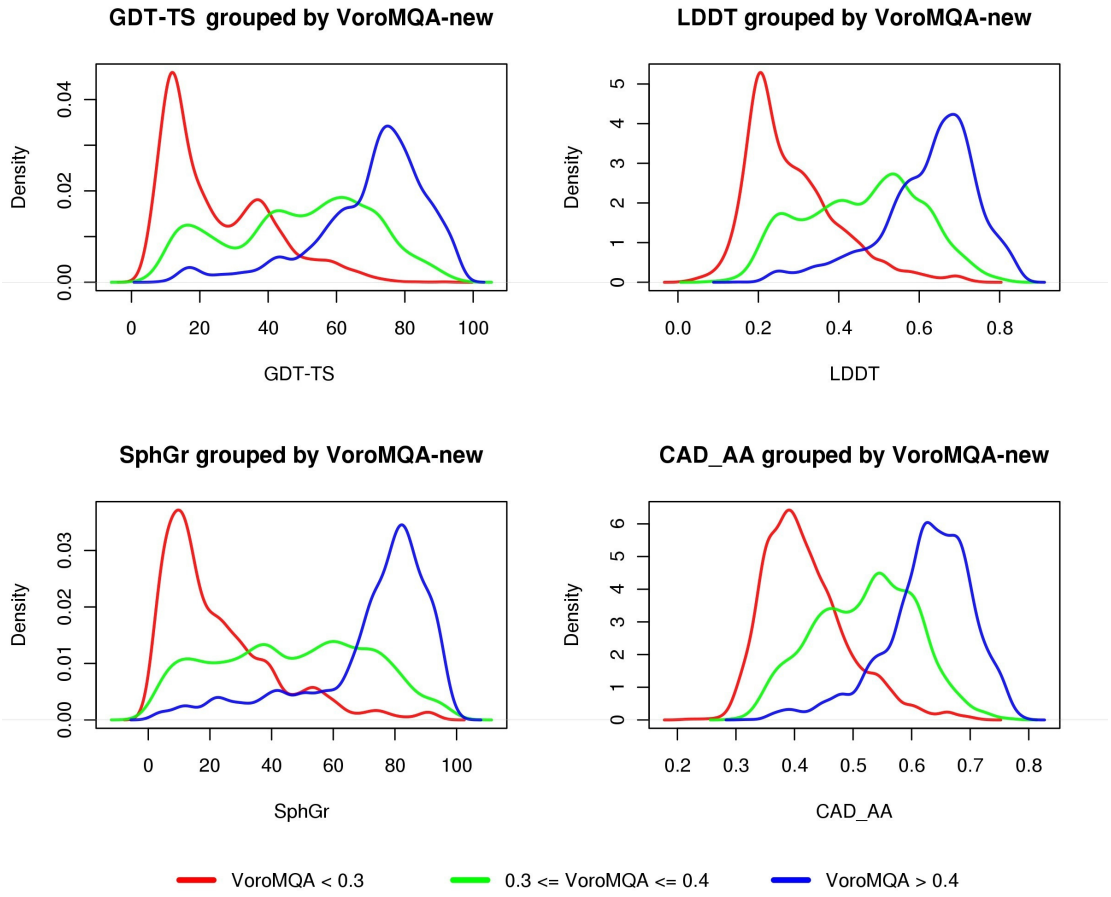


Figure 4.4: Empirical distribution densities of GDT-TS, LDDT, SphGr and CAD_AA scores of the CASP11 models that have VoroMQA-new scores in intervals $(0, 0.3)$, $[0.3, 0.4]$ and $(0.4, 1)$: the corresponding lines are colored in red, green and blue, respectively.

Method	Total %	Mean %
VoroMQA-new-sr	82.50	82.70
VoroMQA-new	81.80	82.16
GOAP	80.11	80.57
VoroMQA-old	79.70	80.48
MULTICOM-NOVEL	79.66	80.25
ProQ2-refine	78.69	79.40
MULTICOM-CLUSTER	78.76	79.21
ProQ2	78.13	78.86
dDFIRE	77.73	78.43
DOOP	76.09	76.57
Wang_SVM	74.42	75.24
Wang_deep_2	72.12	72.83
Wang_deep_3	71.65	72.30
Wang_deep_1	71.57	72.19

Table 4.3: Agreement percentage scores calculated for best150 sets of models from CASP11: second column contains scores for all the targets combined, third column contains mean per-target scores. The table is sorted by the third column.

all the possible 799703 pairs of models, only for 425877 (53%) of them all four reference-based scores (GDT-TS, LDDT, SphGr and CAD_AA) agree which model out of the two is better. The middle column in Table 4.3 shows how often different QA scores agree with the unanimous judgment of all four reference-based scores, the last column shows the average per-target agreement percentages, i.e. mean Agreement-scores. Table 4.4 shows the p-values calculated by applying the Wilcoxon signed-rank test to compare VoroMQA-new-sr with the other methods according to per-target Agreement-scores. Considering the significance level threshold of 0.05, VoroMQA-new-sr significantly outperformed all the others, except for VoroMQA-new (results of the analogous test for VoroMQA-new are presented in Supplementary Table S2[10]).

Next, we asked each available QA method to select a single model from the best150 set of every target. The mean Wins-score, GDT-TS, LDDT, SphGr and CAD_AA values of the selected models are presented in Ta-

Method	p-value (two-sided)	p-value (one-sided)
VoroMQA-new	0.20292	0.10146
VoroMQA-old	0.01182	0.00591
MULTICOM-NOVEL	0.00558	0.00279
ProQ2-refine	0.00041	0.00020
GOAP	0.00024	0.00012
ProQ2	0.00016	0.00008
MULTICOM-CLUSTER	0.00004	0.00002
Wang_SVM	0.00000	0.00000
Wang_deep_3	0.00000	0.00000
Wang_deep_2	0.00000	0.00000
Wang_deep_1	0.00000	0.00000
DOOP	0.00000	0.00000
dDFIRE	0.00000	0.00000

Table 4.4: Results of the Wilcoxon signed-rank test applied to compare the agreement percentage scores achieved by the VoroMQA-new-sr method for best150 sets of models from CASP11 with the corresponding Agreement-scores achieved by the other methods. The table is sorted by the middle column. All the p-values are rounded up to the five decimal places. Gray background is used to indicate methods that performed significantly worse than VoroMQA-new-sr.

ble 4.5 along with the corresponding mean z-scores. Wins-scores and z-scores were calculated considering only the models from best150 sets. While Table 4.5 shows that VoroMQA-new-sr and VoroMQA-new performed relatively well, the achieved advantage over other methods is mostly not significant. Supplementary Table S3[10] shows the p-values calculated by applying the Wilcoxon signed-rank test to compare the Wins-scores of the models selected by the VoroMQA-new-sr method with the corresponding results achieved by the other methods (results of the analogous test for VoroMQA-new are presented in Supplementary Table S4[10]). The p-values greater than 0.05 indicate that all the scores from VoroMQA, ProQ2 and MULTICOM families, as well as GOAP and DOOP scores, demonstrate very similar model-selection abilities when analyzing models from best150 sets using our tournament-based methodology.

Additionally, we performed the analysis based on Agreement-scores and Wins-scores on BZQ15 sets (Supplementary Tables S5–S10[10]). The overall trends are similar to those of best150 sets but there are some differences. Most notably, VoroMQA-new-sr performed significantly better than VoroMQA-new indicating that side-chain rebuilding was particularly beneficial for scoring models from BZQ15 sets. This is consistent with the fact that BAKER-ROSETTASERVER differs considerably from Zhang-Server and QUARK in the side-chain positioning quality.[126] Therefore, rebuilding side-chains before scoring apparently helps to level significant differences in side-chain packing leading to improved results. Also, in the test based on Agreement-score, VoroMQA-new-sr did not significantly outperform MULTICOM-NOVEL.

Finally, we analyzed sel20 sets, and the detailed results are presented in Supplementary Tables S11–S16.[10] Both VoroMQA-new and VoroMQA-new-sr performed significantly worse than ProQ2-refine and MULTICOM-NOVEL in Agreement-score-based testing and significantly worse than ProQ2-refine and ProQ2 in Wins-score-based testing. Over-

Method	Wins-score	GDT-TS	LDDT	SphGr	CAD_AA
Wins-score	1/2.66	55/2.11	0.544/2.19	58.3/2.26	0.586/2.23
LDDT	0.939/2.42	54.3/1.93	0.549/2.33	57.8/2.14	0.584/2.13
CAD_AA	0.902/2.29	53.5/1.72	0.536/1.99	57.3/2.08	0.594/2.46
SphGr	0.861/2.1	54.1/1.85	0.522/1.7	59.7/2.55	0.58/2.04
GDT-TS	0.857/2.1	56.2/2.4	0.522/1.7	56.7/1.97	0.575/1.85
VoroMQA-new-sr	0.735 / 1.63	50 / 1.17	0.497/ 1.28	53 / 1.54	0.566/ 1.67
MULTICOM-CLUSTER	0.717/1.58	48.9/0.94	0.495/ 1.28	51/1.27	0.563/1.57
VoroMQA-new	0.713/1.56	50 /1.09	0.496/1.26	51.8/1.32	0.56/1.48
VoroMQA-old	0.704/1.52	48.9/1.03	0.492/1.25	51.1/1.34	0.559/1.51
ProQ2-refine	0.703/1.51	49.3/1	0.495/1.26	52.1/1.39	0.562/1.56
MULTICOM-NOVEL	0.695/1.49	49.4/1.07	0.49/1.19	51.8/1.38	0.564/1.64
GOAP	0.694/1.49	49.8/0.89	0.501 /1.27	52.2/1.22	0.568 /1.66
ProQ2	0.691/1.46	49.9/0.98	0.496/1.21	52.2/1.34	0.561/1.51
DOOP	0.681/1.46	48.8/0.81	0.499/1.27	50.7/1.05	0.564/1.55
Wang_SVM	0.616/1.2	47.5/0.73	0.474/0.83	49.4/1.02	0.546/1.11
Wang_deep_2	0.568/0.99	47.2/0.65	0.471/0.72	49.8/0.96	0.545/1.05
Wang_deep_1	0.546/0.91	46.3/0.49	0.464/0.6	48.8/0.87	0.542/0.95
dDFIRE	0.542/0.97	46.1/0.33	0.471/0.75	48.4/0.78	0.553/1.28
Wang_deep_3	0.519/0.8	46.6/0.46	0.463/0.53	48.7/0.78	0.542/0.93

Table 4.5: Mean per-target scores of the models selected by various QA methods from best150 sets of models from CASP11. Each numeric cell contains two slash-separated values: the mean reference-based score of the selected models and the corresponding mean z-score. The top five rows show the results obtained using reference-based scores, i.e. results that are close to ideal. The table is sorted by the Wins-score values. Gray background is used to indicate the greatest values in each numeric column.

all, for sel20 sets, every method that is based solely on analyzing geometric features and applying statistical potentials (GOAP, DOOP, dDFIRE and all the VoroMQA variations) achieved worse results than the best-performing composite methods incorporating evolutionary information in the form of predicted features such as secondary structure or solvent accessibility: this was definitely not the case for best150 and BZQ15 sets. We thus asked if additional information can be decidedly beneficial when evaluating sel20 models. Considering that every sel20 set was formed to contain models as different from each other as possible, there may be cases when incorrect models can be identified simply by being significantly different from a reasonably reliable model produced by some homology-based structure prediction server. To test this surmise we defined a simple QA method, dubbed "HHpred-agreement", that evaluates models by comparing them with a model produced by the HHpred server[127] (HHpredA in CASP11) using TM-score.[46] Higher TM-scores were considered to represent better models. Supplementary Tables S17–S19[10] show how HHpred-agreement performed in selecting models from sel20, best150 and BZQ15 sets: HHpred-agreement performed very similarly as ProQ2-refine and ProQ2 for sel20, but much worse than all the other tested QA scores for best150 and BZQ15. We also defined a meta-score, named "VoroMQA-new-and-HHpred-agreement", which is simply an unweighted geometric mean of VoroMQA-new and HHpred-agreement scores. An analogous meta-score was also defined for VoroMQA-new-sr. As shown in Supplementary Tables S17–S19,[10] the two meta-scores achieved top spots in the Wins-score-based ranking of QA methods for sel20 sets and performed relatively well (although not as well as the original VoroMQA-new-sr) for best150 and BZQ15 sets. To sum up, for sets of models similar to sel20, using just VoroMQA may not be as effective as using it in conjunction with additional information derived using sequence homologs. Results of our analysis also raise concerns about whether sel20 sets in CASP11 represent real-life model selection challenges, because the relatively good performance of the HHpred-

agreement score suggests that it may be more efficient just to get a single model from HHpred (or some other well-performing homology-based server) instead of selecting a model from a small but very diverse (sel20-like) set.

In our analysis we concentrated on the ability of considered QA methods to identify best models and so far neglected the correlation analysis, i.e. the calculation of coefficients of correlation between QA scores and reference-based model evaluation scores. Correlation analysis alone is a poor indicator of the method's performance, however, it may provide useful insights and is traditionally used for CASP data.[64] For consistency, we also performed correlation analysis for best150, sel20 and BZQ15 sets (Supplementary Tables S20–S22[10]). Both VoroMQA-new and VoroMQA-new-sr showed top results for best150 sets, but not for sel20. Also, VoroMQA-new-sr showed top results for BZQ15 sets. Overall, the results of the correlation-based analysis are consistent with those of tournament-based tests. In addition, correlation analysis showed a positive trait of Wins-score: the four reference-based scores (GDT-TS, LDDT, SphGr and CAD_AA) correlate better with Wins-score than with each other.

4.3.5 Local scoring

VoroMQA global scores are directly derived from the atom-level VoroMQA scores, so while testing global VoroMQA scores we also indirectly tested VoroMQA local scoring capabilities, at least the cumulative effect of atomic VoroMQA scores. Another possible way of testing local scoring is investigating how local VoroMQA scores conform to some reference-based local scores. However, due to the nature of our method, this approach is not easily applicable as illustrated with the local scoring example in Figure 4.5. The figure shows residue-level VoroMQA scores of a native (target) structure and its two models, the first model is bet-

ter than the second one according to all four reference-based scores. The VoroMQA global scores correctly rank both models by their deviation from the native one. The color-coded VoroMQA local scoring profiles support the judgment of the global scores, because for most of the residue positions the local VoroMQA scores get lower as the global model quality gets lower. However, low absolute values of local VoroMQA scores do not necessarily correspond to low reference-based local scores. For a simple example, let us consider just the native structure. Its local VoroMQA scores are not homogeneous despite all the residue positions being correct. This is so, because different residues are not in equally favorable contact environments. Similarly, when considering a modeled structure different from the native one, low VoroMQA score for a single residue does not necessarily mean that the structural position of the residue is incorrect. This is illustrated in the bottom part of Figure 4.5 with the plots of residue distance deviations obtained from LGA[103] structural alignments and colored by the corresponding VoroMQA scores: some of the well-aligned residues have low VoroMQA scores. Another observation from the same plots is that the positions of the residues with higher VoroMQA scores tend to be well-predicted. To check if this is true in general, we used data from the CAMEO project (`www.cameo3d.org`).[19]

We had the latest version of VoroMQA entered to the CAMEO model quality estimation category under name "VoroMQA_v2" since August 2015, thus we were able to download "1-year" (weeks from 2015.10.17 to 2016.10.08) dataset and analyze it to investigate the relations between VoroMQA local scores and the corresponding LDDT and CAD-score local scores. We looked at the empirical distribution of VoroMQA local scores that correspond to the three classes of reference-based local scores: low, average and high. Similarly, we looked at the distributions of reference-based local scores that correspond to low, average and high local VoroMQA scores. The results, presented in Figure 4.6, prompt us to make two important observations: if the local VoroMQA score for

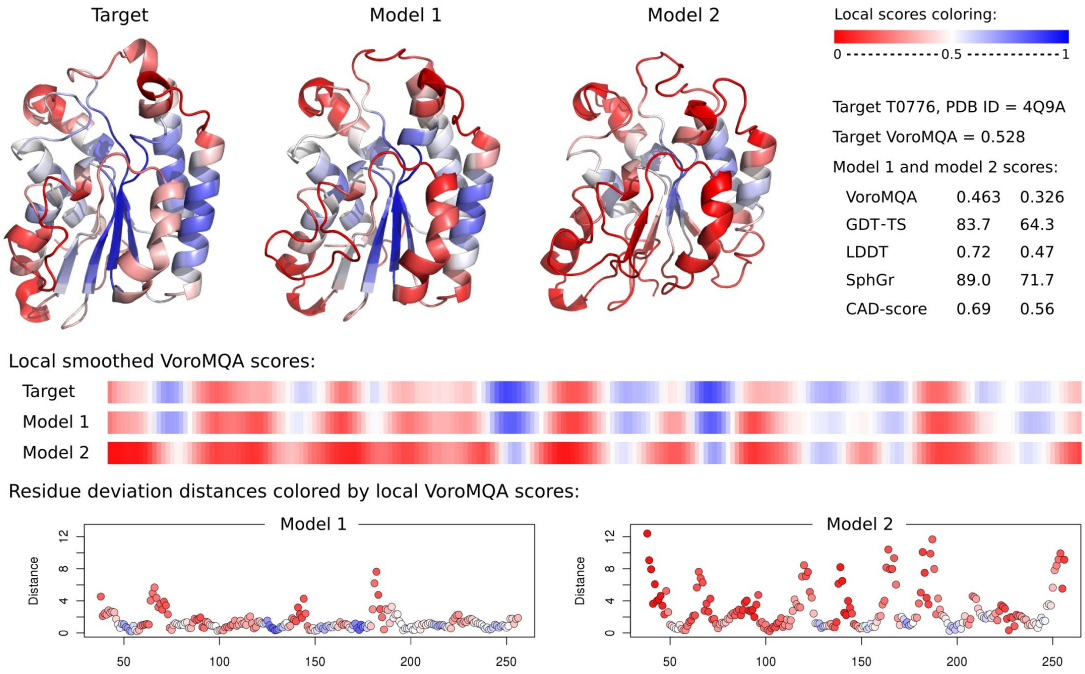


Figure 4.5: Local VoroMQA scores calculated for T0776 target structure and two its models using VoroMQA web-server. The cartoon structural representations are colored by smoothed per-residue VoroMQA scores. The corresponding one-dimensional color-coded profiles are shown in the middle part of the figure. Residue distance deviations (in angstroms) colored by smoothed per-residue VoroMQA scores are plotted in the bottom part of the figure.

a residue in the model is greater than 0.5, the residue is likely well-predicted; if the local VoroMQA score is low (0.25 or less), the accuracy of the residue position is uncertain, because it may mean either incorrectly predicted position or correctly predicted position in unfavorable environment. The latter point is illustrated in Figure 4.5 showing that even a native structure can have regions with relatively low local VoroMQA scores.

Overall, VoroMQA local scores are most useful when analyzed along with the manual inspection of the protein structure. For example, let us inspect models 1 and 2 in Figure 4.5. In model 1 most of the low-scoring residues correspond to solvent-accessible regions while many of the high-scoring ones are buried in the core of the structure, in model 2 the low-scoring regions cover larger parts of the structure and the core is scored much lower than in model 1. These observations allow us to conclude that model 1 is better than model 2 even without considering global scores.

4.4 Discussion

VoroMQA is an all-atom knowledge-based protein structure scoring method. It is important to emphasize that the scoring function of VoroMQA was not optimized or trained in any way to better correspond to any of the reference-based protein structure accuracy measures such as RMSD, GDT-TS or CAD-score. Only an unsupervised learning procedure was applied taking experimentally determined structures of protein biological units (assemblies) as the source of structural information. Also, VoroMQA does not use any additional predictive features, e.g. predicted secondary structure or solvent accessibility that are typically derived using multiple sequence homologs. In other words, only protein 3D structure is needed for its assessment with VoroMQA. Accordingly, VoroMQA falls into the category of statistical energy potentials. How-

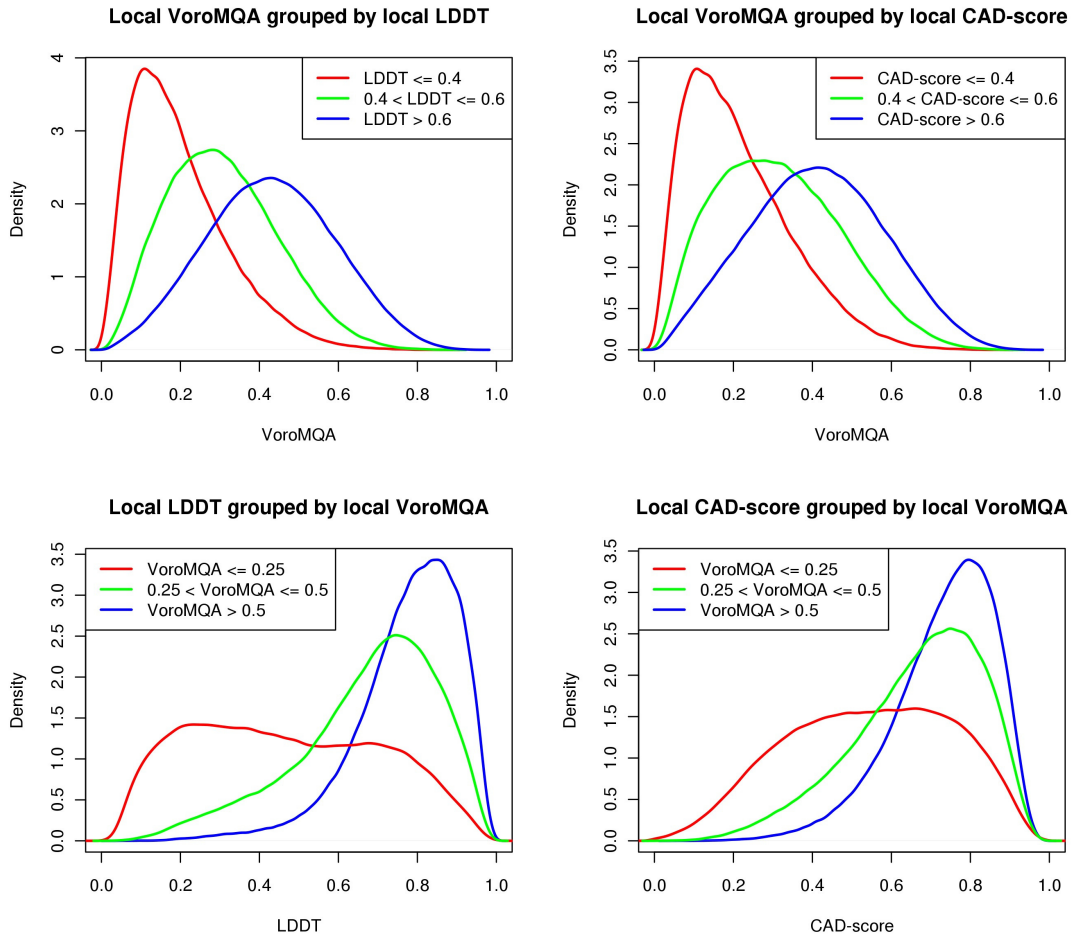


Figure 4.6: Empirical distribution densities of scores obtained from the CAMEO "1-year" dataset. Top row: VoroMQA local scores grouped by the corresponding reference-based local scores, i.e. LDDT and CAD-score. Bottom row: Reference-based local scores grouped by the corresponding VoroMQA local scores.

ever, in contrast to most statistical potentials that are distance-based, our method uses interatomic contact areas. The choice of contact areas offers several advantages. Contact areas not only define physical interactions but also implicitly take into account their strength. Moreover, contact areas make it possible to treat interactions within the protein structure and interactions with solvent in the same way. Interactions of protein atoms with solvent are considered as just another type of contacts. In addition, the use of contact areas allows efficient normalization of pseudo-energy values, so that they can be converted into quality estimates ranging from 0 to 1. This means that the VoroMQA scores are largely independent of the type or the size of an input protein structure.

We tested the performance of VoroMQA by the ability to identify the native structure among the decoys (computational models) in a test typical for statistical potentials. In addition, we explored how well VoroMQA is able to select models by their similarity to the native structure according to different scenarios. Whereas the task of selecting native structure is unambiguous, the evaluation of model selection by their similarity to the native structure is not. There are at least two reasons why evaluation of methods for model selection is not trivial, especially in cases when differences between models are small. One of the reasons is the uncertainty of any reference score.[124] Another reason is that it is quite common for different reference scores to disagree about the exact model ranking. To test the ability of VoroMQA and other methods to select models closest to the native structure we chose the same four reference scores used by the official assessment of model accuracy estimation methods in CASP11,[64] namely, a rigid-body measure (GDT-TS) and three local-structure-based scores (LDDT, CAD-score, and SphereGrinder). However, instead of analyzing the results of these four scores separately,[64] we devised a simple procedure that enabled us to combine all four scores and in so doing to avoid the two problems mentioned above. The main idea of this procedure is that one model is considered to be better (closer to the native

structure) than the other one only if all four reference scores agree to that unanimously. Based on this idea we introduced two scores, Agreement-score and Wins-score, and used them throughout the study for performance comparison of different methods. We believe that by including multiple reference scores simultaneously this procedure provides a robust way for comparing model quality estimation methods. We also believe that this evaluation scheme might be useful for comparing other type of prediction methods as well.

In our tests VoroMQA consistently outperformed DOOP, GOAP and dDFIRE that are similarly based on all-atom statistical potentials, but use distances rather than contact areas. The outcome of these tests is rather unexpected, taking into account that both GOAP and dDFIRE feature orientation-dependent potentials whereas DOOP potentials include the dependence on the backbone torsion angles. In contrast, VoroMQA does not include any terms associated with either conformation preferences of the main chain or orientation-dependence of side chains. This may suggest that contact areas are perhaps more suitable compared to distances in identifying native structure and scoring near-native conformations.

We also tested VoroMQA alongside with model quality assessment methods that in addition to the actual structure utilize various predictions derived using evolutionary information and rely heavily on using machine learning to predict reference-based model quality scores. As VoroMQA does not use any additional information, comparison with such composite methods puts VoroMQA at disadvantage. Despite this, the tests showed that VoroMQA often outperformed these composite methods, especially in the one-out-of-two model selection scenario. VoroMQA achieved top results when tested on the roughly pre-filtered sets of CASP11 models, i.e. the sets comprised of models produced by the top three prediction servers (BZQ15 sets defined in this paper) or the sets comprised of models selected using a simple consensus-based algorithm (best150 sets provided by the CASP11 organizers). It has been previously

observed that the side chain remodeling may lead to improved model selection.[128] Indeed, the rebuilding of side chains for best150 and BZQ15 model sets has further improved VoroMQA results suggesting that significant differences in side chain packing may conceal the main chain similarities.

The only case where VoroMQA (with or without side chain remodeling) was more prone to make mistakes compared to the composite QA scores was when faced with the CASP11 sel20 sets that were composed to contain not better models, but models as different from each other as possible. Such sets, however, hardly represent any real-life model selection scenario. Moreover, we found that the relatively poor performance of VoroMQA in this type of setting could be rescued by simple combination of the VoroMQA score and the evolutionary information in the form of HHpred template-based models. This observation suggests that VoroMQA can be easily incorporated into composite scoring functions.

VoroMQA global scores are directly derived from atom-level scores, so the relatively good results achieved by our method in model selection tests are direct implications of the VoroMQA local scoring capabilities. However, it should be emphasized that local VoroMQA scores of a structural model indicate how energetically favorable or unfavorable the local region is, and not how much it deviates from the corresponding region in the native structure. A native protein structure has a combination of both energetically favorable (e.g. hydrophobic core) and unfavorable regions (e.g. active sites, protein-protein binding sites or solvent-exposed loops). Therefore, even a very accurate structural model will have regions with low VoroMQA scores that will closely reproduce the pattern observed for the native structure (see Figure 4.5). In general our tests indicate that high local VoroMQA scores usually correspond to accurate structural regions. In contrast, low local VoroMQA scores do not necessarily imply that the corresponding region is unrealistic. It may just be one of the regions in a less favorable environment. In other words, VoroMQA local scores could

be used to classify the structure into the accurate regions and those with the uncertain accuracy. In practice, the VoroMQA local scoring perhaps would be most useful in qualitative analysis performed in conjunction with the manual inspection of the protein structure.

In summary, VoroMQA computes meaningful local and global scores and shows robust performance both in recognition of the native structure among decoys and in selecting best models. The use of interatomic contact areas instead of distances might be one of the reasons for relatively good results. Thus, VoroMQA might be a valuable addition to the available set of model quality assessment methods, not only because of strong performance, but also because of its orthogonality to the existing scores.

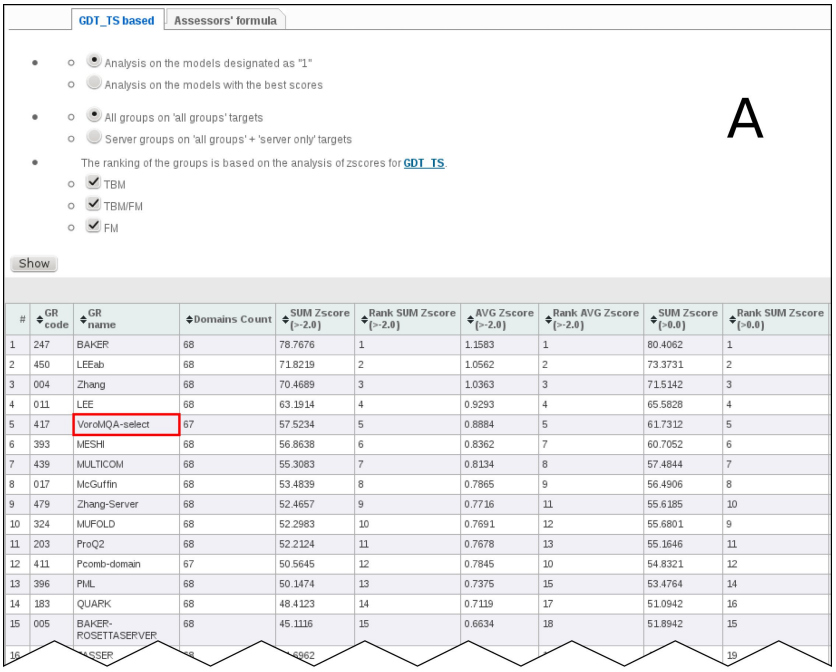
4.5 VoroMQA application in CASP12 and CAPRI experiments

We tested the recently developed VoroMQA method in blind mode during the 2016 world-wide CASP12 experiment. We entered the main CASP category, i.e. the tertiary structure prediction category, as a human group called VoroMQA-select (group members: Kliment Olechnovič and Česlovas Venclovas). The method behind VoroMQA-select is a simple model selection protocol. In short, we used VoroMQA (with and without side-chain rebuilding) to evaluate models available from various automated servers, and submitted the best 5 of them as our predictions. The VoroMQA method was also used to determine if model structures contained unstructured terminal regions that could be removed prior to evaluation: this part was not fully automatic and required manual intervention for deciding the exact cutting locations.

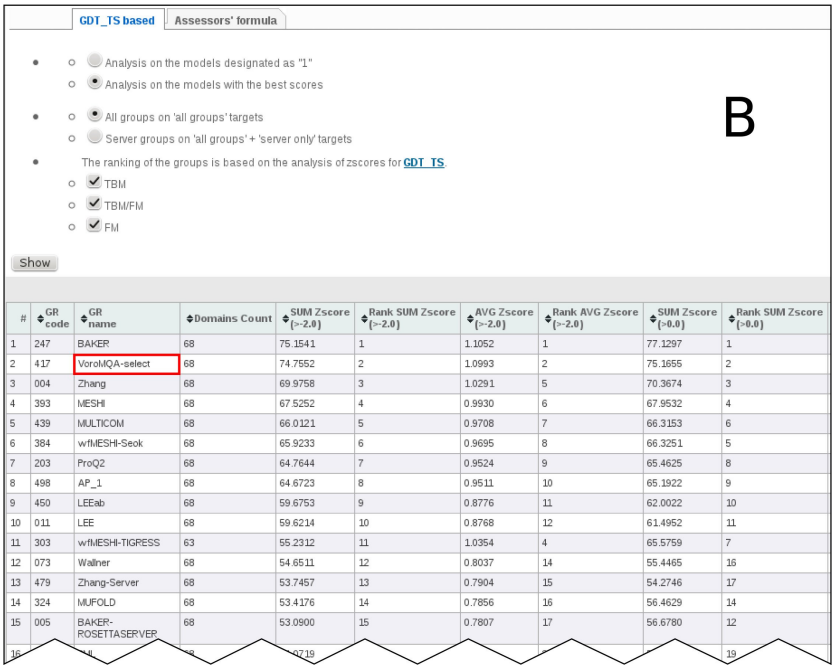
According to the official CASP12 results, available at `www.predictioncenter.org/casp12/zscores_final.cgi`, VoroMQA-select was

5th (out of 128 groups) in the overall ranking if the most confident model (model designated as first by a predictor) was considered (see Figure 4.7 A). More impressively, VoroMQA-select was second if the best-of-five model was considered (see Figure 4.7 B), and the difference with the 1st place was relatively minor. In any case, the groups that outranked ours used modeling methodologies that went beyond the capabilities of modern automated structure prediction servers. There were other groups who, like VoroMQA-select, used QA methods to select best models from server-produced ones, but none of them outperformed VoroMQA-select according to the official CASP12 results.

In addition, the combination of VoroMQA with PPI3D (briefly described in section 3.5) enabled our team ("Venclovas") to produce the best results in the 2016 CAPRI experiment, which was organized in conjunction with CASP12 and was focused on modeling quaternary structures, i.e. protein complexes. The official CAPRI ranking is available at `www.predictioncenter.org/casp12/doc/presentations/CASP12_CAPRI_Lensink.pdf`. It should be noted that our CAPRI team had three members (Justas Dapkūnas, Kliment Olechnovič and Česlovas Venclovas) and the major credit for its performance should go to Dr. Dapkūnas who actually produced the models. As of the role of VoroMQA, it was used to select models of monomers to assemble complexes from and to evaluate and rank modelled complexes.



#	GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)
1	247	BAKER	68	78.7676	1	1.1583	1	80.4062	1
2	450	LEEab	68	71.8219	2	1.0562	2	73.3731	2
3	004	Zhang	68	70.4689	3	1.0363	3	71.5142	3
4	011	LEE	68	63.1914	4	0.9293	4	65.5828	4
5	417	VoroMQA-select	67	57.5234	5	0.8884	5	61.7312	5
6	393	MESHI	68	56.8638	6	0.8362	7	60.7052	6
7	439	MULTICOM	68	55.3083	7	0.8134	8	57.4844	7
8	017	McGuffin	68	53.4839	8	0.7865	9	56.4906	8
9	479	Zhang-Server	68	52.4657	9	0.7716	11	55.6185	10
10	324	MUFOLD	68	52.2983	10	0.7691	12	55.6801	9
11	203	ProQ2	68	52.2124	11	0.7678	13	55.1646	11
12	411	Pcomb-domain	67	50.5645	12	0.7845	10	54.8321	12
13	396	PML	68	50.1474	13	0.7375	15	53.4764	14
14	183	QUARK	68	48.4123	14	0.7119	17	51.0942	16
15	005	BAKER-ROSETTASERVER	68	45.1116	15	0.6634	18	51.8942	15
16		NSSER		0962					19



#	GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)
1	247	BAKER	68	75.1541	1	1.1052	1	77.1297	1
2	417	VoroMQA-select	68	74.7552	2	1.0993	2	75.1655	2
3	004	Zhang	68	69.9758	3	1.0291	5	70.3674	3
4	393	MESHI	68	67.5252	4	0.9930	6	67.9532	4
5	439	MULTICOM	68	66.0121	5	0.9708	7	66.3153	6
6	384	wfMESHI-Seok	68	65.9233	6	0.9695	8	66.3251	5
7	203	ProQ2	68	64.7644	7	0.9524	9	65.4625	8
8	498	AP_1	68	64.6723	8	0.9511	10	65.1922	9
9	450	LEEab	68	59.6753	9	0.8776	11	62.0022	10
10	011	LEE	68	59.6214	10	0.8768	12	61.4952	11
11	303	wfMESHI-TIGRESS	63	55.2312	11	1.0354	4	65.5759	7
12	073	Wallner	68	54.6511	12	0.8037	14	55.4465	16
13	479	Zhang-Server	68	53.7457	13	0.7904	15	54.2746	17
14	324	MUFOLD	68	53.4176	14	0.7856	16	56.4629	14
15	005	BAKER-ROSETTASERVER	68	53.0900	15	0.7807	17	56.6780	12
16		ull		0719					19

Figure 4.7: Screenshots from `www.predictioncenter.org/casp12/zscores_final.cgi` showing the official CASP12 rankings of human tertiary structure prediction groups. Only top 15 places (out of total 128) are shown here. (A) Results considering the most confident model (model 1). (B) Results considering the best-of-five model.

Conclusions

The ultimate result of the studies comprising this dissertation is a collection of novel effective methods for the analysis and evaluation of biomolecular structures. The presented methods construct and utilize the Voronoi tessellation of atomic balls. The usage of the tessellation-derived interatomic contact areas to analyze structural models is the main feature that sets the presented methods apart from the traditional distance-based structure analysis methods. The conclusions related to each of the three developed methods are presented below.

- The first presented method, Voronota, is a method for computing the vertices of the Voronoi diagram of balls that is particularly well-suited for processing three-dimensional structures of biological macromolecules. It takes the advantage of the observation that in macromolecular structures the overwhelming majority of triples of neighboring atomic balls have two tangent planes. When processing each such triple, the Voronota algorithm efficiently combines the knowledge of the search space partitioned by the two tangent planes with the use of hierarchical spatial indexing to find the Voronoi vertices related to the triple. Triples without two tangent planes are extremely rare in macromolecular structures, thus they can be processed using simple brute-force approach without sacrificing the overall processing time. Voronota also features a simple procedure for finding the first valid triple that enabled the parallelization of the algorithm in a straightforward manner. Large-scale tests showed that Voronota is a fast and reliable tool for processing both experimentally determined and computationally modeled macromolecular structures, thus Voronota can serve as a core component for developing other tools that exploit the Voronoi diagram of balls.

- The second presented method, CAD-score (Contact Area Difference Score), is a method for the comparison of different conformations of macromolecules, for example, native and modeled structures. CAD-score works by assessing physical contacts derived from the Voronoi tessellation of atomic balls and computing contact area differences. The method definition is simple, it does not include arbitrary parameters, the defined output value range is $[0, 1]$. The method can directly evaluate the accuracy of both full structures and inter-domain or inter-subunit interfaces. The universal nature of CAD-score allows it to be applied for any major type of biological macromolecular structures (proteins, nucleic acids and various complexes), the method effectively evaluates various nucleic acid-specific subsets of atoms and interaction types (stacking, pairing). CAD-score was tested extensively using protein structural models from CASP experiments. The testing results showed that CAD-score is not only a robust measure for evaluating and ranking single-domain models, it also has advantages over traditional rigid-body superposition-based methods: CAD-score promotes the physical realism of structural models, it provides a balanced assessment of the inter-domain arrangement accuracy in models for multi-domain proteins, thus removing the necessity to split multi-domain proteins into domains for model evaluation purposes. The tests performed using data from RNA-puzzles experiments showed that CAD-score is also effective for the reference-based evaluation of RNA structural models on both global and local levels. Additionally, CAD-score offers an alternative to the superposition-based structure clustering: this possibility was successfully exploited when developing structural data redundancy handling for the PPI3D method of searching and analyzing protein-protein interactions.

- The third presented method, VoroMQA (the Voronoi diagram-based Model Quality Assessment), is a method for the referenceless esti-

mation of protein structure quality. VoroMQA combines the idea of knowledge-based statistical potentials with the advanced use of atom-level solvent-accessible surface areas and interatomic contact areas derived from the Voronoi tessellation of atomic balls. It does not use any additional predictive features, e.g. predicted secondary structure or solvent accessibility. VoroMQA produces local and global quality scores that lay in $(0, 1)$ and are largely independent of the type or the size of an input structure. VoroMQA global scores can be used not only for model selection, but also for deducing if a model is similar to the native structure. VoroMQA local scores can be used to classify the structure into the accurate regions and those with the uncertain accuracy. The tests performed on CASP8-CASP11 data show that VoroMQA generally performs better than other statistical potential-based methods, it also often outperforms methods that use additional evolutionary information. VoroMQA-based model selection protocol was blindly tested in CASP12 structure prediction experiment and showed top results, outperforming other methods that were also based on the idea of selecting best models from automatic prediction server using structure quality assessment methods. VoroMQA also played an important role in achieving best results in protein-protein complex structure modeling experiment CAPRI in 2016.

Overall, the main conclusion of the presented studies is that Voronoi tessellation-derived contact areas capture important structural features of biological macromolecules and are useful as a foundation for new effective methods for the analysis and assessment of three-dimensional structures of proteins and nucleic acids.

Bibliography

[1] J. Kuriyan, B. Konforti, and D. Wemmer, *The molecules of life: physical and chemical principles*. New York, NY: GS, Garland Science, 2013. OCLC: 779577263.

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, pp. 235–242, Jan. 2000.

[3] The UniProt Consortium, "UniProt: a hub for protein information," *Nucleic Acids Res.*, vol. 43, pp. D204–D212, Jan. 2015.

[4] R. B. Altman and J. M. Dugan, "Defining Bioinformatics and Structural Bioinformatics," in *Methods of Biochemical Analysis* (P. E. Bourne and H. Weissig, eds.), pp. 1–14, Hoboken, NJ, USA: John Wiley & Sons, Inc., Jan. 2005.

[5] A. Tramontano, *Protein structure prediction: concepts and applications*. Weinheim: Wiley-VCH, 2006. OCLC: 181462508.

[6] K. Olechnovič, E. Kulberkytė, and C. Venclovas, "CAD-score: a new contact area difference-based function for evaluation of protein structural models," *Proteins*, vol. 81, pp. 149–162, Jan. 2013.

[7] 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.

[8] K. Olechnovič and C. Venclovas, "The use of interatomic contact areas to quantify discrepancies between RNA 3D models and reference structures," *Nucleic Acids Res.*, vol. 42, pp. 5407–5415, May 2014.

[9] K. Olechnovič and C. Venclovas, "The CAD-score web server: contact area-based comparison of structures and interfaces of proteins, nucleic acids and their complexes," *Nucleic Acids Res.*, vol. 42, pp. W259–263, July 2014.

[10] K. Olechnovic and C. Venclovas, "VoroMQA: Assessment of protein structure quality using interatomic contact areas," *Proteins*, Mar. 2017.

[11] J. Dapkunas, A. Timinskas, K. Olechnovic, M. Margelevicius, R. Diciunas, and C. Venclovas, "The PPI3D web server for searching, analyzing and modeling protein–protein interactions in the context of 3D structures," *Bioinformatics*, vol. 33, pp. 935–937, Mar. 2017.

[12] R. A. Abagyan and M. M. Totrov, "Contact area difference (CAD): a robust measure to evaluate accuracy of protein models," *J. Mol. Biol.*, vol. 268, pp. 678–685, May 1997.

[13] A. Zemla, C. Venclovas, J. Moult, and K. Fidelis, "Processing and analysis of CASP3 protein structure predictions," *Proteins*, vol. Suppl 3, pp. 22–29, 1999.

[14] G. Voronoi, "Nouvelles applications des parametres continus a la theorie des formes quadratiques," *J. Reine Angew. Math.*, vol. 134, pp. 198–287, 1908.

[15] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. Wiley Series in Probability and Statistics, Wiley, 2000.

[16] N. B. Leontis and E. Westhof, eds., *RNA 3D structure analysis and prediction*. No. v. 27 in Nucleic acids and molecular biology, Heidelberg ; New York: Springer, 2012. OCLC: ocn759177599.

[17] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP)–round x," *Proteins*, vol. 82 Suppl 2, pp. 1–6, Feb. 2014.

[18] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction: Progress and new directions in round XI," *Proteins*, vol. 84 Suppl 1, pp. 4–14, Sept. 2016.

[19] J. Haas, S. Roth, K. Arnold, F. Kiefer, T. Schmidt, L. Bordoli, and T. Schwede, "The Protein Model Portal–a comprehensive resource for protein structure and model information," *Database*, vol. 2013, p. bat031, 2013.

[20] E. D. Scheeff and J. L. Fink, "Fundamentals of Protein Structure," in *Methods of Biochemical Analysis* (P. E. Bourne and H. Weissig, eds.), pp. 15–39, Hoboken, NJ, USA: John Wiley & Sons, Inc., Jan. 2005.

[21] K. A. Dill, "Dominant forces in protein folding," *Biochemistry*, vol. 29, pp. 7133–7155, Aug. 1990.

[22] G. D. Rose, P. J. Fleming, J. R. Banavar, and A. Maritan, "A backbone-based theory of protein folding," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, pp. 16623–16633, Nov. 2006.

[23] L. Brocchieri and S. Karlin, "Protein length in eukaryotic and prokaryotic proteomes," *Nucleic Acids Res.*, vol. 33, no. 10, pp. 3390–3400, 2005.

[24] S. Neidle, B. Schneider, and H. M. Berman, "Fundamentals of DNA and RNA Structure," in *Methods of Biochemical Analysis* (P. E. Bourne and H. Weissig, eds.), pp. 41–73, Hoboken, NJ, USA: John Wiley & Sons, Inc., Jan. 2005.

[25] S. Neidle, *Nucleic acid structure and recognition*. Oxford ; New York: Oxford University Press, 2002.

[26] C. C. Correll and P. A. Rice, "Chapter 1. Introduction," in *Protein-Nucleic Acid Interactions* (P. A. Rice and C. C. Correll, eds.), pp. 1–12, Cambridge: Royal Society of Chemistry, 2008.

[27] A. Poupon, "Voronoi and Voronoi-related tessellations in studies of protein structure and interaction," *Curr. Opin. Struct. Biol.*, vol. 14, pp. 233–241, Apr. 2004.

[28] F. M. Richards, "The interpretation of protein structures: total volume, group volume distributions and packing density," *J. Mol. Biol.*, vol. 82, pp. 1–14, 1974.

[29] B. J. Gellatly and J. L. Finney, "Calculation of protein volumes: an alternative to the Voronoi procedure," *J. Mol. Biol.*, vol. 161, pp. 305–322, 1982.

[30] A. Goede, R. Preissner, and C. Frömmel, "Voronoi cell: new method for allocation of space among atoms: elimination of avoidable errors in calculation of atomic volume and density," *J. Comput. Chem.*, vol. 18, pp. 1113–1123, 1997.

[31] D.-S. Kim, Y. Cho, and D. Kim, "Euclidean Voronoi diagram of 3D balls and its computation via tracing edges," *Comput. Aided Design*, vol. 37, pp. 1412–1424, Nov. 2005.

[32] J. D. Boissonnat and M. Teillaud, *Effective Computational Geometry for Curves and Surfaces (Mathematics and Visualization)*. Springer-Verlag New York, Inc., 2006.

[33] Y. Cho, D. Kim, H. Lee, J. Park, and D. S. Kim, "Reduction of the search space in the edge-tracing algorithm for the Voronoi diagram of 3D balls," *Lect. Notes. Comput. Sc.*, vol. 3980, pp. 111–120, 2006.

[34] M. Manák and I. Kolingerová, "Fast Discovery of Voronoi Vertices in the Construction of Voronoi Diagram of 3D Balls," in *2010 International Symposium on Voronoi Diagrams in Science and Engineering*, pp. 95–104, 2010.

[35] N. N. Medvedev, V. P. Voloshin, V. A. Luchnikov, and M. L. Gavrilova, "An algorithm for three-dimensional Voronoi S-network," *J. Comput. Chem.*, vol. 27, pp. 1676–1692, 2006.

[36] N. Lindow, D. Baum, and H. C. Hege, "Voronoi-based extraction and visualization of molecular paths," *IEEE T. Vis. Comput. Gr.*, vol. 17, pp. 2025–2034, 2011.

[37] D. S. Kim, D. Kim, Y. Cho, and K. Sugihara, "Quasi-triangulation and interworld data structure in three dimensions," *Comput. Aided Design*, vol. 38, pp. 808–819, 2006.

[38] D. S. Kim, Y. Cho, J. K. Kim, and J. Ryu, "QTF: Quasi-triangulation file format," *Comput. Aided Design*, vol. 44, pp. 835–845, 2012.

[39] B. Delaunay, "Sur la sphere vide," *Izvestiya Akademii Nauk SSSR, Otdelenie Matematicheskikh i Estestvennykh Nauk*, vol. 7, pp. 793–800, 1934.

[40] K. Olechnovic, M. Margelevicius, and C. Venclovas, "Voroprot: an interactive tool for the analysis and visualization of complex geometric features of protein structure," *Bioinformatics*, vol. 27, pp. 723–724, Mar. 2011.

[41] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallogr. Sect. A*, vol. 32, pp. 922–923, Sept. 1976.

[42] J. Moult, J. T. Pedersen, R. Judson, and K. Fidelis, "A large-scale experiment to assess protein structure prediction methods," *Proteins*, vol. 23, pp. ii–iv, Nov. 1995.

[43] J. Moult, T. Hubbard, S. H. Bryant, K. Fidelis, and J. T. Pedersen, "Critical assessment of methods of protein structure prediction (CASP): round II," *Proteins*, vol. Suppl 1, pp. 2–6, 1997.

[44] A. Zemla, n. Venclovas, J. Moult, and K. Fidelis, "Processing and evaluation of predictions in CASP4," *Proteins*, vol. Suppl 5, pp. 13–21, 2001.

[45] N. Siew, A. Elofsson, L. Rychlewski, and D. Fischer, "MaxSub: an automated measure for the assessment of protein structure prediction quality," *Bioinformatics*, vol. 16, pp. 776–785, Sept. 2000.

[46] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins*, vol. 57, pp. 702–710, Dec. 2004.

[47] D. Cozzetto, A. Kryshtafovych, K. Fidelis, J. Moult, B. Rost, and A. Tramontano, "Evaluation of template-based models in CASP8 with standard measures," *Proteins*, vol. 77 Suppl 9, pp. 18–28, 2009.

[48] R. I. Sadreyev, S. Shi, D. Baker, and N. V. Grishin, "Structure similarity measure with penalty for close non-equivalent residues," *Bioinformatics*, vol. 25, pp. 1259–1263, May 2009.

[49] P. Aloy, A. Stark, C. Hadley, and R. B. Russell, "Predictions without templates: new folds, secondary structure, and contacts in CASP5," *Proteins*, vol. 53 Suppl 6, pp. 436–456, 2003.

[50] L. N. Kinch, J. O. Wrabl, S. S. Krishna, I. Majumdar, R. I. Sadreyev, Y. Qi, J. Pei, H. Cheng, and N. V. Grishin, "CASP5 assessment of fold recognition target predictions," *Proteins*, vol. 53 Suppl 6, pp. 395–409, 2003.

[51] J. Kopp, L. Bordoli, J. N. D. Battey, F. Kiefer, and T. Schwede, "Assessment of CASP7 predictions for template-based modeling targets," *Proteins*, vol. 69 Suppl 8, pp. 38–56, 2007.

[52] D. A. Keedy, C. J. Williams, J. J. Headd, W. B. Arendall, V. B. Chen, G. J. Kapral, R. A. Gillespie, J. N. Block, A. Zemla, D. C. Richardson, and J. S. Richardson, "The other 90% of the protein: assessment beyond the Calphas for CASP8 template-based and high-accuracy models," *Proteins*, vol. 77 Suppl 9, pp. 29–49, 2009.

[53] V. Mariani, F. Kiefer, T. Schmidt, J. Haas, and T. Schwede, "Assessment of template based protein structure predictions in CASP9," *Proteins*, vol. 79 Suppl 10, pp. 37–58, 2011.

[54] R. Jauch, H. C. Yeo, P. R. Kolatkar, and N. D. Clarke, "Assessment of CASP7 structure predictions for template free targets," *Proteins*, vol. 69 Suppl 8, pp. 57–67, 2007.

[55] M. Ben-David, O. Noivirt-Brik, A. Paz, J. Prilusky, J. L. Sussman, and Y. Levy, "Assessment of CASP8 structure predictions for template free targets," *Proteins*, vol. 77 Suppl 9, pp. 50–65, 2009.

[56] S. Shi, J. Pei, R. I. Sadreyev, L. N. Kinch, I. Majumdar, J. Tong, H. Cheng, B.-H. Kim, and N. V. Grishin, "Analysis of CASP8 targets, predictions and assessment methods," *Database*, vol. 2009, p. bap003, 2009.

[57] J. F. Atkins, R. F. Gesteland, and T. Cech, eds., *RNA worlds: from life's origins to diversity in gene regulation*. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory Press, 2011. OCLC: ocn649700128.

[58] M. A. Jonikas, R. J. Radmer, A. Laederach, R. Das, S. Pearlman, D. Herschlag, and R. B. Altman, "Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters," *RNA*, vol. 15, pp. 189–199, Feb. 2009.

[59] C. E. Hajdin, F. Ding, N. V. Dokholyan, and K. M. Weeks, "On the significance of an RNA tertiary structure prediction," *RNA*, vol. 16, pp. 1340–1349, July 2010.

[60] E. Capriotti, T. Norambuena, M. A. Marti-Renom, and F. Melo, "All-atom knowledge-based potential for RNA structure prediction and assessment," *Bioinformatics*, vol. 27, pp. 1086–1093, Apr. 2011.

[61] M. Parisien, J. A. Cruz, E. Westhof, and F. Major, "New metrics for comparing and assessing discrepancies between RNA 3D structures and models," *RNA*, vol. 15, pp. 1875–1885, Oct. 2009.

[62] P. Lukasiak, M. Antczak, T. Ratajczak, J. M. Bujnicki, M. Szachniuk, R. W. Adamiak, M. Popenda, and J. Blazewicz, "RNAlyzer–novel approach for quality analysis of RNA structural models," *Nucleic Acids Res.*, vol. 41, pp. 5978–5990, July 2013.

[63] A. Kryshtafovych, A. Barbato, K. Fidelis, B. Monastyrskyy, T. Schwede, and A. Tramontano, "Assessment of the assessment: evaluation of the model quality estimates in CASP10," *Proteins*, vol. 82 Suppl 2, pp. 112–126, Feb. 2014.

[64] A. Kryshtafovych, A. Barbato, B. Monastyrskyy, K. Fidelis, T. Schwede, and A. Tramontano, "Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11," *Proteins*, vol. 84 Suppl 1, pp. 349–369, Sept. 2016.

[65] A. Ray, E. Lindahl, and B. Wallner, "Improved model quality assessment using ProQ2," *BMC Bioinformatics*, vol. 13, p. 224, 2012.

[66] P. Benkert, S. C. E. Tosatto, and D. Schomburg, "QMEAN: A comprehensive scoring function for model quality assessment," *Proteins*, vol. 71, pp. 261–277, Apr. 2008.

[67] M. J. Sippl, "Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins," *J. Mol. Biol.*, vol. 213, pp. 859–883, June 1990.

[68] M. J. Sippl, "Recognition of errors in three-dimensional structures of proteins," *Proteins*, vol. 17, pp. 355–362, Dec. 1993.

[69] H. Zhou and Y. Zhou, "Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction," *Protein Sci.*, vol. 11, pp. 2714–2726, Nov. 2002.

[70] M.-y. Shen and A. Sali, "Statistical potential for assessment and prediction of protein structures," *Protein Sci.*, vol. 15, pp. 2507–2524, Nov. 2006.

[71] J. Zhang and Y. Zhang, "A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction," *PLoS ONE*, vol. 5, no. 10, p. e15386, 2010.

[72] M.-H. Chae, F. Krull, and E.-W. Knapp, "Optimized distance-dependent atom-pair-based potential DOOP for protein structure prediction," *Proteins*, vol. 83, pp. 881–890, May 2015.

[73] Y. Yang and Y. Zhou, "Specific interactions for ab initio folding of protein terminal regions with secondary structures," *Proteins*, vol. 72, pp. 793–803, Aug. 2008.

[74] M. Lu, A. D. Dousis, and J. Ma, "OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing," *J. Mol. Biol.*, vol. 376, pp. 288–301, Feb. 2008.

[75] H. Zhou and J. Skolnick, "GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction," *Biophys. J.*, vol. 101, pp. 2043–2052, Oct. 2011.

[76] B. J. McConkey, V. Sobolev, and M. Edelman, "Discrimination of native protein structures using atom-atom contact scoring," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, pp. 3215–3220, Mar. 2003.

[77] K. Rother, P. W. Hildebrand, A. Goede, B. Gruening, and R. Preissner, "Voronoia: analyzing packing in protein structures," *Nucleic Acids Res.*, vol. 37, pp. D393–395, Jan. 2009.

[78] J. Esque, S. Léonard, A. G. de Brevern, and C. Oguey, "VLDP web server: a powerful geometric tool for analysing protein structures in their environment," *Nucleic Acids Res.*, vol. 41, pp. W373–378, July 2013.

[79] A. Zomorodian, L. Guibas, and P. Koehl, "Geometric filtering of pairwise atomic interactions applied to the design of efficient statistical potentials," *Comput. Aided Geom. D.*, vol. 23, pp. 531–544, Aug. 2006.

[80] M. Mirzaie and M. Sadeghi, "Delaunay-based nonlocal interactions are sufficient and accurate in protein fold recognition," *Proteins*, vol. 82, pp. 415–423, Mar. 2014.

[81] R. Jafari, M. Sadeghi, and M. Mirzaie, "Investigating the importance of Delaunay-based definition of atomic interactions in scoring of protein-protein docking results," *J. Mol. Graph. Model.*, vol. 66, pp. 108–114, May 2016.

[82] P. Su and R. L. S. Drysdale, "A comparison of sequential Delaunay triangulation algorithms," *Comput. Geom.*, vol. 7, pp. 361–385, 1997.

[83] T. Masaharu, T. Ogawa, and N. Ogita, "A new algorithm for three-dimensional voronoi tessellation," *J. Comput. Phys.*, vol. 51, pp. 191–207, 1983.

[84] A. Maus, "Delaunay triangulation and the convex hull ofn points in expected linear time," *BIT Numer. Math.*, vol. 24, pp. 151–163, 1984.

[85] D. S. Kim, Y. Cho, and K. Sugihara, "Quasi-worlds and quasi-operators on quasi-triangulations," *Comput. Aided Design*, vol. 42, pp. 874–888, 2010.

[86] M. L. Gavrilova and J. Rokne, "Updating the topology of the dynamic Voronoi diagram for spheres in Euclidean d-dimensional space," *Comput. Aided Geom. D.*, vol. 20, pp. 231–242, 2003.

[87] W. Degen, "Cyclides," in *Handbook of computer aided geometric design*, pp. 575–601, Elsevier, 2002.

[88] L. Druoton, L. Garnier, R. Langevin, H. Marcellier, and R. Besnard, "Blending Planes and Canal Surfaces Using Dupin Cyclides," *Comm. Com. Inf. Sc.*, vol. 167, pp. 406–420, 2011.

[89] J. Spillmann, M. Becker, and M. Teschner, "Efficient updates of bounding sphere hierarchies for geometrically deformable models," *J. Vis. Commun. Image R.*, vol. 18, pp. 101–108, 2007.

[90] J. L. Bentley, "Multidimensional Binary Search Trees Used for Associative Searching," *Commun. ACM*, vol. 18, pp. 509–517, 1975.

[91] D. S. Kim, Y. Cho, J. Ryu, J. K. Kim, and D. Kim, "Anomalies in quasi-triangulations and beta-complexes of spherical atoms in molecules," *Comput. Aided Design*, vol. 45, pp. 35–52, 2013.

[92] W. Kahan, "Pracniques: Further Remarks on Reducing Truncation Errors," *Commun. ACM*, vol. 8, p. 40, 1965.

[93] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge University Press, 2002.

[94] A. Bondi, "van der Waals Volumes and Radii," *J. Phys. Chem.*, vol. 68, pp. 441–451, 1964.

[95] J. Moult, K. Fidelis, A. Kryshtafovych, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP)–round IX," *Proteins*, vol. 79 Suppl 10, pp. 1–5, 2011.

[96] 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.

[97] B. J. McConkey, V. Sobolev, and M. Edelman, "Quantification of protein surfaces, volumes and atom-atom contacts using a constrained Voronoi procedure," *Bioinformatics*, vol. 18, pp. 1365–1373, Oct. 2002.

[98] A. Dietrich and B. Maigret, "Program for the visualization of inorganic crystals," *J. Mol. Graph.*, vol. 9, pp. 85–90, 97–99, June 1991.

[99] A. Tramontano, D. Cozzetto, A. Giorgetti, and D. Raimondo, "The assessment of methods for protein structure prediction," *Methods Mol. Biol.*, vol. 413, pp. 43–57, 2008.

[100] A. Kryshtafovych, J. Moult, S. G. Bartual, J. F. Bazan, H. Berman, D. E. Casteel, E. Christodoulou, J. K. Everett, J. Hausmann, T. Heidebrecht, T. Hills, R. Hui, J. F. Hunt, J. Seetharaman, A. Joachimiak, M. A. Kennedy, C. Kim, A. Lingel, K. Michalska, G. T. Montelione, J. M. Otero, A. Perrakis, J. C. Pizarro, M. J. van Raaij, T. A. Ramelot, F. Rousseau, L. Tong, A. K. Wernimont, J. Young, and T. Schwede, "Target highlights in CASP9: Experimental target structures for the critical assessment of techniques for protein structure prediction," *Proteins*, vol. 79 Suppl 10, pp. 6–20, 2011.

[101] L. N. Kinch, S. Shi, H. Cheng, Q. Cong, J. Pei, V. Mariani, T. Schwede, and N. V. Grishin, "CASP9 target classification," *Proteins*, vol. 79 Suppl 10, pp. 21–36, 2011.

[102] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, "MolProbity: all-atom structure validation for macromolecular crystallography," *Acta Crystallogr. D Biol. Crystallogr.*, vol. 66, pp. 12–21, Jan. 2010.

[103] A. Zemla, "LGA: A method for finding 3D similarities in protein structures," *Nucleic Acids Res.*, vol. 31, pp. 3370–3374, July 2003.

[104] J. L. MacCallum, A. Pérez, M. J. Schnieders, L. Hua, M. P. Jacobson, and K. A. Dill, "Assessment of protein structure refinement in CASP9," *Proteins*, vol. 79 Suppl 10, pp. 74–90, 2011.

[105] J. A. Cruz, M.-F. Blanchet, M. Boniecki, J. M. Bujnicki, S.-J. Chen, S. Cao, R. Das, F. Ding, N. V. Dokholyan, S. C. Flores, L. Huang, C. A. Lavender, V. Lisi, F. Major, K. Mikolajczak, D. J. Patel, A. Philips, T. Puton, J. Santalucia, F. Sijenyi, T. Hermann, K. Rother, M. Rother, A. Serganov, M. Skorupski, T. Soltysinski, P. Sripakdeevong, I. Tuszynska, K. M. Weeks, C. Waldsich, M. Wildauer, N. B. Leontis, and E. Westhof, "RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction," *RNA*, vol. 18, pp. 610–625, Apr. 2012.

[106] P. Gendron, S. Lemieux, and F. Major, "Quantitative analysis of nucleic acid three-dimensional structures," *J. Mol. Biol.*, vol. 308, pp. 919–936, May 2001.

[107] S. Lemieux and F. Major, "RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire," *Nucleic Acids Res.*, vol. 30, pp. 4250–4263, Oct. 2002.

[108] I. Tuszynska and J. M. Bujnicki, "DARS-RNP and QUASI-RNP: new statistical potentials for protein-RNA docking," *BMC Bioinformatics*, vol. 12, p. 348, Aug. 2011.

[109] M. Sokolowska, M. Kaus-Drobek, H. Czapinska, G. Tamulaitis, R. H. Szczepanowski, C. Urbanke, V. Siksnys, and M. Bochtler, "Monomeric Restriction Endonuclease BcnI in the Apo Form and in an Asymmetric Complex with Target DNA," *J. Mol. Biol.*, vol. 369, pp. 722–734, June 2007.

[110] K. Katoh, K.-i. Kuma, H. Toh, and T. Miyata, "MAFFT version 5: improvement in accuracy of multiple sequence alignment," *Nucleic Acids Res.*, vol. 33, no. 2, pp. 511–518, 2005.

[111] D. Butina, "Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets," *Journal of Chemical Information and Computer Sciences*, vol. 39, pp. 747–750, July 1999.

[112] J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, and A. Tramontano, "Critical assessment of methods of protein structure prediction - Round VIII," *Proteins*, vol. 77 Suppl 9, pp. 1–4, 2009.

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

[114] V. Mariani, M. Biasini, A. Barbato, and T. Schwede, "lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests," *Bioinformatics*, vol. 29, pp. 2722–2728, Nov. 2013.

[115] P. L. M. Antczak, T. Ratajczak, J. Blazewicz, P. Lukasiak, and J. Blazewicz, "SphereGrinder - reference structure-based tool for quality assessment of protein structural models," pp. 665–668, IEEE, Nov. 2015.

[116] J. Li, R. Cao, and J. Cheng, "A large-scale conformation sampling and evaluation server for protein tertiary structure prediction and its assessment in CASP11," *BMC Bioinformatics*, vol. 16, p. 337, 2015.

[117] R. Cao and J. Cheng, "Protein single-model quality assessment by feature-based probability density functions," *Sci. Rep.*, vol. 6, p. 23990, 2016.

[118] K. Uziela and B. Wallner, "ProQ2: estimation of model accuracy implemented in Rosetta," *Bioinformatics*, vol. 32, pp. 1411–1413, May 2016.

[119] T. Liu, Y. Wang, J. Eickholt, and Z. Wang, "Benchmarking Deep Networks for Predicting Residue-Specific Quality of Individual Protein Models in CASP11," *Sci. Rep.*, vol. 6, p. 19301, 2016.

[120] G. G. Krivov, M. V. Shapovalov, and R. L. Dunbrack, "Improved prediction of protein side-chain conformations with SCWRL4," *Proteins*, vol. 77, pp. 778–795, Dec. 2009.

[121] D. E. Kim, D. Chivian, and D. Baker, "Protein structure prediction and analysis using the Robetta server," *Nucleic Acids Res.*, vol. 32, pp. W526–531, July 2004.

[122] Y. Zhang, "I-TASSER server for protein 3D structure prediction," *BMC Bioinformatics*, vol. 9, p. 40, 2008.

[123] D. Xu and Y. Zhang, "Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field," *Proteins*, vol. 80, pp. 1715–1735, July 2012.

[124] W. Li, R. D. Schaeffer, Z. Otwinowski, and N. V. Grishin, "Estimation of Uncertainties in the Global Distance Test (GDT_ts) for CASP Models," *PLoS ONE*, vol. 11, no. 5, p. e0154786, 2016.

[125] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, p. 80, Dec. 1945.

[126] V. Modi, Q. Xu, S. Adhikari, and R. L. Dunbrack, "Assessment of template-based modeling of protein structure in CASP11," *Proteins*, vol. 84 Suppl 1, pp. 200–220, Sept. 2016.

[127] A. Hildebrand, M. Remmert, A. Biegert, and J. Söding, "Fast and accurate automatic structure prediction with HHpred," *Proteins*, vol. 77, no. S9, pp. 128–132, 2009.

[128] B. Wallner, "ProQM-resample: improved model quality assessment for membrane proteins by limited conformational sampling," *Bioinformatics*, vol. 30, pp. 2221–2223, Aug. 2014.

Publications by the author

Papers that this dissertation is based on

1. Kliment Olechnovič, Eleonora Kulberkytė and Česlovas Venclovas. *CAD-score: a new contact area difference-based function for evaluation of protein structural models*. Proteins (2013) 81 (1): 149-162.

2. Kliment Olechnovič and Česlovas Venclovas. *Voronota: A fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls*. J. Comput. Chem. (2014) 35 (8): 672-681.

3. Kliment Olechnovič and Česlovas Venclovas. *The use of interatomic contact areas to quantify discrepancies between RNA 3D models and reference structures*. Nucleic Acids Res. (2014) 42 (9): 5407-5415.

4. Kliment Olechnovič and Česlovas Venclovas. *The CAD-score web server: contact area-based comparison of structures and interfaces of proteins, nucleic acids and their complexes*. Nucleic Acids Res. (2014) 42 (W1): 259-263. *NAR Breakthrough Article*.

5. Justas Dapkūnas, Albertas Timinskas, Kliment Olechnovič, Mindaugas Margelevičius, Rytis Dičiūnas and Česlovas Venclovas. *The PPI3D web server for searching, analyzing and modeling protein-protein interactions in the context of 3D structures*. Bioinformatics (2017) 33 (6): 935-937.

6. Kliment Olechnovič and Česlovas Venclovas. *VoroMQA: Assessment of protein structure quality using interatomic contact areas*. Proteins (2017) 10.1002/prot.25278.

Other papers

1. Kliment Olechnovič, Mindaugas Margelevičius and Česlovas Venclovas. *Voroprot: an interactive tool for the analysis and visualization of complex geometric features of protein structure*. Bioinformatics (2011) 27 (5): 723-724.

International conference presentations in 2012–2016

1. EMBO Conference on Critical Assessment of Protein Structure Prediction (Gaeta, Italy, 2012.12.9-12). Poster presentation: *CAD-score: a new method for the evaluation of protein structural models*. Best poster award, selected for an oral presentation.

2. SocBiN: Society for Bioinformatics in Northern European countries (Torun, Poland, 2013.06.26-29). Poster presentation: *The use of interatomic contact areas for the assessment of RNA 3D structural models*. Best poster award.

3. Intelligent Systems for Molecular Biology (Berlin, Germany, 2013.07.19-23). Poster presentation: *The use of interatomic contact areas for the assessment of RNA 3D structural models*.

4. European Conference on Computational biology (Strasbourg, France, 2014.09.7-10). Poster presentation: *The CAD-score webserver: contact area-based comparison of structures and interfaces of proteins, nucleic acids and their complexes*.

5. 11$^{\text{th}}$ Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (Riviera Maya, Mexico, 2014.12.7-10). Poster presentation: *Quality assessment of single protein structure models using inter-atom contact areas derived from the Voronoi diagram of atomic balls*.

6. European Conference on Computational biology (Hague, Netherlands, 2016.09.3-7). Poster presentation: *Estimation of protein structure quality using contact areas derived from the Voronoi tessellation of atomic balls*.

7. 12$^{\text{th}}$ Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (Gaeta, Italy, 2016.12.10-13). Poster presentation: *VoroMQA: assessment of protein structure quality using interatomic contact areas derived from the Voronoi tessellation of atomic balls*. Best poster award, selected for an oral presentation.

Curriculum Vitae

Personal information

Name:	Kliment Olechnovič
Birth:	1987.08.18, Vilnius, Lithuania
Phone:	+370-6049-1418
Email:	`kliment@ibt.lt`

Education

2012 – 2016	PhD student in Computer Science, Vilnius University.
2010 – 2012	M.S. (*Magna Cum Laude*) Computer Science, Vilnius University.
2005 – 2009	B.S. Bioinformatics, Vilnius University.
1993 – 2005	L. Karsavin secondary school, Vilnius.

Work

2016 – now	Junior researcher at Department of Bioinformatics, Institute of Biotechnology, Vilnius University.
2010 – 2016	Research engineer at Department of Bioinformatics, Institute of Biotechnology, Vilnius University.
2009 – 2010	Research assistant at Department of Bioinformatics, Institute of Biotechnology.
2007 – 2009	C++ software developer, 4Team Corporation, Vilnius.

Publications

7 peer-reviewed papers in international journals.
12 presentations at international conferences (3 awards for best poster).

Awards

2015	Lithuanian Academy of Sciences award for best works by young researchers in 2014.
2013–2015	The Research Council of Lithuania scholarship for PhD students actively conducting scientific research.
2013	INFOBALT incentive scholarship for young scientists.

List of abbreviations

CAD	Contact Area Difference
CASP	community wide experiment on the Critical Assessment of techniques for protein Structure Prediction
DI	Deformation Index
GDT	Global Distance Test
GDT-TS	Global Distance Test Total Score
INF	Interaction Network Fidelity
PDB	Protein Data Bank
PPI	Protein-Protein Interactions
QA	Quality Assessment
RMSD	Root Mean Square Deviation
SAS	Solvent-Accessible Surface
TBM	Template-Based Modeling
VDW	van der Waals radius

Abstract in Lithuanian (Santrauka)

Disertacijoje aprašyti trys nauji metodai, skirti baltymų ir nukleorūgščių struktūroms analizuoti ir vertinti. Pristatyti metodai konstruoja ir naudoja atomų rutulių Voronojaus diagramą. Tarpatominių kontaktų plotų, išvedamų iš Voronojaus diagramos, panaudojimas struktūrų analizei yra pagrindinis bruožas, skiriantis naujus metodus nuo tradicinių, aprašančių sąveikas remiantis atstumais. Pirmasis metodas, Voronota, skirtas rutulių Voronojaus diagramos viršūnėms konstruoti. Jis efektyviai apdoroja makromolekulių struktūras išnaudodamas žinias apie dažnai pasitaikančias atomų rutulių erdvinio išsidėstymo konfigūracijas. Voronota yra efektyvus įrankis tarpatominėms sąveikoms identifikuoti. Antrasis metodas, CAD-score, skirtas makromolekulių skirtingoms konformacijoms lyginti. Jis efektyviai sprendžia struktūrinių modelių vertinimo esant etalonui užduotį, ir išvengia tradiciniams etaloninio vertinimo metodams būdingų problemų naudodamas kontaktų plotus. CAD-score gali efektyviai analizuoti ir lyginti visų pagrindinių biologinių makromolekulių (baltymų, nukleorūgščių ir jų kompleksų) struktūras. Trečiasis metodas, VoroMQA, skirtas baltymų struktūrinių modelių tikslumui nusakyti nežinant etaloninės struktūros. Jis efektyviai derina empirinio statistinio potencialo idėją su tarpatominių kontaktų plotų, gaunamų iš atomų rutulių Voronojaus diagramos, panaudojimu. VoroMQA sistemingai sprendžia modelių kokybės vertinimo užduotis geriau negu kiti statistiniais potencialais paremti metodai. Svarbiausia disertacijos išvada yra tai, kad atomų Voronojaus diagramos pagrindu sukonstruoti kontaktai ir jų plotai atspindi svarbias biologinių makromolekulių ypatybes ir gali būti sėkmingai naudojami kaip pagrindas naujiems efektyviems baltymų ir nukleorūgščių struktūrų analizės ir vertinimo metodams kurti.