

Contact Map Prediction based on Cellular Automata and Protein Folding Trajectories

Nestor Milciades Diaz Mariño

Supervisor:

Ph. D. Irene Tischer

Thesis submitted for the degree of Doctor in Engineering – Emphasis Computing Science

Universidad del Valle

Engineering Faculty

School of Systems Engineering and Computation

October 2019

Abstract

In Structural Bioinformatics, it is necessary to know the protein's tertiary structure because its specific shape is central in its interaction with binding molecules. Being experimental tertiary structure determination a highly expensive process, computational protein structure prediction becomes an alternative option aimed toward cost and technical limitations reduction.

In the last decade, residue-residue protein contact prediction (PCP) has taken broad consideration. Currently, PCP has become a common subtask of computational structure prediction. Residue-residue interactions can constraint the space of possible protein conformations, improving protein structure determination. Despite the recent improvements in PCP, the high rate of false positive predicted contacts hinders the applicability of existing PCP tools. To reduce the false positive rate in PCP, we developed a novel approach based on cellular automata (CAs), which determines residue-residue contacts that are likely to be actual contacts. Our approach exploits the local interactions found in protein contact maps and the iterative refinement provided by CAs. Our CAs were identified using a parallel genetic algorithm which used for training the PSICOV data set (150 proteins). To benchmark our approach, we used the CASP12 data set (Critical Assessment of Techniques for Structure Prediction, year 2016). Our best CA outperformed the ten PCP tools compared in the benchmark. However, a more

detailed analysis using non-parametric Friedman's statistical test revealed that our tool does not excel the performance of prominent PCP tools such as MetaPSICOV and RaptorX-Contact. Although our CA-based approach for PCP was successful, the precision for long-range contacts (sequence separation > 24 amino acids) was hard to improve. To enrich local interactions, we proposed a multiclass contact map representation that can improve long-range PCP. Our multiclass contact map was obtained using a large-scale comparison of decision trees. The next step to follow is to reformulate our CA-based approach to incorporate multiclass contacts and repeat the overall process to obtain a new PCP tool.

Table of Contents

Table of Contents	i
List of Tables	v
List of Figures	vi
List of Annexes	ix
Chapter I. Introduction.....	1
1. Thesis Overview	5
2. References	7
Chapter II. Referential Framework	9
1. Glossary	10
Inter-Residue Contact.	10
Contact Range.	10
Sequence Separation.	11
Contact Map.....	12
Reference Databases.	13
Template Based Modeling and Prediction.	13
Free Modeling.	13
Correlated Mutation.	13
Direct Coupling Analysis.....	15
Mutual Information.	16
Critical Assessment of Protein Structure Prediction (CASP).	16

2.	State of the Art	17
2.1	Evaluation Protocol in CASP.....	17
2.2	Performance Measures in CASP12.....	18
2.3	Target Proteins in CASP12.....	21
2.4	Top Predictors in CASP12.....	23
	MetaPSICOV	24
	DeepFold_ Contact, iFold_1, and naive	27
	RaptorX-Contact	28
	MULTICOM-CONSTRUCT.....	29
	PConsC31 and PCons-net.....	30
	FALCON_COLORS.....	32
	IGBTeam.....	32
3	Cellular Automata and PCP	33
4	Conclusions.....	36
5	References.....	37
Chapter III. GACAI-PCP: Cellular Automata based Tool for Contact Map Prediction ...		44
	Abstract	44
1.	Introduction.....	44
2.	Materials and Methods.....	47
2.1.	Initial Conditions Dataset	47
2.2.	Cellular Automata Identification Framework.....	49
2.3.	Genetic Algorithm for Identification of CAs that evolve	
	Contact Maps	49

2.4.	Density Classification Task as Benchmark for GACAI-PCP.....	51
3.	Discussion	52
4.	Conclusion	55
5.	References.....	56
Chapter IV. Contact Map Prediction based on Cellular Automata.....		58
	Abstract	58
1.	Introduction.....	59
2.	Materials and Methods.....	63
2.1.	Databases and Training Data Sets.....	63
2.2.	Initial Conditions Generation.....	64
2.3.	Standard Tools	72
2.4.	CA-based PCP Method Description	72
3.	Prediction Results	76
3.1.	Validation Data Set.....	76
3.2.	PCP-CA Model 1	76
3.3.	Prediction Example using PCP-CA	79
3.4.	PCP-CA Evaluation	80
4.	Discussion	82
5.	Conclusion	84
6.	References.....	85
Chapter V. Multiclass Protein Contact Maps		89
	Abstract	89
1.	Introduction.....	90

2.	Methods.....	92
2.1	Contact Ranges Combinations.....	94
2.2	Datasets Generation	96
2.3	Decision Trees Building and Evaluation	97
2.4	Top Models Selection	98
2.5	Top Decision Trees Comparison	100
3.	Results.....	104
3.1	Contact Refinement	105
4	Conclusions.....	108
5.	References.....	109
Chapter VI. Conclusions and Future Work.....		112
References		115

List of Tables

Chapter II. Referential Framework

Table 1. Consolidated results in CASP12 official residue-residue contact predictions. .. 20

Table 2. Top 10 predictors CASP12 by an average of performance measures. 25

Chapter IV. Contact Map Prediction based on Cellular Automata

Table 1. All performance measures for predictors MetaPSICOV and RaptorX-Contact. 71

Table 2. PCP-CA rule description. 77

Table 3. Precision by target/predictor CASP12 top 10 plus PCP-CA, for full list contacts and medium/long range sequence separation. 81

Table 4. Average precision for all predictors at several sizes of list of contacts for all target in CASP12. 83

Chapter V. Multiclass Protein Contact Maps

Table 1. Average rankings of DTs for top contact range combinations. 103

Table 2. MCC for each protein trajectory in the test set for contact refinement. 107

Table 3. Wilcoxon test for contact refinement comparison. 108

List of Figures

Chapter I. Introduction

Figure 1. Project breakdown structure.	3
---	---

Chapter II. Referential Framework

Figure 1. Native structure for protein 1vii.	12
Figure 2. Four contact maps for protein 1vii, contact range 8 Å.	14
Figure 3. Native structure and contact map for protein 1vii.	15
Figure 4. CASP12 PCP targets. First half.	23
Figure 5. CASP12 PCP targets. Second half.	24
Figure 6. Proposed pipeline for PCP training and CA-based PCP prediction.	37

Chapter III. GACAI-PCP: Cellular Automata based Tool for Contact Map Prediction

Figure 1. RaptorX-Contact predicted contact map and actual contact map for target protein T0900 (CASP12).	47
Figure 2. Pipeline for the generation of the dataset of initial conditions.	48
Figure 3. Genetic Algorithm Process for CA identification.	50
Figure 4. Examples of Moore's CA neighborhood.	53
Figure 5. Performance of the five CAs on the CASP12 evaluation dataset.	53
Figure 6. Dominance graph analysis with statistical significance comparison by Friedman's test and post-hoc Nemenyi's test.	55

Chapter IV. Contact Map Prediction based on Cellular Automata

Figure 1. MetaPSICOV2 predicted contact and actual contact map for target protein T0866 (CASP12).	62
Figure 2. Dominance graph for top 10 contact predictors in CASP12, FM targets, top L/2 contacts, full + medium sequence separation, and precision as performance measure.	67
Figure 3. Dominance graph for top 10 contact predictors in CASP12, FM targets, top L contacts, full + medium sequence separation, and precision as performance measure.	68
Figure 4. Dominance graph for top 10 contact predictors in CASP12, FM targets, FL, full + medium sequence separation, and precision as performance measure.	70
Figure 5. Dominance graph for top 10 contact predictors in CASP12, FM targets, FL, full + medium sequence separation, and accuracy as performance measure.	71
Figure 6. Pipeline for a conventional PCP approach.	74
Figure 7. Pipeline for CA-based PCP.	75
Figure 8. PCP-CA Neighborhood.	76
Figure 9. Examples of transitions in the PCP-CA rule.	78
Figure 10. Contact map predicted by PCP-CA / MetaPSICOV2 for CASP12 target protein T0891.	79
Figure 11. Dominance graph for top 10 contact predictors in CASP12 plus PCP-CA for CASP12 targets, full list of contacts predicted, full + medium sequence separation, and precision as a performance measure.	82
 <i>Chapter V. Multiclass Protein Contact Maps</i>	
Figure 1. Method RoadMap. Three phases, inputs/outputs, and processes.	94
Figure 2. EM clustering of DTs performance measures for 413 contact ranges combinations.	99

Figure 3. Frequencies by rule ranges in the dataset.	101
Figure 4. Nemenyi’s test results.....	103
Figure 5. The rule in a 2-state contact map vs. 4-state contact map for Model 1.....	104
Figure 6. Average results for contact map refinement.....	106

List of Annexes

Annex 1: Generic framework for mining cellular automata models on protein-folding simulations (Published Paper).

Annex 2: GACAI-PCP: Cellular Automata based Tool for Contact Map Prediction (Published Paper).

Annex 3: Data mining approach to improve protein contact maps (Submitted Paper).

Annex 4: Mining Stochastic Cellular Automata to Solve Density Classification Task in Two Dimensions (Submitted Paper).

Chapter I.

Introduction

Proteins are a primary component of the vital processes for all kind of living beings, including the human being. The native structure of a protein determines the biological function that it will fulfill, and for this reason, it is necessary to know this structure. For several decades, multidisciplinary researchers have been working on methods that allow predicting or identifying a near-native structure for a protein from its amino acid sequence only. Despite all existing approaches to protein structure prediction, the problem remains unsolved, given the low quality of predicted structures especially for proteins with few homologs. In the last decade, residue-residue contacts have become a common source of information about constraints in the space of possible conformations (Ovchinnikov et al., 2018). Contacts stabilize protein folds, at all levels including short range interactions as well as long range.

The high rate of false contacts that predictors usually produce (Kosciolek & Jones, 2014), is the topic addressed in our approach. Our solution exploits the mechanism based on local information that cellular automata (CAs) provide. Our CA-based approach improves the performance of residue-residue contact prediction. CAs fit in the task related to improving the full list of contacts predicted because they can generate global coordination from local information only. CAs do not require global control and computation is done from a set of local transitions

defined in a small window around each cell of the contact map. Chapter II introduces the reader to the concepts related to protein contact map prediction (PCP) and CA.

A CA in 2D arranges information in a matrix similar to a contact map. However, they provide two complementary features: 1) a set of local transitions, called rule; 2) and the capability to evolve the initial matrix state using the CA rule. Our approach requires to identify the rule of a CA able to evolve an initial predicted contact map to a more realistic one.

In this thesis, we develop the hypothesis that CAs can improve predicted contact maps. However, the CAs modeling must be accomplished by a machine learning approach because the mechanism able to evolve a predicted contact map into a closer to the real one, is unknown. In the context of our approach, the main goal was to predict contact maps with better accuracy. As mentioned above, one current issue in protein contact prediction (PCP) is the high proportion of false positives (FP), so that our approach was focused in the reduction of FP to improve accuracy. In Figure 1, we use a project breakdown structure (Becker et al., 2014) to depict how the thesis development was structured. To accomplish our main goal of predicting contact maps, we developed four specific objectives, and for each one, we mention in Figure 1 the methods, the resources, and the outcomes.

The first specific objective was the implementation of a genetic algorithm (GA) that models CAs for PCP. The GA task is to design a CA from the information provided by pairs of initial configurations (IC) – final configurations (FC), this kind

of approach is called inverse design (Back et al., 2005). An IC is the initial state or initial prediction, and the FC is the desired final global state where the GA must identify the CA rule that evolves the IC to the desired state or improved prediction. We describe the software architecture of our approach in (Díaz & Tischer, 2016) (Annex 1 “Generic framework for mining cellular automata models on protein-folding simulations”). The specific details of the GA for PCP are described in Chapter III “GACAI-PCP: Cellular Automata based Tool for Contact Map Prediction” and was published in (Diaz & Tischer, 2019).

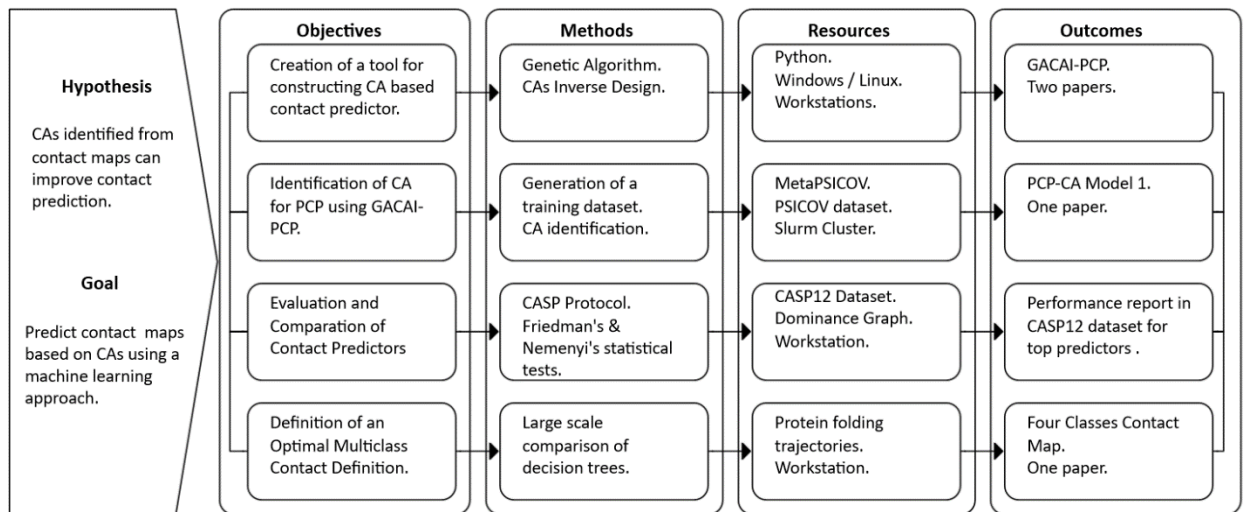


Figure 1. Project breakdown structure.

The thesis was organized in a way that objectives are accomplished through the methods, which utilize the resources to achieve the outcomes that support the hypothesis about the use of CAs in PCP.

Further, we showed that the software framework could solve diverse problems related to the inverse design of CAs, such as the density classification task

(DCT) (De Oliveira, 2014). Our results for DCT are described in Annex 4 “Mining Stochastic Cellular Automata to Solve Density Classification Task in Two Dimensions.”

The second specific objective involved the generation of the training dataset comprised of ICs obtained from MetaPSICOV (Buchan & Jones, 2018) predictions and their native contact maps (desired final state). MetaPSICOV was the best option, as we concluded through the statistical comparison of the top predictors that we describe in Chapter III. The set of protein sequences used for training was the one described in (Jones et al., 2012). The identification of CAs is a computationally expensive task; therefore, it was necessary to run the GA in a high-performance computing cluster.

The third specific objective aims for the evaluation of PCP results by following the CASP (Critical Assessment of Protein Structure Prediction) protocol (Monastyrskyy et al., 2011). The validation dataset contains the target proteins published for CASP round 12 (CASP12). For a more detailed comparison of our predictor and the CASP top-10, we propose a graphical tool called Dominance Graph that displays the rankings and dominance relationships for predictors in an all vs. all comparison. The details of the results and the comparison with the top-10 predictors for the best identified CA are described in Chapter IV.

The last specific objective opens the way to a complementary approach in contact map prediction. The binary definition of contacts in traditional approaches is an inherent source of complexity for PCP. The binary definition (contact / no –

contact) implies that several inter – secondary structure interactions are hard to predict because these are scarce in the contact map. As a complement to the work with binary contact maps, we developed a large – scale comparison of multiclass contact maps that convey richer information for scarce contacts. These results were submitted to the Journal of Machine Learning Research under the title “Data mining approach to improve protein contact maps.” We are currently developing an algorithm to generate ICs for optimal multiclass contact map representation because none of the existing predictors is compatible with our proposed definition.

In this thesis, we explored the use of CAs to improve PCP, and satisfactorily proved that its characteristic emergent global coordination was helpful to reduce the excessive proportion of FP that the state of art predictor MetaPSICOV produces.

1. Thesis Overview

Chapter II starts with a description of the relevant concepts in PCP and CAs. Then we survey the results for PCP in CASP₁₂, we select and describe the top-10 predictors. The chapter concludes, showing the relationship between CAs and PCP.

Chapter III describes our approach for inverse design of CAs, the CA definition we used, and preliminary results. Finally, we compare our identified CA for PCP with other CAs that use traditional neighborhoods instead of the machine learning designed neighborhood of our approach.

In Chapter IV, we describe, analyze, and survey our best CA in the context of the top-10 predictors defined in Chapter II. In particular, we focused our analysis on Friedman’s framework and Nemenyi’s tests and dominance graphs.

Chapter V reports our approach to find an optimal definition of multiclass contact map. Our optimal multiclass definition was selected from a set of plus four – hundred combinations of contact ranges. We evaluated each candidate multiclass definition using decision trees and performance measures adapted for this problem. Finally, Friedman’s and Nemenyi’s tests allowed us to select an optimal definition.

In Chapter VI, we resume the main contributions of this thesis and indicate further work.

Annex 1 contains our published paper “Generic framework for mining cellular automata models on protein-folding simulations,” which describes the software architecture at the core of our approach.

Annex 2 comprises our published paper “GACAI-PCP: Cellular Automata based Tool for Contact Map Prediction.” GACAI-PCP is our GA designed to identify CAs for PCP.

In Annex 3, we include our submitted paper “Data mining approach to improve protein contact maps,” which describe the methods and results of our approach for multiclass contact map definition.

Annex 4 contains our submitted paper “Mining Stochastic Cellular Automata to Solve Density Classification Task in Two Dimensions,” which describes our solution to DCT that allowed us to tune the parameters for GACAI-PCP.

2. References

- Back, T, Breukelaar, R, & Willmes, LBT-UPP. (2005). Inverse Design of Cellular Automata by Genetic Algorithms: An Unconventional Programming Paradigm. In *Unconventional Programming Paradigms* (pp. 161–172). Springer. Retrieved from http://dx.doi.org/10.1007/11527800_13
- Becker, M, Schütt, B, & Amini, S. (2014). *Proposal Writing for International Research Projects: A Guide for Teachers*.
- Buchan, DWA, & Jones, DT. (2018). Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins: Structure, Function and Bioinformatics*. <http://doi.org/10.1002/prot.25379>
- De Oliveira, PPB. (2014). On Density Determination With Cellular Automata : Results, Constructions and Directions. *Journal of Cellular Automata*, 9(5–6), 357–385.
- Díaz, N, & Tischer, I. (2019). GACAI-PCP : Cellular Automata based Tool for Contact Map Prediction. In *17th LACCEI Latin American and Caribbean Conference for Engineering and Technology* (pp. 1–5).
- Díaz, N, & Tischer, I. (2016). Generic framework for mining cellular automata models on protein-folding simulations. *Genetics and Molecular Research*, 15(2), 1–16. <http://doi.org/10.4238/gmr.15028654>
- Jones, DT, Buchan, DW a, Cozzetto, D, & Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics (Oxford, England)*, 28(2), 184–90. <http://doi.org/10.1093/bioinformatics/btr638>
- Kosciolek, T, & Jones, DT. (2014). De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PLoS ONE*, 9(3), 1–15. <http://doi.org/10.1371/journal.pone.0092197>
- Monastyrskyy, B, Fidelis, K, Tramontano, A, & Kryshchuk, A. (2011). Evaluation of residue–residue contact predictions in CASP9. *Proteins*, 79(Suppl 10), 119–125. <http://doi.org/10.1002/prot.23160>.Evaluation

Ovchinnikov, S, Park, H, Kim, DE, DiMaio, F, & Baker, D. (2018). Protein structure prediction using Rosetta in CASP12. *Proteins: Structure, Function and Bioinformatics*, 86(July 2017), 113–121. <http://doi.org/10.1002/prot.25390>

Chapter II.

Referential Framework

Protein contact prediction (PCP) is a research area which originates at least four decades ago. The first approach that exploits the benefits of contact prediction is the one described in (Tanaka & Scheraga, 1975). Tanaka and Scheraga propose a 3-step algorithm for protein folding, where each step incorporates information related with short, medium and long contact ranges, respectively. Later, Gromiha and Selvaraj make an extended review (Gromiha & Selvaraj, 2004), where they consolidate several decades of knowledge about inter-residue contacts and their relevance in protein structure studies. Nowadays, an important quantity of research groups is working in PCP, some in standalone predictors and others in pipelines of native protein structure predictors where PCP has an important role.

Two main approaches allow classifying the best contact predictors (Sun, Huang, Wang, Zhang, & Shen, 2015): mutual information (MI) and evolutionary coupling (EC). Since 1996, the Critical Assessment of Structure Prediction (CASP) has given special attention to PCP in its biennial competition (Monastyrskyy, D'Andrea, Fidelis, Tramontano, & Kryshchuk, 2015). After several years of development, EC has become perhaps the approach with the most significant improvements in the CASP competition.

Even though the improvement in average precision is vast, PCP requires further advancement. In CASP₁₂ (2016), the performance was around twice as good as the results in CASP₁₀ (2014), but the overall precision is still under 50% (Schaarschmidt, Monastyrskyy, Kryshchuk, & Bonvin, 2017). In this thesis, we propose and test some novel algorithms based on cellular automata which allow us to improve PCP precision by adding a new prediction layer to the results of well-known efficient predictors (e.g., MetaPSICOV (Buchan & Jones, 2018), PSICOV (Jones, Buchan, Cozzetto, & Pontil, 2012), and EvFold (Morcos et al., 2011)). Our approach uses cellular automata as the core for machine learning of rules for PCP refinement.

In this chapter, we describe the conceptual framework for PCP (as glossary) in the section 1. We describe the state of the art in section 2 and a statistical analysis of contact predictors performance in section 3. In section 4 we describe how PCP is related to cellular automata. Finally, we draw conclusions in section 5.

1. Glossary

Inter-Residue Contact. The inter-residue contact defines the closeness between a pair of residues of a protein by specifying a contact range. It allows identifying structural characteristics of a tertiary structure of a protein. Inter-residue contact is synonymous of residue-residue contact (RRC).

Contact Range. The contact range is generally a sphere placed in an atom of a residue's backbone and if the same atom of another residue of the protein is found

inside this sphere the two residues are in contact. In PCP, predictors use carbon alpha ($C\alpha$) or carbon beta ($C\beta$) as reference atoms for contact analysis. The common contact range is 8 Å, and for CASP evaluation, the reference atom is the $C\beta$ ($C\alpha$ for glycine). In some cases, the contact range may also be based on the centroid of the residue or the mass center of the side chain.

Sequence Separation. In PCP, the sequence separation between a pair of residues is a relevant factor in the precision evaluation because medium and long ranges are the hardest targets for prediction, although they are crucial for providing structural constraints (Monastyrskyy et al., 2015). For evaluation purposes there are three sequence separation ranges (Holland, Pan, & Grigoryan, 2018): short-range, for sequence separation greater than five and smaller than twelve residues; medium-range, for separation greater than eleven and below twenty-four residues; and long-range separation for pairs of residues with separation above twenty-three. PCP performance evaluation excludes sequence separation inferior to six residues because contacts at this range are situated mostly inside secondary structure elements (Monastyrskyy, D'Andrea, Fidelis, Tramontano, & Kryshtafovych, 2014). In Figure 1 we show examples of contacts for protein ivii, at medium and long range separations. In Figure 1, medium-range contacts define contacts between secondary structure elements: Coil 1 – Alpha helix 2; Coil 2 – Alpha helix 3; and Coil 2 – Coil 4. At long-range sequence separation contacts appear for Coil 1 – Coil 4 and Coil 2 – Coil 4.

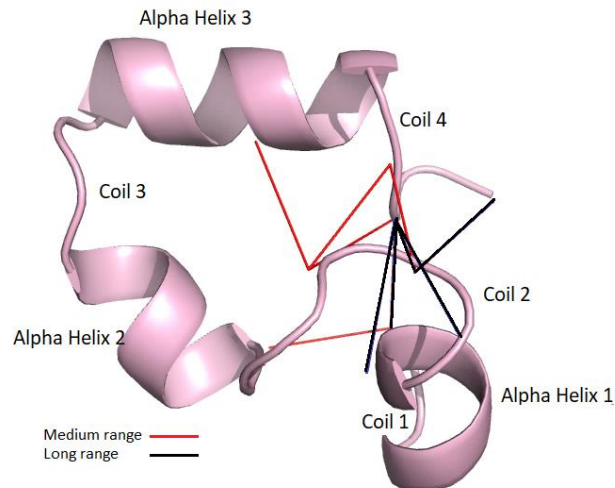


Figure 1. Native structure for protein 1vii.

The figure displays contacts in the contact range 8 Å for medium and long-range sequence separation. Lines indicate contacts (i.e., gray for medium-range and black for long-range).

Contact Map. A contact map (CM) is a binary matrix that in row i denotes which residues of the protein are in contact range of residue i . In Figure 2, we display four contact maps for protein 1vii. This protein has a sequence of thirty-six amino acids, so its contact map is a binary matrix of size 36×36 . Each panel in Figure 2 is a contact map that exemplifies the differences between four common definitions of contact range. Panels I and II of Figure 2, based on the $C\alpha$ and $C\beta$ atoms, respectively, are among the more popular contact range definitions. Currently, the de facto standard is the contact range definition using $C\beta$ atoms and a threshold of 8 Å, because the predictors that participate in CASP competitions, or that at least use CASP targets and evaluation methodology as a benchmark, require to use it. In Figure 3, we show the $C\beta$ contact map at 8 Å threshold for protein 1vii and its 3D native structure. In the native structure, residues in contact are shown linked by a straight line. Arrows in Figure 3 connect contacts for residues VAL₁₀-LEU₂ and THR₁₄-ASP₄, whose contact distance is 5.9 Å and 7.9 Å, respectively.

Reference Databases. Reference databases are sets of protein sequences that are used commonly in sequence analysis in PCP. Some reference databases are: PDB70 that is constructed from PSI-BLAST (Altschul et al., 1997) alignments of sequences extracted from the Protein Data Bank (PDB) with sequence similarity inferior to 70%; Uniref100 (Suzek, Wang, Huang, McGarvey, & Wu, 2015) that provides clusters of sequences extracted from UniProt Knowledgebase (Bateman et al., 2017) where of a number of repeated subsequences, only one unique sequence is stored to avoid redundancy. Uniref70, Uniref50 and Uniref20 contain the same sequences as Uniref100 but they are grouped at 70%, 50% and 20% of sequence identity.

Template Based Modeling and Prediction. Some protein structures are modeled using as initial structure another protein structure. In prediction, many predictors require homolog structures to make a precise approximation of the desired structure.

Free Modeling. Free modeling (FM) is a category of prediction used to indicate that a protein lacks a template or homolog structure to be used in structure prediction or any other kind of analysis. Prediction in the FM situation is hardest because Multisequence Alignments (MSAs), sequence-related data and local structure information provide important support for prediction.

Correlated Mutation. Correlated Mutation is one of the most important concepts for the current success of contact prediction. The idea behind it is that there is a relationship between sequence mutations and structure, where a mutation drives to a correlated mutation because of physical constraints (Göbel, Sander, Schneider, & Valencia, 1994). Several methods integrate this concept and are categorized into two

main approaches: Direct Coupling Analysis (DCA) and Mutual Information (MI) explained next.

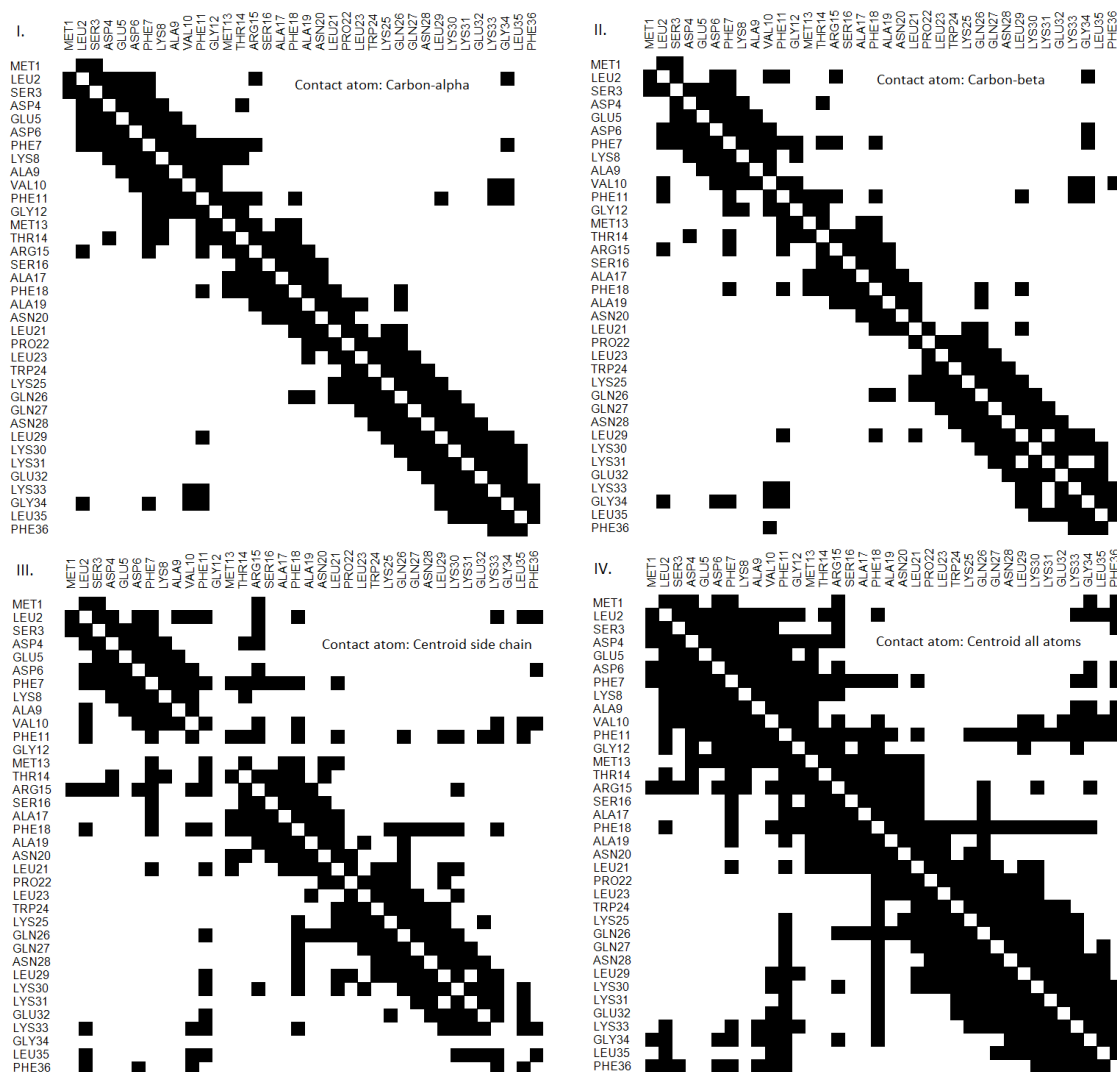


Figure 2. Four contact maps for protein 1vii, contact range 8 Å.

Panel I. Distances were calculated for atom $C\alpha$ of each residue pair. *Panel II.* Distances were calculated for atom $C\beta$ of each residue pair. *Panel III.* Distances were calculated for the centroid of the residue side chain for each residue pair. *Panel IV.* Distances were calculated for the centroid of all the residue atoms for each residue pair.

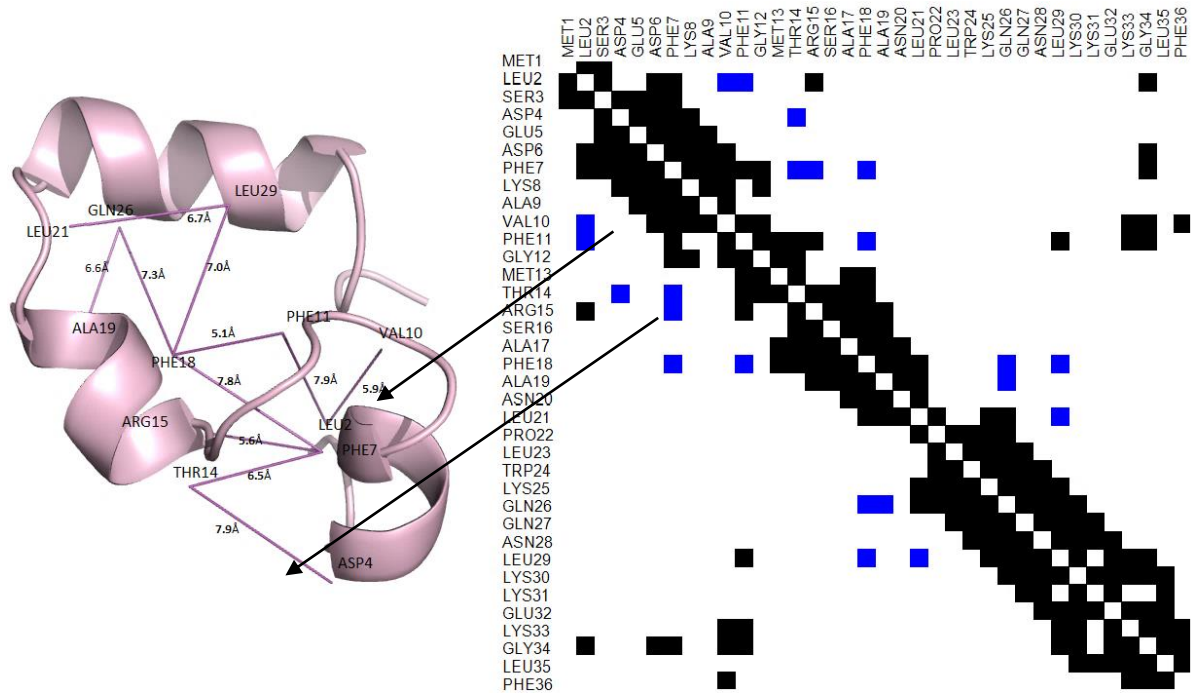


Figure 3. Native structure and contact map for protein 1vii.

Contacts with a sequence separation greater than six residues and below 12 residues are shown in blue.

Direct Coupling Analysis. DCA is a method for extracting information from sequence related data (Morcos et al., 2011). DCA allows inferring contacts based on amino acid composition at some sites of the protein sequence. DCA is one of the main approaches used by successful predictors. This kind of approach is useful, when it is able to discern where correlations are caused by direct relationships between amino acids and not by indirect inference. Message-passing DCA (mpDCA) is an important example of the DCA category. mpDCA locally uses MI and later applies a statistical model that incorporates frequency of amino acids at specific positions and direct couplings for those amino acids (Weigt, White, Szurmant, Hoch, & Hwa, 2009). mpDCA implements a message-passing approach to reduce computational complexity, but it is still an expensive method. Mean field DCA (mfDCA) is a

computationally more efficient approach for DCA (Morcos et al., 2011). mfDCA reduces the computational cost by using analytical methods instead of the iterative one used in mpDCA. mfDCA is implemented by the open source Freecontact (Kaján, Hopf, Kalaš, Marks, & Rost, 2014). Freecontact offers an additional implementation for PSICOV (Jones et al., 2012), which is a DCA based method that uses sparse inverse covariance based on the lasso concept for neighborhood selection in sparse matrices (Meinshausen & Bühlmann, 2006). Other approaches such as plmDCA (Ekeberg, Lökvist, Lan, Weigt, & Aurell, 2013), GREMLIN (Kamisetty, Ovchinnikov, & Baker, 2013) and CCMpred (Seemayer, Gruber, & Söding, 2014), include pseudolikelihood optimization as the core to avoid spurious indirect correlations.

Mutual Information. Mutual Information (MI) (Martin, Gloor, Dunn, & Wahl, 2005), implements the concept that the identity of an amino acid in a position is able to predict which amino acid will be in another position. We mention two approaches for MI, row-column weighting (RCW) (Gouveia-Oliveira & Pedersen, 2007) and average product correction (APC) (Dunn, Wahl, & Gloor, 2008).

Critical Assessment of Protein Structure Prediction (CASP). CASP is a well-established biennial competition that focuses on diverse problems in the field of protein structure. In its website (predictioncenter.org), there is information about all competitions held since 1994 (CASP₁). Currently, CASP develops five prediction categories: template-based modeling, ab-initio modeling, contact prediction, refinement, and data assisted modeling. In the PCP category, predictors are applied to a protein sequences dataset of unknown structures, and assessors, chosen

specifically for each competition, perform the evaluation. The evaluation protocol implies to select a top of residues pairs in contact. The top may be selected in dependence of the length (L) of the protein sequence. Evaluators commonly use the top 10, top $L/5$, top $L/2$, L and full list (FL) of predicted contacts to assess a predictor. To differentiate best predictions, the result for each pair of residues in contact must have a probability for such contact, which is defined by the predictor itself. Higher probability indicates a better likelihood for that pair of residues to be in contact in the native structure.

2. State of the Art

In this section, we describe and compare the top contact predictors using the results of CASP₁₂ because this round is the last one with full results published until now. The next round of the competition (CASP₁₃) is scheduled to run from April - 2018 to December - 2018 and the full results will be published later.

2.1 Evaluation Protocol in CASP

As predictors based on machine learning advance in quality performance, the CASP evaluation protocol includes measures that are very common in machine learning. Precision, recall, Matthews correlation coefficient (MCC) and area under the precision-recall curve (AUC_{PR}) were used in CASP₁₁ (Kryshtafovych, Monastyrskyy, & Fidelis, 2016). Predictors are evaluated in two ways: reduced list (RL) and full list (FL) (Monastyrskyy et al., 2015). RL evaluation ensures that all

predictors are assessed on the same quantity of contacts. The typical RL sizes are top-5, top-10, top L/5, top L/2, top L and top 2L (Adhikari, Nowotny, Bhattacharya, Hou, & Cheng, 2016), where L is the length of the protein sequence.

2.2 Performance Measures in CASP12

The performance measures that evaluators selected in CASP12 were: precision (Equation 1), harmonic mean (F1) (Equation 3), recall (Equation 2) and a pair of measures for entropy: ES (Equation 4) and ES_ext (Equation 4). Entropy measures try to capture the dispersion of predicted contacts along the overall structure (Schaarschmidt et al., 2017). Alternative adaptations of precision, F1 and recall that try to incorporate the probability associated for a predictor to each RRC predicted were also included. Being based on probabilities, these measures are skewed by the training data and the scoring method used for each predictor, so we are excluding those measures in our analysis. This topic was addressed previously by Jinbo Xu from the RaptorX-Contact team (S. Wang, Sun, & Xu, 2018) in a Web entry (<http://ttic.uchicago.edu/~jinbo/WhyF1probsIncorrect.htm>, visited: November 2018).

$$precision = \frac{TP}{TP+FP}, TP: \text{True positives}; FP: \text{False positives}. \quad (1)$$

$$recall = \frac{TP}{P}, P: \text{Number of contacts in the protein}. \quad (2)$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (3)$$

$$ES = 100 \times \frac{Entropy|0 - Entropy|C}{Entropy|0} \quad (4)$$

Entropy|0: entropy of the protein without distance restrictions imposed by contacts.

Entropy|C: entropy of the protein with a set of restrictions defined by predicted contacts.

Entropy|x = $\frac{1}{N} \sum_{i=1}^N H_i$, N: sequence length,

$$H_i = \frac{1}{N-1} \sum_{j=1, j \neq i}^N \log(u_{ij} - l_{ij}), u_{ij}: \text{upper limit}, l_{ij}: \text{lower limit} \quad (5)$$

$u_{ij} = 8 \text{ \AA}$, for contacts

$$u_{ij} = \begin{cases} 5.54 \text{ \AA} \times N^{0.34} \text{ \AA}, & \text{for non contacts in ES} \\ 3.8 \text{ \AA} \times N, & \text{for non contacts in ES_ext} \end{cases}$$

$l_{ij} = 3.2 \text{ \AA}$

Some other measures applied by assessors in CASP 12 can be used only for FL contacts because true negatives (TN) and false negatives (FN) need to be known. TN is the number of non-contacts pairs that are not listed in the predictions file. FN is the number of true contacts that are not predicted. One of these measures is the Matthews correlation coefficient (MCC) (Equation 6). MCC is useful when we have predictions with binary decisions (e.g., contact vs. non-contact) because it recognizes biases to contacts or non-contacts predictions. MCC has values in the range [-1,1], where a value closer to zero indicates randomness in the predictions. MCC identifies a fully correct predictor at value 1, and a predictor that just makes all wrong predictions at value -1.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \quad (6)$$

At CASP12 official results web page

(predictioncenter.org/casp12/rrc_avrg_per_target_results.cgi, visited: March 2018),

we can obtain the results for predictions on 38 targets. The results are consolidated in Table 1. We can appreciate that precision is greater for the RL's L/5 and L/2 than for the FL, which means that more contacts on RL are real contacts than in FL. However, recall in RL points out that the overall rate of predicted contacts covered by predictors is low for both top L/5 (average recall = 4.32%) and L/2 (average recall = 8.73%). By comparing recall and precision for RL in Table 1 we conclude, that a better choice is using RL predictions for top L/2 since recall for L/2 is twice as big as recall for L/5. This means that by losing about 6% in precision, we can obtain 100% more TP. For FL predictions, quite the opposite occurs (bigger recall, smaller precision), the reason behind this issue is that even when the average recall is greater than 60%, there is an excessive number of FP that reduces the average precision to less than 7%. From Table 1 we deduce that in PCP there is room for improvement specifically in approaches that make predictions with a lower proportion of FP.

Table 1. Consolidated results in CASP12 official residue-residue contact predictions. Scores belong to sequence separation at medium and long range.

	TOP L/5 PREDICTIONS				TOP L/2 PREDICTIONS				FULL LIST PREDICTIONS			
MEASURE	Min	Max	Average	Std. Desv.	Min	Max	Average	Std. Desv.	Min	Max	Average	Std. Desv.
F1	1.36	14.70	7.65	3.75	2.60	28.63	13.27	6.67	2.11	19.98	9.39	3.97
PRECISION	3.69	77.68	35.41	20.15	3.17	70.19	29.09	16.93	1.18	14.98	6.63	3.37
RECALL	0.79	8.31	4.32	2.09	2.20	17.98	8.73	4.23	28.98	82.12	63.44	13.59
ES	0.02	0.55	0.19	0.15	0.04	0.99	0.35	0.28	0.32	6.32	2.22	1.79
ES(EXT)	0.80	11.94	4.44	3.54	1.03	17.51	6.69	4.94	3.25	35.52	15.93	10.43
RECALL_SS	0.00	35.48	14.89	8.39	0.00	67.74	24.20	12.95	0.00	95.16	74.56	16.72

A few years ago, (Kim, DiMaio, Wang, Song, & Baker, 2014) found that it is possible to model protein topology with one TP for every twelve residues in the protein. Developing this conception with a recall of around 8% is enough to achieve

this amount of contacts. However, this is only possible with high precision since in an actual application of contact predictions to protein structure modeling the researcher can no discriminate true contacts from FP. In Table 1 with RL truncated at $L/2$, the average recall is above 8% while average precision is below 30%, which implies that around two of three contacts predicted are FP. For FL predictions, the situation for structure modeling is the opposite, because average recall is above 60% and average precision is inferior to 7%. Therefore, for FL is very hard to pick the real contacts from the set of predictions. We think that novel approaches could be able to increase the balance in predictions, that is, keeping the recall rate but improving precision. In the published results for PCP in CASP12 (Schaarschmidt et al., 2017), assessors performed a similar experiment where they used the contact predictions submitted and analyzed the impact of several RL sizes in structure modeling. They observed that in some cases RL top $L/5$ provide scarce TP to enhance structure modeling. For results with predictors such as RaptorX-Contact and MetaPSICOV at $L/2$, L , $1.5L$ and $2L$ sizes of RL, the improvements are largest. For predictions at RL top $3L$, there is no improvement or even modeling is worse than without contact information. We believe that this is due to the fact that the number of FP increases with the list size.

2.3 Target Proteins in CASP12

Targets in CASP are classified into three groups: Template-Based Modeling (TBM), Free Modeling (FM), and TBM with hard to discover templates (TBM/FM).

For contact predictor evaluations it is more interesting to assess performance in FM targets, because in the absence of templates, contacts can provide us with constraints to explore the conformational space (Schaarschmidt et al., 2017).

FM targets in the CASP₁₂ dataset have low similarity. We try to use FastaHerder₂ (Mier & Andrade-Navarro, 2016) to cluster sequences by similarity and size but it was not possible to obtain clusters because there are vast differences for all instances in the dataset. The full dataset has high variability in sequence length (standard deviation: 78.73 amino acids), with a minimum and maximum sequence size of 66 and 356 amino acids, respectively. In Figures 4 and 5 we show 3D structures for targets in CASP₁₂. For each target, we show the cartoon representation and its CM. We notice that there are many differences between targets, which allow us to say that, as benchmark dataset, CASP₁₂ targets are illustrative for the main protein families defined in SCOP (Andreeva et al., 2008).

For top L/5, L/2 and FL contacts, the minimum for precision belongs to target To863 which is the largest protein in the dataset and an all- α protein. This target was hard to predict because it has many α -helix/ α -helix contacts at long sequence separation. RaptorX-Contact (S. Wang et al., 2018) and MetaPSICOV (Buchan & Jones, 2018) were the best predictors that succeeded predicting a few of these contacts (recall for secondary structure contacts (recall_{ss}) 13.15% and 10.81%, respectively). The protein with best predictions was the α + β target To886, that has a more elongated structure. The average for FL precision, recall and recall_{ss} were 13.36%, 67.9%, and 83.2%. Recall and recall_{ss} indicate us that almost every

predictor was able to predict many contacts in the protein, though the low precision is the result of many FP. For the top L/5 and L/2 contacts, the precision is greater because the list of predictions only includes some of the contacts that are more likely to be real contacts.

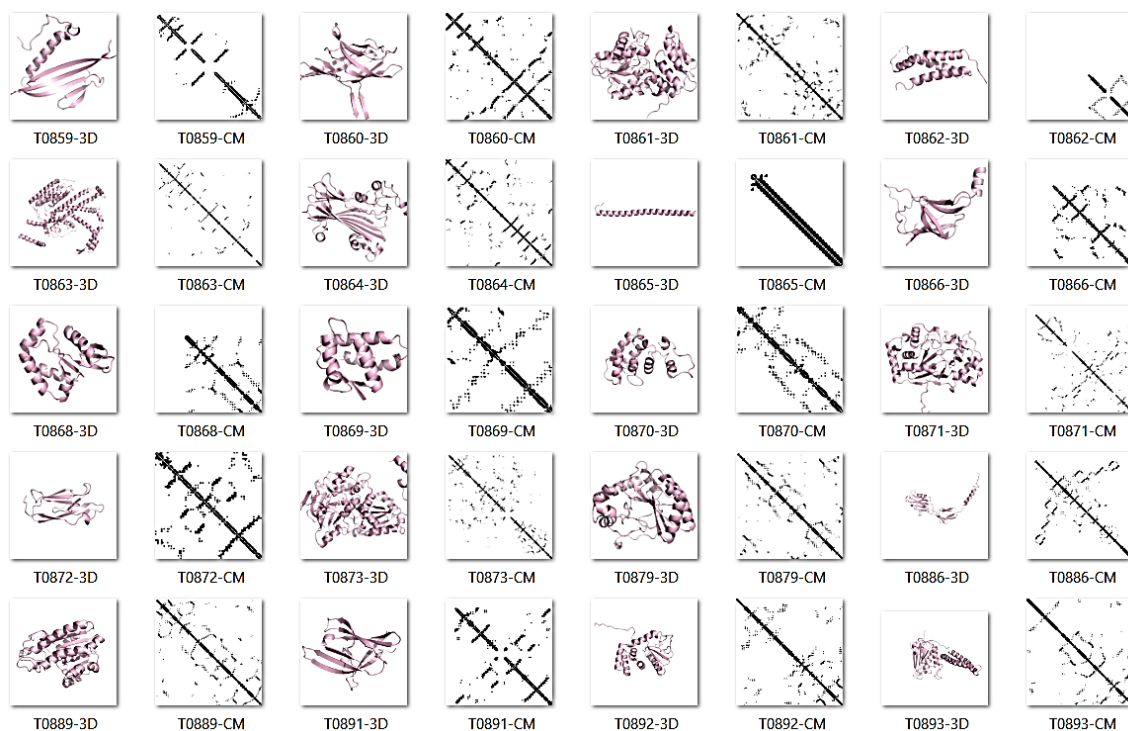


Figure 4. CASP12 PCP targets. First half.

For each target, we display the 3D structure with a cartoon representation followed by its contact map.

2.4 Top Predictors in CASP12

We used prediction results from the official CASP12 results web page (predictioncenter.org/casp12/rrc_avrg_per_target_results.cgi, visited: November 2018) to select the top 10 predictors. From this top 10 we tried to identify the best predictors that would be an appropriate base for our novel predictor. The ranking

score we applied was the average of F1, precision, recall, ES, ES_ext, MCC, recall_SS, MCC_SS y AUC_PR. To calculate the score, we scaled every measure by dividing it by its maximum value. Finally, the average was penalized by the proportion of targets submitted for the competition (for full submissions there is no penalization). In Table 2 we show the top 10 predictors ranked by our score.

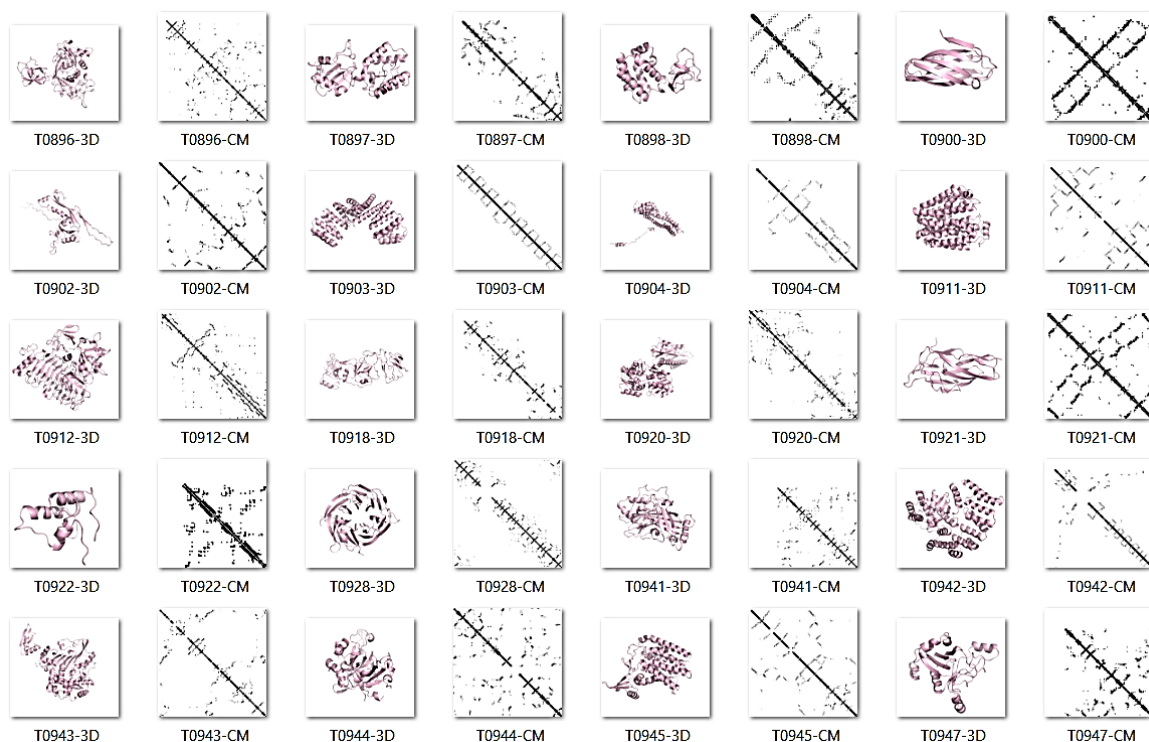


Figure 5. CASP12 PCP targets. Second half.

For each target, we display the 3D structure with a cartoon representation followed by its contact map.

MetaPSICOV

MetaPSICOV (Jones, Singh, Kosciolk, & Tetchner, 2015) combines predictions from three coevolution methods: PSICOV (Jones et al., 2012), mFDCA (Kaján et al., 2014) and CCMpred (Seemayer et al., 2014). MetaPSICOV performed as one of the top predictors by precision in RL top L/5 in CASP11 (Monastyrskyy et al.,

2015). In MetaPSICOV2 the pipeline uses MSAs from the reference databases PDB70, Uniref20, and Uniref100. The first step is to obtain alignments for the protein sequence against PDB70 using HHblits where subsequences with similarity greater than 98% are treated as template-based prediction and subsequences with lower similarity are handled as single domains. In the second step, MetaPSICOV2 tries to build a new alignment against Uniref20, and if the number of hits is below 2.000, a new alignment using JackHMER against Uniref100 is realized, followed of a new run of HHblits on this last result. The following steps are the same as for the previous version: local and global sequence features, coevolution scores, MI scores, and contact potential for columns i and j in the alignment.

Table 2. Top 10 predictors CASP12 by an average of performance measures.

The scores were reported in CASP12 website.

Predictor	Score
MetaPSICOV	0.775
iFold_1	0.767
RaptorX-Contact	0.762
Deepfold-Contact	0.752
MULTICOM-CONSTRUCT	0.751
Pcons-net	0.745
PconsC31	0.743
FALCON_COLORS	0.739
IGBteam	0.735
Naïve	0.726

Local sequence features are calculated in three windows centered at positions i , j and $(i+j)/2$ and include amino acid composition in each window position; secondary structure probabilities; and solvent accessibility. Coevolution scores are

obtained from PSICOV, mfDCA, and CCMpred. Two MI (Mutual Information) scores are used as defined in (Dunn et al., 2008). An averaged contact potential for columns i and j in the alignment is also incorporated. The last local features are associated with sequence separation for amino acids i, j at several ranges of separation. Global sequence features are calculated for the whole sequence and MSAs, including amino acid composition, secondary structure composition, average solvent accessibility, sequence length score, alignment size score, score for number of effective sequences in the MSAs, global average Shannon entropy.

A set of 624 sequences extracted from the PDB with length in the range [50, 500] residues was used as input for training prediction models. Sequences were selected with identity lower than 25%, resolution at 1.5 Å and proteins that had similarity with proteins in PSICOV dataset (Jones et al., 2012) were identified and removed. The PSICOV dataset (150 proteins) and an additional dataset of 434 proteins with resolution at 2.0 Å and sequence length in the range [50, 400] residues were used for testing. In both, training and test datasets, overlapping sequences were excluded using HHsearch (Söding, 2005).

MetaPSICOV uses the training dataset to build a new training dataset with the big set of features described above. This dataset of features (672 features $\times$ 634 proteins) is the input for the first stage of contact prediction. The training of the prediction model for this stage is realized using an architecture based on a feedforward neural network (two hidden layers), using for testing a subset of 10% of the initial training set.

The contact map obtained by the first stage of training is reprocessed in the second stage by extracting an 11×11 window centered at positions i, j . The set of features described for the first stage is recalculated removing the window centered at position $(i+j)/2$ and expanding the windows around positions i and j . This new dataset of features (731 features $\times$ 634 proteins) is treated in the same way as in the first stage creating a new prediction model.

DeepFold_Contact, iFold_1, and naive

The DeepContact Framework (Liu, Palmedo, Ye, Berger, & Peng, 2018) provides the architecture that is used for three predictors in the top 10 in Table 2: iFold_1, DeepFold_Contact and naive. The feature set used by DeepContact is similar to the one used by MetaPSICOV. The differences are that local features are calculated pairwise instead of windows around amino acids i and j and that there is no global score for sequence length. The features are used to create a training dataset from a subset extracted from the ASTRAL reference database (Chandonia, 2004) filtered at 40% of sequence identity. iFold_1, DeepFold_Contact and naive follow the prediction pipeline specified by DeepContact, but there are some differences in parameters and datasets.

For training, DeepContact splits the training dataset into three bins: 80% for the training of prediction models, 10% for testing, and 10% for validation. To ensure data independence, all the sequences that belong to a superfamily are assigned to the same bin. Therefore, sequences in different bins can share only class and fold. The architecture used for training the prediction models implements convolutional

neural networks (CNN). CNN are widely used for image processing, speech recognition, drug design and genetic code analysis (Lecun, Bengio, & Hinton, 2015). In DeepContact, CNN are used in a two-fold way: Processing contact maps converted to images (2D features) and extracting statistical information related to amino acids (1D features).

Deepfold_Contact implements an extra step that identifies domains and obtains MSAs using HHblits and JackHMMER. Naive does not identify domains and trains the prediction model on a separate dataset. iFold_1 generates a first MSA by means of HHBlits and if the MSA's size is less than 1,000 sequences, it generates a larger alignment using JackHMMer.

RaptorX-Contact

RaptorX-Contact (S. Wang et al., 2018) is another predictor in our top 10 (Table 2), it is one of the first predictors to depend on deep learning for training PCP models, and it was already ranked in the top 10 in CASP11 (Monastyrskyy et al., 2015), just as MetaPSICOV. RaptorX-Contact implements an architecture similar to the one defined by DeepContact: one CNN for 1D features and another one for 2D features. The main feature in the design of RaptorX-Contact is considering the contacts as part of a whole (contact map) that are related to each other in contrast to many predictors that perform predictions for each possible amino acid pair in the protein sequence as if they were not part of a bigger structure.

RaptorX-Contact was trained on a PDB25 subset that is obtained by removing any sequence with E-value < 0.1 when compared with the testing dataset. The testing data set included 150 Pfam families, CASP11 proteins, CAMEO proteins and a dataset of nonredundant membrane proteins. The training dataset size was 6300 proteins, and more than 400 proteins were used for testing. The first step of the training process converts 1D sequence features to a 2D representation. This 2D representation is used as input for a CNN which is fed to the 2D CNN. Finally, the previous 2D CNN includes as input 2D predictions for MI and a contact potential derived from CCMpred. For training, the dataset is divided into small bins of proteins of similar size and a few iterations are performed on each bin. The output of the last CNN is the probability for each residue pair to be in contact.

During prediction, RaptorX-Contact generates four MSAs for each target using the reference database Uniprot20 (versions November-2015 and February-2016) and E-values 0.001 and 1. The four predictions (one for each MSA) are averaged to obtain the final prediction. The in-house predictor RaptorX-Property (S. Wang, Li, Liu, & Xu, 2016) predicts local sequence features (secondary-structure, solvent accessibility, and disorder regions) and tries to model interrelationships between adjacent properties and sequence-structure.

MULTICOM-CONSTRUCT

The fifth predictor in our top 10, MULTICOM-CONSTRUCT, is part of a family of predictors that includes MULTICOM-NOVEL and MULTICOM-CLUSTER (Adhikari, Hou, & Cheng, 2018). MULTICOM-NOVEL integrates DNCON contact

predictions (Eickholt & Cheng, 2013) as the core of its pipelines. DNCON implements several sets of deep belief networks (DBN) that train models for several sequence window sizes in the range [7, 19] of residues. DNCON assembles the trained DBN using boosting and the set of features includes only sequence related and local features and excludes correlated mutations information. As in other approaches, DNCON centers windows on residues i and j and computes features such as secondary structure, sequence profile, solvent accessibility and a set of less common features derived from the statistical characterization of residues proposed by Atchley et al. (Atchley, Zhao, Fernandes, & Druke, 2005).

For prediction, MULTICOM-CONSTRUCT generates MSAs using HHblits (reference database Uniprot20) and JackHMMER (reference database Uniref90) iteratively by relaxing parameters until the alignment reaches 2.5L sequences. MULTICOM-CONSTRUCT uses these MSAs as input for CCMpred, PSICOV, and FreeContact and then combines the predictions in the same way as for MetaPSICOV.

PConsC31 and PCons-net

PConsC₃₁ and PCons-net are the next two predictors in our top 10 (Table 2). They are implemented based on PConsC₂ (Skwark, Raimondi, Michel, & Elofsson, 2014) and PConsC₃ (Skwark, Michel, Menéndez Hurtado, Ekeberg, & Elofsson, 2016).

PConsC2 performs sixteen predictions for each input sequence, eight predictions using PSICOV and eight predictions using plmDCA, based on eight MSAs generated by HHblits and JackHMMER. The sixteen DCA predictions are integrated using the concept of neighborhood and that the contacts or non-contacts depend on the frequency of contacts/non-contacts in a 2D window with sizes in the range from 3×3 up to 11×11 . PConsC2 trains prediction models using a deep learning approach of three layers of random forest with one hundred trees and integrates secondary structure prediction, solvent accessibility and sequence profiles using the PSICOV dataset.

PConsC3 uses for training the PSICOV dataset extended with 30 small families. The testing dataset was the same of PConsC2. PConsC3 uses HHblits on Uniprot20 to generate MSAs for DCA using plmDCA and GaussDCA (Baldassi et al., 2014). For local and sequence related features PConsC3 relies on PhyCMAP (Z. Wang & Xu, 2013) that uses MI to estimate the contact probability of an amino acid pair and includes secondary structure and a distance statistical potential. PhyCMAP implements a prediction model supported on random forest and integer linear programming, which is less sensitive to the size of the MSAs. PConsC3 also integrates solvent accessibility and secondary structure using an algorithm analogous to CNN but replacing neural nets by random forest. The individual contact pair predictions are consolidated in a 2D 11×11 window. PConsC3₁ is the same as PConsC3 but using the best MSA generated by PConsC2.

PCons-net is a combination that takes PConsC3 predictions as input for PConsFold, a pipeline for 3D structure prediction that uses the Rosetta folding protocol (Leaver-Fay et al., 2011) which refines a protein structure from short fragments.

FALCON_COLORS

FALCON_COLORS (Zhang et al., 2016) is the eighth predictor in our CASP12 top 10. FALCON_COLORS main difference is that it implements a special method (low-rank and sparse decomposition) to remove indirect coupling from correlation matrices obtained by using MI and mfDCA. FALCON_COLORS estimates parameters for low-rank and sparse decomposition using the PSICOV dataset.

IGBTeam

IGBTeam (Magnan & Baldi, 2014) is our ninth predictor and is another example of the success of deep learning, combining DCA information (CCMpred and FreeContact), in PCP. IGBTeam complements DCA with two in-house predictors for solvent accessibility (ACCpro) and secondary structure (SSpro).

Considering the top 10 predictors described above, we want to emphasize the role of CCMpred and FreeContact for DCA prediction, the use of HHblits and JackHMMER for MSAs, the deep learning related approaches for training models used by every predictor in our CASP12 top 10 except FALCON_COLORS, and finally the importance of secondary structure and solvent accessibility which are used in

almost every predictor. In Chapter IV, section 2.2, we analyze the performance of the ten predictors.

3 Cellular Automata and PCP

A cellular automaton (CA) (Bhattacharjee, Naskar, Roy, & Das, 2018) provides us with a way to represent discrete dynamical models that allows depicting interconnections that affect the state of a system. The discrete space of a CA can be a matrix of finite integer values (states), and the interconnections of the system that it implements are expressed in terms of a window around each cell (neighborhood) in the matrix. CAs were devised by John von Neumann in the 1950s and have been since then widely studied and used in several knowledge domains. Wolfram (Wolfram, 2002) developed an extensive study on CAs dynamics for the simplest kind of CA: a matrix of 1 row $\times$ N columns; two states; neighborhood of radius 1 (size 3). Wolfram's findings are well known, especially his behavior classification for dynamics in CAs.

Bioinformatics and protein related phenomena have several successful cases in which CAs have been used to model dynamic behavior. A full recount of several applications related to genomics and proteomics was made by Chaudhuri et al. in a book that intends to summarize the application of CA models for biomolecules (Chaudhuri, Ghosh, Dutta, & Choudhury, 2018). Although there are no known applications of CA models in PCP, there are some interesting CAs that are used in proteomics. In (Chopra & Bender, 2007) Chopra and Bender designed a CA to

predict secondary structure, using a genetic algorithm (GA). The GA was employed in the design of the CA because the rule that governs the secondary structure classification for a residue requires deep knowledge about how the values of the cell in a window of the sequence determine the secondary structure of each cell. The CA used by Chopra and Bender has a window of radius 5 (size 11) in a matrix of 1 row $\times$ L columns and three states (i.e., helix, beta sheet, and coil). The initial state for the matrix is assigned by DSSP (Kabsch & Sander, 1983) and propensities derived from Chou-Fasman parameters (P. Y. Chou & Fasman, 1974). The GA optimizes the weight associated with each neighbor in the window. In contrast to any other classification model, a CA achieves the final prediction after a successive application of the rule for several iterations, which allows the emergence of global implicit coordination and transference of information for the whole matrix.

Chaudhuri et al. describe in (Chaudhuri, Parimal Pal Soumyabrata, Dutta, & Choudhury, 2018) a CA for modeling protein chains which uses a binary matrix of size 1 row $\times$ 8L columns. Each amino acid is represented by eight cells. For the CA representation of a protein, a set of sixty-four rules can be applied, and several features can be extracted from the evolution (iterative application of a rule) of the CA. The protein representation and the set of rules comprise a framework for analysis and experimentation of sequence manipulation. The analysis of binary patterns that emerge in the CA evolution allows obtaining measures for homogeneity, correlation, contrast, and entropy. The details of the CAs used by Chaudhuri et al. are described in (Ghosh, Maiti, & Chaudhuri, 2014).

Neural-CA (Varela & Santos, 2018) is a very interesting approach that merges CAs and neural networks. Varela and Santos use a matrix 2D where each amino acid occupies the center of the edges in a unit cubic so that every amino acid has twelve neighbors. The neural-CA defines its states according to the hydrophobic polar (HP) model and uses evolutionary algorithms to optimize the neural-CA parameters. The neural-CA rule considers the energy landscape instead of the local spatial relationships in traditional CAs. The main aim of this approach is to create a framework for protein folding. However, the published tests only train models for specific targets, and their generalization capability is undetermined.

Diao et al. describe an approach to predict the topology of transmembrane proteins in (Diao et al., 2008). This approach combines protein pseudo amino acid composition (K. C. Chou, 2011) with CAs. The representation of amino acids uses a binary chain of five bits and the CA's matrix size is 1 row $\times$ 5L columns. This approach predicts the segment class in three categories (α -helix, β -strand, and non-trans membrane). The prediction process involves evolving the CA for twenty-six iterations to obtain a matrix of 26 rows $\times$ 5L columns (1 initial row plus 25 evolved rows). Lempel-Ziv complexity and pseudo amino acid composition are used to characterize the class for each segment.

We hypothesize that CAs are able to evolve PCP from predictors such as CCMpred, FreeContact, and MetaPSICOV obtaining predictions with a less FP proportion. The idea is that a contact map is analogous to a 2D lattice of a CA with binary states and that we can learn the window (neighborhood) that defines

interaction for contacts in the contact map and that the CA rule can evolve an initial contact map, predicted for CCMpred or any other predictor, towards a more native-like contact map. To identify the CA's neighborhood and rule, we have developed previously a GA (Díaz & Tischer, 2016) which we tested identifying a CA to reproduce a contact map folding trajectory from protein 2f4k. In Figure 6 we present a high-level representation of the pipeline for training and PCP based on CAs.

4 Conclusions

We examined CASP12 results and selected a top 10 of predictors that were reviewed. The set of predictors that we selected represents the state of the art in PCP; we observe that every predictor in Table 2 shares these characteristics:

1. MSAs are generated using HHblits and JackHMMER adjusting parameters and selecting reference databases to optimize results.
2. Correlated mutations are fundamental for every predictor, though are used in diverse ways in each prediction pipeline. CCMpred and FreeContact are the main tools that predictors use to obtain MI or coevolution predictions.
3. Secondary structure, solvent accessibility, and sequence profiles provide complementary information to enhance PCP.
4. Sequence windows or contact map 2D windows allow predictors to consolidate profiles of features and to integrate patterns associated with contact identification.

- Machine learning is the core for PCP models training. Deep learning and neural nets are the most popular approaches that predictors use in their training process.

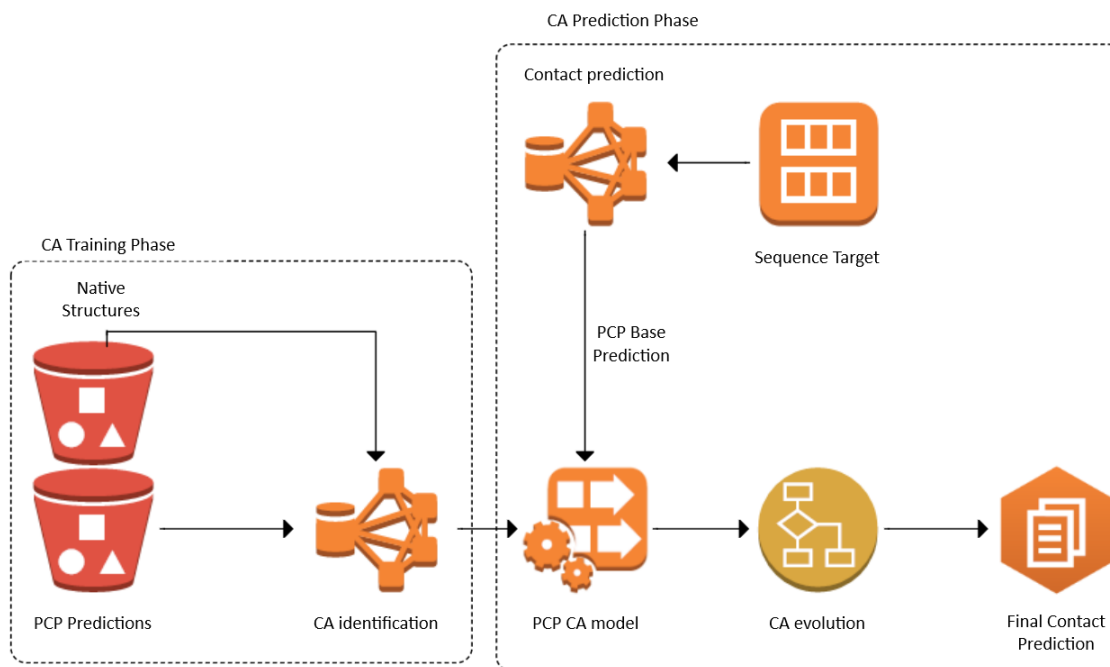


Figure 6. Proposed pipeline for PCP training and CA-based PCP prediction.

Structure modeling can be improved by using contact predictions. However, the current results require to improve the fraction of FP to allow modelers to enrich their result with information about real contacts.

5 References

Adhikari, B., Hou, J., & Cheng, J. (2018). Protein contact prediction by integrating deep multiple sequence alignments, coevolution and machine learning. *Proteins: Structure, Function and Bioinformatics*, 86(June 2017), 84–96. <https://doi.org/10.1002/prot.25405>

- Adhikari, B., Nowotny, J., Bhattacharya, D., Hou, J., & Cheng, J. (2016). ConEVA: a toolbox for comprehensive assessment of protein contacts. *BMC Bioinformatics*, 17(1), 517. <https://doi.org/10.1186/s12859-016-1404-z>
- Altschul, S., Thomas, M., Alejandro, S., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389–3402.
- Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S. E., Hubbard, T. J. P., Chothia, C., & Murzin, A. G. (2008). Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Research*, 36(SUPPL. 1), 419–425. <https://doi.org/10.1093/nar/gkm993>
- Atchley, W. R., Zhao, J., Fernandes, A. D., & Druke, T. (2005). Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences*, 102(18), 6395–6400. <https://doi.org/10.1073/pnas.0408677102>
- Baldassi, C., Zamparo, M., Feinauer, C., Procaccini, A., Zecchina, R., Weigt, M., & Pagnani, A. (2014). Fast and accurate multivariate Gaussian modeling of protein families: Predicting residue contacts and protein-interaction partners. *PLoS ONE*, 9(3), 1–12. <https://doi.org/10.1371/journal.pone.0092721>
- Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., ... Zhang, J. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, 45(D1), D158–D169. <https://doi.org/10.1093/nar/gkw1099>
- Bhattacharjee, K., Naskar, N., Roy, S., & Das, S. (2018). A survey of cellular automata: types, dynamics, non-uniformity and applications. *Natural Computing*, 8, 1–29. <https://doi.org/10.1007/s11047-018-9696-8>
- Buchan, D. W. A., & Jones, D. T. (2018). Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins: Structure, Function and Bioinformatics*. <https://doi.org/10.1002/prot.25379>
- Chandonia, J.-M. (2004). The ASTRAL Compendium in 2004. *Nucleic Acids Research*, 32(90001), 189D–192. <https://doi.org/10.1093/nar/gkho34>
- Chaudhuri, Parimal Pal Soumyabrata, G., Dutta, A., & Choudhury, S. P. (2018). Cellular Automata (CA) Model for Protein. In *A New Kind of Computational Biology* (pp. 291–325). Singapore: Springer.
- Chaudhuri, P. P., Ghosh, S., Dutta, A., & Choudhury, S. P. (2018). *A New Kind of Computational Biology*. Singapore: Springer. <https://doi.org/10.1007/978-981-13-1639-5>

- Chopra, P., & Bender, A. B. T.-I. S. B. (2007). Evolved cellular automata for protein secondary structure prediction imitate the determinants for folding observed in nature. *In Silico Biology*, 7(1), 87–93.
- Chou, K. C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of Theoretical Biology*, 273(1), 236–247. <https://doi.org/10.1016/j.jtbi.2010.12.024>
- Chou, P. Y., & Fasman, G. D. (1974). Prediction of Protein Conformation. *Biochemistry*, 13(2), 222–245. <https://doi.org/10.1021/bio0699a002>
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1), 1–30. <https://doi.org/10.1016/j.jecp.2010.03.005>
- Diao, Y., Ma, D., Wen, Z., Yin, J., Xiang, J., & Li, M. B. T.-A. A. (2008). Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel-Ziv complexity. *Amino Acids*, 34(1), 111–117. <https://doi.org/10.1007/s00726-007-0550-z>
- Díaz, N., & Tischer, I. (2016). Generic framework for mining cellular automata models on protein-folding simulations. *Genetics and Molecular Research*, 15(2), 1–16. <https://doi.org/10.4238/gmr.15028654>
- Dunn, S. D., Wahl, L. M., & Gloor, G. B. (2008). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3), 333–340. <https://doi.org/10.1093/bioinformatics/btm604>
- Eickholt, J., & Cheng, J. (2013). A study and benchmark of DNcon: a method for protein residue-residue contact prediction using deep networks. *BMC Bioinformatics*, 14(Suppl 14), 1–8. <https://doi.org/10.1186/1471-2105-14-S14-S12>
- Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., & Aurell, E. (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Physical Review E*, 87(1), 1–19. Retrieved from <http://pre.aps.org/abstract/PRE/v87/i1/e012707>
- Ghosh, S., Maiti, N. S., & Chaudhuri, P. P. (2014). Cellular automata model for protein structure synthesis (PSS). *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8751, 268–277. https://doi.org/10.1007/978-3-319-11520-7_28

- Göbel, U., Sander, C., Schneider, R., & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Genetics*, 18(4), 309–317. <https://doi.org/10.1002/prot.340180402>
- Gouveia-Oliveira, R., & Pedersen, A. G. (2007). Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. *Algorithms for Molecular Biology*, 2(1), 1–12. <https://doi.org/10.1186/1748-7188-2-12>
- Gromiha, M. M., & Selvaraj, S. (2004). Inter-residue interactions in protein folding and stability. *Progress in Biophysics and Molecular Biology*, 86(2), 235–277. <https://doi.org/10.1016/j.pbiomolbio.2003.09.003>
- Holland, J., Pan, Q., & Grigoryan, G. (2018). Contact prediction is hardest for the most informative contacts, but improves with the incorporation of contact potentials. *PLoS ONE*, 13(6), 1–20. <https://doi.org/10.1371/journal.pone.0199585>
- Jones, D. T., Buchan, D. W. a, Cozzetto, D., & Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics (Oxford, England)*, 28(2), 184–190. <https://doi.org/10.1093/bioinformatics/btr638>
- Jones, D. T., Singh, T., Kosciolk, T., & Tetchner, S. (2015). MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, 31(7), 999–1006. <https://doi.org/10.1093/bioinformatics/btu791>
- Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12), 2577–2637. <https://doi.org/10.1002/bip.360221211>
- Kaján, L., Hopf, T. a, Kalaš, M., Marks, D. S., & Rost, B. (2014). FreeContact: Fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics*, 15(85), 1–6. <https://doi.org/10.1186/1471-2105-15-85>
- Kamisetty, H., Ovchinnikov, S., & Baker, D. (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences of the United States of America*, 110(39), 15674–15679. <https://doi.org/10.1073/pnas.1314045110>
- Kim, D., DiMaio, F., Wang, R. Y.-R., Song, Y., & Baker, D. (2014). One contact for every twelve residues allows robust and accurate topology level protein

- structure modeling. *Proteins: Structure, Function, and Bioinformatics*, 82 Suppl 2(April), 208–218. <https://doi.org/10.1002/prot.24374>
- Kryshtafovych, A., Monastyrskyy, B., & Fidelis, K. (2016). CASP11 statistics and the prediction center evaluation system. *Proteins: Structure, Function, and Bioinformatics*, 84(Suppl 1), 15–19. <https://doi.org/10.1002/prot.25005>
- Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., ... Bradley, P. (2011). Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology*, 487, 545–574. <https://doi.org/10.1016/B978-0-12-381270-4.00019-6>
- Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Liu, Y., Palmedo, P., Ye, Q., Berger, B., & Peng, J. (2018). Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. *Cell Systems*, 6(1), 65–74.e3. <https://doi.org/10.1016/j.cels.2017.11.014>
- Magnan, C. N., & Baldi, P. (2014). SSpro/ACCpro 5: Almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics*, 30(18), 2592–2597. <https://doi.org/10.1093/bioinformatics/btu352>
- Martin, L. C., Gloor, G. B., Dunn, S. D., & Wahl, L. M. (2005). Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, 21(22), 4116–4124. <https://doi.org/10.1093/bioinformatics/bti671>
- Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3), 1436–1462. <https://doi.org/10.1214/0090536060000000281>
- Mier, P., & Andrade-Navarro, M. A. (2016). FastaHerder2: Four Ways to Research Protein Function and Evolution with Clustering and Clustered Databases. *Journal of Computational Biology*, 23(00), 270–278. <https://doi.org/10.1089/cmb.2015.0191>
- Monastyrskyy, B., D’Andrea, D., Fidelis, K., Tramontano, A., & Kryshtafovych, A. (2014). Evaluation of residue-residue contact prediction in CASP10. *Proteins: Structure, Function and Bioinformatics*, 82(SUPPL.2), 138–153. <https://doi.org/10.1002/prot.24340>
- Monastyrskyy, B., D’Andrea, D., Fidelis, K., Tramontano, A., & Kryshtafovych, A. (2015). New encouraging developments in contact prediction: Assessment of the CASP11 results. *Proteins*, (October), 1–14. <https://doi.org/10.1002/prot.24943>

- Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., ... Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49), E1293–E1301. <https://doi.org/10.1073/pnas.111471108>
- Schaarschmidt, J., Monastyrskyy, B., Kryshchak, A., & Bonvin, A. M. J. J. (2017). Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. *Proteins: Structure, Function, and Bioinformatics*, (October 2017), 51–66. <https://doi.org/10.1002/prot.25407>
- Seemayer, S., Gruber, M., & Söding, J. (2014). CCMpred - Fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics*, 30(21), 3128–3130. <https://doi.org/10.1093/bioinformatics/btu500>
- Skwark, M. J., Michel, M., Menéndez Hurtado, D., Ekeberg, M., & Elofsson, A. (2016). Accurate contact predictions for thousands of protein families using PconsC3. *BioRxiv Preprint*, 1–18. <https://doi.org/http://dx.doi.org/10.1101/079673>
- Skwark, M. J., Raimondi, D., Michel, M., & Elofsson, A. (2014). Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns. *PLoS Computational Biology*, 10(11). <https://doi.org/10.1371/journal.pcbi.1003889>
- Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21, 951–960. <https://doi.org/10.1093/bioinformatics/bti125>
- Sun, H., Huang, Y., Wang, X., Zhang, Y., & Shen, H.-B. (2015). Improving accuracy of protein contact prediction using balanced network deconvolution. *Proteins: Structure, Function and Bioinformatics*, 83(3), 485–496. <https://doi.org/doi:10.1002/prot.24744>
- Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., & Wu, C. H. (2015). UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6), 926–932. <https://doi.org/10.1093/bioinformatics/btu739>
- Tanaka, S., & Scheraga, H. a. (1975). Model of protein folding: inclusion of short-, medium-, and long-range interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 72(10), 3802–3806. <https://doi.org/Doi 10.1073/Pnas.72.10.3802>
- Varela, D., & Santos, J. (2018). Automatically obtaining a cellular automaton scheme for modeling protein folding using the FCC model. *Natural*

- Computing, 0123456789(Levinthal 1968). <https://doi.org/10.1007/s11047-018-9705-y>
- Wang, S., Li, W., Liu, S., & Xu, J. (2016). RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Research*, 44(W1), W430–W435. <https://doi.org/10.1093/nar/gkw306>
- Wang, S., Sun, S., & Xu, J. (2018). Analysis of deep learning methods for blind protein contact prediction in CASP12. *Proteins: Structure, Function and Bioinformatics*. <https://doi.org/10.1002/prot.25377>
- Wang, Z., & Xu, J. (2013). Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics*, 29(13), 266–273. <https://doi.org/10.1093/bioinformatics/btt211>
- Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., & Hwa, T. (2009). Identification of direct residue contacts in protein–protein interaction by message passing. *PNAS*, 106(1), 67–72. <https://doi.org/https://doi.org/10.1073/pnas.0805923106>
- Wolfram, S. (2002). *A New Kind of Science* (Vol. 1). Champaign, IL: Wolfram Media Inc.
- Zhang, H., Gao, Y., Deng, M., Wang, C., Zhu, J., Li, S. C., ... Bu, D. (2016). Improving residue-residue contact prediction via low-rank and sparse decomposition of residue correlation matrix. *Biochemical and Biophysical Research Communications*, 472(1), 217–222. <https://doi.org/10.1016/j.bbrc.2016.01.188>

Chapter III.

GACAI-PCP: Cellular Automata based Tool for Contact Map Prediction

Abstract

In this chapter we describe the software tool that we developed to build contact map predictors. This novel approach exploits the capabilities of cellular automata to offer complex behavior from local interactions only. Our tool identifies cellular automata able to classify predicted contacts that are more likely real contacts. The core of our cellular automata identification tool is a genetic algorithm that optimizes the prediction of balanced contact maps. With our tool, we proved that is possible to improve contact map prediction by means of cellular automata models.

1. Introduction

In Chapter II we lined out that cellular automata (CAs), are used as a tool to improve protein contact maps prediction (CMP). In this chapter we describe the tools that we developed to help us in CA identification for CMP.

CMP and protein contact prediction (PCP) fields arose in 1970's (Tanaka & Scheraga, 1975). Nowadays, the field has shown a great development particularly since the inception of the analysis of correlated mutations (Monastyrskyy,

D'Andrea, Fidelis, Tramontano, & Kryshtafovych, 2014). Though the first PCP tools implementing correlated mutations were developed more than a decade ago, the issue of false correlations that dampened the success of this kind of tools were solved just recently (Monastyrskyy, D'Andrea, Fidelis, Tramontano, & Kryshtafovych, 2015).

Protein contact prediction (PCP) has acquired importance because of its helpfulness in template-free protein structure prediction (Adhikari & Cheng, 2016). PCP provides spatial constraints derived from the protein chain that can be used in tertiary structure reconstruction or in pipelines that predict more detailed native structures (Zhang, Mortuza, He, Wang, & Zhang, 2018). PCP has been a tool used since the 1970s when Tanaka and Scheraga used protein contacts in an approach for protein folding (Tanaka & Scheraga, 1975). After several decades of advancement, PCP has taken a prominent place in protein folding and protein structure prediction, especially for proteins that have few homologs (Moult, Fidelis, Kryshtafovych, Schwede, & Tramontano, 2018). Despite the current progress, there is room for PCP improvement in the way of enhancing the contacts in the context of the overall protein structure and the contact map that represents the protein. In Figure 1, we show an example of a predicted protein contact map and a real contact map. RaptorX-Contact (Wang, Sun, Li, Zhang, & Xu, 2017) was the tool used to predict the contact map in Figure 1. The protein used for the Figure 1 is part of the benchmark data set of the biennial Critical Assessment of Protein Structure Prediction (CASP) for year 2016 (CASP12). The predicted contact map in Figure 1

includes a high proportion of all the contacts in real contact maps (true positives or TP), but the proportion of false contacts (false positives or FP) is bigger.

We propose a software tool to identify CAs able to transform a PCP to a contact map closer to the real one. Researchers rely on approaches for CA identification when the knowledge about the inherent mechanisms that define transitions is insufficient (Bolt, Baetens, & De Baets, 2015). In the case of PCP, many prediction tools contesting in the CASP use the idea that local effects govern the presence of contacts for amino acid pairs. Usually this concept is implemented by defining a window and the prediction tools analyzes information from MSAs and sequence related features (i.e., solvent accessibility, secondary structure). This window can be seen as a specific neighborhood in a CA, but there is little evidence supporting that it is the optimal neighborhood for determining residue-residue contacts.

In this chapter, we describe the methods we used for CA identification for PCP. In section 2 we describe the databases, datasets, and tools used in our CA identification process. In section 3, we compare a CA identified by our approach with CAs defined by commonly used neighborhoods. Finally, in section 4 we state our conclusions about our methods for CA identification for PCP.



Figure 1. RaptorX-Contact predicted contact map and actual contact map for target protein T0900 (CASP12).

Upper triangular: RaptorX-Contact predicted contacts. Lower triangular: Actual contact map. TP denotes true positives (true contacts predicted); FP indicates false contacts predicted (false positives); FN (false negatives) represents actual contacts that were not predicted.

2. Materials and Methods

2.1. Initial Conditions Dataset

We require several protein reference databases to build MSAs from which we obtain initial PCPs using CCMpred. The reference databases that we used were Uniprot20 (Bateman et al., 2017) and Uniref100 (Suzek, Wang, Huang, McGarvey, & Wu, 2015). To generate MSAs on Uniprot20 we used HHblits (Remmert, Biegert, Hauser, & Söding, 2012) and JackHHMER (Johnson, Eddy, & Portugaly, 2010) on

Uniref100. In Figure 2 we show the pipeline we defined to build a dataset of contact maps. These contact maps are used as initial conditions to extract rules for CAs. The phase of MSAs generation ends with the best MSA to be used as input for CCMpred.

We build the dataset of initial conditions from the 150 proteins dataset described in (Jones, Buchan, Cozzetto, & Pontil, 2012). This set of proteins contains 150 biological macromolecules, with lengths in the range [50, 275] amino acids, with high resolution (≤ 1.9 Å), and unique Pfam domains (El-Gebali et al., 2018).

CCMpred takes as input the optimal MSA for each sequence in the training dataset (Figure 2) and returns a matrix that predicts the coupling scores for each pair of amino acids in the sequence. We use ConKit (Simkovic, Thomas, & Rigden, 2017) to extract a contact from the CCMpred coupling scores matrix.

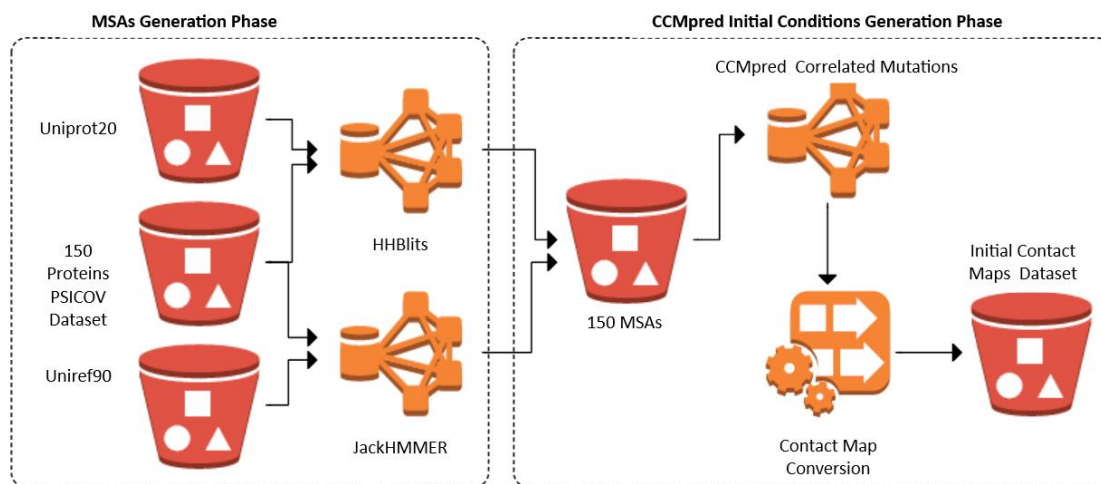


Figure 2. Pipeline for the generation of the dataset of initial conditions.

2.2. Cellular Automata Identification Framework

We used the architectural framework described in (Díaz & Tischer, 2016) to implement a genetic algorithm that identifies CAs that evolve predicted contact maps. This framework provides a core architecture that eases the process of implementation of several kinds of algorithms for CAs identification. By using this framework, we can put the focus on the details of the CAs that we want to obtain, because many tools common in CAs identification are ready to use.

2.3. Genetic Algorithm for Identification of CAs that evolve Contact Maps

CAs identification is an optimization task. In our case we require an algorithm that searches for an optimal CA that improves a predicted contact map by reducing the proportion of false contacts while keeping real contacts. Our approach implements a genetic algorithm (GA). The search space for our GA includes 5.63×10^{14} possible CAs. The size of the search space depends on the size of the maximum neighborhood, which we confined to a 7×7 matrix of around each cell in the lattice.

In Figure 3 we show a high-level description of the steps that implements our GA. The first step is to arrange the training dataset (pairs of predicted/real contact map) in data strata, which allows us to reduce the computational cost for an iteration of the GA. By splitting the data, we can use one stratum in each GA iteration, avoiding model overfitting and increasing the diversity in the set of CA transitions. In each iteration, one half of the contact maps are used for transitions

identification and the other half for testing. When the dataset has been consumed by the GA, it continues using the strata again in the same order.

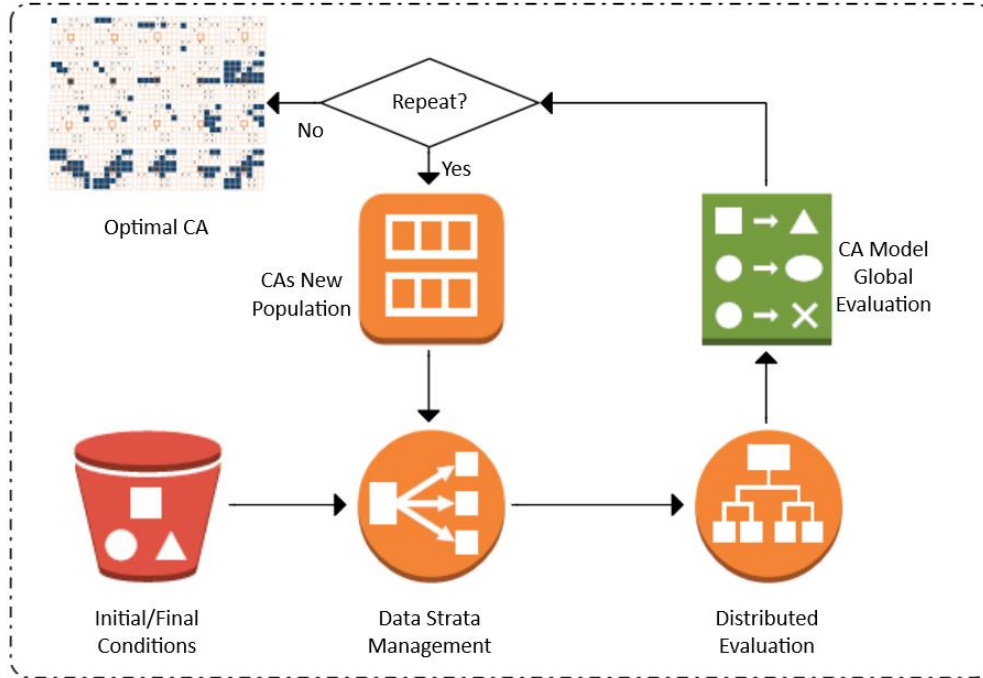


Figure 3. Genetic Algorithm Process for CA identification.

For each iteration, the GA creates a new population of one hundred CAs. For the first iteration the population is generated in a random way, and for the following iterations the new population is generated by GA operators of selection and searching. Each CA in the population is evaluated in parallel, exploiting the high-level library DASK (Rocklin, 2015) that allows to run parallel processes in modern computing clusters. Once each CA is evaluated in a parallel and independent process, the master node retrieves all the results (CAs and individual scores). Then the global evaluation arranges the best individuals to be used in the next iteration.

A new iteration is executed until the stop condition is met, i.e., two-thousand iterations are done, or a perfect CA is found.

For quality evaluation of each CA we used the Matthews Correlation Coefficient (Monastyrskyy et al., 2014), which is insensitive to class bias. PCP is a highly biased problem, because the proportion of non-contacts is very high, so that measures that are sensitive to class bias could search for models that prefers to predict non-contacts in most of the cases.

We named our algorithm GACAI-PCP as an acronym for Genetic Algorithm for Cellular Automata Identification for Protein Contact Prediction.

2.4. Density Classification Task as Benchmark for GACAI-PCP

GACAI-PCP as described in the above section, was adapted from our solution for the density classification task (DCT). GACAI-DCT was the solution we proposed for DCT (Diaz & Tischer, 2019). We used DCT as a benchmark for our approach to DCT, because the two problems are similar in some respects. DCT is a theoretical problem that requires the identification of CAs that can evolve an initial condition that has majority of a class (0 or 1) to a final condition where all the cells in the lattice have the state of the majority class. DCT requires that the CA implements mechanisms of global coordination expressed just by the local transitions included in the CA rule. CA global coordination is a property that emerges when the CA rule is applied to an initial condition iteratively, and it is hard to design manually. For the 2D DCT case, we have a problem similar to PCP: 1) The initial and final

conditions are known; 2) The optimal neighborhood is unknown; 3) The rule transitions can be deterministic or stochastic (there are no explicit restriction about this); 4) The lattice can be a regular $n \times n$ 2D binary matrix; 5) the border condition can be cyclic.

In DCT we can use small lattices; datasets with lattices of 21×21 initial conditions are very common, so that less processing is required for parameters tuning. In contrast, in PCP an average contact map can easily have a size of 300×300 .

In GACAI-PCP we used the same algorithm as for GACAI-DCT and the only important difference is that the MCC used in GACAI-PCP is adjusted to evaluate only the upper triangular of the contact map to reduce the processing time.

3. Discussion

The GACAI-PCP search space includes CA's neighborhoods that are common in CAs modelling. Perhaps the most widely used neighborhood is the Moore's neighborhood of radius one (Figure 4.a). We compare some obvious CAs using the Moore's neighborhood of radius one, two and three (Figure 4.a-c), as well as a random neighborhood (Figure 4.d) against a CA evolved by GACAI-PCP.

For evaluation we use the targets in the CASP12 dataset, which is comprised of 39 proteins (predictioncenter.org). In Figure 5, we show the precision of the five CAs defined by the neighborhoods of the Figure 4, measured for the full list of predicted

contacts. The model identified by GACAI-PCP outperforms every other CA. The only target protein where our model was outperformed was To862. The CAs that use Moore's neighborhood of radius one and two, and the CA with random neighborhood show a similar performance with precision in the range $[0.0, 0.1]$. The CA of radius three is the only model in the comparison that surpasses the threshold precision at 0.1, for twelve out of the 39 proteins in the evaluation dataset. But, in the overall comparison is evident that GACAI-PCP obtains better contact predictors.

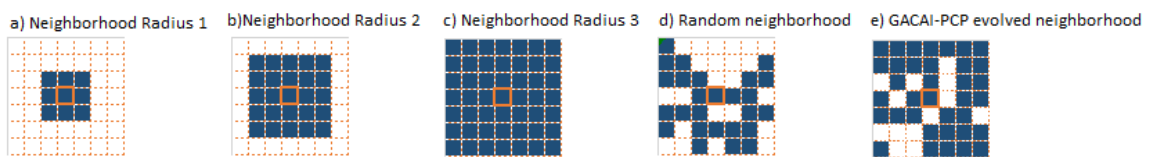


Figure 4. Examples of Moore's CA neighborhood.

a) Radius one Moore's CA neighborhood. b) Radius two Moore's CA neighborhood. c) Radius three Moore's CA neighborhood. d) Random neighborhood. e) GACAI-PCP evolved neighborhood. Blue cells indicate cells that affect CA's transitions. The cell with red border is the one that is updated by the CA evolution. White cells have no effect in the CA transitions.

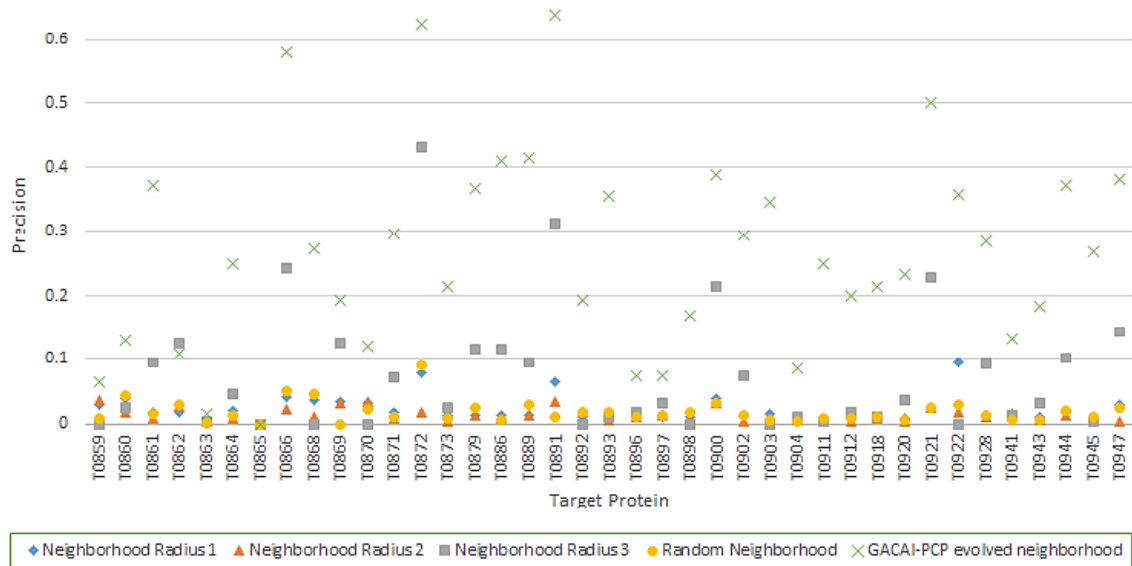


Figure5. Performance of the five CAs on the CASP12 evaluation dataset.

In Figure 5, target protein To862 is an exceptional case where our identified CA is outperformed by the CA with Moore's neighborhood radius three. In contrast, for eight target proteins, the CA with neighborhood radius three obtains the worst prediction result. In Figure 6, we illustrate the differences of the five CAs in the comparison. To assess the performance of the predictors, we used Friedman's test and Nemenyi's post-hoc test. Friedman's test compares the precision achieved for each predictor in each protein target and determines its average ranking. If the differences in rankings are significant, Friedman's test reports a small p-value and rejects the null hypothesis (there is no difference in performance). If the null hypothesis is rejected, it is necessary to perform a post-hoc test to find the predictors that perform better than others.

Nemenyi's post-hoc test compares all predictor pairs and allows us to identify those with significant difference in performance. For this comparison (39 targets and 5 predictors), by Nemenyi's test we conclude that predictors with rank differences greater than the critical difference ($CD \geq 0.9768$), are significantly different and the one with the best ranking (lowest value) is the dominant in the set of target proteins. Figure 6 shows the dominance graph of the predictors based on CAs with different neighborhoods (in Chapter IV, section 2.2 we use the same methodology to compare CASP12's top10 predictors). Our evolved CA dominates all others, which allows to conclude that our approach gets better CAs than those defined with traditional neighborhoods, because its irregular neighborhood shape. The radius three CA surpasses only the radius two CA.

Friedman p-value=4.3288E-11; Critical Difference (CD)=0.9768; $\alpha=0.05$

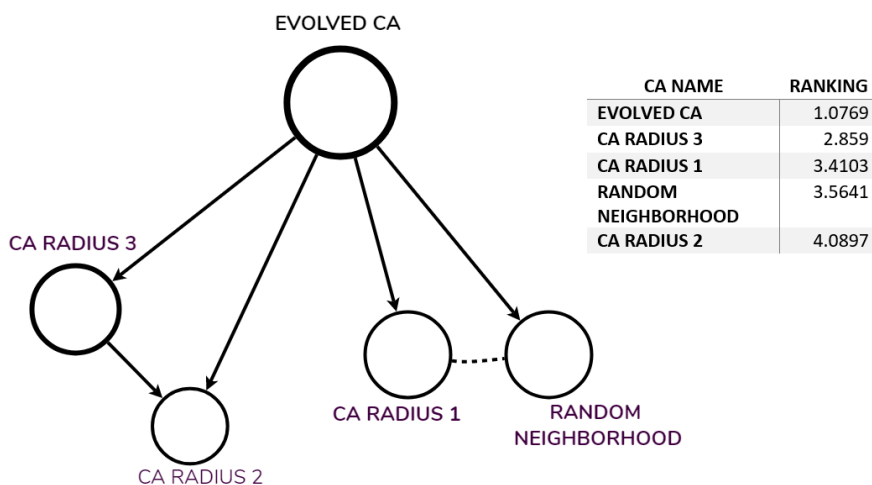


Figure 6. Dominance graph analysis with statistical significance comparison by Friedman's test and post-hoc Nemenyi's test.

4. Conclusions

We proposed a framework based on GAs to identify CAs we used to improve protein contact maps prediction. GACAI-PCP identified a CA that exceeded three obvious CAs and a random CA, providing evidence that our approach identifies optimal CAs for PCP.

We tuned the GA parameters using DCT as a problem that shares similar specifications to PCP. This allowed us to try several configurations and design options in a computationally less expensive setting.

GACAI-PCP evolves the optimal neighborhood shape based on the CA rule performance. The GA's evolutive behavior resulted in irregular neighborhood shapes (i.e. Figure 4.e)

5. References

- Adhikari, B., & Cheng, J. (2016). Protein Residue Contacts and Prediction Methods. In O. Carugo & F. Eisenhaber (Eds.), *Data Mining Techniques for the Life Sciences* (Vol. 1415, pp. 463–476). Humana Press, New York, NY.
<https://doi.org/10.1007/978-1-4939-3572-7>
- Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., ... Zhang, J. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, 45(D1), D158–D169. <https://doi.org/10.1093/nar/gkw1099>
- Bolt, W., Baetens, J. M., & De Baets, B. (2015). An evolutionary approach to the identification of Cellular Automata based on partial observations. 2015 IEEE Congress on Evolutionary Computation, CEC 2015 - Proceedings, 2966–2972. <https://doi.org/10.1109/CEC.2015.7257258>
- Diaz, N., & Tischer, I. (2019). Mining Stochastic Cellular Automata to Solve Density Classification Task in Two Dimensions.
- Díaz, N., & Tischer, I. (2016). Generic framework for mining cellular automata models on protein-folding simulations. *Genetics and Molecular Research*, 15(2), 1–16. <https://doi.org/10.4238/gmr.15028654>
- El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., ... Finn, R. D. (2018). The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(October 2018), 427–432. <https://doi.org/10.1093/nar/gky995>
- Johnson, L. S., Eddy, S. R., & Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 11(431), 1–8. <https://doi.org/10.1186/1471-2105-11-431>
- Jones, D. T., Buchan, D. W. a, Cozzetto, D., & Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics (Oxford, England)*, 28(2), 184–190. <https://doi.org/10.1093/bioinformatics/btr638>
- Monastyrskyy, B., D'Andrea, D., Fidelis, K., Tramontano, A., & Kryshtafovych, A. (2014). Evaluation of residue-residue contact prediction in CASP10. *Proteins: Structure, Function and Bioinformatics*, 82(SUPPL.2), 138–153. <https://doi.org/10.1002/prot.24340>
- Monastyrskyy, B., D'Andrea, D., Fidelis, K., Tramontano, A., & Kryshtafovych, A. (2015). New encouraging developments in contact prediction: Assessment of the CASP11 results. *Proteins*, (October), 1–14. <https://doi.org/10.1002/prot.24943>

- Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2018). Critical assessment of methods of protein structure prediction (CASP)—Round XII. *Proteins: Structure, Function and Bioinformatics*, 86, 7–15. <https://doi.org/10.1002/prot.25415>
- Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2), 173–175. <https://doi.org/10.1038/nmeth.1818>
- Rocklin, M. (2015). Dask: Parallel Computation with Blocked algorithms and Task Scheduling. *SciPy, (Scipy)*, 130–136.
- Simkovic, F., Thomas, J. M. H., & Rigden, D. J. (2017). ConKit: A python interface to contact predictions. *Bioinformatics*, 33(14), 2209–2211. <https://doi.org/10.1093/bioinformatics/btx148>
- Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., & Wu, C. H. (2015). UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6), 926–932. <https://doi.org/10.1093/bioinformatics/btu739>
- Tanaka, S., & Scheraga, H. a. (1975). Model of protein folding: inclusion of short-, medium-, and long-range interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 72(10), 3802–3806. <https://doi.org/Doi 10.1073/Pnas.72.10.3802>
- Wang, S., Sun, S., Li, Z., Zhang, R., & Xu, J. (2017). Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Computational Biology*, 13(1), 1–34. <https://doi.org/10.1371/journal.pcbi.1005324>
- Zhang, C., Mortuza, S. M., He, B., Wang, Y., & Zhang, Y. (2018). Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. *Proteins: Structure, Function and Bioinformatics*, 86, 136–151. <https://doi.org/10.1002/prot.25414>

Chapter IV.

Contact Map Prediction based on Cellular Automata

Abstract

Contact map prediction has become an important step in ab-initio protein structure prediction. However, despite the remarkable improvements that many researchers have achieved with many contact prediction tools, the predicted contact maps still include a high proportion of false contacts that hinder the practical use of these predictions. In this chapter, we describe our novel approach based on cellular automata that allows us to improve predicted contact maps by removing false contacts predicted by renowned predictors such as MetaPSICOV2. We compare our results with the public dataset of target proteins used for the Critical Assessment of Protein Structure Prediction (year 2016, usually referred as CASP12). To study the quality of the predicted contact maps we use statistical analysis to find the dominance of our approach in the top 10 of predictors. The results indicate that our approach is successful when we use the full predicted contact map but that the quality decreases when the evaluation is made with reduced lists of contacts.

1. Introduction

Protein contact prediction (PCP) has acquired importance because of its helpfulness in template-free protein structure prediction (Adhikari & Cheng, 2016). PCP provides spatial constraints derived from the protein chain that can be used in tertiary structure reconstruction or in pipelines that predict more detailed native structures (C. Zhang, Mortuza, He, Wang, & Zhang, 2018). PCP has been a tool used since the 1970s when Tanaka and Scheraga used protein contacts in an approach for protein folding (Tanaka & Scheraga, 1975). After several decades of advancement, PCP has taken a prominent place in protein folding and protein structure prediction, especially for proteins that have few homologs (Moult, Fidelis, Kryshtafovych, Schwede, & Tramontano, 2018).

The use of correlated mutations analysis from multi-sequence alignments (MSAs) has allowed an important improvement in PCP mainly since the incorporation of artificial correlations filtering (Monastyrskyy, D'Andrea, Fidelis, Tramontano, & Kryshtafovych, 2015). In PCP, two core methods are used for correlated mutations analysis: direct coupling analysis (DCA) (Morcos et al., 2011) and mutual information (MI) (Martin, Gloor, Dunn, & Wahl, 2005). MI approaches predict which amino acids are in contact using the identity of each pair in the protein sequence. In DCA there three main branches: mean field DCA (mfDCA (Morcos et al., 2011)), pseudo-likelihood maximization (plmDCA (Ekeberg, Lövkvist, Lan, Weigt, & Aurell, 2013)) and sparse inverse covariance (e.g. PSICOV (Jones, Buchan, Cozzetto, & Pontil, 2012)). Some standard tools that implement these

approaches are used in several of the top predictors such as MetaPSICOV₂ (Buchan & Jones, 2018) and RaptorX-Contact (Wang, Sun, Li, Zhang, & Xu, 2017), which incorporate them in their prediction pipelines. MetaPSICOV₂ includes in its prediction pipeline Freecontact (Kaján, Hopf, Kalaš, Marks, & Rost, 2014) implementations of PSICOV and mfDCA, as well as CCMpred (Seemayer, Gruber, & Söding, 2014) plmDCA implementation. PCP up-to-date predictors exploit the successful implementations that provide Freecontact and CCMpred.

Correlated mutations analysis is carried out from MSA generated using diverse strategies. MSAs are built using three principal tools: PSI-BLAST (Altschul et al., 1997), JackHMMER (Johnson, Eddy, & Portugaly, 2010), and HHBlits (Remmert, Biegert, Hauser, & Söding, 2012). Each approach for PCP selects and integrates some of these alignment tools using reference databases and pipelines that allow getting an optimal MSA that can obtain suitable results for correlated mutations analysis. Some of the more common reference databases in PCP, are PDB70 (Altschul et al., 1997) and Uniref100 (Suzek, Wang, Huang, McGarvey, & Wu, 2015). PDB70 is a subset of sequences extracted from the Protein Data Bank (PDB) with sequence similarity inferior to 70%. Uniref100 contains clusters of unique sequences extracted from UniProt Knowledgebase (Bateman et al., 2017). From PDB70 and Uniref100 we can obtain subsets that reduce the sequence identity, or we can apply any criterion that allows us to enhance PCP.

We can meet a set of the more prominent PCP tools in the residue-residue contact category of the Critical Assessment of Protein Structure Prediction (CASP)

(Moult et al., 2018). In Chapter 2, we selected a top 10 of predictors from the groups that participated in CASP₁₂ (year 2016). From our analysis, we concluded that MetaPSICOV₂ and RaptorX-Contact were the best predictors because they outperform several other predictors while they were not surpassed by any other. Although we can identify predictors that excel in PCP, there are several issues that have to be solved for improving predictions. All predictors keep a high rate of false contact predictions, which results in low precision for hard targets. A common prediction example is the one depicted in Figure 1, where the big proportion of false positives (FP) predicted by MetaPSICOV₂ for the target protein To866 used in the CASP₁₂ is responsible for the low precision around 14%. In Figure 1, all but two of the actual contacts were predicted by MetaPSICOV₂, so that recall is greater than 99%. Predicted false contacts are false positives (FP) and not predicted actual contacts are false negatives (FN). In the example of Figure 1, the overall quality of the predictor falls because the FP are densely placed, which makes it hard to identify real contacts. The discrepancy between the actual and the predicted contact map is easily observed.

PCP must be improved by reducing the FP while preserving the TP. We hypothesize that this goal can be reached using cellular automata (CAs). CAs are a kind of tool that allow us to model discrete systems, where an initial condition is evolved into the desired state employing a set of local transitions. A cellular automaton (CA) is discrete because its global state is limited to a regular lattice where each cell in the lattice can have a state whose value is taken from a

predefined finite set of states. The set of local transitions defines the new state of a cell in the lattice by selecting a subset of near cells to be updated. The subset of cells that the CA uses to evolve cells states is called neighborhood and usually is the same for every cell in the lattice. For our approach, the 2D initial condition is a predicted contact where the set of states is {contact, no contact}. Our goal is to find a neighborhood and a set of transitions (CA rule) that reduce the proportion of FP in the initial condition.

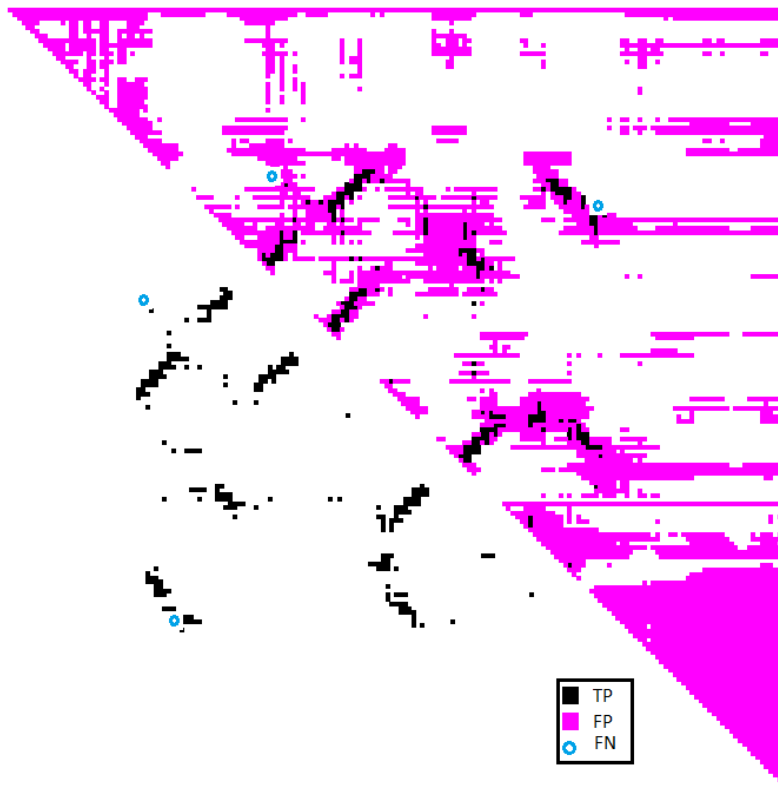


Figure1. MetaPSICOV2 predicted contact and actual contact map for target protein T0866 (CASP12).

Upper triangular: MetaPISCOV2 predicted contacts. Lower triangular: Actual contact map. TP denotes true positives (predicted true contacts); FP indicates no contacts predicted as contacts (false positives); FN (false negatives) represents actual contacts that were not predicted.

In this chapter, we describe the process we defined for PCP based on CA. In section 2 we describe the databases, datasets, and tools used to build our novel PCP tool. In section 3, we describe our prediction results and carry out a statistical analysis comparing the quality of our tool and the top 10 predictors. In section 4, we discuss our results and develop some remarks of PCP based on CAs. Finally, in section 5 we state our conclusions about our tool and PCP based on CAs.

2. Materials and Methods

2.1. Databases and Training Data Sets

We use several reference databases to generate the initial condition for PCP based on CA. For MSAs, we use Uniref reference databases current versions (Suzek et al., 2015). HHblits builds an initial MSA on Uniref20; if the MSA's size is small, a new MSA is obtained utilizing JackHMMer on the Uniref100 database. The bigger MSA is used for correlated mutations identification techniques.

When we use MetaPSICOV2 predictions as an initial condition for CA-based PCP, we obtain an MSA using HHblits on PDB70 to identify protein structural domains, as specified in the MetaPSICOV2 pipeline. PDB70 is a subset of structures in the Protein Data Bank (PDB) (Berman et al., 2002), where the sequences share at most 70% of similarity. The MSAs for structural domains allow running a parallel PCP thread to enhance predictions.

For training, we use the PSICOV data set (Jones et al., 2012), which comprises 150 sequences of lengths in the range [50, 275] amino acids. All the structures in the data set have high resolution (≤ 1.9 Å), are biological macromolecules, and are unique Pfam domains (El-Gebali et al., 2018).

2.2. Initial Conditions Generation

For predictors in Table 2 of Chapter II, we perform a comparison of the performance for each predictor for each target in the category FM used in CASP12. These comparisons allow us to determine if there are differences in the performance of predictors by comparing precision on each target instead of comparing averages of measures. We use Friedman’s test (Pereira et al., 2015) to decide if there are significant differences between predictors and Nemenyi’s posthoc test (Demšar, 2006) to identify the predictors that dominate any other predictors in our top ten. The dominating predictors could be used to generate the initial conditions (IC) to use in the training dataset.

Friedman’s test is a non-parametric statistical test which was proposed to detect the existence of relevant differences in treatments when applied to multiple test subjects. Its use is recommended when five or more predictors need to be compared (Stapor, 2018). The parametric counterpart of Friedman’s test is the ANOVA test with repeated measures. As parametric test ANOVA requires to meet assumptions (Demšar, 2006) that are hard to comply in predictors performance evaluation. The first assumption is that the performance measures are drawn from

normal distributions, and the second assumption is that the performance measures have equal variances. As we cannot ensure these assumptions in CASP₁₂ predictions, we prefer to compare predictors using Friedman's test. For each target, Friedman's test assigns a ranking to each predictor, the best predictor obtains rank 1 and so on. In the case of ties, an average rank is assigned for the predictors with equal performance measure.

Given that Friedman's test can only determine if there are at least two predictors that have significantly different performance, we require to use a posthoc test that allows to find out which of the predictors are better than others. Nemenyi's test is one of the posthoc tests that we can use to compare predictors after Friedman's test rejects the null hypothesis (H_0). H_0 is that there is no difference in the performance of the set of predictors, and the Alternative Hypothesis (H_1) is that at least two predictors have a significant difference in the performance evaluated. Nemenyi's test defines a critical difference (CD) that decides if a predictor is better than another. Predictor i is better than predictor j , if the average rank of predictor i is higher (closer to 1) than the average rank of predictor j and the difference of rankings is greater than the CD.

For the CASP₁₂ contact predictors we included in Table 2 (Chapter II), the protocol that we followed for comparison was: i) we downloaded predictions from CASP₁₂ official results download website (http://predictioncenter.org/download_area/CASP12/, visited March 2019); ii) for downloaded predictions in CASP RR format

(<http://predictioncenter.org/CASP9/index.cgi?page=format#RR>) we calculate precision for medium plus long sequence separation RL top L/2 contacts, top L contacts, and FL, and additionally accuracy for FL; iii) we perform four Friedman's tests (three for precision and one for MCC) at significance level (α) 0.05; iv) for those cases where Friedman's test rejects H_0 we use Nemenyi's posthoc test to identify which predictor dominates any other in precision or accuracy measures.

We use an adaptation of a dominance graph to show the results of Friedman's and Nemenyi's tests. The nodes of the graph represent predictors, and their size denotes the performance of the predictor, the bigger the node, the better the performance of the predictor. Arrows indicate a dominance relationship; the arrow starts at the dominant predictor and ends at the dominated predictor; dashed lines link predictors of equal performance.

Figure 2 illustrates the results of Friedman's and Nemenyi's tests for RL top L/2 contacts. Friedman's p-value (6.22×10^{-9}) indicates rejection of H_0 with a high level of confidence. H_0 rejection allows us to apply Nemenyi's test to find out what predictors are significantly better than the others. In precision for top L/2 contacts, the best predictor is RaptorX-Contact because it obtains the best rank and surmounts CD (2.14) when compared against PConsC31, PCons-net, FALCON_COLORS, IGBTeam, and iFold_1. However, though RaptorX-Contact has a better ranking than DeepFold, naïve, MetaPSICOV and MULTICOM-CONSTRUCT, the difference in each case is lower than CD. Therefore, we can conclude that for precision of top L/2 contacts RaptorX-Contact is not significantly

better than DeepFold, naïve, MetaPSICOV and MULTICOM-CONSTRUCT. DeepFold and naïve tie at the second position and both are significantly better than FALCON_COLORS, IGBTeam, and iFold_1. MetaPSICOV and MULTICOM-CONSTRUCT are not dominated by any other predictor, but neither dominates any other.

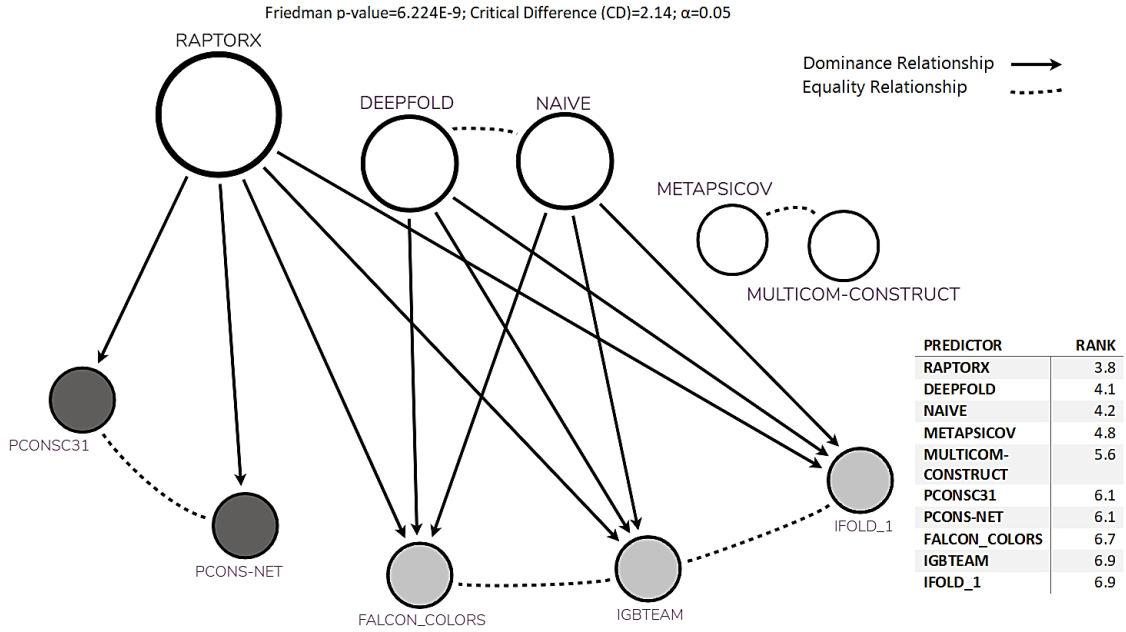


Figure 2. Dominance graph for top 10 contact predictors in CASP12, FM targets, top L/2 contacts, full + medium sequence separation, and precision as performance measure.

Bigger circles represent better rankings.

In Figure 3 we present the results for the statistical comparison of top 10 predictors at RL top L contacts and precision as performance measure. H_0 is rejected by the Friedman's test with p-value 1.174×10^{-10} . Nemenyi's test shows us that the predictor that dominates more predictors than any other is RaptorX-Contact. From Figure 3 we can conclude that RaptorX-Contact is significantly better when used to predict top L contacts for CASP12 FM targets and dominates PConsC31, MULTICOM-CONSTRUCT, PCons-net, iFold_1, FALCON_COLORS, and IGBTeam.

In this comparison, RaptorX-Contact does not show better performance than DeepFold, naïve and MetaPSICOV. In Figure 3 we note that DeepFold is very close in precision performance to RaptorX-Contact except that DeepFold is not better than PConsC31. MetaPSICOV performs the same in top L/2 and L contacts, is not dominated and does not dominate any other predictor. At top L contacts MULTICOM-CONSTRUCT is dominated by RaptorX-Contact and DeepFold, while at top L/2 it is not dominated by any other predictor.

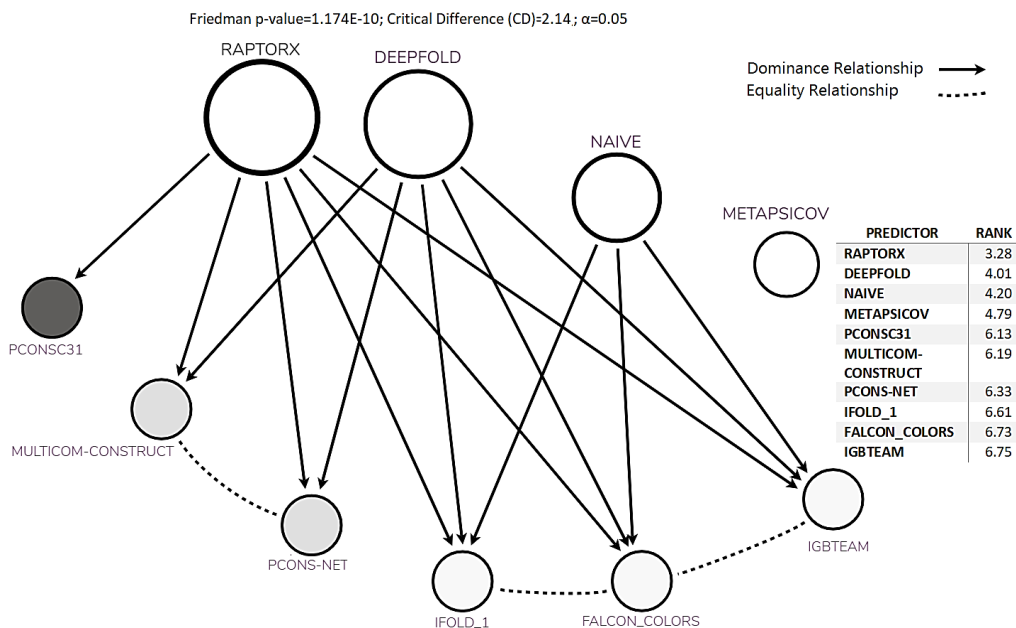


Figure 3. Dominance graph for top 10 contact predictors in CASP12, FM targets, top L contacts, full + medium sequence separation, and precision as performance measure.

Bigger circles represent better rankings.

For FL contacts, Friedman’s test and Nemenyi’s posthoc test give us divergent results if we compare results against L/2 and L contacts. Figure 4 shows that MetaPSICOV dominates in precision every other predictor aside from RaptorX-Contact, and the later one only dominates MULTICOM-CONSTRUCT, FALCON_COLORS, and iFold_1. In Table 3 we contrast performance of

MetaPSICOV and RaptorX-Contact for FL contact and long plus medium sequence separation. We note that average precision is greater, but recall is worse for MetaPSICOV, a possible reason for this is that the proportion of FP is bigger for RaptorX-Contact, which is corroborated in the MCC measure because smaller values can be associated to a bigger proportion of FP. The averages of precision and MCC in Table 1 confirm that MetaPSICOV has better performance than RaptorX-Contact for FL contacts.

When we compare FL contacts predictions using accuracy as performance measure (Figure 5), the results show us a different outcome: MetaPSICOV keeps the first place, but there is a tie with RaptorX-Contact. Both MetaPSICOV and RaptorX-Contact dominate DeepFold, MULTICOM-CONSTRUCT, IGBteam, naïve, iFold_1, PConsC31 and PCons-net. FALCON_COLORS performs significantly better than MULTICOM-CONSTRUCT, IGBteam, iFold_1, PConsC31, and PCons-net. Usually, accuracy is not used in PCP performance evaluation because due to the bias of non-contacts (negatives) a predictor that classifies every residue pair as negative obtains a high accuracy. However, the top 10 predictors evaluated are more prone to predict contacts, so accuracy is valid as a performance measure. MetaPSICOV, RaptorX-Contact, and FALCON_COLORS dominate in accuracy because they have the higher rate, the greater the proportion of TN and less FP. In a comparison by precision, this fact is indiscernible.

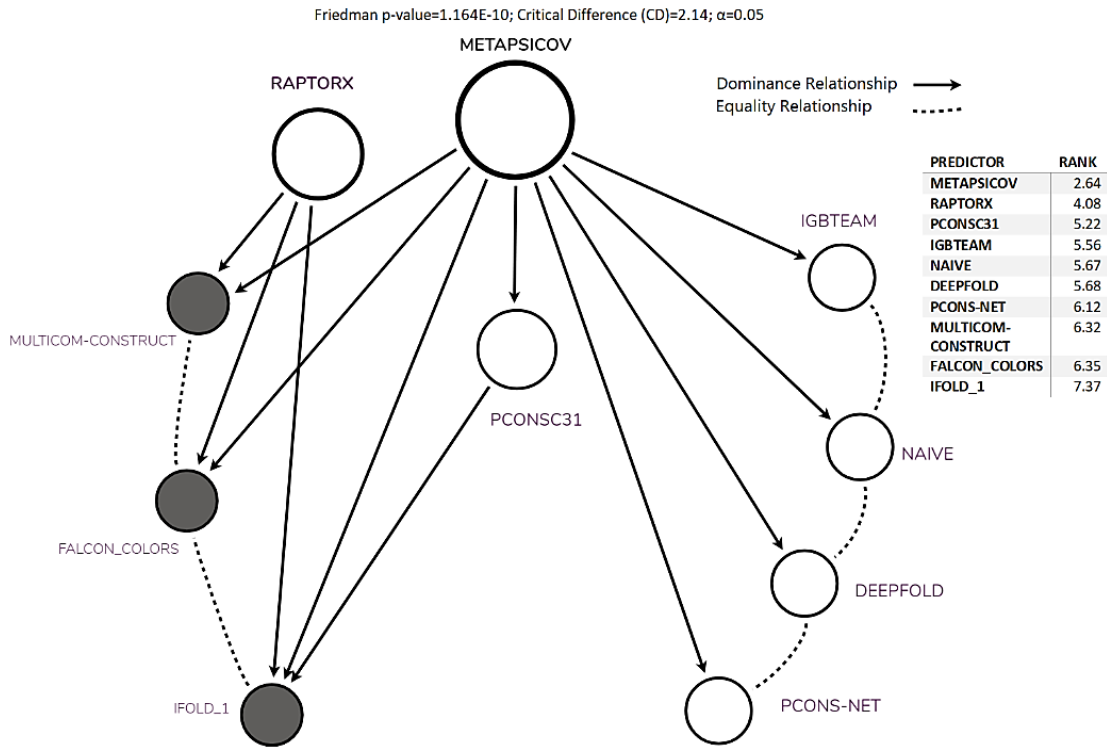


Figure 4. Dominance graph for top 10 contact predictors in CASP12, FM targets, FL, full + medium sequence separation, and precision as performance measure.

Bigger circles represent better rankings.

From this set of statistical comparisons, we conclude that there are some predictors that consistently dominate when we analyze predictions at L/2, L, and FL contacts. RaptorX-Contact is better in L/2 and L contacts and is not significantly better in these cases than MetaPSICOV. MetaPSICOV reverts the situation for FL contacts but does not dominate RaptorX-Contact. Therefore, results in pipelines that integrate PCP results may be more reliable if the predictions came from MetaPSICOV or RaptorX-Contact.

From the results we conclude that the dominating predictors are MetaPSICOV and RaptorX-Contact. We choose MetaPSICOV, because it is available in an open source version. RaptorX-Contact is available online only.

Table 1. All performance measures for predictors MetaPSICOV and RaptorX-Contact. *SS* refers to contacts *between residues in distinct secondary structures*.

Measure	MetaPSICOV	RaptorX-Contact
F1	18.111	13.280
Prec	11.002	7.342
Recall	74.163	76.594
ES	3.553	3.511
ES(ext)	22.731	22.717
Recall(SS Tp)	86.188	87.336
AUC_PR(1)	0.340	0.368
AUC_PR(2)	0.333	0.367
MCC	0.210	0.166
MCC(SS)	0.399	0.333
F1(SS)	65.734	62.003
Prec(SS)	52.899	47.759
Recall(SS)	91.267	93.293

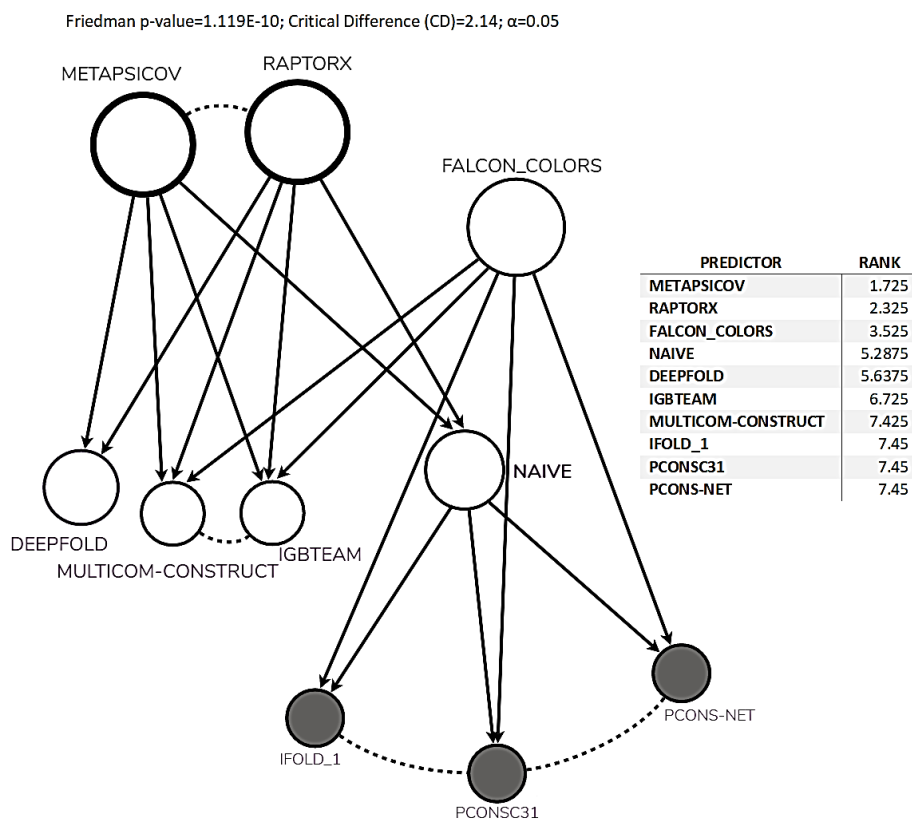


Figure 5. Dominance graph for top 10 contact predictors in CASP12, FM targets, FL, full + medium sequence separation, and accuracy as performance measure.

2.3. Standard Tools

We use several software tools that are very common in the pipelines of many PCP predictors. All these tools are open source and available for several computing systems configurations. Here is a brief description of the more important ones:

JackHMMer (Johnson et al., 2010) implements an optimized algorithm for MSA construction based on hidden Markov models (HMMs). JackHMMER uses a heuristic approach to search homologs iteratively.

HHblits (Remmert et al., 2012) is a tool that allows to build MSAs based on HMMs. HHblits uses a window of thirteen residues around each residue in the query sequence and creates an HMM which is used to search in a database of HMMs iteratively.

CCMpred (Seemayer et al., 2014) and Freecontact (Kaján et al., 2014) are popular in PCP because they implement successful methods for real correlated mutations identification. These tools offer efficient implementations of DCA and are necessary for the prediction pipeline of MetaPSICOV2.

2.4. CA-based PCP Method Description

The common pipeline for PCP involves two main phases (see Figure 6). The initial phase extracts correlated mutations from MSAs that are obtained from several reference databases. The processing starts with the sequence of the target protein that is used to generate MSAs using several reference databases. In the PCP phase,

the predictors use CCMpred, Freecontact or in-house correlated mutations identification tools to extract evolutionary information about contacts from the MSAs.

In some cases, predictors combine several tools for correlated mutations and merge the results. Besides the information provided by the correlated mutation analysis, sequence-related information, such as solvent accessibility, secondary structure and physicochemical properties, is extracted locally in a predefined window in an additional step. To obtain this local information, the predictors define windows around each pair of amino acids in the target sequence. Finally, all information provided by correlated mutations and sequence related features, is refined, and the final PCP is consolidated. The general process, depicted in Figure 6, is followed by predictors such as MetaPSICOV2, RaptorX-Contact, and DeepContact (Liu, Palmedo, Ye, Berger, & Peng, 2018). These predictors use deep learning (Lecun, Bengio, & Hinton, 2015) as machine learning method to merge the contact information obtained previously.

Our approach has a phase of CAs identification that searches the parameters of a CA for PCP (PCP-CA) which are unknown beforehand. The kind of CAs we are using for PCP is taken from the representation of a protein contact map. Specifically, we selected a 2D lattice with binary states and cyclic border condition. The unknown parameters are the neighborhood and the evolution rule. In Figure 7 we show the main phases of our approach. In the PCP-CA training phase, we use as start conditions the PCP of the training data set and we look for a CA that evolves the

initial condition towards the native structures employing a genetic algorithm (GA). We can use this CA to improve PCP of predictors such as MetaPSICOV2. When we use a PCP-CA, the prediction process starts with an initial condition (e.g., MetaPSICOV2 PCP), which is evolved using the CA's rule to a contact map closer to a native contact map and a PCP is generated from this process. Once we have a CA-PCP model and the initial condition, the final prediction is generated in a few seconds.

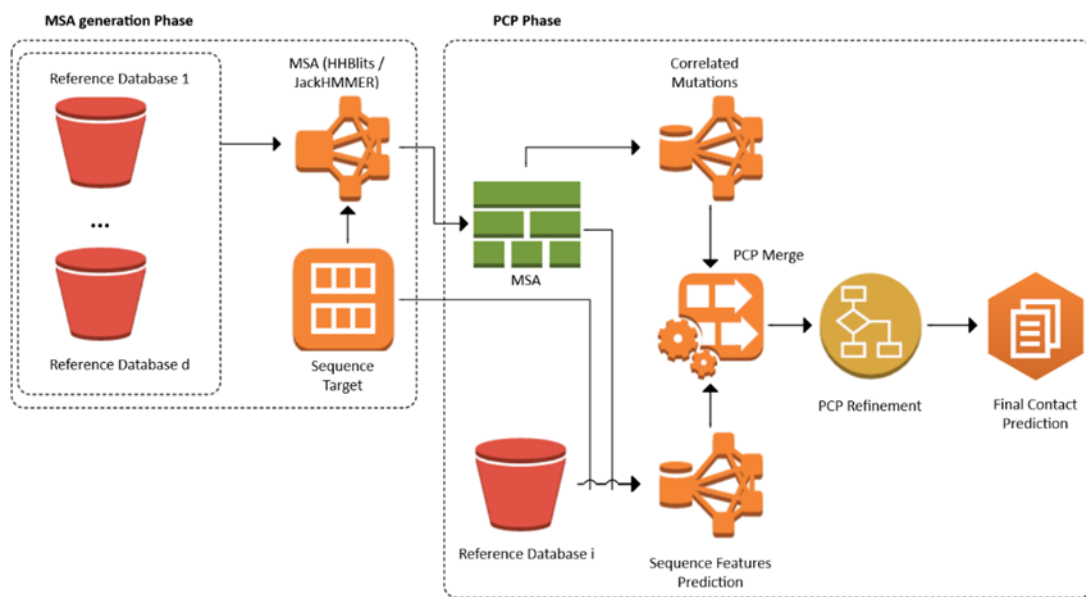


Figure 6. Pipeline for a conventional PCP approach.

MSA generation phase produce an MSA that can be used for correlated mutations analysis and /or prediction of sequence related features. PCP phase implies merging correlated mutations data, MSAs and sequence and structural features to obtain the final prediction

Our approach has a phase of CAs identification that searches the parameters of a CA for PCP (PCP-CA) which are unknown beforehand. The kind of CAs we are using for PCP is taken from the representation of a protein contact map. Specifically, we selected a 2D lattice with binary states and cyclic border condition. The unknown

parameters are the neighborhood and the evolution rule. In Figure 7 we show the main phases of our approach. In the PCP-CA training phase, we use as start conditions the PCP of the training data set and we look for a CA that evolves the initial condition towards the native structures employing a genetic algorithm (GA). We can use this CA to improve PCP of predictors such as MetaPSICOV2. When we use a PCP-CA, the prediction process starts with an initial condition (e.g., MetaPSICOV2 PCP), which is evolved using the CA's rule to a contact map closer to a native contact map and a PCP is generated from this process. Once we have a CA-PCP model and the initial condition, the final prediction is generated in a few seconds.

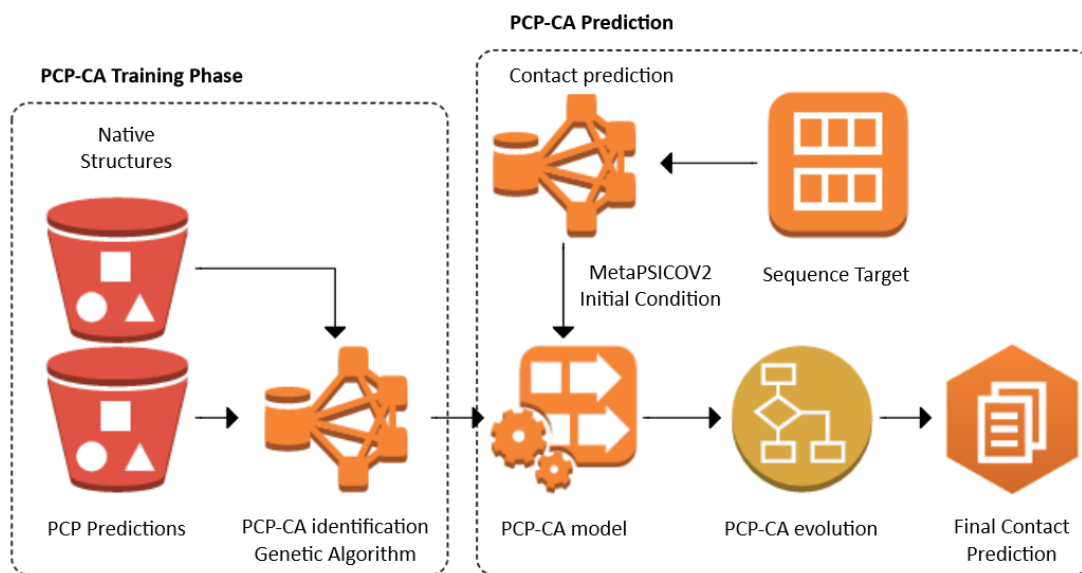


Figure 7. Pipeline for CA-based PCP.

The GA that we used in our approach was implemented utilizing a framework that we developed previously (Díaz & Tischer, 2016).

3. Prediction Results

3.1. Validation Data Set

To validate the results of our approach we used the CASP₁₂ dataset which contains forty proteins belonging to several of the main SCOP families (Andreeva et al., 2008) and with low similarity between them. We use this dataset to evaluate the performance of our approach as well as to compare it to the top 10 predictors in CASP₁₂.

3.2. PCP-CA Model 1

Using our GA for PCP-CA identification, we obtain a CA that can improve MetaPSICOV₂ predictions. The neighborhood decided as optimal by the GA is non-uniform (irregular shape), because it excludes several cells of a complete neighborhood (see Figure 8). The central cell is the one that is updated by the transitions in the CA rule. In the context of CAs, an update can, but not necessarily does, change the state of the cell.

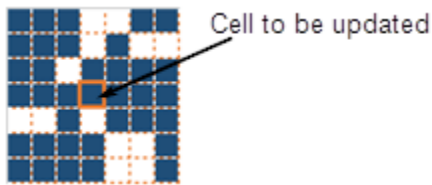


Figure 8. PCP-CA Neighborhood.

The neighborhood max size is a 7×7 2D window. Contacts or no contacts in dark cells affect the transition during state updating of the contact map. The central cell in the neighborhood is the only one that is modified by CA evolution.

For the above neighborhood, a CA rule is extracted by the GA from the data in the training data set. In Table 2 we describe the transitions broken down into four possible results for a cell. The updating process of a contact map can result in state conservation (i.e., a contact state remains as a contact or a no-contact continues as no contact) or in a change of state (i.e., a contact state swaps to no contact or a no-contact switches to contact). The rule used for our PCP-CA contains 45,085 transitions which summarize more than three million of transitions extracted from the contact maps in the training data set. Although the average classification error is low ($< 7\%$) there are more than forty-thousand transitions for contacts that are misclassified.

Table 2. PCP-CA rule description.

Transitions	Number of transitions	Frequency in Training Data Set		Classification Error
		No contact	Contact	
No contact $\rightarrow$ No contact	10,384	3,097,679	42,878	1.37%
Contact $\rightarrow$ No contact	3,207	21,491	1,473	6.41%
No contact $\rightarrow$ Contact	4,854	3,727	18,427	16.82%
Contact $\rightarrow$ Contact	26,640	947	29,901	3.07%
Total	45,085	3,123,844	92,674	6.92%

The transitions in our CA accomplish two main tasks: transitions maintain contacts that are probably real contacts and remove those contacts that appear in the initial prediction but are likely no real contacts. In Figure 5 we show examples of diverse transitions that were frequent in the training data set. In CA-PCP, the evolution process starts with an initial condition that is a contact map extracted from a PCP full prediction (e.g., MetaPSICOV2 prediction), and this initial contact map is evolved by applying the transitions in the CA rule for each cell that represents

contacts for any pair of amino acids i, j in the contact map. Transitions in Figure 9.a apply to cells where there are no contacts, for all the cells that have any of these patterns in its neighborhood; the most likely result of the evolution is that the cell maintains no contact. Transitions in Figure 9.b are very important in PCP because they allow removing contacts that are more likely to be no contacts. We show transitions that add new contacts that were not predicted in the initial condition in Figure 9.c. In Figure 9.d we show transitions that apply to contacts that are more likely to remain in the contact map.

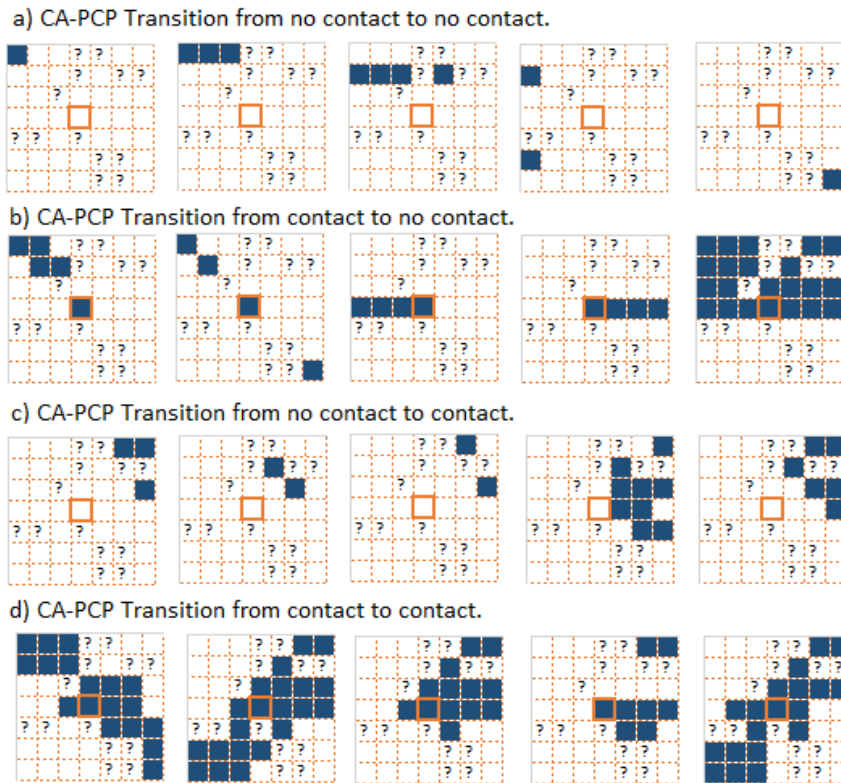


Figure 9. Examples of transitions in the PCP-CA rule.

a) Transitions whose state is no contact and which are maintained in the no-contact state. b) Transitions that change a contact state to a no-contact state. c) Transitions that add new contacts in the contact map. d) Transitions that keep contacts that are already in the contact map. The cells that contain a symbol '?' are not part of the neighborhood, therefore, they do not affect the transition. Blue cells indicate contacts and white cell no contacts.

3.3. Prediction Example using PCP-CA

A prediction using PCP-CA starts with an initial condition. In Figure 10 we show the resulting contact map prediction for the All- β target protein To891 (CASP12). The lower triangular in Figure 6 represents the contact map as predicted by MetaPSICOV2. The initial condition has a high proportion of FP (false contacts), therefore, its precision is low (around 7%). The PCP-CA resulting contact map (Figure 9, upper triangular) keeps a high proportion of all TP (true contacts) and reduce the errors (FP), improving the precision to 51.2%.

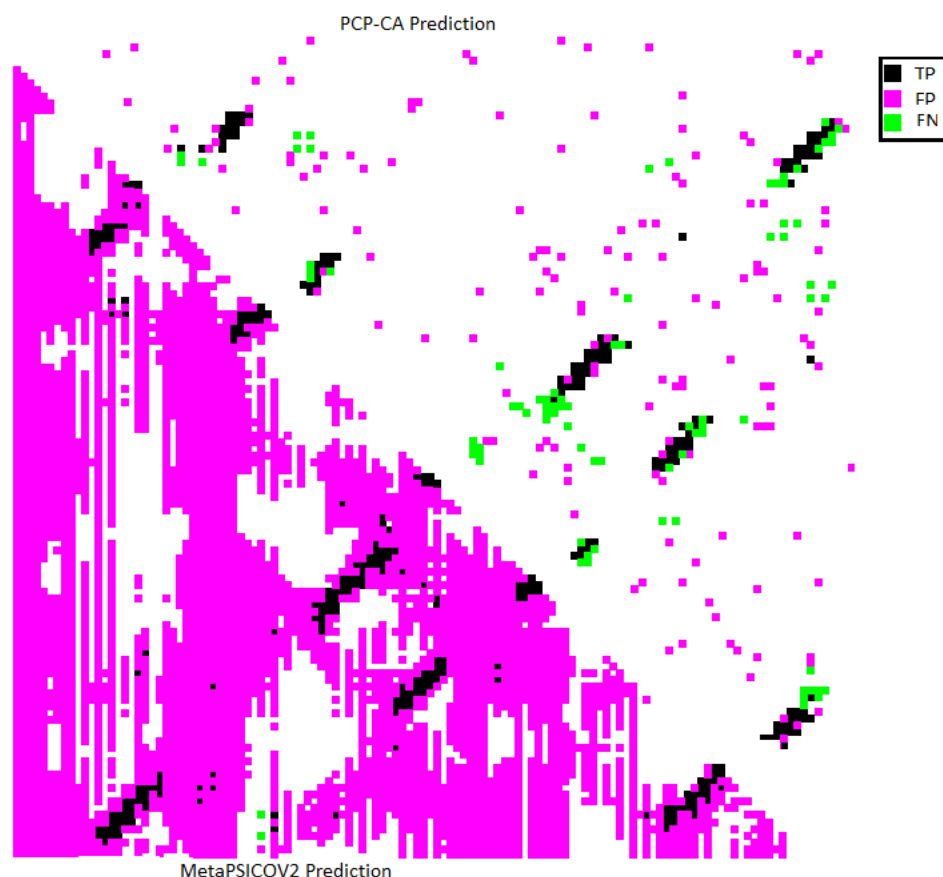


Figure 10. Contact map predicted by PCP-CA / MetaPSICOV2 for CASP12 target protein T0891.

Lower Triangular: initial condition as predicted by MetaPSICOV2. Upper Triangular: Final contact map predicted by PCP-CA.

3.4. PCP-CA Evaluation

We review the results of PCP-CA utilizing a statistical comparison of our results and the results reported for the CASP12 top 10 predictors described in Chapter II. In Table 3 we show the precision for all predictors used in our analysis. Precision in Table 3 was calculated for the full list of contacts predicted with sequence separation at medium and long range (i.e., greater than 12 amino acids).

We compared the overall performance for the full list of predictors/targets employing the Friedman's test and the post-hoc Nemenyi's test as before (see section 2.2 of this chapter). In Figure 10 we display the results of comparing all predictors. The Friedman's test rejects the null hypothesis ($p\text{-value} = 1.35 \times 10^{-10}$) implying that there are significant differences so that we can use Nemenyi's test to identify which are the best predictors. Nemenyi's critical difference defines which is the minimum difference between the ranks of two predictors that allows determining which one of them is better than the other.

In Figure 11, the graph represents dominance relationships of a predictor over another to which an arrow points. Our approach exceeds eight out ten of the other predictors. PCP-CA does not dominate MetaPSICOV and RaptorX-Contact but they themselves dominate every predictor that is dominated by our approach excepting Falcon_Colors (H. Zhang et al., 2016). The comparison of the full list of contacts precision shows that PCP-CA ranks first but it is not good enough to exceed MetaPSICOV and RaptorX-Contacts.

Table 3. Precision by target/predictor CASP12 top 10 plus PCP-CA, for full list contacts and medium/long range sequence separation.

Target	FALCON_COLORS	naive	IGBTeam	iFold_1	Deepfold	MULTICOM-Construct	PconsC31	Pcons-net	RaptorX	Meta-PSICOV	PCP-CA
T0859	3.1%	3.1%	3.0%	3.1%	3.1%	3.1%	3.1%	3.1%	5.0%	3.4%	6.4%
T0860	4.0%	4.0%	4.0%	0.0%	4.0%	4.0%	4.0%	4.0%	16.4%	5.4%	11.7%
T0861	7.8%	4.3%	2.5%	0.0%	4.1%	2.7%	2.9%	2.9%	11.2%	15.3%	25.5%
T0862	5.4%	3.8%	2.8%	2.8%	2.8%	2.8%	2.8%	2.8%	6.4%	9.6%	8.3%
T0863	1.9%	1.8%	1.0%	1.8%	1.8%	1.3%	1.3%	1.3%	1.8%	4.0%	1.3%
T0864	4.7%	2.8%	2.0%	2.9%	2.9%	2.0%	2.0%	2.0%	7.5%	9.2%	18.1%
T0865	0.3%	0.3%	0.3%	0.0%	0.3%	0.3%	0.3%	0.3%	0.5%	0.3%	0.0%
T0866	5.8%	4.1%	4.0%	4.1%	4.1%	4.1%	4.1%	4.1%	8.2%	11.9%	44.2%
T0868	4.7%	3.3%	3.2%	3.3%	3.3%	3.3%	3.3%	3.3%	6.2%	7.1%	18.6%
T0869	3.1%	3.1%	3.1%	3.1%	3.1%	3.1%	3.1%	3.1%	5.5%	3.9%	14.2%
T0870	2.4%	2.4%	2.4%	2.4%	2.4%	2.4%	2.4%	2.4%	4.5%	3.8%	10.1%
T0871	7.2%	5.8%	2.9%	0.0%	5.8%	3.3%	3.0%	3.0%	7.3%	13.6%	23.5%
T0872	5.7%	5.7%	5.6%	5.7%	5.7%	5.7%	5.7%	5.7%	8.4%	5.8%	52.7%
T0873	7.7%	6.0%	3.7%	0.0%	6.0%	4.1%	4.1%	4.1%	7.4%	22.0%	16.9%
T0879	4.9%	2.5%	2.0%	0.0%	2.5%	2.1%	2.1%	2.1%	7.7%	10.7%	28.5%
T0886	7.1%	5.3%	3.2%	5.5%	5.3%	3.6%	3.2%	3.2%	8.1%	17.2%	32.6%
T0889	5.4%	2.8%	1.9%	0.0%	2.8%	2.0%	2.0%	2.0%	8.6%	11.6%	30.3%
T0891	4.6%	4.6%	4.5%	0.0%	4.6%	4.6%	4.6%	4.6%	9.8%	7.8%	51.2%
T0892	2.6%	1.6%	1.6%	1.6%	1.6%	1.6%	1.6%	1.6%	4.2%	5.1%	14.7%
T0893	4.3%	2.2%	1.5%	0.0%	2.2%	1.5%	1.5%	1.5%	6.3%	8.4%	28.1%
T0896	5.3%	3.2%	2.7%	3.5%	3.2%	2.8%	3.0%	2.9%	3.8%	8.8%	6.0%
T0897	4.5%	2.3%	2.0%	3.0%	2.3%	2.3%	2.3%	2.3%	5.0%	6.6%	5.9%
T0898	2.7%	2.1%	2.1%	2.1%	2.1%	2.1%	2.1%	2.1%	5.3%	4.8%	10.2%
T0900	5.4%	5.4%	5.4%	5.4%	5.4%	5.4%	5.4%	5.4%	7.2%	5.6%	33.7%
T0902	5.3%	3.6%	1.9%	0.0%	3.6%	2.0%	2.2%	2.2%	6.7%	11.2%	21.2%
T0903	6.0%	4.6%	1.9%	0.0%	4.6%	2.2%	2.3%	2.3%	6.3%	13.5%	22.4%
T0904	2.8%	2.0%	1.1%	1.5%	1.5%	1.4%	1.5%	1.5%	3.4%	5.3%	7.1%
T0911	7.2%	4.4%	2.3%	5.2%	4.3%	2.6%	2.6%	2.6%	6.2%	15.6%	18.6%
T0912	13.6%	8.5%	5.1%	8.9%	8.5%	5.4%	5.5%	5.5%	10.0%	25.7%	16.1%
T0918	7.4%	8.4%	2.9%	8.3%	8.3%	4.6%	5.2%	5.2%	6.2%	15.8%	17.5%
T0920	7.2%	7.7%	3.3%	0.0%	7.3%	4.7%	4.7%	4.8%	6.7%	26.7%	18.3%
T0921	4.4%	4.3%	4.2%	0.0%	4.3%	4.3%	4.3%	4.3%	8.8%	8.4%	34.3%
T0922	5.2%	5.2%	5.1%	0.0%	5.2%	5.2%	5.2%	5.2%	5.6%	5.2%	27.8%
T0928	11.5%	8.7%	3.9%	0.0%	7.4%	4.6%	4.2%	4.3%	10.0%	19.6%	23.3%
T0941	7.0%	4.2%	3.1%	4.7%	4.7%	3.5%	3.3%	3.1%	4.9%	9.4%	11.5%
T0942	9.0%	7.0%	3.1%	7.0%	7.0%	1.5%	2.7%	2.8%	5.7%	0.0%	0.0%
T0943	7.3%	8.1%	2.9%	0.0%	5.7%	7.3%	3.7%	3.7%	6.2%	14.7%	13.8%
T0944	1.2%	4.4%	2.3%	4.5%	4.6%	2.5%	2.5%	2.5%	8.4%	12.8%	25.4%
T0945	5.8%	3.9%	2.4%	3.8%	3.8%	2.7%	2.5%	2.5%	6.8%	13.4%	17.8%
T0947	6.7%	3.6%	2.9%	3.6%	3.6%	2.9%	2.9%	2.9%	7.3%	11.7%	30.8%
Average	5.5%	4.3%	2.9%	2.4%	4.1%	3.2%	3.1%	3.1%	6.8%	10.3%	20.0%

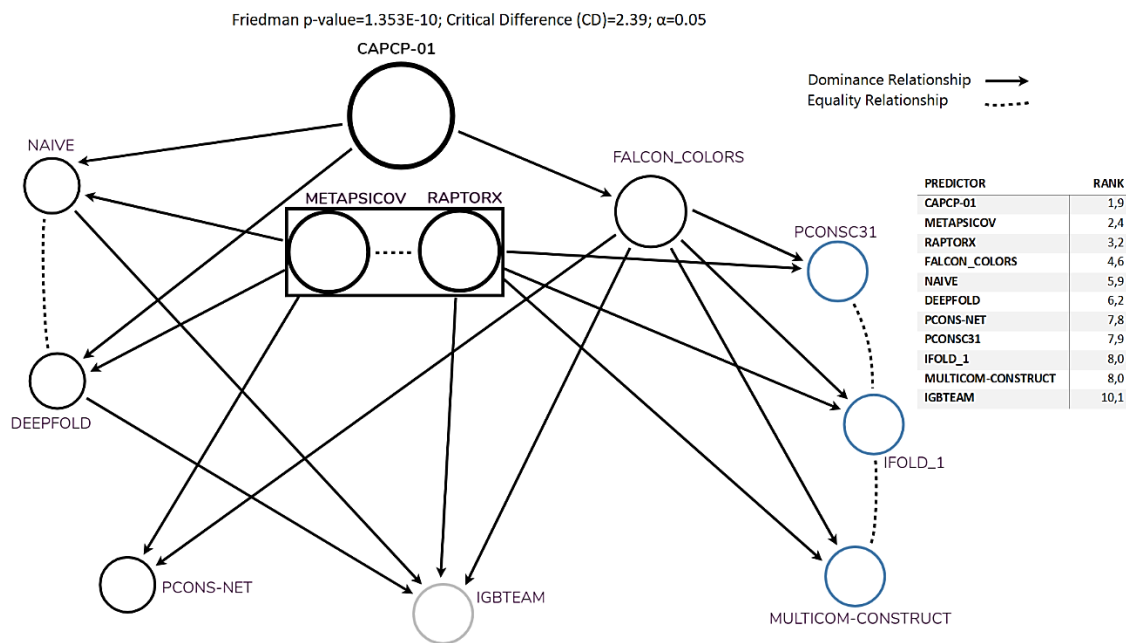


Figure 11. Dominance graph for top 10 contact predictors in CASP12 plus PCP-CA for CASP12 targets, full list of contacts predicted, full + medium sequence separation, and precision as a performance measure.

In Table 4 we show the average precision for all the predictors when we evaluate lists of contacts of different sizes. In PCP, we generally evaluate performance at full list (all predicted contacts), at the top L contacts list (L accounts for the protein sequence size), at the top L/2 contacts list, and the top L/5 contacts list. For reduced lists of contacts, each predictor provides a probability for the occurrence of each contact. PCP-CA performance shows no improvement for reduced lists of contacts, and its average precision is barely at the same level of some predictors such Falcon_Colors and i_Fold_1 (Liu et al., 2018).

4. Discussion

CAs have been used as dynamic models for applications in several areas such as biology, sequence modeling, social behavior, molecular modeling. Though, at the

best of our knowledge they have not been used previously for contact map prediction. We developed a GA was able to identify a CA that improves PCP by mainly reducing the proportion of false contacts that the renowned predictor MetaPSICOV2 produces.

Table 4. Average precision for all predictors at several sizes of list of contacts for all target in CASP12.

Contacts List	FALCON_COLORS	naïve	RaptorX	IGBTeam	iFold_1	Meta PSICOV	Deepfold	MULTICOM-Construct	PconsC31	Pcons-net	PCP-CA
Full List	5.5%	4.3%	6.8%	2.9%	2.4%	10.3%	4.1%	3.2%	3.1%	3.1%	20.0%
Top L	39.9%	52.3%	55.9%	41.8%	24.8%	49.3%	51.5%	45.0%	45.2%	44.0%	45.2%
Top L/2	48.8%	61.3%	63.7%	49.9%	30.1%	58.6%	60.6%	56.3%	55.4%	54.0%	48.8%
Top L/5	57.9%	67.8%	70.8%	59.5%	34.3%	67.3%	67.1%	65.9%	64.3%	64.1%	41.4%

The average precision for the full list of predicted contacts (Table 4) shows a higher precision for PCP-CA. Although PCP-CA precision is 20%, Friedman’s and Nemenyi’s tests concludes that PCP-CA dominates neither MetaPSICOV2 nor RaptorX-Contact.

The decrease of the performance of PCP-CA in reduced lists of predicted contacts (Table 4), occurs because the probability that we assign to our PCP depends solely on the evidences in the training dataset. To improve the probability of predicting a true contact, we require to carry out an extra analysis to determine a better distribution of the probability in order to rise our probability values for real contacts.

The contact maps predicted by PCP-CA resemble real contact maps more than the ones of other predictors because many of the clusters of contacts that those put in their predicted contact maps are reduced using the PCP-CA transitions that

change state from contact to no contact. PCP-CA achieves this because these clusters have a low probability of occurrence in contact maps and the rule transitions remove precisely predicted contacts with low support in real contact maps.

Transitions that maintain the no-contact state have a low classification error (Table 2). However, the raw number of real contacts that classified as no- contacts is bigger than the correct transitions to contact state. This issue is hard to solve in our current approach because the evidences for the CA rule determine that there is support just for the current transitions. One possible solution is to identify a new measure that conserves the actual quality of the contact map and improves the prediction of new contacts, to be used while the GA explore the space of CA models.

5. Conclusions

We proved that with employment of CAs it is possible to improve the quality of contact maps predicted by a top predictor such as MetaPSICOV2. CA transitions can remove false contacts while being able to keep real contacts. Therefore, the full list of predictions keeps a high similarity to the contact map when applied to contacts with medium and long separation ranges.

Friedman's test complemented by a post-hoc test to identify differences in performance of predictors, provides us with a powerful tool that gives a more meticulous analysis of the performance of several predictors. We used it for selecting

the predictor to generate our training dataset as well as to compare our predictor to the CASP12 top 10 predictors.

6. References

- Adhikari, B., & Cheng, J. (2016). Protein Residue Contacts and Prediction Methods. In O. Carugo & F. Eisenhaber (Eds.), *Data Mining Techniques for the Life Sciences* (Vol. 1415, pp. 463–476). Humana Press, New York, NY. <https://doi.org/10.1007/978-1-4939-3572-7>
- Altschul, S., Thomas, M., Alejandro, S., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389–3402.
- Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S. E., Hubbard, T. J. P., Chothia, C., & Murzin, A. G. (2008). Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Research*, 36(SUPPL. 1), 419–425. <https://doi.org/10.1093/nar/gkm993>
- Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., ... Zhang, J. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, 45(D1), D158–D169. <https://doi.org/10.1093/nar/gkw1099>
- Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., ... others. (2002). The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, 58(6), 899–907.
- Buchan, D. W. A., & Jones, D. T. (2018). Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins: Structure, Function and Bioinformatics*. <https://doi.org/10.1002/prot.25379>
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1), 1–30. <https://doi.org/10.1016/j.jecp.2010.03.005>
- Diaz, N., & Tischer, I. (2019). Mining Stochastic Cellular Automata to Solve Density Classification Task in Two Dimensions.
- Díaz, N., & Tischer, I. (2016). Generic framework for mining cellular automata models on protein-folding simulations. *Genetics and Molecular Research*, 15(2), 1–16. <https://doi.org/10.4238/gmr.15028654>

- Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., & Aurell, E. (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Physical Review E*, 87(1), 1–19. Retrieved from <http://pre.aps.org/abstract/PRE/v87/i1/e012707>
- El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., ... Finn, R. D. (2018). The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(October 2018), 427–432. <https://doi.org/10.1093/nar/gky995>
- Johnson, L. S., Eddy, S. R., & Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 11(431), 1–8. <https://doi.org/10.1186/1471-2105-11-431>
- Jones, D. T., Buchan, D. W. a, Cozzetto, D., & Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics (Oxford, England)*, 28(2), 184–190. <https://doi.org/10.1093/bioinformatics/btr638>
- Kaján, L., Hopf, T. a, Kalaš, M., Marks, D. S., & Rost, B. (2014). FreeContact: Fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics*, 15(85), 1–6. <https://doi.org/10.1186/1471-2105-15-85>
- Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Liu, Y., Palmedo, P., Ye, Q., Berger, B., & Peng, J. (2018). Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. *Cell Systems*, 6(1), 65–74.e3. <https://doi.org/10.1016/j.cels.2017.11.014>
- Martin, L. C., Gloor, G. B., Dunn, S. D., & Wahl, L. M. (2005). Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, 21(22), 4116–4124. <https://doi.org/10.1093/bioinformatics/bti671>
- Monastyrskyy, B., D’Andrea, D., Fidelis, K., Tramontano, A., & Kryshtafovych, A. (2015). New encouraging developments in contact prediction: Assessment of the CASP11 results. *Proteins*, (October), 1–14. <https://doi.org/10.1002/prot.24943>
- Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., ... Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49), E1293–E1301. <https://doi.org/10.1073/pnas.1111471108>
- Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2018). Critical assessment of methods of protein structure prediction (CASP)—

- Round XII. Proteins: Structure, Function and Bioinformatics, 86, 7–15.
<https://doi.org/10.1002/prot.25415>
- Pereira, D. G., Afonso, A., & Medeiros, F. M. (2015). Overview of Friedmans Test and Post-hoc Analysis. *Communications in Statistics: Simulation and Computation*, 44(10), 2636–2653.
<https://doi.org/10.1080/03610918.2014.931971>
- Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2), 173–175. <https://doi.org/10.1038/nmeth.1818>
- Seemayer, S., Gruber, M., & Söding, J. (2014). CCMpred - Fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics*, 30(21), 3128–3130.
<https://doi.org/10.1093/bioinformatics/btu500>
- Stapor, K. (2018). Evaluating and Comparing Classifiers: Review, Some Recommendations and Limitations Katarzyna. In M. Kurzynski, M. Wozniak, & R. Burduk (Eds.), *Proceedings of the 10th International Conference on Computer Recognition Systems CORES 2017* (Vol. 578, pp. 12–21). Springer International Publishing AG 2018.
<http://doi.org/10.1007/978-3-319-59162-9>
- Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., & Wu, C. H. (2015). UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6), 926–932.
<https://doi.org/10.1093/bioinformatics/btu739>
- Tanaka, S., & Scheraga, H. a. (1975). Model of protein folding: inclusion of short-, medium-, and long-range interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 72(10), 3802–3806.
<https://doi.org/Doi 10.1073/Pnas.72.10.3802>
- Wang, S., Sun, S., Li, Z., Zhang, R., & Xu, J. (2017). Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Computational Biology*, 13(1), 1–34. <https://doi.org/10.1371/journal.pcbi.1005324>
- Zhang, C., Mortuza, S. M., He, B., Wang, Y., & Zhang, Y. (2018). Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. *Proteins: Structure, Function and Bioinformatics*, 86, 136–151. <https://doi.org/10.1002/prot.25414>
- Zhang, H., Gao, Y., Deng, M., Wang, C., Zhu, J., Li, S. C., ... Bu, D. (2016). Improving residue-residue contact prediction via low-rank and sparse

decomposition of residue correlation matrix. *Biochemical and Biophysical Research Communications*, 472(1), 217–222.
<https://doi.org/10.1016/j.bbrc.2016.01.188>

Chapter V.

Multiclass Protein Contact Maps

Abstract

Protein contact maps are a valuable tool for protein structure prediction. A contact map with just a fraction of true contacts gives a useful representation of the protein structure. However, contacts between amino acids distant in the protein sequence are scarce and hard to predict. Here, we describe an approach that allows us to improve binary contact map representation with regards to native contacts formation, by mean of a multiclass representation. We show that decision trees performance clustering and non-parametric statistical tests allow us to determine the combination of contact ranges that catches native contacts formation. We generate our datasets by a combination of contact ranges and covering arrays to select representative contact ranges. In total, we compare 413 decision trees. Our approach allows to replace the standard representation of just two states (contact, non-contact) by an improved contact map representation: we obtain a representation of four states to describe contacts according to the ranges obtained by the data mining-based process. This representation allows us capturing some details of native contact formation that are not visible in two-state contact maps.

1. Introduction

Proteins are complex molecules that perform essential roles for every living being; their three – dimensional structure determines their specific biological role. Protein contact prediction (PCP) is an active research topic in the area of protein structure determination. In PCP, contact maps (CMs) are an important tool that helps in several tasks related to structure prediction (Gao et al., 2019). The traditional CM representation is a 2D binary matrix. Values are defined, typically by a cutoff range of $[4 \text{ \AA}, 8 \text{ \AA}]$ for residue-residue (RR) contact. We are looking for an extended version CMs that provide more information about the less frequent contacts at medium and large ranges. Each cell in a CM indicates physical proximity for a pair of residue atoms of a protein. Trivial closeness, such as consecutive residues in the protein chain, is irrelevant on a CM. Meaningful contacts are those at chain separation greater than six residues and more notably medium-long range (chain separation greater than 12 residues) (Schaarschmidt et al., 2017). Closeness is measured between the same atoms for each pair of residues. The commonly used residue atoms are carbon-alpha or carbon-beta; in some cases, groups of atoms are used (see Glossary in Chapter II).

There is not a consensus about the cutoff range; in the literature, contact cutoff ranges start at $3.5 \text{ \AA}$ and go up to $9 \text{ \AA}$ (Da Silveira et al., 2009). An assessment of the optimal cutoff range for 3D structure reconstruction is described in (Duarte et al., 2010). For a dataset of 60 proteins, they found that contacts in the cutoff range from $10 \text{ \AA}$ to $12 \text{ \AA}$ enhance the reconstruction accuracy. (Vassura et al., 2008)

conclude that for a cutoff range from 9 Å to 18 Å reconstruction accuracy improves significantly.

A quaternary contact map representation is used in (Walsh et al., 2009) to improve contact prediction accuracy. Contact ranges are $[0 \text{ Å}, 8 \text{ Å})$, $[8 \text{ Å}, 13 \text{ Å})$ and $[13 \text{ Å}, 19 \text{ Å})$. This representation provides more information than the one offered by traditional binary contact maps, specifically for scarce long-range contacts. The main results can be summarized as follows: 1) Quaternary contact maps can achieve accuracy comparable to state-of-the-art contact prediction tools (SVMCon, SAM_To6, ProfCon, Betapro); 2) Structure reconstruction accuracy is improved by quaternary contact maps. A systematic approach to identifying multiple contact ranges is not known at the best of our knowledge. Recently, Kukic et al. try to extend their previously introduced four-class contact maps to distance matrix prediction (Kukic et al., 2014). However, the results are unfavorable for distance matrix prediction and they are significantly worse for proteins with sequence length greater than 200 residues.

Although optimal contact map ranges have been studied, current results are not conclusive. We propose an exploratory methodology that aims to obtain optimal contact ranges. Our methodology involves generation of contact ranges and detailed evaluation of machine learning models trained on an extensive dataset of contact ranges.

Most approaches for evaluation of machine learning models are focused on a single performance measure. However, there is no consensus about a measure to

evaluate rule-based classification methods such as decision trees (DTs). In (Sathyadevan and Nair, 2015), a dataset is used to compare classification accuracy for three classifiers (ID3, C4.5, and Random Forest) and they draw conclusions about which method is better for classification. (Demšar, 2006) uses Area Under the Curve (AUC) to compare DTs built with the C4.5 algorithm, combining several values of confidence and minimum number of instances per leaf. His goal was to identify the DT with the best performance on several datasets.

For some years now, researchers have been tackling the issues of using single performance measures in classifiers comparison ((Provost et al., 1998), (Caruana and Niculescu-Mizil, 2006)). However, there is a consensus that single performance measures are unable to capture all desirable aspects of models' evaluation. Therefore, we treat our problem as a multiple criteria decision analysis (MCDA) problem over DTs. Dealing with a data mining project we define the fundamental goal, and according to it, we select the evaluation criteria that allows us getting the best models for the task (Fayyad et al., 1996). In this chapter, we are addressing an unconventional problem that requires a specific approach. MCDA focused on DT sets is described in (Osei-Bryson, 2004). We adapt that methodology to our problem aiming to select the DT that best represents dynamics at the level of contact ranges.

2. Methods

Our method involves three major phases (Figure 1). Each phase addresses a key task:

1. Generation of a set of contact ranges combinations for analysis and evaluation.
2. Selection of a subset of optimal candidates for contact ranges combinations.
3. Comparison and evaluation of top models.

In Figure 1, we show these phases, their main tasks and their corresponding inputs and outputs. In the first phase, data preprocessing involves the building of datasets for training. Our primary source of data is a dataset of protein trajectories. As we want to select an optimal representation of contact maps with more than two classes, we had to choose some representative contact range sets. We used two approaches to obtain a total of 413 possible representations: covering arrays and combinatorial combination. Phase I ends with a set of 413 trained DTs. In phase II, we used traditional measures for machine learning evaluation and clustering to select a subset of decision trees that stand out. In Phase III, we use an additional set of measures to evaluate in detail the models selected in Phase II. This set of measures allows to assess the performance of DTs in a more detailed way. Phase III finishes with the selection of the best multiclass contact map representation using statistical analysis.

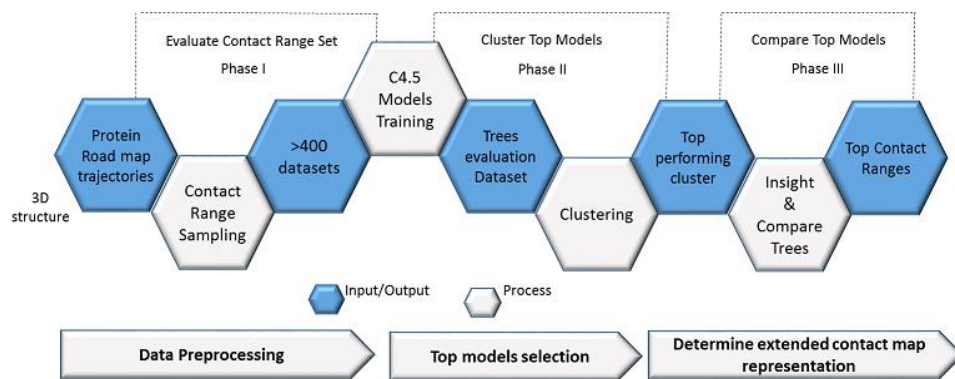


Figure 1. Method RoadMap. Three phases, inputs/outputs, and processes.

The proposed method's final goal is to select the dataset combination that optimally captures events of formation of native contacts. Below we describe the used datasets, the performance measures and the criteria and methods to select optimal datasets.

2.1 Contact Ranges Combinations

To explore a diverse set of contact range combinations, we use two ways of sampling. The first one is a subset of 256 combinations taken from covering arrays (Torres-Jimenez and Izquierdo-Marquez, 2014). We used a subset of possible range combinations because the whole search space is too big to be assessed exhaustively.

Covering arrays are used to minimize the number of test cases, guarantying the coverage of relevant combinations of a predefined size. Commonly, covering arrays are used in the design of experiments for software testing (Torres-Jimenez et al., 2015). In the definition of a covering array, we select several features (Factors), and from them we choose subsets. The size of factors in the subset is called Strength. The advantage of covering arrays is that with a subset of factors combinations, we

can be sure to comprise all combinations of factors at the desired strength. In our case, we use covering arrays to warrant that with a subset of possible range combinations we are covering an optimal subset. Calculation of an optimal covering array is a combinatorial problem (Torres-Jimenez and Izquierdo-Marquez, 2014). For our approach, we used a covering array kindly provided by Ph.D. Jose Torres-Jimenez (http://www.tamps.cinvestav.mx/investigador/_jose/_torres/_jimenez, visited: October: 2017).

For contact range combinations, we use a covering array of 256 combinations, strength 7, and 12 factors (each factor denotes a contact range of width 1 Å, starting at 8 Å). Strength is a parameter that specifies that each combination of 7 ranges is used to generate a dataset. Using 12 factors we cover contact ranges from 8 Å to 20 Å. To assess results in a way that is similar to the protocol used in Critical Assessment of Techniques for Protein Structure Prediction (CASP) (Monastyrskyy et al., 2014), we consider the contact range [4 Å, 8 Å) as mandatory. A combination in this covering array is a binary set of twelve factors that indicate when a range of width 1 Å is part of contact range (value 1), or not (value 0). An example of a row in this covering array is {0,1,1,1,0,0,0,0,1,1,1}, for this combination we built a dataset with contact ranges {[4 Å, 8 Å), [9 Å, 13 Å), [18 Å, 20 Å)}. For each row in the covering array, we proceed in the same way. By including the range [4 Å, 8 Å), our representation allows the extraction of the binary contact map at an 8 Å threshold.

For the second way of sampling contact ranges combinations, we generate 157 combinations of ranges of 2 Å, 3 Å, 4 Å, and 5 Å. We generate every possible combination of this ranges to complete our set, because, covering arrays do not allow us to codify combinations without gaps.

2.2 Datasets Generation

Our primary source is a set of protein folding trajectories selected from a repository created by probabilistic roadmaps methods (Thomas et al., 2005). We choose the folding trajectories of three proteins, one from each major protein folding class (protein 1PRB for class α , for class β we choose protein 1HCC, and protein 1GB1 for structural class $\alpha + \beta$), obtaining 529 structures. For each one of the 413 range combinations, we generate a dataset using the 529 protein structures. The building process for every range combination involves three steps.

1. Generate a contact map for each one of the 529 protein structures, matching a range combination.
2. Transform the 529 contact maps in a table of 50 columns by n^2 rows, where n is the size of the protein sequence. For each cell in the contact map, we put a row in the table. 49 columns are for contact pairs in a radius 3 for each cell i, j in the contact map. The last column contains the contact value of cell i, j in the protein structure that follows in the trajectory. In this way, we keep the dynamic changes that happen in the folding trajectory.

3. Filter rows that do not have contacts in cell i,j for the current contact map or the next contact map in the same folding trajectory.

Carrying out this process for each range combinations, we obtain the corresponding 413 datasets.

2.3 Decision Trees Building and Evaluation

To identify the best contact range combinations that capture native contact formation, we rely on DTs (Sathyadevan and Nair, 2015). DTs provide us with an explicit representation of each possible solution to a decision, based on conditions specified by the data (Stiglic et al., 2012). We can assess DTs by measures that evaluate the quality of the whole model, as well as by measures for each decision node.

To generate the 413 DTs we used Weka platform (Frank et al., 2005). A well-known implementation of DTs is the J48 implementation of algorithm C4.5 (Quinlan, 1993). All DTs were built with the same parameters to ensure a fair evaluation.

We selected 12 performance measures to evaluate DTs built over each contact range combination: area under precision-recall curve, false negative rate, true negative rate, false positive rate, true positive rate, f measure, precision, recall, kappa, mean entropy gain, Matthews correlation, mean prior absolute error, relative absolute error. This set of measures is generally used in machine learning. For

CASP12, the approach for evaluation of protein contact prediction changed and uses now a similar set of measures (F1 score, precision, recall, area under the precision-recall curve, Matthews correlation). These measures allow evaluating with more detail the performance of contact predictors (Adhikari et al., 2016).

After training the 413 DTs, we evaluated each one and created a new dataset of 413 rows and 12 columns. This dataset allows to evaluate global performance for each DT, and we used it to select the top models as we describe in the section below.

2.4 Top Models Selection

To identify a subset of the contact range combinations that show better performance, we used EM clustering algorithm (Sharma et al., 2012). EM algorithm identifies the optimal number of clusters by performing several iterations. In Figure 2 we show clustering results for the 21 optimal clusters. In Figure 2.a) we show a scatter plot of normalized values of relative absolute error vs. Matthews Correlation Coefficient; in the upper left corner, we can observe the top three clusters (clusters 4, 10 and 11). In Figure 2.b) we show a scatter plot of normalized values of precision-recall area under the curve vs. Matthews Correlation Coefficient; in the upper right corner, we can observe the same top three clusters (clusters 4, 10 and 11).

Matthews Correlation Coefficient is a useful measure when we are facing data where classes are biased. The area under the curve for precision-recall is best suited for our case because recall and precision allow us to measure contact prediction performance for the minority classes that represent contacts. DTs for contact ranges

combination in top-3 clusters all are 4-states models (3 contact ranges + 1 non-contact). Similarly, models in clusters with worst performance are all 7-states and 8-states models.

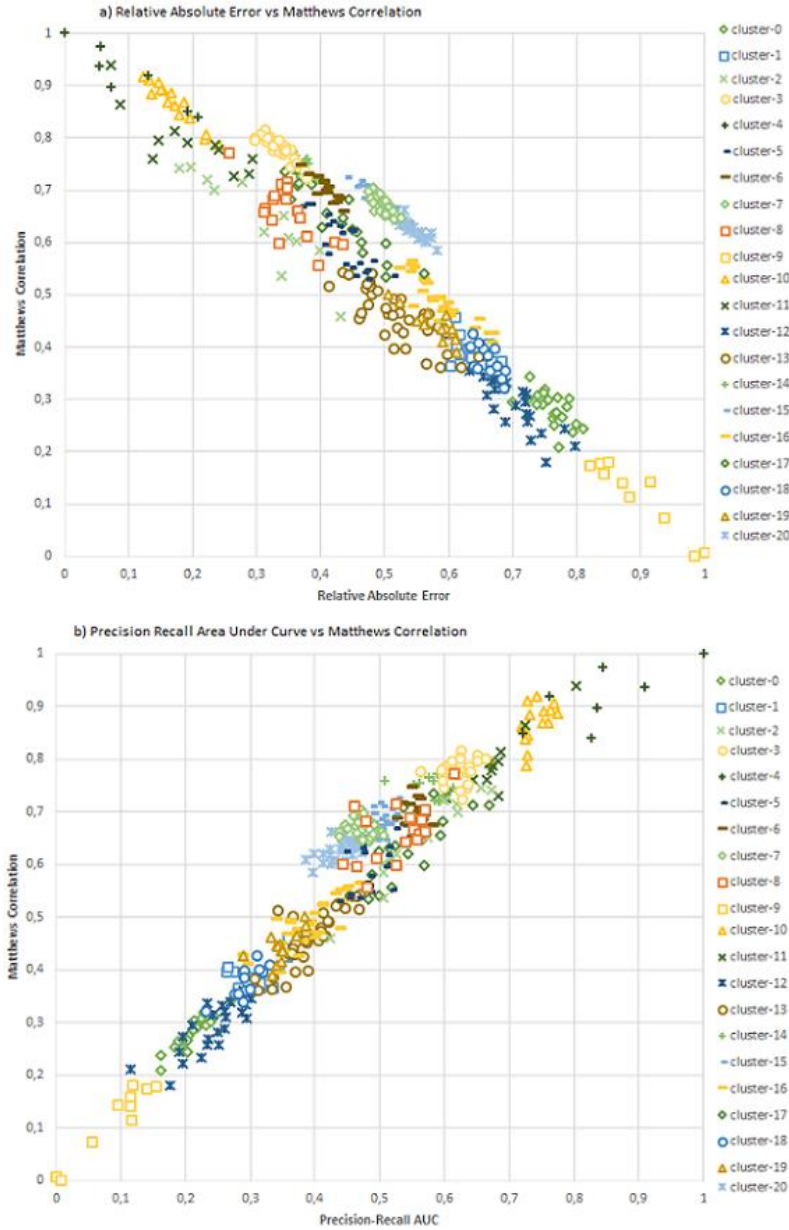


Figure 2. EM clustering of DTs performance measures for 413 contact ranges combinations.

a) Clusters dispersion by Relative Absolute Error vs. Matthews Correlation Coefficient. b) Clusters dispersion by Precision-Recall Area Under the Curve vs. Matthews Correlation Coefficient. Values are normalized.

2.5 Top Decision Trees Comparison

We used KEEL (Alcalá-Fdez et al., 2009) implementation of Friedman test to evaluate if there are some DT that reject the null hypothesis. In our case, rejecting the null hypothesis implies that there is a statistically significant difference among contact range combinations.

We re-evaluated models in the top 3 clusters using additional performance measures, which allow evaluating specific characteristics of our DTs. We adapted these measures from (Osei-Bryson, 2004):

- Coarse Stability: quantifies the similarity between accuracies for training and validation cases. Values close to 1.0 imply higher stability of the DT.

$$STAB_{Coarse} = \text{Min}\left(\frac{Accuracy_{Training}}{Accuracy_{Validation}}, \frac{Accuracy_{Validation}}{Accuracy_{Training}}\right)$$

- Fine Stability: analogously to coarse stability, measures how similar are accuracies for training and validation but quantifying at the level of each leaf for the whole decision tree.

$$STAB_{Fine} = \sum_k \varphi_{Vk} * \sigma_{VTk}, k \in [1, ||leaves||]$$

$$\varphi_{Vk} = \frac{||Cases_{Vk}||}{||Cases_{Validation}||}, \sigma_{VTk} = \text{Min}\left(\frac{\rho_{Vk}}{\rho_{Tk}}, \frac{\rho_{Tk}}{\rho_{Vk}}\right),$$

$$\rho_{Tk} = \frac{TruePositives_{Tk}}{||Cases_{Validation}||}, \rho_{Vk} = \frac{TruePositives_{Vk}}{||Cases_{Validation}||}$$

- Rule Simplicity: measures the weighted depth of each leaf, adjusting it to an expected max leaf depth. Shortest trees have higher rule simplicity.

$$SIMPL_{Rule} = \frac{20 - x_{Mean}}{20},$$

$$x_{Mean} = \sum_k \phi_{V_k} * x_k,$$

$$x_k = depth(leaf_k)$$

- Discriminatory Power: quantifies coverage of leaves that surpass a predefined threshold for precision.

$$DSCPWR = \sum_k \phi_{V_k} * \Psi(k),$$

$$\Psi(k) = \begin{cases} 0, & Precision_{T_k} \geq \tau \\ 1, & Precision_{T_k} < \tau \end{cases}, \tau: Precision \text{ threshold}$$

We used as Precision Threshold (τ) the value 0.7. In Figure 3 we show the frequency of rules by ranges of precision for DTs in the dataset. $\tau = 0.7$ guarantees that the rules that we are using for analysis have a reasonable precision, covering a representative subset of rules.

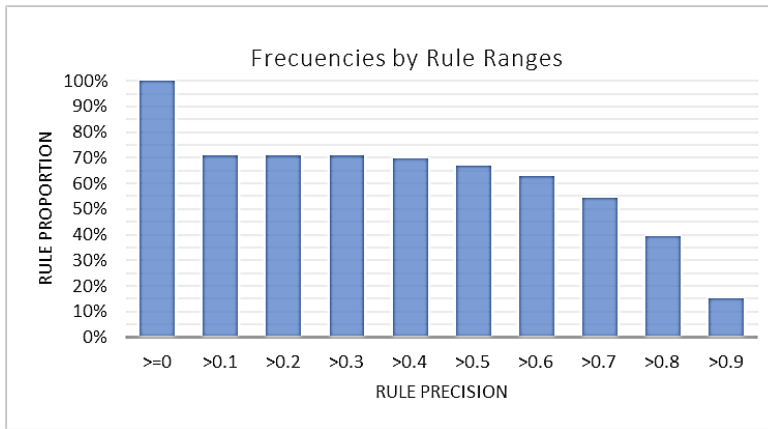


Figure3. Frequencies by rule ranges in the dataset.

Rules with precision above 0.7 covers about 55% of total rules. For precision above 0.8 proportion of rules accounts for 40% of rules.

- Native Contact Precision: measures precision for leaves whose class is a native contact. We propose this measure because native contacts are the main issue to consider in contact prediction. We try to evaluate performance in native contact prediction using this measure.

$$Precision_{NativeContact} = \sum_k \rho_{Tk} * Precision_{Tk} * N(k),$$

$$N(k) = \begin{cases} 0, & Class(k) \neq native\ contact \\ 1, & Class(k) = native\ contac \end{cases}$$

In Table 1 we present the average rankings calculated for all measures by Friedman test. The Friedman test recognizes significant statistical difference (p-value 6.7598^{-10} ; Friedman statistic 150.8165, χ^2 critical value with 30 degrees of freedom 18.49), therefore, it is appropriate to perform post-hoc tests to find contact ranges combinations that are significantly different. We use Nemenyi test and critical distance (CD) to determine significantly different models (Demšar, 2006).

In Figure 4 we depict the results of Nemenyi test. Post-hoc Nemenyi test ($\alpha = 0.05$, CD: 15.3089), allows to identify that just the top ten models outperform models with ranking above 20 (Table 1). For models with rankings greater than 12 and less than 20, null hypotheses are not rejected. Model 1 (contact ranges: [4 Å, 8 Å); [8 Å, 10 Å); and [10 Å, 15 Å)) and Model 5 (contact ranges: [4 Å, 8 Å); [8 Å, 10 Å); and [10 Å, 13 Å)) are the best performers in native contact precision (0.2307 and 0.3867, respectively). Model 1 outperforms model 5 in ranking and the average of normalized values for coarse stability, fine stability, accuracy test, rule simplicity,

discriminative power, f-measure, precision, recall, Matthews coefficient and native contact precision (0.897 vs. 0.597).

Table 1. Average rankings of DTs for top contact range combinations.

Model	Contact Ranges	Ranking	Model	Contact Ranges	Ranking
1	[4,8); [8,10); [10,15)	5.0	17	[4,8); [8,15); [17,18)	16.2
2	[4,8); [8,10); [10,14)	6.6	18	[4,8); [8,9); [12,16)	16.6
3	[4,8); [8,13); [13,20)	6.7	19	[4,8); [8,13); [13,16)	17.5
4	[4,8); [8,12); [12,17)	7.7	20	[4,8); [10,11); [12,18)	17.55
5	[4,8); [8,10); [10,13)	8.5	21	[4,8); [8,13); [13,15)	17.8
6	[4,8); [8,10); [12,19)	9.2	22	[4,8); [9,12); [14,19)	19.5
7	[4,8); [8,11); [11,16)	9.7	23	[4,8); [8,9); [15,18)	20.9
8	[4,8); [8,13); [13,18)	10.1	24	[4,8); [8,12); [16,17)	21.1
9	[4,8); [8,11); [11,15)	11.4	25	[4,8); [8,12); [12,14)	21.45
10	[4,8); [8,13); [19,20)	11.7	26	[4,8); [9,10); [14,18)	22.9
11	[4,8); [8,13); [13,17)	12.4	27	[4,8); [10,12); [19,20)	23.9
12	[4,8); [8,11); [11,14)	14.1	28	[4,8); [11,15); [19,20)	25.4
13	[4,8); [8,12); [12,15)	14.8	29	[4,8); [8,9); [17,20)	26.15
14	[4,8); [8,12); [12,16)	15.5	30	[4,8); [9,15); [18,20)	27.05
15	[4,8); [8,11); [11,13)	15.7	31	[4,8); [10,14); [17,18)	27.1
16	[4,8); [8,10); [10,12)	15.8			

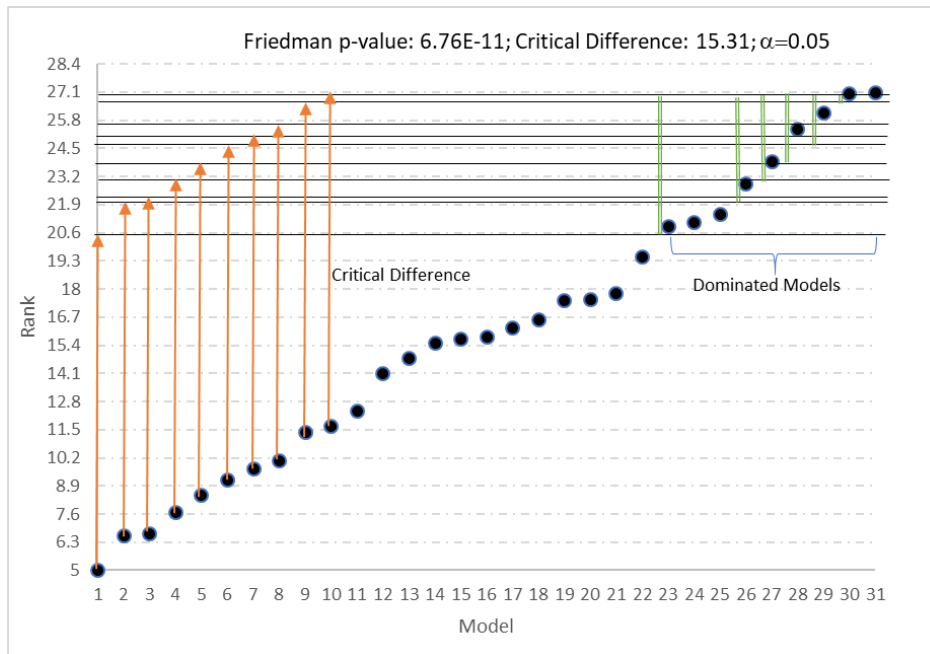


Figure 4. Nemenyi's test results.

Arrows indicate the CD range. Models above the CD of another model are dominated. Model 1 overcomes models 23 to 31. Models 2 and 3 overcome models 26 to 31. Models 4 and 5 dominate models 27 to 31.

3. Results

Model 1 provides some interesting rules that capture native contact formation and that are unlikely to be expressed with the typical representation for binary contact maps. Figure 5 allows us to contrast a specific rule from Model 1 in which a native contact is formed in the final contact map. Figure 5.a) shows cells that are used as decision nodes in the rule (shadowed cells), for a 2-state contact map we cannot see any pattern to express native contact formation. All decision nodes for the rule in Figure 5.a) are in the non-contact state, therefore, is impossible to decide about native contact formation. Figure 5.b) shows a contact map pattern comprised of decision nodes in a rule for native contact formation from Model 1.

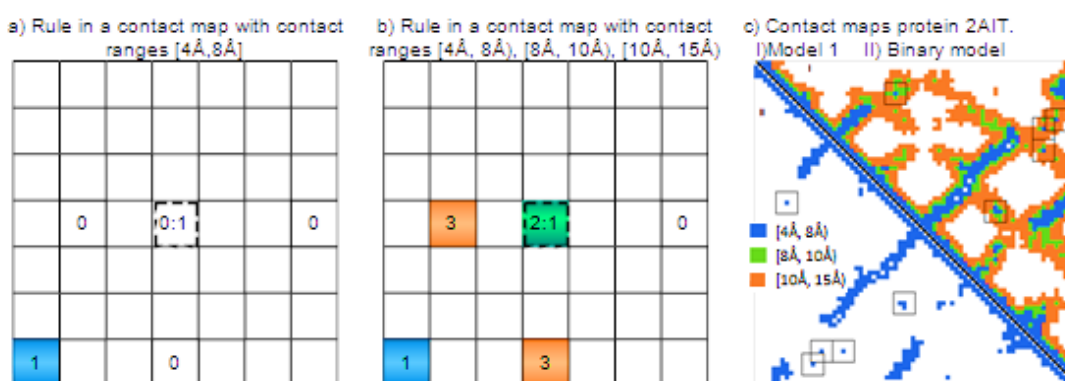


Figure 5. The rule in a 2-state contact map vs. 4-state contact map for Model 1.

Cells with values are those that are involved in the rule. The cell with dashed lines indicates a cell for which in the native contact map appears a native contact. We assigned: value 0 to non-contact; value 1 to contact in the range [4 Å, 8 Å]; value 2 is to contact in the range [8 Å, 10 Å]; value 3 to a contact in the range [10 Å, 15 Å]. c) Contact map for protein 2AIT; Upper triangular is the contact map representation using Model 1; Lower triangular is the binary contact map. Boxes denote long-range contacts between secondary structures for both representations.

In Figure 5.c), we illustrate the contact maps for protein 2AIT (chain A). In the upper triangular matrix, we show contacts using contact ranges used for Model

1. In the lower triangular matrix, we display contacts using the traditional binary contact map contact range. Black framed boxes in Figure 5.c) indicate contacts between amino acid pairs, with sequence separation greater than twenty amino acids and between secondary structures. Long-range contacts are usually hardest to predict because they are scarce (Wang and Xu, 2013). For the binary case, the prediction is more challenging since relevant contacts are isolated in the contact maps. For multiclass contact maps, the same contacts are surrounded by contacts of the other classes, which allow us to learn some patterns that are represented as no contacts.

3.1 Contact Refinement

We define contact refinement as the process of evolving a contact map for an unfolded state of the protein through its folded state.

We test the predictive power of DT for models 1 and 5, in a context related to contact prediction. We use a DT to a contact map of a protein's unfolded state and apply DT rules to refine it to its native state. Additionally, we test a DT (Model o) trained on contact ranges $[4 \text{ \AA}, 8 \text{ \AA})$, $[8 \text{ \AA}, 13 \text{ \AA})$, and $[13 \text{ \AA}, 19 \text{ \AA})$, similar to the ones used in (Walsh et al., 2009). Model o provides a baseline to compare our results to those of a previously proposed multiclass contact map.

The dataset for this testing comprises ten proteins, and for each one, we take the first 50 structures in a trajectory. Proteins in the test set belong to the three main structural classes (α , β , $\alpha+\beta$). For each structure in the trajectory, we predict a new

contact map and the contacts are evaluated similarly to contact prediction in CASP (Monastyrskyy et al., 2014). In Figure 6 we show results for contact map refinement. We averaged precision values for: short-range contacts, sequence separation in range [6, 12); medium-range contacts, separation in range [12, 24); and long-range contacts, sequence separation in range [24, ∞). We also include Matthews correlation coefficient (MCC) (Monastyrskyy et al., 2014), for contacts with sequence separation greater than six amino acids.

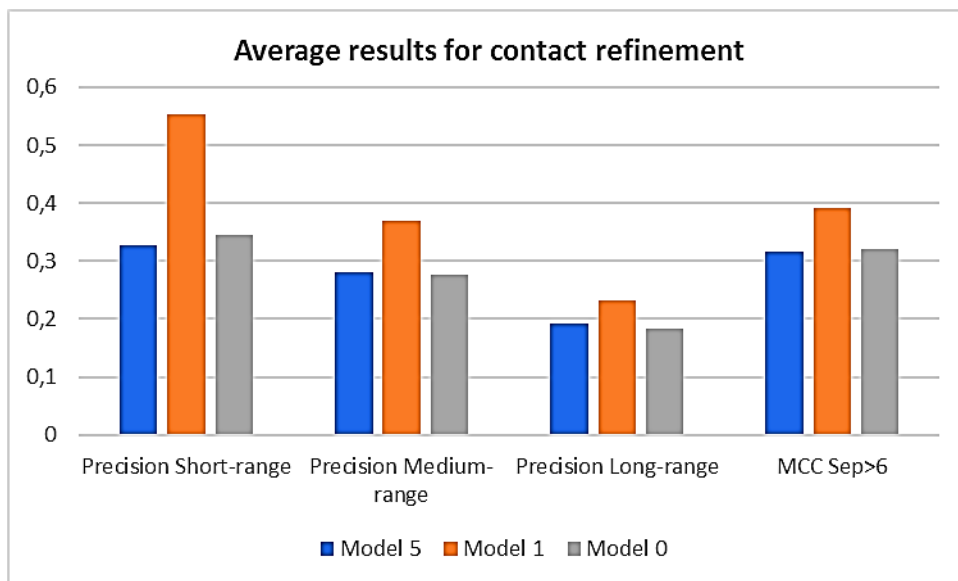


Figure 6. Average results for contact map refinement.

We compared the three models by using MCC for contacts between amino acids with sequence separation greater than 6. In Table 2 we show results for each model in each protein. Column Baseline in Table 2 stands for a DT trained with the common contact range of 8 Å, for binary contact maps.

In Table 3 we show results of Wilcoxon test for comparisons of Model 1 vs. models 0 and 5, as well as for models 0 and 5. In our case, Wilcoxon statistic critical value (W_c) is 3 (significance level 0.005). Wilcoxon statistic (W) for each comparison is the minimal value from R^+ and R^- . We calculate ranks using, MCC as a performance measure for contacts between amino acids with sequence separation greater than 6 (Table 2).

Table 2. MCC for each protein trajectory in the test set for contact refinement. *Baseline column displays values for a model built for the classical binary contact map representation (contact range $[4 \text{ \AA}, 8 \text{ \AA}]$).*

Protein	Structural Class	MCC, Separation > 6			
		Model 5	Model 1	Model 0	Baseline
1SHG	β	0.1132	0.1557	0.1272	0.1457
1NYF	β	0.6636	0.7137	0.6909	0.4789
1SRM	β	0.6074	0.6641	0.6253	0.4888
2AIT	β	0.5659	0.6507	0.5866	0.4573
2YCC	α	0.0217	0.0874	0.0270	0.0445
351C	α	0.2762	0.4462	0.2584	0.1716
3CHY	$\alpha+\beta$	0.1024	0.1753	0.0949	0.0764
1PBA	$\alpha+\beta$	0.3516	0.4319	0.3579	0.2531
1UBI	$\alpha+\beta$	0.1827	0.2472	0.1795	0.1391
1COA	$\alpha+\beta$	0.2818	0.3315	0.2653	0.1975

In Table 3, we can see that Wilcoxon test rejects the Null hypothesis when comparing Model 1 vs. Model 0 and Model 1 vs. Model 5 because $W \leq W_c$ and $p\text{-value} \leq 0.005$. Therefore, we can say confidently that Model 1 outperforms models 0 and 5. For Model 5 vs. Model 0 comparison, we cannot reject the Null hypothesis.

Table 3. Wilcoxon test for contact refinement comparison.

The R+ value accounts for the sum of ranks where the first model outperforms the second model. The R- value accounts for the sum of ranks where the second model outperforms the first model.

Comparison	R+	R-	p-value	W
Model 1 vs. Model 0	55	0	0.002	0
Model 1 vs. Model 5	55	0	0.002	0
Model 5 vs. Model 0	19	36	>0.2	19

4 Conclusions

Protein contact maps are a useful tool for contact prediction and tertiary structure prediction. Nevertheless, prediction of binary contact maps is a problem of difficult solution, mainly because long-range contacts are scarce. Dependencies between amino acids with long sequence separation are particularly hard to learn because they are scarce. We develop a process that allows us to find a multiclass contact map representation capable of learning interactions related to protein contacts at short, medium and long ranges. We focused our approach on the quality of DT rules assessed in a context related to contact prediction, where contacts for residues with sequence separation greater than six amino acids are especially relevant in the quality of the final contact map. In Table 2, we verify that the contact ranges combination $\{[4 \text{ \AA}, 8 \text{ \AA}), [8 \text{ \AA}, 10 \text{ \AA}), [10 \text{ \AA}, 15 \text{ \AA})\}$ is optimal when compared to the multiclass contact map described in (Walsh et al., 2009) and the baseline model (contact range $< 8 \text{ \AA}$).

Our main goal was to find a multiclass contact map that was capable of capture details of protein contact formation in folding trajectories, in a better way

that the traditional representation of binary contact maps. With Model 1 we achieved this. In Table 2, we reported results for contact refinement using both models, and for each case Model 1, could obtain more correct contacts and no-contacts than the binary representation.

Though our approach allows us to find an optimal model; this is not a definitive solution. Evaluating our top ten models by Nemenyi's test and CD, we do not obtain significant differences. However, Model 1 show significantly different behavior than another model previously used in the state of the art for contact prediction (see Table 2). Our next step is to use this multiclass contact map to train a novel residue-residue contact predictor, exploiting the capability of this model to capture information of native contact formation. This novel predictor will be developed using the same methodology of GACAI-PCP.

5. References

- Adhikari, B, Nowotny, J, Bhattacharya, D, Hou, J, & Cheng, J. (2016). ConEVA: a toolbox for comprehensive assessment of protein contacts. *BMC Bioinformatics*, 17(1), 517. <http://doi.org/10.1186/s12859-016-1404-z>
- Alcalá-Fdez, J, Sánchez, L, García, S, del Jesús, MJ, Ventura, S, Garrell, JM, ... others. (2009). KEEL: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, 13(3), 307–318. <http://doi.org/10.1007/s00500-008-0323-y>
- Caruana, R, & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning*, (June 2006), 161–168. <http://doi.org/10.1145/1143844.1143865>
- Da Silveira, CH, Pires, DE V, Minardi, RC, Ribeiro, C, Veloso, CJM, Lopes, JCD, ... Santoro, MM. (2009). Protein cutoff scanning: A comparative analysis of

- cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins: Structure, Function and Bioinformatics*, 74(3), 727–743. <http://doi.org/10.1002/prot.22187>
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1), 1–30. <http://doi.org/10.1016/j.jecp.2010.03.005>
- Duarte, JM, Sathyapriya, R, Stehr, H, Filippis, I, & Lappe, M. (2010). Optimal contact definition for reconstruction of contact maps. *BMC Bioinformatics*, 11, 283. <http://doi.org/10.1186/1471-2105-11-283>
- Fayyad, U, Piatetsky-Shapiro, G, & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3), 37. <http://doi.org/10.1609/aimag.v17i3.1230>
- Frank, E, Hall, M, Holmes, G, Kirkby, R, Pfahringer, B, Witten, IH, & Trigg, L. (2005). WEKA: A Machine Learning Workbench for Data Mining. In *Data Mining and Knowledge Discovery Handbook* (pp. 1305–1314). Springer US. http://doi.org/10.1007/o-387-25465-X_62
- Kukic, P, Mirabello, C, & Tradigo, G. (2014). Toward an accurate prediction of inter-residue distances in proteins using 2D recursive neural networks. *BMC ...*, 15, 6. <http://doi.org/10.1186/1471-2105-15-6>
- Monastyrskyy, B, D'Andrea, D, Fidelis, K, Tramontano, A, & Kryshchuk, A. (2014). Evaluation of residue-residue contact prediction in CASP10. *Proteins: Structure, Function and Bioinformatics*, 82(SUPPL.2), 138–153. <http://doi.org/10.1002/prot.24340>
- Osei-Bryson, K-M. (2004). Evaluation of decision trees: a multi-criteria approach. *Computers & Operations Research*, 31(11), 1933–1945. [http://doi.org/10.1016/S0305-0548\(03\)00156-4](http://doi.org/10.1016/S0305-0548(03)00156-4)
- Provost, F, Fawcett, T, & Kohavi, R. (1998). The Case Against Accuracy Estimation for Comparing Induction Algorithms. In J. Shavlik (Ed.), *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 445–453). San Francisco, CA: Morgan Kaufmann Publishers.
- Quinlan, JR. (1993). *C4. 5: Programs for machine learning*. Elsevier.
- Sathyadevan, S, & Nair, RR. (2015). Comparative Analysis of Decision Tree Algorithms: ID3, C4.5 and Random Forest. In *Computational Intelligence in Data Mining-Volume 1* (Vol. 1, pp. 549–562). Springer India. <http://doi.org/10.1007/978-81-322-2205-7>

- Schaarschmidt, J, Monastyrskyy, B, Kryshtafovych, A, & Bonvin, AMJJ. (2017). Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. *Proteins: Structure, Function, and Bioinformatics*, (October 2017), 51–66. <http://doi.org/10.1002/prot.25407>
- Sharma, N, Bajpai, A, & Litoriya, R. (2012). Comparison the various clustering algorithms of weka tools. *International Journal of Emerging Technology and Advanced Engineering*, 2(5), 73–80.
- Stiglic, G, Kocbek, S, Pernek, I, & Kokol, P. (2012). Comprehensive Decision Tree Models in Bioinformatics. *PLoS ONE*, 7(3), 1–13. <http://doi.org/10.1371/journal.pone.0033812>
- Thomas, S, Song, G, & Amato, NM. (2005). Protein folding by motion planning. *Physical Biology*, 2(4), S148–S155. <http://doi.org/10.1088/1478-3975/2/4/S09>
- Torres-Jimenez, J, & Izquierdo-Marquez, I. (2014). Survey of covering arrays. *Proceedings - 15th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2013*, (1), 20–27. <http://doi.org/10.1109/SYNASC.2013.10>
- Torres-Jimenez, J, Izquierdo-Marquez, I, Kacker, RN, & Richard Kuhn, D. (2015). Tower of covering arrays. *Discrete Applied Mathematics*, 190–191, 141–146. <http://doi.org/10.1016/j.dam.2015.03.010>
- Vassura, M, Margara, L, Di Lena, P, Medri, F, Fariselli, P, & Casadio, R. (2008). Reconstruction of 3D structures from protein contact maps. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4463, 357–367. http://doi.org/https://doi.org/10.1007/978-3-540-72031-7_53
- Walsh, I, Baù, D, Martin, AJM, Mooney, C, Vullo, A, & Pollastri, G. (2009). Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. *BMC Structural Biology*, 9(5), 1–20. <http://doi.org/10.1186/1471-2105-10-195>
- Wang, Z, & Xu, J. (2013). Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics*, 29(13), 266–273. <http://doi.org/10.1093/bioinformatics/btt211>

Chapter VI.

Conclusions and Future Work

In protein tertiary structure determination PCP is currently an habitual step. We proved that CAs can help in PCP. Our approach identifies CAs that are suited for the search of feasible local arrangement of contacts. The neighborhoods of the CAs for PCP obtained by our machine learning based approach (GACAI-PCP) showed an irregular shape. The several irregular neighborhood shapes capture the relationships which improves the identification of real contacts from false positives. We used two neighborhoods in this work: In Chapter III we contrasted one of our irregular neighborhoods to standard regular ones; and in Chapter IV we compared predictions made by our best neighborhood against the results of the CASP12 top10 contact predictors. However, we are planning to explore several other promising neighborhoods.

Currently, the prominent machine learning approach to PCP is deep learning. However, our CA based approach was successful in improving predictions by MetaPSICOV. The main reason for the improvement is the flexibility that provides the search of the CA's neighborhood, instead of using a predefined neighborhood. Compared to the state of art, our results were outstanding for FL but not competitive for RL. For FL, CAs can exploit the effect of global coordination.

That is not possible for RL; the neighborhoods increase the proportion of non-contacts, making the contact map less similar to a real one. To improve our performance in RL, we are exploring how to determine a better distribution of probability where true contacts obtain higher probability.

Friedman's and Nemenyi's tests provided us with better insights into the compared predictors. Friedman's test compares predictions protein by protein instead of averaging performance measures. In CASP predictors comparison, Nemenyi's test allowed us to conclude that CAs perform in a similar way to well-known top predictors. Even though CA based prediction revealed higher average precision, the Friedman's comparison performed target by target, expose that performance lacks statistical significance to dominate the best contact predictors in CASP. In order to compare performance visually, we used dominance graphs as a tool that eases the visualization of dominance relationships between each pair of predictors. In a dominance graph is easier to distinguish groups of predictors with similar performance from those that are significantly different. We are planning to develop a tool that creates dominance graphs for comparison of predictors.

We used the methodology based on Friedman's and Nemenyi's tests and dominance graphs in two scenarios. First, it allowed us to select the from the top₁₀ CASP₁₂ predictors the most appropriate one to use in our training dataset. Second, we used it to compare our results in the context of the top₁₀ performers in CASP₁₂. We selected the top₁₀ from CASP₁₂ official results using a pondered average of all CASP₁₂ performance measures to all participating contact predictors.

GACAI-PCP is a tool that allows to identify new CAs for PCP. It was developed in Python language and is publicly available in GitHub (<https://github.com/nediazo1/GACAI-PCP>). GACAI-PCP is a framework that permits discovering new CAs from diverse datasets of initial conditions. Initial conditions datasets generation for GACAI-PCP is computationally expensive, so that we planned to have an automated tool to train several CAs for PCP as new datasets are available. In the future work we want to use GACAI-PCP to identify CA's tailored for improvement of predictions from other PCP tools. The results described in Chapter IV are based on MetaPSICOV predictions and are improved by applying GACAI-PCP. Our next step is to follow the same approach for RaptorX-Contact predictions.

GACAI-PCP is currently at version 1.0 with a full set of functionalities implemented. We are exploring several improvements to reduce the computational time for CA identification. One of these options is about the representation of initial conditions. We are designing a tuple-based representation of global configurations to reduce the dynamical exploration of lattices cell by cell.

We obtained a multiclass contact map that enhances the scarceness of information around long range isolated native contacts (Diaz & Tischer, 2019). The methodology we described in Chapter V, uses a huge set of contact ranges and compares them by deep evaluation of decision trees. We obtained an optimal set contact ranges that allowed us to propose a novel multiclass contact map. We verified these optimal ranges and obtained richer neighborhoods for contacts

isolated in binary contact maps. Our multiclass contact map proposal seems promising, but it requires a generator of initial conditions to be implemented in GACAI-PCP. We are planning to implement a generator of initial conditions from structural features generated by a novel predictor such as NetSurfP (Klausen et al., 2019).

We are exploring the idea of using new states based on contacting amino acid pairs. To apply our methodology, we have to search an optimal neighborhood, and determine support and frequency of contact.

Finally, as a result of an internship developed in the Computational Biology & Data Mining Group (Mainz University, Germany), we are developing new tools aiming to the application of our approach to tandem repeat proteins structure analysis and prediction.

References

Diaz, N., & Tischer, I. (2019). Data mining approach for selecting contact ranges for multiclass protein contact maps.

Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K., Jurtz, V. I., Sønderby, C. K., ... Marcatili, P. (2019). NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*. <https://doi.org/10.1002/prot.25674>