X ENCONTRO ACADÊMICO
MODELAGEM COMPUTACIONAL

CONTACT ASSISTED PROTEIN STRUCTURE PREDICTION USING A GENETIC ALGORITHM WITH PHENOTYPIC CROWDING

Karina B. Santos

Gregório K. Rocha

Fábio L. Custódio

Laurent E. Dardenne

karinabs@lncc.br

gregorio@lncc.br

flc@lncc.br

dardenne@lncc.br

National Laboratory for Scientific Computing (LNCC)

Av. Getúlio Vargas, 333 - Quitandinha, 25651-075, RJ, Petrópolis, Brazil

Abstract. *The use of protein contact maps in protein structure predictions methodologies has proved promising and was highlighted in the last CASP editions (CASP10 and CASP11). The goal of this work is the investigation and the development of strategies to use the contact map's information to improve the quality of the protein models predicted by the program GAPF (Genetic Algorithm for Protein Folding). GAPF uses a steady-state genetic algorithm and a parental replacement by phenotypic crowding. This strategy provides a better exploration of the energy hypersurface allowing the identification of multiple minima and preserving the population diversity of the generated structures. Concerning the GAPF predictive ability, it's algorithm needs some improvement in prediction of proteins that have complex structures, with several β-strands or with more than 100 residues in the sequence. To overcome these difficulties, we incorporated the protein contact map information into the GAPF program by developing and implementing a new term in the fitness function, associated with the contacts restrictions. The methodology was evaluated using a test set of six proteins belonging to different classes (i.e., preferably α, α/β and preferably β). The results show that using contact map knowledge associated with a new term in GAPF's fitness function increased the capacity of the algorithm to obtain better protein models. For almost all target proteins the best-predicted models had their RMSD below 4.30 Å and GDT-TS above 57%.*

Keywords: *Protein structure prediction, protein contact map, genetic algorithm*

1 INTRODUCTION

The protein structure prediction (PSP) is still being one of the most important purposes of computational molecular biology research area, and its goal is to discover the native three-dimensional (3D) protein structure using the information stored in the amino acid sequence (Michel et al., 2014). Since 1994, with the CASP event (*The Critical Assessment of Techniques for Protein Structure Prediction*), it has been possible to note a significant advance in computational methods to predict protein structures (Dill and MacCallum, 2012). In the last two CASP editions (CASP10 and CASP11), the use of the knowledge about predicted sets of residue pairs that are closer to each other in the native protein structure, *i.e.* that are in "contact", combined with *de novo* PSP strategies showed promising results (Taylor et al., 2014; Kosciolek and Jones, 2014; Kinch et al., 2016a). The information about which residues should be in contact, represented as a contact map, is commonly used in the form of a distance restraint term in the fitness function (Joo et al., 2016; Kim et al., 2014; Kosciolek and Jones, 2015; Wang et al., 2016). During CASP10 experiment, models predicted using contact maps information (contact-assisted predictions) had GDT-TS (*Global Distance Test Total Score*) scores 40 points higher than the best-predicted models, for the same proteins, without contact information (contact-unassisted prediction) (Taylor et al., 2014). The contact-assisted structure modeling experiments in CASP11 significantly improved structure predictions, with a correct *de novo* prediction of proteins with much larger domains than it was possible to be modeled on previous CASP editions (Kinch et al., 2016b).

In a contact map, two residues are in "contact" if the Euclidean distance between its C-β is less than or equal to a predefined distance threshold, usually, 8 Å (Wang and Xu, 2013). Residue–residue contacts can be predicted from analysis of correlated evolution mutations obtained from multiple sequence alignments of a sufficiently large number of homologous proteins (Michel et al., 2014).

This work proposes the development of a strategy to use the information from protein contact maps to improve the predictive ability of PSP methodologies. We also focuses on the GAPF PSP program, a phenotypic crowding-based steady-state genetic algorithm, *i.e.*, an algorithm that work with a population of candidate solutions whose parental replacement method forces competition between the most similar individuals. GAPF also uses protein fragment libraries and secondary structure predictions (obtained from a PSIPRED horizontal format file, Jones (1999)) to guide the protein structure prediction (Custódio et al., 2010, 2014; Rocha et al., 2015; Santos et al., 2015).

2 METHODS

2.1 The Fitness Function

A coarse-grained representation is used in GAPF, where the side chain atoms are replaced by a super-atom located at its geometric center and the fitness function is based on the energy from the interaction between the atoms of the protein. This energy is calculated using a molecular force field made up of a dihedral potential ($Dihed$) from GROMOS96 force field (van Gunsteren et al., 1996), four hydrogen bonds terms, wich one is fully independent of secondary structure prediction (HB_{dist_NHO}) and tree dependent on secondary structure prediction (HB_{hx} - for α-helix; HB_{st} - for β-sheet; HB_{att_st} - for attracting the backbone dipoles of β-strands),

an atomic repulsive term (A_{rep}) and a hydrophobic compaction term (C_{pk}) (Rocha, 2015). The proposed strategy consists of (I) incorporate into the GAPF fitness function a residue-residue contact term to model the knowledge contained in the protein contact map generated for the protein being folded (II) assess its influence on the final quality of the predicted models.

Contact Map Term. Typically, for each amino acid pair described in the contact map, there is a confidence value (γ) associated with the probability of such residues being in contact. Often this confidence value ranges between 0-1.

The Contact Map Term proposed in this work deals with the residue-residue contacts described in the contact map in the form of distance restraints. Thus, for each protein model generated, the distance of two residues (($\sigma(a_i, a_j)$) is calculated on Cβ atoms (Cα for Glycine). According to the distance value, a contributing value ($\lambda_{(ai,aj)}$) is computed (Eq. (1)). Equation 1 is applied for every residue-residue pair found in the contact map, and each contribution value is assigned to the contact term score (E_{rr}) (Eq. (2)).

$$\lambda_{(ai,aj)} = \begin{cases} \gamma * 1000 & if 2.0 \leq \sigma(a_i, a_j) \leq 8 \\ \\ \gamma/2 * 500 & if 8 < \sigma(a_i, a_j) \leq 10 \end{cases} \tag{1}$$

$$E_{rr} = \sum_{rr-pairs} \lambda_{(ai,aj)} \tag{2}$$

$$E_{total} = Dihed + HB_{dist\_NHO} + HB_{hx} + HB_{st} + HB_{att\_st} + A_{rep} + C_{pk} - E_{rr} \tag{3}$$

Thus, in the Contact Map Term, the more residues are respecting the distance constraints described in the contact map, the greater their contribution to the Fitness Function.

Equation 3 shows the complete fitness function used in GAPF program.

2.2 "Experimental" Contact Map

To validate Contact Map Term embedded in the GAPF fitness function, for each target protein, a "experimental" contact map was generated using the PDB file from the respective experimental protein structure. In this case, for each residue pair with at least three residue separation along the protein sequence, the Euclidean distance between its C-β (Cα for Glycine) was calculated. So, the "experimental" contact map consists of all residue pairs that have distance value less than or equal to 8 Å. The threshold value assigned to each residue pair in the "experimental" contact map is 1.0. Table 1 shows the total number of contact pairs described in the "experimental" contact map for each target protein.

2.3 Test Set

To assess the effects of Contact Map Term in the quality of predicted structures, a set of six proteins with 3D experimentally determined structure, with 56-108 amino acid residues in the sequence, was used (Table 1).

Table 1: Test set and contact map

ID (PDB/CASP)	Class	Sequence Lenght*	"Experimental" contact pairs
3FIL	α/β	56	128
2N2U (CASP ID - T0773-D1)	α/β	67	189
——** (CASP ID - T0820-D1)	α	90	186
1FNA	β	91	212
2MQ8 (CASP ID - T0769-D1)	α/β	97	295
4Q53 (CASP ID - T0766-D1)	α/β	108	317

*Number of amino acids in sequence.
**Protein without identifier in the Protein Data Bank. Experimentally determined structure obtained directly from CASP11 server (http://www.predictioncenter.org/casp11/).

2.4 Parameters and Tests Configurations

The parameters for the GAPF program during the tests were: population size of 200 individuals, a maximum number of 2,000,000 fitness function evaluations and 20 independent runs per sequence.

At the end of each run, the quality of the 200 generated structures was evaluated after comparing it to the reference structure using the RMSD (*Root-mean-square deviation of atomic positions*) and the GDT-TS.

The quality of the models predicted using the contact map strategy (Contact Map Protocol) was compared with those predicted by the GAPF running with the original protocol, i.e., without the contact map term.

3 RESULTS AND DISCUSSION

3.1 Impact of the use of Contact Maps

The use of distance constraints imposed by contact maps allowed the algorithm to generate models with structural conformations closer to that found in the native protein. That's because satisfying these restrictions, increases the probability of correctly reproduce the 3D arrangement of the predicting protein.

Table 2 shows the RMSD and GDT-TS values of the best models and of the best energy models predicted in 20 independent runs of GAPF program. According to Kryshtafovych et al. (2005), are considered good predictive ability RMSD values bellow 4.0 Å. Martí-Renom et al. (2000) believes that RMSD values between 3.0 and 4.0 Å means that the model can correctly

represent the protein folding. RMSD values above 5.0 indicate little information about the three-dimensional protein structure (Unger, 2004). A GDT-TS value equal to 100% means that the distance between the structural alignments is no more than 1.0 Å. GDT-TS values above 50% indicate good predictive ability, values below 50% infer that proteins do not have the same folding, and scores between 10-20% represent only random overlaps between the structures (Xu and Zhang, 2010).

Table 2: Comparison between the original and the contact map protocols using the best RMSD and GDT-TS values obtained in 20 independent runs of the GAPF program.

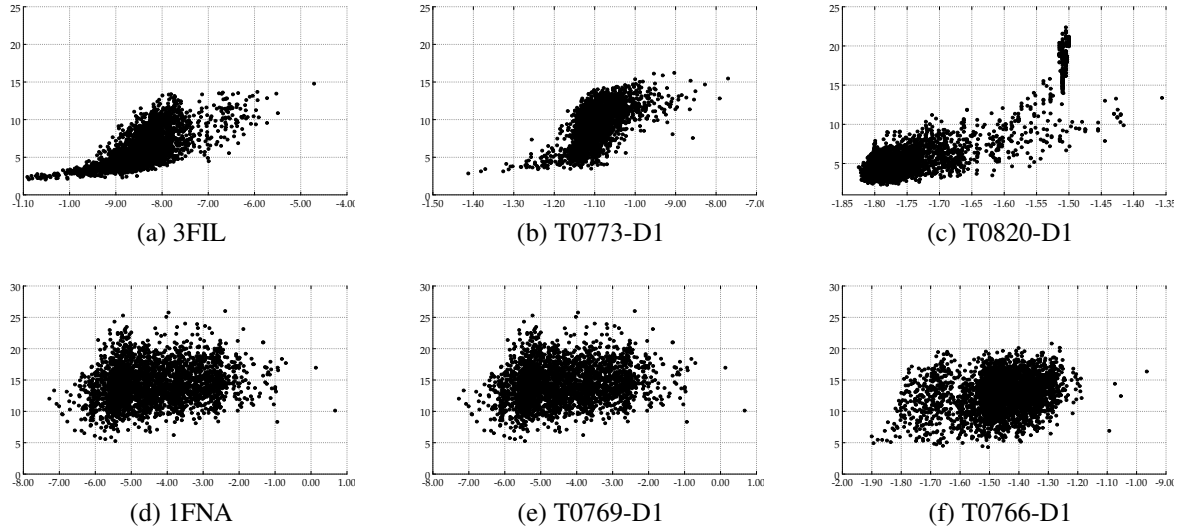
	Original Protocol				**Contact Map Protocol**			
ID (PDB/CASP)	RMSD (Å)		GDT-TS (%)		RMSD (Å)		GDT-TS (%)	
	Best model	Best energy	Best model	Best energy	Best model	Best energy	Best model	Best energy
3FIL	5.33	11.59	51.82	37.73	2.03	2.38	58.64	51.82
T0773-D1	7.40	11.85	41.79	24.25	2.84	2.84	69.03	69.03
T0820-D1	9.60	13.78	40.28	26.94	2.28	4.53	75.56	57.22
1FNA	8.77	15.52	21.43	12.09	5.26	12.02	17.86	13.74
T0769-D1	7.78	14.61	42.01	20.88	3.91	7.15	60.82	55.41
T0766-D1	10.63	17.45	31.25	21.99	4.29	6.00	57.18	52.55

The relationship between the energy and RMSD values of the models generated using the Contact Map Protocol is summarize in Figure 1. Our data shows that, for most targets, the use of contact term in the fitness function provides best structural quality models (best RMSD value) among those of lower energy (best energy). One common behavior observed from all sequences is that models with same energy values can present very distinct RMSD values, this illustrates the multiple minima characteristic of the GAPF algorithm.

Table 3 indicates the average of the best models in terms of RMSD and GDT-TS values and in terms of energy in each run. In Tables 2 and 3, it can be observed that when executed using the original protocol, the quality of solutions decreases for longer sequences, e.g., proteins with more than 90 residues in the sequence. Furthermore, the use of the "experimental" contact maps improved the results generating models for the same sequences with RMSD bellow 4.30 Å and GDT-TS value above 57% (Table 2).

Figure 2 shows the best models generated for each target using the original protocol and the contact map protocol. The compelling advantage of the new approach is the improvement of the predictive ability for all secondary structure type, with GAPF being able to generate α-helix and β-strands in the correct place. Even for 1FNA protein, that had RMSD and GDT-TS values representative of random overlaps between structures in both experiments (Original and Contact Term protocols), its 3D representation shows the formation of the β-barrel (Fig. 3).

For T0769-D1 protein, the shortened β-sheet can be explained by the secondary structure predicted by the PSIPRED program (Jones, 1999) that shows some differences between the



(a) 3FIL (b) T0773-D1 (c) T0820-D1

(d) 1FNA (e) T0769-D1 (f) T0766-D1

Figure 1: Results for 20 independent runs for each sequence. Energies in the x-axis are in Kcal/mol and are divided by 10^4 (lower is better). Structural variations (measured with the RMSD) of the predicted models relative to the native reference structure for each protein in the y-axis are in Å.

Table 3: Comparison between the original and the contact map protocols using the average of the best RMSD and GDT-TS values obtained in 20 independent runs.

ID (PDB/CASP)	Original Protocol				Contact Map Protocol			
	RMSD (Å)		GDT-TS (%)		RMSD (Å)		GDT-TS (%)	
	Best model	Best energy	Best model	Best energy	Best model	Best energy	Best model	Best energy
3FIL	6.41	9.55	43.59	36.61	2.74	3.46	55.97	51.61
T0773-D1	8.68	11.45	38.17	30.91	3.57	4.02	63.26	58.99
T0820-D1	11.44	14.91	34.70	26.38	2.69	3.96	70.50	62.59
1FNA	11.41	14.63	16.77	13.18	7.18	10.46	16.34	13.40
T0769-D1	10.79	14.74	32.40	25.54	4.90	5.60	56.03	52.21
T0766-D1	11.86	16.37	24.52	20.37	5.75	8.43	48.85	41.75

secondary structure assigned by DSSP (Kabsch and Sander, 1983) (Fig. 4), and also shows low confidence values at the end of each predicted secondary structure type which could have hindered a better fragment insertion.

4 CONCLUSION

The use of contact maps in the form of distance constraints is a promising strategy for *de novo* PSP methodologies. Using "experimental" contact maps composed by real contacts found in the native structure permitted to assess the potential of this technique, isolating the problem of the correctness of the contacts assignement.

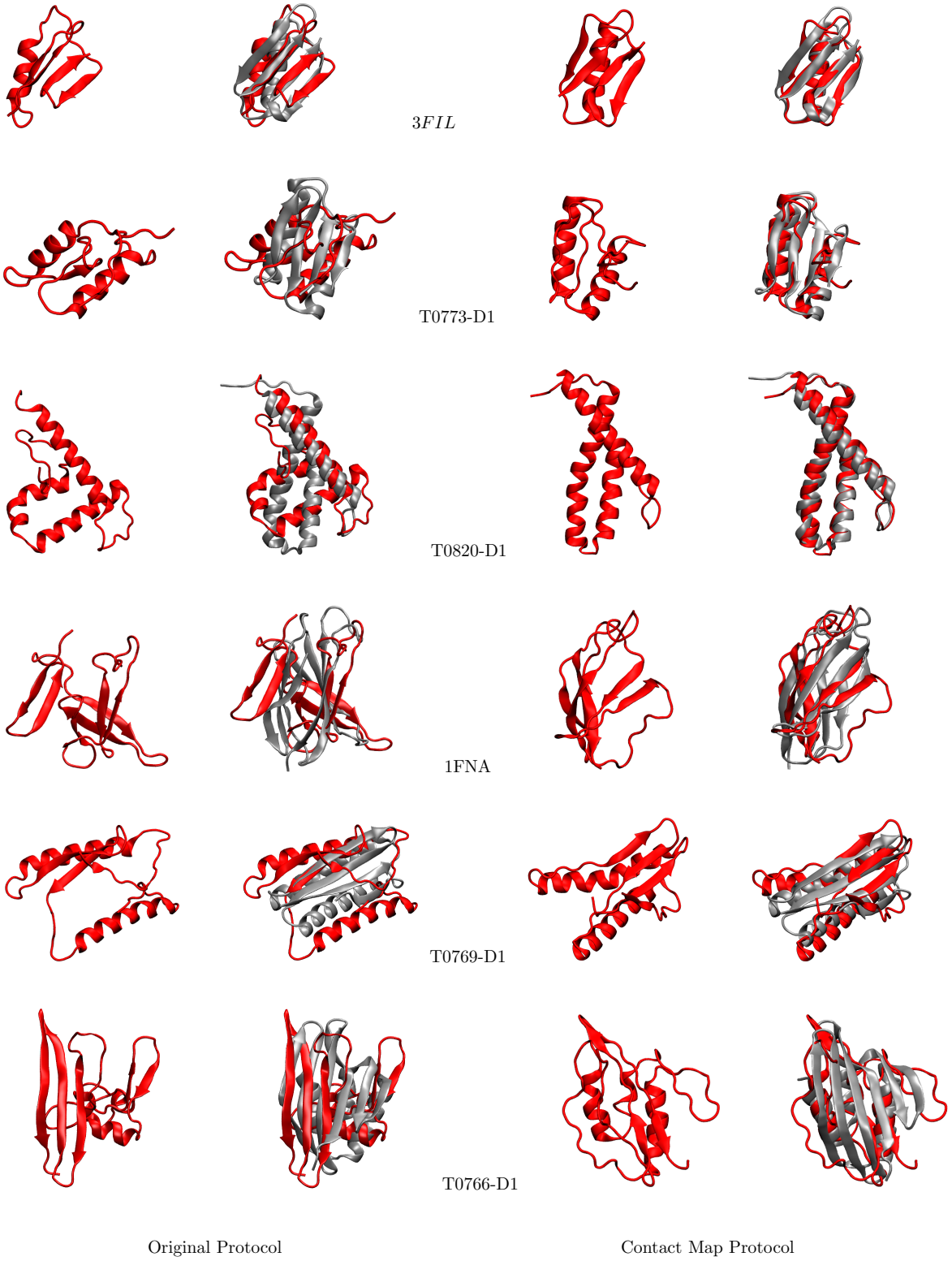
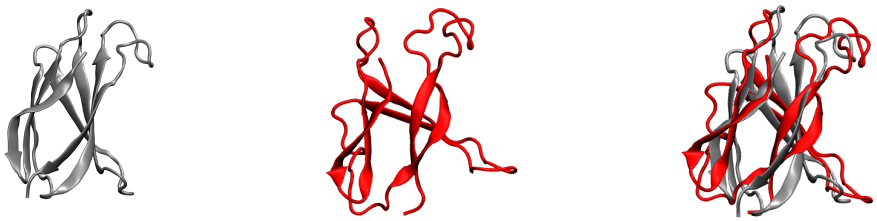


Figure 2: Comparison between the best models generated with GAPF program (red structures). On the left are the models generated using the original GAPF protocol, on the right created using Contact Map Term. Each model was aligned with the experimentally determined structure that was used as reference for calculating the RMSD and GDT-TS (silver structures).



(a) (b) (c)

Figure 3: 1FNA protein seen for another angle that emphasizing the formation of β-barrel.(a) Experimentally determined structure, (b) Model generated using Contact Map Protocol and (c) Alignment between experimental and predicted structure.

```
   Conf: 92589999971287405899999988899999986189844688764069639999985O
PSIPRED: CCEEEEEEEEECCCCHHHHHHHHHHHHHHHHHHHHHHCCCCEEEEEEECCCEEEEEEEEC
   DSSP: CEEEEEEEEEECCCCHHHHHHHHHHHHHHHHHHHHHHCCCCEEEEEEEEEECCEEEEEEEEE
     AA: MLTVEVEVKITADDENKAEEIVKRVIDEVEREVQKQYPNATITRTLTRDDGTVELRIKVK

   Conf: 68367899999999999999963399973565321 9
PSIPRED: CCCHHHHHHHHHHHHHHHHHHHHCCCCCEEEEEECC
   DSSP: ECCHHHHHHHHHHHHHHHHHHHCCCCCCEEEEEECC
     AA: ADTEEKAKSIIKLIEERIEEELRKRDPNATITRTVRT
```

Figure 4: T0769-D1 - Psipred Secondary structure prediction compared with DSSP determined secondary structure. The letters 'C', 'H' and 'E' means Coil, helix-α and β-sheet respectively. In red are highlighted the disagreement secondary structures. 'Conf' term refers to confidence in this secondary structure prediction by PSIPRED.

However, when the protein 3D structure is unknown, it is necessary to use other techniques to predicted residue-residue contacts. Protein contact maps can be predicted using coevolution-based methods. As an example of programs to generate contact maps are METAPSICOV (Jones et al., 2015) and GREMLIN (Ovchinnikov et al., 2014). METAPSICOV and GREMLIN contact maps, usually show the probability of all residues being in contact with all residues. In this case, these maps can contain various incorrect contact predictions. The lack of full confidence in the prediction of such contacts, points as very necessary to develop some strategies to filter and enhance the information of predicted contacts. In this way, the next step towards the improvement of the use of a Contact Map Term should be the development and introduction of a contact map filter that identifies which predicted contacts, from METAPSICOV, GREMLIN or other contact map predictor, are really valuable or not for PSP.

ACKNOWLEDGMENT

The authors would like to thank the support from CAPES and FAPERJ (grant n°. E26/010.001229/2015).

REFERENCES

Custódio, F. L., Barbosa, H. J., and Dardenne, L. E. Full-atom ab initio protein structure prediction with a genetic algorithm using a similarity-based surrogate model. *IEEE Congress on Evolutionary Computation*, pages 1–8, 2010.

Custódio, F. L., Barbosa, H. J., and Dardenne, L. E. A multiple minima genetic algorithm for protein structure prediction. *Applied Soft Computing*, 15:88–99, 2014.

Dill, K. A. and MacCallum, J. L. The protein-folding problem, 50 years on. *Science*, 338 (6110):1042–1046, 2012.

Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, 292(2):195–202, 1999.

Jones, D. T., Singh, T., Kosciolek, T., and Tetchner, S. Metapsicov: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, 31(7):999–1006, 2015.

Joo, K., Joung, I., Cheng, Q., Lee, S. J., and Lee, J. Contact-assisted protein structure modeling by global optimization in casp11. *Proteins: Structure, Function, and Bioinformatics*, 2016.

Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

Kim, D. E., DiMaio, F., Yu-Ruei Wang, R., Song, Y., and Baker, D. One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins: Structure, Function, and Bioinformatics*, 82(S2):208–218, 2014.

Kinch, L. N., Li, W., Monastyrskyy, B., Kryshtafovych, A., and Grishin, N. V. Assessment of casp11 contact-assisted predictions. *Proteins: Structure, Function, and Bioinformatics*, 2016a.

Kinch, L. N., Li, W., Monastyrskyy, B., Kryshtafovych, A., and Grishin, N. V. Evaluation of free modeling targets in casp11 and roll. *Proteins: Structure, Function, and Bioinformatics*, 2016b.

Kosciolek, T. and Jones, D. T. De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PloS one*, 9(3):e92197, 2014.

Kosciolek, T. and Jones, D. T. Accurate contact predictions using covariation techniques and machine learning. *Proteins: Structure, Function, and Bioinformatics*, 2015.

Kryshtafovych, A., Milostan, M., Szajkowski, L., Daniluk, P., and Fidelis, K. Casp6 data processing and automatic evaluation at the protein structure prediction center. *Proteins: Structure, Function, and Bioinformatics*, 61(S7):19–23, 2005.

Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F., and Šali, A. Comparative protein structure modeling of genes and genomes. *Annual review of biophysics and biomolecular structure*, 29(1):291–325, 2000.

Michel, M., Hayat, S., Skwark, M. J., Sander, C., Marks, D. S., and Elofsson, A. Pconsfold: improved contact predictions improve protein models. *Bioinformatics*, 30(17):i482–i488, 2014.

Ovchinnikov, S., Kamisetty, H., and Baker, D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *Elife*, 3:e02030, 2014.

Rocha, G. K., Custódio, F. L., Barbosa, H. J. C., and Dardenne, L. E. A multiobjective approach for protein structure prediction using a steady-state genetic algorithm with phenotypic crowding. *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on*, pages 1–8, 2015.

Rocha, G. K. *Desenvolvimento de Metodologias Para Predição de Estruturas de Proteínas Independente de Moldes.* Tese de Doutorado, PhD thesis, Laboratório Nacional de Computação Científica (LNCC), Petrópolis, RJ., 2015.

Santos, K. B., Custódio, F. L., Barbosa, H. J., and Dardenne, L. E. Genetic operators based on backbone constraint angles for protein structure prediction. *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on*, pages 1–8, 2015.

Taylor, T. J., Bai, H., Tai, C.-H., and Lee, B. Assessment of casp10 contact-assisted predictions. *Proteins: Structure, Function, and Bioinformatics*, 82(S2):84–97, 2014.

Unger, R. The building block approach to protein structure prediction. pages 177–188. Springer, 2004.

van Gunsteren, W. F., Billeter, S., Eising, A., Hünenberger, P. H., Krüger, P., Mark, A. E., Scott, W., and Tironi, I. G. Biomolecular simulation: the {GROMOS96} manual and user guide. 1996.

Wang, S., Li, W., Zhang, R., Liu, S., and Xu, J. Coinfold: a web server for protein contact prediction and contact-assisted protein folding. *Nucleic acids research*, pages gkw307, 2016.

Wang, Z. and Xu, J. Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics*, 29(13):i266–i273, 2013.

Xu, J. and Zhang, Y. How significant is a protein structure similarity with tm-score= 0.5? *Bioinformatics*, 26(7):889–895, 2010.