Recent Constraint/Logic Programming based advances in the solution of the Protein Folding Problem

Agostino Dovier
Dip. di Matematica e Informatica,
Univ. di Udine, Italy
agostino.dovier@uniud.it

2011

Abstract

In this paper, we summarize the contribution of our research group to the field of Bioinformatics. In particular, we present our approach to the *ab-initio* solution of the protein structure prediction problem based on constraint/logic programming techniques.

1 Introduction

In the last years we have witnessed the birth and the rapid growth of a new research area whose results have a positive impact on traditional and fundamental disciplines such as biology, chemistry, physics, medicine, agriculture, or industry. This area, known as *Bioinformatics*, uses algorithms and methodological techniques developed by Computer Sciences to solve challenging problems in all the above areas. Moreover, new emerging problems spur for Computer Sciences to develop new algorithms and methods. Bioinformatics deals with recognition, analysis, and organization of DNA sequences, with biological systems simulations, with problems of prediction of the spatial conformation of a biological polymer, among others.

We have worked in this field during the last years with the double effort of solving real problems and spreading known techniques, methods, and languages to researchers of the motivating areas above. In this spirit, we organized the workshops WCB (Constraint-Based Methods for Bioinformatics) 2005–2010, co-located with the major meetings of Logic and Constraint Programming ICLP, CP, and CPAIOR (see, e.g., `http://wcb10.dimi.uniud.it/`); we have organized the International Summer Schools BCI (Biology, Communication, and Information) in Dobbiaco and Trieste (see, e.g., `http://www.dmi.units.it/bci2010/`); and we edited a special issue of the journal *Constraints* on these topics [16].

As far as the technical contribution is concerned, we have worked on the Protein Structure Prediction problem using, whenever possible, techniques coming from logic programming and constraint programming (briefly, CLP). In the rest of this paper we shortly review this challenging problem and give an overview of our results.

2 The Protein Structure Prediction Problem

The *Primary structure* of a protein is a linked sequence of amino acids. There are 20 kinds of aminoacids, identified by a letter code in a set $\mathcal{A} \subset \{A, \ldots, Z\}$. The primary structure of a protein is therefore a string $s_1 \cdots s_n$ with $s_i \in \mathcal{A}$. Local 3D conformations involving subsequences of $s_1 \cdots s_n$, called α-helices and β-sheets, are often present in protein conformations. The set of these local structures is known as the *secondary structure* of a protein. The *Tertiary Structure* (or native state) of the protein is a 3D conformation associated to the entire primary structure. The *protein structure prediction problem* is the problem of predicting the tertiary structure, given the primary structure. *Quaternary structures*, i.e. arrangements of several folded proteins in a more general complex, are not discussed in this paper.

Amino acids differ from each other by a set of atoms called side chain. Nevertheless, all amino acids own a particular carbon atom, called $C\alpha$, that represents, in a sense, the center of the amino acid. One of the most important structural features of the 3D conformation of proteins is that the distance between the $C\alpha$ atoms of two consecutive amino acids is fixed and is roughly 3.8 Å. We will refer to this distance as d in this paper.

In Figure 1 we report the primary and the tertiary structure of the protein 2K2P deposited in April 2008. In the top picture all atoms of the amino acids are

A G L S F H V E D M T C G H C A G V I K G
A I E K T V P G A A V H A D P A S R T V V
V G G V S D A A H I A E I I T A A G Y T P E

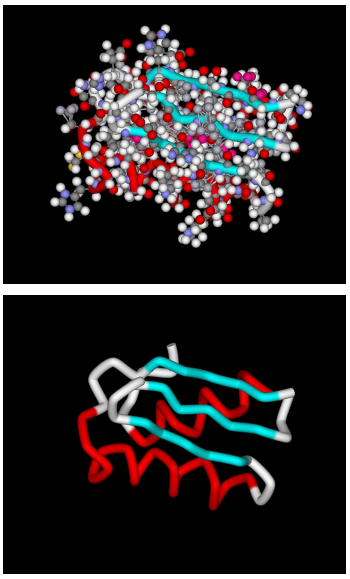

Figure 1: Primary and Tertiary structures (all-atoms and C_α–C_α structure) of Protein 2K2P (amino acids 22–85). Observe the presence of 2 α-helices (in red—dark gray) and 3 β-sheets (in cyan—light gray)

represented. In the bottom picture we report the abstract structure obtained linking the C_α atoms. This abstraction allows us to observe three β-sheets and two α-helices and the regularity forced by the constant distance between two consecutive amino acids.

Let $\mathcal{D}$ be a set of admissible points for (the $C\alpha$ atoms of) the amino acids. A function $\omega : \{1, \ldots, n\} \longrightarrow \mathcal{D}$ is said a *folding* if

- for $i, j \in \{1, \ldots, n\}$ if $i \neq j$ then $|\omega_i - \omega_j| \geq d$.
- for $i \in \{1, \ldots, n-1\}$ it holds: $|\omega_i - \omega_{i+1}| = d$.

Let c be a fixed distance (c will be of the same order as d). For two points

$p, q \in \mathcal{D}$, we define two versions of the function dist as follows:

$$\begin{aligned}
\mathsf{dist}_{\mathsf{discr}}(p,q) &= \begin{cases} 1 & \text{if } |p-q| \leq c \\ 0 & \text{otherwise} \end{cases} \\
\mathsf{dist}_{\mathsf{cont}}(p,q) &= \begin{cases} 1 & \text{if } |p-q| \leq c \\ \frac{c^2}{(p-q)^2} & \text{otherwise} \end{cases}
\end{aligned}$$

The former is used on discrete spatial domains, the latter on continuous spatial domains. These two energy functions are of course approximations of more complex rules that consider interactions between all the atoms forming the amino acids of the protein.

Let Pot be a function from pairs of amino acids to integer numbers. The *free energy of a folding* $E(\omega)$ is computed as follows:

$$E(\omega) = \sum_{\substack{1 \leq i < n \\ i+2 \leq j \leq n}} \mathsf{dist}_*(\omega_i, \omega_j)\mathsf{Pot}(s_i, s_j)$$

where $*$ can be either discr or cont.

In this context, the *Protein Structure Prediction Problem (PSP)* is the problem of determining the folding(s) ω with minimum free energy. The problem contains some symmetries that can be avoided by symmetry breaking search (see e.g. [2]). Some symmetries can be removed by fixing the positions of the first two amino acids (ω_1 and ω_2).

Two main approximations can be made: (1) *space*: the set of admissible points, and (2) *energy*: the details of the Potential function used. It is well-known that lattice-based models are realistic approximations of the set of the admissible points for the C_α atoms of a protein [27]. Lattices are basically 3D graphs with repeated patterns. For instance the *face centered cube (FCC—c.f. Fig. 2)* lattice is defined as: $\mathcal{D} = \{(x,y,z) \in \mathbb{N}^3 \ : \ x+y+z \text{ is even}\}$, $E = \{(p,q) \in \mathcal{D}^2 \ : \ |p-q| = \sqrt{2}\}$. Thus, $d = \sqrt{2}$. In this model, typically, c is set to 2. This means that two amino acids are assumed in contact either if their distance in the lattice is of one or two lattice units. This allows us to partially recover artificial constraints due to the rigidity of the lattice structure. A "toy" 2D lattice used to prove complexity limits is the $\mathbb{N}^2$ lattice: $\mathcal{D} = \mathbb{N}^2, E = \{(p,q) \in \mathcal{D}^2 \ : \ |p-q| = 1\}, c = d = 1$—c.f. Fig. 2.

The discrete energy function $\mathsf{dist}_{\mathsf{discr}}$ is used in lattice models. Moreover, the constraint $|\omega_i - \omega_j| \geq d$ can be relaxed to $\omega_i \neq \omega_j$. Three are the main contact energy models used in literature for Pot: the HP model [18], the HPNX model [4], and the 20×20 model [6].

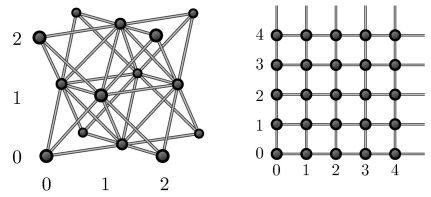

Figure 2: The main cell of a FCC lattice and a portion of the $\mathbb{N}^2$ lattice.

3 Related Work

We focus here only on the logic/constraints based approaches to the problem. For a recent general survey we refer to [23].

In the HP model [18], amino acids are split in two families: hydrophobic (H) and polar (P). Two hydrophobic amino acids in contact contribute -1 to the energy. The other contacts are not relevant. The NP-completeness even in the simple spatial model constituted by the $\mathbb{N}^2$ lattice is proved in [8]. In particular, it is proved that the problem: *Given a sequence of P and H, stating the existence of a folding with at least k contacts between H* is NP-complete.

Backofen and Will solved this problem using constraint programming for proteins of length 160 and more on the FCC (see [3, 1, 2]). Efficiency is obtained using a clever symmetry breaking and the notion of *core*. Basically, the folding is analyzed layer by layer and the various conformations of each layer that maximize contacts are pre-computed. This kind of approach is unapplicable to more detailed energy models and with the addition of other structural constraints (e.g., known α-helices and β-sheets). Slightly more complex energy models have been proposed by the same group for the protein structure prediction problem. In [4] they consider an energy model in which amino acids are split into 4 families. Other researchers (e.g. [26]) approximate the solution to the same problem using local search and refined meta-heuristics.

Barahona and Kripphal, instead, work on off-lattice space model where space is discretized into small cubes. They also deal with protein docking and develop the tool Chemera, commonly used by biochemists in their research [22, 5].

4 Our Contribution

The contribution of our group to the PSP problem covers both lattice and off-lattice approaches.

In our work on lattice models we have used FCC as the space model, the 20×20 statistical potential contact energy model [6], and the function $\mathsf{dist}_{\mathsf{discr}}$. In [19] we encoded the problem using the library clpfd of SICStus Prolog. Since contact energy is not suited to predict α-helices and β-strands in the FCC lattice, we pre-computed these secondary structure elements using other well-known tools, and exploited these pre-computations as constraints within the main code. In this first encoding the number of admissible angles for secondary structure elements was too limited. We relaxed this restraint in [9] where a more general and precise handling of secondary structure constraints was implemented. However, the exponential growth of the search space w.r.t. protein length made impossible to explore the whole search space even using state-of-the-art constraint solvers for proteins of length greater than 30/40. Therefore, we proposed an ad-hoc constraint search with biologically motivated heuristics and we introduced the data structure *potential matrix* that allowed us to reduce calculations during this phase. This approach was then extended by relaxing some constraints and developing other search heuristics [10].

In all these approaches we used a double representation for the tertiary structure: a cartesian one, based on the set of points, and a polar one, based on the torsional angles generated by the protein during the folding process. The cartesian representation is useful for defining the notion of folding and the constraint-based energy function. The polar representation simplifies the encoding of secondary structure constraints. However, several extra constraints need to be introduced so as to manage the conversion between the two representations. This badly scales on large proteins (the constraint solvers used were close to their memory limit for protein of length 60). Thus, in [13] we decided to abandon the polar representation and to impose secondary structure using cartesian constraints only. This way, we loose the chirality property (clockwise/counterclockwise) of helices but the overall definition become simpler. In the same paper we also developed a search heuristic (Bounded Block Fail—BBF). The list of variables associated to positions of the amino acids is dynamically split into blocks of k variables that will be instantiated together. When the variables in the block B_i are instantiated to an apparently admissible solution, the search moves to the successive block B_{i+1}, if any. If all the assignments of the block B_{i+1} violate some constraints or produce a worsening of the partial energy value, the search backtracks to the block B_i. Now, there are two options: if the number of times that B_{i+1} has failed is below a fixed threshold, then the process continues, by generating one more solution to B_i and re-entering B_{i+1}. Otherwise, the heuristic generates a failure for B_i as well and backtracks to B_{i-1}. The key idea is that most of small local changes do not

ID–n	[19]	[9]	[10]	[13]	[15]
1LE3–16	2m	1.6s	25s	1.2 s	0.2 s
1ZDD–34	13m	7m	1m	138s	0.4 s
2GP8–40	30m	10s*	8h	72s	0.2 s
1ENH–54	40m	40m*	>10h	10h	80s

Figure 3: Running time of various proposals on some small proteins

change the form of a protein too much. When we try a sufficient number of close conformations and we fail, we can freely abandon that research branch.

In [12] we developed an ad-hoc constraint solver written in C, named COLA (COnstraint solving on LAttices). In COLA the lattice point is an elementary element, associated with a 3D domain (a box) rather than a triple of FD variables as in the previous approaches. We developed and implemented ad-hoc constraint propagation techniques and the BBF heuristic. This approach with a further parallelization was then presented in [15].

Just to give an idea of the evolution of our proposals, we report the running times of the systems on the prediction of some small proteins in Fig. 3. Timings have been recomputed on an AMD Opteron 2.2GHz Linux machine, using SICStus Prolog 4 for the first 4 columns and COLA 2.1 for the last one. The solutions found with different techniques are not always the same, but (save for the first column related to a too strict encoding) they have comparable energy and shape. More importantly, the solutions are very close to the real tertiary structure (see the original papers for details). Times marked by a "*" are worse than those in [10] and [13], since some extra structural constraints among secondary structure were added in the original papers. However, [15] improves also the original results even without extra information. The protein 1FVS of Figure 1 is predicted by COLA 2.1 with BBF in less than one hour. All codes are available from `www.dimi.uniud.it/dovier/PF`.

We have also investigated how to use model checking results in order to analyze the folding process [17] and how to model the PSP as a planning problem using a variant of the action description language $\mathcal{B}$ [21] and as an Answer Set Program [20].

The *ab-initio* approach used by COLA is still computationally infeasible when applied to the prediction of protein structures with more than hundred amino acids. Only the presence of other kind of partial information (e.g., known folds

for sub-blocks picked from the protein data bank) can speed up the search significantly. This is however in line with what done by other prediction tools (like e.g. Rosetta [24]), where partial information are retrieved from the protein databank from similar structures/substructures and only small subsequences need to be arranged. Thus, we have started a systematic study of what kind of *global constraints* are needed in a solver for structure predictions in lattice models. In particular, in [14], we have studied the definition and the complexity of testing satisfiability and applying propagation for the constraints `alldifferent`, `contiguous`, `self avoiding walk`, `alldistant`, `chain`, and `rigid block`, and we have studied a global constraint that accounts for partial information coming from density maps. These global constraints have been incorporated in COLA to profit as much as possible from partial information coming from known proteins and from partial predictions. However, exploiting knowledge of real rigid substructures in a discrete lattice is infeasible, due to the errors introduced by the approximations imposed by the discretized representation of the search space. On the other hand, the use of a less constrained space model leads to search spaces that are intractable. These two considerations lead to a new off-lattice approach based on CLP, to predict the 3D conformation of a protein via fragments assembly [11]. The primary structure is split into consecutive 4-tuples of amino acids. For each of these tuples a set of possible 3D conformations (fragments) is associated. The fragments are extracted by a preprocessor from a database of known protein structures that clusters and classifies the fragments according to similarity and frequency. The problem of assembling fragments into a complete conformation minimizing the free energy (using the continuous version $\mathsf{dist}_{\mathsf{cont}}$—c.f. Sect. 2) is mapped to a constraint solving problem and solved using CLP; speed-up in the searching stage is obtained using Large Neighboring search (see, e.g., [25]) implemented in Prolog.

We also presented an approach to the protein folding problem using multi-level Agent-Based (concurrent) simulation in [7].

5 Conclusions and future work

This work represents a typical effective use of a declarative programming paradigm for problem solving. The problem can be encoded easily and solutions (for small inputs) can be computed by built-in mechanisms of (constraint) logic programming. Heuristics and alternative encodings can be easily programmed and tested. When the encoding becomes stable, speed-up can be obtained by less declara-

tive methods. The results obtained are promising for the application of the same approach to other challenging problems of Bioinformatics.

Acknowledgements. The research summarized in this paper would not have been possible without the help of my valuable colleagues and real friends Alessandro Dal Palù, Federico Fogolari, and Enrico Pontelli. I would also like to thank Luca Bortolussi, Raffaele Cipriano, Elisabetta De Maria, Andrea Formisano, Angelo Montanari, and Carla Piazza for their collaboration, and Rolf Backofen, Francois Fages, and Sebastian Will for the help in WCB's organizations.

The research has been partially supported by the FIRB RBNE03B8KK, PRIN 20089M932N, and by INdAM-GNCS projects.

References

[1] R. Backofen. The protein structure prediction problem: A constraint optimization approach using a new lower bound. *Constraints* 6:223–255, 2001.

[2] R. Backofen and S. Will. Excluding Symmetries in Constraint-Based Search. *Constraints* 7(3–4):333–349, 2002.

[3] R. Backofen and S. Will. A Constraint-Based Approach to Fast and Exact Structure Prediction in 3-Dimensional Protein Models. *Constraints* 11(1):5–30, 2006.

[4] R. Backofen, S. Will, and E. Bornberg-Bauer. Application of constraint programming techniques for structure prediction of lattice proteins with extended alphabets. *Bioinformatics* 15(3): 234–242, 1999.

[5] P. Barahona and L. Krippahl. Constraint Programming in Structural Bioinformatics. [16]:3–20.

[6] M. Berrera, H. Molinari, and F. Fogolari. Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. *BMC Bioinformatics* 4(8), 2003.

[7] L. Bortolussi, A. Dovier, and F. Fogolari. Agent-based Protein Structure Prediction. *Multiagent and Grid Systems* 3(2):183–197, 2007.

[8] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis. On the Complexity of Protein Folding, *Journal of Computational Biology*, 5(3):423–466, 1998.

[9] A. Dal Palù, A. Dovier, and F. Fogolari. Protein Folding in CLP(FD) with Empirical Contact Energies. Proc. of CSCLP03:250–265, LNCS 3010, 2004.

[10] A. Dal Palù, A. Dovier, and F. Fogolari. Constraint Logic Programming approach to Protein Structure Prediction. *BMC Bioinformatics* 5(186), 2004.

[11] A. Dal Palù, A. Dovier, F. Fogolari, and E. Pontelli. CLP-based protein fragment assembly. *Theory and Practice of Logic Programming*. 10(4-6):709–724, 2010.

[12] A. Dal Palù, A. Dovier, and E. Pontelli. A Constraint Logic Programming Approach to 3D Structure Determination of Large Protein Complexes. Proc. of LPAR05:48–63, LNCS 3835, Springer, 2005.

[13] A. Dal Palù, A. Dovier, and E. Pontelli. Heuristics, Optimizations, and Parallelism for Protein Structure Prediction in CLP(FD). Proc. of PPDP05:230–241, ACM, 2005.

[14] A. Dal Palù, A. Dovier, and E. Pontelli. Computing Approximate Solutions of the Protein Structure Determination Problem using Global Constraints on Discrete Crystal Lattices. *Int'l Journal of Data Mining and Bioinformatics* 4(1):1-20, 2010.

[15] A. Dal Palù, A. Dovier, and E. Pontelli. A constraint solver for discrete lattices, its parallelization, and application to protein structure prediction. *Software Practice and Experience* 37(13):1405–1449, 2007.

[16] A. Dal Palù, A. Dovier, and S. Will (eds.) Special issue on Constraint Based Methods for Bioinformatics. *Constraints* 13(1–2), 2008.

[17] E. De Maria, A. Dovier, A. Montanari, and C. Piazza. Exploiting Model Checking in Constraint-based Approaches to the Protein Folding. Proc. of WCB06, pp.46-54, Nantes, 2006.

[18] K. A. Dill. Dominant forces in protein folding. Biochemistry 29:7133-7155, 1990.

[19] A. Dovier, M. Burato, and F. Fogolari. Using Secondary Structure Information for Protein Folding in CLP(FD). Proc. of WFLP02, ENTCS 76, 2002.

[20] A. Dovier, A. Formisano, E. Pontelli. A comparison of CLP(FD) and ASP solutions to NP-complete problems. In M. Gabbrielli and G. Gupta eds., *Proc of ICLP 2005*:67–82, LNCS 3668, Springer, 2005.

[21] A. Dovier, A. Formisano, and E. Pontelli. Multivalued Action Languages with Constraints in CLP(FD). *Theory and Practice of Logic Programming* 10(2):167–235, 2010.

[22] L. Krippahl and P. Barahona. PSICO: Solving Protein Structures with Constraint Programming and Optimisation. *Constraints*, 7:317–331, 2002.

[23] A. Kryshtafovych and K. Fidelis. Protein structure prediction and model quality assessment. *Drug Discov Today* 14(7–8):386-393, 2009.

[24] S. Raman R. Vernon, J. Thompson, M. Tyka, R. Sadreyev, J. Pei, D. Kim, E. Kellogg, F. DiMaio, O. Lange, L. Kinch, W. Sheffler, B.-H. Kim, R. Das, N. V. Grishin, and D. Baker. Structure prediction for casp8 with all-atom refinement using Rosetta. *Proteins 77,* S9, 89–99, 2009.

[25] P. Shaw. Using constraint programming and local search methods to solve vehicle routing problems. In *Proc. of CP98*:417–431, LNCS 1520, Springer, 1998.

[26] A. Shmygelska and H. H. Hoos. An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem. *BMC Bioinformatics* 6(30), 2005.

[27] J. Skolnick and A. Kolinski. Reduced models of proteins and their applications. *Polymer*, 45:511–524, 2004.