Recenti progressi basati su Programmazione Logica e/o a Vincoli sulla soluzione del protein folding

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

Agostino Dovier

SOMMARIO/*ABSTRACT*

In questo articolo desidero illustrare il contributo del mio gruppo di ricerca alla disciplina Bioinformatica, con particolare riferimento alla risoluzione del problema della predizione di struttura di una proteina usando metodologie di programmazione logica e a vincoli.

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

Keywords: Logic Programming, Constraint Programming, Bioinformatics

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 (briefly denoted globally as "Bio"). This area, known as *Bioinformatics* uses algorithms and methodological techniques developed by Computer Sciences to solve challenging problems in "Bio" areas. Moreover, new emerging problems produce stimuli 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 in the last years with the double effort of solving real problems and of spreading known techniques, methods, and languages to "Bio" researchers.

In this spirit, we have been organizers of the workshops WCB (Constraint-Based Methods for Bioinformatics) associated with ICLP in 2005 and 2007, with CP in 2006, and with CPAIOR in 2008 (see, e.g., `http://wcb08.dimi.uniud.it`); we have organized the International Summer Schools BCI (Biology, Communication, and Information) in Dobbiaco and Trieste (see, e.g., `http://bioinf.dimi.uniud.it/bci2006`; and we have been guest editors of a special issue of the journal *Constraints* on these topics [17].

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. In the rest of this paper we briefly introduce 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 aminoacids. There are 20 kinds of aminoacids, identified by a letter. For the scope of this paper, the primary structure of a protein is a string $s_1 \cdots s_n$ with $s_i \in \{A, \ldots, Z\} \setminus \{B, J, O, U, X, Z\}$.

The *Tertiary Structure* (native state) of the protein is a 3D conformation associated to the primary structure. The protein structure prediction problem is the problem of predicting the tertiary structure, given the primary structure.

The Tertiary Structure usually assumes two types of local conformation: α-helices and β-sheets. In Figure 1 we report the primary and the tertiary structure of the protein 2K2P deposited in April 2008. In the top figure all atoms of the amino acids are represented. In the lower figure we report the abstract structure obtained linking the C_α atoms (briefly, a central atom of each aminoacid). With this abstraction the secondary structure elements (three β-sheets and two α-helices) are evident.

Let $\mathcal{D}$ be a set of admissible points for the amino acids. Let c, d two fixed distances. For two points $p, q \in \mathcal{D}$, we say that $\mathsf{next}(p, q)$ if and only if $|p - q| = d$.[1] For two

[1] For real proteins, $d = 3.8\mathring{A}$ corresponding to the distance between two consecutive C_α in the sequence

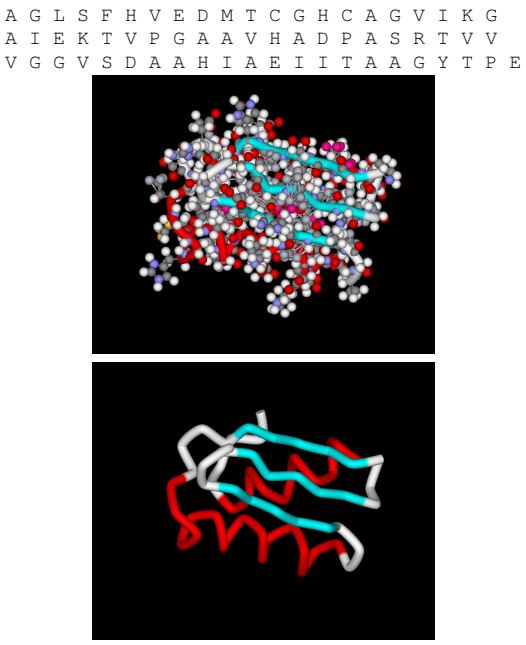

Figure 1: Primary and Tertiary structures (all-atoms and C_α–C_α structure) of Protein 2K2P (amino acids 22–85). Observe the presence of 2 α-helices (in red—dark gray) and 3 β-sheets (in cyan—light gray)

points $p, q \in \mathcal{D}$, we define the Boolean function contact as follows: $\mathsf{contact}(p, q) = 1$ if and only if $|p - q| \leq c$.

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 $\omega_i \neq \omega_j$
- for $i \in \{1, \ldots, n-1\}$ it holds that $\mathsf{next}(\omega_i, \omega_{i+1})$

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:

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

The *protein structure prediction problem (PSP)* is the problem of determining the folding(s) ω with minimum energy. The problem contains some symmetries that can be avoided by symmetry breaking search (see e.g. [2]). The simplest way to remove some symmetries is to fix the positions of the first two points (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 [24]. Lattices are basically 3D graphs with repeated patterns. For instance the *face centered cube (FCC)* 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, $d = \sqrt{2}, c = 2$.

Three are the main contact energy models used in literature for Pot: the HP model [19], the HPNX model [4], and the 20x20 model [6].

3 Related Work

In the HP model [19], 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[2] is proved in [9]. 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 protein 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 precomputed. This kind of approach is unapplicable to a more detailed energy models and with the adding 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. [23]) instead approximated 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

In all our works, we have used FCC as the space model, and the 20x20 statistical potential contact energy model presented in [6].

CLP(FD) encoding. In [20] we encoded the problem using the library clpfd of SICStus Prolog. Since contact energy is not suited to predict helices and sheets in the FCC lattice, we pre-computed secondary structure elements (α-helices and β-strands) using other well-known tools. The results of these pre-computations were then used 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 restrain in [10] 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

[2] I.e., $\mathcal{D} = \mathbb{N}^2, E = \{(p, q) \in \mathcal{D}^2 : |p - q| = 1\}, c = d = 1$.

even using state-of-the-art constraint solvers for proteins of length greater than 30/40. Therefore, we proposed an ad-hoc labeling search with biologically motivated heuristics and we introduced 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 [11].

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. The cartesian representation is useful for defining the notion of self-avoiding walk and the notion of constraint-based energy function. The polar representation simplifies the encoding of secondary structure constraints. However, a lot of extra constraints need to be introduced 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 we decided (in [13]) to abandon the polar representation and to impose secondary structure constraints only using cartesian constraints. This way, we loose the chirality property of helices but the overall definition becomes simpler.

In the same paper we also developed a search heuristics (Bounded Block Fail—BBF). The list of variables is dynamically split into blocks of k variables that will be labeled 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 the labeling of the block B_{i+1} fails, 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 certain threshold, then the process continues, by generating one more solution to B_i and re-entering B_{i+1}. Otherwise, the heuristics generates a failure for B_i as well and backtracks to B_{i-1}. The key idea is that small local changes do not change too much the form of a protein. When we tried a sufficient number of close conformations and we fail, we can freely abandon that research branch (with fail we consider either no admissible foldings or admissible foldings with energy greater than the local minimum already found).

Ad-hoc constraint solver. In [12] we developped an ad-hoc constraint solver written in C, named COLA (COnstraint solving on LAttices). In the previous approaches each 3D point was viewed as a triple of FD variables $\langle X, Y, Z\rangle$. In COLA, instead, the lattice point is an elementary element, associated with a 3D domain (a box). We developed and implemented ad-hoc constraint propagation techniques and the BBF heuristics. This approach with a further parallelization was then presented in [16].

Just to give a taste of the evolution of our proposals, we report the running times of the systems on the prediction of some small proteins in Figure 2. Timings are taken from the published papers (the machine used for the leftmost column is roughly 3x slower than the machine for the rightmost). The solutions found with various techniques are not always the same, but (save for the first column related to a too strict encoding) they have comparable energy and form. And, more important, the form is very close to their real tertiary structure. The protein 2K2P of Figure 1 is predicted by COLA 3.1 with BBF in less than one hour.

ID–n	[20]	[10]	[11]	[13]	[16]
1LE3–16	12.5m	5s	2.5s	1.5s	0.5s
1ZDD–34	47m	41s	17.5s	2m	0.1s
2GP8–40	6.5h	9m	10.5h	1.5m	0.5s
1ENH–54	3.5h	13m	24h	55m	49.5s

Figure 2: Running time of the various approaches on some small proteins

Towards generalition and integration. 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 subblocks picked from the protein data bank) can speed-up significantly the search. This is however in line with what done by other prediction tools (like e.g. ROSETTA), where partial information is picked from the protein data-bank 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 lattice models structure predictions. In particular 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` constraint in [14]; we have studied a global constraint that accounts for partial information coming from density maps in [15]. These global constraints will be incorporated in COLA so as to obtain a tool able to profit as much as possible of partial information coming from known proteins and from partial predictions.

We have also studied how to use model checking results for analyzing the folding process [18] and how to model the protein folding problem as a planning problem using a variant of the well-known action description language $\mathcal{B}$ [21]. An alternative approach to the protein folding problem based on Agent-Based simulation is proposed in [7].

5 Conclusions and future work

This work represents a typical use of logic 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 success of 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, Elisabetta De Maria, Andrea Formisano, Angelo Montanari, and Carla Piazza for their collaboration. The research has been partially supported by the FIRB RBNE03B8KK, and by PRIN and 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. [17]: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] R. Cipriano, A. Palù, and A. Dovier. A hybrid approach mixing local search and constraint programming applied to the protein structure prediction problem. Proc. of WCB08, Paris, 2008.

[9] 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.

[10] 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.

[11] A. Dal Palù, A. Dovier, and F. Fogolari. Constraint logic programming approach to protein structure prediction. *BMC Bioinformatics* 5(186), 2004.

[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, pp. 48–63, 2005.

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

[14] A. Dal Palù, A. Dovier, and E. Pontelli. Global constraints for Discrete Lattices. Proc. of WCB06, Nantes, pp. 55–68, 2006.

[15] A. Dal Palù, A. Dovier, and E. Pontelli. The density constraint. Proc. of WCB07, Porto, pp. 10–19, 2007.

[16] 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*, DOI: `10.1002/spe.810`.

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

[18] 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.

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

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

[21] A. Dovier, A. Formisano, and E. Pontelli. Multivalued Action Languages with Constraints in CLP(FD). Proc. of ICLP07, LNCS 4670, pp. 255–270, 2007.

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

[23] 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.

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

6 Contacts

Agostino Dovier
Dip. di Matematica e Informatica
Univ. di Udine
Via delle Scienze 206, 33100 Udine (UD)
Tel: +39 0432 558494
E-mail: dovier@dimi.uniud.it

7 Biography

Agostino Dovier received his PhD in Computer Science from the University of Pisa in 1996 and he is an Associate Professor of Computer Science at the University of Udine. His current research interests include the development and the application of declarative programming languages with constraints and Bioinformatics. He is member of AI*IA and of the EC of GULP and ALP and he has published over 60 international referred publications. He served as program chair or in the program committee of several conferences and workshops of logic and constraint programming, as guest editor of special issues of international journals, and he has coordinated some research projects in the area of Constraint Logic Programming and Bioinformatics. He is the general chair of ICLP08 (International Conference on Logic Programming).