Constraint Programming and Biology: Introduction

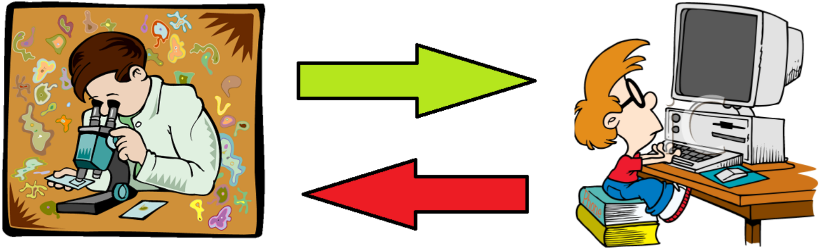
Agostino Dovier

Dept. Math and Computer Science, Univ. of Udine, Italy

ACP Summer School in Constraint Programming
Wrocław, September 2012

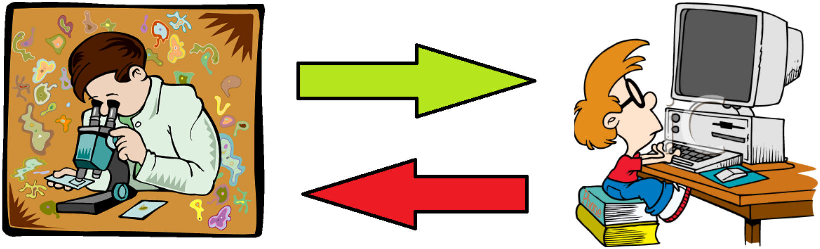
Introduction

- Biology is an incredible source of challenging problems for computer science
- Problems are often hidden or confused and emerge only after long discussions with biologist, physics, chemists, physicians, and so on (briefly, biologist)



Introduction

- Biology is an incredible source of challenging problems for computer science
- Problems are often hidden or confused and emerge only after long discussions with biologist, physics, chemists, physicians, and so on (briefly, biologist)



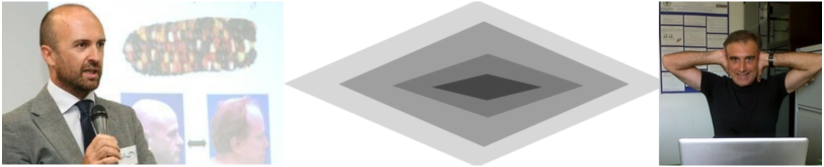
- Solving one of these problems can be of unpredictable importance for life sciences and medicine

Introduction

- Some problems are of little interest for computer science but of great importance for biologist (eg developing scripts for automatization of sequences of simple tasks).

Introduction

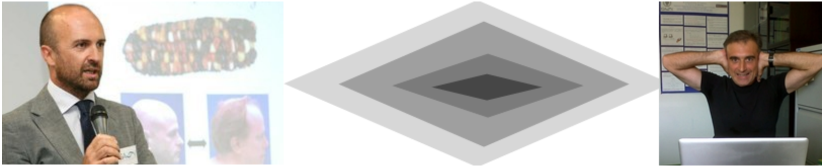
- Some problems are of little interest for computer science but of great importance for biologist (eg developing scripts for automatization of sequences of simple tasks).



✓ Since we don't want to hear these problems we don't solve them

Introduction

- Some problems are of little interest for computer science but of great importance for biologist (eg developing scripts for automatization of sequences of simple tasks).

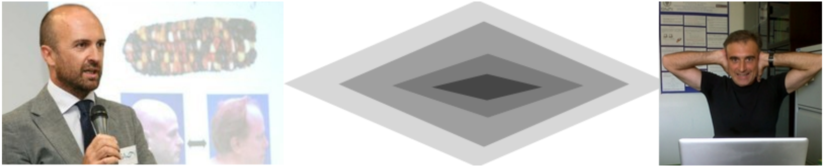


✓ Since we don't want to hear these problems we don't solve them

- Some problems are polynomial time solvable but the input size is huge (e.g., a DNA string). These problems require fast string matching algorithms. They are important and challenging but the constraint programming approach is not the best suited for them.

Introduction

- Some problems are of little interest for computer science but of great importance for biologist (eg developing scripts for automatization of sequences of simple tasks).



✓ Since we don’t want to hear these problems we don’t solve them

- Some problems are polynomial time solvable but the input size is huge (e.g., a DNA string). These problems require fast string matching algorithms. They are important and challenging but the constraint programming approach is not the best suited for them.

✓ We will not deal with the two kinds of problems above in these lectures

Introduction

- There is a large set of bio problems that we can prove they are intractable (NP complete or worse) even with simplifications.

WANTED
$P \neq NP$
Wanted for crimes against the 250cc Bikes, He is a 2-Stroke Killer. Outlaw is known to be extremely dangerous and should be approached with caution.
$1,000,000
REWARD

Introduction

- There is a large set of bio problems that we can prove they are intractable (NP complete or worse) even with simplifications.

- We love studying and solving these simplified models and at the end we win (at least for little inputs, and using some “reasonable” heuristics) but the risk is that with these simplifications our solutions are useless for biologists.

Introduction

- There is a large set of bio problems that we can prove they are intractable (NP complete or worse) even with simplifications.

WANTED
P ≠ NP
Wanted for crimes against the 250cc Bikes, He is a 2-Stroke Killer. Outlaw is known to be extremely dangerous and should be approached with caution.
$1,000,000
REWARD

- We love studying and solving these simplified models and at the end we win (at least for little inputs, and using some “reasonable” heuristics) but the risk is that with these simplifications our solutions are useless for biologists.

✓ We will focus on this family. CP techniques are perfect for NP problems. And sometimes our solutions are not useless!

Problems for Bioinformatics

Bioinformatics can be seen as the area of computer science that deal with modeling and solving problems for Biology.

We have several families of problems.

- Those concerning DNA and genes
- Those concerning the transcription DNA $\mapsto$ RNA and the structure of RNA
- Those concerning the translation RNA $\mapsto$ proteins and the structure of proteins
- Those concerning the interaction between molecules and the behavior/interaction of systems of molecules (e.g. cells), till the modeling of living organisms.

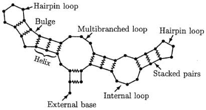
Areas of Bioinformatics

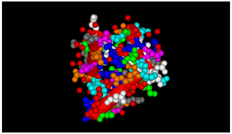
1. Genomics. Study of the genomes. Huge amount of data, fast algorithms (not always), limited to sequence analysis.

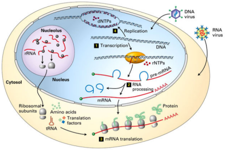
```
. . .   G   A   T   C   T   G   T   A   C   T   G   A   G   T   . . .
. . .   G   A   T   C   T   G   T   A   C   T   G   A   A   T   . . .
```

2. Structural Bioinformatics. Study of the folding process of bio-molecules. Less structural data than sequence data available.

⇑ ⇑







⇓

3. Systems Biology. Study of complex interactions in biological systems. High level of representation.

Why Constraint Programming?

(At least) two main reasons:

Why Constraint Programming?

(At least) two main reasons:

- Models are rarely stable (and also the problems change quickly). Modifying a CP-modeling is easy and fast.

Why Constraint Programming?

(At least) two main reasons:

- Models are rarely stable (and also the problems change quickly). Modifying a CP-modeling is easy and fast.
- Linear Programming is not enough (in particular for modeling energy models)

What we'll see in mode details

We'll focus on some challenging problems and how modeling them using constraints:

- Genomics:
  - ✓ Haplotype Inference
  - ✓ Phylogenetic trees
- Systems Biology:
  - ✓ Reasoning on Biological Networks
- Structural Bioinformatics:
  - ✓ RNA secondary structure prediction
  - ✓ protein structure prediction (on/off lattice)

⇒ For these problems I have prepared the encodings in CLP(FD) (tested with BProlog—free). Link in my home page.

Some introductory references

- P. Clote and R. Backofen. *Computational Molecular Biology*. An Introduction. Wiley, 2000.
- Nice introductory slides by Sebastian Will (MIT) `http://math.mit.edu/classes/18.417/Slides/intro.pdf`
- A movie on DNA replication `http://www.youtube.com/watch?v=teV62zrm2P0`
- A movie on DNA transcription `http://www.youtube.com/watch?v=5MfSYnItYvg`
- A movie on Protein synthesis `http://www.youtube.com/watch?v=lpb5s2F1pyM&feature=related`
- A movie on Systems Biology `http://www.youtube.com/watch?v=HNP1EAYLhOs&feature=fvwrel`

Some references on Constraints and Bioinformatics

- P. Barahona, L. Krippahl, and O. Perriquet. *Bioinformatics: A Challenge to Constraint Programming*. In Hybrid Optimization – The Ten Years of CPAIOR, Springer, 2011.
- Workshops on Constraint-based methods for Bioinformatics: WCB05 (Sitges), WCB06 (Nantes), WCB07 (Porto), WCB08 (Paris), WCB09 (Lisbon), WCB10 (Edinburgh), WCB11 (Perugia), WCB12 (Budapest).
  Formerly: *Workshops on Constraints and Bioinformatics/ Biocomputing* in CP'97 and CP'98.
- Constraints, Volume 13. Special Issue on Bioinformatics and Constraints, 2008.
- Algorithms for Molecular Biology 7:15–17 (Thematic Series of AMB on Constraints and Bioinformatics), 2012.

Acknowledgments

(in advance)

- School organizers: ACP, Krzysztof Apt, Witold Charatonik, Leszek Pacholski, . . .
- School participants (you)
- My main collaborators in Bioinformatics: Alessandro Dal Palù, Federico Fogolari, Enrico Pontelli, Sebastian Will, Rolf Backofen, Francois Fages, Federico Campeotto, Ferdinando Fioretto, . . .
- Those that helped me directly or indirectly in preparing/checking these slides: Martin Gebser, Giuseppe Lancia, Simone Scalabrin, Esra Erdem, . . .