The Density Constraint

Alessandro Dal Palù Agostino Dovier
Enrico Pontelli

Dip. di Matematica, Univ. di Parma

Dip. di Informatica, Univ. di Udine

Dept. Computer Science, New Mexico State University

WCB07 — Porto, Sept. 13th, 2007

Outline

Global Constraints

- Given n vars $X_1, \ldots, X_n$, with domains $D_1, \ldots, D_n$, a *global constraint* C on $X_1, \ldots, X_n$ is a subset $C \subseteq D_1 \times \cdots \times D_n$
- We are mainly interested in verifying two properties:
  - consistency (CON): $C \neq \emptyset$
  - generalized arc consistency (GAC):

$$\forall i \in \{1, \ldots, n\} \, \forall a_i \in D_i \\ \exists a_1 \in D_1 \cdots \exists a_{i-1} \in D_{i-1} \exists a_{i+1} \in D_{i+1} \cdots \exists a_n \in D_n \\ (a_1, \ldots, a_n) \in C$$

- filtering is the process of removing values from the domains of variables in order to obtain an equivalent constraint which is *GAC*.
- We are interested in the complexities of these problem, and in approximation algorithms for them whenever they are NP hard
- In lattice models atomic values are 2D/3D points.

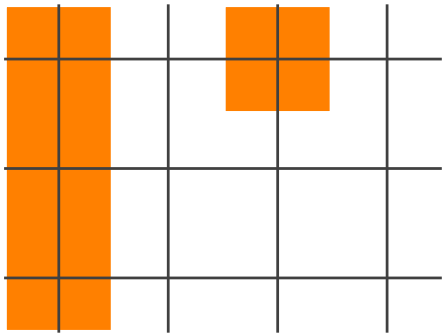
Global Constraints

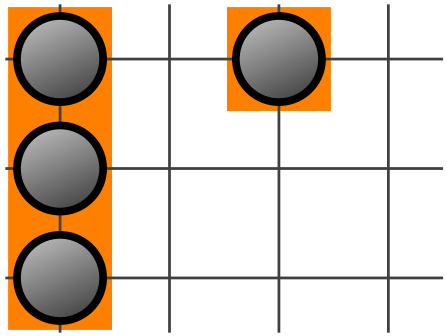
- Given n vars $X_1, \ldots, X_n$, with domains $D_1, \ldots, D_n$, a *global constraint* C on $X_1, \ldots, X_n$ is a subset $C \subseteq D_1 \times \cdots \times D_n$
- We are mainly interested in verifying two properties:
  - consistency (CON): $C \neq \emptyset$
  - generalized arc consistency (GAC):

$$\forall i \in \{1, \ldots, n\} \, \forall a_i \in D_i \\ \exists a_1 \in D_1 \cdots \exists a_{i-1} \in D_{i-1} \exists a_{i+1} \in D_{i+1} \cdots \exists a_n \in D_n \\ (a_1, \ldots, a_n) \in C$$

- filtering is the process of removing values from the domains of variables in order to obtain an equivalent constraint which is *GAC*.
- We are interested in the complexities of these problem, and in approximation algorithms for them whenever they are NP hard
- In lattice models atomic values are 2D/3D points.

Constraints on lattices
alldifferent





$D_{X_1} = \cdots = D_{X_4} =$ Orange points

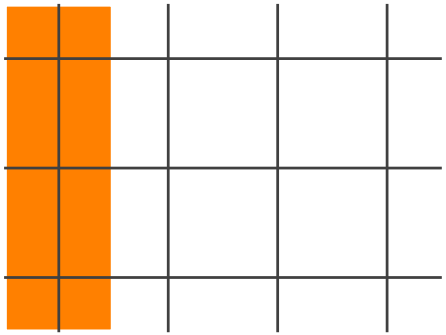
A solution

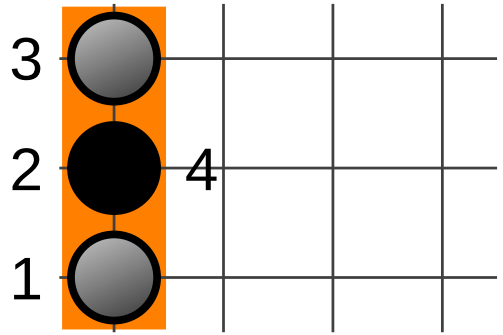
- Let $X_1, \ldots, X_n$ be variables with domains $D_1, \ldots, D_n$:

$$\texttt{alldifferent}(X_1, \ldots, X_n) = (D_1 \times \cdots \times D_n) \setminus \{(a_1, \ldots, a_n) \in (D_1 \times \cdots \times D_n) : \exists i, j.\ (1 \leq i < j \leq n \wedge a_i = a_j)\}$$

- CON and GAC are polynomial

Constraints on lattices
contiguous





$D_{X_1} = \cdots = D_{X_4} =$ Orange points A solution

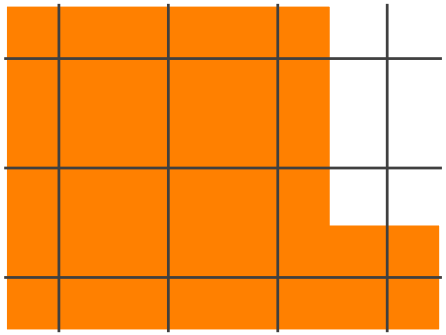
- Let $X_1, \ldots, X_n$ be variables with domains $D_1, \ldots, D_n$:

$$\texttt{contiguous}(X_1, \ldots, X_n) = (D_1 \times \cdots \times D_n) \setminus \{(a_1, \ldots, a_n) \in (D_1 \times \cdots \times D_n) : \exists i.\ (1 \leq i < n \wedge (a_i, a_{i+1}) \notin E)\}$$

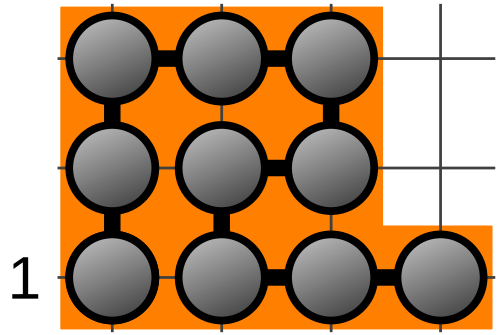
  where E is the set of lattice edges.

- CON and GAC are polynomial

Constraints on lattices
self avoiding walk



$D_1 = \cdots = D_{10} =$ Orange points



A solution

- Given n variables $X_1, \ldots, X_n$, with domains $D_1, \ldots, D_n$, the global constraint saw is the following:

$$\texttt{saw}(X_1, \ldots, X_n) = \texttt{alldifferent}(X_1, \ldots, X_n) \cap \texttt{contiguous}(X_1, \ldots, X_n)$$

- CON (and GAC) are NP-complete.

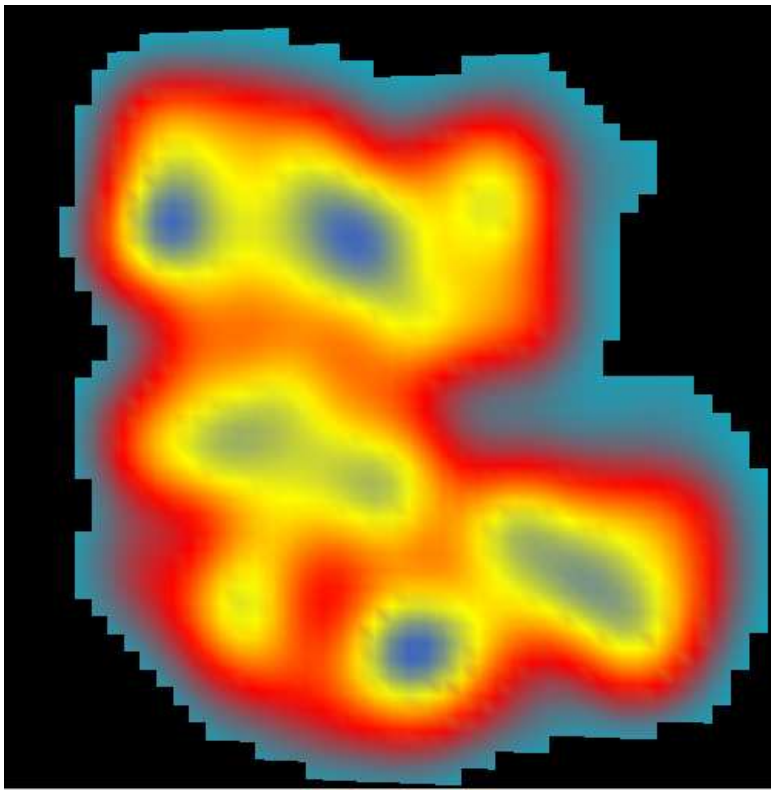
Constraints on lattices
Other constraints

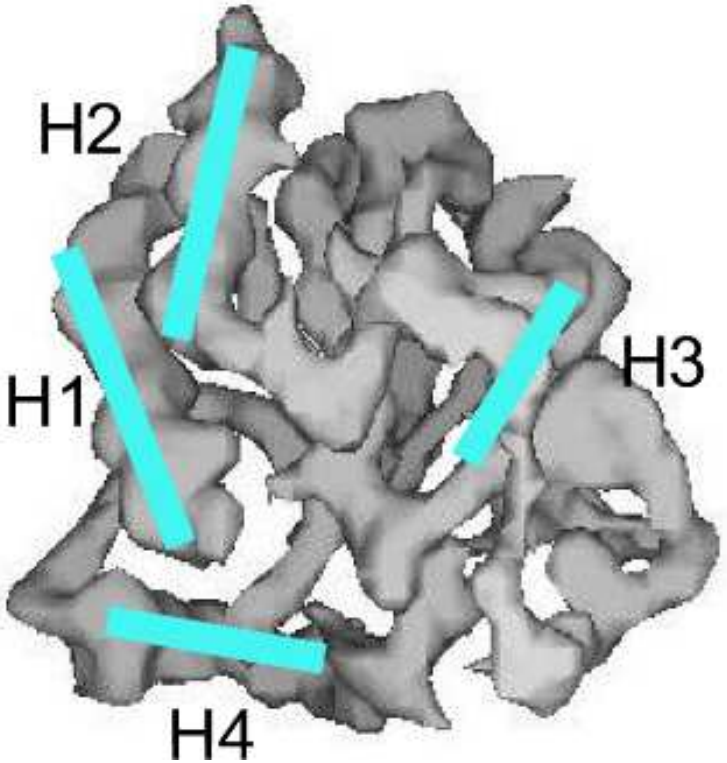
- all distant (NP)
- chain (NP)
- rigid block (P)
- These constraints are built-in in the tool COLA, in order to predict proteins in the FCC lattice with a 20×20 contact energy function.

Density Maps
Introduction

- A density map D is a 3D data set obtained through electron cryomicroscopy analysis.
- It represents the electron density of molecules in a given portion of space (with resolution 6Å–12Å).
- The density map is sampled at uniform rate R in the 3D space.
- This generates a discretization of the space into cubes with fixed size.
- Each cube contains a certain amount of density that is associated to a real number.

Density Maps
Examples





Density maps
Reconstruction of D

- Each component of a molecule can be approximated by a function $\mathcal{F} : \mathbb{N}^3 \times \mathbb{N}^3 \rightarrow \mathbb{R}^+ \cup \{0\}$.
- A choice for $\mathcal{F}$ could be a *Gaussian* contribution:

$$\mathcal{F}(\vec{x}, \vec{p}) = \mathcal{G}_{a,\sigma}(\vec{x}, \vec{p}) = ae^{-\frac{|\vec{x}-\vec{p}|^2}{2\sigma^2}}$$

- $\vec{p} \in \mathbb{N}^3$ is the reference point for the object $a \in \mathbb{R}^+$ is the intensity of the map, and $\sigma \in \mathbb{R}^+$ is the decay control parameter.
- a and σ can be estimated according to the type of the component, by generating and analyzing density maps for the single components
- Other $\mathcal{F}$ can be used, of course

Density maps
Reconstruction of D

- Each component of a molecule can be approximated by a function $\mathcal{F} : \mathbb{N}^3 \times \mathbb{N}^3 \rightarrow \mathbb{R}^+ \cup \{0\}$.
- A choice for $\mathcal{F}$ could be a *Gaussian* contribution:

$$\mathcal{F}(\vec{x}, \vec{p}) = \mathcal{G}_{a,\sigma}(\vec{x}, \vec{p}) = ae^{-\frac{|\vec{x}-\vec{p}|^2}{2\sigma^2}}$$

- $\vec{p} \in \mathbb{N}^3$ is the reference point for the object $a \in \mathbb{R}^+$ is the intensity of the map, and $\sigma \in \mathbb{R}^+$ is the decay control parameter.
- a and σ can be estimated according to the type of the component, by generating and analyzing density maps for the single components
- Other $\mathcal{F}$ can be used, of course

Density maps
Reconstruction of D

- Given a molecule chemical description (components and chemical bonds) and a density map D,
- the molecule is decomposed into its n components
- E.g., a protein can be decomposed into its amino acids, or, with higher precision, into its set of atoms.
- Each component $i \in 1 \ldots n$ if placed in the position $\vec{p}_i$ provides a specific contribution function $\mathcal{F}_i(\vec{x}, \vec{p}_i)$
- The goal is to find the possible placements of the components $[\vec{p}_1, \ldots, \vec{p}_n]$ so that they globally produce a density map close to D:

$$(\forall \vec{x} \in \mathbb{N}^3) \left(\sum_{i=1}^{n} \mathcal{F}_i(\vec{x}, \vec{p}_i) \approx D(\vec{x}) \right)$$

The Density Constraints

- We propose two global constraints.
- Both of them relies on the idea of discretization of the space into regions (e.g., cubes of size $d \in \mathbb{N}$).
- In the first one, called average density constraint, *regions* of space are “big” i.e., capable of containing more than one component.
- The second constraint, called punctual density constraint, considers instead “small” (point-wise) regions. The center of a component is placed exactly on a discrete point in the space, and it also affects the neighboring points.
- Atomic values are points in $\mathbb{N}^3$.

Average Density Constraint

Let us consider the following components:

- k disjoint regions $i = 1, \ldots, k$, each described by:
  - the set of points $R_i \subseteq \mathbb{N}^3$ of the region, and
  - a (density) value $s_i \in \mathbb{R}^+ \cup \{0\}$

  Let $\vec{R} = (R_1, \ldots, R_k)$ and $\vec{s} = (s_1, \ldots, s_k)$.
- n molecules $1, \ldots, n$, each of them characterized by the corresponding contribution function $\mathcal{F}_i$, where $\mathcal{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_n)$,
- n vars $\vec{X} = (X_1, \ldots, X_n)$ with domains $O_1, \ldots, O_n$.

The global constraint a_density$(\vec{X}, \mathcal{F}, \vec{R}, \vec{s})$ is satisfied by all the n-tuples $\langle p_1, \ldots, p_n \rangle \in O_1 \times \cdots \times O_n$ s.t. for all $i = 1, \ldots, k$ it holds that:

$$\sum_{j=1}^{n} \sum_{\vec{x} \in R_i} \mathcal{F}_i(\vec{x}, p_j) \leq s_i \qquad (1)$$

Punctual Density Constraint

Let us consider

- a *discrete* density map $D(x, y, z) \in \mathbb{R}^+ \cup \{0\}$, where $x, y, z \in \mathbb{N}$
- n molecules $1, \ldots, n$, with contribution function $\mathcal{F}_i$, where $\mathcal{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_n)$
- n vars $\vec{X} = (X_1, \ldots, X_n)$ with domains $O_1, \ldots, O_n$

The global constraint p_density($\vec{X}, \mathcal{F}, D$) is satisfied by the set of n-tuples $\langle p_1, \ldots, p_n \rangle \in O_1 \times \cdots \times O_n$ s.t. for all $\vec{x} \in \mathbb{N}^3$ it holds that:

$$\sum_{i=1}^{n} \mathcal{F}_i(\vec{x}, p_i) \leq D(\vec{x}) \tag{2}$$

Average and Punctual Density Constraint
Complexity

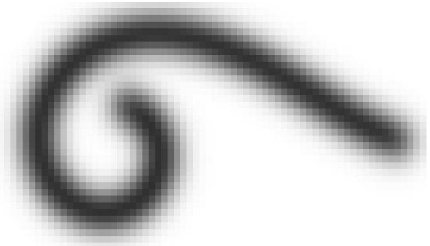
- Simplifications of the two density constraints are proved to be NP-complete (bin packing)
- If the number α of possible molecule is known and the maximum number of molecules per region is bounded by β (realistic hypothesis) then there is a polynomial algorithm $O(k^q)$ where k is the number of regions and $q = \frac{(\alpha+\beta)!}{\alpha!\beta!}$

(Some) Experimental Results

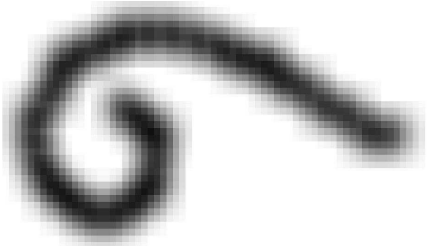
- We have included density constraints in a CSP framework to determine the protein structure in $\mathbb{N}^3$.
- Given a protein sequence $s_1 \dots s_n$, its density map D, $s_i \in \{a_1, \dots, a_{20}\}$, the j-th amino acids density map parameters a_j and a Gaussian function $\mathcal{G}_j$, $j \in \{1, \dots, 20\}$, vars X_i with domains $D_i \subseteq \mathbb{N}^3$ ($i \in \{1, \dots, n\}$).
- The constraints added to the problem are:
  - p_density($\vec{X}, \vec{a}, \vec{\mathcal{G}}, D$)
  - contiguous($\vec{X}$) which imposes that each pair of consecutive amino acids are placed at Euclidean distance 3.3Å–5.2Å (3.8Å+ ε).
- For discretization and Gaussian approximation problems, D is slightly enlarged

(Some) Experimental Results
Test 1

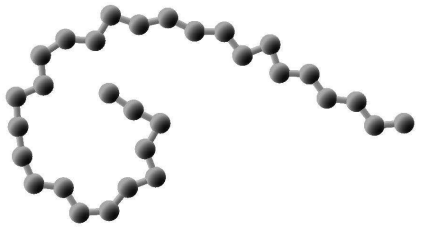
- We implemented a first filtering: a domain point $\vec{p}$ for amino acid S_i has support only if the presence of S_i does not violate D. Moreover there must be a compatible assignment of S_{i-1} and S_{i+1} that respects both the contiguous and the p_density constraints.
- We arrange a chain of 30 amino acids with a spiral shape, respecting the distance of 3.8Å between consecutive elements.



D



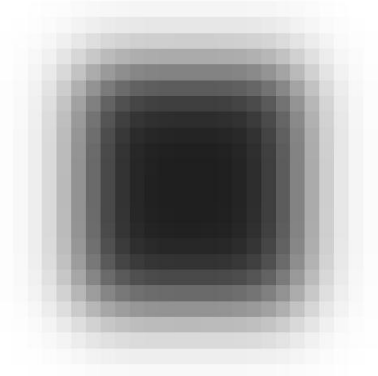
Enlarged D



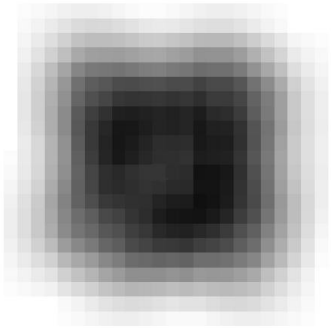
A solution

(Some) Experimental Results
Test 2

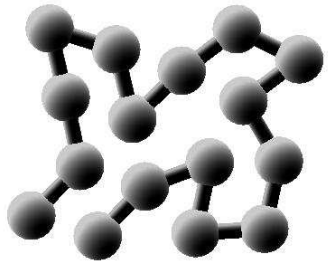
- *D* is a uniform density map (obtained by a square tiling arrangement of 16 amino acids on the plane)



D



Enlarged *D*



A solution

(Some) Experimental Results
Timings

	Amino acids	Filtering	Domain size	Search	Nodes
Spiral	30	119.1 s	16–387	0.13 s	103
Square	16	23.8 s	214	95.6 s	124,663

Question: Does the approach work in 3D?

Answer: The point is not xD but the number of points and the forms of the density maps

Question: Have you tried with real proteins?

Answer: Not yet (see next slide)

(Some) Experimental Results
Timings

	Amino acids	Filtering	Domain size	Search	Nodes
Spiral	30	119.1 s	16–387	0.13 s	103
Square	16	23.8 s	214	95.6 s	124,663

Question: Does the approach work in 3D?

Answer: The point is not xD but the number of points and the forms of the density maps

Question: Have you tried with real proteins?

Answer: Not yet (see next slide)

Conclusions and Future Work

- We have identified a global constraint for predicting molecule structure
- Of course, there is a lot to do
- First of all, improving the implementation and the integration with the other global constraints
- We wish to test PFP on real-size proteins (where we could also benefit, e.g., by secondary structure prediction or partial homology information, and even NMR constraints)
- There are, maybe, some similarities between the density constraint and the cellular phone/wireless networks antenna location problem. If you know relevant references (in particular for complexity, approximations, constraint approaches) please send us an email! We'll pay a coffee in Udine or Parma for each suggestion.

Conclusions and Future Work

- We have identified a global constraint for predicting molecule structure
- Of course, there is a lot to do
- First of all, improving the implementation and the integration with the other global constraints
- We wish to test PFP on real-size proteins (where we could also benefit, e.g., by secondary structure prediction or partial homology information, and even NMR constraints)
- There are, maybe, some similarities between the density constraint and the cellular phone/wireless networks antenna location problem. If you know relevant references (in particular for complexity, approximations, constraint approaches) please send us an email! We'll pay a coffee in Udine or Parma for each suggestion.