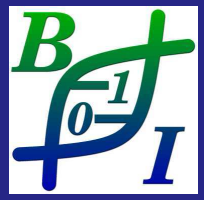
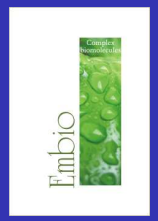




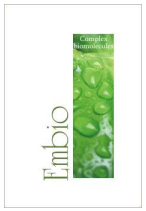
Optimal Structure Prediction and Application

Freiburg Bioinformatics Group

Martin Mann, Sebastian Will and Rolf Backofen

Albert-Ludwigs-University Freiburg
Bioinformatics at the Department of Computer Science

EMBIO Meeting Leipzig 2007



Native structure prediction

Martin Mann
Freiburg

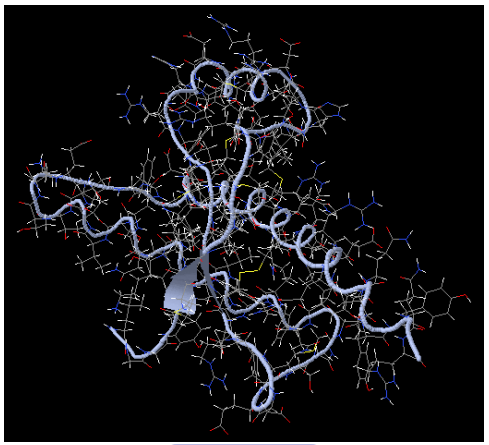
The goal

- Prediction of the native structure given an AA-sequence

LGGYMLGSA. . . RESQAYYQR

⇓

Human prion 1HJM



Zoom in

Assumptions

- Sequence determines structure
- Native structure has lowest energy

Problems

- too complex energy function
- too many degrees of freedom

Native structure prediction

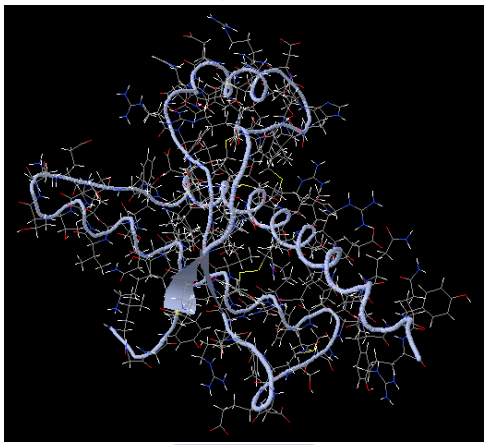
Martin Mann
Freiburg

The goal

- Prediction of the native structure given an AA-sequence

LGGYMLGSA. . . RESQAYYQR

Human prion 1HJM



▸ Zoom in

Assumptions

- Sequence determines structure
- Native structure has lowest energy

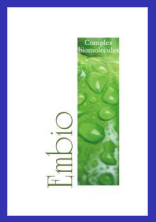
Problems

- too complex energy function
- too many degrees of freedom

Computationally not capable!

Simplified Off-Lattice Protein Models

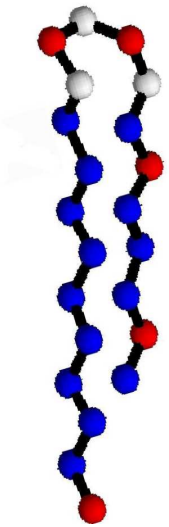
One possible abstraction



Martin Mann
Freiburg

Backbone Structure

C_α sequence only



Full 3D Space

all angles etc. allowed

Reduced Alphabet

e.g. HP, HPNX, . . .

Contact Energy Function

e.g.

	H	P
H	-1	0.5
P	0.5	-0.5

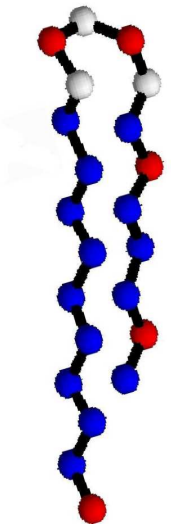
Simplified Off-Lattice Protein Models

One possible abstraction

Martin Mann
Freiburg

Backbone Structure

C_α sequence only



Reduced Alphabet

e.g. HP, HPNX, . . .

Full 3D Space

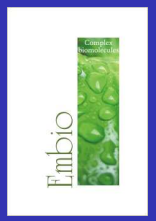
all angles etc. allowed

Contact Energy Function

e.g.

	H	P
H	-1	0.5
P	0.5	-0.5

$\Rightarrow$ **Still too many degrees of freedom . . .**



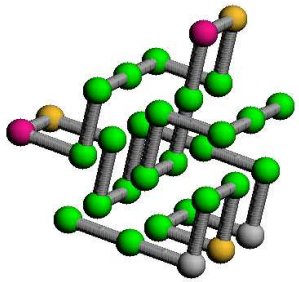
Martin Mann
Freiburg

Backbone Structure

C_α sequence only

Reduced Alphabet

e.g. HP, HPNX, . . .



Lattice Space

e.g. cubic, fcc, . . .

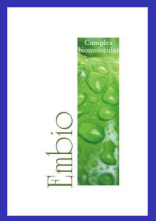
Contact Energy Function

e.g.

	H	P
H	-1	0
P	0	0

Simplified Lattice Protein Models

An other possible discrete abstraction



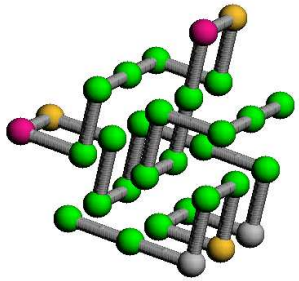
Martin Mann
Freiburg

Backbone Structure

C_α sequence only

Reduced Alphabet

e.g. HP, HPNX, ...



Lattice Space

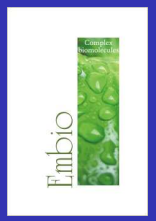
e.g. cubic, fcc, ...

Contact Energy Function

e.g.

	H	P
H	-1	0
P	0	0

Discrete, Enumerable, Computationally Capable



Martin Mann
Freiburg

Applications e.g.

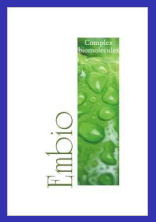
- Neutral nets and protein evolution
- Exploring energy landscapes and protein kinetic
- Base for more complex protein models
- . . .

Therefore you need:

Prediction of optimal structures

- NP-complete in 3D-lattice (Berger & Leighton, 1998) (even in 2D)
- can be solved by Constraint Programming !
  (Backofen & Will, 2006)

. . .



Martin Mann
Freiburg

Applications e.g.

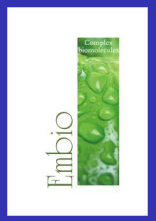
- Neutral nets and protein evolution
- Exploring energy landscapes and protein kinetic
- Base for more complex protein models
- ...

Therefore you need:

Prediction of optimal structures

- NP-complete in 3D-lattice (Berger & Leighton, 1998) (even in 2D)
- can be solved by Constraint Programming !
  (Backofen & Will, 2006)

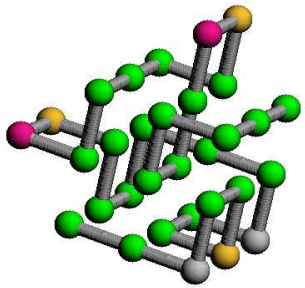
...



Martin Mann
Freiburg

Constraint-based Protein Structure Prediction[1]

- **Short introduction**
  - **Optimal structure prediction (CPSP)**
- **Applications in**
  - **Protein stability**
  - **Inverse folding problem**



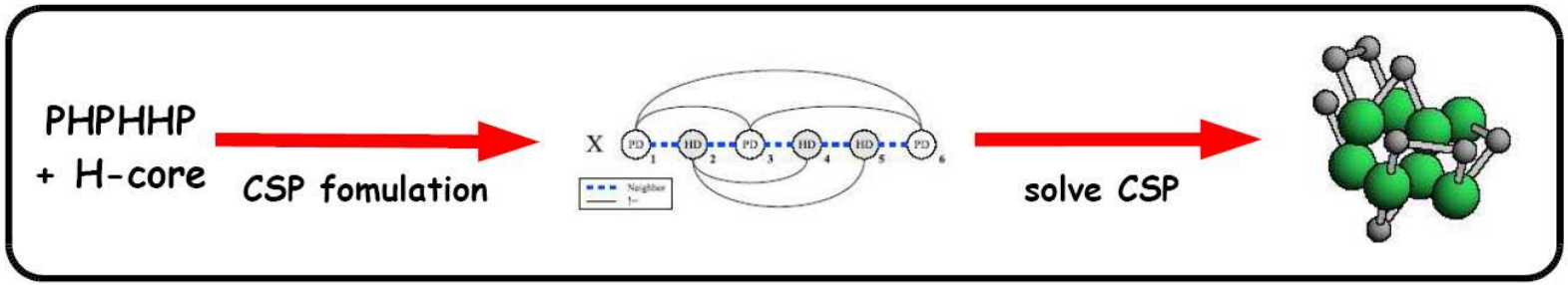
[1] R. Backofen and S. Will

The CPSP Approach

Martin Mann
Freiburg

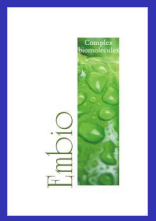
Constraint-based Protein Structure Prediction (CPSP)

An Approach for optimal structure prediction
in the HP-lattice-model



Rolf Backofen and Sebastian Will
'A constraint-based approach to fast and exact structure prediction

in three-dimensional protein models' 2006



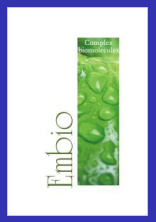
Martin Mann
Freiburg

The HP-Model

- simplest energy function available
- focus on hydrophobicity (hydrophobic cores)
- maximizing HH-contacts $\leftrightarrow$ minimizing surface

CPSP - The central idea

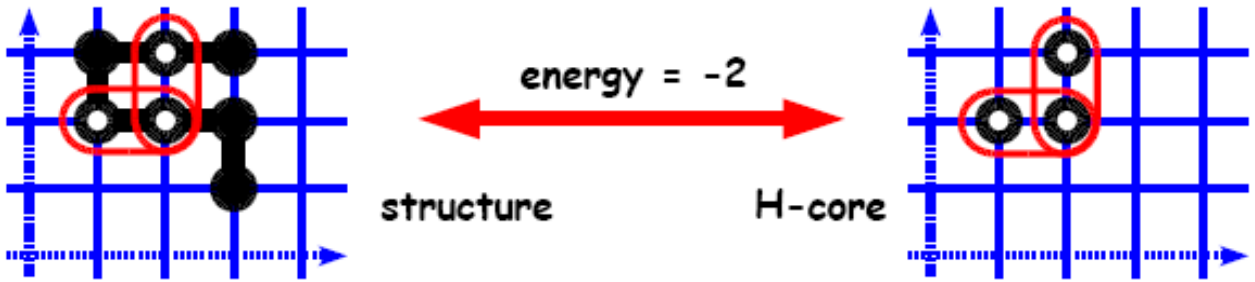
- optimal H-monomer distribution is *sequence independent*
- precalculate such optimal and suboptimal so called *H-cores*
- try to find a mapping of a given sequence to H-cores
  $\Rightarrow$ *Sequence-Threading*



Martin Mann
Freiburg

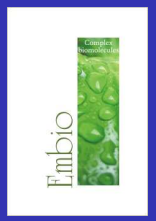
H-Core of a given structure

- H-Core = set of H-monomer positions
- core energy $\leftrightarrow$ structure energy (only HH-contacts important)



- optimality implies optimal structure energy
- candidates can be precomputed based on H-number
- hard problem too $\rightarrow$ (solved via CP)

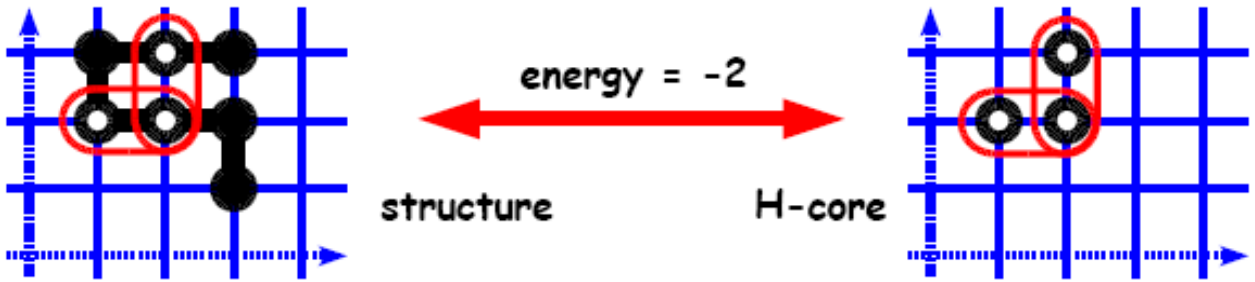
$\Rightarrow$ for now used as black box and given in a DB ... !



Martin Mann
Freiburg

H-Core of a given structure

- H-Core = set of H-monomer positions
- core energy $\leftrightarrow$ structure energy (only HH-contacts important)



- optimality implies optimal structure energy
- candidates can be precomputed based on H-number
- hard problem too $\rightarrow$ (solved via CP)

$\Rightarrow$ for now used as black box and given in a DB ... !

The CPSP Approach

Sequence Threading : The Question to solve . . .

Martin Mann
Freiburg

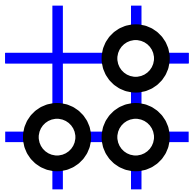
Given

an HP-sequence

P-H-P-H-H-P-P

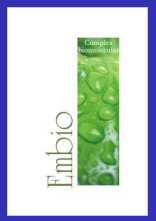
and

an optimal H-core



there is the question:

Exists a structure, so that all H-monomers are placed on H-core positions?



Martin Mann
Freiburg

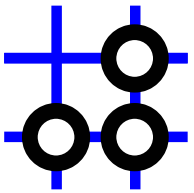
Given

an HP-sequence

P-H-P-H-H-P-P

and

an optimal H-core



there is the question:

Exists a structure, so that all H-monomers are placed on H-core positions?

Martin Mann
Freiburg

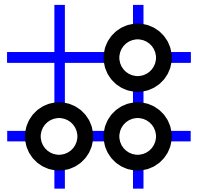
Given

an HP-sequence

P-H-P-H-H-P-P

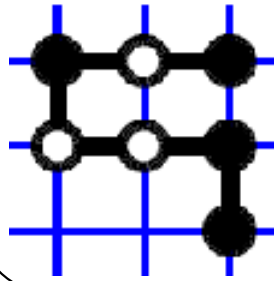
and

an optimal H-core



there is the question:

Exists a structure, so that all H-monomers are placed on H-core positions?



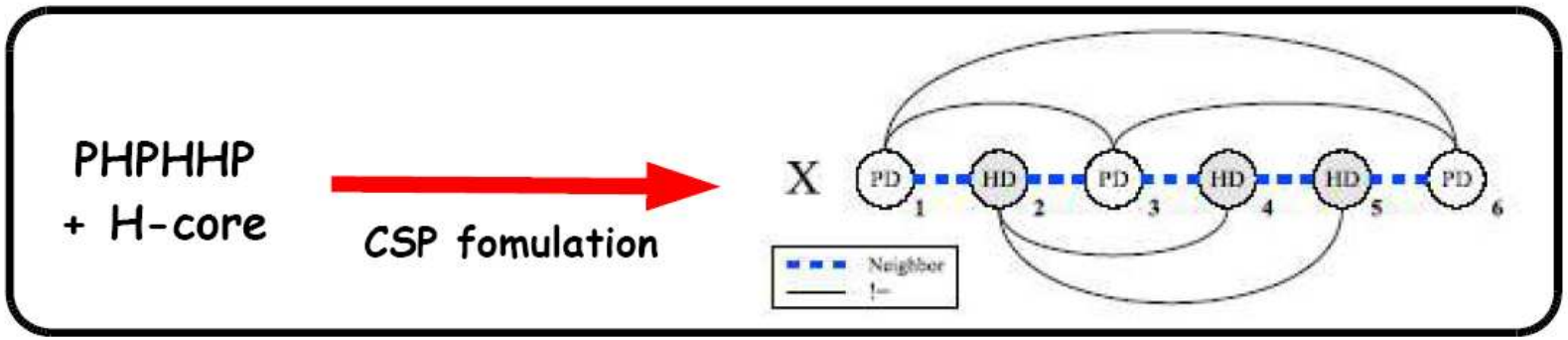
YES!

The CPSP Approach

Solving a Constraint Satisfaction Problem (CSP)

Martin Mann
Freiburg

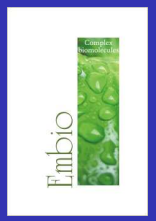
Modelling of the question as CSP



From Solution to structure

- fast standard CP-solvers can be applied for solving
- a CSP solution assigns a lattice position to each monomer
- solution = structure, and optimal due to H-core !
- usually a huge number of solutions / optimal structures

CSP Formulation



Martin Mann
Freiburg

The CPSP-Approach

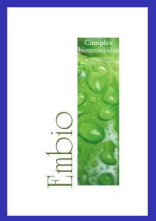
- proven optimal structures via precalculated H-cores
- fast (first hit within seconds)
- deterministic (no stochastic structure space exploration)
- CP yields a very flexible, extensibel modelling

Extensibility

- lattices (cubic, face centered cubic, . . .)
- energy functions (HP, HPNX, . . .)
- exclusion of symmetric solutions during enumeration
- solution space sampling via distance constraints (D-10)
- advanced CP-techniques for solution counting (D-9)
- . . . future: side chain models, structure shapes, . . .

The CPSP Approach

Fast and very flexible . . .



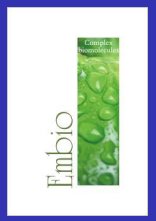
Martin Mann
Freiburg

The CPSP-Approach

- proven optimal structures via precalculated H-cores
- fast (first hit within seconds)
- deterministic (no stochastic structure space exploration)
- CP yields a very flexible, extensibel modelling

Extensibility

- lattices (cubic, face centered cubic, . . .)
- energy functions (HP, HPNX, . . .)
- exclusion of symmetric solutions during enumeration
- solution space sampling via distance constraints (D-10)
- advanced CP-techniques for solution counting (D-9)
- . . . future: side chain models, structure shapes, . . .



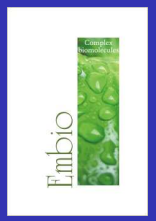
Martin Mann
Freiburg

The CPSP-Approach

- Prediction of optimal structures without folding simulation
- Exhaustive enumeration possible

Use

- Verifying folding simulation results
- Creation of interesting test sets for analysis
- Clustering of sequences into proteinlike or not
- Enumeration of the low energy part of the landscape
  $\longrightarrow$ base for kinetic studies
- Base to solve other problems e.g. sequence evolution, inverse folding / designability



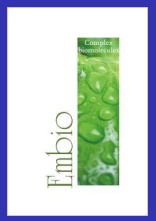
Martin Mann
Freiburg

The CPSP-Approach

- Prediction of optimal structures without folding simulation
- Exhaustive enumeration possible

Use

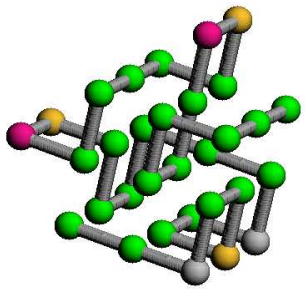
- Verifying folding simulation results
- Creation of interesting test sets for analysis
- Clustering of sequences into proteinlike or not
- Enumeration of the low energy part of the landscape
  $\longrightarrow$ base for kinetic studies
- Base to solve other problems e.g. sequence evolution, inverse folding / designability



Martin Mann
Freiburg

Constraint-based Protein Structure Prediction

- Short introduction
  - Optimal structure prediction (CPSP)
- Applications in
  - Protein stability
  - Inverse folding problem



What needs an AA-sequence to be a protein?

Martin Mann
Freiburg

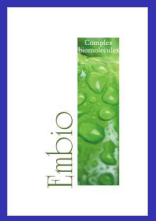
Observations

- Not all possible AA-sequences used
- Fast folding process
  (folding funnel hypthesis)
- Usually ***one*** stable, native structure
- ...

Human prion 1HJM

Degeneracy

A measure of protein stability



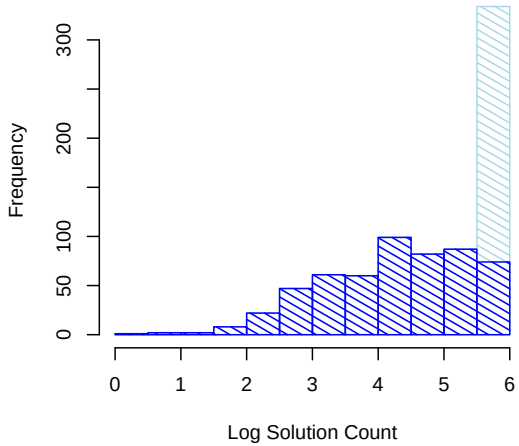
Martin Mann
Freiburg

Degeneracy

- = the number of optimal structures
- important for protein stability
- deg. 1 = indicator for being stable

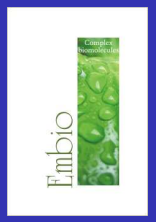
Degeneracy in Lattice-Models

- is high for most sequences
- due to simple energy function
- assumption: deg. 1 = stable



log_{10}(degeneracy) histogram
of 809 HP-sequences (H:P=1:1)
(32% with deg. > 10^6)

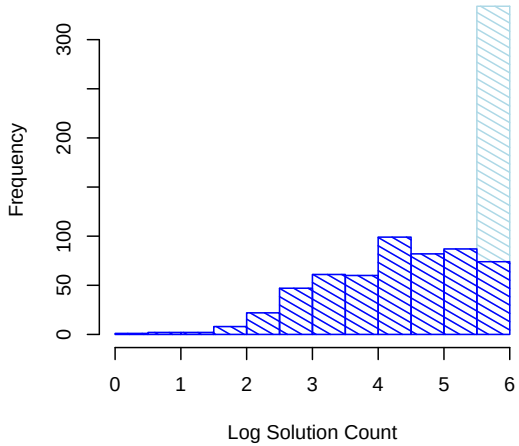
$\Rightarrow$ **exact determinable via CPSP approach**



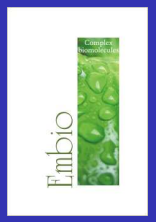
Martin Mann
Freiburg

Degeneracy

- = # of optimal structures
  $\longrightarrow$ measure for protein stability
- exactly determinable using CPSP
- usually very high in HP-Model
- e.g. to distinguish proteinlike or random sequences
- counting can be improved (’06)



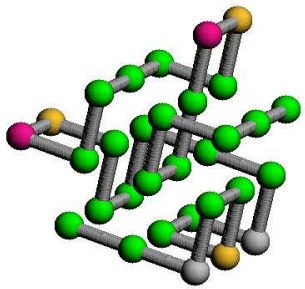
log_{10}(degeneracy) histogram
of 809 HP-sequences (H:P=1:1)
(32% with deg. > 10^6)



Martin Mann
Freiburg

Constraint-based Protein Structure Prediction

- Short introduction
  - Optimal structure prediction (CPSP)
- Applications in
  - Protein stability
  - Inverse folding problem



The Inverse Folding Problem

Design of proteinlike Sequences

Martin Mann
Freiburg

The goal

- design of sequences for a given structure

Sequence constraints

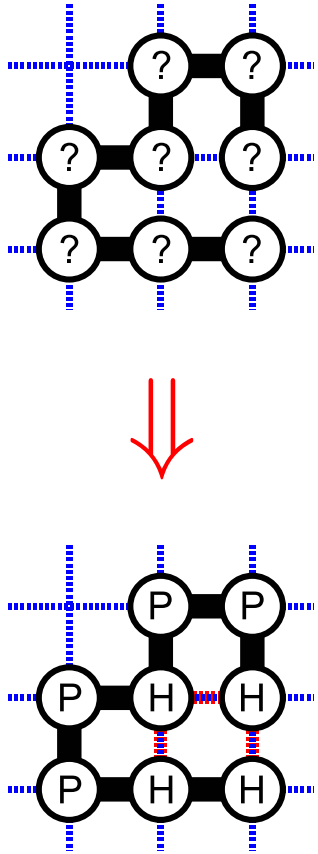
- proteinlike (low degeneracy)
- forms structure as optimal one

Problem

- # of sequences is expontentially in length

Addressed Questions (Designability)

- Is a structure X codeable?
- How many sequences code X?



The Inverse Folding Problem

Design of proteinlike Sequences

Martin Mann
Freiburg

The goal

- design of sequences for a given structure

Sequence constraints

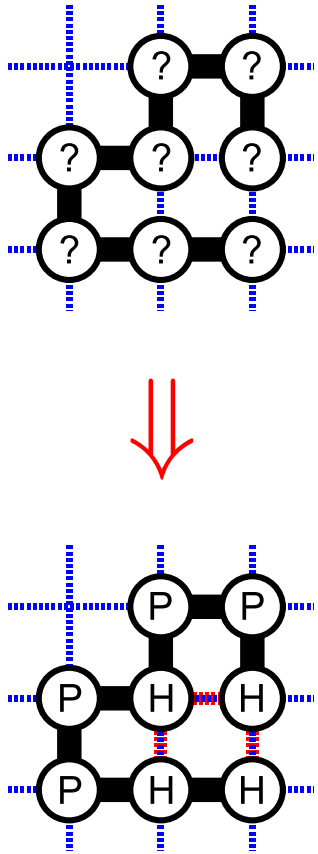
- proteinlike (low degeneracy)
- forms structure as optimal one

Problem

- # of sequences is expontentially in length

Addressed Questions (Designability)

- Is a structure X codeable?
- How many sequences code X?



$\Rightarrow$ **Solved using H-cores and CPSP . . .**

HPdesign

A 'Generate and Test' Workflow

Martin Mann
Freiburg

Input

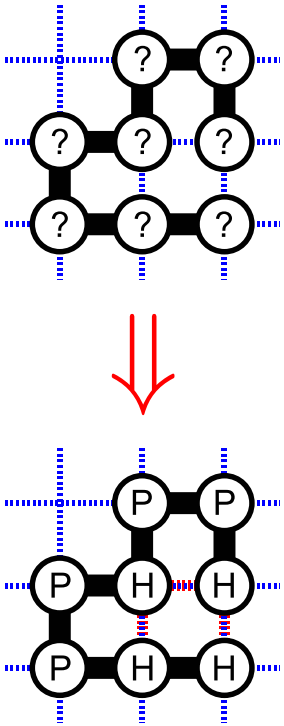
- a structure X

HPdesign workflow

1. Generate good candidates
2. Validate the sequences

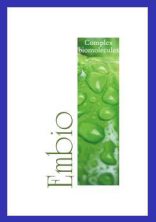
Output

- a set of sequences that:
  - form X as their optimum
  - are stable (degeneracy 1)



?

HPPHPPHH
HPHHPHHH
...



Martin Mann
Freiburg

Observation

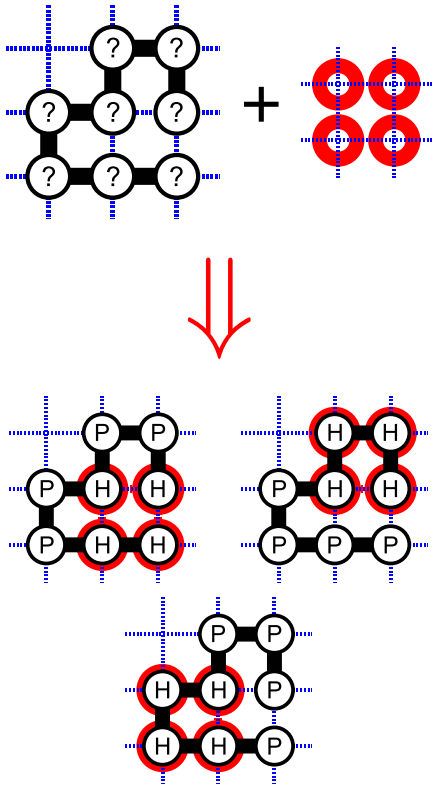
- optimal structure $\Leftrightarrow$ optimal H-core

Generation

- take an arbitrary optimal H-core *C*
- shift *C* through structure *X*
- store resulting sequence for each hit

Result

- sequences with high chance to form *X* as their optimal structure



HPdesign

Step 2 : Sequence Validation via CPSP Approach

Martin Mann
Freiburg

Task

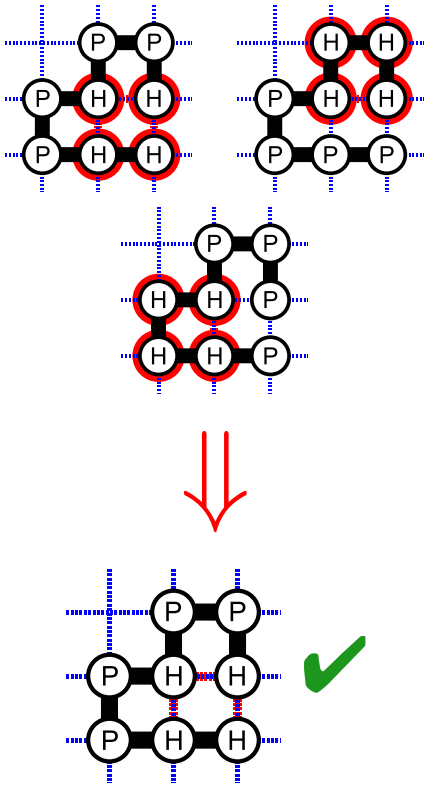
- Check for each sequences if it
  - is stable (degeneracy 1)
  - forms X as optimal structure

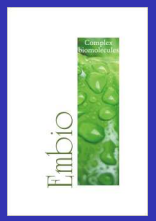
Workflow using CPSP

1: **for all** sequences S **do**
2: $\quad \mathscr{X} \leftarrow \text{CPSP}(S, max = 2)$
3: $\quad$ **if** $(|\mathscr{X}| = 1 \wedge X \in \mathscr{X})$ **then**
4: $\quad\quad$ STORE(S)
5: $\quad$ **end if**
6: **end for**

Result of Filtering

- stable S that form X as optimum

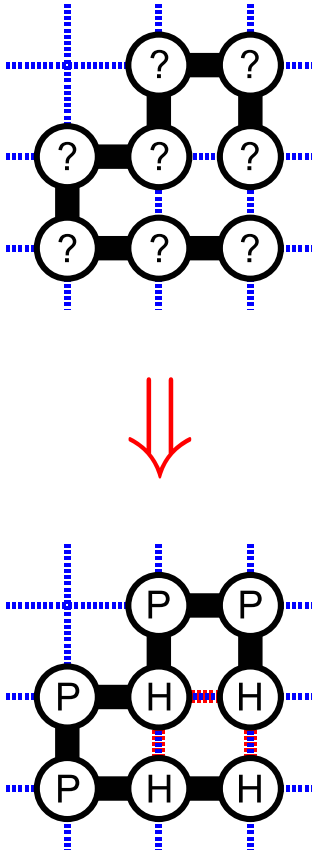


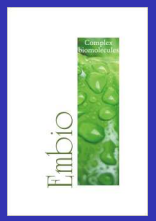


Martin Mann
Freiburg

The Inverse Folding Problem

- Design of stable sequences that form a given structure as optimum is a hard task
- Can be solved using 'Generate and Test'
- Candidate set can be shrinked (H-cores)
- Validation via CPSP approach possible

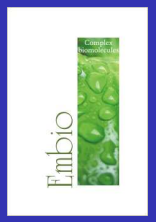




Martin Mann
Freiburg

Open . . .

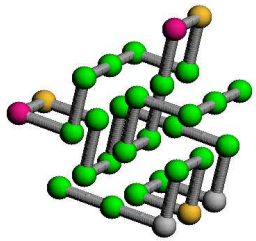
- What distinguishes proteinlike and random sequences?
- Relations to real protein sequence properties?
- Are there common patterns in stable structures?
- How big are the bassins of attraction of stable optima?
- Leads one optimal structure to a folding funnel?
- What makes a structure designable?
- . . .

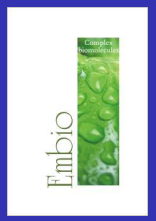


Martin Mann
Freiburg

Constraint-based Protein Structure Prediction

- **CPSP Approach**
  - **Enumeration of optimal structures**
  - **Fast and extensible**
- **Degeneracy**
  - **Important measure of protein stability**
- **Inverse folding problem**
  - **Find stable sequences that form a structure X as optimum**



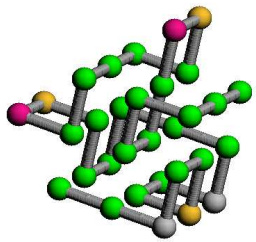


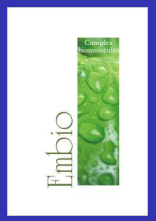
Martin Mann
Freiburg

Constraint-based Protein Structure Prediction

CPSP-tools

- **HPstruct:** CPSP Approach
  - Enumeration of optimal structures
  - Fast and extensible
- **HPdeg:** Degeneracy
  - Important measure of protein stability
- **HPdesign:** Inverse folding problem
  - Find stable sequences that form a structure X as optimum
- ...





Martin Mann
Freiburg

CPSP-tools

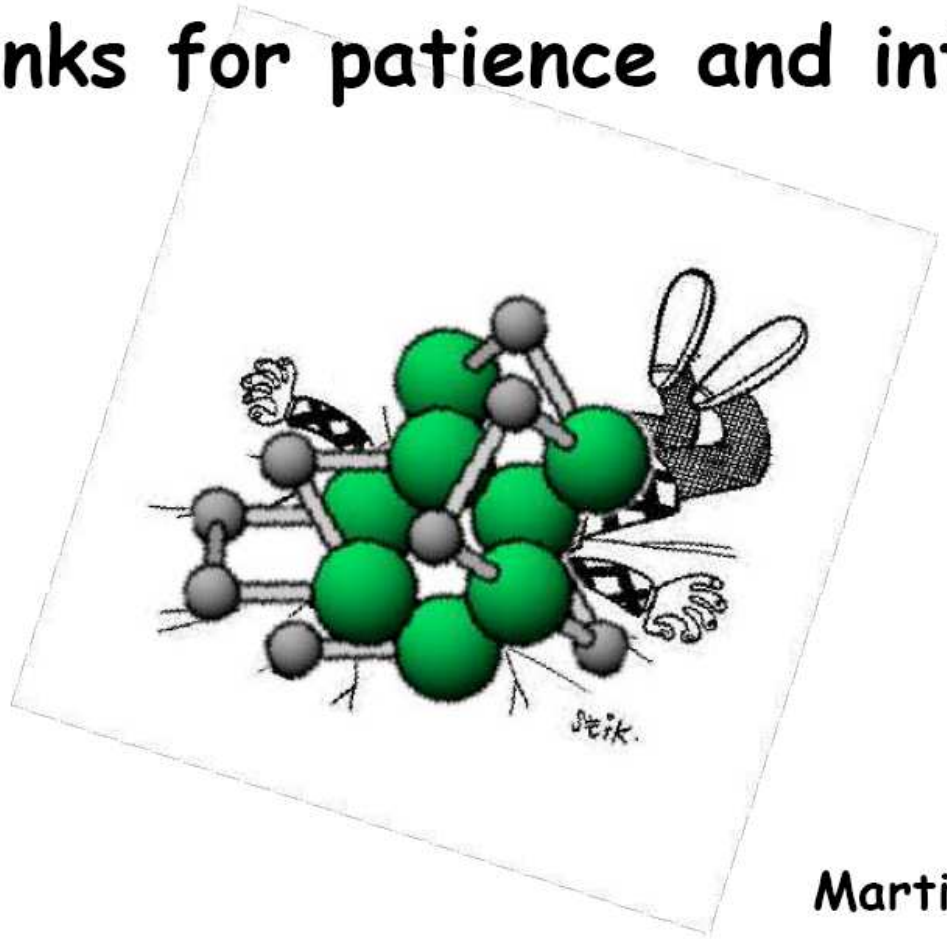
- Implements the CPSP approach etc.
- Object oriented C++
- Library of core classes and functionality
- Completely documented / API
- Freely available

http://www.bioinf.uni-freiburg.de/sw/cpcp/

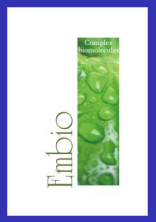
That's all folks!

Martin Mann
Freiburg

Thanks for patience and interest



Martin Mann

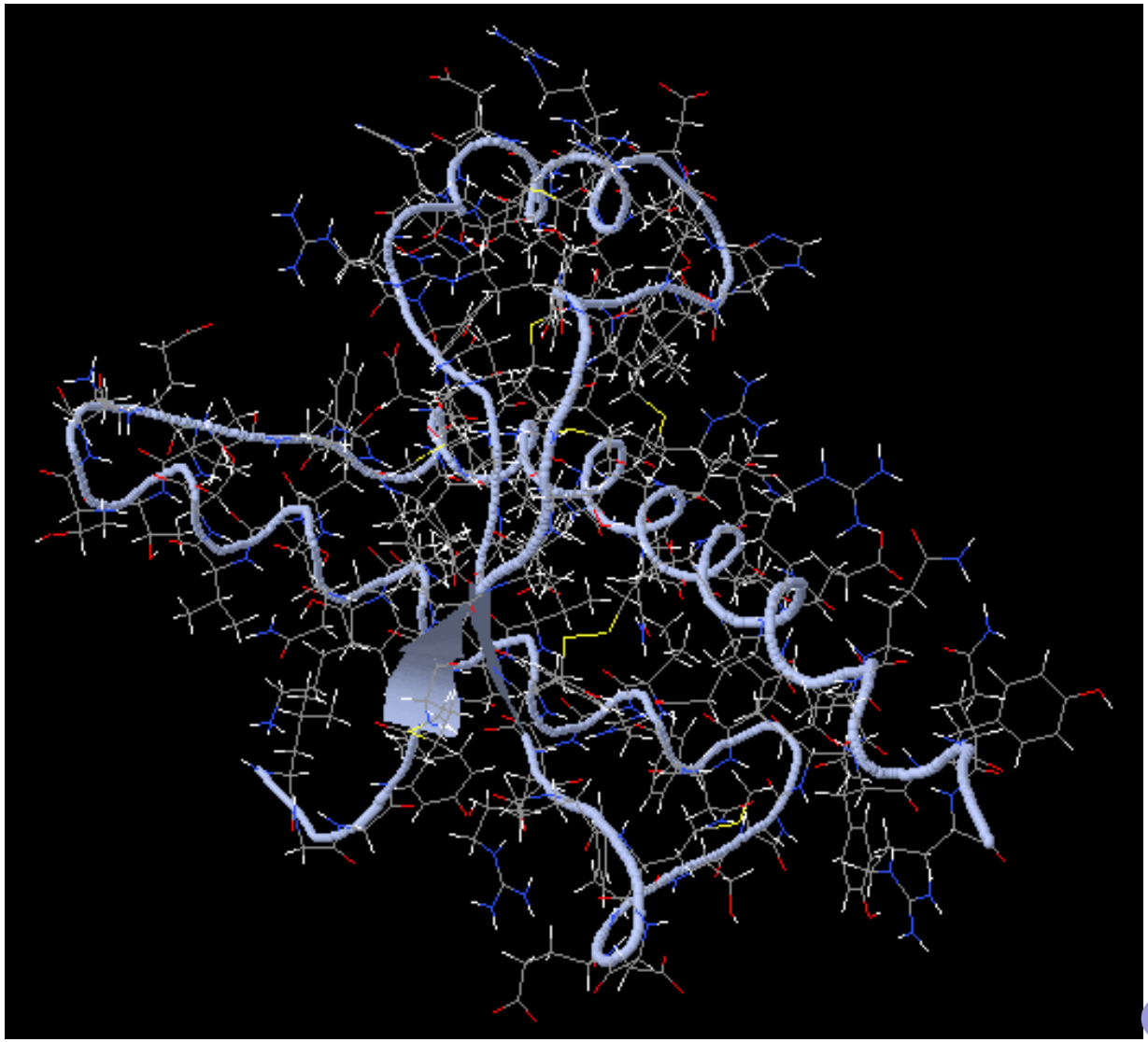


Martin Mann
Freiburg

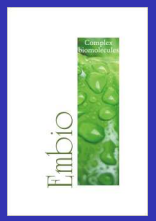
Appendix section

Protein folding and native structure prediction

Martin Mann
Freiburg



Zoom out



Martin Mann
Freiburg

Advantages

- discrete $\longrightarrow$ full enumeration for result validation
- computationally capable
- folding dynamics similar to real proteins (time scale)
- unique folders

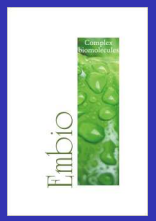
Possible Critics

- Energy function (HP) $\Rightarrow$ HPNX, ...
- Lattice type (3D-cubic) $\Rightarrow$ FCC, ...
- Lattice vs. angles $\Rightarrow$ discrete angle model, ...

Return

The CSP

A very very simple Formulation



Martin Mann
Freiburg

For a given HP-sequence and an optimal H-core:

Variables

- one for each sequence monomer

Domains = sets of lattice positions

- H-Monomers: H-core positions (ensures optimality)
- P-Monomers: remaining lattice

Constraints

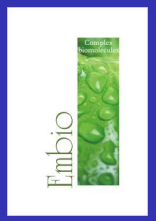
- binary Neighboring constraints along the chain (backbone)
- one global Alldifferent constraint (selfavoiding structure)

$\Rightarrow$ encodes the selfavoiding walk

Return

The CSP

A very very simple Formulation



Martin Mann
Freiburg

For a given HP-sequence and an optimal H-core:

Variables

- one for each sequence monomer

Domains = sets of lattice positions

- H-Monomers: H-core positions (ensures optimality)
- P-Monomers: remaining lattice

Constraints

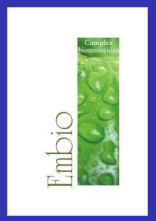
- binary Neighboring constraints along the chain (backbone)
- one global Alldifferent constraint (selfavoiding structure)

$\Rightarrow$ encodes the selfavoiding walk

Return

The CSP

A very very simple Formulation



Martin Mann
Freiburg

For a given HP-sequence and an optimal H-core:

Variables

- one for each sequence monomer

Domains = sets of lattice positions

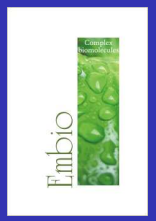
- H-Monomers: H-core positions (ensures optimality)
- P-Monomers: remaining lattice

Constraints

- binary Neighboring constraints along the chain (backbone)
- one global Alldifferent constraint (selfavoiding structure)

$\Rightarrow$ encodes the selfavoiding walk

Return



Martin Mann
Freiburg

For a given HP-sequence and an optimal H-core:

Variables

- one for each sequence monomer

Domains = sets of lattice positions

- H-Monomers: H-core positions (ensures optimality)
- P-Monomers: remaining lattice

Constraints

- binary Neighboring constraints along the chain (backbone)
- one global Alldifferent constraint (selfavoiding structure)

$\Rightarrow$ encodes the selfavoiding walk