BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 5 2009, pages 676–677
doi:10.1093/bioinformatics/btp034

*Structural bioinformatics*

# CPSP-web-tools: a server for 3D lattice protein studies

Martin Mann[1,*], Cameron Smith[1], Mohamad Rabbath[1], Marlien Edwards[2], Sebastian Will[1] and Rolf Backofen[1,*]

[1]Bioinformatics Group, University of Freiburg, Georges-Köhler-Allee 106, 79016 Freiburg, Germany and [2]German University in Cairo, Department of Computer Science, 5th Settlement New Cairo City, Cairo, Egypt

Received on December 23, 2008; revised on January 9, 2009; accepted on January 10, 2009

Advance Access publication January 16, 2009

Associate Editor: Dmitrij Frishman

**ABSTRACT**

**Summary:** Studies on proteins are often restricted to highly simplified models to face the immense computational complexity of the associated problems. Constraint-based protein structure prediction (CPSP) tools is a package of very fast algorithms for *ab initio* optimal structure prediction and related problems in 3D HP-models [cubic and face centered cubic (FCC)]. Here, we present CPSP-web-tools, an interactive online interface of these programs for their immediate use. They include the first method for the direct prediction of optimal energies and structures in 3D HP side-chain models. This newest extension of the CPSP approach is described here for the first time.

**Availability and Implementation:** Free access at http://cpsp.informatik.uni-freiburg.de

**Contact:** cpsp@informatik.uni-freiburg.de

## 1 INTRODUCTION

Lattice models are a common abstraction of protein structures to enable high-throughput studies in sequence and structure space (Huard *et al.*, 2006; Jacob *et al.*, 2007; Wolfinger *et al.*, 2006). This is achieved by a simplified representation of amino acids with one or a few monomers each, and a discretization of the structure space that restricts the positions of the monomers to nodes on a regular lattice. The modelling accuracy of structures depends on the underlying lattice (e.g. 3D cubic or 3D face-centered-cubic (FCC) and the number of monomers per amino acid. Arbitrary accuracy comes at the cost of computational complexity. Even in the simplest lattice model, the 2D-square HP model, problems of interest as *ab initio* optimal structure prediction and inverse folding stay computationally demanding (NP-complete) (Berger and Leighton, 1998; Berman *et al.*, 2004). Solving such problems for complex full atom models is currently completely infeasible.

The HP model is widely used in literature and groups amino acids into (H)ydrophobic and (P)olar to focus on hydrophobic forces in protein structures. This is done by maximizing (HH) contacts between neighbored H-monomers resulting in a compact hydrophobic core (Dill, 1985). In the original model each amino acid is represented by a single monomer but, as discussed later, different models also representing side chains have been developed.

Here, we present a web interface to constraint-based protein structure prediction (CPSP)-tools (Mann *et al.*, 2008b), a package of exact and complete methods to cope with tasks in the field of HP lattice protein studies. The algorithms are based on the approach by Backofen and Will (2006) utilizing advanced techniques like constraint programming (see CPSP section). This approach enables the calculation of one or all optimal structures (not possible using stochastic approaches) and has opened up new vistas for protein studies. The CPSP-tools package is designed for fast high-throughput experiments in 3D lattices (Mann *et al.*, 2008a; Wolfinger *et al.*, 2006). All CPSP-tools support 3D lattices such as the 3D-FCC lattice. The latter was shown to yield highly accurate fits of real protein structures (Park and Levitt, 1995).

Our CPSP-web-tools enable direct instant access to the CPSP-tools when no high-throughput experiments are needed. The package serves as a platform for research and teaching in the field of HP protein models. The combination of interactive result visualization including 3D views of the structures, the interlink of the different tools for chained applications and the fast runtimes of the CPSP-tools results in a useful service for the end-user.

In particular, we present the latest extension of the CPSP approach that enables for the first time the prediction of optimal structures in the 3D HP side-chain model (Bromberg and Dill, 1994). Here, each amino acid is represented by two monomers, one for the side-chain residue group and one for the backbone atoms. While the backbone is considered to be neutral, the side-chain monomers are distinguished into (H)ydrophic and (P)olar and, following the original model, a maximization on HH-contacts between H-side-chain monomers is sought. This more detailed model is closer to real protein structures, while still focusing on a compact hydrophobic core. Still, the extended CPSP approach is able to tackle the increased complexity and calculates optimal energies and structures within seconds.

## 2 CPSP-WEB-TOOLS

The CPSP-web-tools provide a direct and interactive web platform for the programs contained in the CPSP-tools package (Mann *et al.*, 2008b). It focuses on *ad hoc* and experimental usage of the programs as needed when one is only interested in a few experiments or teaching. The JavaServer Pages based platform supports workflow oriented operations by providing direct tool application on the results of a previous request. For example, an optimal structure calculated by HPstruct can be forwarded into the sequence design front end of HPdesign to expand afterwards the corresponding neutral network (HPnnet). It is the first and only online service that interfaces

*To whom correspondence should be addressed.

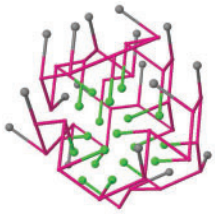

**Fig. 1.** An optimal structure (energy −55) of HPPHHPPPHPHHPHHPPH-PHPPHHHPHHPPHPHPH in the 3D FCC side-chain HP model.

programs for these tasks and is able to handle HP sequences of more than 100 amino acids in length. Extensive javascripting is utilized to provide interactive input validation to the user. All CPSP-web-tools provide detailed help on the required parameters. Further help and information concerning the CPSP approach, the tool package itself and general concepts in the field of lattice proteins is provided by a collection of frequently asked questions. The interactive 3D visualization of lattice protein structures utilizes the Jmol molecule viewer (Herrez, 2006). For an example see Figure 1. Programs and tasks currently part of the CPSP-web-tools interface are:

- HPstruct: optimal energy and structure calculation for backbone and side-chain model (see next section).
- HPview: interactive visualization of lattice protein structures.
- HPconvert: conversion between different structure formats like 3D-coordinates, PDB, CML, etc.
- HPdeg: calculation of a sequence's degeneracy, i.e. the number of optimal structures it can adopt.
- HPnnet: expansion of the neutral network of a given sequence, i.e. the interlinked fraction of sequence space that stably share a common structure.
- HPdesign: design of sequences that fold stably into a given structure.

The CPSP-web-tools interface is being constantly expanded such that further tools and features will become available.

## 3 CPSP-APPROACH AND SIDE-CHAIN MODELS

The CPSP-approach (Backofen and Will, 2006) is based on a database of so called *H-cores* following the observation that optimal structures show a compact placement of H-monomers. An optimal H-core is a set of H-monomer positions allowing for the maximal number of HH-contacts.

For a concrete sequence $S$ the approach systematically examines the list of H-cores compatible with $S$ in decreasing contact number. For each core, it attempts to find a placement of the monomers of $S$ in a self-avoiding walk such that all H-monomers are elements of the given H-core and all P-monomers are outside of the core. Since the H-cores are considered in the order of decreasing contacts, the first successful threading results in a structure with global minimal energy. Note that at this point the algorithm has *proven* that there is no structure of $S$ that forms more HH-contacts.

Technically, the threading of a sequence through a core is performed by a constraint program. A constraint satisfaction problem (CSP) is formulated that constrains the H-monomers to positions in the H-core. Further, it enforces successive monomers along the sequence to be neighbored in the lattice and prohibits the multiple use of a single position. The constraint-programming machinery allows for the enumeration of all valid placements according to the given constraints. In this way, all optimal structures for a given sequence can be calculated. For details of the CSP definition and the mechanisms for solving this model see Backofen and Will (2006).

Here, we describe the newest extension of the CPSP-approach, to enable the prediction of *optimal structures in side-chain HP-models*. As introduced above, the side-chain model focuses on a compact placement of H-side-chain monomers. The extension consists of the formulation of a new CSP that constrains only the H-side-chains to positions in the H-core. Additionally, successive backbone monomers as well as the side-chain and backbone monomer of each amino acid have to be neighbored in the lattice. The new CSP is exploited within the original CPSP framework analogously to the standard CSP. Thus, the advanced constraint programming machinery previously developed is reused in the calculation of optimal structures in 3D side-chain HP models.

## 4 CONCLUSION

The CPSP-web-tools are the online interface of the newest version of the CPSP-tools package (Mann *et al.*, 2008b) including algorithms for lattice protein studies. While the CPSP-tools are made for high-throughput experiments, the web-tools focus on *ad hoc* usage for research and teaching. The user can immediately work with the CPSP-tools 'out of the box' which are combined with an interactive, interlinked result management system. The outcome of this is a the first web service that answers the question for optimal energy and structures, degeneracy, neutral networks and other properties in the field of HP-models in 3D lattices.

Furthermore, we have presented the latest extension of the CPSP approach that enables the prediction of optimal energy and structures in the side-chain HP-model. This is the first method that is able to calculate all optimal side-chain structures of a given sequence, while proving their optimality. This extension is the base for new studies in more realistic protein models that are still computationally feasible.

## ACKNOWLEDGEMENTS

Thanks to Stefan Jankowski for his help with the server setup.

*Funding*: EU project EMBIO (EC contract number 012835).

*Conflict of Interest*: none declared.

## REFERENCES

Backofen,R. and Will,S. (2006) A constraint-based approach to fast and exact structure prediction in three-dimensional protein models. *Constraints*, **11**, 5–30.

Berger,B. and Leighton,T. (1998) Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *J. Comp. Biol.*, **5**, 27–40.

Berman,P. *et al*. (2004) The protein sequence design problem in canonical model on 2D and 3D lattices. In *Proc. of Combinatorial Pattern Matching*. 3109, LNCS, Springer, pp. 244–253.

Bromberg,S. and Dill,K.A. (1994) Side-chain entropy and packing in proteins. *Protein Sci.*, **3**, 997–1009.

Dill,K.A. (1985) Theory for the folding and stability of globular proteins. *Biochemistry*, **24**, 1501–1509.

Herrez,A. (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem. Educ.*, **34**, 255–261.

Huard,F.P.E. *et al*. (2006) Modelling sequential protein folding under kinetic control. *Bioinformatics*, **22**, e203–e210.

Jacob,E. *et al*. (2007) Different mechanistic requirements for prokaryotic and eukaryotic chaperonins: a lattice study. *Bioinformatics*, **23**, i240–i248.

Mann,M. *et al*. (2008a). Classifying protein-like sequences in arbitrary lattice protein models using LatPack. *HFSP J.*, **2**, 396.

Mann,M. *et al*. (2008b) CPSP-tools – exact and complete algorithms for high-throughput 3D lattice protein studies. *BMC Bioinformatics*, **9**, 230.

Park,B.H. and Levitt,M. (1995) The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.*, **249**, 493–507.

Wolfinger,M. *et al*. (2006) Exploring the lower part of discrete polymer model energy landscapes. *Europhysics Lett.*, **74**, 725–732.