

# Team meeting 2015

**Sune S. Nielsen**

Computer Science and Communications (CSC)

CSC | COMPUTER SCIENCE  
AND COMMUNICATIONS  
RESEARCH UNIT

# Outline

- 1 **Introduction**
  - My Thesis
- 2 **Inverse protein folding**
  - Overview
- 3 **Optimisation**
  - Main Objective Function
  - Diversity Objective
  - Diversity-As-Objective Genetic Algorithm
  - Quantile constraint
  - Preference Based Genetic Algorithm
- 4 **Validation Experiments**
  - Overview
  - Validation of tertiary structure
- 5 **NK Model for IFP Benchmarks**
  - NK Model
  - Proposed NKL Model setup
- 6 **Future work**
- 7 **Bibliography**

# Outline

- 1 **Introduction**
  - My Thesis
- 2 **Inverse protein folding**
  - Overview
- 3 **Optimisation**
  - Main Objective Function
  - Diversity Objective
  - Diversity-As-Objective Genetic Algorithm
  - Quantile constraint
  - Preference Based Genetic Algorithm
- 4 **Validation Experiments**
  - Overview
  - Validation of tertiary structure
- 5 **NK Model for IFP Benchmarks**
  - NK Model
  - Proposed NKL Model setup
- 6 **Future work**
- 7 **Bibliography**

# Thesis overview

- Decision-theoretic Fine Tuning of Multi-objective Co-Evolutionary Algorithms
  - Prof. Pascal Bouvry, Prof. Nikos Vlassis, Dr Grégoire Danoy
  - Started 15. September 2011
  - Part of Evoperf work package 3
- Interdisciplinary research
  - Luxembourg Centre for Systems Biomedicine (LCSB)
- Changes to plan meanwhile
  - Prof. Nikos Vlassis has been replaced by Prof. Reinhard Schneider (September)
  - Six months of parental leave (January-June)
  - One year extension of AFR grant

# Thesis work areas

- Co-evolutionary optimisation
  - Fitness evaluation is based on interaction of multiple individuals, sub-populations or agents
- Multi-modal and diversity preserving optimisation
  - Multiobjectivization with diversity as objective
  - Quantile Constraint approach to maintain certain diversity level
  - Preference based Genetic Algorithm (PBGA)
  - Adaptive diversity control ongoing
- Bio-informatics problem
  - Inverse Folding Problem (IFP) for proteins
- Benchmark problem
  - Mimics the properties of the IFP problem

# Outline

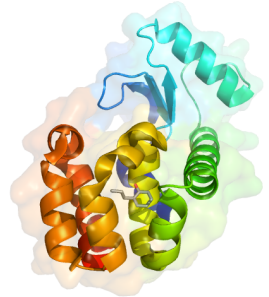
- 1 **Introduction**
  - My Thesis
- 2 **Inverse protein folding**
  - Overview
- 3 **Optimisation**
  - Main Objective Function
  - Diversity Objective
  - Diversity-As-Objective Genetic Algorithm
  - Quantile constraint
  - Preference Based Genetic Algorithm
- 4 **Validation Experiments**
  - Overview
  - Validation of tertiary structure
- 5 **NK Model for IFP Benchmarks**
  - NK Model
  - Proposed NKL Model setup
- 6 **Future work**
- 7 **Bibliography**

# Motivation

- New coarse grained approach to old problem: inverse protein folding problem [BE93]
  - Explore sequence space leading to similar folded protein structure
  - Based on secondary structure prediction
- Simplification of the **ab initio** protein folding problem
  - Structure is known and compatibility of sequences is to be determined
- Ultimately the results may be used for protein design
  - Here the designer provides desired structure, program compatible sequences
  - Improved enzymes for waste-water treatment or biomass production etc.
  - Optimised antibodies designed specific towards known targets
  - New medical intervention paths

# Protein structure

- Primary structure
  - The sequence composed of the 20 amino acids
  - Sequence is typically 50 to 1000 amino acids long
- Secondary structure
  - Organisation or annotation of alpha-helices and beta-sheets
- Tertiary structure
  - The actual three dimensional placement of atoms in space
- Quarternary structure
  - The sub-units placements



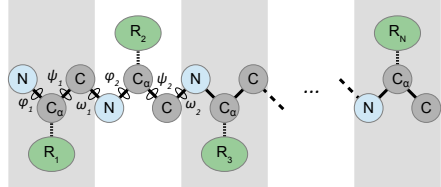
Primary structure

$aa_1$	$aa_2$	$aa_3$	...	$aa_N$
--------	--------	--------	-----	--------

Secondary structure

$T_1$	$T_2$	$T_3$	...	$T_N$
-------	-------	-------	-----	-------

Tertiary structure - backbone and sidechains shown





# Outline

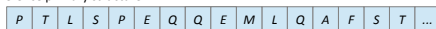
- 1 **Introduction**
  - My Thesis
- 2 **Inverse protein folding**
  - Overview
- 3 **Optimisation**
  - Main Objective Function
  - Diversity Objective
  - Diversity-As-Objective Genetic Algorithm
  - Quantile constraint
  - Preference Based Genetic Algorithm
- 4 **Validation Experiments**
  - Overview
  - Validation of tertiary structure
- 5 **NK Model for IFP Benchmarks**
  - NK Model
  - Proposed NKL Model setup
- 6 **Future work**
- 7 **Bibliography**

# Secondary structure prediction

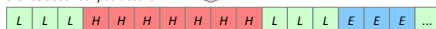
- Per amino acid  $aa_i$  at location  $i \in \{1..N\}$  in a sequence  $A$  we obtain
- Using ReProf / PROFphd [RS94]
  - Likely secondary structure type  $T_{pred}(i)$  with reliability  $R_{pred}(i) \in \{1..10\}$

$$F_{sec}(A) = \frac{\Sigma_{max} - \sum_{i=1}^N M_i}{\Sigma_{max}} . \quad (1)$$

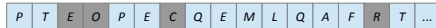
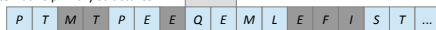
Reference primary structure



Reference secondary structure



Alternative primary structures



# Diversity preservation and niching

- Problem specific motivation
  - Finding very **diverse** nucleotide sequences which produce **similar** structures
  - Contrary to finding a set of sequences with little diversity
- Algorithm specific motivation
  - High diversity is a valuable feature to prevent premature convergence
  - Multi-modality of the problem

## Diversity-as-objective - method

- Definition of an additional diversity objective

- Hamming distance  $d_{Ham}(A, A')$  between amino acid sequences  $A$  and  $A'$ :

$$d_{Ham}(A, A') = \sum_{i=1}^N d_i, \quad d_i = \begin{cases} 0 & \text{if } aa_i = aa'_i \\ 1 & \text{otherwise} \end{cases} . \quad (2)$$

- Diversity objective is calculated as follows:

$$F_{div}(A) = N - \frac{1}{M-1} \sum_{i=1}^{M-1} d_{Ham}(A, A_i) . \quad (3)$$

- Described with words: the average inverted distance to all the other individuals

# Quantile constraint - motivation and method

- Complexity of the two objectives is very unbalanced
  - Sequences with optimal diversity objective can easily be obtained
  - Sequences with good predicted structure match are hard to find
- Quantile constraint allows to adjust the priority of one objective over the other(s)
  - High population diversity  $\Rightarrow$  Low average fitness with high std. deviation
  - Adjust trade-off between exploration vs. exploitation
- Applied each generation of the bi-objective algorithm
  - Remove doubles because identical sequences lead to poor diversity
  - Divide population according to  $F_{sec}(A)$  into a  $C_q\%$  and a  $100 - C_q\%$  quantile
  - Assign a constraint to the  $C_q\%$  quantile of less fit individuals and prevent the worst individuals from mating next generation

# Overview

- The users preference described through a Weighted Sum Model (WSM) is used to maintain a best fulfilling population.

$$WSM_{score}(P) = -W_{fit} \cdot F_{fit}(P) + W_{div} \cdot F_{div}(P) \quad (4)$$

- Iteratively, the weakest individuals from the combination of parent and offspring populations are determined and removed until the desired population size is achieved

## Pseudo code

---

**Algorithm 1:** Preference-Based Genetic Algorithm

---

```
1: Initialise ( $P_0$ )
2:  $t \leftarrow 0$ 
3: while  $t < t_{max}$  do
4:    $Q_t \leftarrow \text{makeNewOffspringPop}$  ( $P_t$ )
5:    $R_t \leftarrow P_t + Q_t$ 
6:   while  $|R_t| > |P_t|$  do
7:      $I \leftarrow \text{getWeakestIndividual}$  ( $R_t$ )
8:      $R_t \leftarrow R_t - I$ 
9:   end while
10:   $P_t \leftarrow R_t$ 
11:   $t \leftarrow t + 1$ 
12: end while
```

---

# Outline

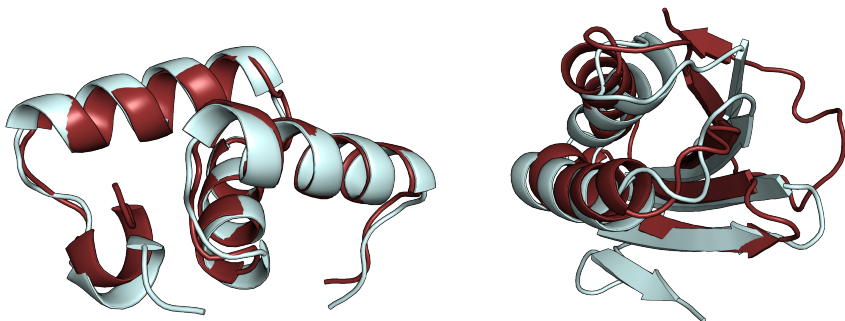
- 1 **Introduction**
  - My Thesis
- 2 **Inverse protein folding**
  - Overview
- 3 **Optimisation**
  - Main Objective Function
  - Diversity Objective
  - Diversity-As-Objective Genetic Algorithm
  - Quantile constraint
  - Preference Based Genetic Algorithm
- 4 **Validation Experiments**
  - Overview
  - Validation of tertiary structure
- 5 **NK Model for IFP Benchmarks**
  - NK Model
  - Proposed NKL Model setup
- 6 **Future work**
- 7 **Bibliography**



# Workflow overview

- Select 5 best sequences of 30 optimisation runs for further study.
- Predict their tertiary structure with I-TASSER4.0 [YYR<sup>+</sup>15]  
⇒ total of 300 I-TASSER runs, almost two years of CPU-time.
- Estimate structure similarity
  - Compare secondary structure
  - Local and global structure alignment scores with LGA tool [Zem03]
  - TM-Score[ZS04] measuring structure similarity between two structures
    - Score above 0.5 can be considered the same fold[XZ10]
  - Visual superposition of best generated sample and the reference

## Tertiary comparison by superpositioning



- Dark red: I-TASSER predicted model, light blue: reference structure

# Tertiary comparison by superpositioning

- Summary of tertiary structure prediction match:

Protein	$\mu_{TM-Score}$	$\sigma_{TM-Score}$	$N_{TM>0.2}$	$N_{TM>0.4}$	$N_{TM>0.5}$	$N_{TM>0.6}$	$N_{TM>0.7}$	$N_{TM>0.8}$
1OAI	0.493	0.135	150	102	51	32	18	4
1URR	0.416	0.061	150	91	10	0	0	0

- Very similar structures were generated
- 1 in 5 and 1 in 15 resp. can be regarded having the same fold

# Outline

- 1 **Introduction**
  - My Thesis
- 2 **Inverse protein folding**
  - Overview
- 3 **Optimisation**
  - Main Objective Function
  - Diversity Objective
  - Diversity-As-Objective Genetic Algorithm
  - Quantile constraint
  - Preference Based Genetic Algorithm
- 4 **Validation Experiments**
  - Overview
  - Validation of tertiary structure
- 5 **NK Model for IFP Benchmarks**
  - NK Model
  - Proposed NKL Model setup
- 6 **Future work**
- 7 **Bibliography**

# Original definition

- Introduced by Kaufmann [KW89]
  - A tunable rugged fitness function designed to model complex epistatic links among variables
  - Used to study topics such as gene-interaction

$$F_{NK}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N F_i(x_i; x_{i_1}, \dots, x_{i_k}), \mathbf{x} \in \{0, 1\}^N$$

- Decomposition
  - $\mathbf{x}$  is a vector of  $N$  bits
  - Each position in  $\mathbf{x}$  interacts with  $K$  neighbours
  - This interaction is dictated by the neighbourhood definition
    - E.g.  $K$  adjacent loci or  $K$  random
  - Function  $F_i$  takes  $K + 1$  bits and returns a predefined uniformly distributed value in  $[0, 1]$ .

## Modified NKL definition

- Li **et al.** extended the NK Model to continuous and mixed integer solution spaces [LEE<sup>+</sup>06]
- The **nominal discrete NKL model** is of particular interest
  - $L = 2$  defines the binary case corresponding to the original model
  - $L = 20$  provides an encoding similar to a amino-acid chain of proteins
- Additional modification to the NKL model
  - the  $i$ th position is made optional

$$F(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N F_0(x_{i1}, \dots, x_{ik}), \mathbf{x} \in \{0, L\}^N$$

# Benchmark model design

- Combine two NKL Models with different neighbourhoods
- Model 1
  - $F^A$ : a  $K = 4$  semi-adjacent circular neighbourhood is designed as follows:  $\{x_{i-2}, x_{i-1}, x_{i+1}, x_{i+2}\}$ , omitting the central position  $x_i$ .
  - $F^B$ : a  $K = 3$  neighbourhood of uniform random distribution.
- Model 2
  - $F^A$ : a  $K = 4$  neighbourhood as Model 1.
  - $F^B$ : a  $K = 5$  neighbourhood of uniform random + 20 positions wide triangular distribution.

Available for download at:

<http://nk-ifp-bench.gforge.uni.lu/>

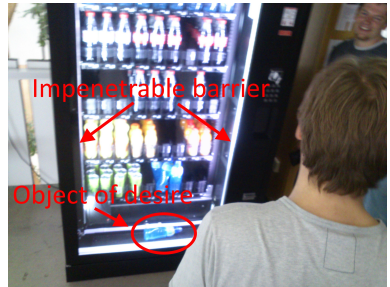
# Outline

- 1 **Introduction**
  - My Thesis
- 2 **Inverse protein folding**
  - Overview
- 3 **Optimisation**
  - Main Objective Function
  - Diversity Objective
  - Diversity-As-Objective Genetic Algorithm
  - Quantile constraint
  - Preference Based Genetic Algorithm
- 4 **Validation Experiments**
  - Overview
  - Validation of tertiary structure
- 5 **NK Model for IFP Benchmarks**
  - NK Model
  - Proposed NKL Model setup
- 6 **Future work**
- 7 **Bibliography**



# Future work

- Thesis writing and defense
  - Hopefully with more success than other peoples lemonade purchase! ⇨
- Adaptive Diversity Control
  - With the help of student Christof



## We made it in 10 minutes!

- Thanks for listening
- Questions?
- Non-research related contributions
  - A little sister for Mick!



# Outline

- 1 **Introduction**
  - My Thesis
- 2 **Inverse protein folding**
  - Overview
- 3 **Optimisation**
  - Main Objective Function
  - Diversity Objective
  - Diversity-As-Objective Genetic Algorithm
  - Quantile constraint
  - Preference Based Genetic Algorithm
- 4 **Validation Experiments**
  - Overview
  - Validation of tertiary structure
- 5 **NK Model for IFP Benchmarks**
  - NK Model
  - Proposed NKL Model setup
- 6 **Future work**
- 7 **Bibliography**

**James U. Bowie and David Eisenberg.**

Inverted protein structure prediction.

*Current Opinion in Structural Biology*, 3(3):437–444, June 1993.

**Stuart A Kauffman and Edward D Weinberger.**

The nk model of rugged fitness landscapes and its application to maturation of the immune response.

*Journal of theoretical biology*, 141(2):211–245, 1989.

**Rui Li, Michael TM Emmerich, Jeroen Eggermont, Ernst GP Bovenkamp, Thomas Bäck, Jouke Dijkstra, and Johan HC Reiber.**

Mixed-integer nk landscapes.

In *Parallel Problem Solving from Nature-PPSN IX*, pages 42–51. Springer, 2006.

**B. Rost and C. Sander.**

Combining evolutionary information and neural networks to predict protein secondary structure.

*Proteins*, 19(1):55–72, May 1994.

**Jinrui Xu and Yang Zhang.**

How significant is a protein structure similarity with tm-score= 0.5?

*Bioinformatics*, 26(7):889–895, 2010.

**Jianyi Yang, Renxiang Yan, Ambrish Roy, Dong Xu, Jonathan Poisson, and Yang Zhang.**

The i-tasser suite: protein structure and function prediction.

*Nature methods*, 12(1):7–8, 2015.

**A. Zemla.**

LGA: A method for finding 3D similarities in protein structures.

*Nucleic acids research*, 31(13):3370–3374, July 2003.

**Yang Zhang and Jeffrey Skolnick.**

Scoring function for automated assessment of protein structure template quality.

*Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.