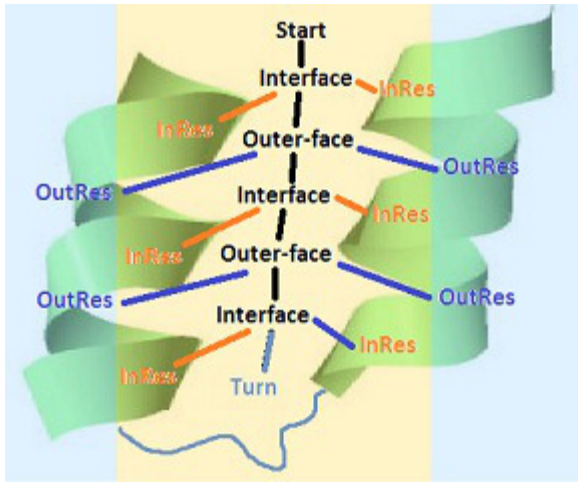
Wrocław University of Technology

Institute of Biomedical Engineering and Instrumentation

KotulskaLab

Towards 3D modeling of interacting transmembrane helix pairs based on classification of helix pair sequence



Witold Dyrka, Jean Christophe-Nebel, Małgorzata Kotulska

witold.dyrka@pwr.wroc.pl

Nijmegen, NL, 23.09.2010

Transmembrane (TM) proteins

Transmembrane proteins represent:

- 30% of human genome
- 50-60% of drug targets
- but only 1% of known structures in Protein Data Bank

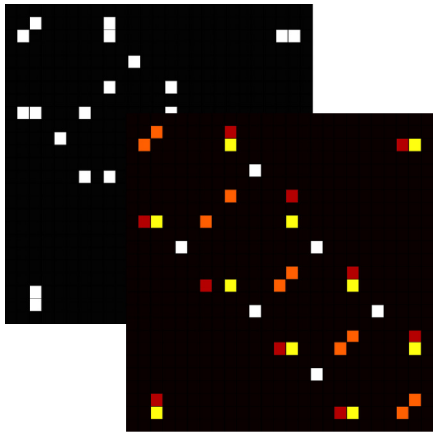
Structure prediction methods:

- Homology-based depend on availability of known structures
- Ab initio are successfull for proteins up to 200 amino-acids

for longer chains→ search space needs to be constrained

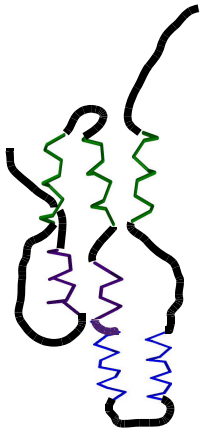
Motivation

- Contact maps are state-of-the-art constraints for ab initio protein structure prediction
- We present the method to enhance transmembrane helix-helix contact maps by adding information on the interaction conformation class

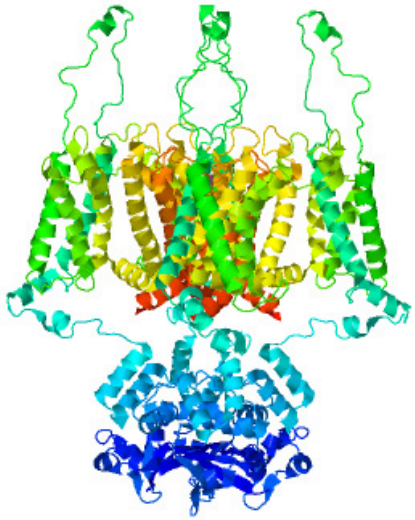


Enhance contact map

Create spatial constraints



Constrain ab inito prediction



Motivation

- Contact maps are state-of-the-art constraints for ab initio protein structure prediction
- We present the method to enhance transmembrane helix-helix contact maps by adding information on the interaction conformation class

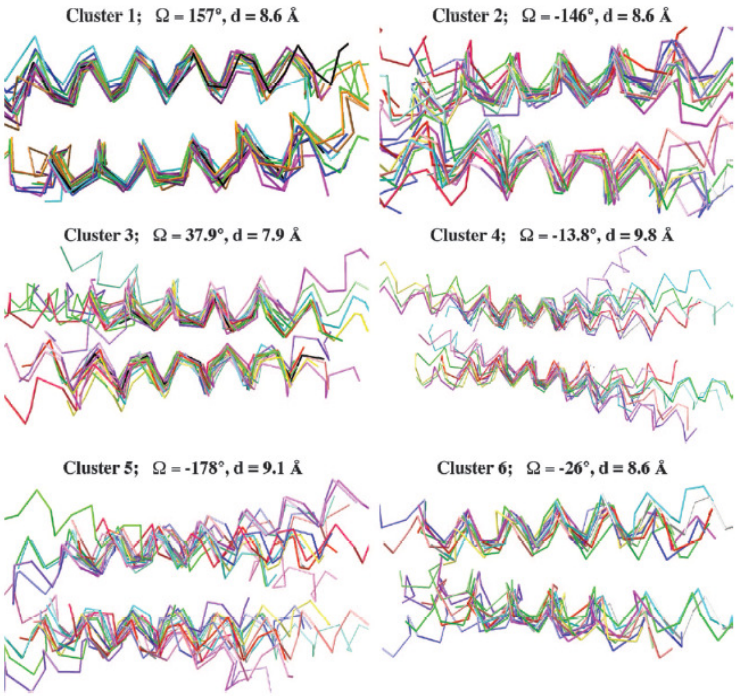
- Characterise sequence-structure relation in α-helical transmembrane proteins
- Understand the nature of polymeric molecules: analogies with natural language sentences

Patterns in helix-helix interfaces

Structure-level classes

- 90% of helix-helix interactions in TM proteins can be accurately (rmsd ≤1.5Å) represented by a set of 8 templates, which differ in 3D shapes

(Walters and DeGrado, PNAS 2006)



Walters and DeGrado, PNAS 2006

Sequence-level motifs

- Some short sequences or motifs, eg. GX_3G are common in helices, which form interfaces

(Russ and Engelman, JMB 2000)

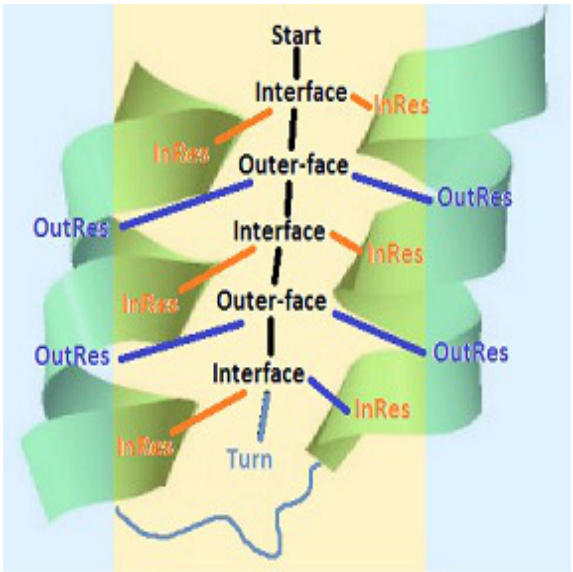
Memotif db. Marsico et al. NAR 2010

Context-Free grammatical model of helix-helix interactions

Pairwise nested dependencies between residues can be represented by Context-Free Grammar production rules*:

Start → Any Interface Any | Any Outer-face Any
Outer-face → OutsideResidues1 Interface OutsideResidues2| Any
Interface → InsideResidues1 Outer-face InsideResidues2 | Any

based on Waldispuehl J, Steyaert J-M. TCS 335:67-92 (2005)



* W Dyrka, JC Nebel, A Stochastic Context Free Grammar based Framework..., BMC Bioinformatics 2009 10:323)

Stochastic Context-Free Grammar induction – Evolutionary Algorithm

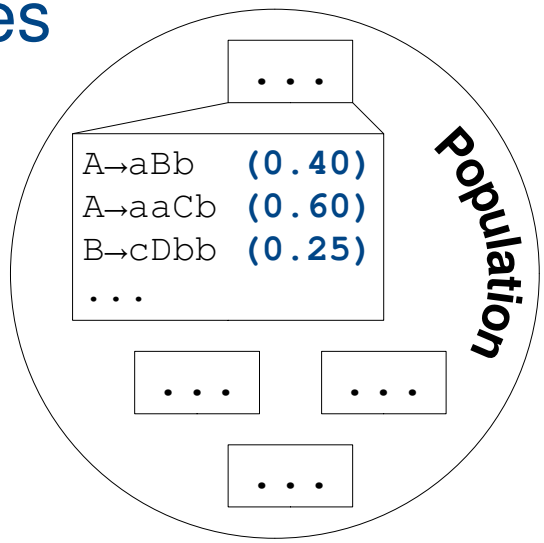
Input:

- Set of Context-Free Grammar rules
- Positive training set of amino-acid sequences of helix pairs representing a given class of conformation

Output: Probabilities for grammar rules

Population: ~200 individuals

- Individual is a real number vector representing rule probabilities



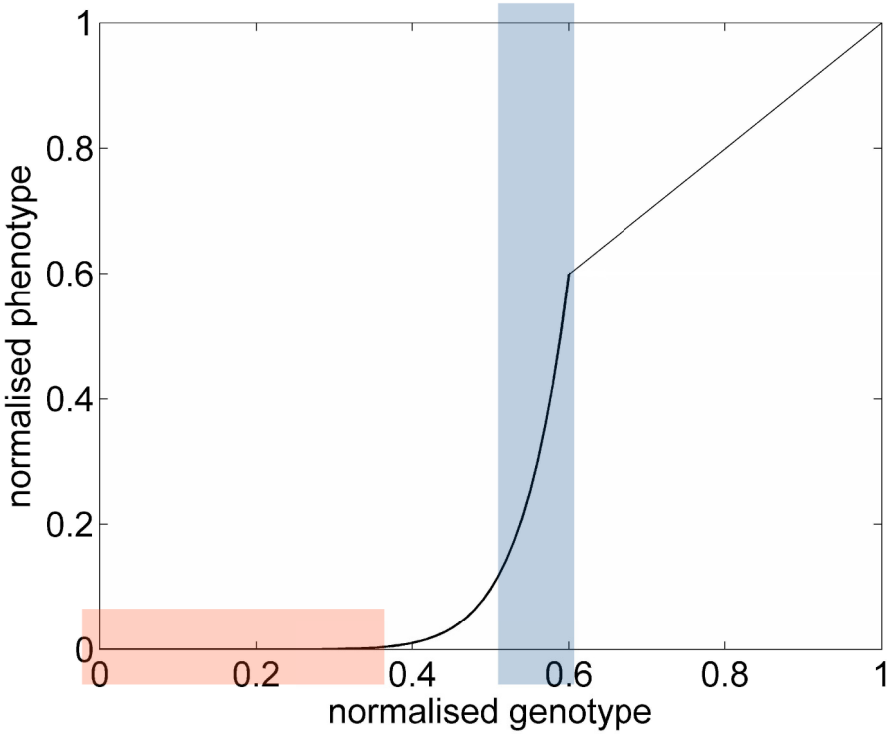
Objective function:

$$f(x) = \frac{\sum_{i}^{|S|} \log \Pr(W_i|G)}{|S|},$$

Pr(Wi|G) – log probability that sequence Wi from training set S belongs to language described by grammar G

Evolutionary Algorithm features

- Parallel processing of the training sample
- Novel **genotype** to **phenotype** function:
  - enhanced exploratory capabilities
  - faster computation

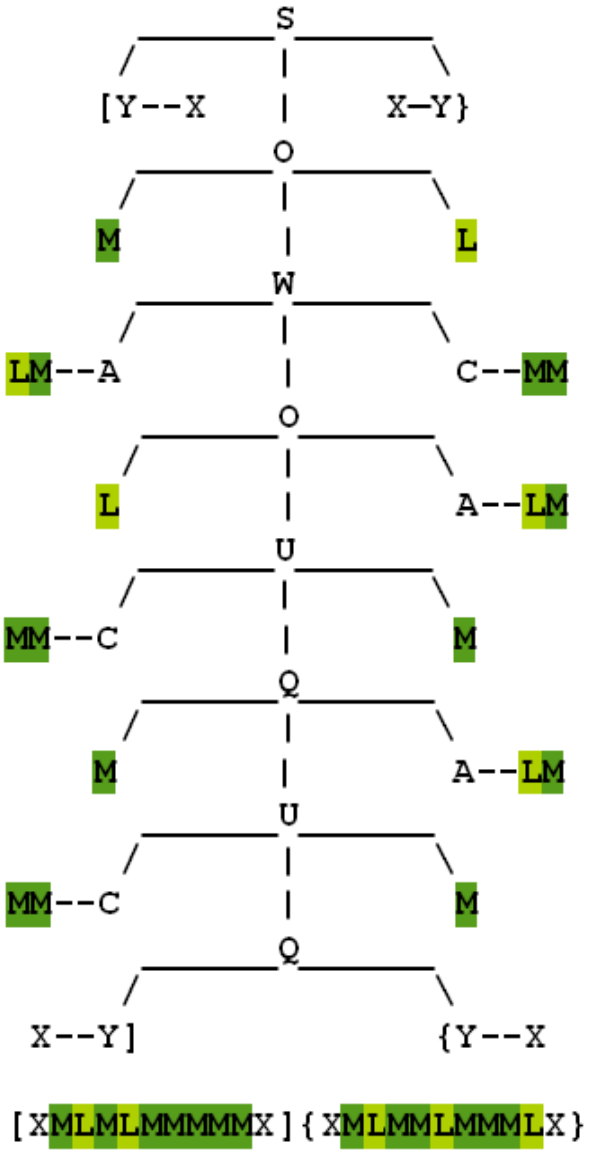



```
if          gene(A→BCD) >2*gene_avg(A)
then        tmpval(A→BCD) =gene(A→BCD)
else        tmpval(A→BCD) =gene(A→BCD)^10/(2*gene_avg(A))^9
phene(A→BCD) = tmpval(A→BCD) /Σ_XYZ tmpval(A→XYZ)
```

Sample: Stochastic Context-Free Grammar representation of helix-helix interaction

Grammar rules (21 relevant rules out of 201 initial):
(van der Waals volume based grammar induced for class 2)

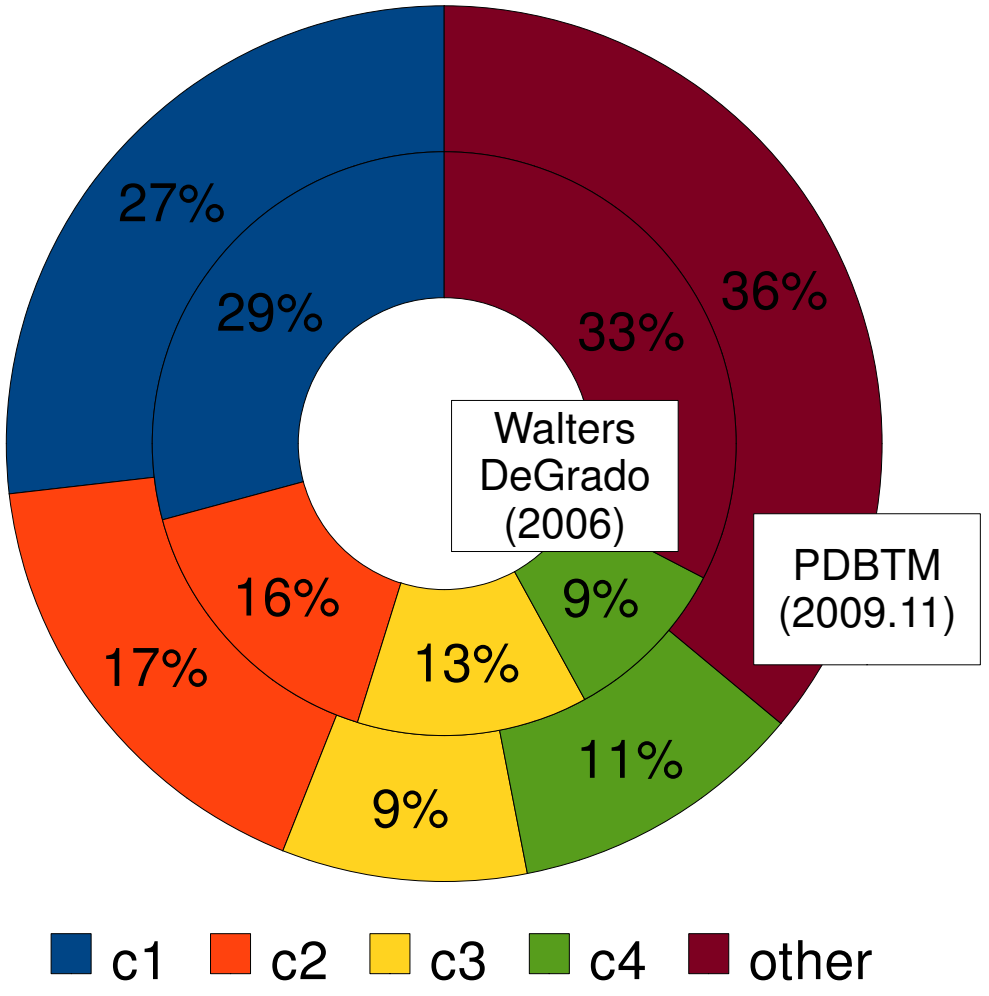
```
S → [YOY} (1.0)

O → LUA (0.40) | MUC (0.21) | MWL (0.39)

Q → LUB (0.13) | MUA (0.44) | Y]{Y (0.43)

U → AQM (0.19) | MQH (0.10) | CQL (0.16)
                | CQM (0.55)

W → AOC (0.80) | COA (0.20)

Y → XY (0.37) | e (0.64)

A → LM (0.42) | ML (0.39) | MM (0.19)

B → LL (0.27) | MM (0.73)

C → MM (1.0)
```



Parse tree built using rules of highest probability:

Results: Datasets of helix-helix contact fragment amino-acid sequences

- **Walters&DeGrado** (2006) standardized to lengths 10-10 non-redundant (homology <40%)
  c1 – 92, c2 – 49,
  c3 – 37, c4 – 27 pairs
- **PDBTM** (2009) lengths 10-10 non-redundant (homology <40%)
  c1 – 100, c2 – 60,
  c3 – 25, c4 – 42 pairs
- WDG and PDBTM mutually **non-homologous (<40%)**



Results: Benchmark 1

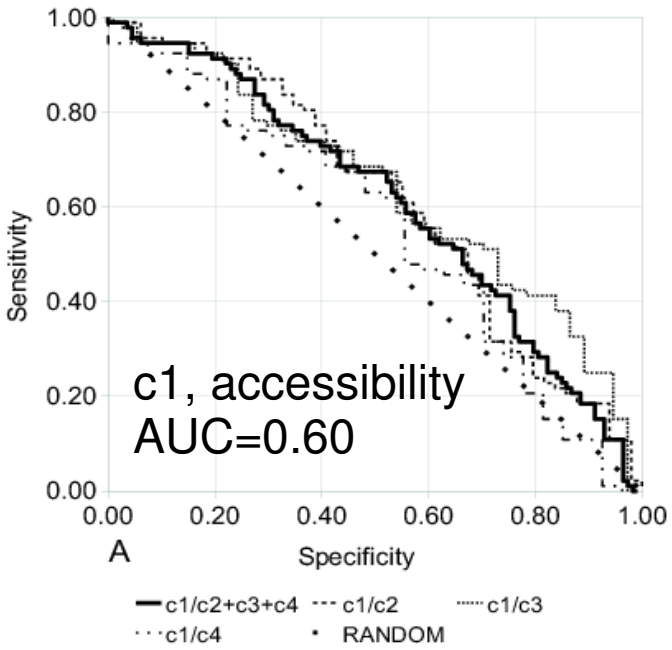
- **Task**
  Assign a helix-helix pair to one of four classes according to the parsing score that reflects how helix-helix contact fragment sequence fits the grammar induced for a given class

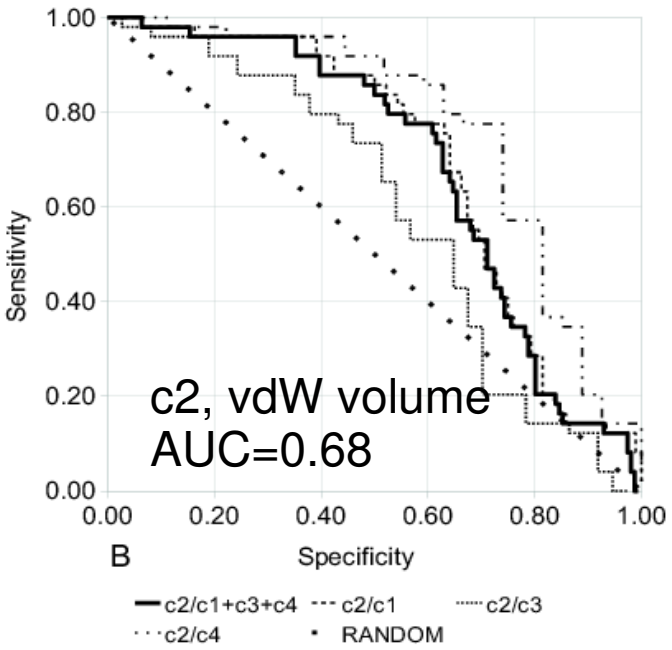
- **Training**
  20 helix-helix contact fragments from PDBTM set closest to each class template in terms of rmsd; *positive training set only here, homologous sequences were accepted*

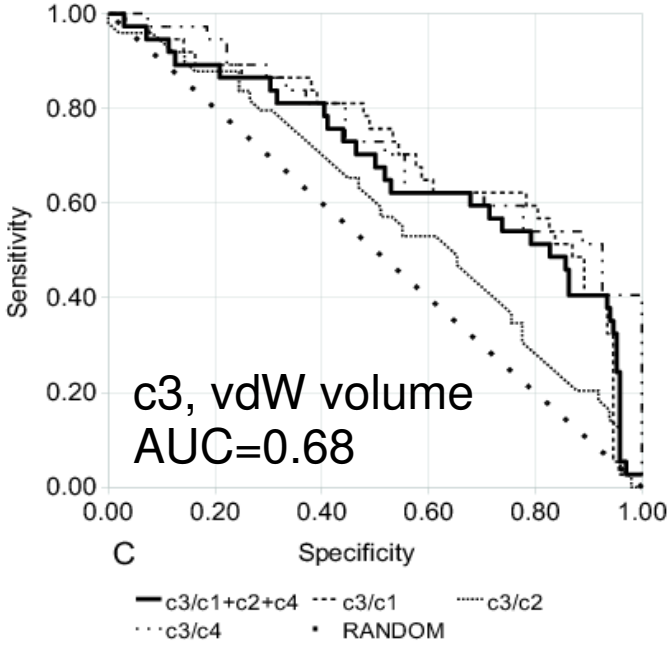
- **Validation**

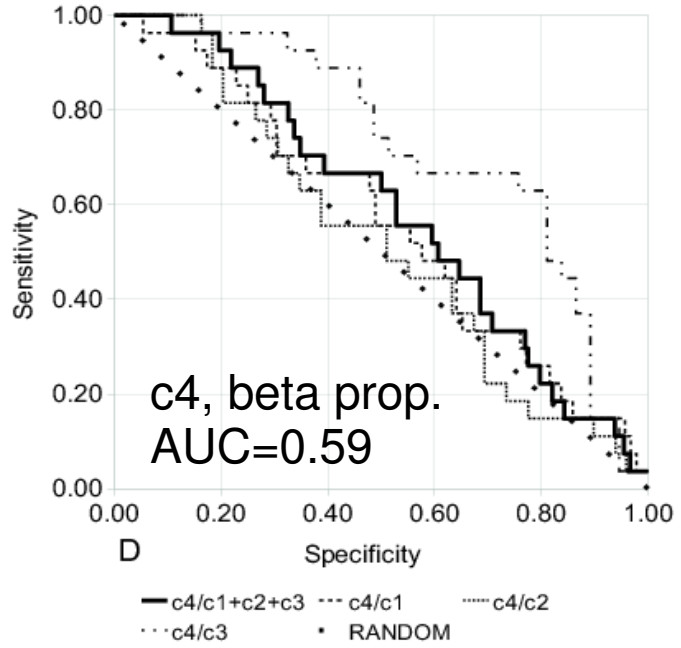
  Helix-helix contact fragments from WDG set; *non-homologous and independent from the training set*

ROC curves


1.00
0.80
0.60
0.40
0.20
0.00
Sensitivity
0.00 0.20 0.40 0.60 0.80 1.00
Specificity
A
c1, accessibility
AUC=0.60
c1/c2+c3+c4
c1/c2
c1/c3
c1/c4
RANDOM


1.00
0.80
0.60
0.40
0.20
0.00
Sensitivity
0.00 0.20 0.40 0.60 0.80 1.00
Specificity
B
c2, vdW volume
AUC=0.68
c2/c1+c3+c4
c2/c1
c2/c3
c2/c4
RANDOM


1.00
0.80
0.60
0.40
0.20
0.00
Sensitivity
0.00 0.20 0.40 0.60 0.80 1.00
Specificity
C
c3, vdW volume
AUC=0.68
c3/c1+c2+c4
c3/c1
c3/c2
c3/c4
RANDOM


1.00
0.80
0.60
0.40
0.20
0.00
Sensitivity
0.00 0.20 0.40 0.60 0.80 1.00
Specificity
D
c4, beta prop.
AUC=0.59
c4/c1+c2+c3
c4/c1
c4/c2
c4/c3
RANDOM

Most useful amino-acid properties

Properties used by best class-by-class classifiers:

	c1		c2		c3		c4	
c1			accessibility	0.61	accessibility	0.63	frequency	0.57
c2	VdW volume	0.70			frequency	0.64	vdW volume	0.77
c3	vdW volume	0.71	vdW volume	0.59			vdW volume	0.73
c4	beta prop.	0.56	accessibility	0.59	beta prop.	0.73		

Class-by-class classification of helix-helix pair contact fragments performance measured by Area under ROC curve using independent test set.

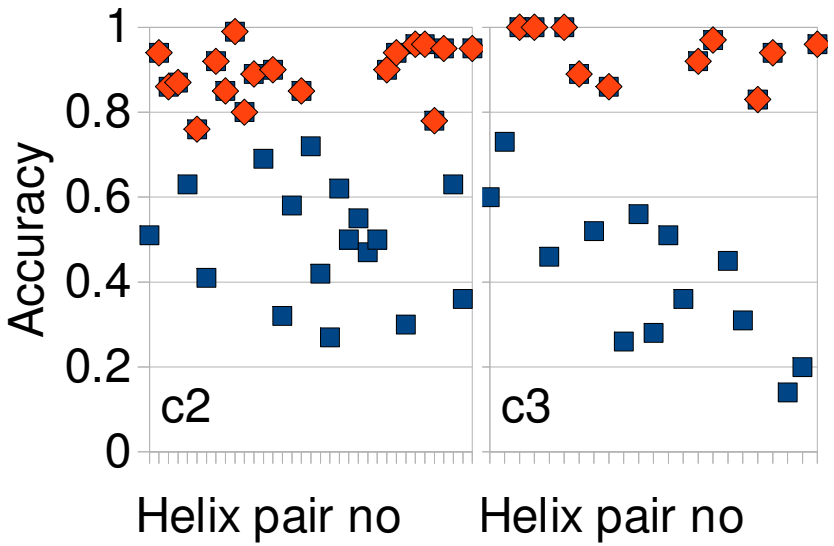
Results: Benchmark 2

- **Leave-one-out cross-validation**
  on helix-helix contact fragments from WDG; *positive training set only; non-homologous sequences only*

- **Task (the same)**
  Assign a helix-helix pair to one of four classes according to the parsing score that reflects how helix-helix contact fragment sequence fits the grammar induced for a given class

Benchmark 2 results

class	property	full set %pairs	full set avg. Accu.	High accuracy subset %pairs	High accuracy subset avg.Accu.
c1 vs c234	accessibility	1.00	0.54	0.30	0.89
c2 vs c134	vdW volume	1.00	0.70	0.51	0.89
c3 vs c124	vdW volume	1.00	0.64	0.43	0.94
c4 vs c123	accessibility	1.00	0.58	0.47	0.90

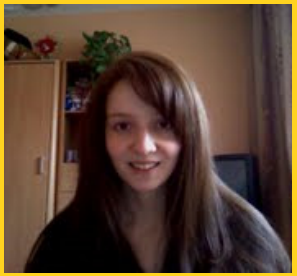


Subsets of helix-helix interaction classes more prone to structural description (red)

Conclusions

- We developed the first method for prediction of conformation of helix-helix interaction classes which explicitly represents dependecies between two helices
- The method was tested using independent validation set
  - achieved AUC ROC 0.59 – 0.68, Accuracy 0.52-0.75
- Accessibility and van der Waals volume were the most informative amino-acid physico-chemical properties
- The performance of the method can be improved in terms of Accuracy from 0.55-0.70 up to around 0.89-0.94 for certain subclasses containing 30-51% of helix-helix pairs

Thank you!



Bioinformatics & Nanopore Biophysics Group @WUT **(KotulskaLab):**

Małgorzata Kotulska
Bogumił Konopka
Monika Rybicka
Paweł Gąsior
Roksana Kowalska

Bratka Deryło
Joanna Basałyga
...

Bioinformatics & Genomic Signal Processing @KU

Jean-Christophe Nebel

This work is partially funded by:

MINISTRY OF SCIENCE AND HIGHER EDUCATION
BRITISH COUNCIL
DOLNY ŚLĄSK
HUMAN CAPITAL
NATIONAL COHESION STRATEGY