Striped Sheets and Protein Contact Prediction

Bob MacCallum

Stockholm Bioinformatics Center
Stockholm University
Sweden

In an ideal world...

We would simulate protein folding with

- ultra-fast computers
- accurate force-fields

But in reality we have to try shortcuts

Shortcuts for ab initio prediction

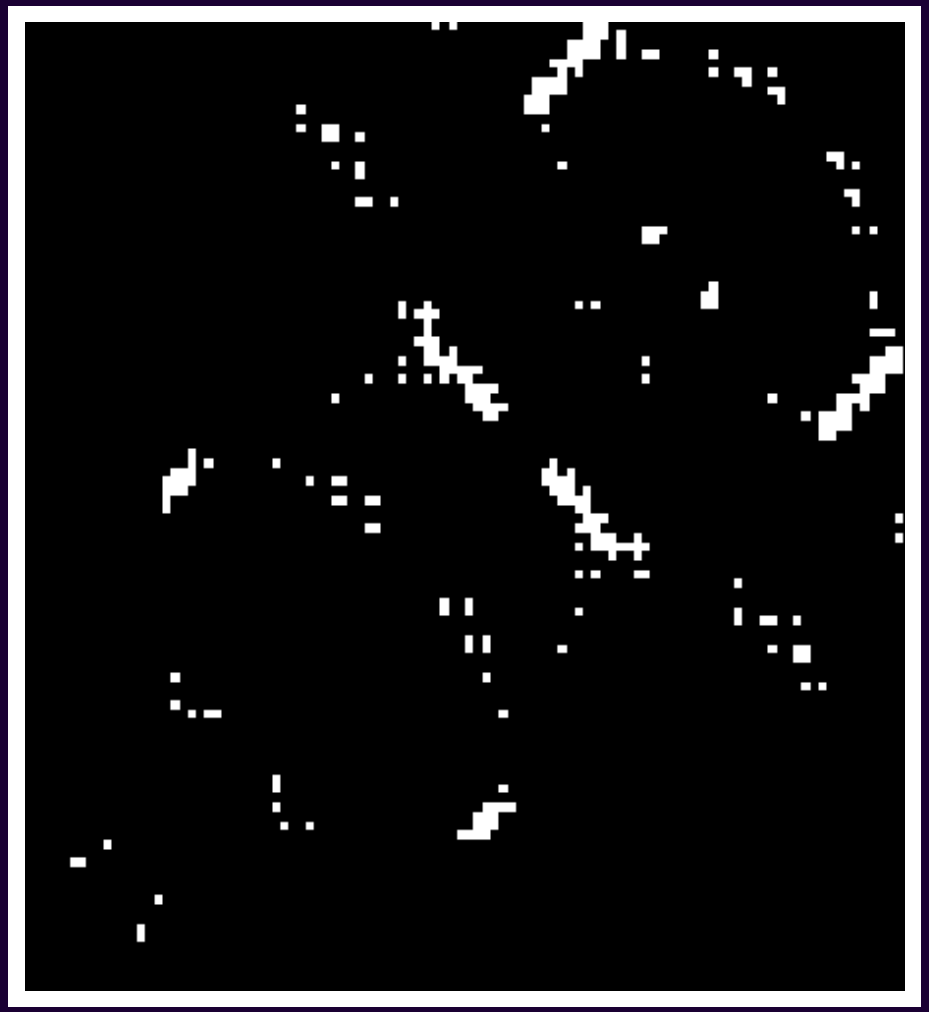
Simplified models

Fragment assembly

Secondary structure prediction

Contact map prediction

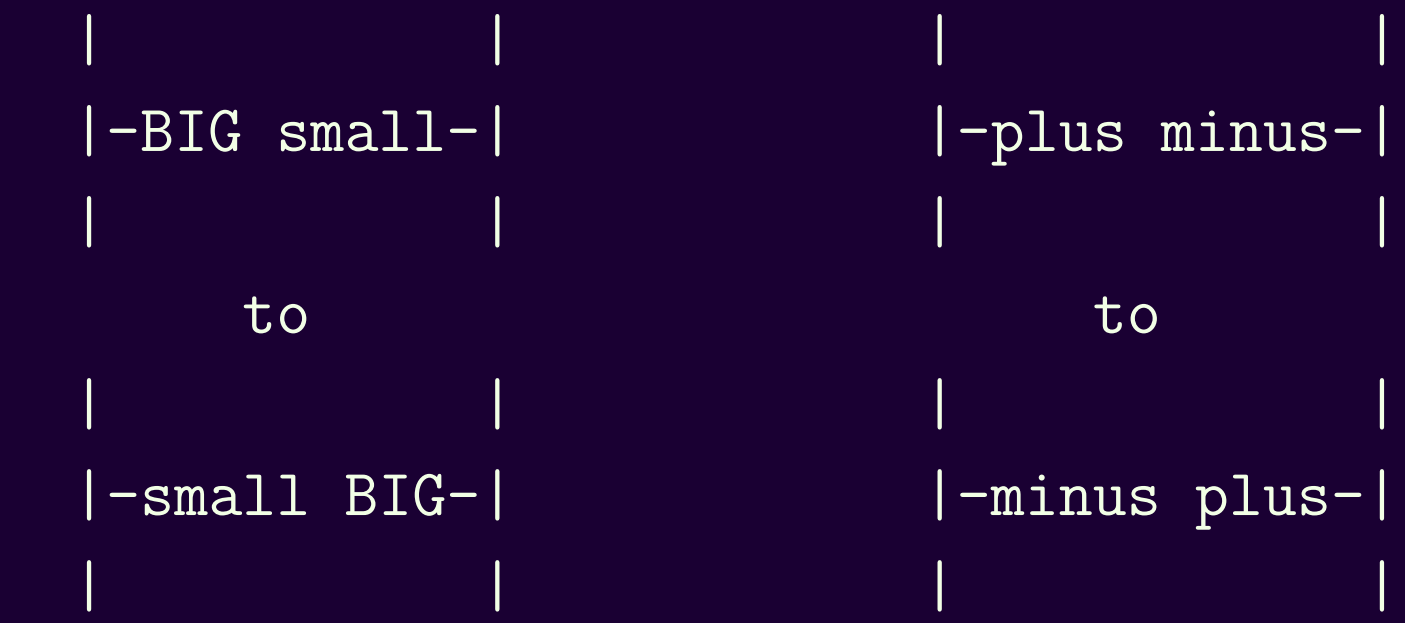
Contact maps



Current approaches to contact map prediction

Look in multiple sequence alignment for

compensatory/correlated mutations



Current approaches to contact map prediction

Useful additional information from

- residue conservation
- profile vectors
- predicted secondary structure
- pair statistics

ANNs or SVMs most commonly used to assimilate the data

Current performance of contact map prediction

Generally poor

Not ready for 3D

Very weak signal in correlated mutations

Similar performance with fold recognition and *ab initio*

Room for improvement...

Outline for rest of talk...

Visualisation of PSI-BLAST profiles with self-organising maps

A new approach to contact prediction (?)

Motivation for sequence profile visualisation

Predicted secondary structure useful in other structure prediction approaches

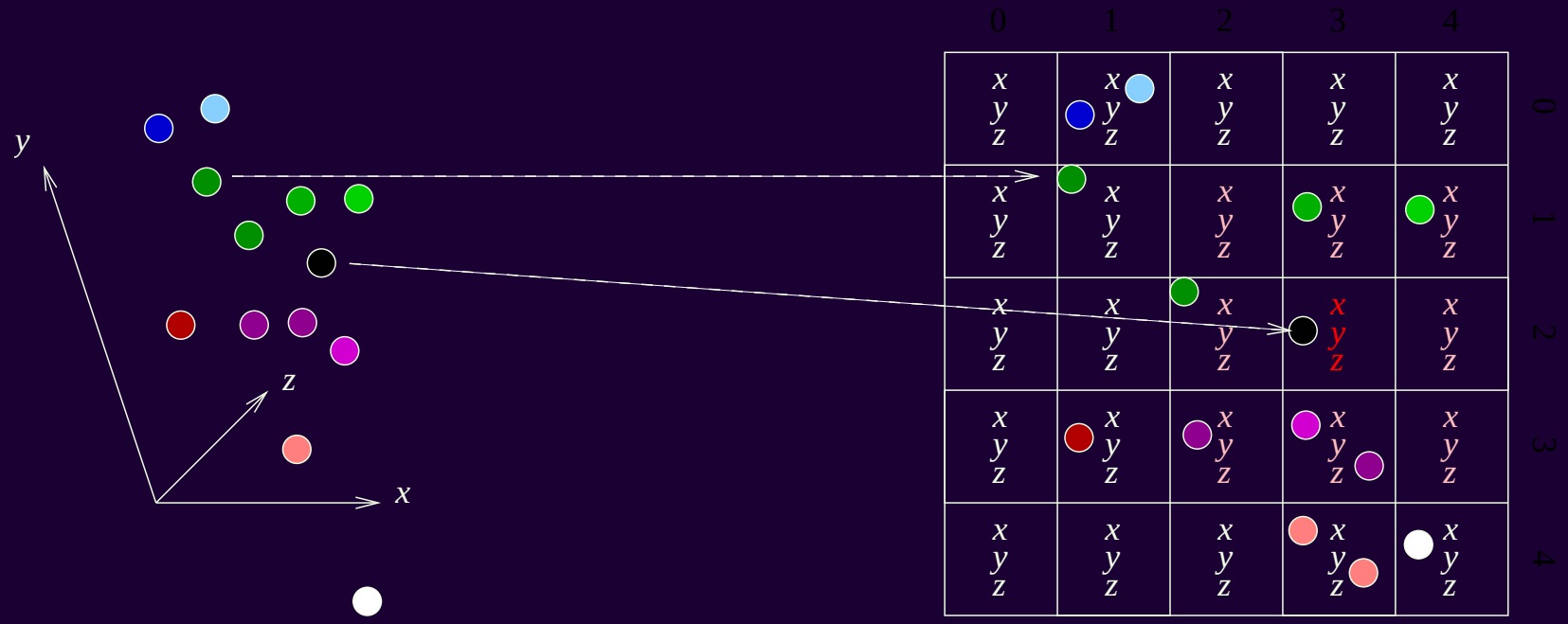
Widespread arbitrary use of helix, strand and coil

Another alphabet or scheme?

Windows of profiles contain a lot of information about SS

Can we cluster profile windows in some meaningful way?

Self-organising maps (SOMs)



Iterated competitive clustering technique

High dimensional data $\rightarrow$ low dimensional grid

Preserves relationships / "flattens"

Mapping sequence profile windows

Input:

Raw profiles from PSI-BLAST (3 iterations; NR)

Window width: $w \in \{1, 5, 9, 15\}$

20 amino acids (and one mystery column)

$\rightarrow w \times 21$ matrices

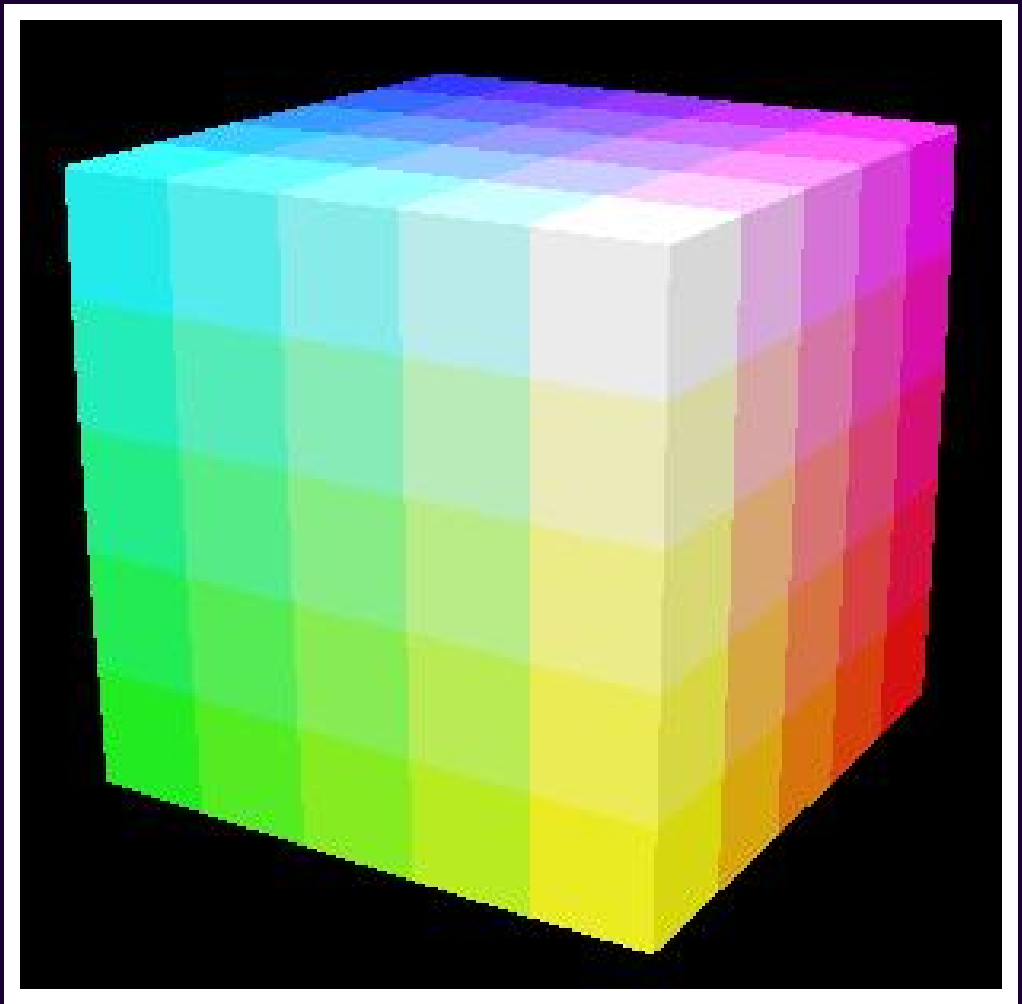
Output:

Locations on a 3D SOM grid (most are 2D)

Dimensions: $6 \times 6 \times 6$

216 discrete locations $\rightarrow$ colours for visualisation

3D SOM locations $\rightarrow$ colours



SOM procedure

Training

1573 SCOP domains

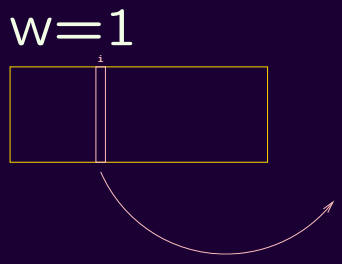
Train SOMs on sample of 100000 profile windows for 6 epochs

Initial training rate 0.1 ; initial radius 3

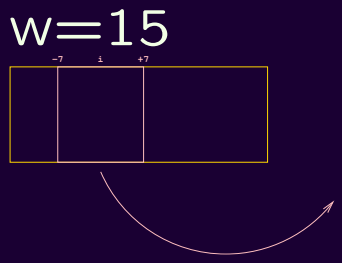
Application

All profile windows from *any* protein can be mapped to a location in the SOM, and are **assigned a colour**

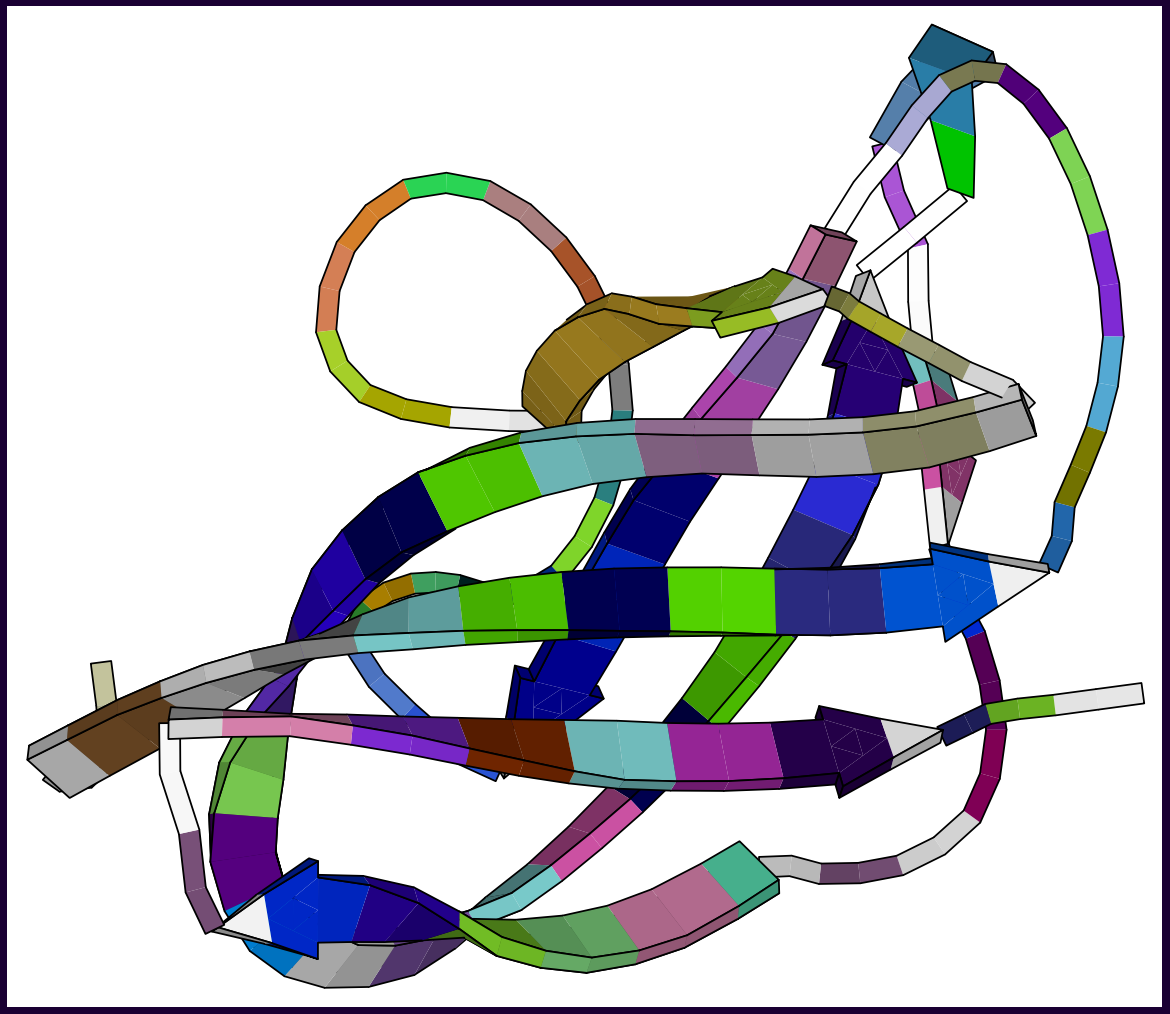
1ekjA



1ekjA

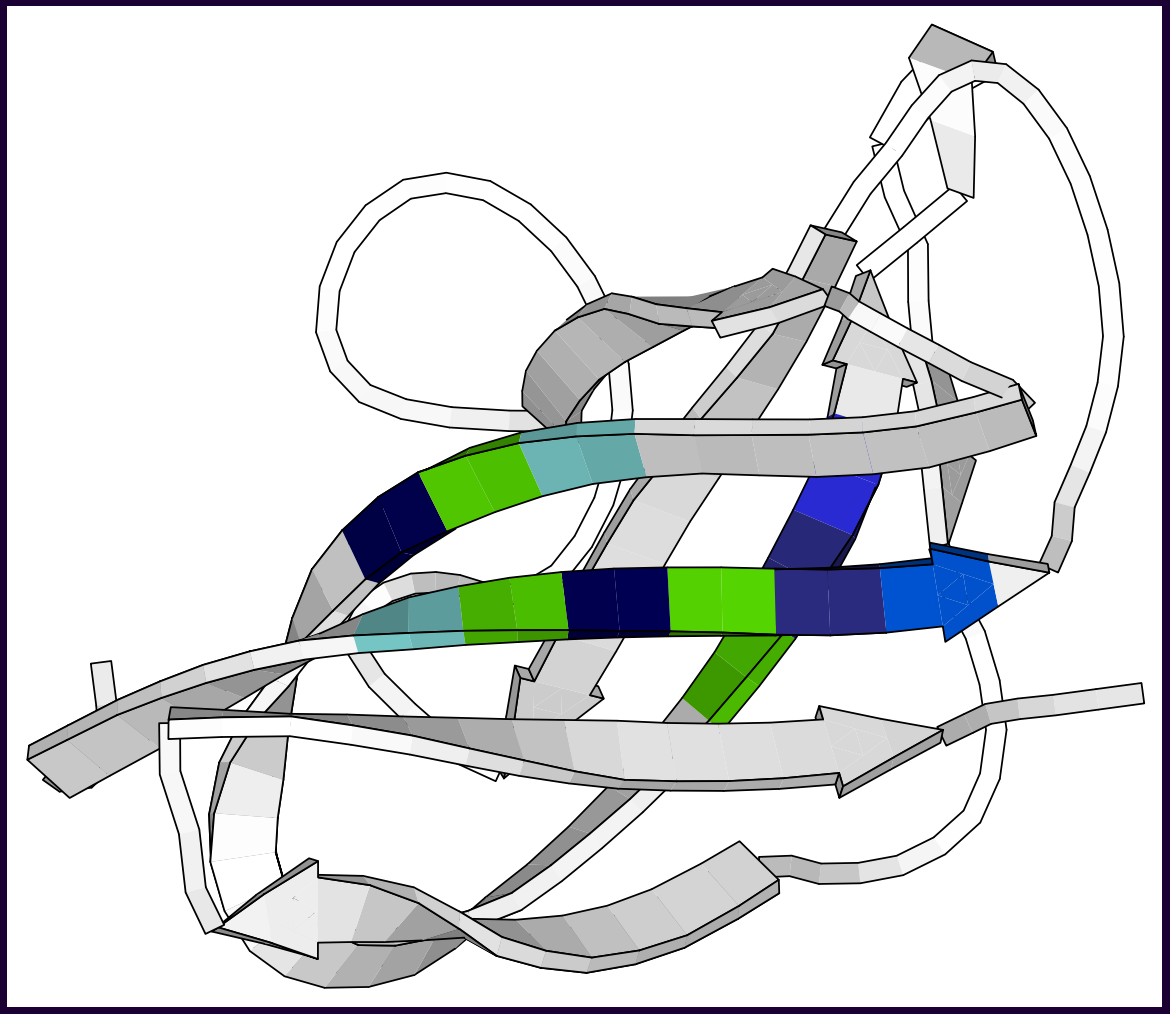


Anti-parallel and facing strands



1qhoA2

Anti-parallel and facing strands



1qhoA2

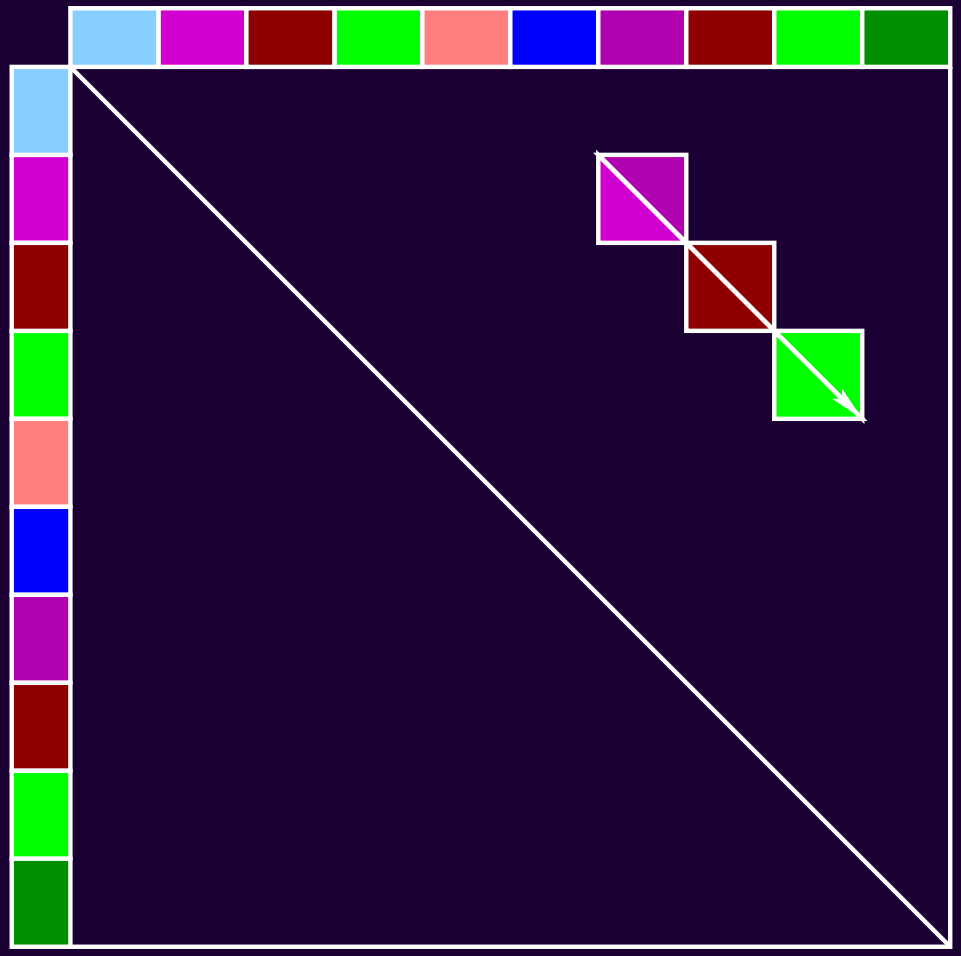
What does it mean?

My theory:

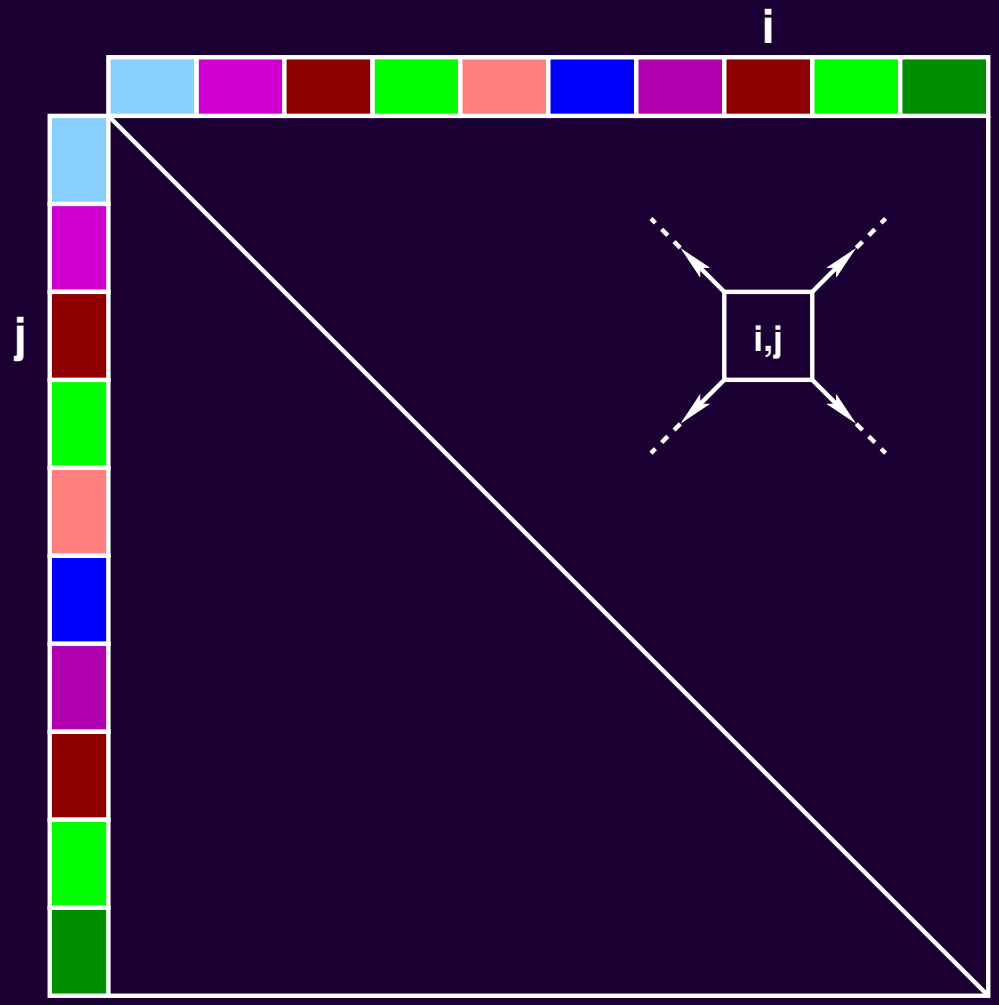
1. Sequence profile windows reflect local structural environment
2. SOM colours summarise these environments
3. Neighbouring strands pass through similar environments
4. Similar colour patterns expected...?

Can we use the SOM mappings for predicting strand pairs and other contacts?

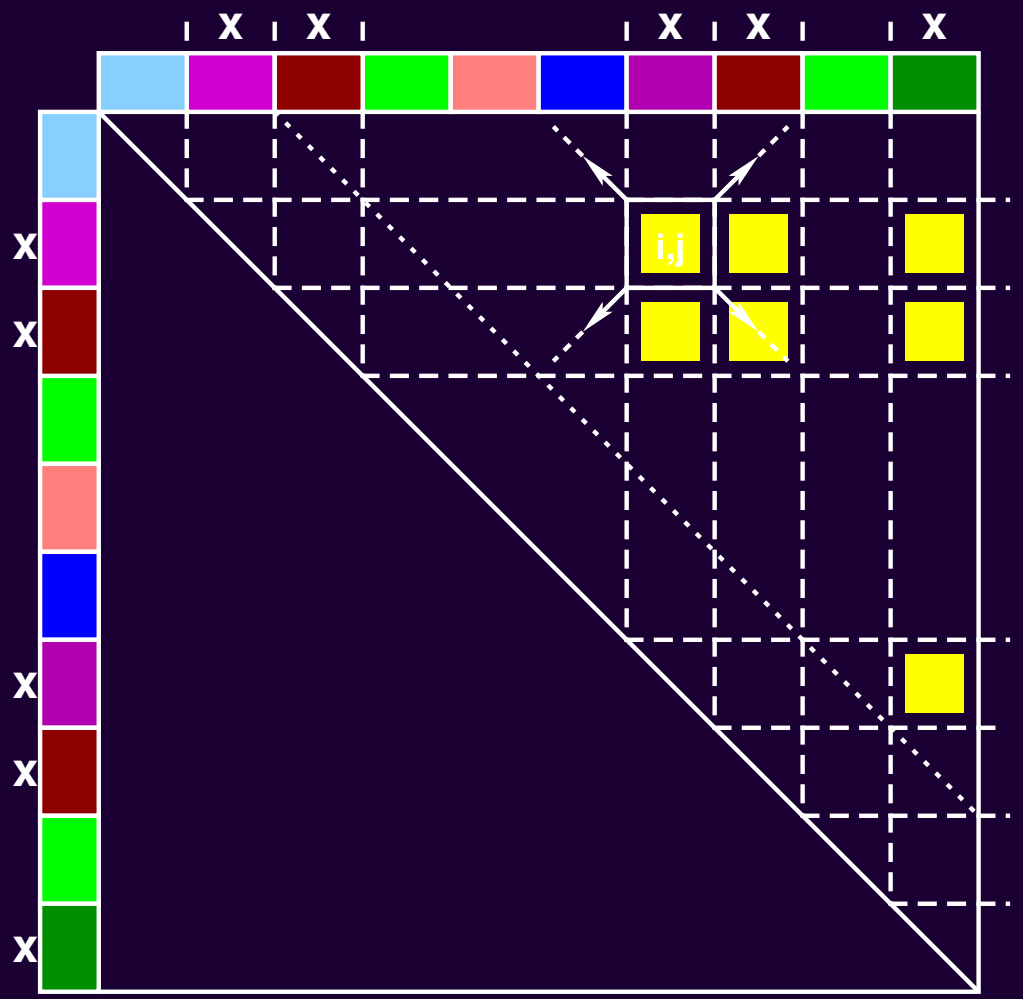
A simple approach



Check all pairs i,j



Filter prior to pairwise calculations



Designing the predictor

- Criteria for pre-filter?
- What is the extent of colour pattern similarities?
- Parallel or anti-parallel?
- What is the influence of distance from diagonal?
- Which SOM mapping(s) to use? ($w \in \{1, 5, 9, 15\}$)
- Which combination of all these...?

Fictional example

Residue filter

```
IF (MP15[i,red] > 3 AND MP15[i,blue] < 4 AND
    MP1[i,green]/(4 + MP1[i,blue]) > 0.3)
THEN ACCEPT residue i
```

Pair distance

```
dist(i,j) = MP9[j,green] * (3 - MP5[i,blue]) +
            parallel_col_dist(MP15,i,j,3) +
            antiparallel_col_dist(MP1,i,j,5) *
            log(abs(i-j)) - 7
```

Genetic programming (GP)

We can evolve the rules using GP

Like a GA, but produces code of any size & shape

Supervised

CPU hungry

But otherwise straightforward

PerlGP – http://perlgp.org

GP training

Fitness function: N_c/N_p for $L/10$ contacts

(C-$\beta \leftrightarrow$ C-$\beta < 8.0$Å ; residue separation $|i - j| \geq 8$)

SCOP domains:

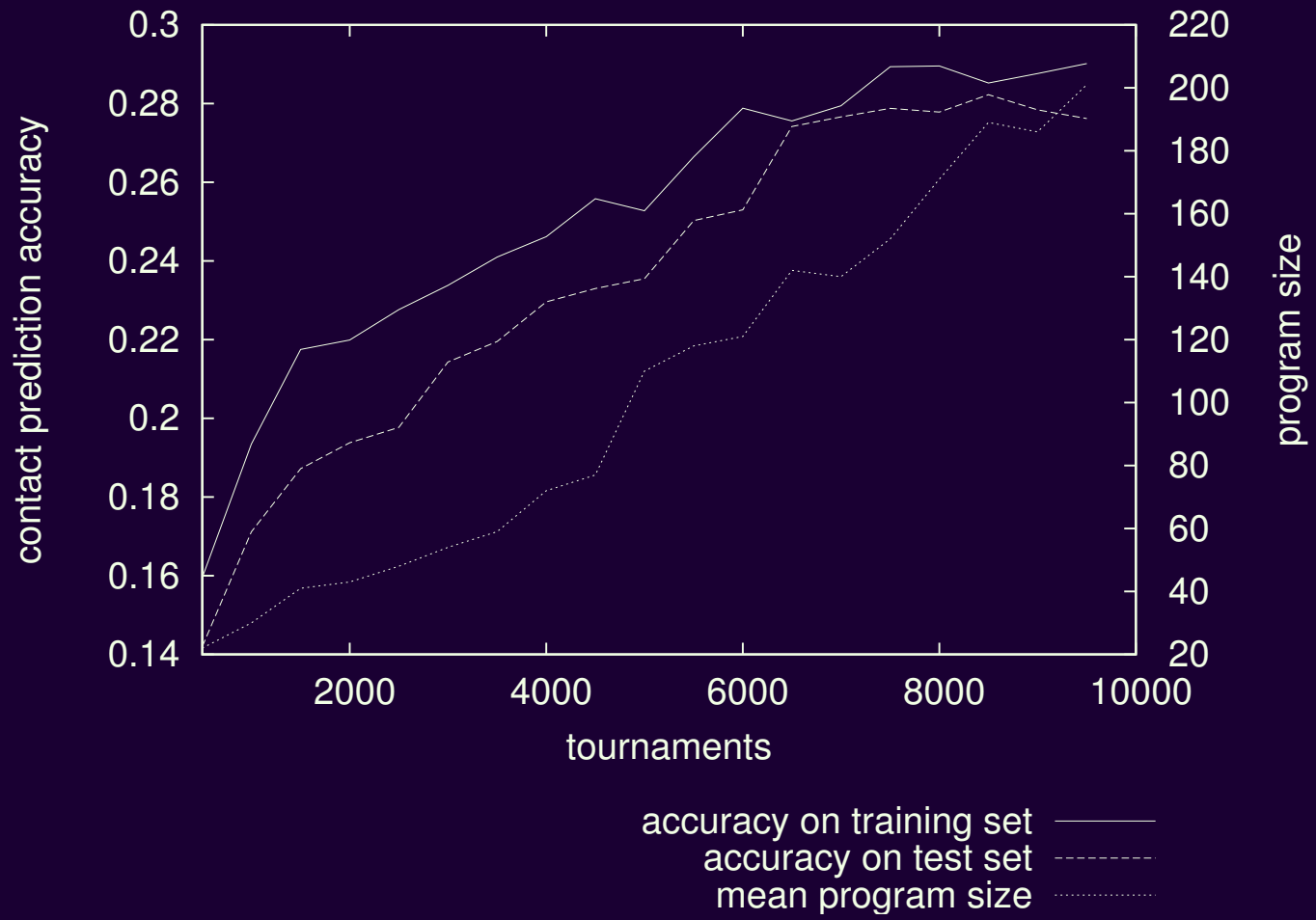
- 451 training
- 227 testing
- 170 final validation

($\leq 10\%$ id, one representative per superfamily)

Re-sample 100 domains for training

20 populations of 2000 individuals

GP training


0.3
0.28
0.26
0.24
0.22
0.2
0.18
0.16
0.14
contact prediction accuracy
220
200
180
160
140
120
100
80
60
40
20
program size
2000
4000
6000
8000
10000
tournaments
accuracy on training set
accuracy on test set
mean program size

An evolved contact predictor

```
sub filter {
  my ($mpssms, $cma, $i, $n) = @_;
  my $j = ($i + -1) % $n;

  return ($mpssms->[0]->at(($i + 0) % $n, 2) - $mpssms->[0]->at(($i + 0) % $n, 1));
}

sub distance {
  my ($mpssms, $cma, $i, $j, $n) = @_;

  return (((-2 > 0.7596 ? -5 : ((eucwin($mpssms->[3], $i, $j, 5, 1, $n) -
    (($mpssms->[2]->at(($i + 0) % $n, 1) > 0.4003 ?  eucwin($mpssms->[3], $i, $j, 2, -1, $n) :
    ($mpssms->[0]->at(($j + 0) % $n, 1) - 1)) - eucwin($mpssms->[0], $i, $j, 0, 1, $n))) -
    ($mpssms->[0]->at(($i + 0) % $n, 1) - $mpssms->[3]->at(($j + 0) % $n, 1)))) -
    ($mpssms->[1]->at(($i + 3) % $n, 1) - ((((($mpssms->[1]->at(($i + 3) % $n, 1) -
    (((5 - $j) + (eucwin($mpssms->[3], $i, $j, 5, 1, $n) + $i)) -
    ($mpssms->[1]->at(($i + 0) % $n, 1) - eucwin($mpssms->[3], $i, $j, 0, -1, $n)))/
    $mpssms->[0]->at(($i + 0) % $n, 1)) - $j) + ($cma->at(($i + 0) % $n, ($j + 0) % $n) + $i)) -
    ($mpssms->[1]->at(($i + 0) % $n, 1) - eucwin($mpssms->[0], $i, $j, 0, -1, $n)))/
    ($mpssms->[1]->at(($i + 0) % $n, 1) + abs($mpssms->[0]->at(($i + 0) % $n, 1))))) -
    ($mpssms->[0]->at(($j + 0) % $n, 1) - eucwin($mpssms->[0], $i, $j, 1, -1, $n)));
}
```

Hand-picked prediction

1bvyF

"L/10" Accuracy by SCOP class

		min. separation		
Validation subset	n	8	16	24
Full set	156	**27.1**	24.6	**20.6**
SCOP class				
all-α	30	19.9	17.9	15.3
all-β	35	30.5	24.7	20.2
$\alpha + \beta$	40	25.9	24.2	21.2
α/β	18	31.5	36.7	34.9
Membrane assoc.	8	0.1	0.1	0.1
Multi-domain	4	23.5	18.6	18.4
Small proteins	21	39.2	32.9	22.0
Containing α and β	58	27.6	28.1	25.4
Containing β	97	28.5	26.5	23.3
No Small or Membrane	127	**26.5**	24.5	**21.4**

Retrospective CASP5 predictions

	number of pairs predicted		
Method	$L/10$	$L/5$	$L/2$
Bystroff	19	16	12
CMAPpro	16	14	13
CORNET	21	18	13
This work	**27**	**21**	**14**

All data used pre-dates CASP5

15 targets not similar to PDB

Low-coverage performance looks promising

Summary

New sequence profile visualisation approach

Striped sheets – sequence constrained by structural environment

Competitive contact prediction

No correlated/compensatory mutations – only mapped profiles and sequence separation

Low coverage is a feature

Food for thought

Novelty

Lund *et al* (1997) used windows of 18 – more recent methods use very small windows

Convergence with Shao & Bystroff's recent approach using I-sites

Improvements

More sophisticated detection of compensatory/correlated mutations (Benner lab)

Applications

Screening models for CASP6, SSP, ...

Acknowledgements

Stockholm Bioinformatics Center

Swedish Foundation for Strategic Research

Swedish Parallel Computing Center (PDC)

Organisers

Audience