

Protein Domain Prediction

David Jones

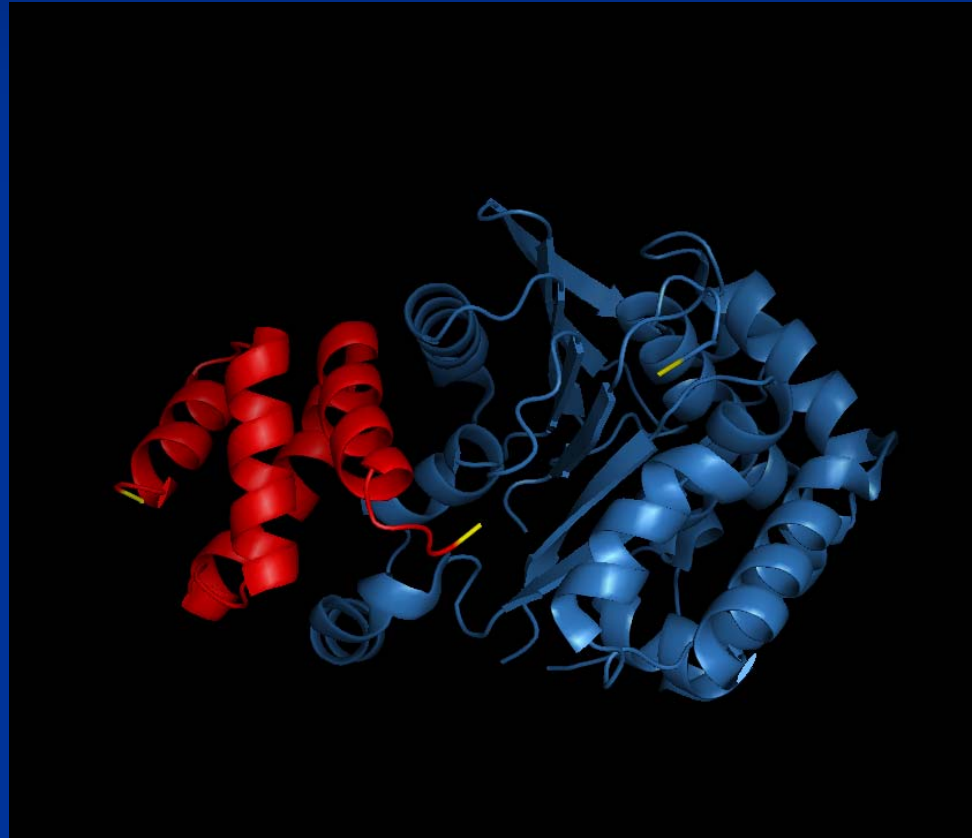
University College London

<http://bioinf.cs.ucl.ac.uk>

Domains

- Structural View
 - Compact independent folding units
- Evolutionary View
 - Evolutionary unit i.e. functional domain or module
- Combination
 - Independent folding unit with a well defined function and/or evolutionary history

Typical 2-domain protein T0233 (CASP6)



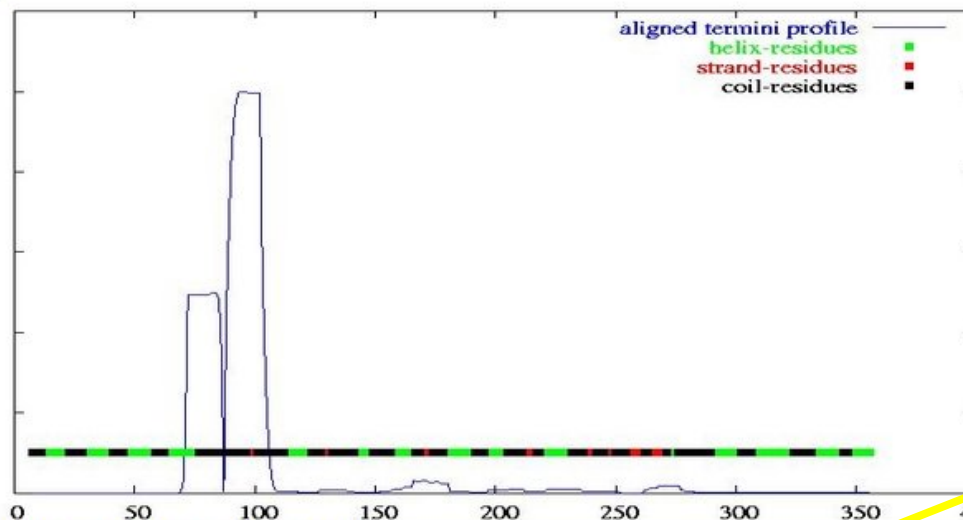
2 Domains: 14-79, 93-362

Marsden home >
 Bryson home >
 McGuffin home >
 DomPred home >

DomPred Prediction Results For T0233

Sequence length 362 residues

PSI-BLAST HELP
 Alignment
 Profile



Number PSIBLAST hits = 1000

[Show PSI-BLAST Output](#)

[Show Parsed PSI-BLAST Output](#)

Putative domain boundaries located in PSI-BLAST alignment profile:

Number of predicted domains by DPS: 2

Domain boundaries predicted by DPS: 94

DPS predicts domain boundary at 94.

DomSSEA HELP
 Results

Score	Match	SSEA	No. Doms	Boundaries	SCOP code
0.847	1kgzB	Show SSEA	2	97	a.46.2.1 , c.27.1.1
0.839	1v8gB	Show SSEA	2	97	a.46.2.1 , c.27.1.1
0.835	1khdB	Show SSEA	2	97	a.46.2.1 , c.27.1.1
0.835	1gxbD	Show SSEA	2	97	a.46.2.1 , c.27.1.1
0.834	1khdD	Show SSEA	2	97	a.46.2.1 , c.27.1.1

DomSSEA predicts a boundary at 97

Current DomPred Result for T0233

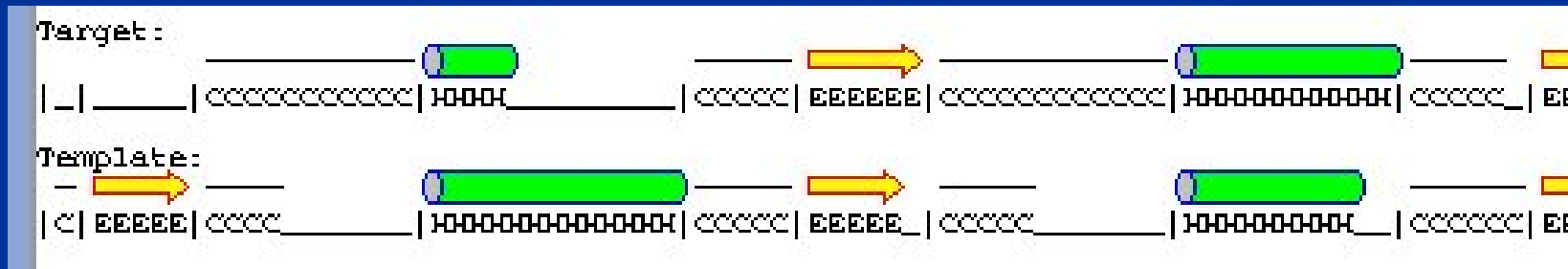
CASP Manual:
 14-79, 93-362

Domains from Secondary Structure Element Alignments (DomSSEA)

Target



PSIPRED



SCOP Database



DSSP Database

SSEA Scoring

\mathcal{R}	C_1	H_1	E_1
C_2	$\min(\text{len}(C_1), \text{len}(C_2))$	$0.5\min(\text{len}(H_1), \text{len}(C_2))$	$0.5\min(\text{len}(E_1), \text{len}(C_2))$
H_2	$0.5\min(\text{len}(C_1), \text{len}(H_2))$	$\min(\text{len}(H_1), \text{len}(H_2))$	0
E_2	$0.5\min(\text{len}(C_1), \text{len}(E_2))$	0	$\min(\text{len}(E_1), \text{len}(E_2))$

HCEC--HCECEC-
HCECHCECECH

DOMSSEA - Protein Domain Prediction

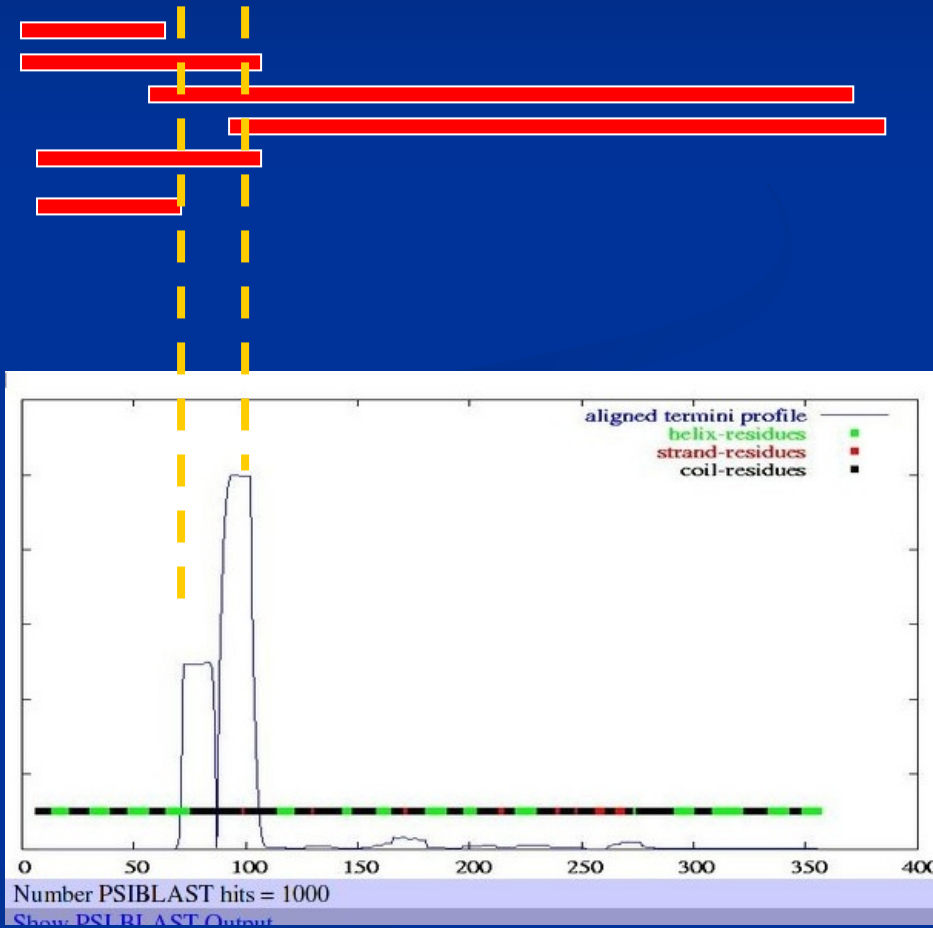
- Method
- Align (PSIPRED) predicted secondary structure pattern of target to observed secondary structure patterns in proteins of known 3D structure
- Use of Secondary Structure Element Alignment methods (SSEA) acts as a crude fold recognition algorithm
- Domain number and boundary predictions taken from top-scoring alignment
- Template library of known structures taken from CATH domain database
 - single and multi-domain chains <30% sequence identity

DOMSSEA - Results

- DomSSEA used for all-against-all alignment of chains in template library. Top scoring hits filtered by PSIBLAST to remove any possible matches due to sequence similarity
- Prediction of domain number (single, two domain, three+ domains)
 - 82% correctly assigned as single domain
 - 46% correctly assigned as two-domain
 - 37% correctly assigned as three-or-more domain
- Prediction of domain boundaries (+/- 20 residues)
- Sensitivity of 31% (percentage of correctly assigned boundaries)
- Selectivity of 39% (percentage of predictions made that are correct)
- Overall, 25% of multi-domain chains have correct assignment of domain number and corresponding boundaries

Domains Predicted from Sequence (DPS)

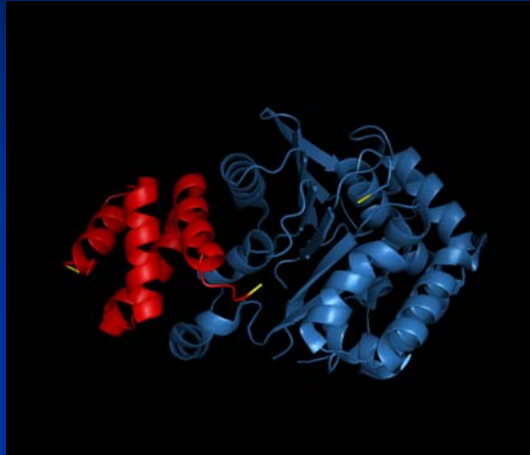
- PSI-BLAST hits
- Histogram N- & C-terminals
- Smooth and z-score



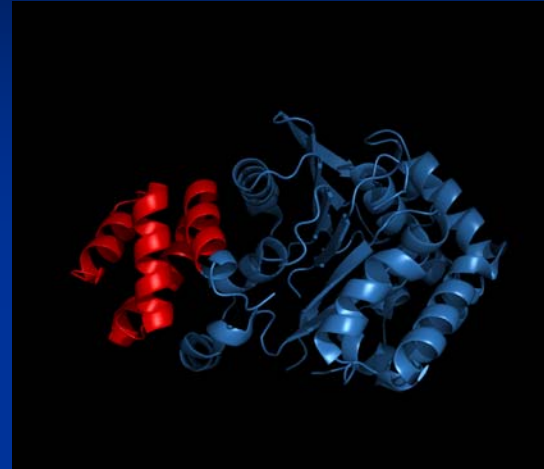
DomPred Server

- A combined approach to domain assignment
 - Sequence based method, post-processes PSI-BLAST sequence alignments, assigns domain boundaries from distribution of aligned sequence termini to target sequence
 - DomSSEA output for harder targets where sequence comparison finds no matches
- Combined methods used for prediction of domain boundaries (same targets) (+/- 20 residues)
- Sensitivity of 55% (percentage of correctly assigned boundaries)
- Selectivity of 45% (percentage of predictions made that are correct)

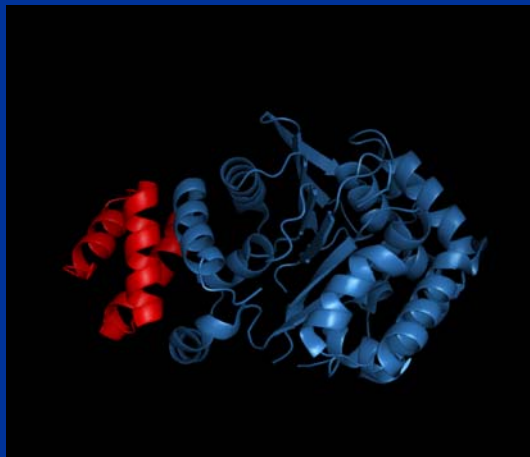
Original T0233 Predictions (CASP6)



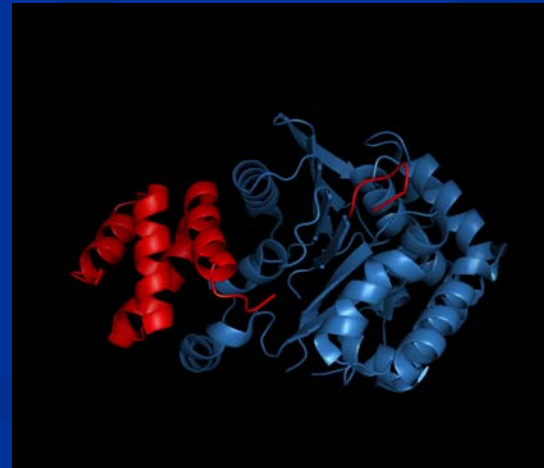
2 Domains: 14-79, 93-362



DPS: Domain cut at 73



Ginzu: Domain cut at 60



DomSSEA: Domain cut at 98

CASP 6 Scoring Scheme: NDO



	D1 (1-20)	L (21-30)	D2 (31-100)	Sum
d1 (1-20)	20	0	0	20
d2 (21-100)	0	10	70	60
Sum	20		70	170

Final NDO Score = $170/180 = 94\%$

CASP6 Results

TABLE IV. Prediction Scores of Each Group for All 63 Targets

Prediction group	Np ^a	NDO score		Z-score		Top-1 ^b		Top-score ^c	
		Mean	Rank	Mean	Rank	Count	Rank	Count	Rank
P0018	8	86.11	10	0.00	12	3	17	6	17
P0061	1	58.39	22	-1.30	22	0	20	0	21
P0063	56	60.94	20	-1.15	20	8	14	9	15
P0089	63	88.07	6	0.37	6	31	5	44	6
P0096	62	89.22	3	0.45	3	30	8	46	2
P0237	7	89.82	2	0.70	1	0	20	3	19
P0461	58	74.29	15	-0.35	15	18	11	22	12
P0536	63	90.73	1	0.51	2	36	1	50	1
P0590	61	87.17	8	0.33	7	31	5	40	8
P0667	56	65.57	17	-0.74	19	11	13	17	14
P0682	33	58.48	21	-1.26	21	3	17	3	19
Human	468	77.16		-0.22		171		240	
P0019	51	84.47	12	0.19	11	28	9	33	10
P0283	21	64.78	18	-0.60	17	1	19	4	18
P0289	2	63.66	19	-0.60	18	0	20	0	21
P0290	56	77.79	14	-0.21	14	8	14	22	12
P0309	61	82.40	13	0.00	12	16	12	31	11
P0353	63	88.74	5	0.39	4	33	3	45	3
P0381	60	86.16	9	0.25	9	28	9	40	8
P0421	63	88.76	4	0.39	4	34	2	45	3
P0435	63	85.08	11	0.20	10	31	5	43	7
P0436	63	87.69	7	0.30	8	32	4	45	3
P0638	40	72.03	16	-0.49	16	4	16	8	16
Server	543	80.14		-0.02		214		316	

^aNumber of predictions each group submitted.

^bNumber of times the prediction was best among all predictions for a given target.

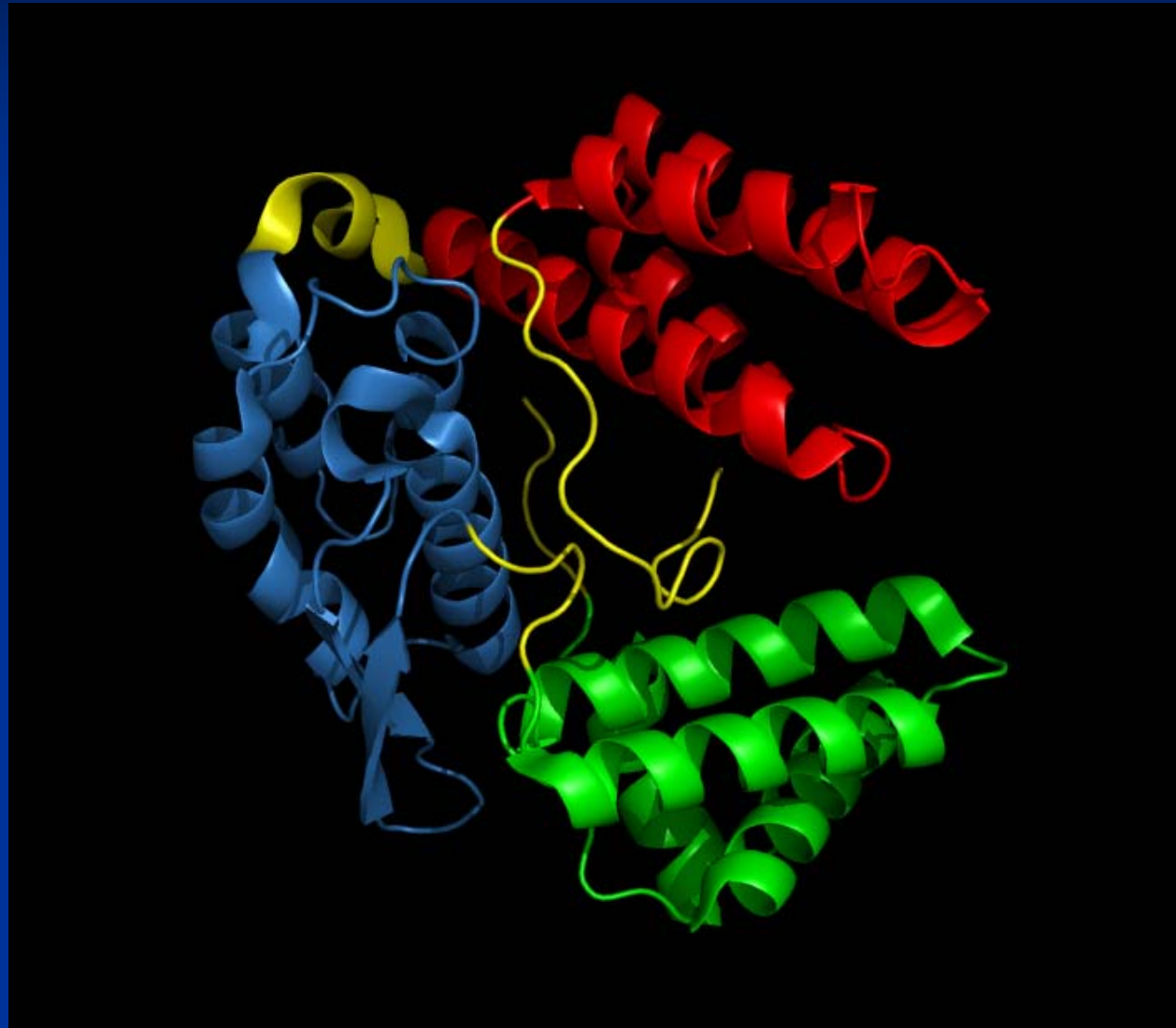
^cNumber of times the prediction score was within 5% of the best score for a given target.

Three domain T0248 (CASPP6)

Domain 1 (FR/A):
21-99

Domain 2 (NF):
107-193

Domain 3 (FR/A):
199-285



Current DomPred Result for T0248

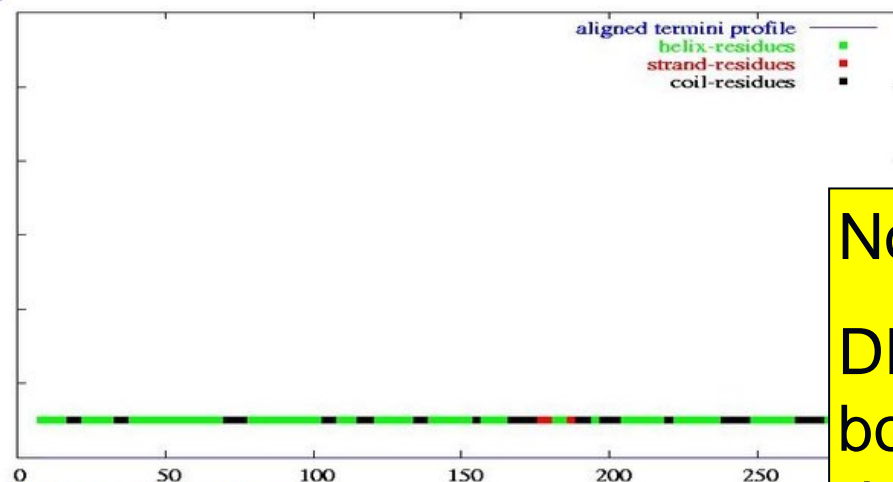
Bioinformatics Unit 

[Marsden home >](#)
[Bryson home >](#)
[McGuffin home >](#)
[DomPred home >](#)

DomPred Prediction Results For T0248

Sequence length 294 residues

[PSI-BLAST](#) HELP
Alignment
Profile



No Signal !
DPS & DomSSEA
both predict single
domain

Number PSIBLAST hits = 7

[Show PSIBLAST Output](#)

[Show Parsed PSIBLAST Output](#)

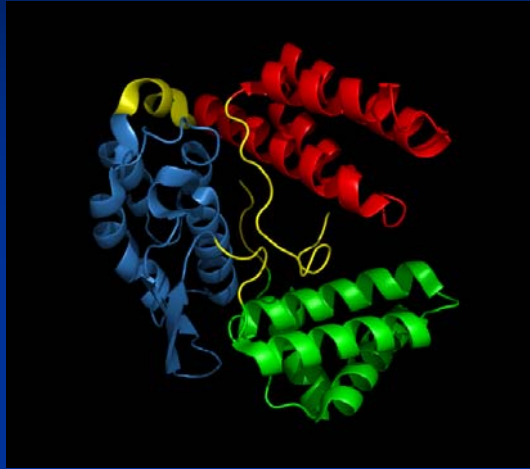
Number of predicted domains by DPS: 1

Domain boundaries predicted by DPS:

[DomSSEA](#) HELP
Results

Score	Match	SSEA	No. Doms	Boundaries	SCOP code
0.810	1qjbB	Show SSEA	1	-	a.118.7.1
0.801	1qjbA	Show SSEA	1	-	a.118.7.1
0.798	1qjaB	Show SSEA	1	-	a.118.7.1

Original T0248 Predictions (CASP6)



21-99 & 107-193 & 199-285



DPS, DomSSEA & Ginzu
all predict single domain



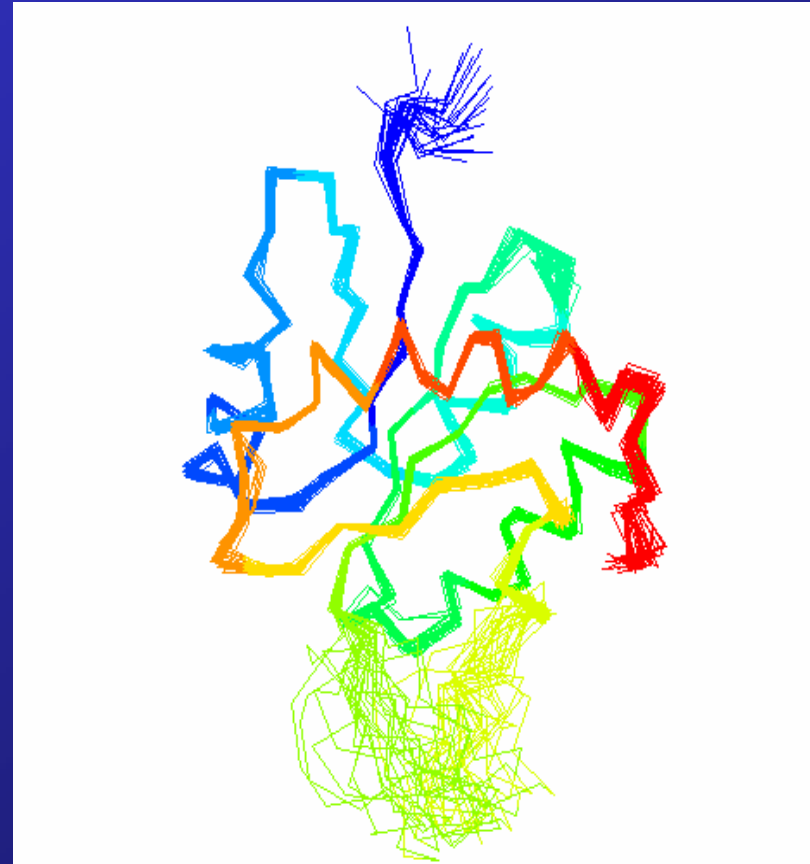
RosettaDom gets one domain cut at 94.

Overview

- Fully automated domain boundary prediction is a hard unsolved problem
 - On larger benchmark sets “completely correct” predictions are obtained for only ~35% of *multi-domain* proteins
- Computer-assisted prediction more tractable
- DomPred2 – available after CASP7
 - Better sequence-based scoring scheme inc. full fold recognition (mGen-3D)
 - Specific domain-linker prediction including disorder prediction
 - Machine learning to combine features

Predicting Disordered Regions in Proteins

KaiA N-terminal domain, *S. elongatus* (NMR data)



Experimental determination of disorder

- X-rays detect disorder as missing electron density or high B-value regions
- NMR spectroscopy reliably determines disorder
- Other techniques: circular dichroism, protease digestion, hydrodynamic parameters (e.g. Stoke's radius)

Disordered proteins

- Some proteins are partly or completely unfolded *in vitro* and yet they are known to be functional *in vivo*
- Ensemble of configurations rather than a unique, ordered structure
- Disordered states can be compact (molten globule) or extended (random-coil)
- They become folded (ordered) upon binding

Protein Disorder

- The dominant view of protein structure-function has been that an amino acid sequence specifies a (mostly) fixed 3-D structure that is a prerequisite to protein function.
- In contrast to the dominant view, it is now becoming clear that many proteins display functions *requiring* the disordered state.

How common is disorder?

	Dunker et al. >40	Ward et al. >30	Ward et al. >50
bacteria	16-45%	2.0%	0.7%
archaea	26-51%	4.2%	1.6%
eukaryotes	52-67%	33.0%	19.6%

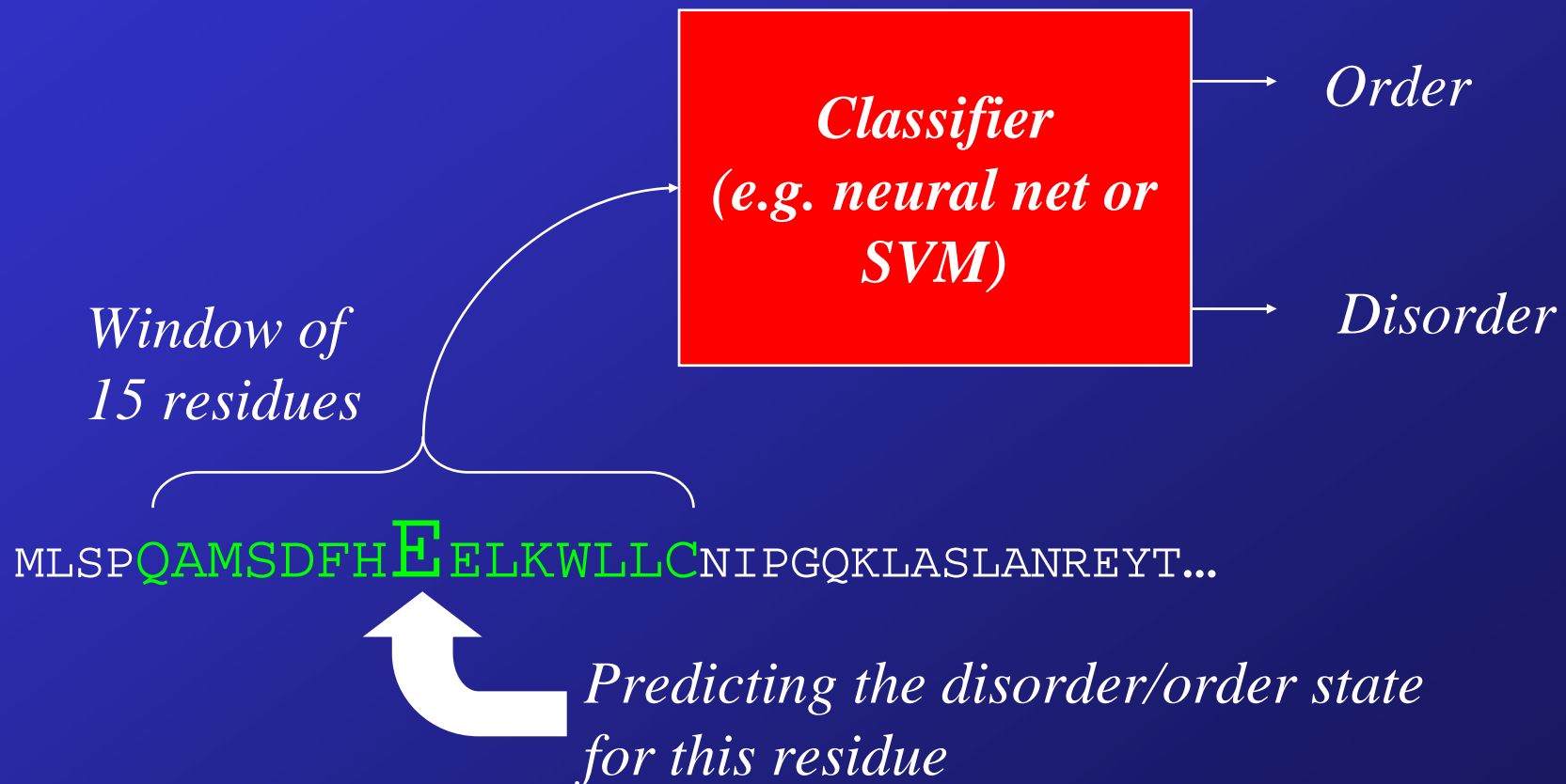
- Eukaryotes have greater need for signalling, control or regulation
- No compartment protection from proteolysis in prokaryotes

Possible advantages of disorder

- Allows recognition and binding of several targets with high specificity and low affinity?
- Allows quick response to environmental changes?
- Facilitates assembly of multi-protein complexes?

Disorder is usually conserved across families

Basic Scheme for Predicting Disorder by Machine Learning

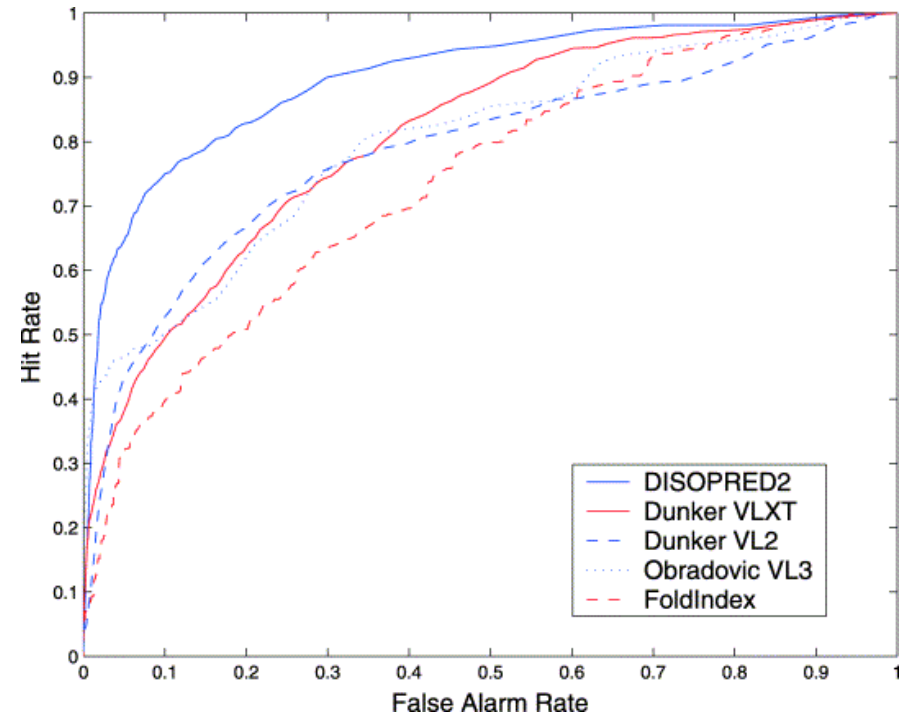
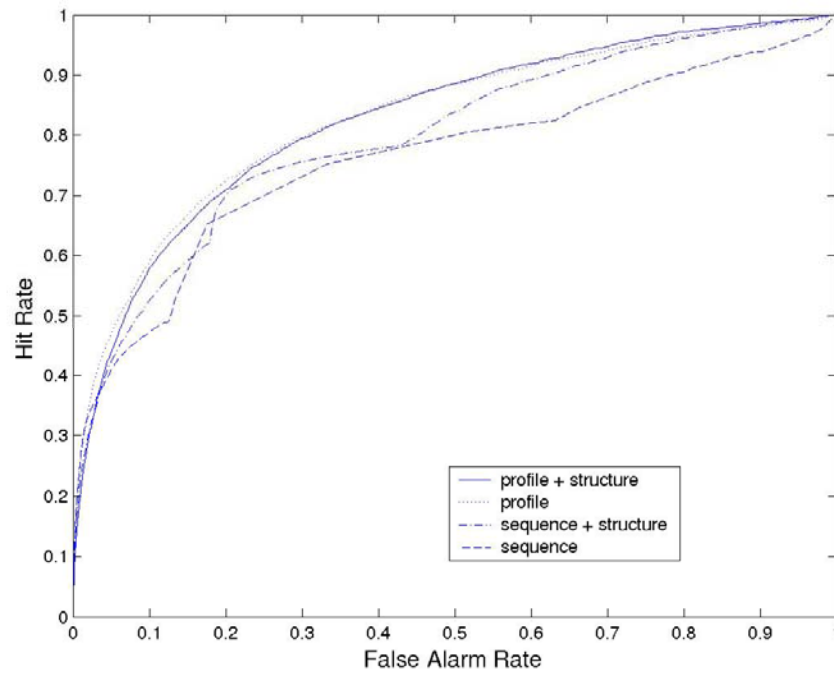


Training an SVM Classifier to Recognize Native Disorder

- DISOPRED2 trained on non-redundant set of crystal structures (725 structures)
- High resolution structures ($< 1.7 \text{ \AA}$)
- Disorder defined by residues in the sequence records that do not have atomic co-ordinates.

Disorder Prediction

95% accuracy but most sites are ordered!



DISOPRED2 Accuracy

- 40% of disordered residues can be predicted with an error rate of 1%
- 60% of residues predicted with an error rate of 10%
- HOWEVER ~90% of long regions (30 residues or more) can be found with < 0.1% error



Bioinformatics Unit

[Home](#)

The DISOPRED Prediction of Protein Disorder Server

Information

Dynamically disordered protein chains do not have stable secondary structures and have high flexibility in solution. A description of DISOPRED and the relevance of disorder to protein function can be found [here](#).

Input Sequence

Input sequence (single letter code) [Help](#)

```
IMWRNAKRQSDRFYDEDVF INGEGLPEQDTRGVNAHMVTNHHALRSRDNIY  
TKTLASKAHTD TTSLRSPSSLAMTQKSSSQASLKGISLKETNGHLVKQSER  
NGSIAKVASPPVVEKRLLQPLSSTPVTQLQAEPKRVP TAAASVSGSSRSTTPV  
THTTTATLSSQPAAQPRRTHLVEGVPQTSV
```

Prediction Options

[Help](#)

Acceptable False Positive Rate: 2%

Include PSIPRED secondary structure prediction

Output Options

[Help](#)

- Don't return PSI-BLAST output
- Return PSI-BLAST hits only
- Return PSI-BLAST hits and alignments

Warning: PSI-BLAST can produce very large output files - please be sure you are able to receive very long e-mail messages if you use these options.

Submit Sequence

[Help](#)

E-mail address

Short name for sequence

<http://bioinf.cs.ucl.ac.uk/disopred>

Ward JJ and Jones DT (2003) Evolutionary information improves prediction of protein disorder. *Bioinformatics*.

In preparation.

globin

PINE 4.44 MESSAGE TEXT

Folder: INBOX Message 2,069 of 2,069 70%

Date: Tue, 14 Oct 2003 00:21:32 +0100 (BST)
 From: j.ward@cs.ucl.ac.uk
 To: D.Jones@cs.ucl.ac.uk
 Subject: Results for sequence: T0170

Results of DISOPRED:

DISOPRED predictions for a false positive rate threshold of: 2%

AA: IMMWRNAKRQSDRFYDEDUFINGEGLEPEQDTRGVDNAHMUTNHHALRSRDNIEYRSPS
 pred: *****.....
 10 20 30 40 50 60

AA: TKTLASKAHTDTTSLRSPSSLAHTQKSSSQASLKSGISLKETNGHLUKQSERAAATPRSQQ
 pred: ..*****.....
 70 80 90 100 110 120

AA: NGSIAKUASPPVEEKRLQLPLSSTPUTQLQAEPAKRUPTAASUSGSSRSTTTPUSARSTT
 pred: *****.....
 130 140 150 160 170 180

AA: THTTATLSSQPAAPRRTHLVEGUPQTSU
 pred: *****.....
 190 200 210

Asterisks (*) represent disordered predictions and dots (.) prediction of order. Links to disorder profile plots and plain text files containing classifier outputs are given below.

Quantitative results are available in plain text format at:
http://bioinf.cs.ucl.ac.uk/domout/disopred/0_18_45_586_14-10-2003.txt

A postscript (.ps) version of the disordered profile can be found at:
http://bioinf.cs.ucl.ac.uk/domout/disopred/0_18_45_586_14-10-2003.ps

portable document format (.pdf) version at:
http://bioinf.cs.ucl.ac.uk/domout/disopred/0_18_45_586_14-10-2003.pdf

jpeg graphics at:
http://bioinf.cs.ucl.ac.uk/domout/disopred/0_18_45_586_14-10-2003.jpg

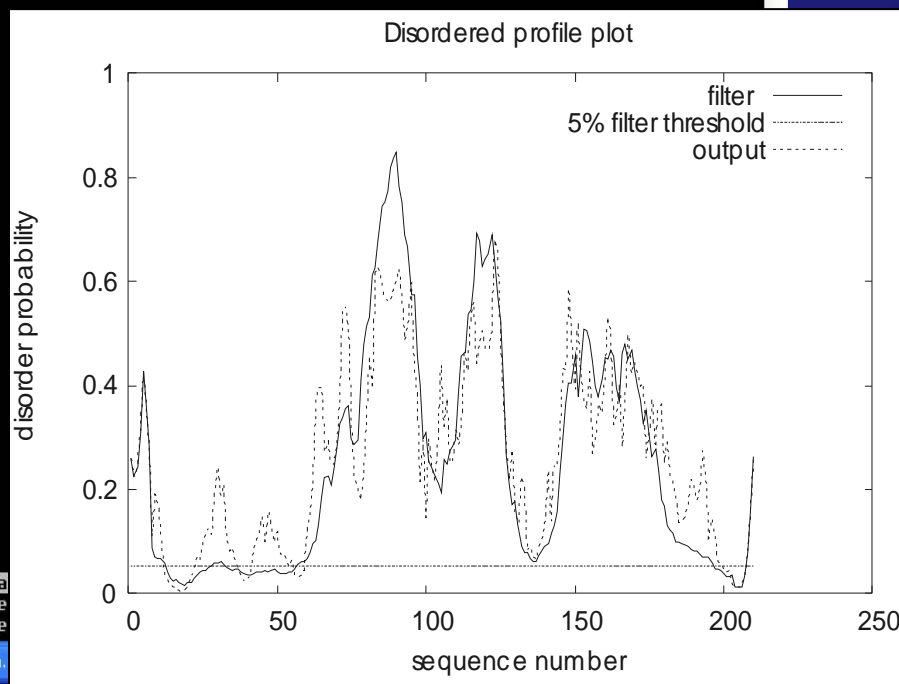
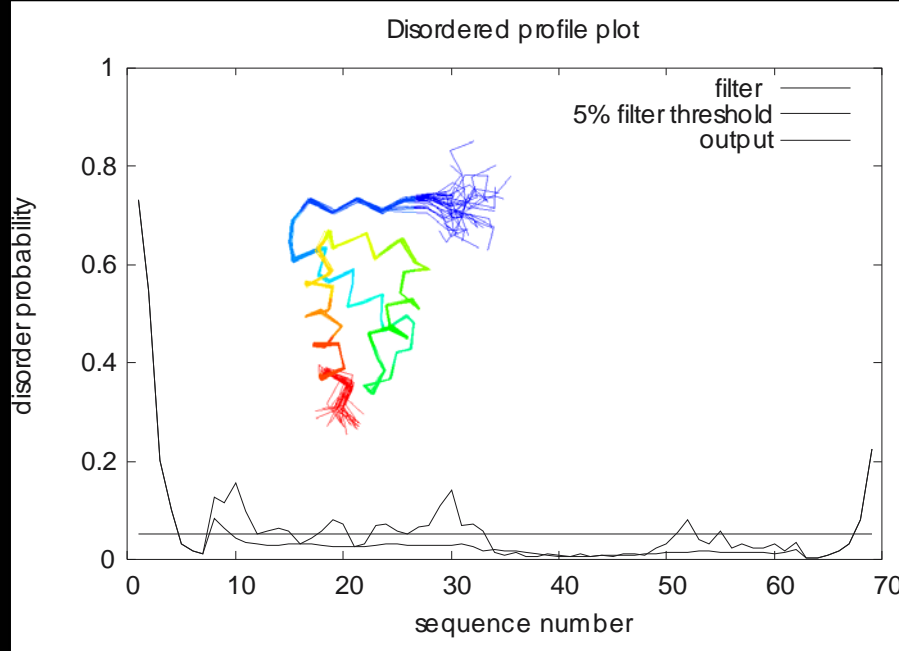
Results from PSIPRED:

Conf: 9755554211122100001467888401102220223321111247896225515874
 Pred: CCCCCCCCCCCCCCHCCCCCCCCCHHHCCCCCHHHHHCCCCCCCCCEEEEECCCC
 AA: IMMWRNAKRQSDRFYDEDUFINGEGLEPEQDTRGVDNAHMUTNHHALRSRDNIEYRSPS
 10 20 30 40 50 60

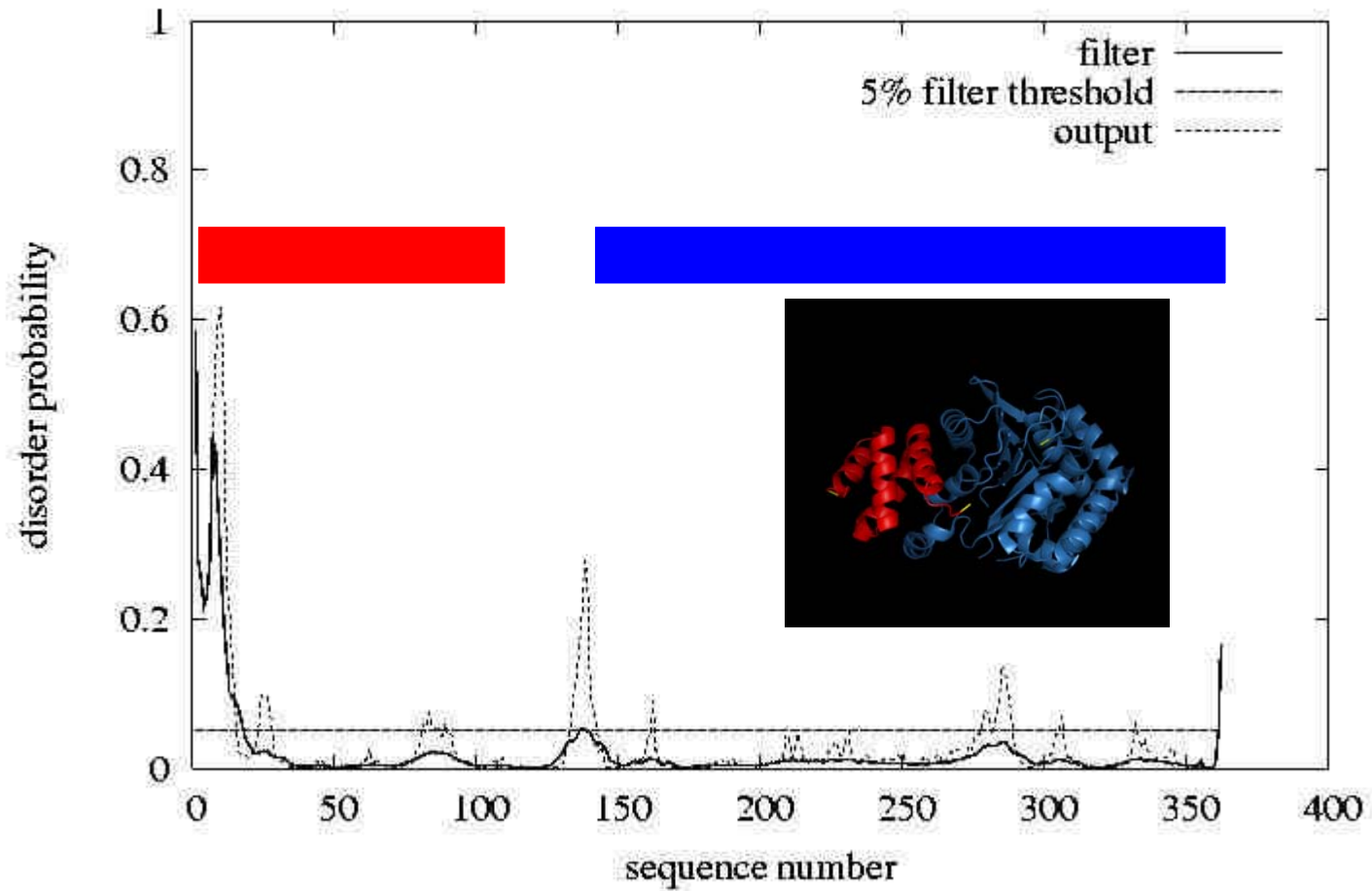
Conf: 32111347888863337765222104655100012112445663103325444874334
 Pred: CCCCCCCCCCCCCCEEECCCCCCCCCCCCCCCCCEEECCCCCCCCCCCC
 AA: TKTLASKAHTDTTSLRSPSSLAHTQKSSSQASLKSGISLKETNGHLUKQSERAAATPRSQQ

Help OTHER CHDS MsgIndex ViewAttch PrevMsg NextMsg Spc NextPage

start DigiGuide 6.0 globin D_18_45_586_14-10-... Submission SuccessFu

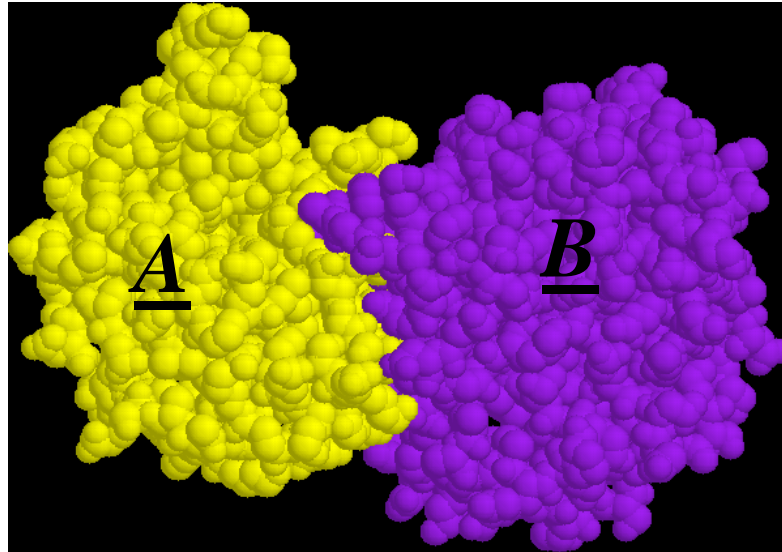


Disordered profile plot

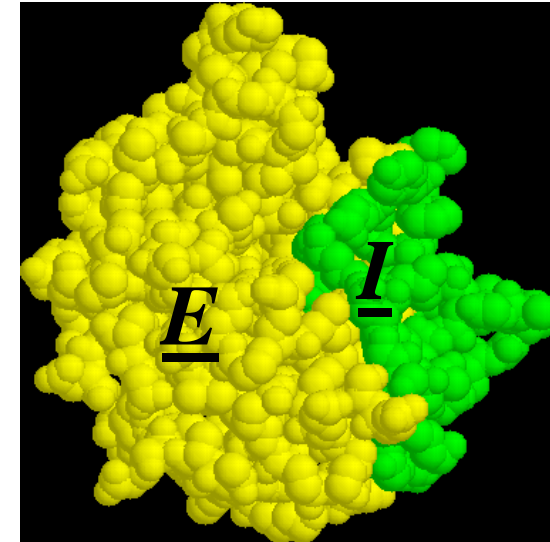


Predicting Domain-Domain Interactions

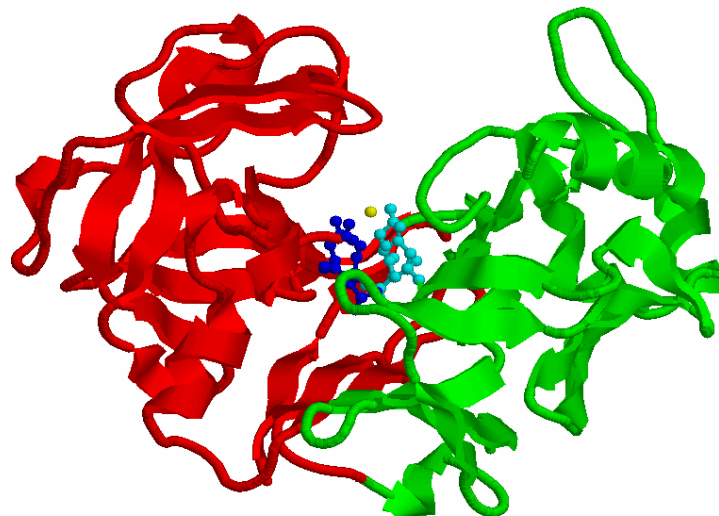
Protein-protein interactions



Obligate



Non-obligate



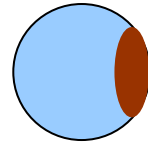
Multidomain
(obligate)

Overall Goals of Knowledge-based Prediction of Protein-Protein/Domain-Domain Interactions

- Predict likely interactions between chains and domains in genomes
- Predictions should be applicable to modelled structures
- Predictions should be relatively fast
- Predictions may be useful for ranking potential protein-protein complexes prior to full docking or structure determination

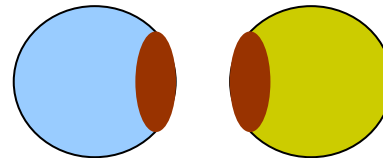
Overview of Methods

● Unilateral Methods



- Patch analysis
 - Predicting interface regions using physical characteristics of patches e.g. hydrophobicity or planarity
- Machine learning approaches
 - Pattern recognition of interface regions (neural nets/SVMs)

● Bilateral Methods

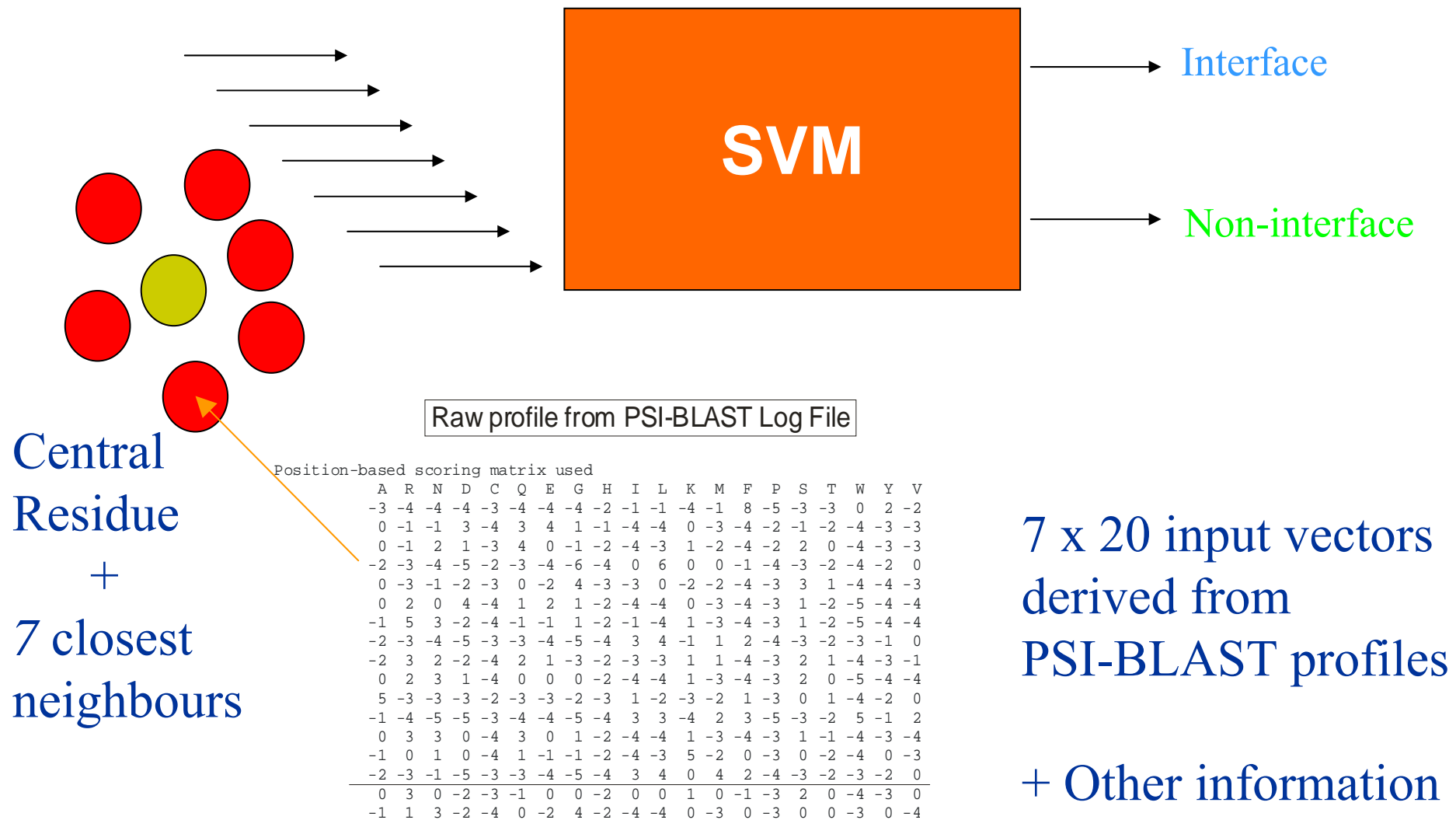


- Full atomic docking
- “Complex Recognition”
 - Recognizing complexes using fold recognition-like techniques (pair and solvation potentials)
- A Hybrid Approach – docking in “Contact Space”
 - Combining patch analysis, complex recognition and machine learning to find possible interaction sites and estimate affinity

Unilateral Methods:

**Machine Learning Prediction of
Interfaces**

Predicting Interface Surfaces with SVMs

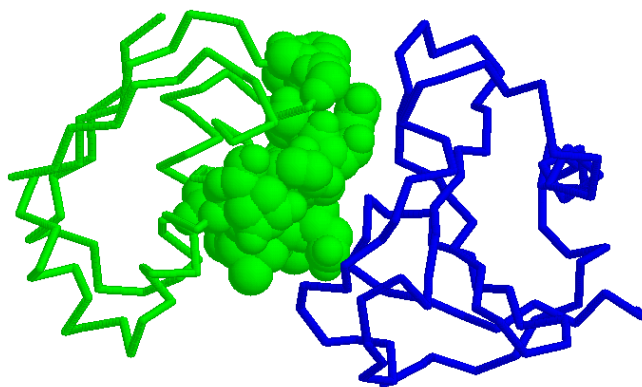


SVM PREDICTION

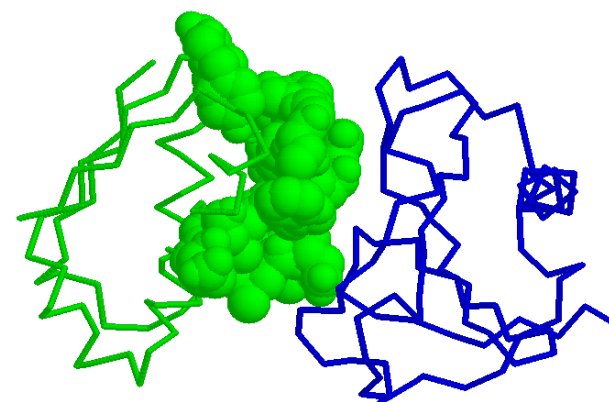
1ay7A (Ribonuclease sa complex with barstar)

Correct predictions: 38/43 (88%)

observed interface residues
residues



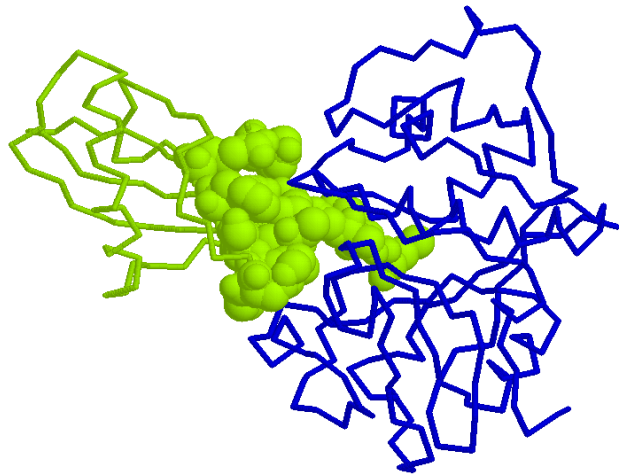
predicted interface



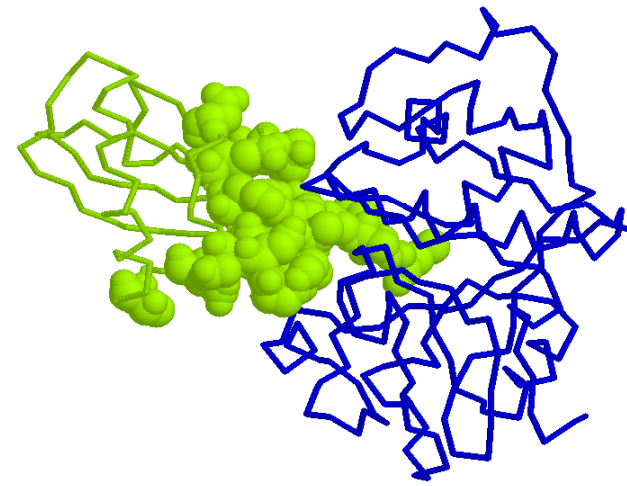
1stfI (stefin B inhibitor complexed with papain)

Correct predictions: 47/55 (85%)

observed interface residues
residues



predicted interface



Bilateral Methods:

Docking Domains in Contact Space

(Lise S., Walker-Taylor A., Jones D.T. BMC Bioinformatics 2006, 7:310)

Pros & Cons

- PROs
 - Simple representation
 - No distances/angles need to be calculated
 - Any interaction pattern can be represented – even allowing for conformational changes
 - Could even be applied
- CONs
 - Very large configuration space
 - Many patterns not embeddable in 3-D space!
 - Does not produce 3-D structure directly

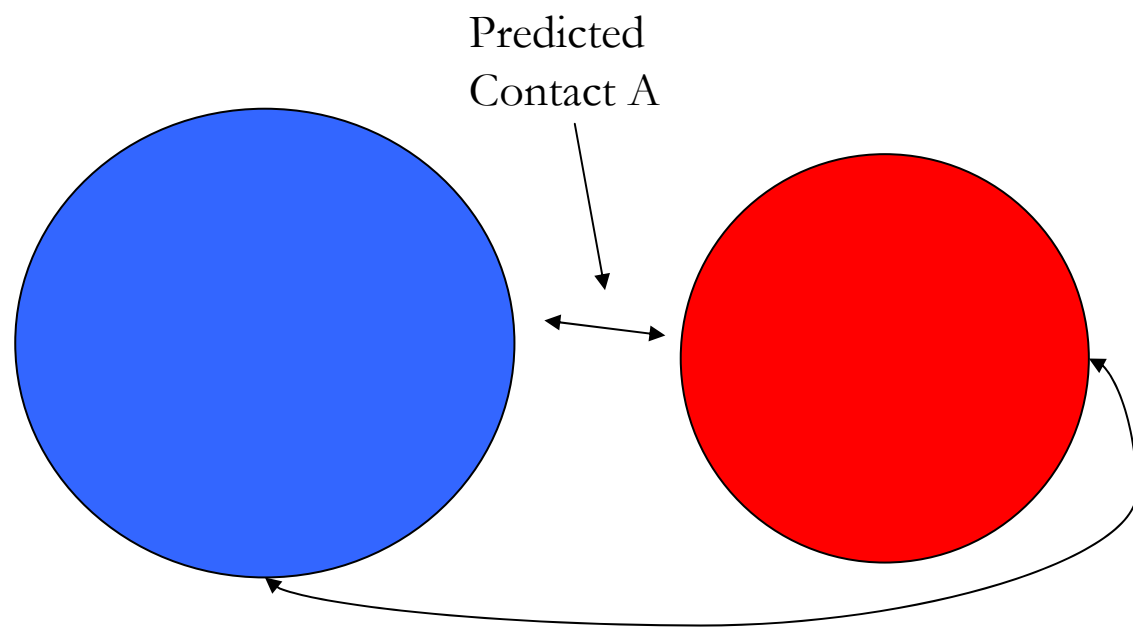
Rules of the Game

- Only accessible residues are considered
- Each residue can make a maximum of n contacts ($n = 2$ or 3 typically)
- Total number of contacts $< K$
 - Can be estimated based on size of proteins
- All contacts must satisfy geometric constraints (i.e. be 3-D embeddable)
- “Game” is played using simulated annealing

Search Space

- Consider two proteins with 100 accessible residues each
- Contact map has $100^2 = 10000$ cells
- Total of 2^{10000} configurations! However...
- Assuming ~10 interacting residues & 3 contacts per residue
- Total configurations $\sim 10000^{30} = 10^{120}$
- Achievable by simulated annealing search

Geometric Compatibility

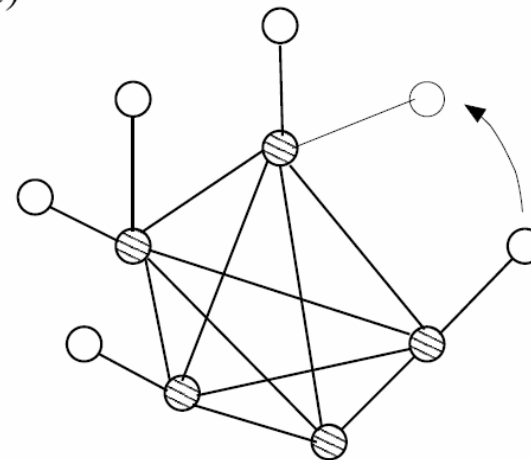


Predicted Contact B is NOT
Geometrically compatible

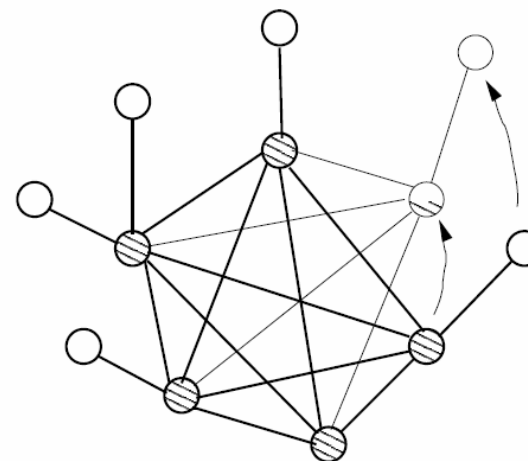
Annealing moves based on a random graph

- Define a random graph with $m \ll K$ nodes representing top scoring contacts and edges representing geometric compatibility between contacts
- Identify initial clique of 4 or 5 nodes with 5-6 peripheral nodes
- Moves involving cliques can be local (a) or large-scale (b)

(a)



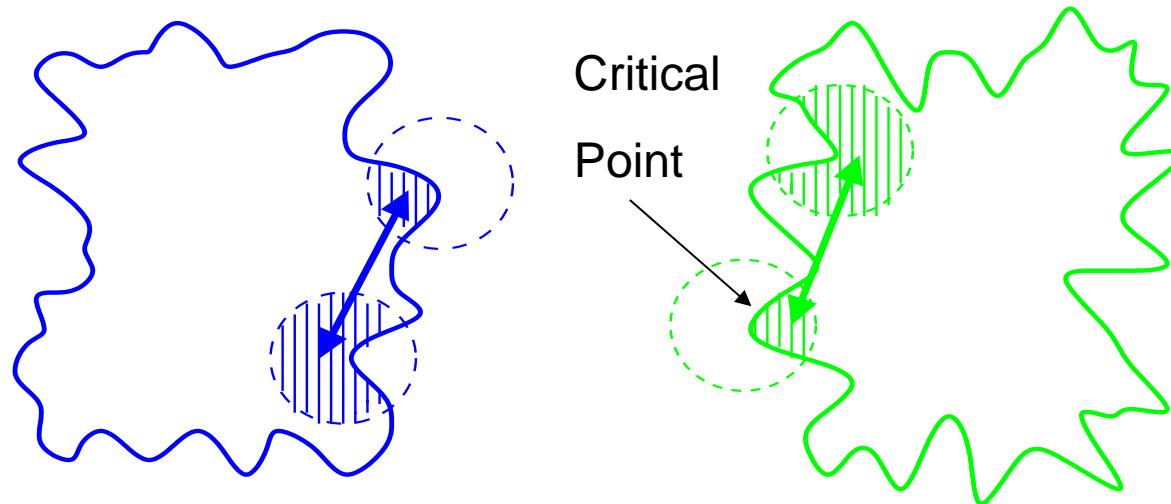
(b)



Objective Function

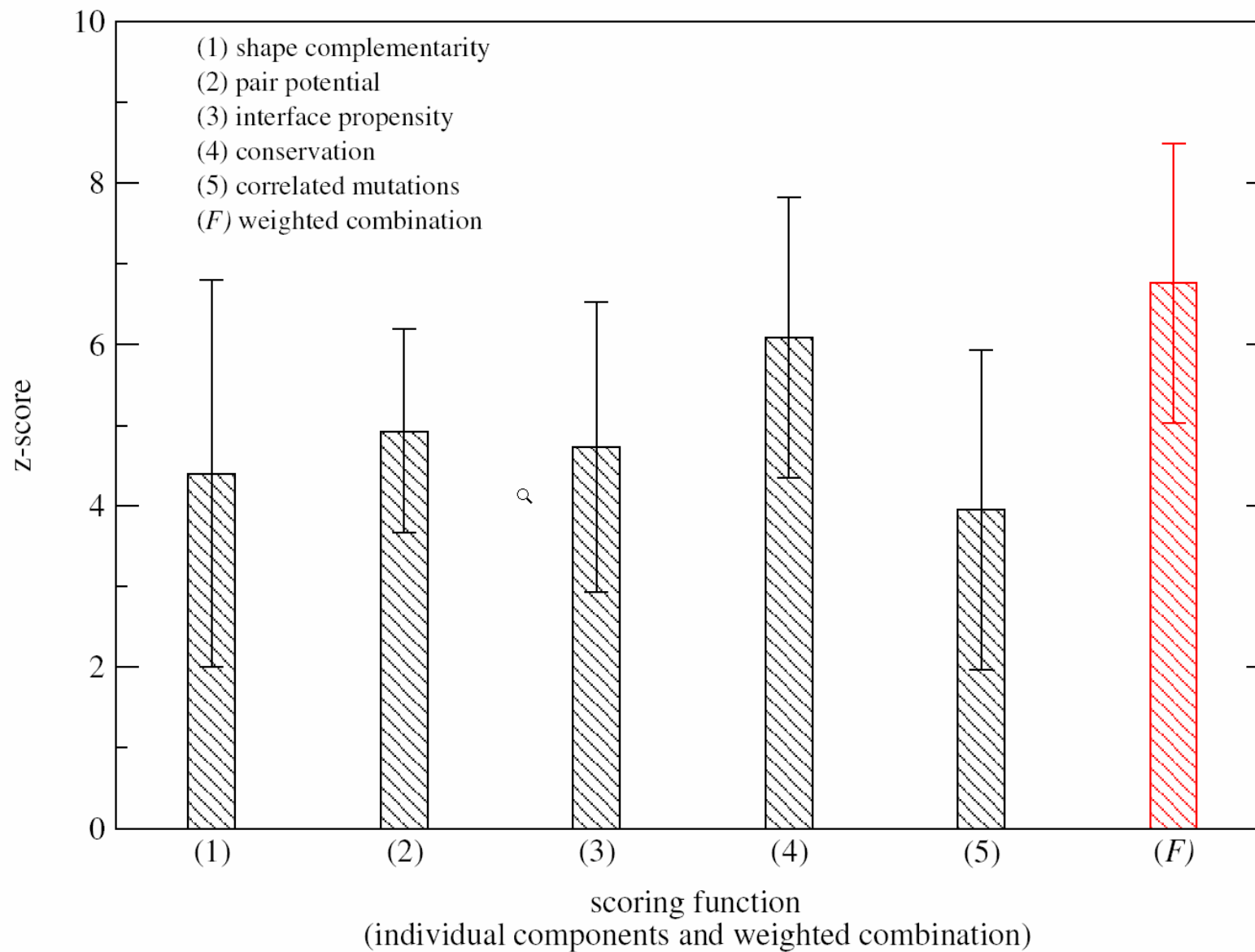
- Residue conservation score
 - Pair potentials
 - Interface propensity
 - Shape complementarity
 - Correlated mutation score
-
- Jack-knifed grid search used to find optimal linear combination of scores

Shape complementarity



- A knob matches a hole (and vice-versa)
- Consider spheres centred on each residue
- A knob “fits” a hole IF the sphere volume not occupied by protein in A equals the sphere volume occupied by protein in B
- Distances between pairs of critical points should also be similar

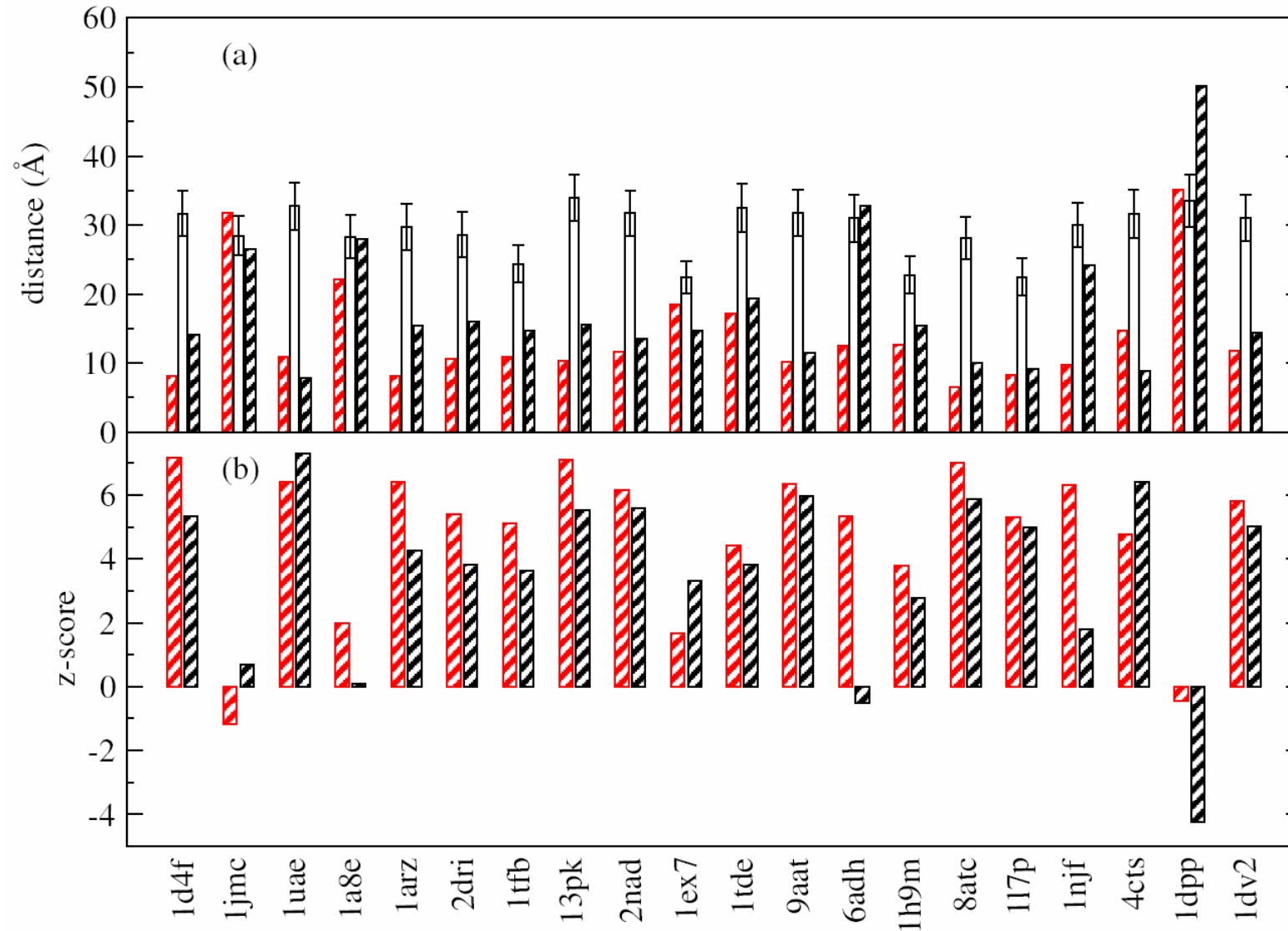
Performance of terms in recognizing known contacts



Proof of principle

- Selected domain pairs from multi-domain protein where BOTH an open and closed conformation is available in PDB
- Managed to find 20 pairs
- Task is to predict the contacts in the closed form based on the individual domains taken from the open form

Average distance between predicted contacts compared to random



Acknowledgements

Jon Ward

Stefano Lise

Liam McGuffin

Kevin Bryson

Russell Marsden

Alice Walker-Taylor

Prof. Bernard Buxton