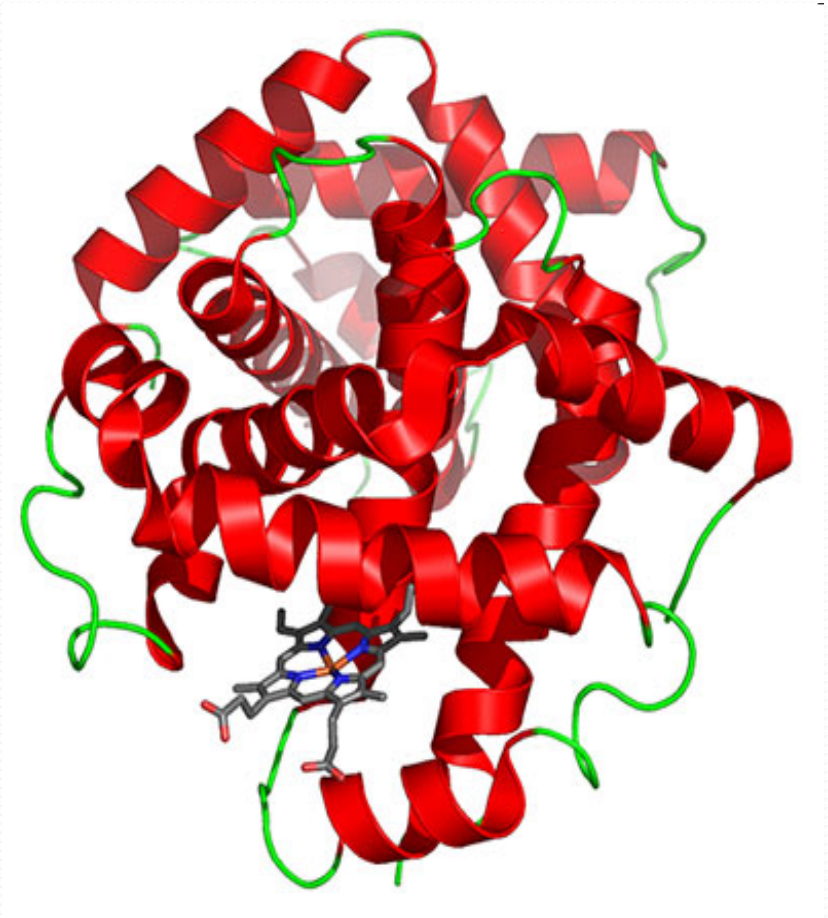
Local descriptors of protein structure: A universal approach to protein structure representation

Torgeir R. Hvidsten
Umeå Plant Science Centre (UPSC) and
Computational Life Science Cluster (CLiC)

Structure - function

3D shape determines function



Hemoglobin

Hammer

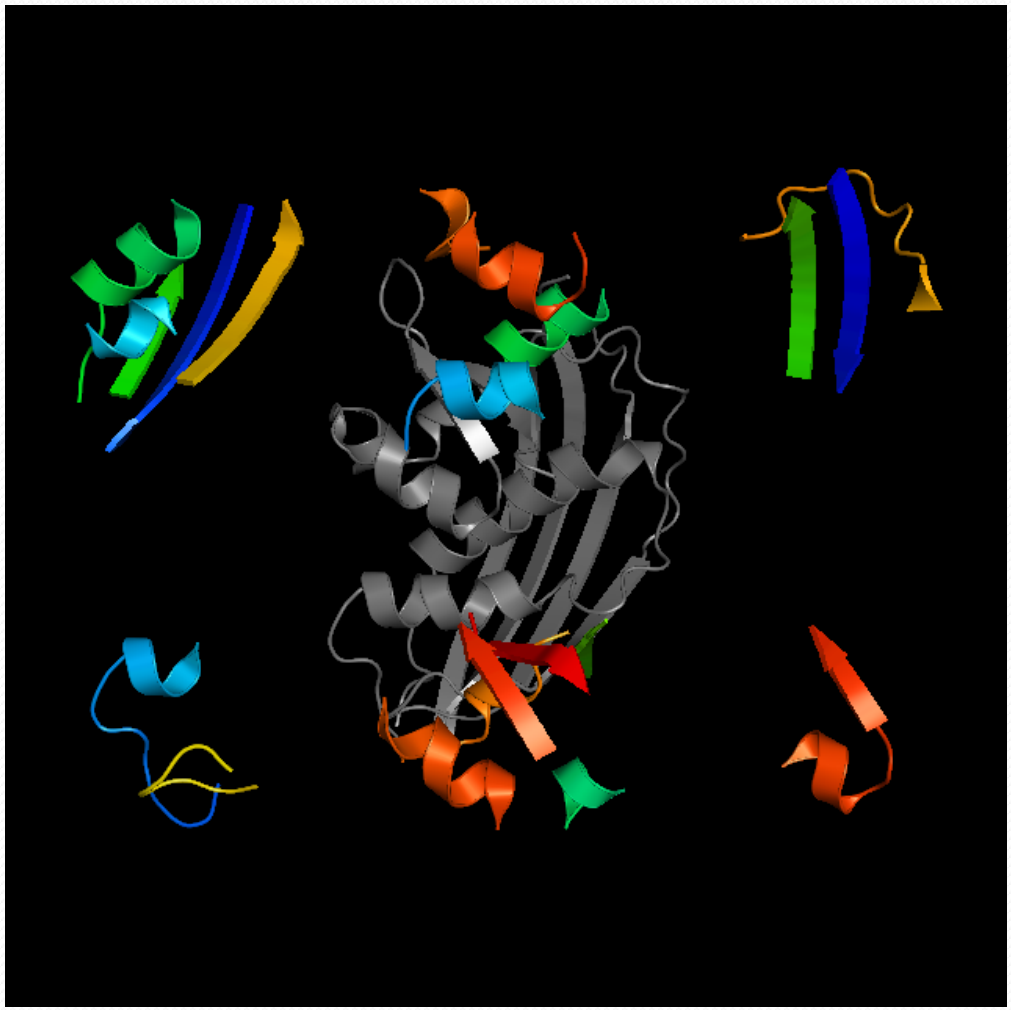
Machine learning using structure

- How do we represent a protein structure as training examples ?

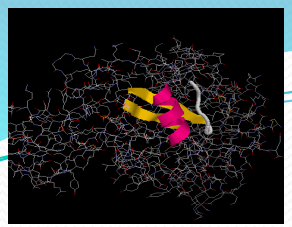
- We don't! Nearest neighbor prediction (e.g. global structural alignment)

- We take the challenge! Need to define a library of common structural properties (features, attributes)

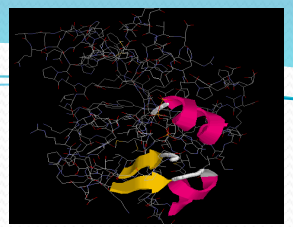
Local descriptors are building blocks for protein structure



IF



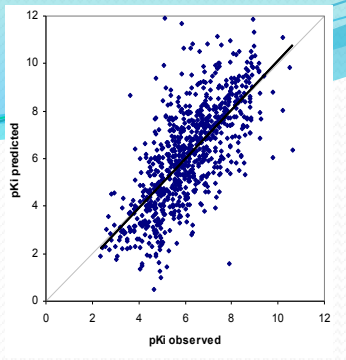
AND



THEN GO:0016646

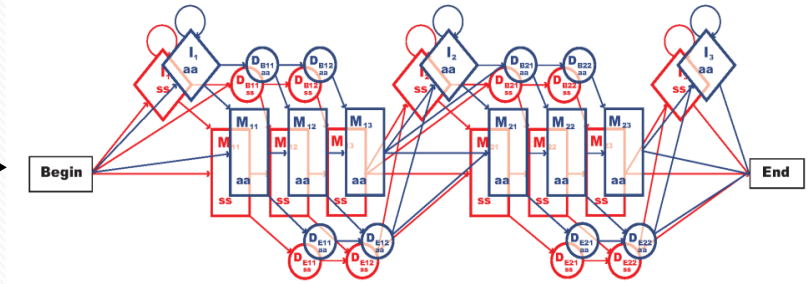
Prediction of function

Prediction of molecular interaction

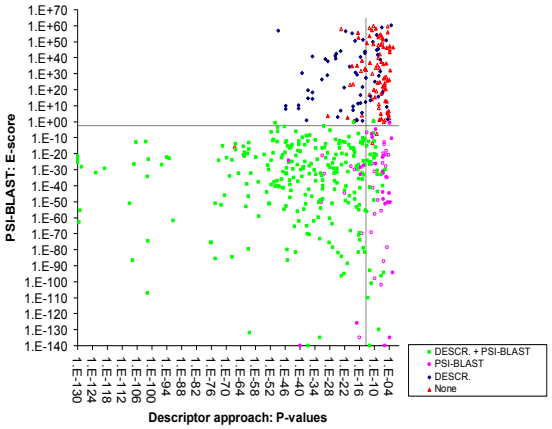


Structure prediction

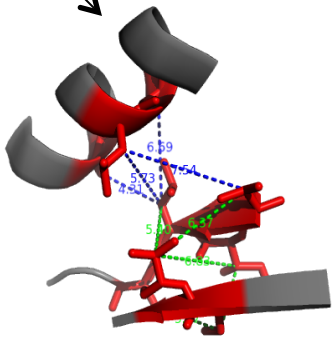
Sequence



Fold recognition

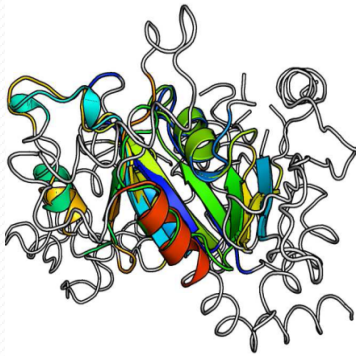


Contact prediction



1. Fold classification
2. Ab inito structure prediction
3. Domain prediction

Structure alignment

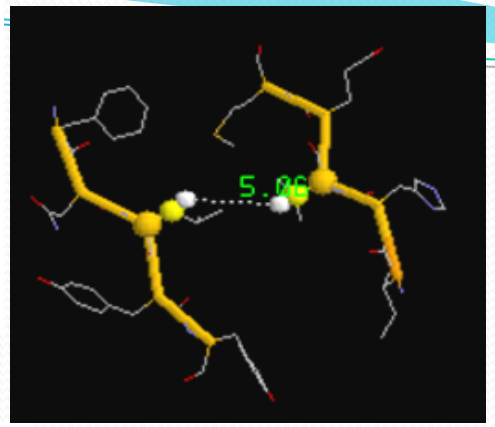


Local descriptors of protein structure

1. T. R. Hvidsten, A. Kryshtafovych, J. Komorowski and K. Fidelis. A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins, *Bioinformatics* 19 Suppl 2 (European Conference on Computational Biology): II81-II91, 2003.
2. T. R. Hvidsten*, A. Kryshtafovych* and K. Fidelis. Local descriptors of protein structure: A systematic analysis of the sequence-structure relationship in proteins using short- and long-range interactions, Proteins: Structure, Function, and Bioinformatics 75 (4): 870-884, 2009.

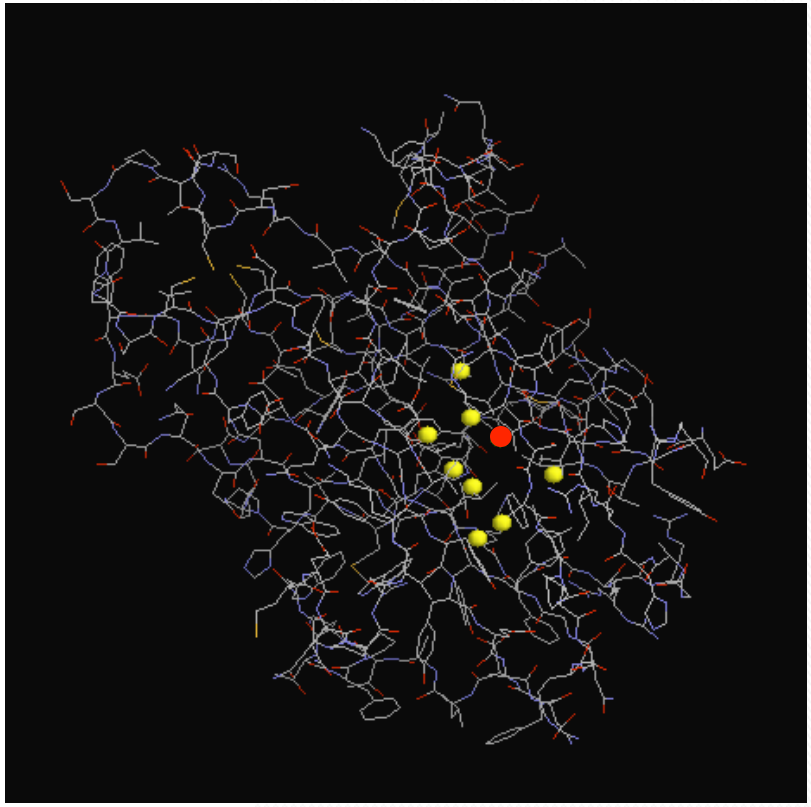
Local descriptors of protein structure

Distance definition

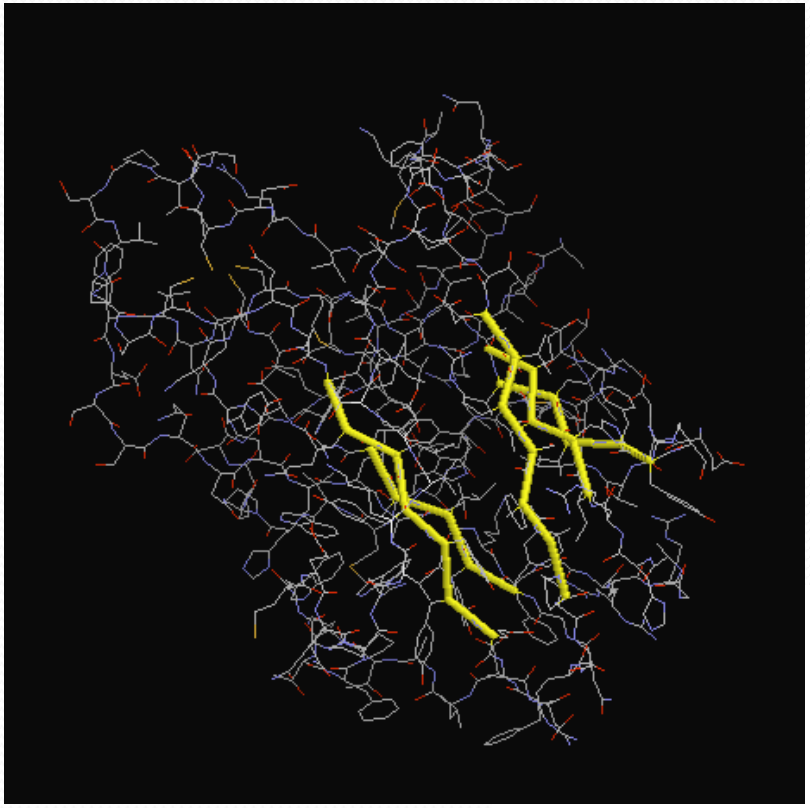


$$\left| C_{\beta x}^{(i)} - C_{\beta x}^{(j)} \right| < 6.5 \mathring{A}$$

Local neighborhood



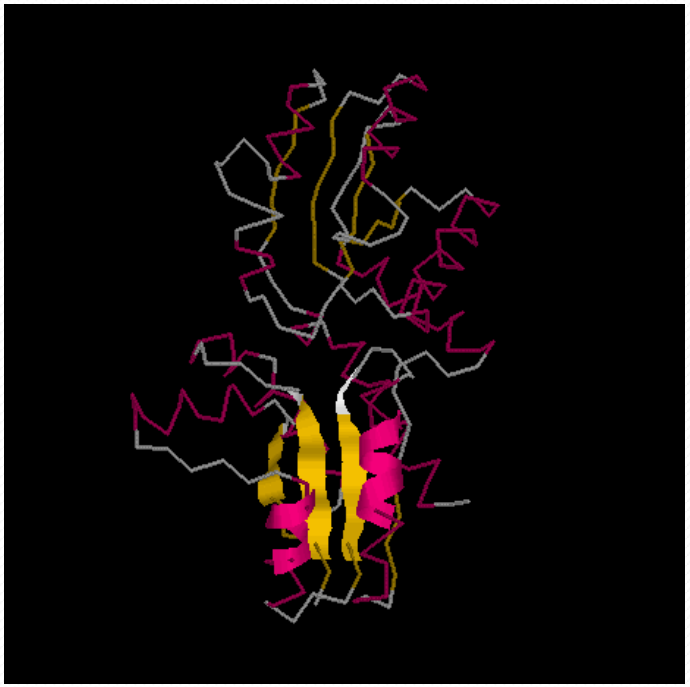
Descriptor: 1nuk_A#96



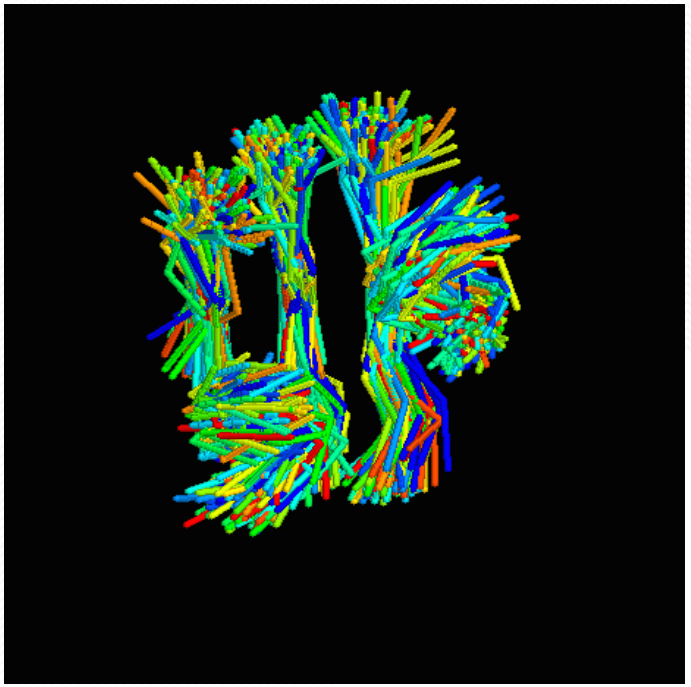
Descriptor group

Descriptor	Segment 1		Segment 2		Segment 3		Segment 4		Segment 5	
1qgoa_#8	4-10	ALLVVSF	39-43	FRAFT	63-67	LQALQ	77-83	VAIQSLH	91-95	EKIVR
1qo2a_#78	74-80	EHIQIGG	47-51	IHVVD	67-71	EKLSE	97-101	--RRQIV	89-93	EKLRK
1rpxa_#73	69-75	LPLDVHL	40-44	IHVDV	58-62	LVVDS	93-97	--DIVSV	85-89	PDFIK
1nsj__#82	78-84	NAVQLHG	58-62	GVFVN	66-70	EKILD	98-104	ILVIKAV	89-93	ELCRK
1mla_1#7	3-9	QFAFVFP	87-91	MMAGH	262-266	EYMAA	270-276	EHLYEVG	283-287	GLTKR
1qfja2#108	104-110	PMILIAG	134-138	TIYWG	183-187	TAVLQ	195-201	HDIYIAG	207-211	KIARD
1efvb1#8	4-10	LRVLVAV	119-123	LVLLG	47-51	EEAVR	59-65	KEVIAVS	76-80	RTALA
1iow_1#8	4-10	KIAVLLG	38-42	YPVDP	48-52	TQLKS	56-62	QKVFIAL	70-74	GTLQG
1yaca_#57	53-59	PTILTTS	80-84	PYIAR	97-101	VKAVK	14-20	AVLLVDH	120-124	AFPAL
1ig0a2#188	184-190	ISLLALG	40-44	TLLIL	128-132	TKCVN	216-222	FKLCYMT	200-204	VHSIT

Central descriptor: 1qgoa_#8

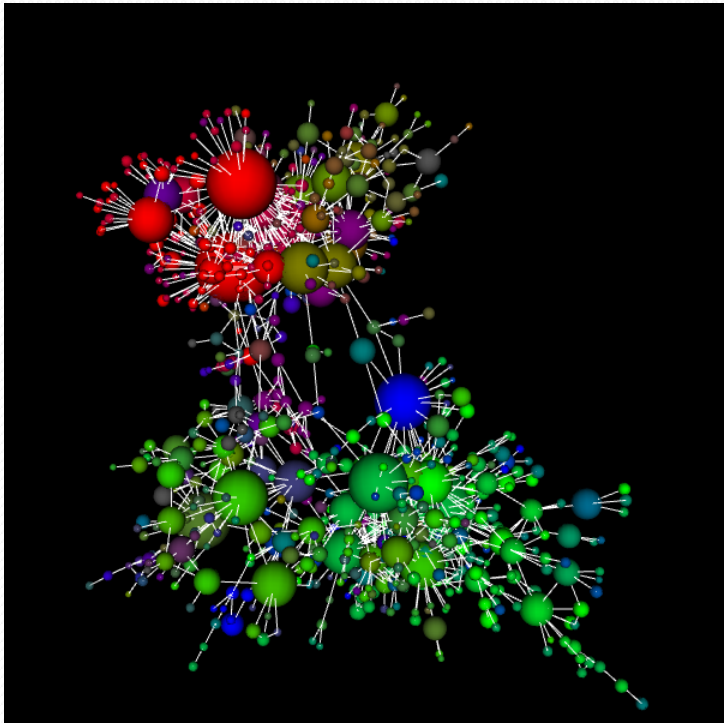


Structurally similar descriptors



Building a library

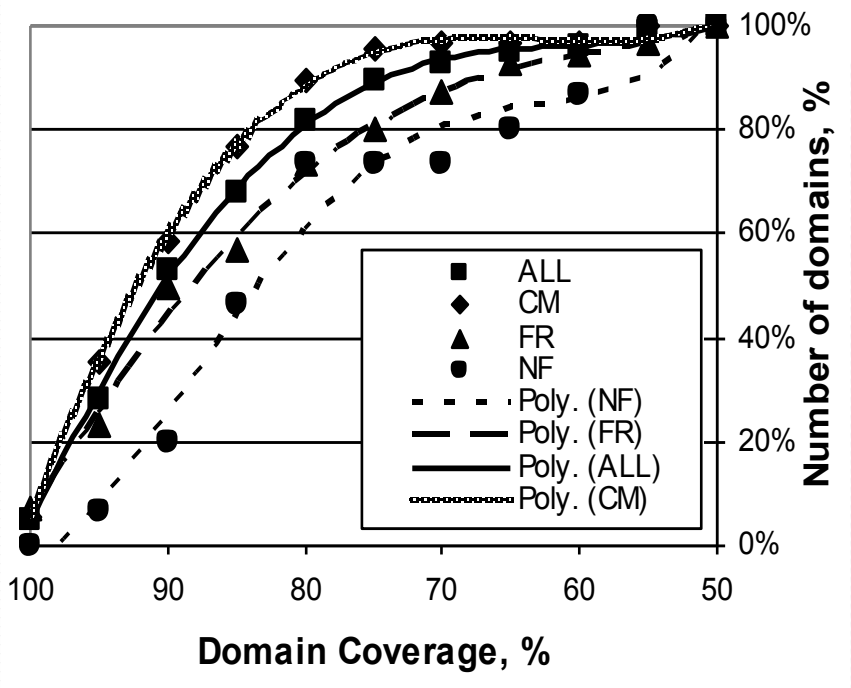
1. Construct descriptor groups for all local descriptors in a representative subset of PDB (ASTRAL 40%)
2. Filter groups based on popularity, number of segments
3. Select a representative set of groups
   i. Regular clustering
   ii. Iteratively selecting the ”best” groups: e.g. best novel coverage in terms of structure, contacts, function, etc.

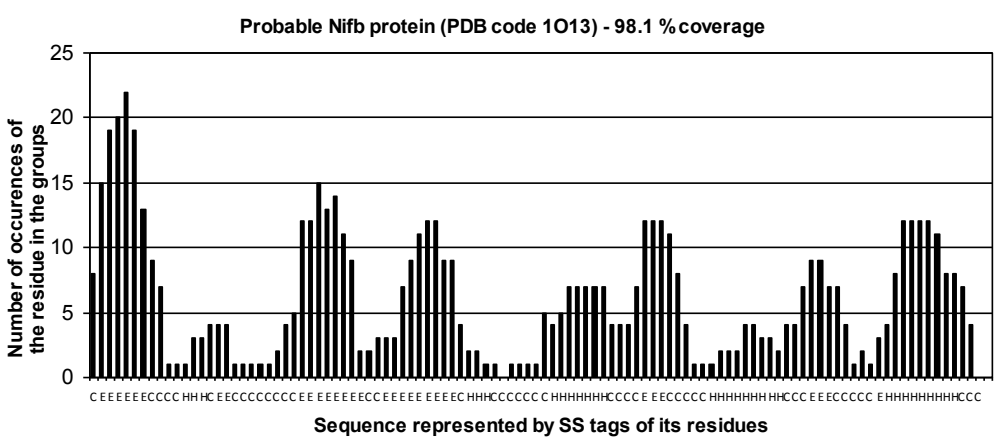


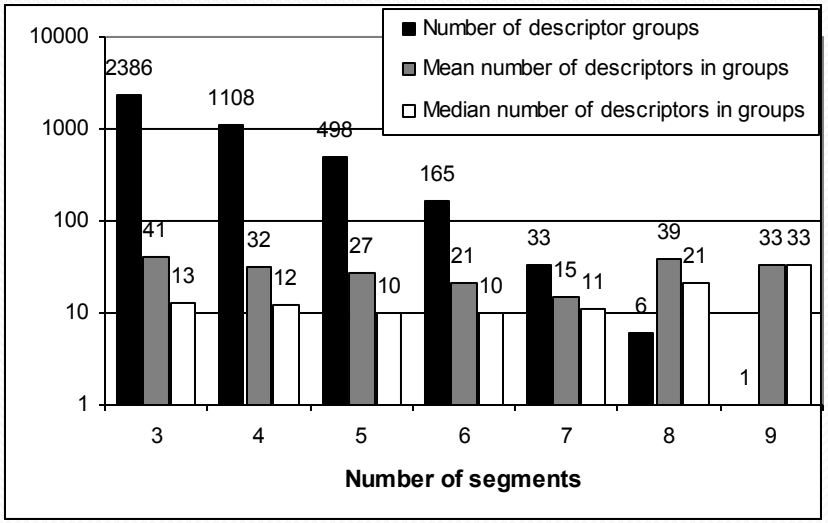
A library of groups comprises a set of building block for protein structure assembly

Library: ~4000 local descriptor groups

Can re-assemble most parts of almost all protein structures in PDB, including new folds

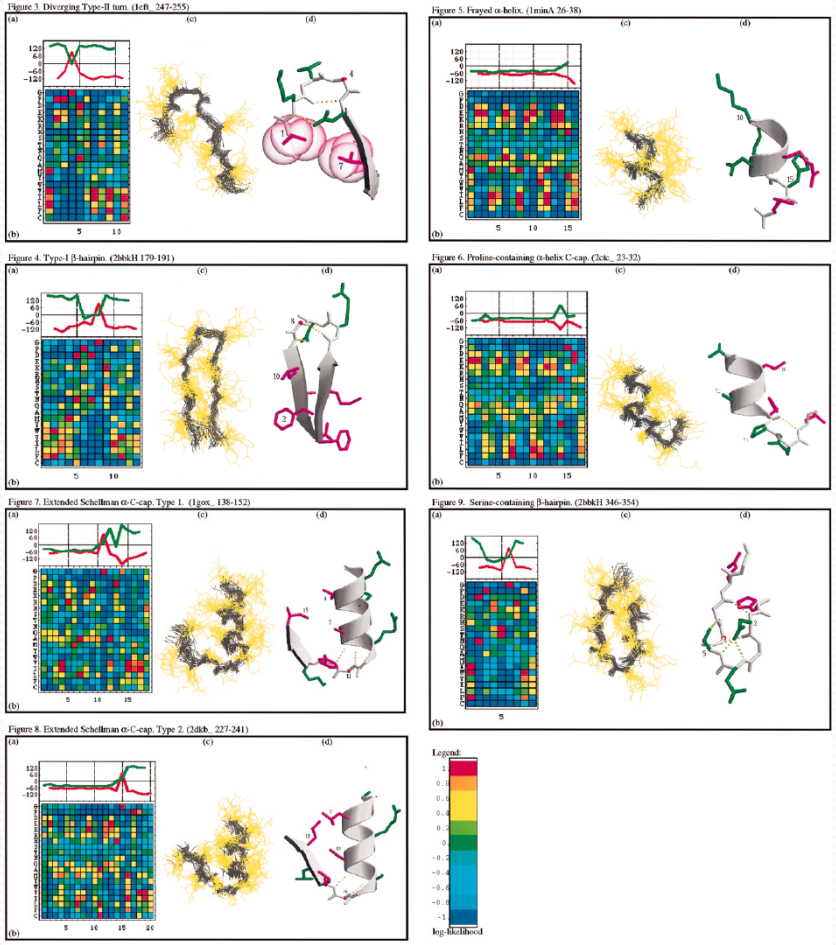






Fragment based methods

- Multi-fragment methods do exist in structure comparison and interaction studies
- Single-fragment methods dominate structure prediction: combined with other methods (e. g. free energy minimization - ROSETTA)
- The lack of correct long-range contacts is a major problem in structure prediction



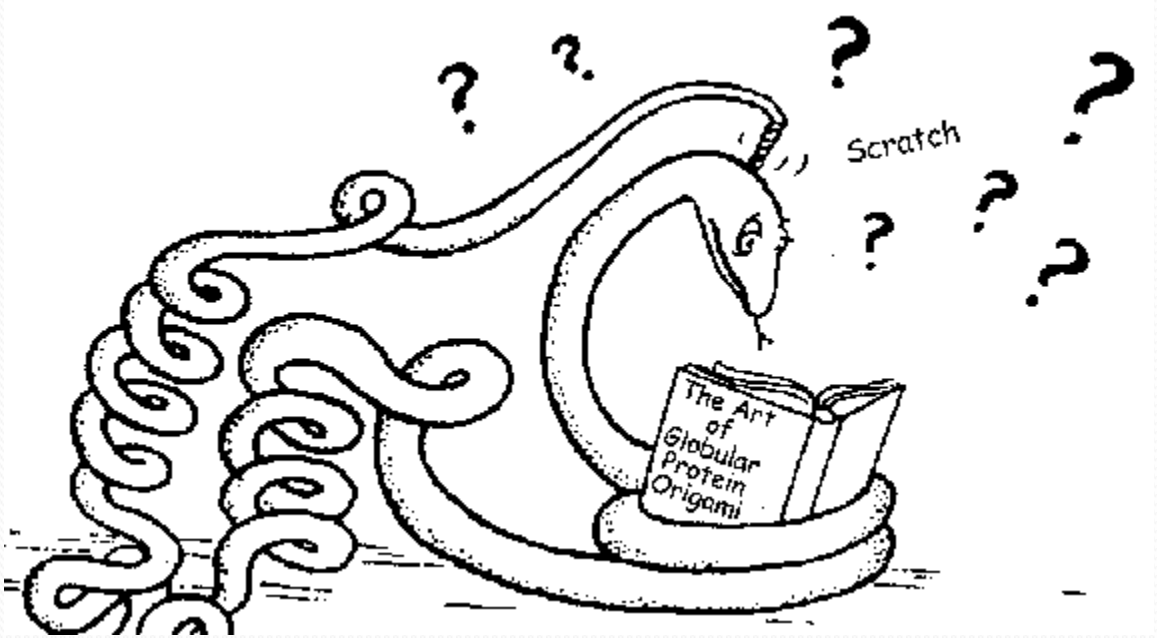
Local descriptor library

- The library contains a systematic collection of all local similarities in PDB
- Local descriptors are large enough to allow identifying meaningful sequence and structure similarity between proteins in PDB
- Local descriptors are small enough to abundantly re-occur in proteins that are not related by sequence homology
- Local descriptors include both short- and long-range interactions – describes the entire neighborhood of the central amino acid

Structure prediction

1. T. R. Hvidsten, A. Kryshtafovych, J. Komorowski and K. Fidelis. A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins, *Bioinformatics* 19 Suppl 2 (European Conference on Computational Biology): II81-II91, 2003.
2. T. R. Hvidsten*, A. Kryshtafovych* and K. Fidelis. Local descriptors of protein structure: A systematic analysis of the sequence-structure relationship in proteins using short- and long-range interactions, Proteins: Structure, Function, and Bioinformatics 75 (4): 870-884, 2009.
3. P. Björkholm, P. Daniluk, A. Kryshtafovych, K. Fidelis, R. Andersson and T. R. Hvidsten. Using multi-data hidden Markov models trained on local neighborhoods of protein structure to predict residue-residue contacts. Bioinformatics 25: 1264-1270, 2009.

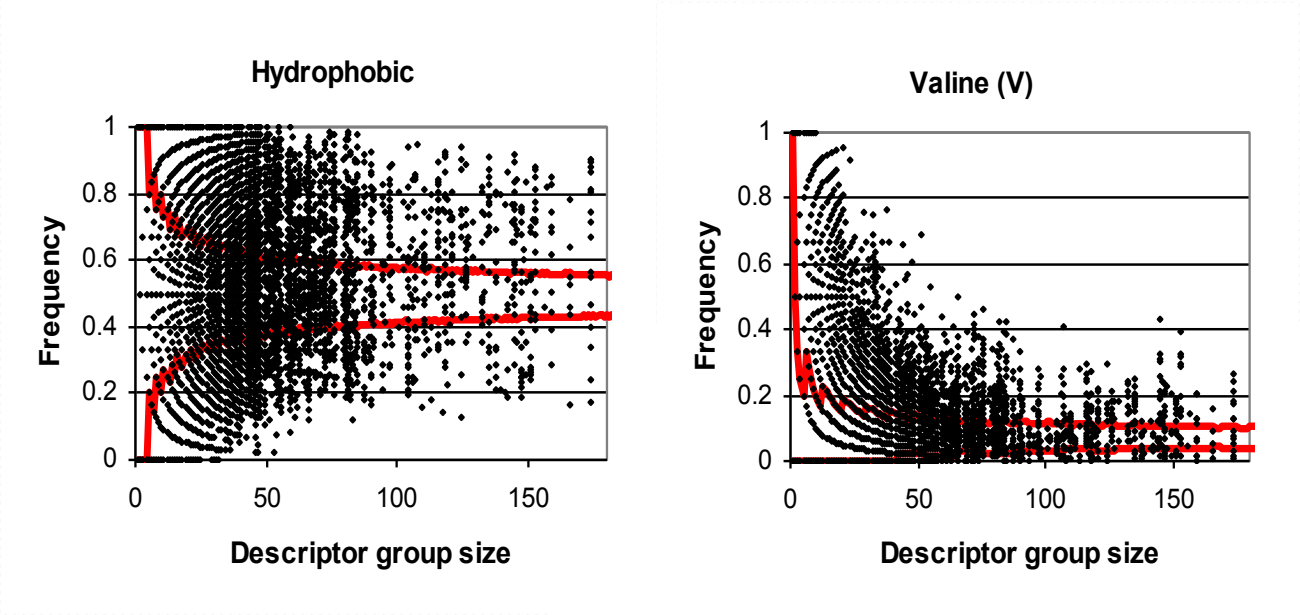
The protein folding problem



Signal extraction

```
DESCRIPTOR GROUP
1ay7_B#81   ITIIL   LVLEW--   SVLQVFREA
1b9y_C#114  GFVYE   VKFCKIR   ALNSSLEC-
1bpm#143    VSAKL   PSVEVDP   ---EGAVLG
1cjm_A#190  VLYLF   VVYVARN   ---QEWWEL
1cy4_A#5    KALVI   DYVVKSS   SELKQLAEK
1d4g_A#382  VGVHF   VPAINVN   TGVHNLYKM
1dos_A#11   PGVIT   VPVILHT   SGAHHVHQM
```

Signal extraction aims at identifying sequence patterns in groups and is based on the observed frequencies of amino acids and amino acid substitution groups in specific positions



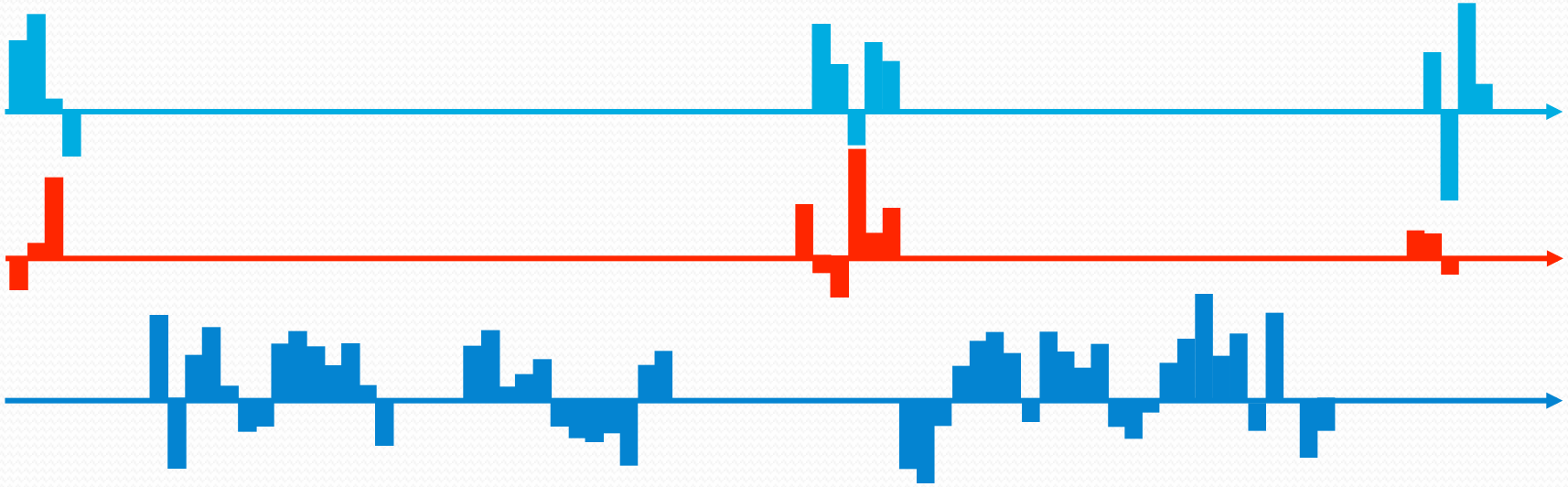
Group assignment

DESCRIPTOR GROUP			
1ay7_B#81	ITIIL	LVLEW--	SVLQVFREA
1b9y_C#114	GFVYE	VKFCKIR	ALNSSLEC-
1bpm#143	VSAKL	PSVEVDP	---EGAVLG
1cjm_A#190	VLYLF	VVYVARN	---QEWWEL
1cy4_A#5	KALVI	DYVVKSS	SELKQLAEK
1d4g_A#382	VGVHF	VPAINVN	TGVHNLYKM
1dos_A#11	PGVIT	VPVILHT	SGAHHVHQM
SS	*EEEEE*	*EEEEEEE*	*HHHHHHHHH*

$\vec{S}_1$ $\vec{S}_2$ $\vec{S}_3$

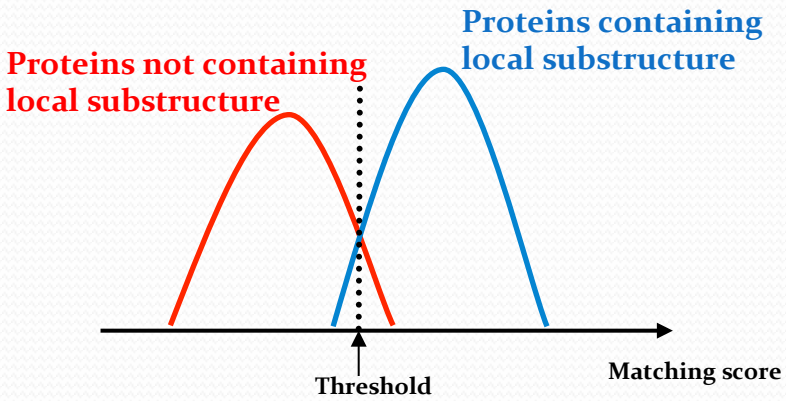
Target

KKAVINGEQIRSISDLHQTLKKELALPEYYGENLDALWDCLTGWVEYPLVLEWRQFEQSKQLTENGAESVLQVFREAKAEGCDITIILS
KHCTISGRAVHSLDELYDEIARQLPLPDYFGRNLDALWDVLSTDIEGPVELIWEDSEHSKRSMGKDYERVVALLKDLTEEREDFRIV
IIGSKIYTEQDFHNQISKIFSIQDYYGNNLDALWDLLSTNVERPITLVWKDAMFSKNQLENIFIEIVNVLERVKKQDED
QSKQEVLETIATSFLFPKHFGKNYDALYDCLTDLVQFVIVLE--QLPVAQKFDKEGRETLLDVFREA
CEEEEECCCCCCHHHHHHHHHHHCCCCHHHCCCHHHHHHHHHCCCCCCEEEEECCHHHHHHHHHHHHHHHHHHHHHCCCEEEEEC



Fold recognition

- Match each group (local protein structure) to the target sequence
- Assign groups with a score higher than the threshold
- Rank folds according to P-values



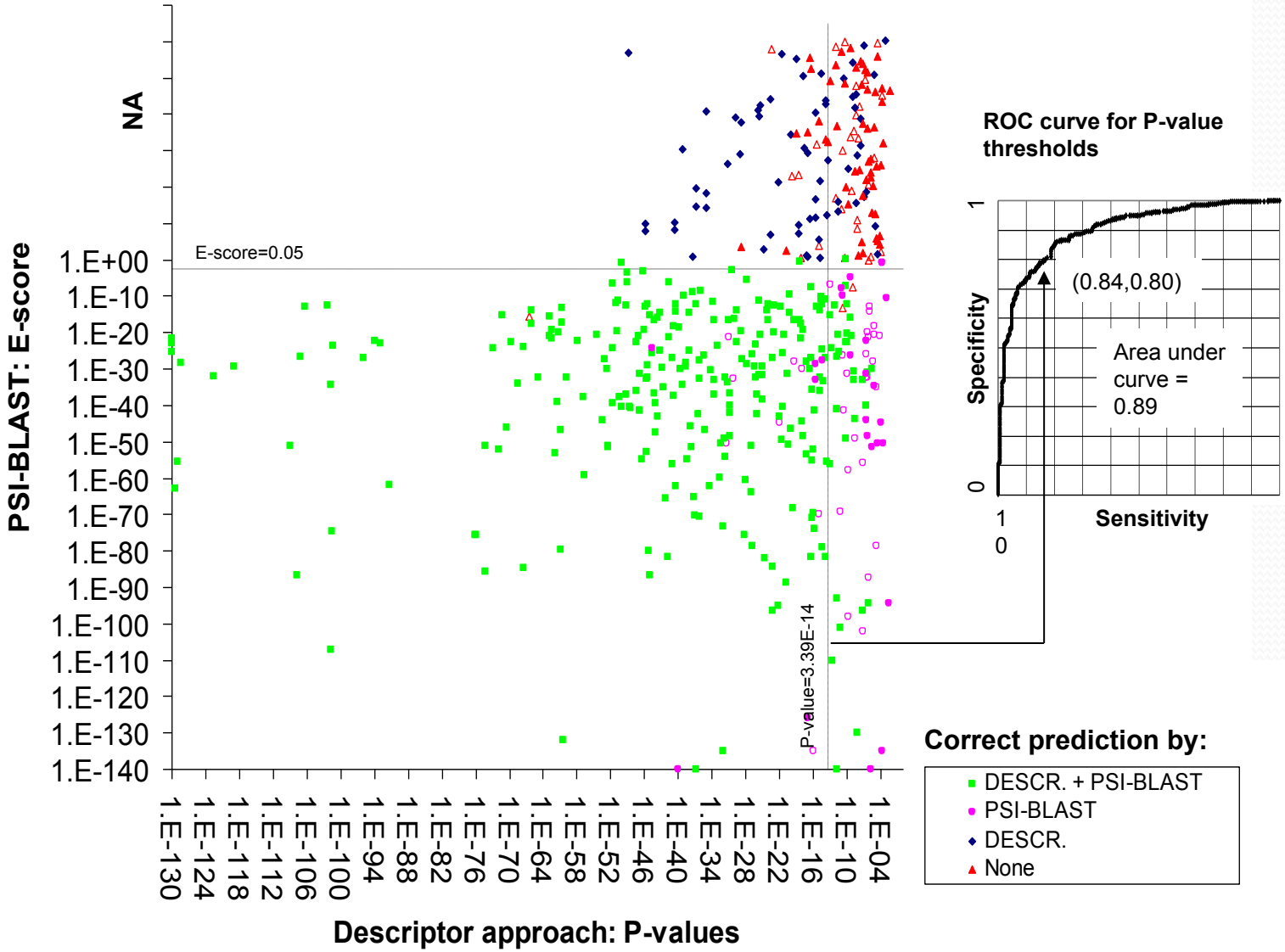
Domain: **1e9ra_**

SCOP fold: P-loop containing nucleoside triphosphate hydrolases (**c37**)

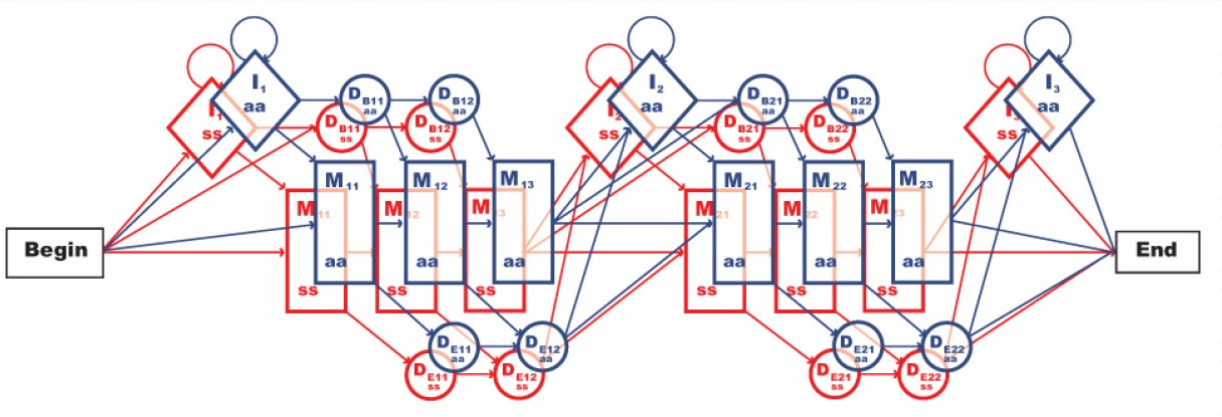
Fold	**Assignment**	**P-value**
1. c37	41/113	5.324e-36
2. d159	3/7	0.0008695
3. c66	5/40	0.0066181
4. e7	2/8	0.0226421
5. c2	112/186	0.0240892
6. b82	3/26	0.0425178
…		
30. d153	1/77	0.9088401

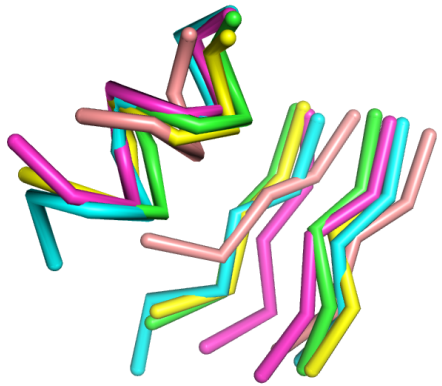
Fold recognition

479 domains with less than 40% sequence identity to the training set



Hidden Markov Models





```
GROUP: 1io7a_#13 : 5
1cpt__#39  36-42    DEQPLAM  54-61    ATKHADVM  326-332  EVRGQNI  a.104 Pseudomonas sp.
1io7a_#13  10-16    KKDPVYY  23-30    VFSYRYTK  266-272  KLGDQTI  a.104 Archaeon Sulfolobus solfataricus
1jfba_#30  27-33    ATNPVSQ  45-52    VTKHKDVC  299-305  MIGDKLV  a.104 Fungus
1jipa_#28  25-31    ETAPVTP  42-49    VTGYDEAK  300-306  EIGGVAI  a.104 Saccaropolyspora erythraea
1n40a_#30  27-33    TREPIRK  45-52    VSSYALCT  293-299  QVGDVLV  a.104 Mycobacterium tuberculosis
```

- One HMM per local descriptor group
- Incoperation of both sequence and secondary structure information

Group assignment quality

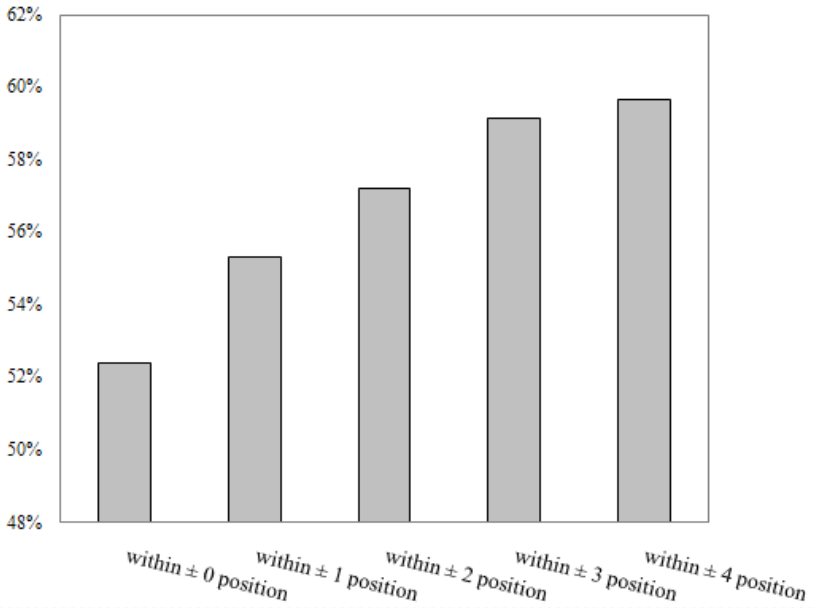
```
>protein2
RGLPGSFIKEVQKGGPADLAGLEDEDVIIEVNGVNVLDEPYEKVVDRIQSSGKNVTLLVCGKKAQD
IIIIIIIMMMMMMMIIIIIIIIIIMMMMMMMMMIIIIIIIIIIIIIIIIMMMMMMMMMIIIIIIIIIII Prediction
IIIIIIIMMMMMMMIIIIIIIIIMMMMMMMMMIIIIIIIIIIIIIIIMMMMMMMMMIIIIIIIIIIIII True alignment

No. Correctly aligned match-states = 19
No. Match states= 23

Result = 19 / 23 = ~83
```

- Targets are represented as multiple alignments (homology information)
- Secondary structure is predicted
- Results for the ”best match” using the Viterbi algorithm

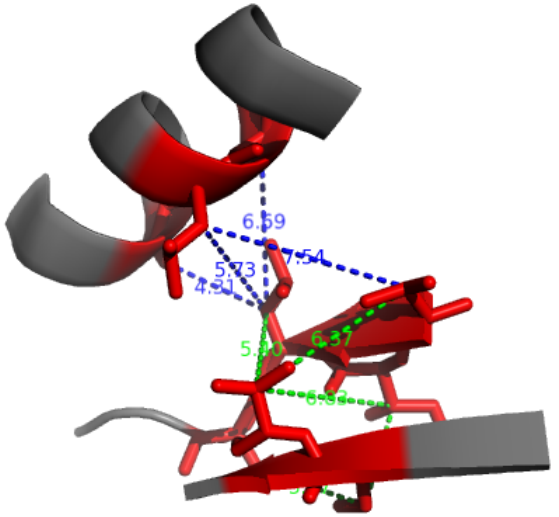
Cross validation on ASTRAL 40%



Contact prediction method

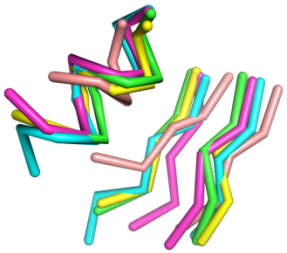
- Groups with a significant score is assigned
- Contacts are transferred form the group
- Contacts predicted by several groups are given more weight

Recombinational repair protein RecR (PDB code 1vdd, chain A), **new fold**



n

Group



```
GROUP: 1io7a_#13 : 5
1cpt__#39   36-42    DEQPLAM   54-61    ATKHADVM   326-332   EVRGQNI
1io7a_#13   10-16    KKDPVYY   23-30    VFSYRYTK   266-272   KLGDQTI
1jfba_#30   27-33    ATNPVSQ   45-52    VTKHKDVC   299-305   MIGDKLV
1jipa_#28   25-31    ETAPVTP   42-49    VTGYDEAK   300-306   EIGGVAI
1n40a_#30   27-33    TREPIRK   45-52    VSSYALCT   293-299   QVGDVLV
```

Contact prediction results

Previously published results suggest that contact prediction accuracy of >22% should improve *ab inito* structure prediction methods

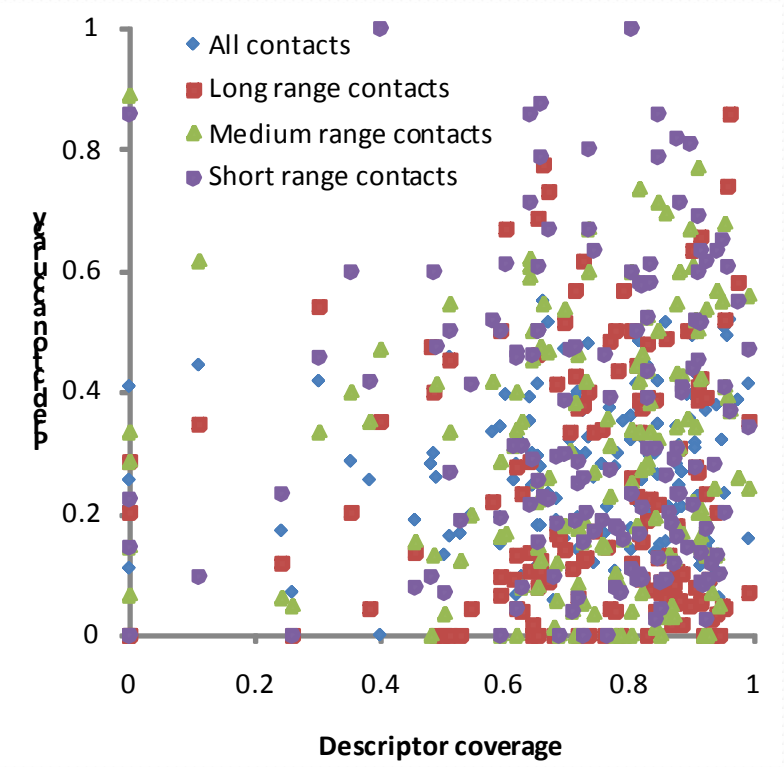
Test set 1: BLAST E-score < 0.05, known fold:

		Long-range	Medium-range	Short-range	All-range
Pct = 0.2	**accuracy**	27.6%	34.1%	37.7%	33.1%
	coverage	5.6%	16.1%	23.6%	11.0%
Pct = 0.5	**accuracy**	21.1%	24.3%	25.4%	23.6%
	coverage	10.5%	28.4%	39.6%	19.6%
Pct = Spline	**accuracy**	16.2%	25.9%	32.0%	21.1%
	coverage	16.2%	26.9%	31.2%	21.3%

Test set 2: BLAST E-score < 0.05, new fold:

		Long-range	Medium-range	Short-range	All-range
Pct = 0.2	**accuracy**	22.8%	28.8%	34.1%	28.6%
	coverage	5.1%	13.7%	22.0%	9.7%
Pct = 0.5	**accuracy**	17.4%	20.7%	22.7%	20.3%
	coverage	9.7%	24.5%	36.0%	17.3%
Pct = Spline	**accuracy**	12.6%	22.0%	28.2%	17.8%
	coverage	13.6%	22.9%	27.9%	18.1%

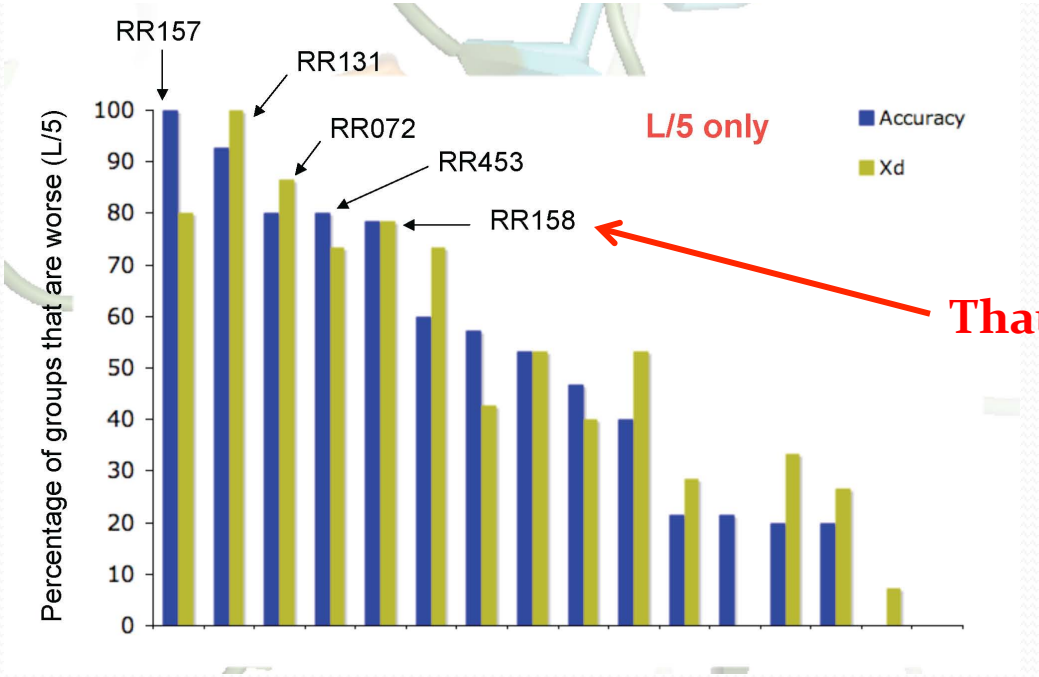
Contact prediction results



- New folds only
- Pct: 0.2
- Descriptor coverage: fraction of structure matched by at least one group in the library (average 73%)

CASP8 performance

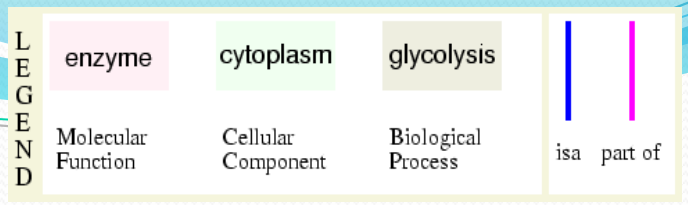
CASP: structure prediction based on sequence followed by evaluation after the structure is solved and published



Predicting molecular function from local descriptors of protein structure

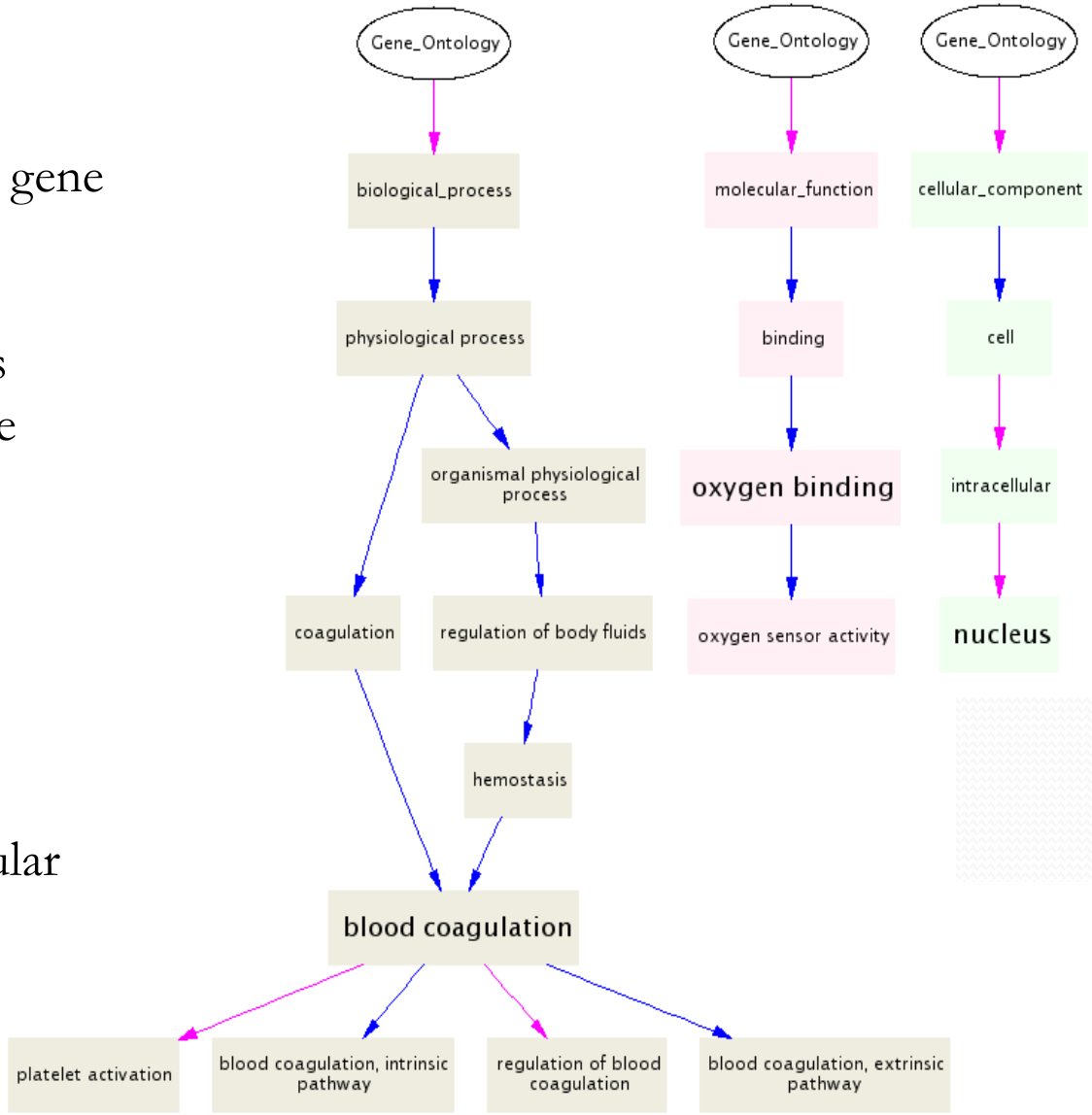
T. R. Hvidsten, A. Lægreid, A. Kryshtafovych, G. Andersson, K. Fidelis and J. Komorowski. A comprehensive analysis of the structure-function relationship in proteins based on local structure similarity. PLoS ONE 4(7): e6266, 2009.

Gene Ontology



Ordered controlled vocabulary organized in a taxonomy for describing the molecular role of gene products

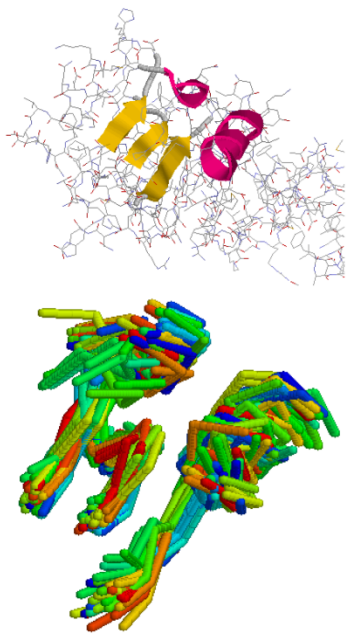
- Molecular function: the tasks performed by individual gene products
- Biological process: broad biological goals that are accomplished by ordered assemblies of molecular functions
- Cellular component: subcellular structures, locations, and macromolecular complexes



Gene Ontology and local descriptors

Gene Ontology	Number of proteins/annotations	Number of classes/proteins/annotations
Molecular function	2549/4963	113/1747/2815
Biological Process	2477/5082	139/1533/2573
Cellular Component	1379/1978	30/561/688

A. Descriptor group: structure



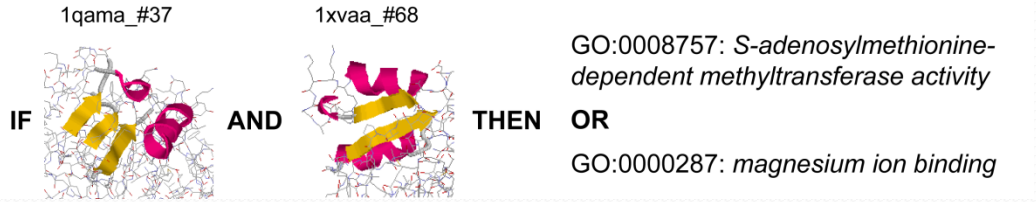
B. Descriptor group: sequence

DESCRIPTOR	SEGMENT 1	SEGMENT 2	SEGMENT 3	SEGMENT 4	SEGMENT 5
1qama_#37	35-40 FEIGSG	56-60 TAIEI	83-7 KDILQ	96-102 YKIFGNI	108-16 TDIIRKIVF
1g38a_#46	44-9 LEPACA	68-72 VGVEI	88-92 ADFLL	100-6 DLILGNP	144-52 GAFLEKAVR
1g55a_#9	7-12 LELYSG	31-5 AAIDV	55-9 KTIEG	71-7 DMILMSP	100-4 ----LHILD
1hdoa_#9	7-12 AIFGAT	31-5 TVLVR	53-7 GDVLQ	69-75 DAVIVLL	88-96 SEGARNIVA
1booa_#272	270-5 VDIFGG	291-5 ISFEM	33-7 GDSLE	48-54 SLVMTSP	77-85 LSFAKVVNK
1i9ga_#106	104-8 LEAGA-	128-32 ISYEQ	160-4 SDLAD	175-9 --AVLDM	183-91 WEVLDAVSR
1eg2a_#249	247-51 LDFFA-	268-72 ICTDA	45-9 CDCLD	60-4 QLIIC--	86-94 KRWLAEAER
1ek6a_#8	6-11 LVTGGA	30-4 VVIDN	65-9 MDILD	83-9 MAVIHFA	109-17 LTGTIQLLE
1bxka_#7	5-10 LITGGA	30-4 VVVDK	58-62 VDICD	78-82 --VMHLA	102-10 IVGTYTLLE
1qrra_#7	5-10 MVIGGD	29-33 CIVDN	74-8 GDICD	92-8 DSVVHFG	121-9 VIGTLNVLF
...	... ...	... ...	... ...	... ...	... ...

C. Molecular function annotations

Gene Ontology (GO:0003673)
- Molecular function (GO:0003674)
  - catalytic activity (GO:0003824)
    - oxidoreductase activity (GO:0016491)
      - oxidoreductase activity, acting on CH-OH group of donors (GO:0016614)
        - lactate dehydrogenase activity (GO:0004457)
          Matching proteins: 9 of 10, P-value: 2.14×10^{-14}
        - oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor (GO:0016616)
          Matching proteins: 25 of 59, P-value: 1.07×10^{-26}
    - transferase activity (GO:0016740)
      - transferase activity, transferring one-carbon groups (GO:0016741)
        - methyltransferase activity (GO:0008168)
          - S-adenosylmethionine-dependent methyltransferase activity (GO:0008757)
            Matching proteins: 14 of 24, P-value: 1.45×10^{-17}

D. Rule example



- Most descriptor groups contains significant over-representation of GO-class
- Combinations in terms of IF-THEN rules are needed to achieve accepteable prediction quality

The local descriptor match 68 proteins of which 14 are annotated with GO:0008757

The rule match 12 proteins annotated with GO:0008757 (two of which also are with GO:0000287)

Classification

IF ... THEN ...
IF ... THEN ...
IF ... THEN ...
IF ... THEN ...
IF ... THEN ...
IF ... THEN ...
IF ... THEN ...

IF 1gsa_2#218 AND 1ra9__#62 THEN GO:0016646 - *oxidoreductase activity*)

IF ... THEN
IF ... THEN ...
IF ... THEN ...

+7

Molecular Function	Votes
oxidoreductase activity	7
hydrolase activity	3
adenyl nucleotide binding	2
ligase activity	1
..	..

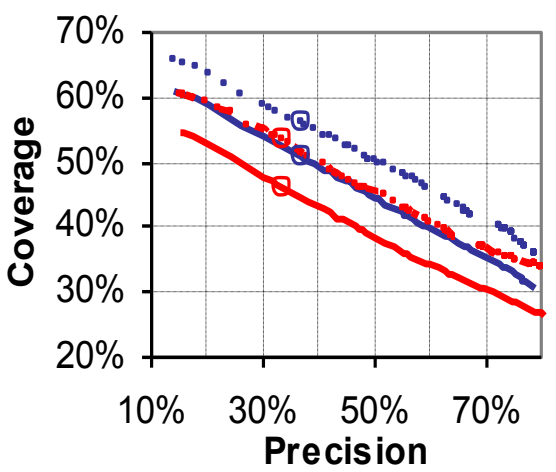
1. Calibrated E-values are calculated for each function according to the votes
2. A decision threshold is found for each function defining a prediction using ROC analysis

Cross validation results: summary

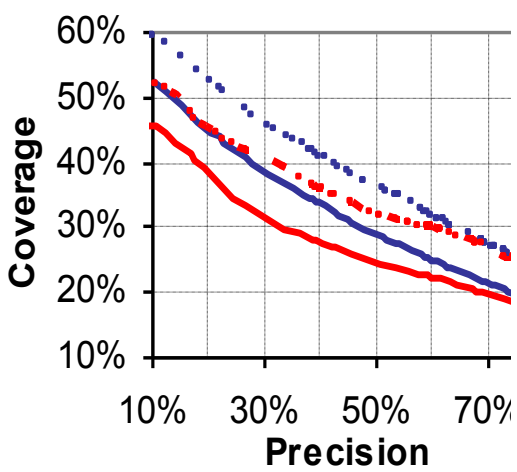
Given decision thresholds:

- *coverage for proteins* is the fraction of proteins with at least one correct prediction
- *coverage for annotations* is the fraction of annotations correctly predicted
- *precision* is the fraction of predictions that are correct.

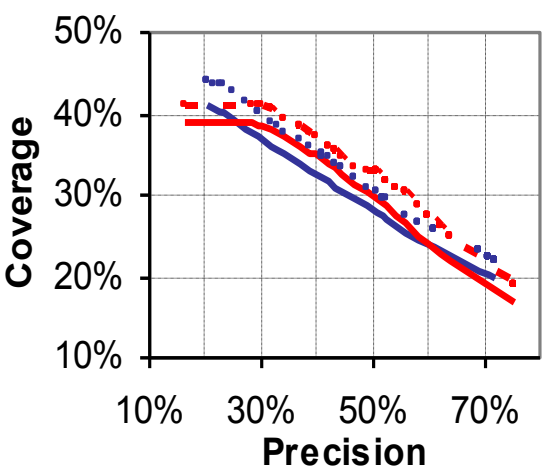
Molecular function

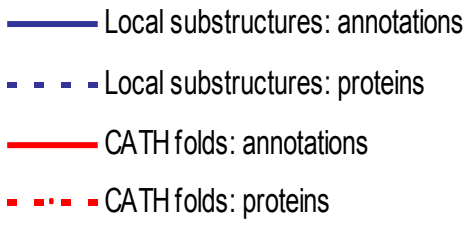


Biological process



Cellular compartment





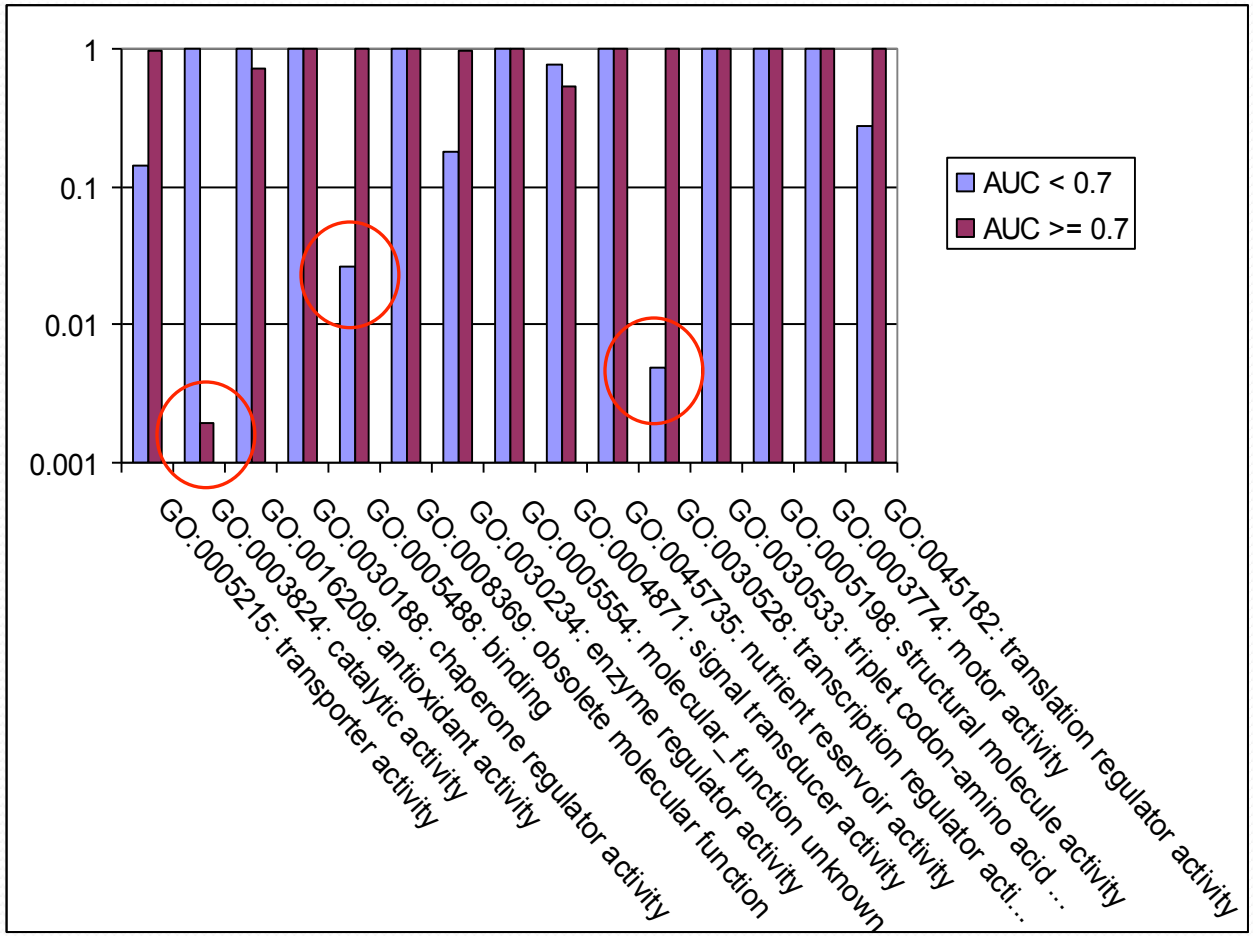
CATH fold: protein domains with similar orientation and connectivity of the main secondary structure elements

Functionally versatile folds

- 69% of 169 proteins with the Rossmann fold had one function correctly predicted by the local substructure method (precision=27%), compared to only 17% for CATH (precision=9%)
- Corresponding numbers for the 50 TIM barrel proteins were 66% (precision=21%) for local substructures and 50% (precision=12%) for CATH.
- Clearly, the use of local descriptors increases the resolution and allows us to functionally discriminate proteins with the same fold.

Cross validation analysis

- Overrepresentation of molecular functions that are easy/hard to predict in the most general level of GO.
- AUC >=0.7 is statistically unlikely ($p < 0.01$, computed by random reshuffling experiments)



Protein disorder

Significant correlation between prediction performance of molecular function class and the degree of disorder in proteins from these classes

- Correlation coefficient of –0.36 ($P<9.9\times10-5$)
- GO:0046983: dimerization activity
  (AUC = 0.69, disorder = 9.8%)
- GO:0005261: cation channel activity
  (AUC = 0.42, disorder = 8.1%)
- GO:0003713: transcription coactivator activity
  (AUC = 0.58, disorder = 8.1%).

External test set 1

- 429 protein structures (with 634 annotations)
- E-score > 0.05 for at least one domain to the training set (as determined by PSI-BLAST)
- Local substructures: Coverage for proteins: 53%, Coverage for annotations: 45%, Precision: 29%
- Local substructures + PSI-BLAST: Coverage for proteins: 76%, Coverage for annotations: 73%, Precision: 30%
  - PSI-BLAST: predictions derived from the annotations of the closest sequence-neighbor in the training set
  - 46 of all the correct predictions (to 44 proteins) were made exclusively by local descriptors

External test set 2

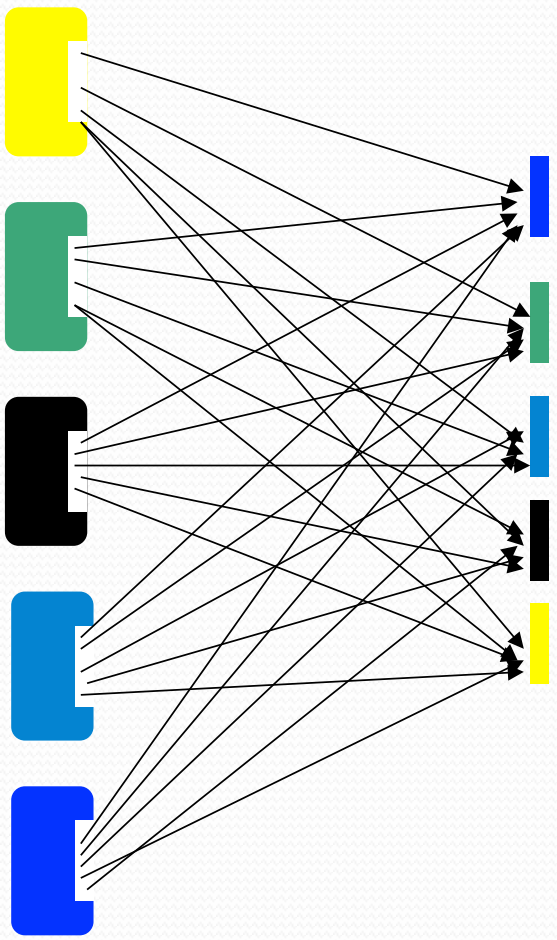
- 167 proteins (with 224 annotations)
- E-score > 0.05 for all domains
- Most annotations are based on sequence; hence a literature study was needed
- Local substructures: of 190 predictions to 93 proteins, 91 predictions to 57 proteins found some support in the literature
- CATH folds: not applicable because most proteins were not classified yet (partly manual assignments)
- Example: alanyl-tRNA synthetase
  - Four correct predictions (only two were annotated)
    - GO:0000049: *tRNA binding*
    - GO:0000287: *magnesium ion binding*
    - GO:0005524: *ATP binding*
    - GO:0004812: *tRNA ligase activity*
  - Fold not represented in the training set!

Generalized modeling of enzyme-ligand interactions

1. H. Strömbergsson, A. Kryshtafovych, P. Prusis, K. Fidelis, J. E. S. Wikberg, J. Komorowski and T. R. Hvidsten. Generalized modeling of enzyme-ligand interactions using proteochemometrics and local protein substructures, Proteins: Structure, Function and Bioinformatics 65: 568-579, 2006.
2. H. Strömbergsson, P. Daniluk, A. Kryshtafovych, K. Fidelis, J. E. S. Wikberg, G. J. Kleywegt and T. R. Hvidsten. Interaction Model Based on Local Protein Substructures Generalizes to the Entire Structural Enzyme-Ligand Space, Journal of Chemical Information and Modeling 48 (11): 2278–2288, 2008.

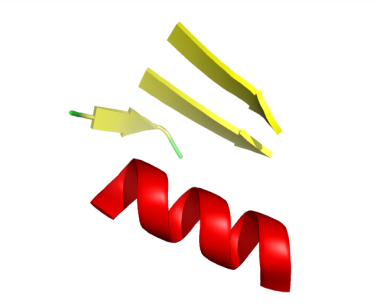
Proteochemometrics

- Many proteins – many ligands
- Experimental binding affinity for parts of the interaction space
- Model the experimental data using descriptors from proteins and ligands
- Use machine learning methods for modeling
- Use the model to predict the ”missing values” in the interaction space
- Generalizations to
  - interact/not interact-type data
  - protein-protein interactions
  - protein-DNA interactions



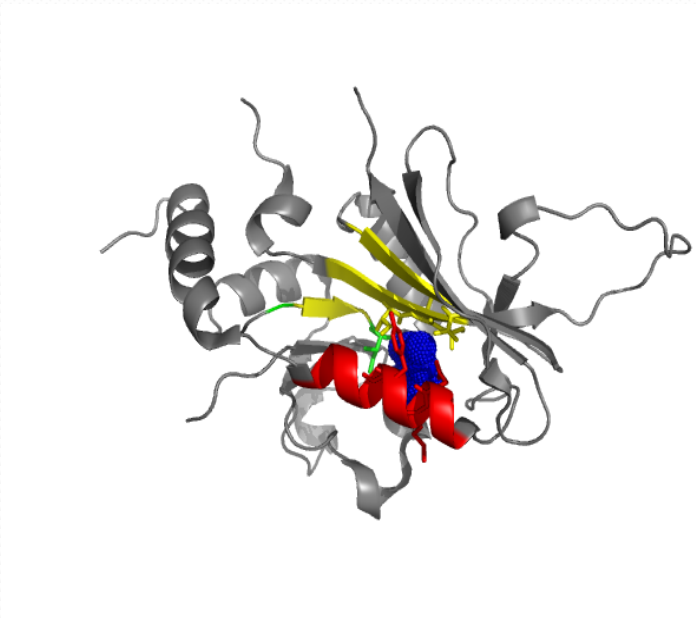
Local descriptors describe binding pockets

(A) Local descriptor



(B) Structural alignment

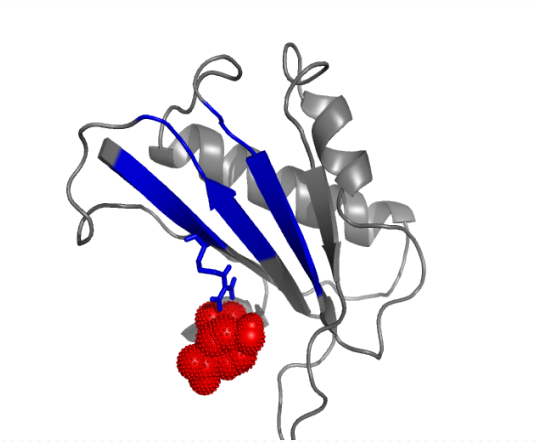
(C) Local descriptor in protein



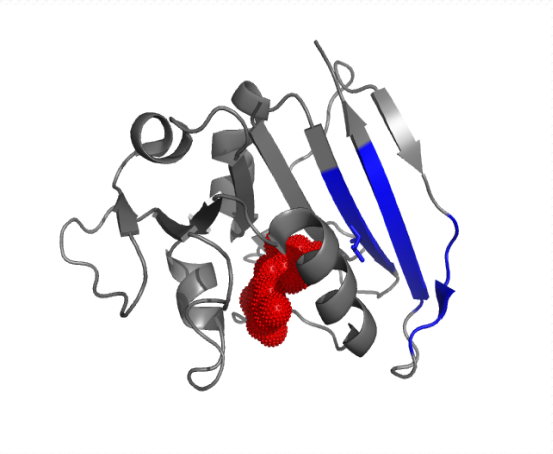
9 out 16 contacting amino acids in the protein is described by the local descriptor

… in different folds

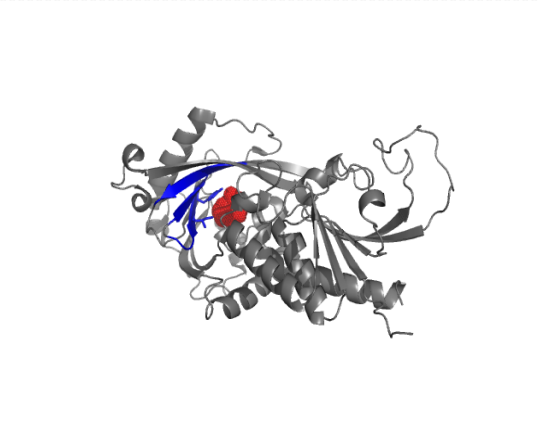
(A) 1com-PRE



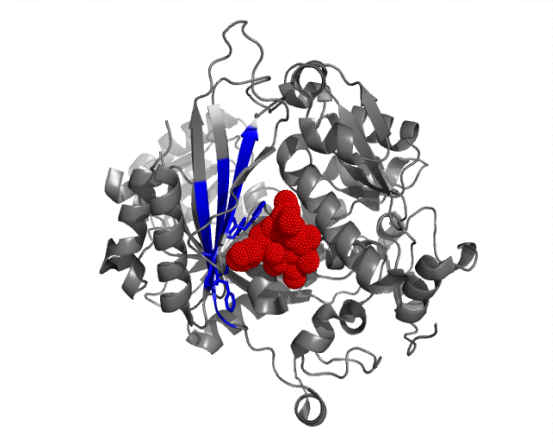
(B) 1dg5-TOP



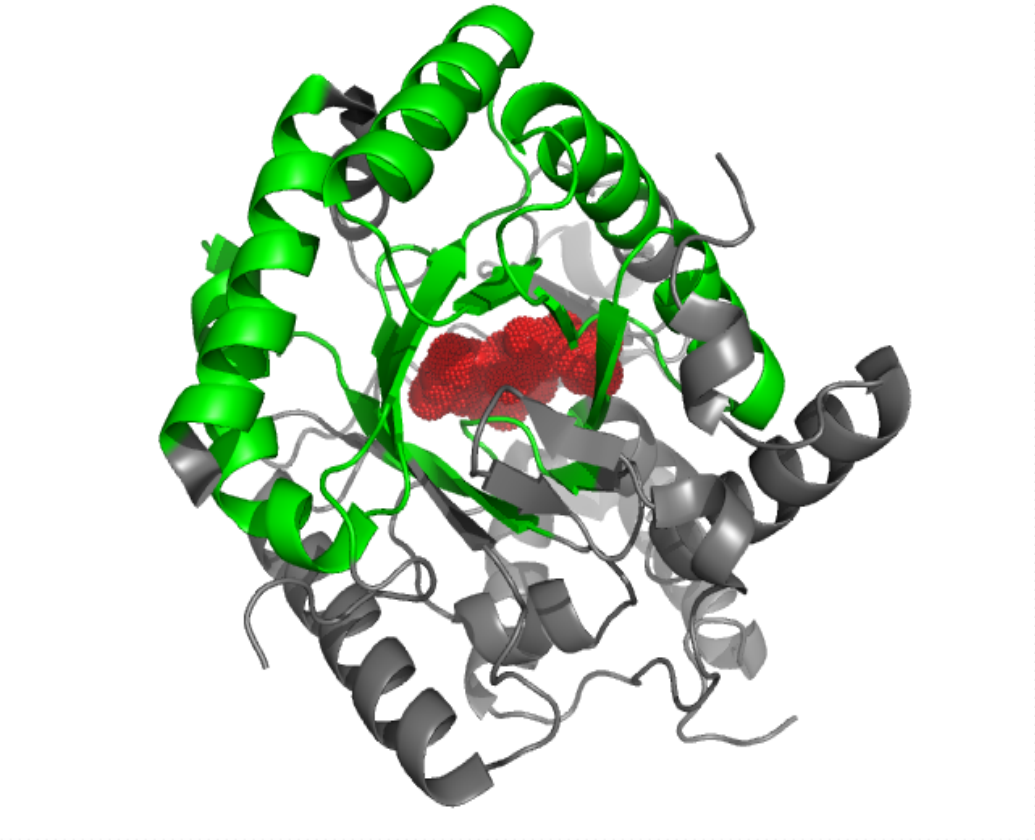
(C) 1coi-BE2



(D) 1w96-S1A



… all matching local descriptors

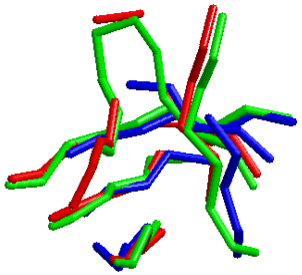


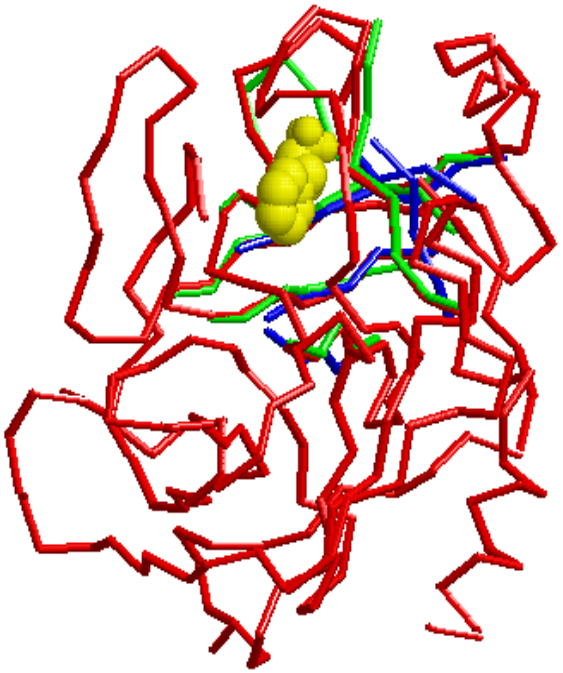
… describe ~75% of the contacts in the training set.

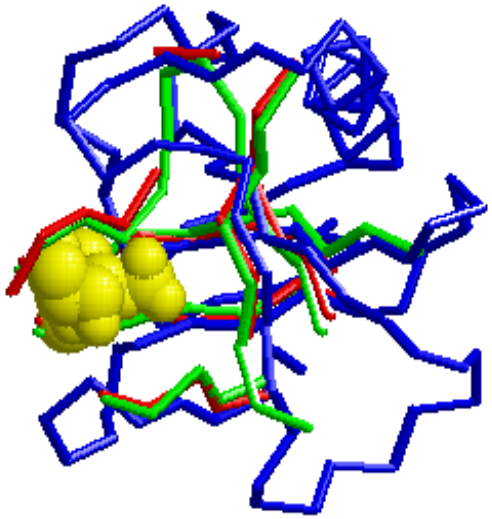
Study I

- Data
  - Training set: 104 hydrolases and lyases
  - Test set: ligands binding to the human form of cathepsin D, ligands binding to influenza virus A neuraminidase, ligands binding to influenza virus B neuraminidase and ligands binding to the human form of carbonic anhydrase II.
  - Local descriptor library: 4197 “general groups”

Alignment independent







Enzyme	Fold	Descriptor	Contacts
1tnj	b47	1tnj__#160	8
1tnl	**b47**	**1tnl__#160**	**9**
1tng	b47	1tng__#160	8
1tni	b47	1tni__#160	8
1tnh	b47	1tnh__#160	8
1tnk	b47	1tnk__#160	8
1fax	g3 b47	1fax_A#A160	8
1rnt	d1	1rnt__#77	11
1rgk	**d1**	**1rgk__#77**	**7**
6rnt	d1	6rnt__#77	7
1ghb	b47	1ghb_E#E160	10
1yyy	b47	1yyy_1#1160	8
1tpp	b47	1tpp__#160	7
1mtw	b47	1mtw__#160	7
1pph	b47	1pph_E#E160	7
3ptb	b47	3ptb__#160	8
1ppc	b47	1ppc_E#E160	7

Rule model

Rule 1: **IF** MW-1e4ea2#215([388.32, ∞)) **AND** MLOGP-1fh0a_#74([0.03, ∞))
THEN Binding(high)
Support: 9 ($p < 2.1E-4$)

Rule 2: **IF** RBN-1f97a1#62 ([0.50, 8.50)) **AND** MLOGP-1pfza_#156([-0.01, 0.05))
THEN Binding(low)
Support: 13 ($p < 5.4E-4$)

MW – molecular weigh
MLOGP – octanol-water partioning coef.
RBN - number of rotatable bonds

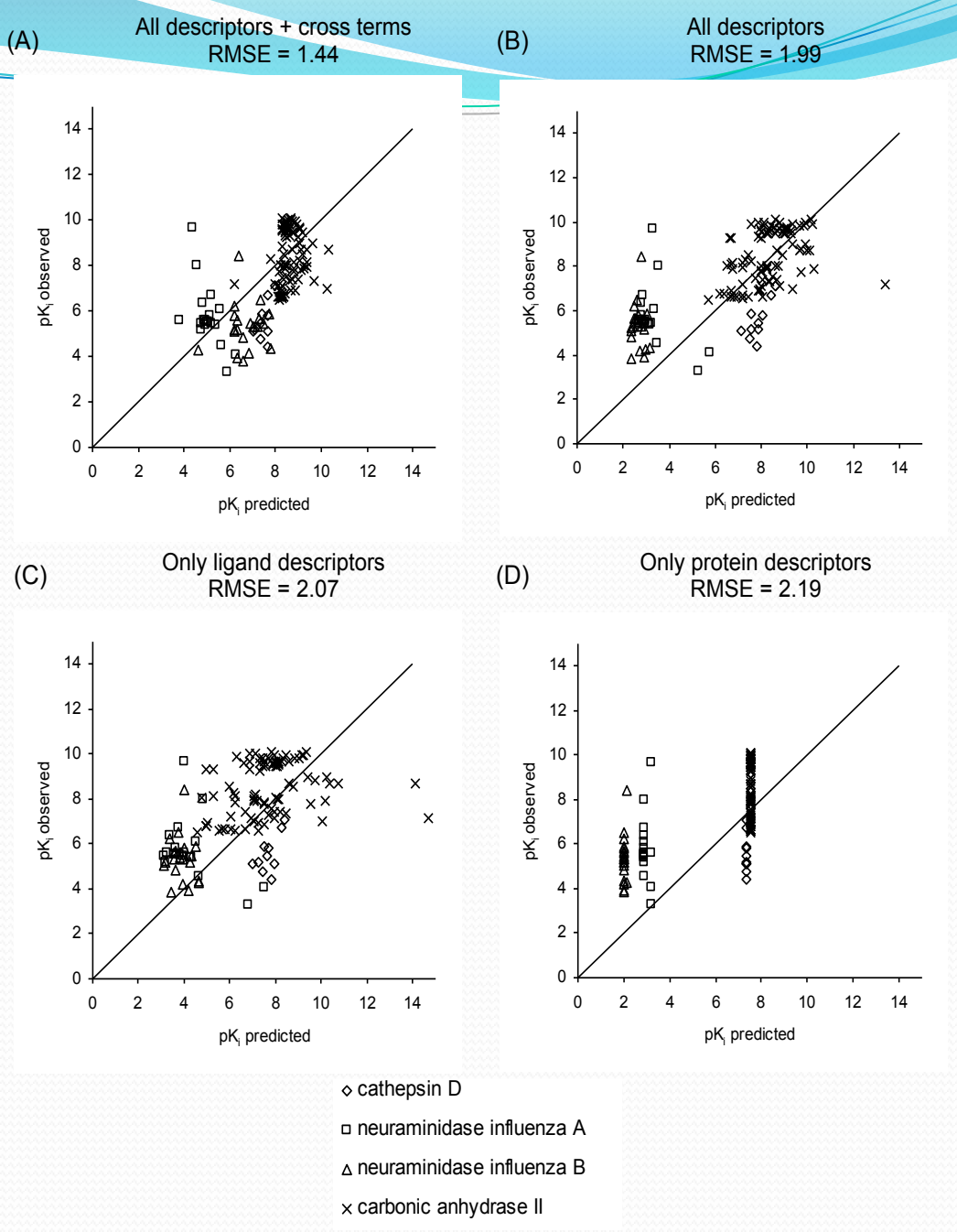
Analyzing the rules

- The rule model included 1914 rules of which 592 were significant ($p < 0.05$) for one of the classes;
  - 118 rules for "low binding"
  - 474 rules for "high binding"
- The rules predicting low binding had a significant overrepresentation of descriptors describing the absence of local substructures
- A conclusion from this analysis is that "low binding" is best modeled by observing which local substructures that are *not* present in the protein.

10-fold cross validation on training set + prediction of the external test set

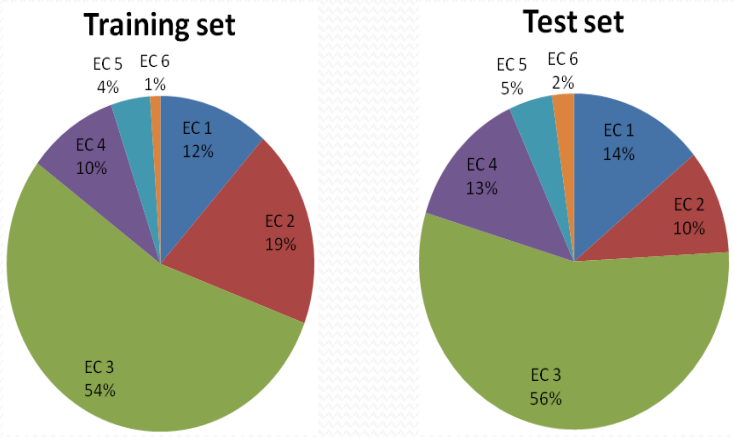
- Cross validation on training set: 77 correct predictions out of 97 predictions for the 104 protein-ligand complexes
  - precision=77/97=0.79
  - coverage=77/104=0.74
- For the 138 external test complexes, 84 out of 121 predictions were correct
  - precision=84/121=0.69
  - coverage=84/138=0.61

Partial least squares (PLS)

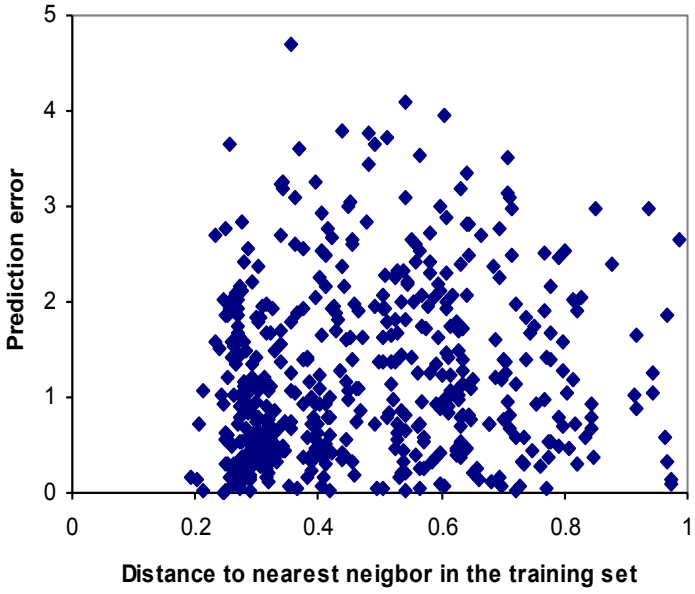


Study II

- Data:
  - 826 co-crystallized with drug-like ligands
  - Test set: 542 complexes (not cocrystalized, many series)
- Local descriptor library of 405 groups describing binding pockets

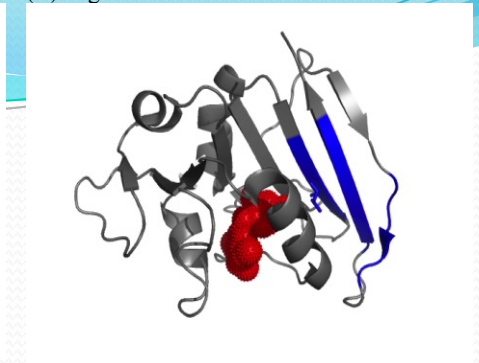


Training set

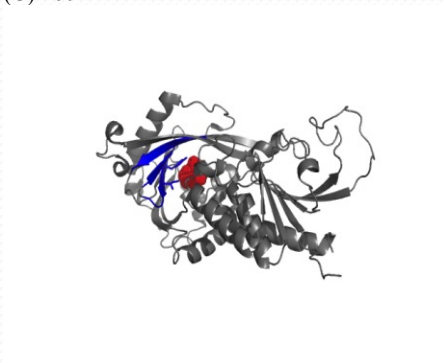


(A) 1com-PRE

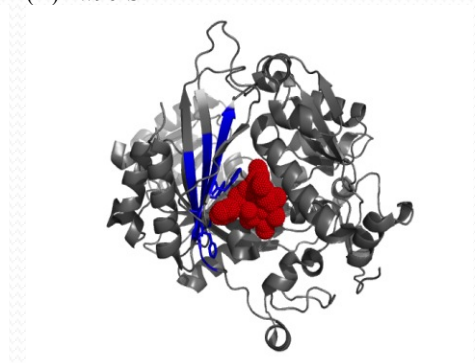
(B) 1dg5-TOP



(C) 1c0i-BE2



(D) 1w96-S1A



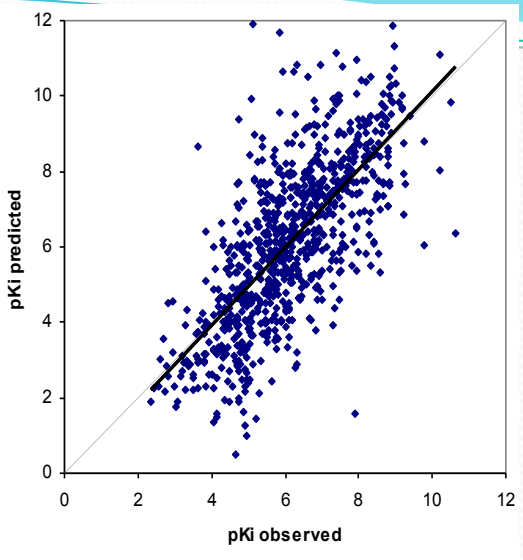
(E) Local sequence alignment of all matches in the training set

Descriptor	Segment1		Segment2		Segment3		Lig	SCOP fold/EC family
1coma #94	3-7	..IRGIR	42-48	VVQMLLS	91-96	VMMTVQ	PRE	Bacillus chorismate mutase-like/5.4.99.5
1dg5a #153	111-115	..EVTEV	141-145	..EWRFS	150-155	RYRLYS	TOP	Dihydrofolate reductases/1.5.1.3
1dy4a #417	88-94	GNSLSIG	130-134	DVDVS..	414-419	AKVTFS		Concanavalin A-like lectins/glucanases/3.2.1.91
1c0ia2#1233	1199-1205	GQTVLVK	1223-1229	KIIPRPG	1231-1235	.EVICG	BE2	FAD-linked reductases, C-terminal domain/1.4.3.3; 1.4.3.4; 1.14.13.2; 1.5.3.1;
1c01a2#1233	1199-1205	GQTVLVK	1223-1229	KIIPRPG	1231-1235	.EVICG	FLA	
1el5a2#267	223-227	..QVVGF	252-257	IYMGFP.	264-269	LKLGYH	DMG	
1el9a2#267	223-227	..QVVGF	252-257	IYMGFP.	264-269	LKLGYH	MTG	
1elia2#267	223-227	..QVVGF	252-257	IYMGFP.	264-269	LKLGYH	PYC	
1ojaa2#343	292-298	GSVIKCI	326-330	..YTLDD	340-345	IMGFIL	ISN	
1phh 2#222	183-189	FGHLGLL	208-214	FAICSCR	219-224	SHYYVQ	DHB	
1qwua2#932	524-528	..FTLDD	554-560	HVVMHNT	929-934	LDKFIF		Supersandwich/3.2.1.114
1w96a1#508	454-460	CTACRIT	487-491	MGYFS..	505-510	QFGHIF	S1A	Barrel-sandwich hybrid/6.4.1.2
1w96a1#510	452-458	GHCTACR	487-491	..MGYFS	507-512	GHIFAF	S1A	
2a06e#194	71-77	MSKIEIK	184-190	SYEFTSD	192-196	.MVIVG		-/1.10.2.2
2web #318	177-183	GSLTYTG	307-311	YVVFD..	315-320	PQLGFA		Acid proteases/3.4.23.20
2wec #318	177-183	GSLTYTG	307-311	YVVFD..	315-320	PQLGFA		
2wed #318	177-183	GSLTYTG	307-311	YVVFD..	315-320	PQLGFA		

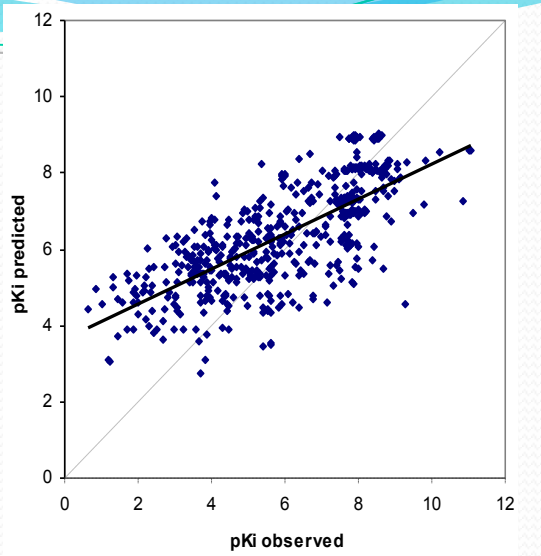
Prediction

- Support vector machine regression
- Proteins varying greatly in terms of sequence and structure
- All enzyme structures with a experimentally measured binding affinity to a ligand: cover all enzymes classes

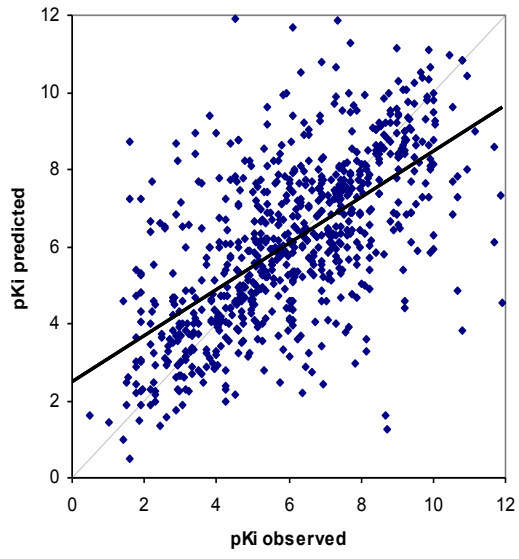
(A) Support vector machine (SVM) predictions, training set (r^2: 0.51, RMSEP: 1.5)



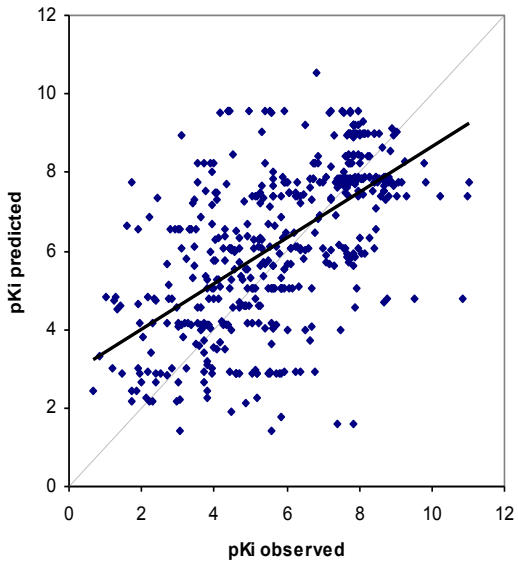
(B) SVM predictions, test set (r^2: 0.53, RMSEP: 1.5)



(C) Control: Nearest neighbor (NN) prediction, training set (r^2: 0.41, RMSEP: 1.9)



(D) Control: NN prediction, test set (r^2: 0.35, RMSEP: 1.9)



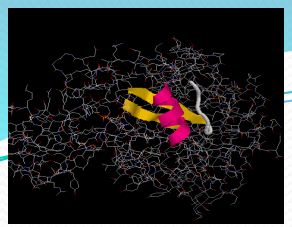
Summary

- Local descriptors of protein structure represent a novel approach to view protein structure and have found application
- Articles and student theses:
  http://www.trhvidsten.com/LocalDescriptors.html

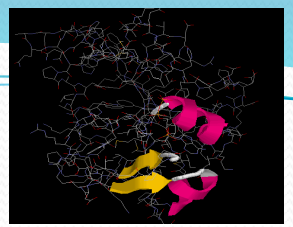
Some existing tools

- Contact prediction:
  http://predictioncenter.org/Services/FragHMMent/
- Structure alignment:
  http://www.bioexploratorium.pl/EP/DSA
- Near future: A common resource for structure analysis using local descriptors

IF



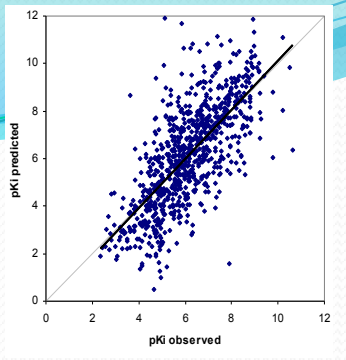
AND



THEN GO:0016646

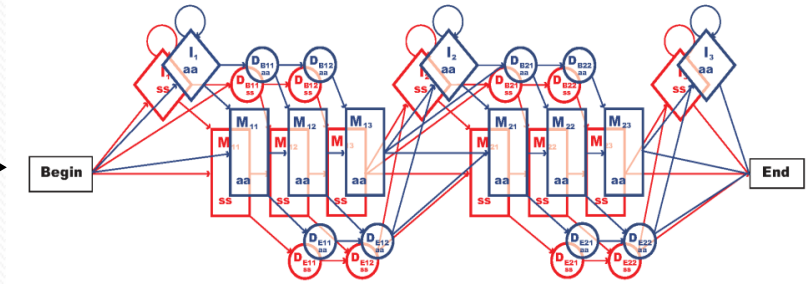
Prediction of function

Prediction of molecular interaction

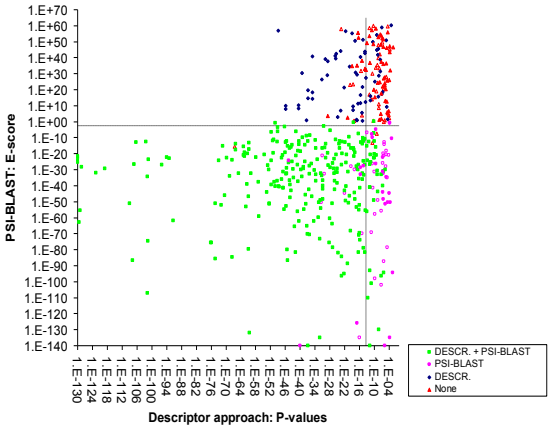


Structure prediction

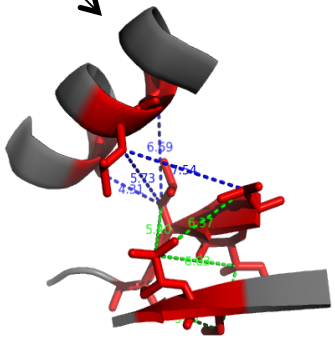
Sequence



Fold recognition

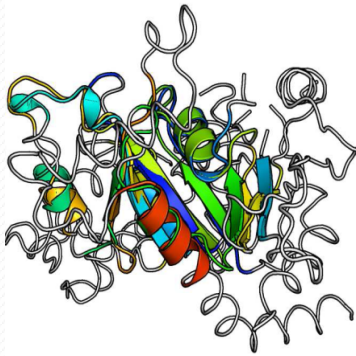


Contact prediction



1. Fold classification
2. Ab inito structure prediction
3. Domain prediction

Structure alignment



Acknowledgments

1. H. Strömbergsson
2. A. Kryshtafovych
3. P. Prusis
4. K. Fidelis
5. J. E. S. Wikberg
6. J. Komorowski
7. P. Daniluk
8. A. Lægreid,
9. G. Andersson
10. P. Björkholm
11. R. Andersson