

Workshop on the definition of protein domains
and their likelihood of crystallization

University of Vienna

28th-30th June 2006

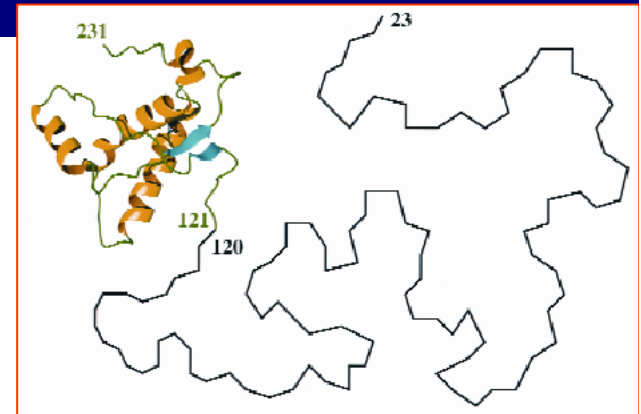
Prediction of domain boundaries and unfolded regions in protein chain

Oxana V. Galzitskaya

Nikita V. Dovidchenko

Sergiy O. Garbuzynskiy

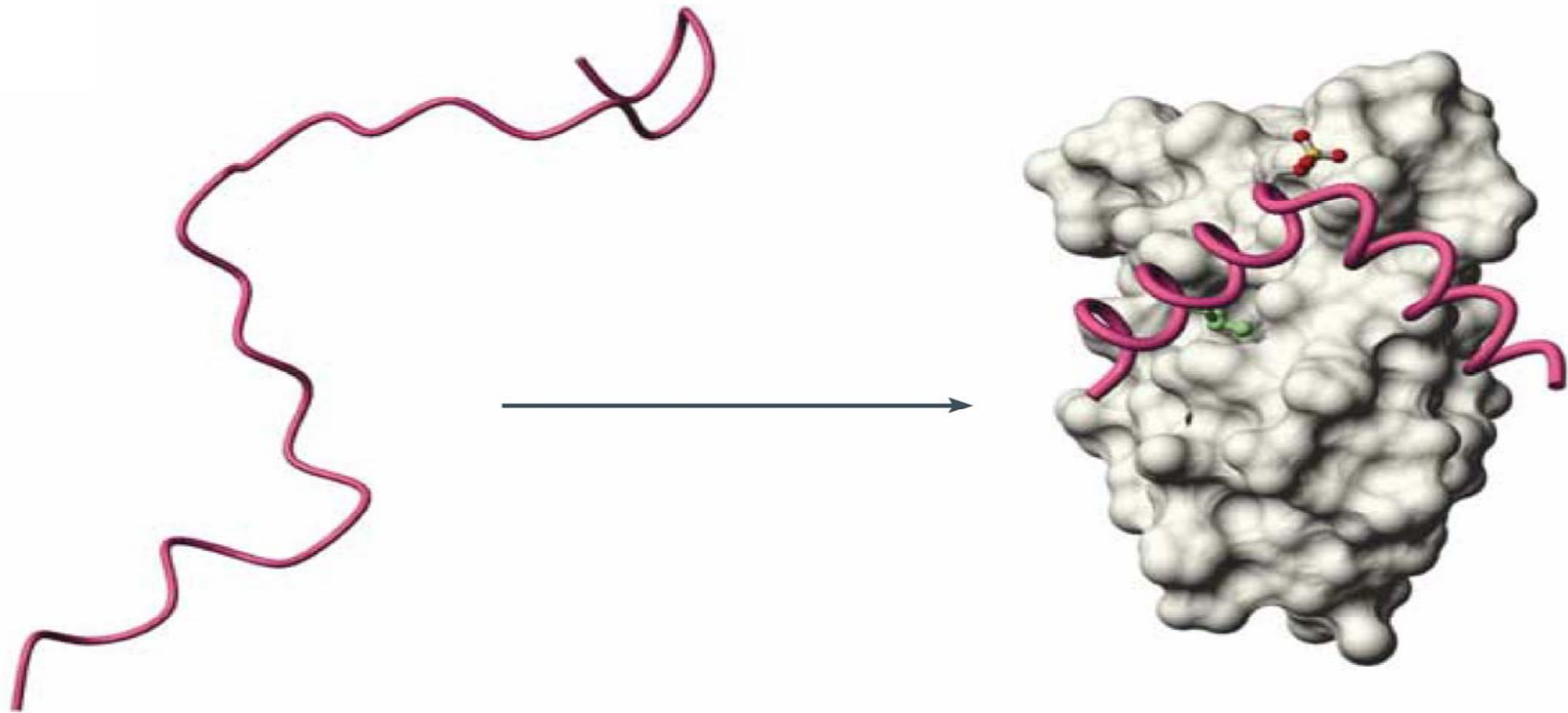
Michail Yu. Lobanov



Institute of Protein Research, Russian Academy of Sciences, 142290,
Pushchino, Moscow Region, Russia

pKID domain of the transcription factor CREB

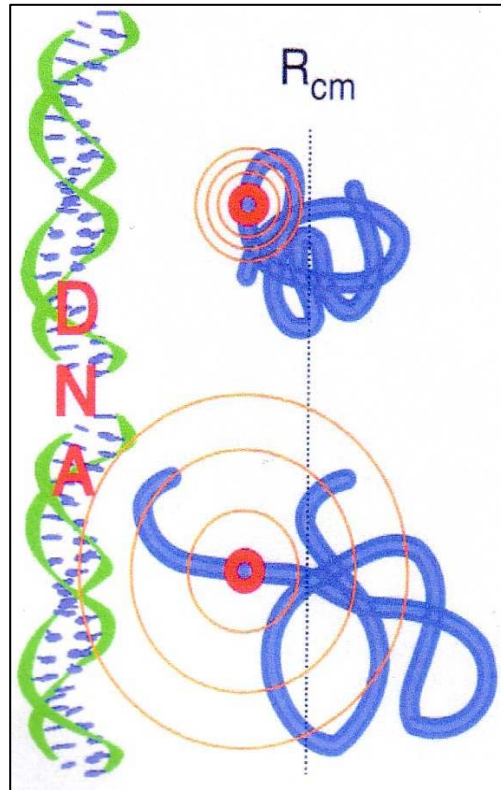
(adopted from *Dyson & Wright, 2005*)



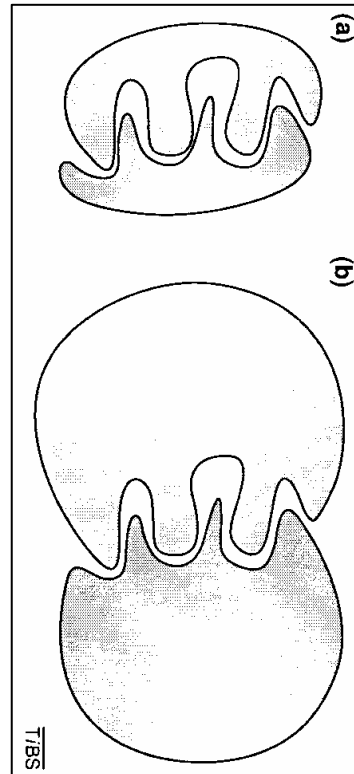
Intrinsically Unstructured Proteins are proteins or domains that, in their native state, are either completely disordered or contain large disordered regions. More than 400 such proteins are known, including Prions, proteins Tau, Bcl-2, p53, 4E-BP1, eIF1A, etc. (Tompa, 2002, Uversky, 2002).

Why does eukaryotic cell prefer unstructured proteins?

Acceleration of molecular recognition




Large area of the interface under smaller sizes



One protein – several functions

Conformation of protein is determined by partner of interaction, but not amino acid sequence itself, as it is typical for globular proteins.



Protein disorder is important for:

Understanding protein function
Protein folding pathways
De novo design of proteins

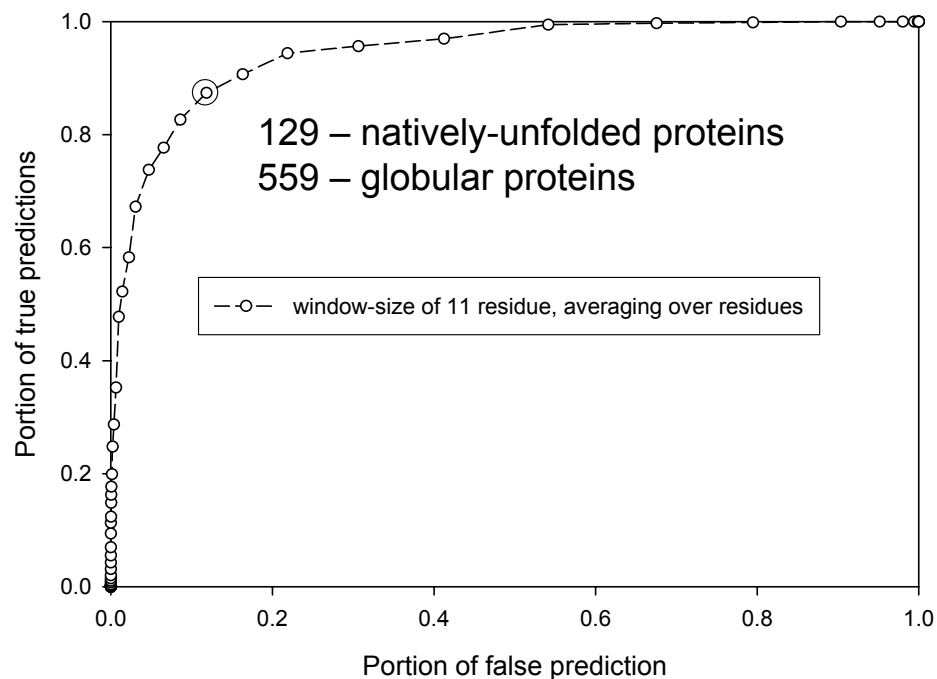
- **Attempts to predict disorder:**
- Finding regions with low complexity, like SEG (Wootton, 1994) and CAST (Promponas et al., 2000).
- Finding regions without regular secondary structure, like NORSp (Liu et al., 2002).
- Combination of low overall hydrophobicity and a large net charge, Uversky et al., 2000 .

Our method assigns ordered/disorder status to residues on the basis of their ability to form sufficient number of contacts in globular state .

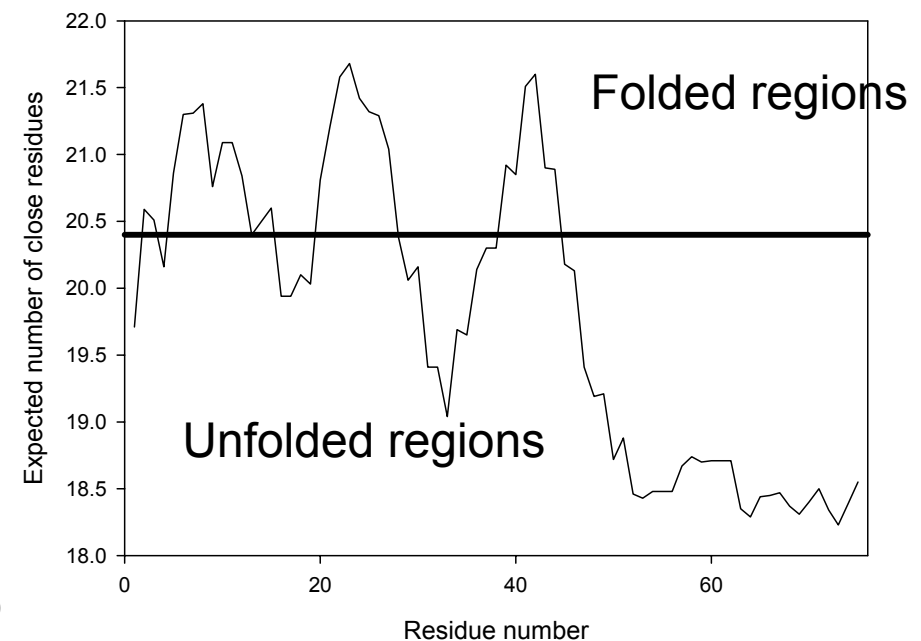
Average number of close amino acid residues (located within a given distance 8Å) to the given residue in the globular proteins

G	P	A	D	E	K	S	N	Q	T
17.11±0.2	17.43±0.03	19.89±0.02	17.41±0.03	17.46±0.02	17.67±0.02	18.19±0.03	18.49±0.03	19.23±0.04	19.81±0.03
R	H	C	V	M	L	I	Y	F	W
21.03±0.3	21.72±0.05	23.52±0.05	23.93±0.03	24.82±0.06	25.36±0.02	25.71±0.03	25.93±0.04	27.18±0.04	28.48±0.07

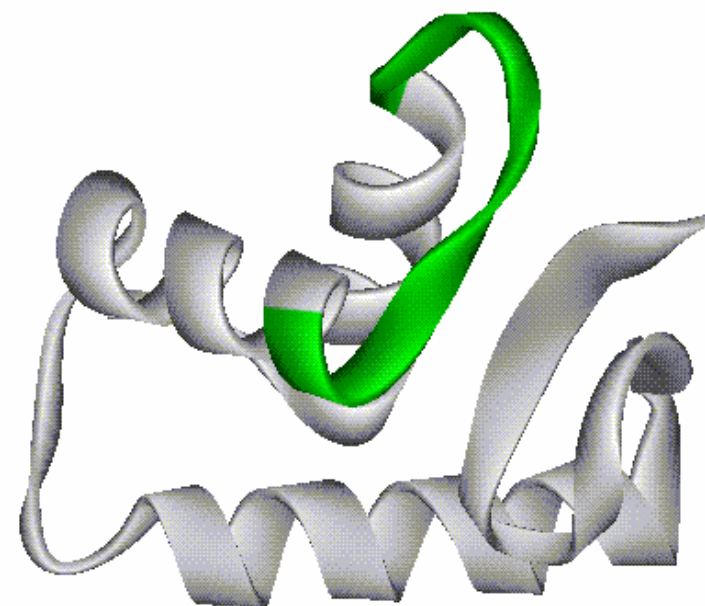
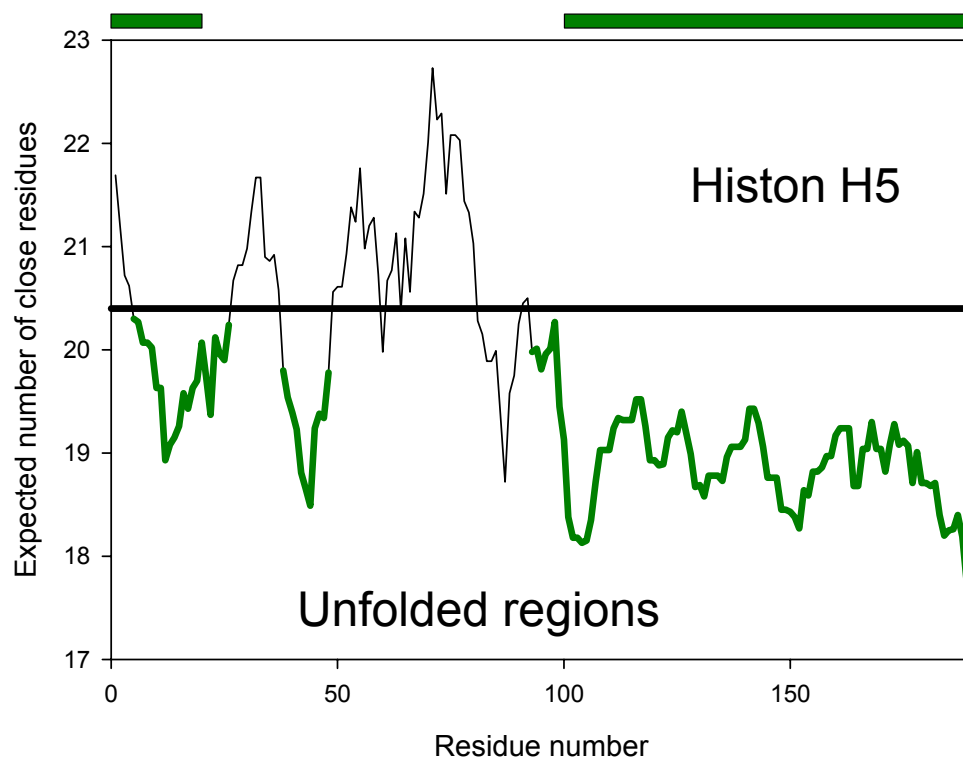
Receiver operator characteristic curve



Profiles of the expected number of close residues



Profiles of the expected number of close residues



1hst.ent

Comparison of performance of disorder prediction methods for 129 natively – unfolded and 559 globular proteins. Averaging is done over residues. Data about other methods are given from Dosztanyi et al., 2005 J. Mol. Biol., 347, 827.

Method	Portion of true predictions (averaging over residues)	Portion of false predictions (averaging over residues)
Fold Unfold (Galzitskaya et al., 2006, Molecular Biology, 40, 341)	0.85	0.05
IUPred (Dosztanyi <i>et al.</i>, 2005, J. Mol. Biol., 347, 827)	0.76	0.05
PONDRVL3H (Obradovic <i>et al.</i>, 2003, Proteins, 53, 566)	0.66	0.05
DISOPRED (Ward <i>et al.</i>, 2004, Bioinformatics, 20, 2138)	0.66	0.05
GlobPlot (Linding <i>et al.</i>, 2003, Structure, 11, 1453)	0.33	0.18

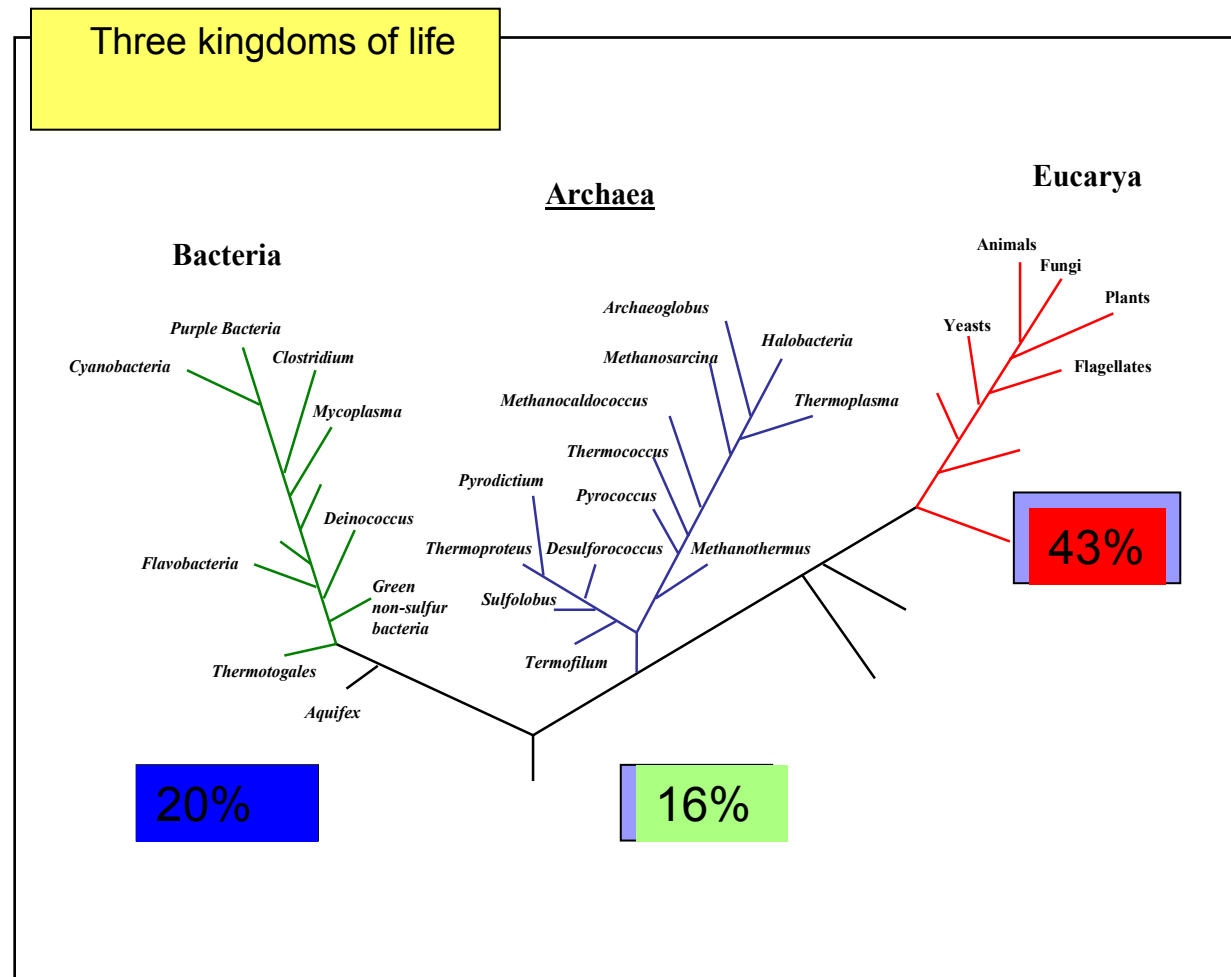
- What is the percentage of totally unstructured proteins in various proteomes?

19 archaean proteomes

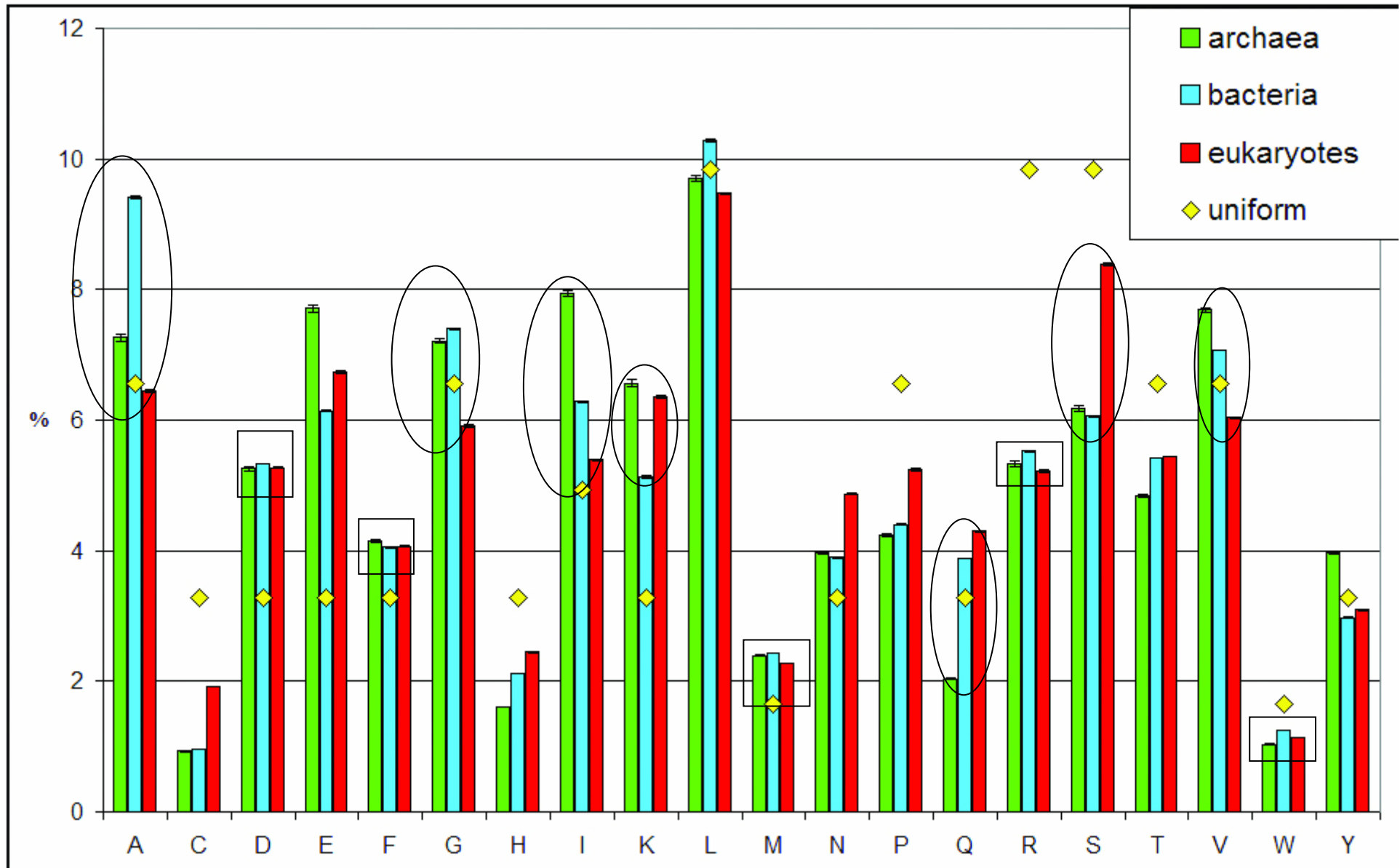
159 bacterial proteomes

17 eukaryotic proteomes

- According to our estimates, 12%, 3% and 2% of the proteins in eukaryotic, bacterial and archaean proteomes are totally disordered, and long (>41 residues) disordered segments are found to occur in 16% of archaean, 20% of eubacterial and 43% of eukaryotic proteins for 19 archaean, 159 bacterial and 17 eukaryotic proteomes, respectively.

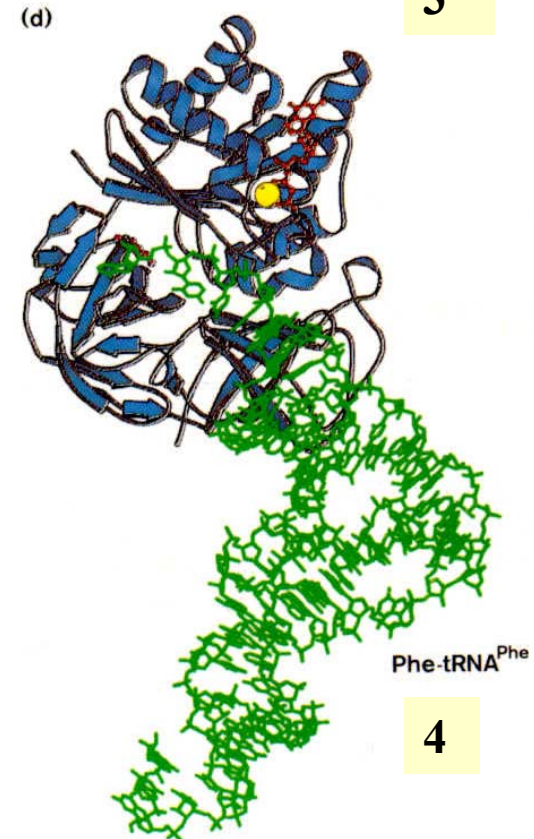
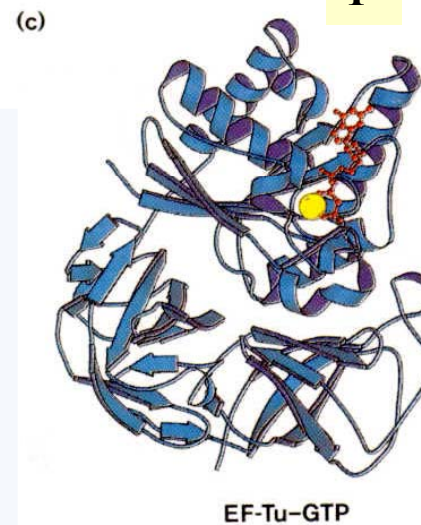
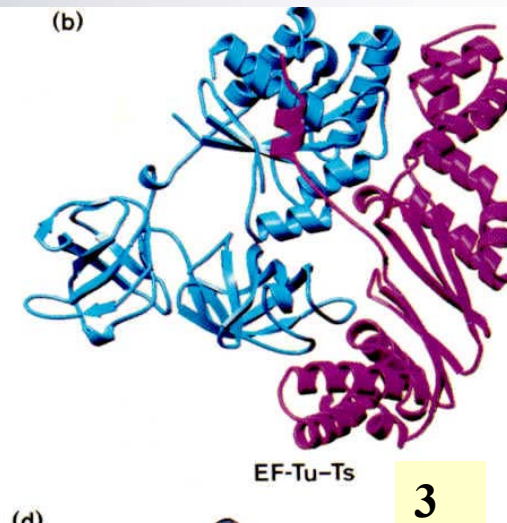
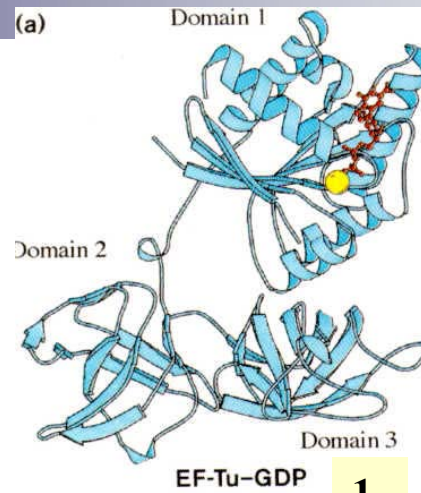


Frequencies of amino acids in proteins for three main taxa
 calculated from the known proteomes : 19 archaea, 159 bacteria, 17 eukaryotes



Crystal structures of procaryotic translation elongation factors in isolated state: with GDP (1), with GTP (2) in complex with ligands EF1A-EF1B (3) and EF1A-tRNA (4).

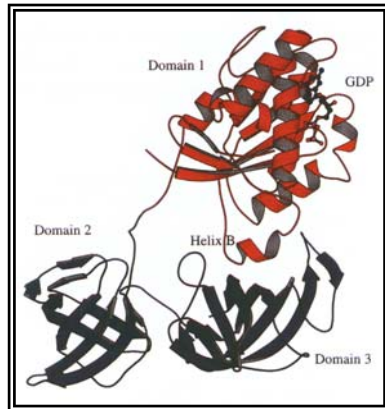
Crystal structure of yeast elongation factor eEF1A in complex with EF1B α (5)



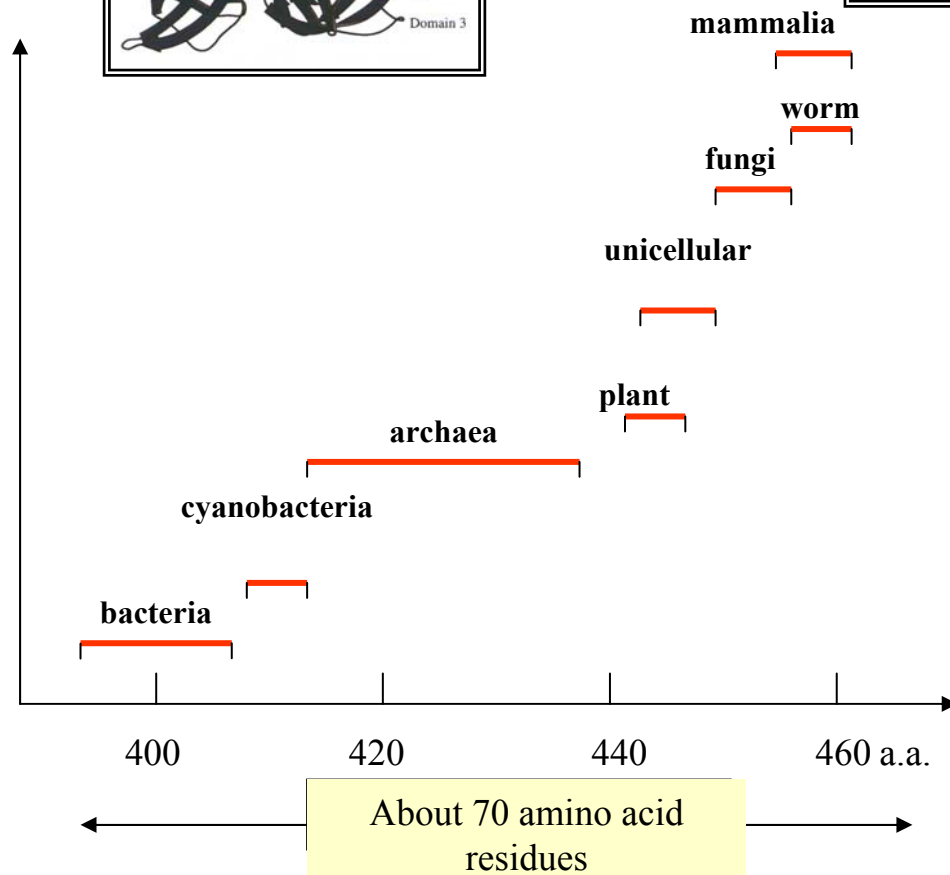
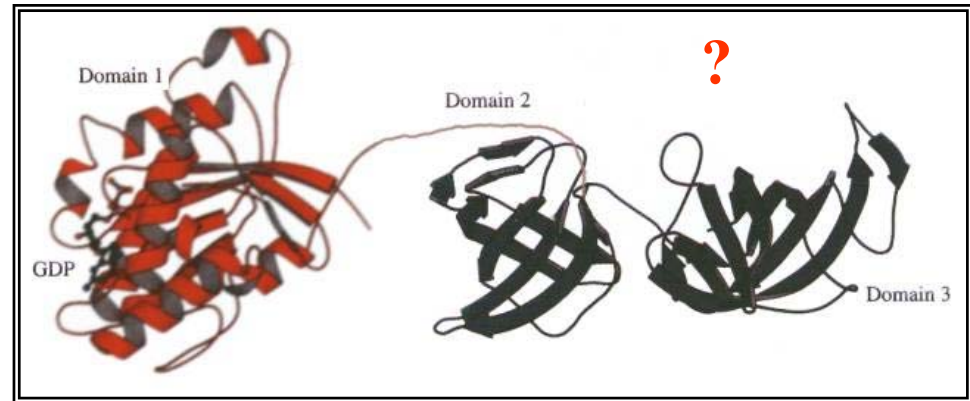
5



EF1A from *S. solfataricus* with GDP



Working model EF1A from *O. cuniculus*

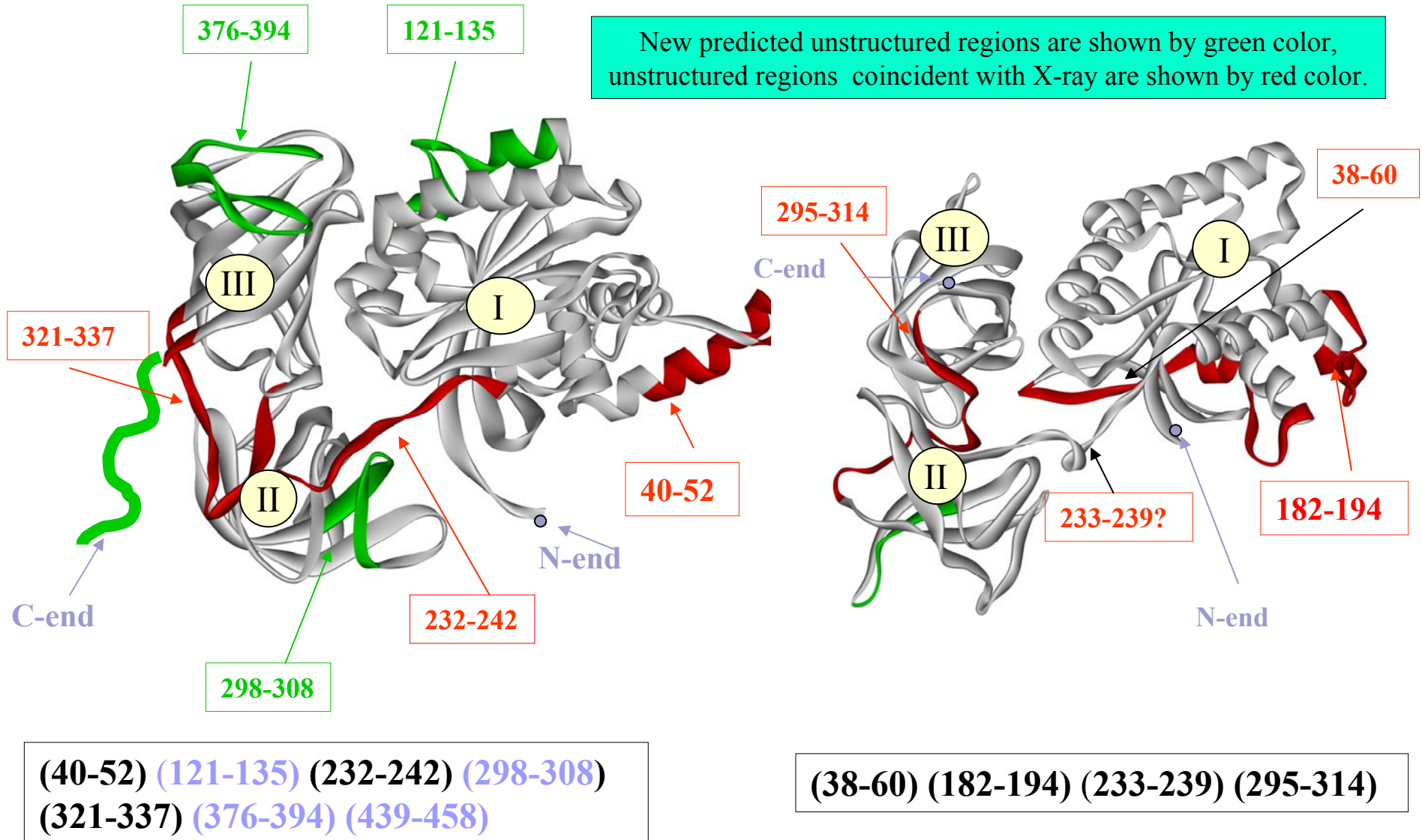


1. What is amino acid composition of prokaryotic and eukaryotic elongation factors?
2. Are there specific amino acid residues in eukaryotic elongation factors?
3. What is the distribution of additional 70 amino acid residues on the length under transition from prokaryotic to eukaryotic factors?

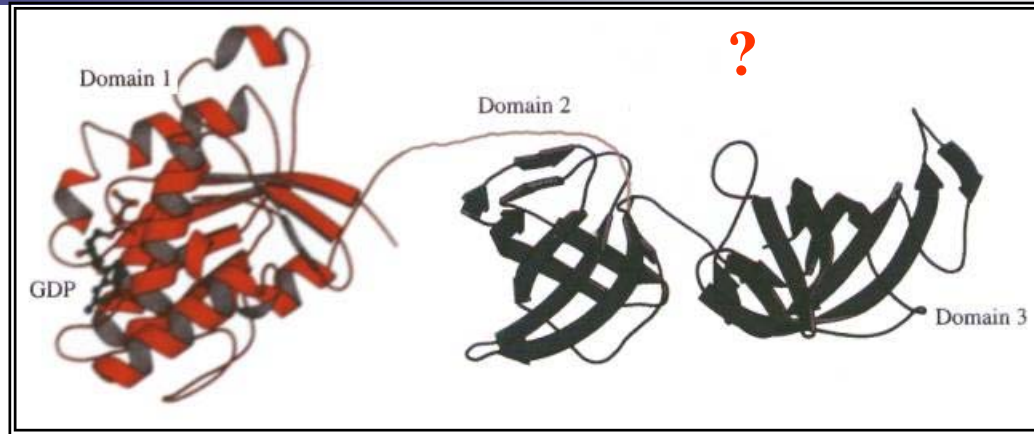
% inclusion of separate amino acids in elongation factors from different organisms

Amino acid	Un Structured	Globular	% inclusion of separate amino acids in elongation factors from different organisms										
	%	%	<i>Pro</i>	<i>Pro</i>	<i>Pro</i>	<i>Pro</i>	<i>Pro</i>	<i>Pro</i>	<i>Eu</i>	<i>Eu</i>	<i>Eu</i>	<i>Eu</i>	<i>Eu</i>
			391	393	394	398	405	405	453	458	458	462	462
Ala A	7,15	8,15	8,7	7,12	7,61	7,29	6,91	6,91	8,61	8,08	8,08	8,44	8,44
Arg R	4,21	4,61	5,63	5,85	5,08	5,28	6,67	6,17	3,97	3,49	3,93	3,68	3,68
Asn N	2,06	4,66	2,05	1,78	2,54	2,76	2,72	2,72	3,09	3,49	3,49	3,25	3,46
Asp D	5,05	5,78	6,14	6,11	7,61	6,78	5,93	5,93	5,08	5,46	5,24	5,63	5,63
Cys C	0,61	1,64	0,77	0,76	0,25	1,76	0,25	0,25	1,1	1,31	1,53	1,3	1,3
Glu E	4,46	3,69	8,44	9,41	9,39	8,29	9,14	9,38	7,51	6,77	6,77	5,84	5,84
Gln Q	14,26	5,98	2,56	2,04	2,28	1,76	2,22	1,98	2,43	1,97	2,62	2,16	2,16
Gly G	4,31	7,99	9,21	10,18	9,14	10,3	9,63	9,63	8,39	9,39	9,17	9,52	9,52
His H	1,51	2,33	2,56	2,8	3,05	3,02	2,96	2,96	2,65	2,4	2,4	2,16	2,16
Ile I	3,67	5,43	7,67	7,38	6,35	5,28	5,19	5,68	7,28	6,55	6,55	6,49	6,49
Leu L	5,44	8,37	6,14	7,12	6,35	6,28	6,67	6,91	4,42	5,02	5,24	5,63	5,63
Lys K	10,43	6,05	5,37	5,85	4,57	6,28	4,94	5,68	9,49	11,1	10,7	10,1	10,17
Met M	1,3	2,03	2,81	2,54	3,55	2,76	2,72	2,72	3,09	1,97	1,75	2,6	2,6
Phe F	1,66	3,95	3,84	3,56	3,05	2,76	2,96	2,96	3,09	3,28	3,71	3,03	3,03
Pro P	12,07	4,61	5,12	5,09	4,57	5,03	5,68	5,43	5,52	5,02	5,02	5,41	5,41
Ser S	6,91	6,31	3,58	2,54	4,06	4,02	1,73	1,48	4,86	4,37	4,59	5,19	4,98
Thr T	5,14	6,15	7,42	7,63	7,61	7,29	7,65	7,9	5,96	6,55	6,11	6,28	6,28
Trp W	0,32	1,55	0	0,25	0	0	0,49	0,49	0,88	1,31	1,31	1,08	1,08
Tyr Y	1,42	3,64	2,3	2,54	2,79	2,76	2,72	2,72	2,87	2,18	1,75	2,6	2,6
Val V	8,02	7,00	9,72	9,41	10,15	10,3	12,8	12,1	9,71	10,2	10,0	9,52	9,52

3D structures of eucaryotic translational elongation factors 1A from *S. cerevisiae* (in complex with fragment EF-1B, not shown, from the left) and procaryotic factor from *T. aquaticus* (from the right)



	Procaryote	Eukaryote				Archaea
Loop A (121 - 134)	Absent for all	Present in all, except lower plant GEFEAGISKNGQTR				Only for sulphuric bacteria KGEYEAGMSAEG
Loop B (155 - 166)	Absent for all	Metazoa			unicellular	Only for hyperthermophile MDATEPPFSEK
		mammalia	Higher plant	higher fungi		
		MDSTEPPYS	Absent for all			
Loop C (182 - 195)	thermophile MHKNPKTKRGEN cyanobacteria TENPETKPGDNK	Absent for all				Only for acid-loving bacteria APDGDNVTH
Loop D (290 - 310)	Absent for all	Present in all, except lower plant PGDNVGF				Present in all PGDNIGF
Loop E (364-381)	Absent for all	DRRSGK or DRRTGK				Present in all sulphuric bacteria DPRTGQEAENPQ hyperthermophile LNPKDGTTLKDK
Loop F (C- end)	Absent for all	Present in all, except lower plant				Absent for all

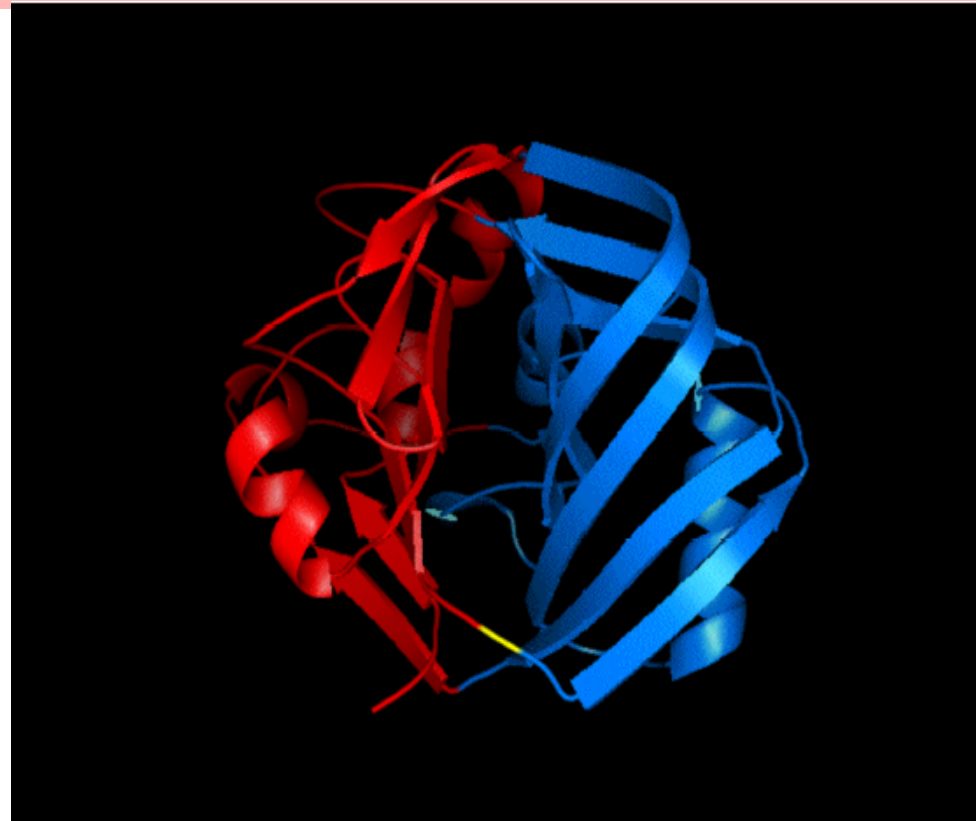


- The absence of a rigid structure and the possibility of large conformational change upon interaction with the partner molecule might explain the well-known ability of the eukaryotic factors to interact with different ligands
- (actin, tubulin, calmodulin and many others) besides the translational components.

Prediction of number and position of domain boundaries in multi-domain proteins using only amino acid sequence alone

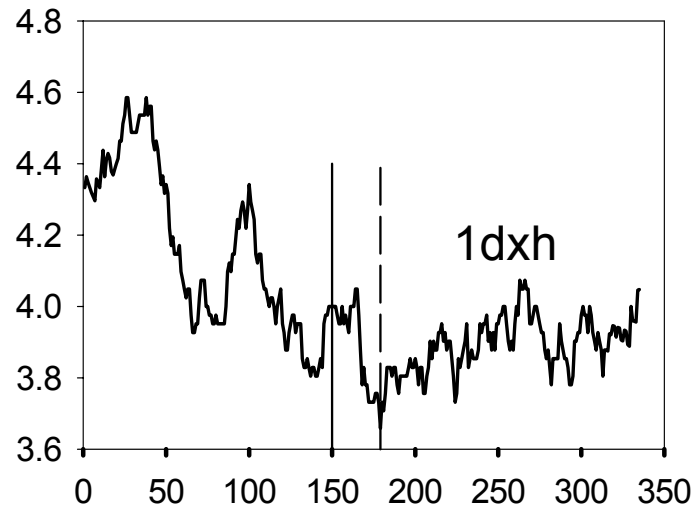
To predict 3D structure of a protein it is necessary to predict the boundaries of its domains.

The domain is defined as a compact structural unit that is capable of independent folding.



CASP6 (T0241)

Prediction of protein domain boundaries from sequence alone

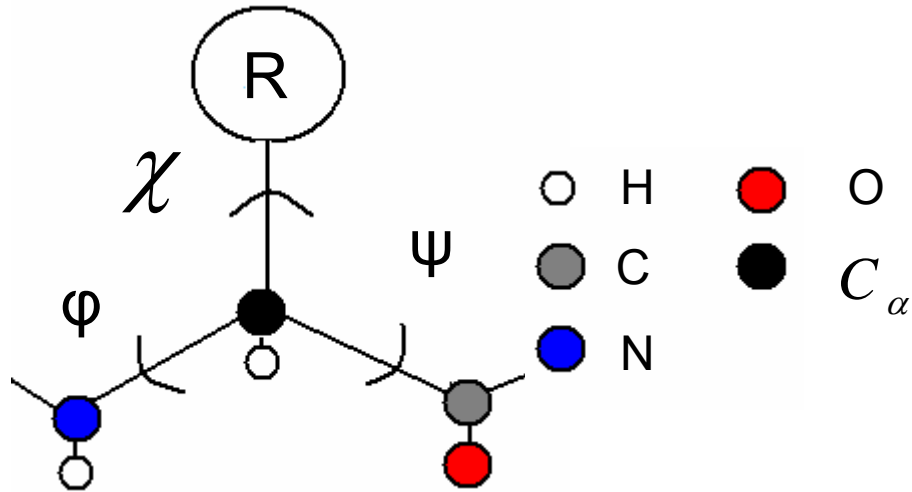
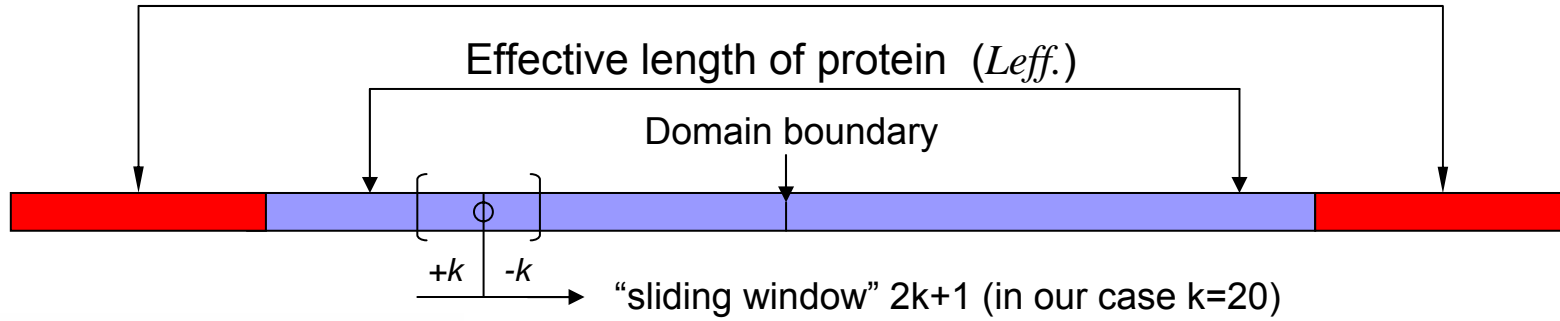


Our method is based on the hypothesis that a high-side chain entropy of a region in a protein chain must be compensated by a high-residue interaction energy within the region, which could correlate with a well-structured part of the globule, that is, with a domain unit.

For protein domains, this means that the domain boundary is conditioned by amino acid residues with a small value of side chain entropy.

Construction of an entropy profile

We do not consider domains of less than 50 residues



a	A	E	Q	D	N	L	G	K	S	V	R	T	P	I	M	F	Y	C	W	H
a																				
n	2	5	5	4	4	4	3	6	4	3	6	4	1	4	5	4	5	4	4	4

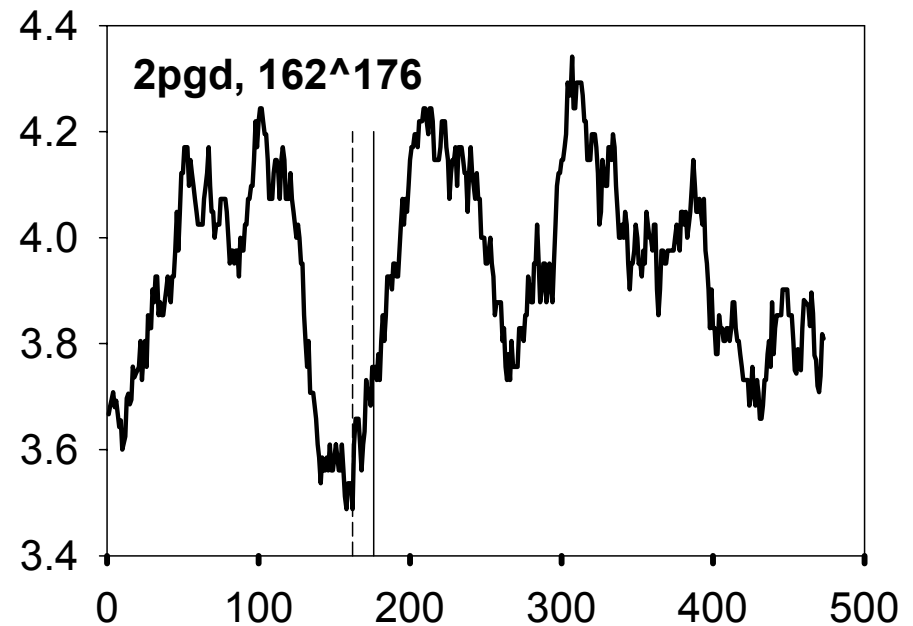
The method was applied to 646 proteins with two contiguous domains extracted from the SCOP 1.59 database (homology <80%) with a success rate of 63%.

Success of prediction strongly depends on the distance between the predicted and the real boundaries at which the prediction is still considered to be correct

$$\langle P_{random} \rangle = \sum_{i=1}^{646} \frac{N_i}{L_i - 100} = 47\%$$

$\langle P_{random} \rangle$ – average probability of random coincidence of the real and predicted boundaries for the given distance from the domain boundary

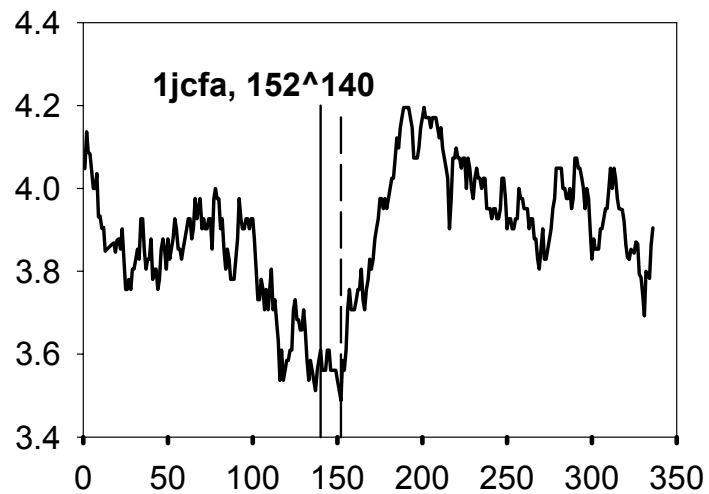
N – the double maximal allowable distance between the real and predicted boundaries;
 L – protein length



----- predicted boundary, — the real boundary

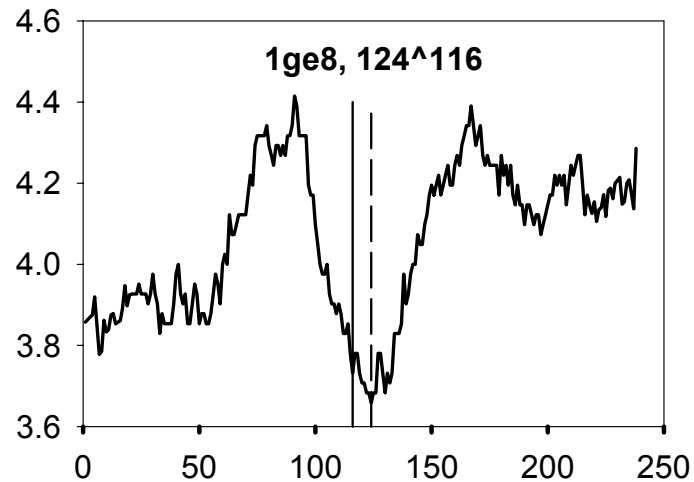


Actin-like ATPase domain



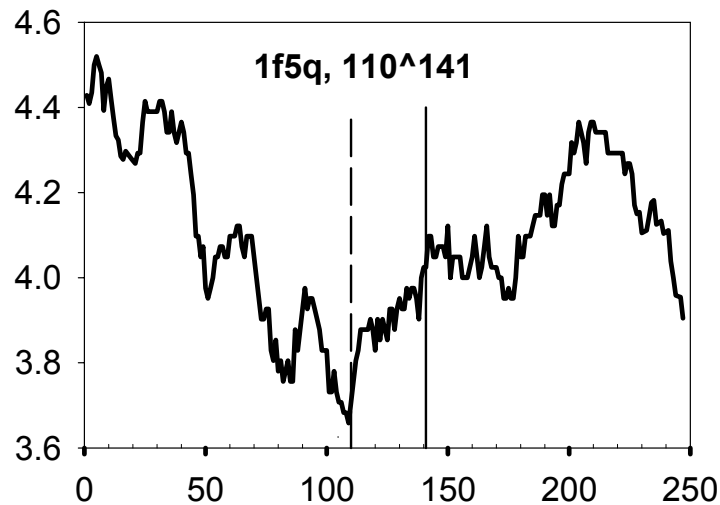
Ala
Pro
Val
Gly

DNA clamp



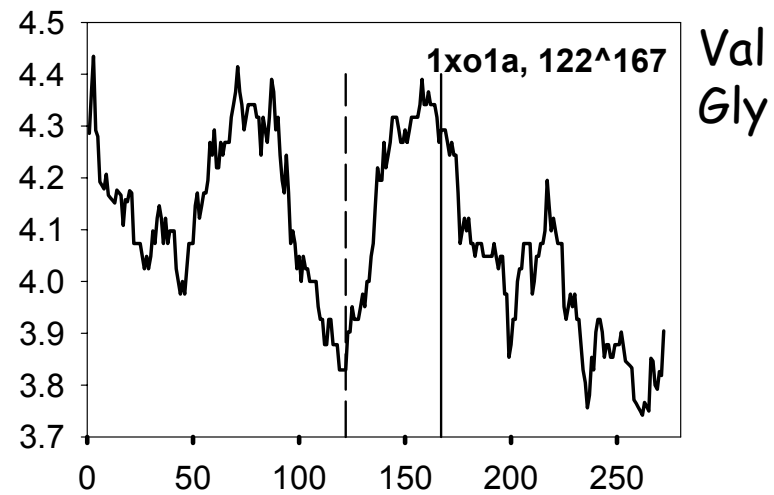
Pro
Val

Cyclin-like



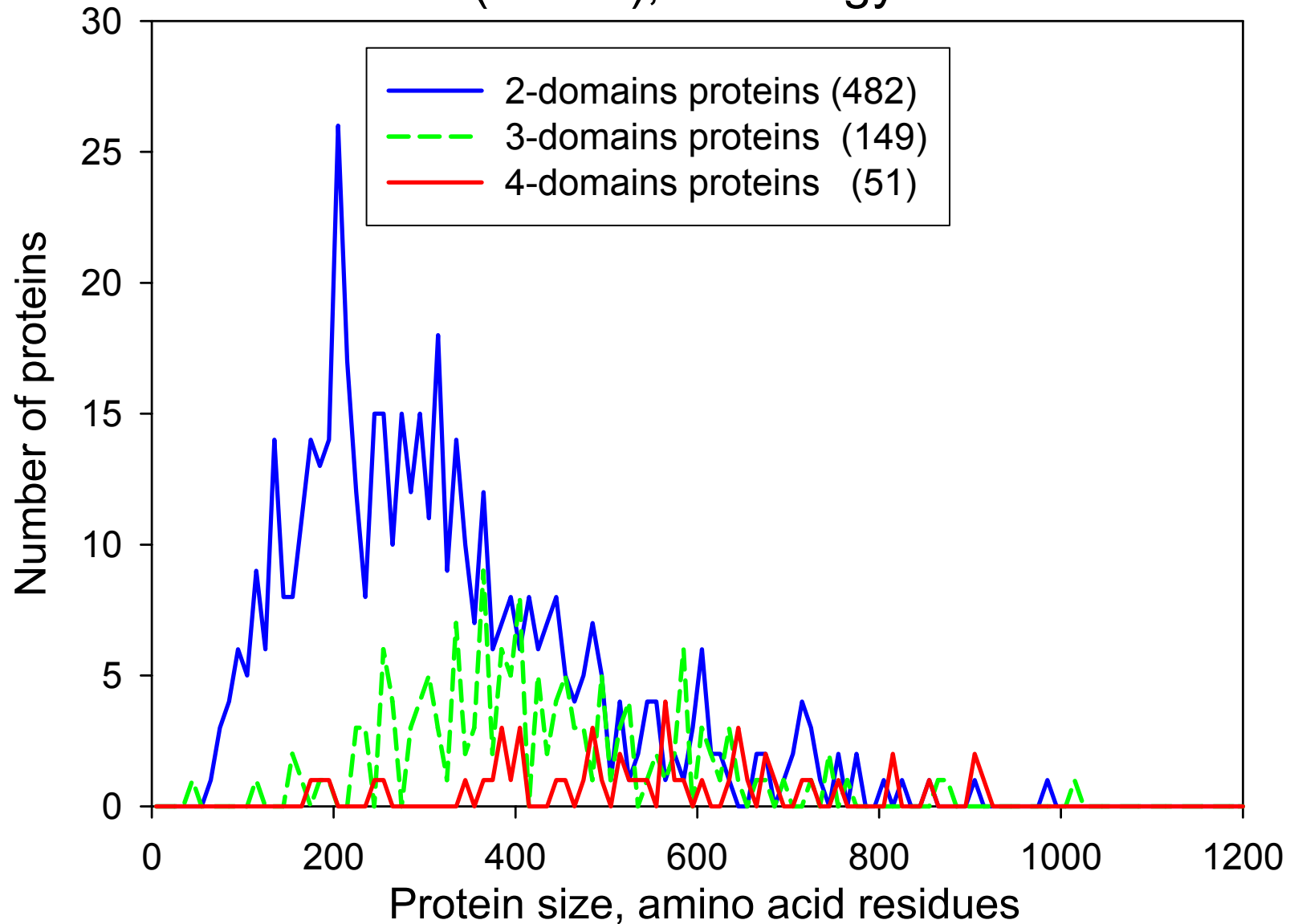
Pro

5' to 3' exonuclease, C-terminal subdomain Resolvase-like

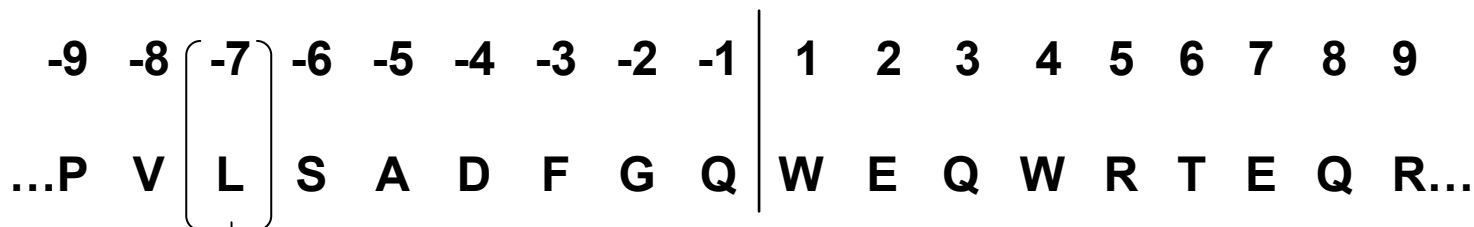


Val
Gly

Database of multi-domains proteins:
SCOP (v.1.63), homology <25%



Statistics of amino acid residues at domain boundaries

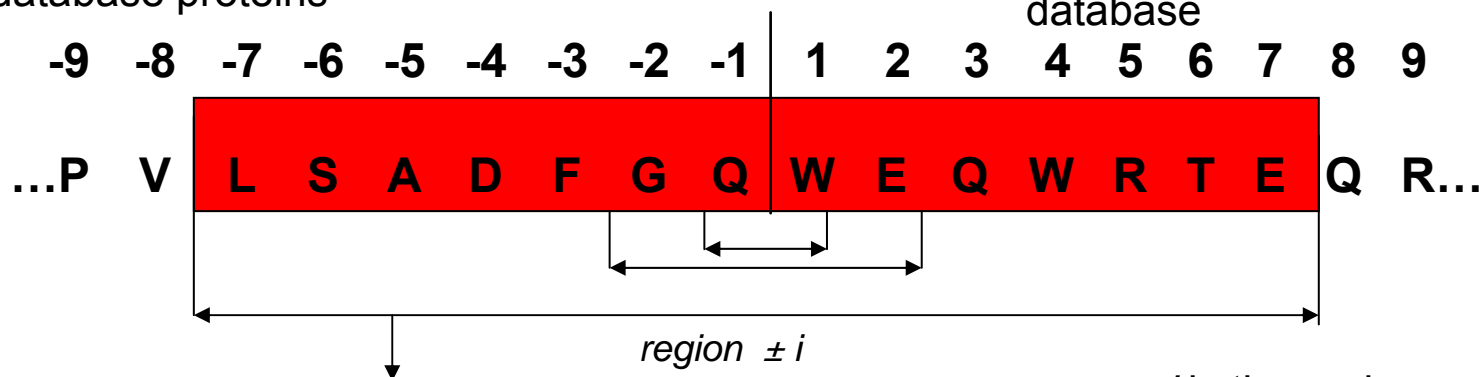


The appearance of a particular amino acid residue in position i relative to the total appearance of this residue in the database proteins

$$\rightarrow P_i^X = \frac{W_i^X}{W^X}$$

The appearance of residue X in position i relative to the domain boundary

the appearance of each residue was computed for the total database



The appearance of a particular residue in the regions of domain boundaries was computed as

$$\rightarrow P_{\pm i}^X = \frac{\sum_{-i}^i P_i^X}{2i}$$

$\pm i$ is the region containing $2i$ amino acid residues and a domain boundary at position 0

Small and hydrophilic residues appear at domain boundaries more often than in total proteins

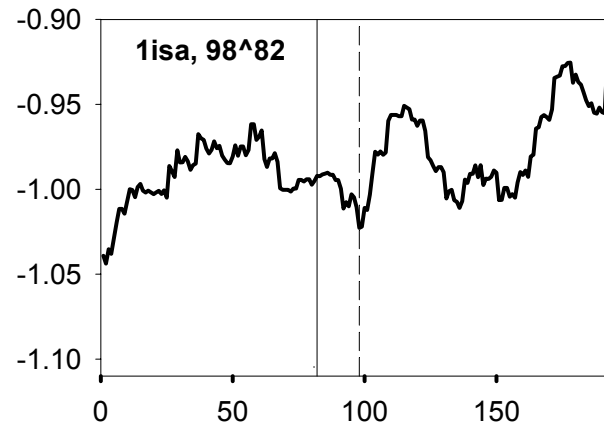
<i>X</i>	Pro	Gly	Gln	Cys	Asp	Ser	Glu	Thr	Arg	Lys
<i>P</i>	2.11	1.46	1.45	1.43	1.21	1.17	1.12	1.10	1.07	1.05

Hydrophobic residues appear at domain boundaries rarer than in total proteins

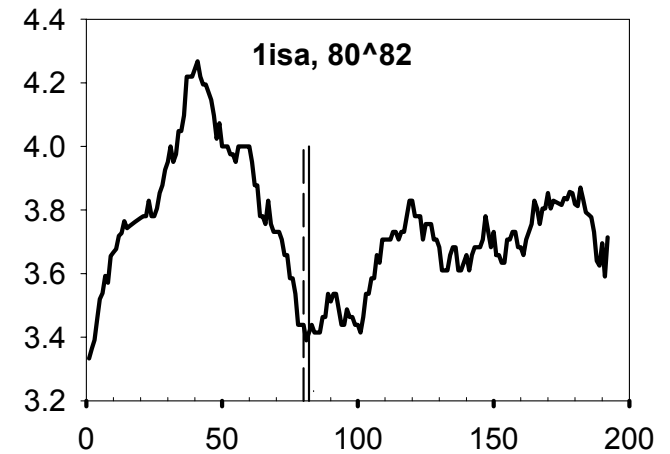
<i>X</i>	Asn	Tyr	Ala	Val	Phe	His	Ile	Trp	Met	Leu
<i>P</i>	0.90	0.85	0.78	0.72	0.71	0.64	0.64	0.62	0.60	0.60

Comparison of the predicted and real domain boundaries for 28 groups of two-domain proteins

Fe, Mn superoxide
Dismutase (SOD)
N-terminal domain
C-terminal domain

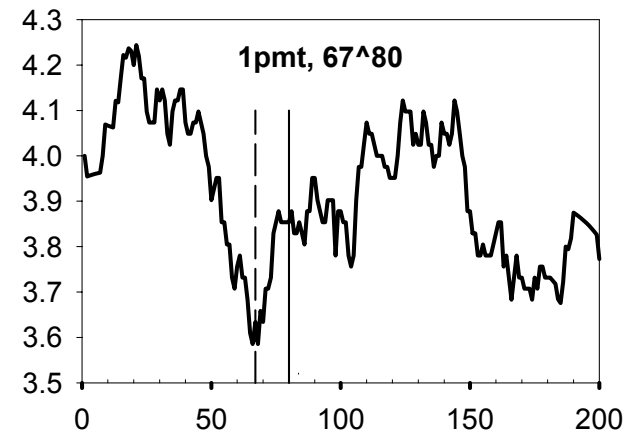
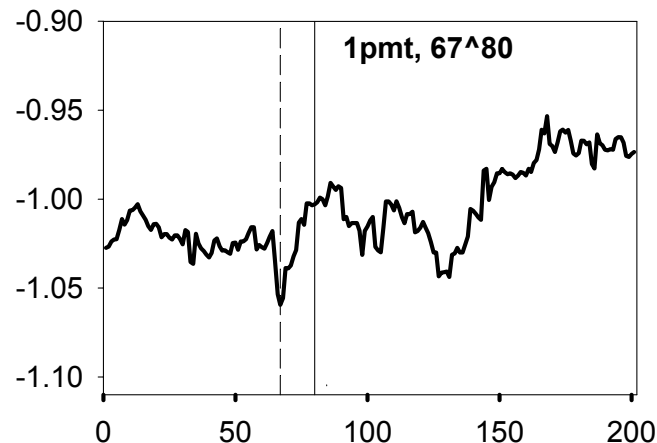



Probability profile



Entropy profile

Glutation
S-transferases,
C-terminal domain
Thioredoxin-like





Prediction of number and position of domain boundaries in multi-domain proteins using only amino acid sequence alone

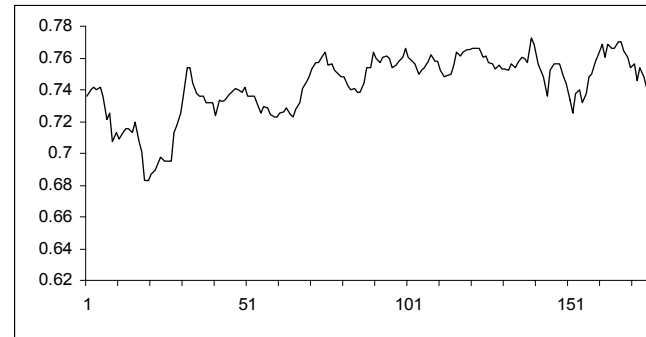
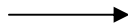
Different methods agree only for ~81% of all domains (Liu and Rost, 2004)

Success of prediction strongly depends on the distance between the predicted and the real boundaries at which the prediction is still considered to be correct

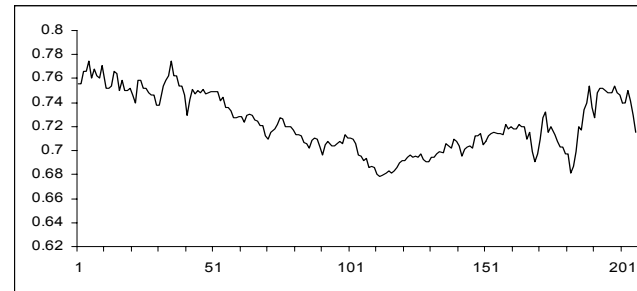
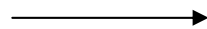
Our method used an optimized scale of appearance of amino acid residues on the domain boundaries.

Examples of character profiles for

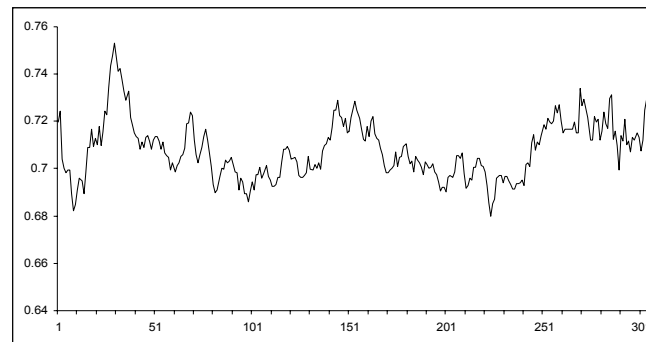
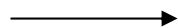
1-domain



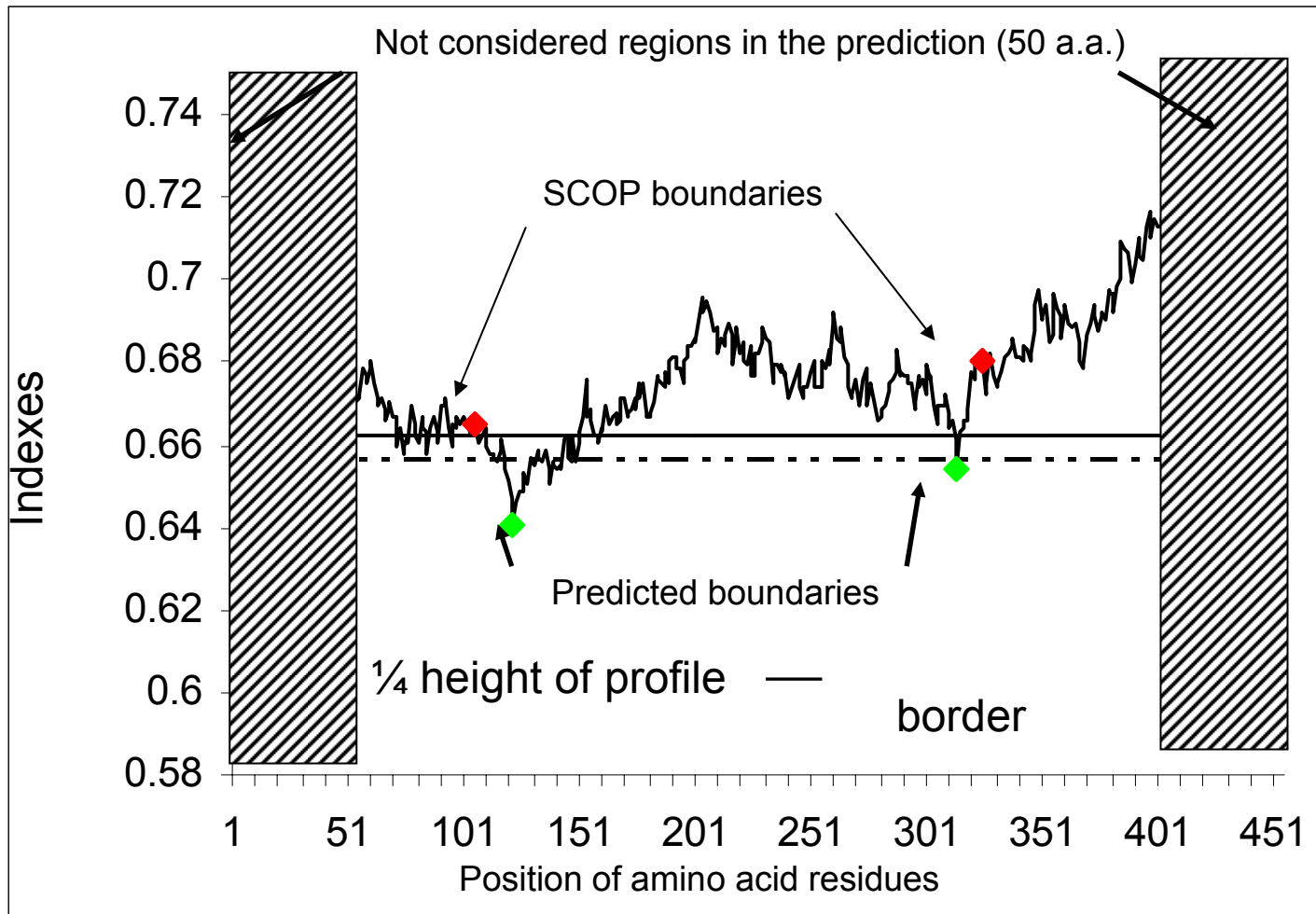
2 domains



3 domains

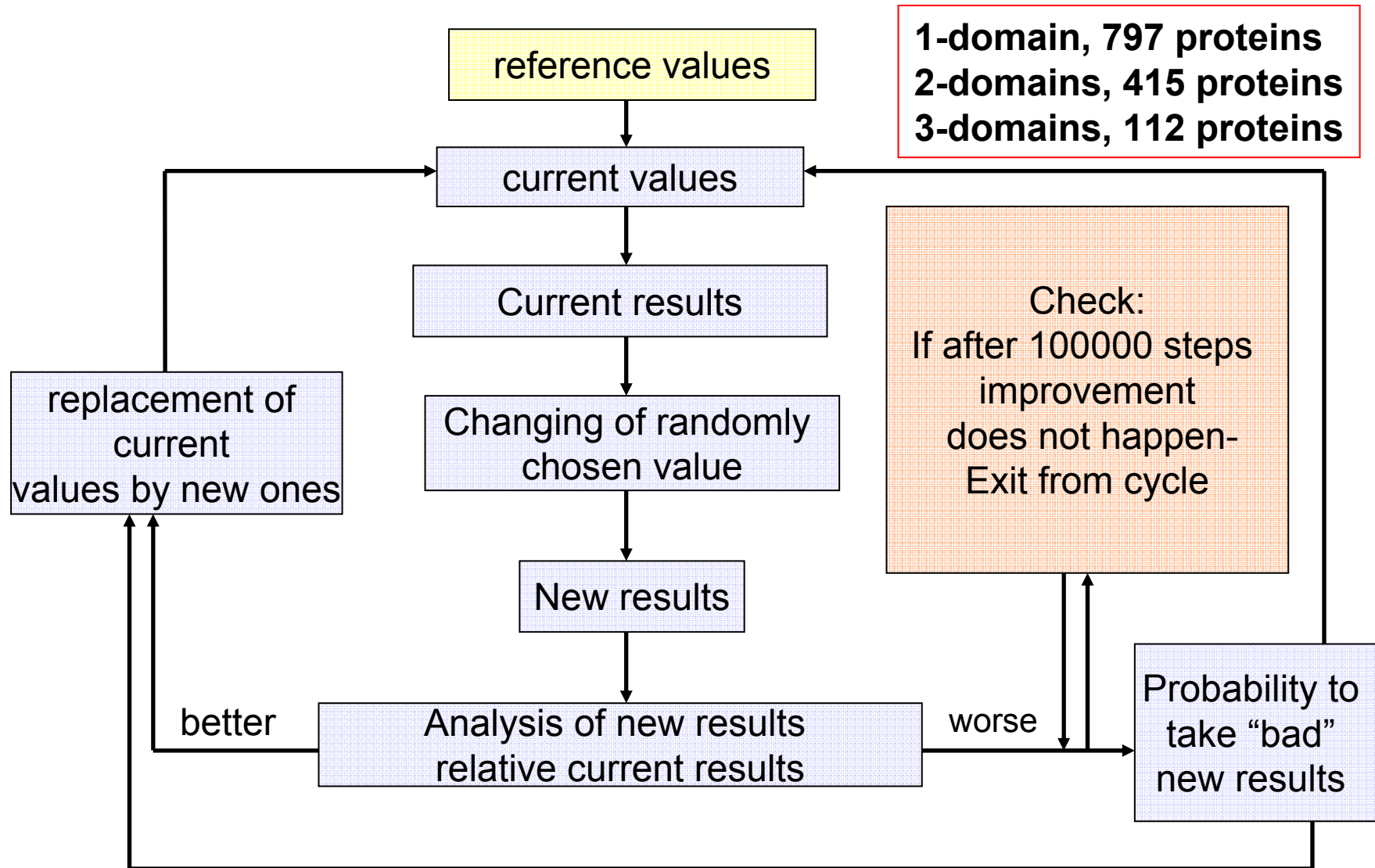


Averaging over minima and determination both numbers and domain boundaries



3 domains protein 1cjc - oxidoreductase (adrenodoxin reductase)

Optimization of parameters by Monte-Carlo algorithm



Optimized scale of appearance of amino acid residues on the domain boundaries

Results of predictions number of domains after optimization of parameters by Monte-Carlo algorithm

Number of domains in protein	Predicted number of domains in protein / starting meaning			
	1	2	3	Larger, or equals to 4
1 (797 proteins)	57 / 53%	31 / 32%	8 / 11%	4 / 4%
2 (415 proteins)	36 / 44%	49 / 34%	10 / 11%	5 / 12%
3 (112 proteins)	29 / 22%	34 / 41%	23 / 21%	14 / 16%

Comparison of our method **ADP**
with method **CHOPnet** (J. Liu и B.Rost, 2004)

CHOPnet is based on neural networks and relied exclusively on information available for all proteins

Number of domain observed	Percentage of proteins with correctly predicted number of domains		Percentage of proteins with correctly predicted number and location of domains (± 20 residues)	
	ADP	CHOPnet	ADP	CHOPnet
1	57%	73%	-	-
2	49%	41%	17%	19%
3	23%	21%	6%	0%
			5%	0%

Percentage of proteins with correctly predicted number of domains for CASP6

		RosettaDom (Kim et al., 2005)		
		1	2	больше
1 domain – 47 proteins		<u>85</u>	15	0
		14	<u>77</u>	9
2 domains – 22 proteins		DomSSEA (Marsden et al., 2002)		
		1	2	больше
		<u>91</u>	9	0
		73	<u>23</u>	4
		DomPred (Marsden et al., 2002)		
		1	2	больше
		<u>83</u>	15	2
		50	<u>45</u>	5

Armadillo (Dumontier et al., 2005)		
1	2	больше
<u>15</u>	68	17
0	<u>55</u>	45
ADP (Galzitskaya et al., 2006)		
1	2	больше
<u>79</u>	19	2
41	<u>50</u>	9

The main advantages of our method are as follows: it is very simple, fast and uses minimal number of parameters in comparison with other methods.



Thanks for your attention!



Acknowledgements:


This work was supported by:

the RAS program “Physical and Chemical Biology”,
the Russian Foundation for Basic Research,
the INTAS grant

http://skuld.protres.ru/~mlobanov/bm_og/bm_og.cgi.

<http://skuld.protres.ru/~mlobanov/og3/og3.cgi>

<http://skuld.protres.ru/~mlobanov/ogu/>



The Z-score is $(M - \langle M \rangle) / \sigma$, where M is the number of correctly predicted domain boundaries by our method and $\langle M \rangle$ is the average number of expected successful random predictions in our method which is equal to the summation of probabilities p_i , where i changes from 1 to the considered number of the proteins. σ is the standard deviation.



**Scale has been obtained after optimization
and normalized relative Proline value**

Pro	Ala	Val	Cys	Asn	Gln	Glu	Trp	Leu	Thr
1	0.8	0.8	0.8	0.7	0.7	0.7	0.7	0.7	0.6

Ile	Tyr	His	Met	Phe	Ser	Arg	Asp	Gly	Lys
0.6	0.6	0.6	0.6	0.6	0.6	0.5	0.5	0.4	0.4