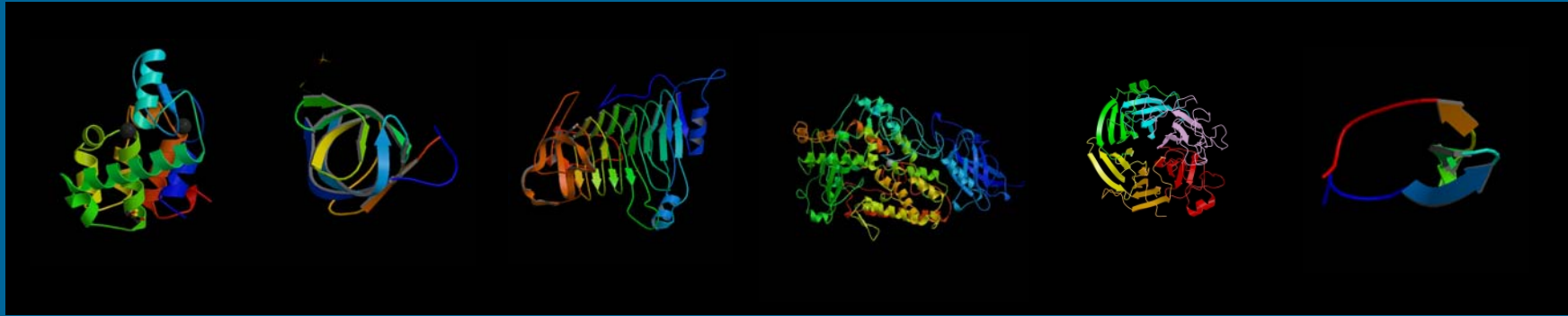


**Dissecting proteins into
ordered and disordered parts using
estimated pairwise energies**

Zsuzsanna Dosztányi

Institute of Enzymology, BRC, Budapest, Hungary



1. IUPred: prediction of protein disorder based on estimated pairwise energies
2. Protein disorder at the domain level
3. Large scale identification of IUPs by a novel 2D gel electrophoresis technique

Pairwise energy of globular proteins

Statistical potentials

The interactions observed more frequently in real structures are more favorable

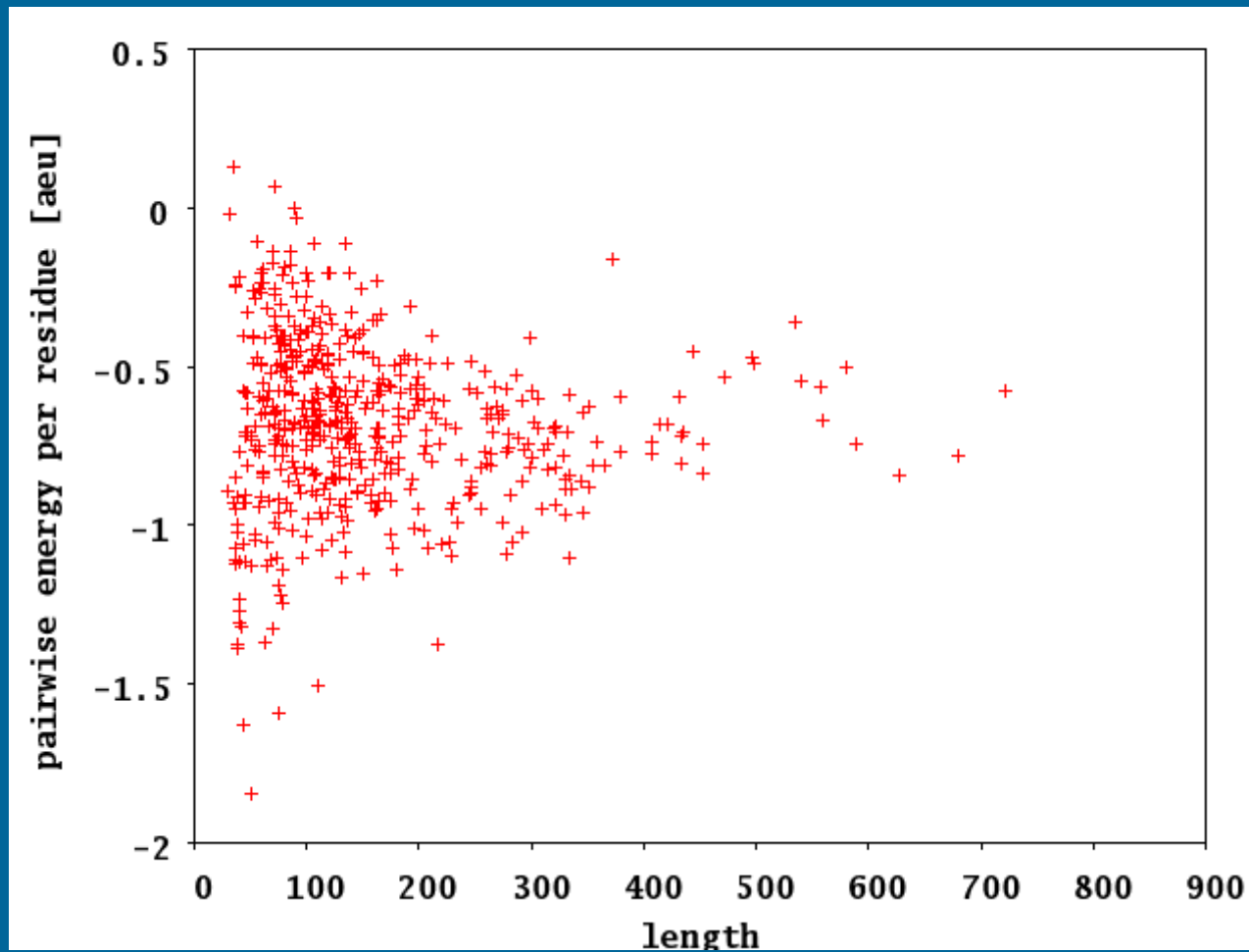
Calculate the observed frequencies of amino acid pairs in native proteins relative to a reference state

Convert this ratio into energies using the Boltzmann relation

To calculate energy, count contacts between each amino acid pair in the structure weighted by the appropriate energy

Calculation is based on the energy matrix of Thomas and Dill (1996)
PNAS, 93, 11628-11633

Pairwise energy calculated from structure



Structure

MODEL	1							
ATOM	1	N	MET	A	23	2.191	28.312	-4.381
ATOM	2	CA	MET	A	23	2.394	27.327	-3.305
ATOM	3	C	MET	A	23	3.514	26.377	-3.706
ATOM	4	O	MET	A	23	3.589	25.977	-4.867
ATOM	5	CB	MET	A	23	1.128	26.503	-3.033
ATOM	6	CG	MET	A	23	0.025	27.305	-2.344
ATOM	7	SD	MET	A	23	-1.456	26.318	-2.038
ATOM	8	CE	MET	A	23	-2.566	27.602	-1.402
ATOM	9	1H	MET	A	23	2.034	27.828	-5.254
ATOM	10	2H	MET	A	23	1.397	28.910	-4.199
ATOM	11	3H	MET	A	23	3.017	28.882	-4.497



Calculated
energy per
residue

Sequence

MKVPPHSIEA	EQSVLGGLML
DNERWDDVAE	RVVADDFYTR
PHRHIFTEMA	RLQESGSPID
LITLAESLER	QGQLDSVGGF
AYLAELSKNT	PSAANISAYA
DIVRERAVVR	EMIS

Amino acid
composition
(*n*)



A	10.5
C	0.0
D	7.0
E	9.6
F	2.6
G	5.3
H	2.6
I	6.1
K	1.8
L	8.8
M	3.5
N	2.6
P	4.4
Q	3.5
R	7.9
S	8.8
T	3.5
V	7.9
W	0.9
Y	2.6



Estimated
energy per
residue

$$E(\text{estimated}) / L$$

Estimation of pairwise energies from amino acid compositions

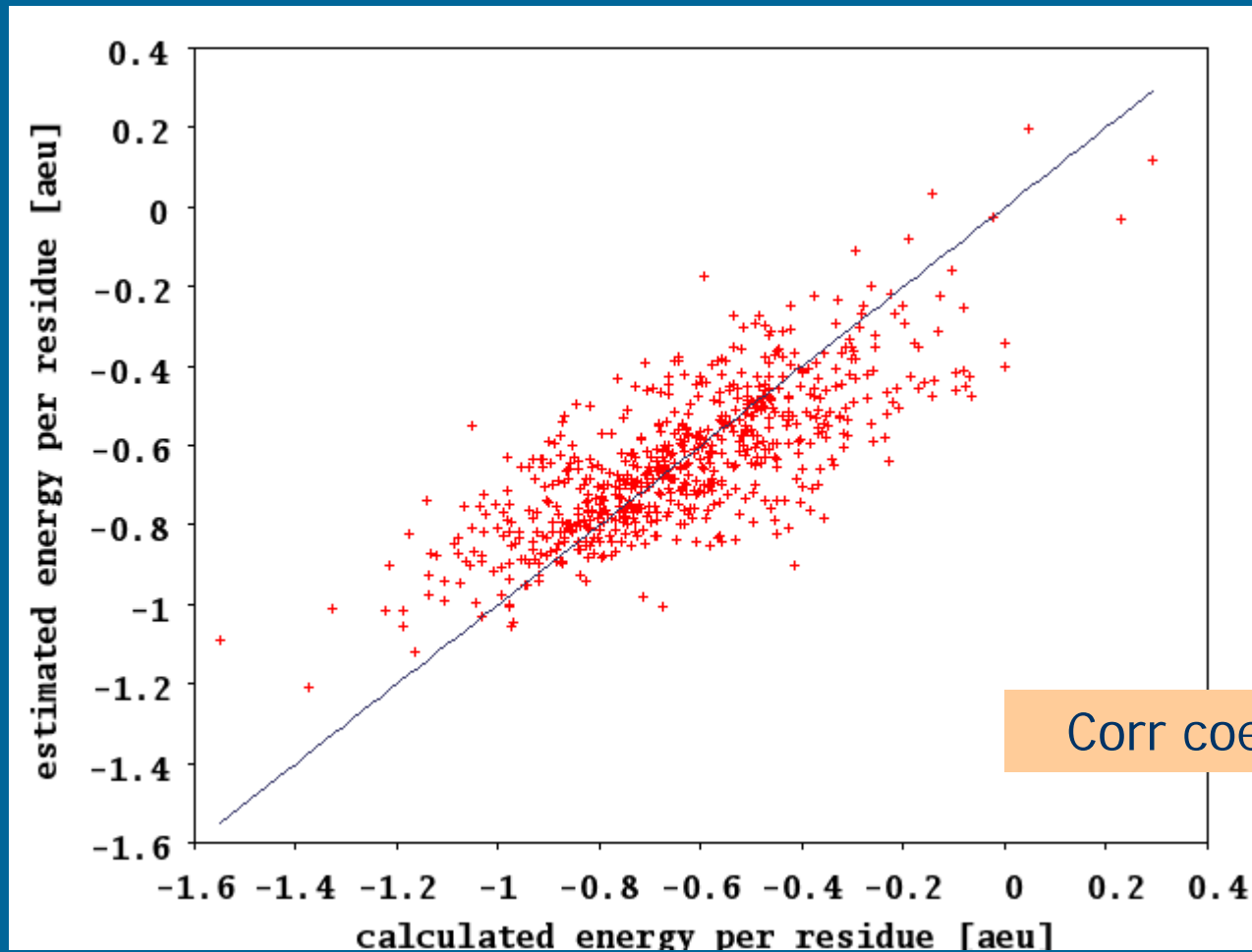
To take into account that the contribution of amino acid i depends on its interaction partners, we need a quadratic form in the amino acid composition

$$E(\text{estimated}) / L = (n_A \quad n_C \quad \dots \quad n_Y) \begin{pmatrix} P_{AA} & P_{AC} & \dots & P_{AY} \\ P_{CA} & P_{CC} & & \\ \vdots & & \ddots & \\ P_{YA} & \dots & \dots & P_{YY} \end{pmatrix} \begin{pmatrix} n_A \\ n_C \\ \vdots \\ n_Y \end{pmatrix}$$

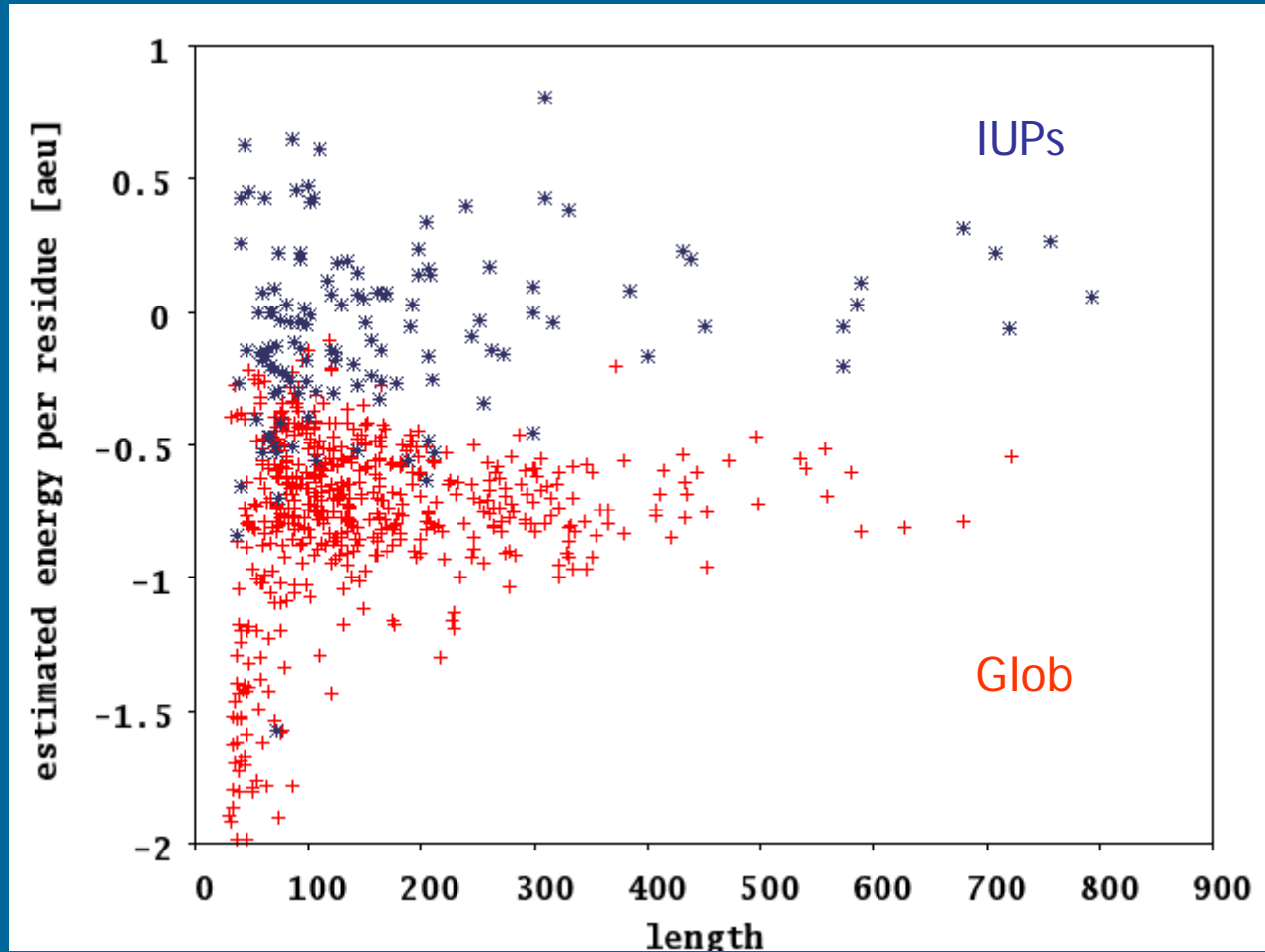
The connection between composition and energy is encoded by the 20x20 energy predictor matrix: P_{ij}

$$P_{ij} : \sum_k^{\text{Globular proteins}} (E_k(\text{calc}) - E_k(\text{est}))^2 \rightarrow \min$$

Estimated energies correlate with calculated energies



Estimated pairwise energies of globular proteins and IUPs



IUPred

- Characterizes the tendency of a residue to fall into disordered or ordered region
- Calculation is limited to a predefined sequence range (2-100 residues apart)
- Energy predictor matrix is recalculated accordingly
- Position specific score is smoothed over a window of 21
- Cutoff at 5% false positive rate (-0.2)

Performance of disorder prediction methods

(%)	True positive rate	False negative rate
	IUP list: 129 proteins	Glob list: 559 proteins
IUPred	76.0	5.0
PONDR VL3H	66.3	5.0
DISOPRED2	63.4	5.0
GlobPlot	33.0	18.7

IUPred: no training on disordered regions !

IUPred: Dosztanyi et al. JMB. 347, 827-839

PONDR VL3H: Obradovic et al. Proteins, 53, 566-572

DISOPRED2: Ward et al. JMB. 337, 635-645

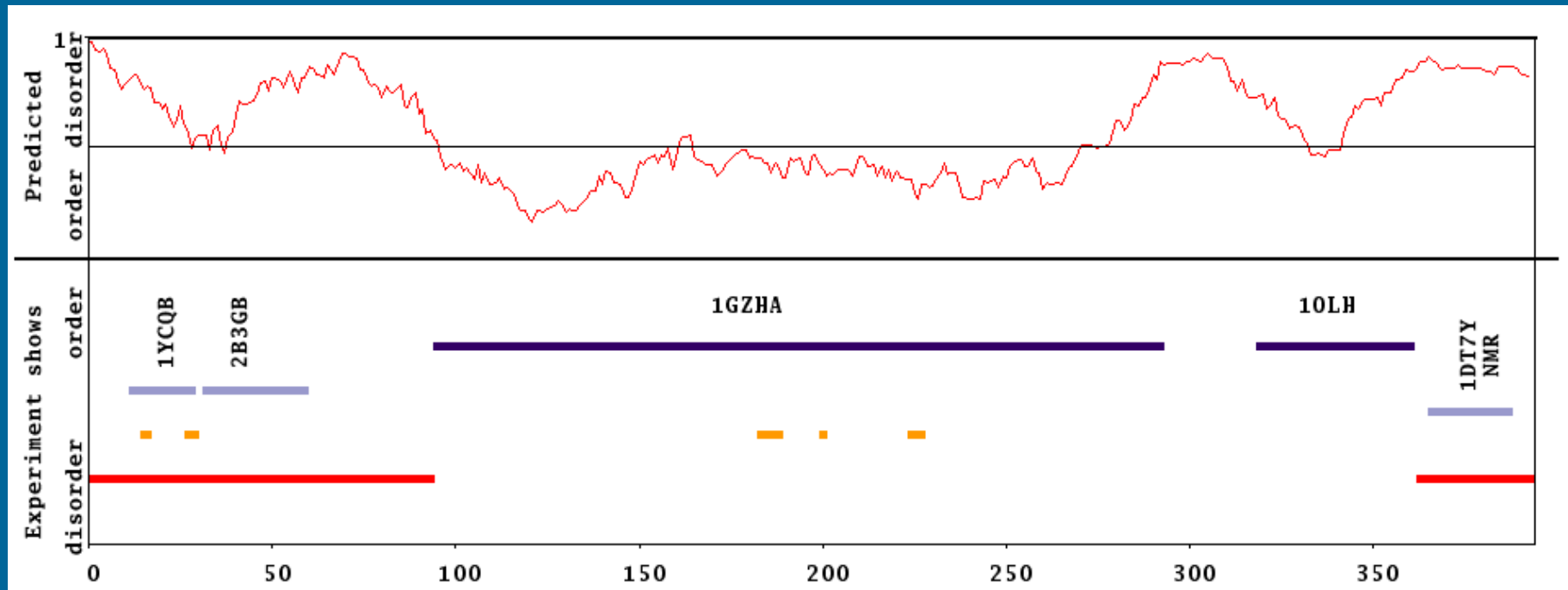
Summary /1

- Predicting pairwise energy content allowed by the amino acid composition
- For globular proteins the predicted energy is favorable
- The predicted energy for IUPs is unfavorable
- These are not random sequences
- The lack of a well defined 3D structure of these proteins is their *intrinsic* property



IUPred: <http://iupred.enzim.hu>

P53 Tumor antigen



Order and disorder at the domain level

- Ordered domains

capable of forming a well-defined structure independently
(one or more structural domains)

- Disordered domains

do not contain ordered domains
not part of ordered domain

- Local order and disorder is not relevant

- Boundaries between ordered and disordered domains

Specific datasets, algorithms, evaluation criteria

Building a database

Existing datasets: fully ordered or disordered



In real proteins these are mixed



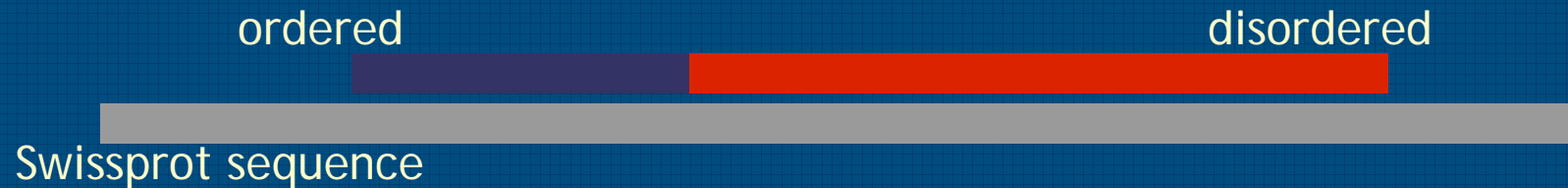
Selecting protein fragments with two or three adjacent ordered and disordered domains



Other regions are thrown out (grey areas)

Database based on experimental data

Disordered domains from DISPROT database (<http://www.disprot.org>)



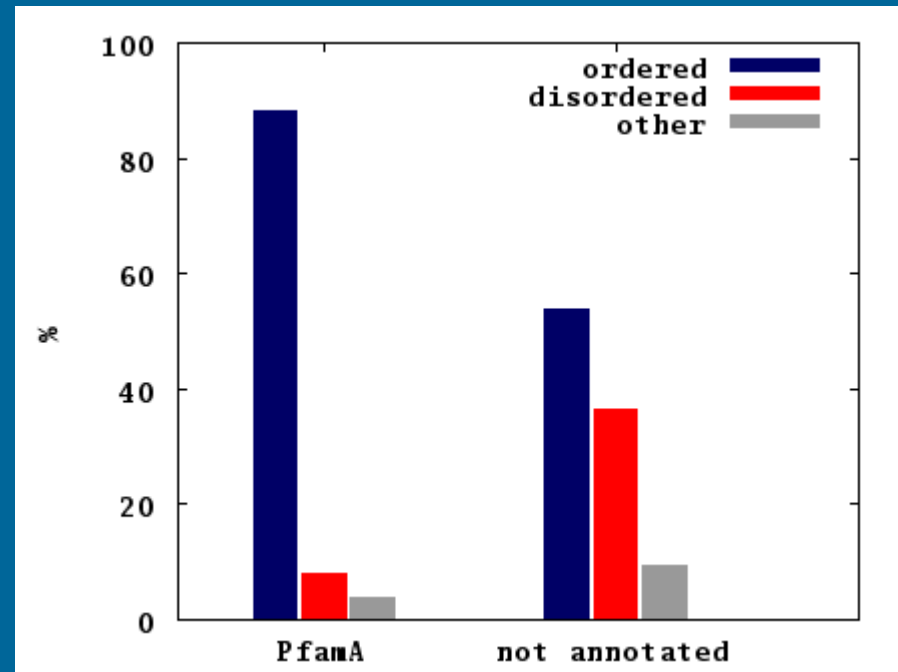
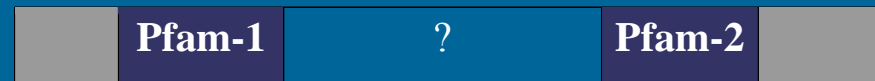
Ordered domains from PDB database (30 AA)

found by aligning the swissprot sequence with PDB sequences

Disordered domains which are part of PDB structures are omitted

Very small database: 34 entries, 2 with 3 domains

Disorder in Pfam domains and not annotated areas evaluated by IUPred



Mostly disordered region:

- no 30 residue long ordered region
- more than 50% of residues predicted disordered

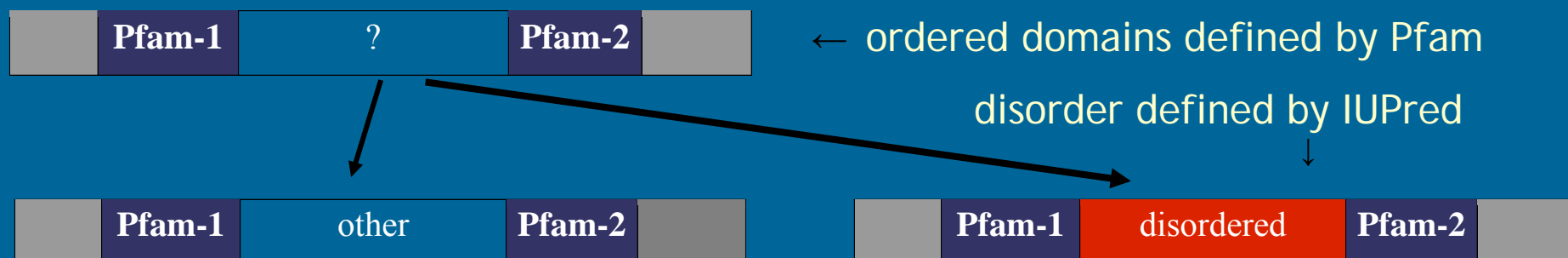
Mostly ordered region:

- no 30 residue long disordered region
- more than 50% of residues predicted ordered

Database 2

Starting from swissprot sequences annotated by PFAM domains

Not annotated areas are potential disordered domains



Much larger, but less reliable dataset (>1000 fragments)

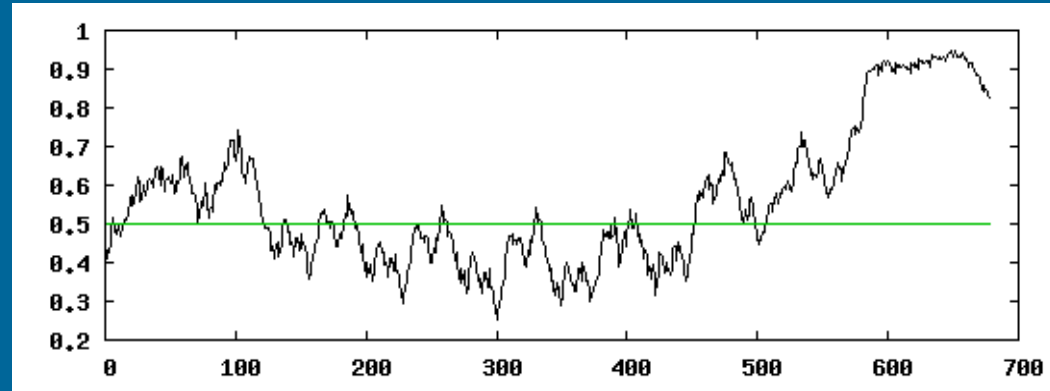
IUPred specific on disorder

Measure improvements over starting prediction by IUPred

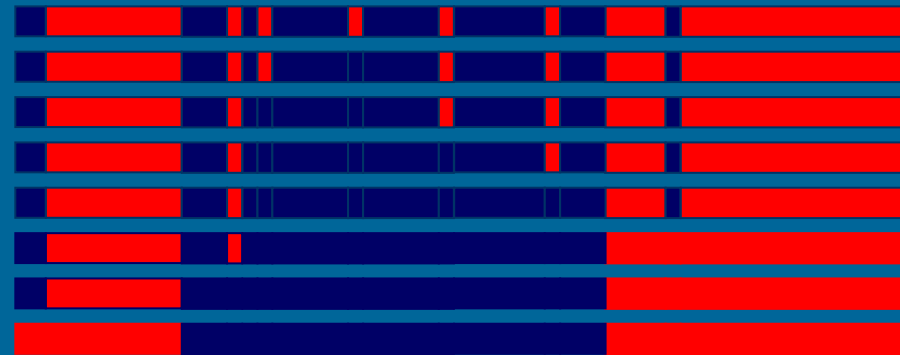
Can be used to try various algorithms, and to tune the optimal parameters

Algorithm 1

Based on length



Variation on Globplot-type
domain merging
(Linding et al. 2003.
NAR 31, 3701.)



Eliminate short regions

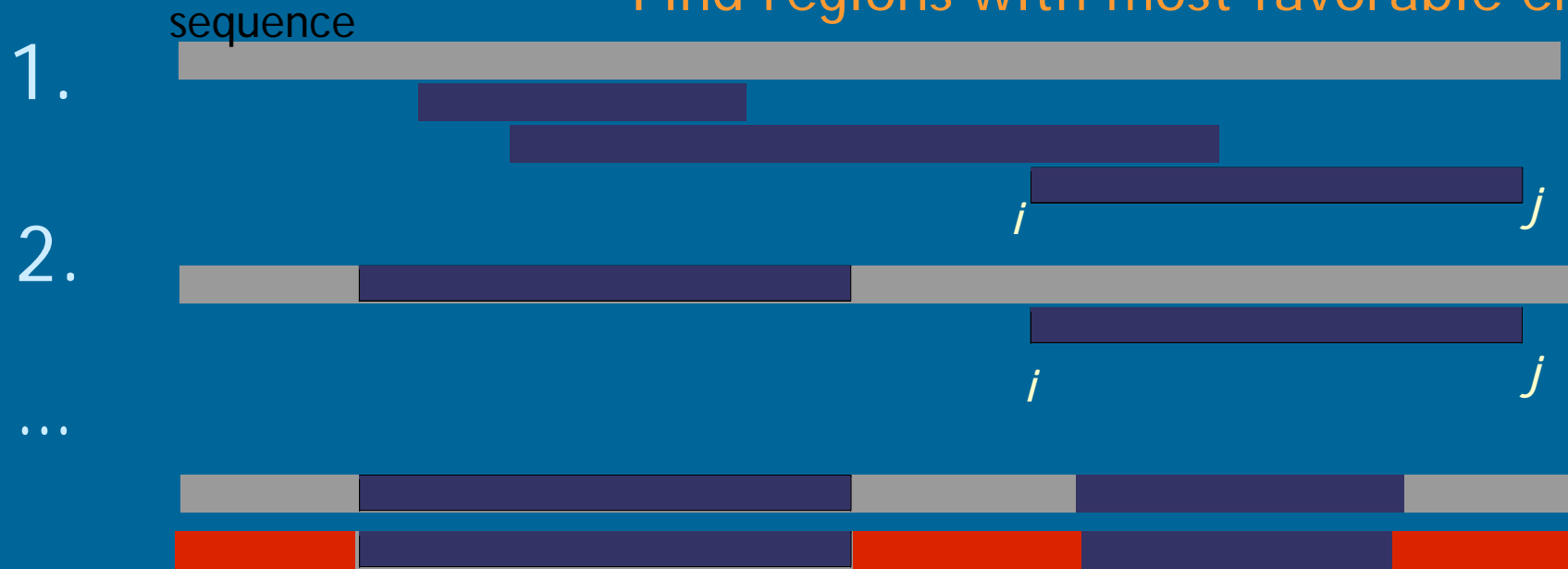
- minimum length of disordered domains (12)
- minimum length of ordered domains (31)

Algorithm 2

Variation on Scooby-domain type domain finding
George et al, 2005, NAR 33, W160.

Based on energy

Find regions with most favorable energy



2 parameters:

- Cutoff energy for ordered positions (-0.2)
- Minimum energy for ordered domain (-24)

Using lengths constraints (31, 12)

Results

A domain is correctly identified:

- Overlap between predicted and observed domain is 50 %
- Predicted start and end positions are within 30 residues of the observed ones

Percentage of correctly identified domains compared to the number of domains		
(%)	Database 1 (PDB+DisProt)	Database 2 (Pfam+IUPred)
IUPred (pos. spec.)	11.1	20.0
A1	48.6	46.0
A2	51.4	52.7

Summary / 2

- Evaluation at the domain level requires specific algorithms
- The algorithms can be applied to any kind of disorder prediction
- Algorithms based on energies have more potential
- The two types of datasets give similar results



A Novel Two-dimensional Electrophoresis Technique for the Identification of Intrinsically Unstructured Proteins*[§]

Veronika Csizmók, Edit Szöllösi, Peter Friedrich, and Peter Tompa‡

Intrinsically unstructured proteins (IUPs) lack a well defined three-dimensional structure under physiological conditions. They constitute a significant fraction of various proteomes, but only a handful of them have so far been identified. Here we report the development of a two-dimensional electrophoresis technique for their *de novo* recognition and characterization. This technique consists of the combination of native and 8 M urea electrophoresis of heat-treated proteins where IUPs are expected to run into the diagonal, whereas globular proteins either precipitate upon heat treatment or unfold and run off the diagonal in the second dimension. This behavior was born out by a collection of 10 known IUPs and four globular proteins. By running *Escherichia coli* and *Saccharomyces cerevisiae* extracts, several novel IUPs were also identified by mass spectrometric analysis of spots at or near the diagonal. By comparing this novel method to several other techniques, such as the PONDR[®] predictor, hydrophobicity-net charge plot, CD analysis, and gel filtration chromatography, it was shown to provide dependable global assessment of disorder even in dubious cases. Overall the reproducibility and ease of performance of this technique may promote the proteomic scale recognition and characterization of protein disorder. *Molecular & Cellular Proteomics* 5:265–273, 2006.

the more compelling as IUPs¹ play essential physiological and pathological roles (1, 2, 4, 8).

IUPs have so far been identified by the chance observation of the structural anomaly of proteins studied for their functional interest. We reasoned that a straightforward technique to separate IUPs from globular proteins in a cellular extract could be established by the combination of a native gel electrophoresis of heat-treated proteins followed by a second, denaturing gel containing 8 M urea. The rationale for the first dimension is that IUPs are very often heat-stable as demonstrated for Csd1 (9), MAP2 (10), NACP (11), stathmin (12), and p21^{Cip1} (13) for example. Heat treatment thus results in a good initial separation from globular proteins, most of which aggregate and precipitate. In the native gel, IUPs and rare heat-stable globular proteins will then be separated according to their charge/mass ratios. Combining this first dimension with an 8 M urea second step is rationalized by the usual structural indifference of IUPs to chemical denaturation by trichloroacetic acid, guanidine HCl, or urea as reported for Csd1 (9), NACP (11), β -casein (14), stathmin (12), and p21^{Cip1} (13) for example. As urea is uncharged and IUPs are just as "denatured" in 8 M urea as under native conditions, they are expected to run the same distance in the second dimension and end up along the diagonal. Heat-stable globular proteins

Molecular and Cellular Proteomics (2005) 5, 265-273

Known IUPs

< 500 (from DisProt <http://www.disprot.org>)

Estimated IUPs

many thousands

IUPs: lack a well-defined 3D structure

insensitive to denaturation

heat

urea, GdCl

A novel 2D-PAGE for identifying IUPs

1. Heat-treatment of extract (yeast, human cell-line)



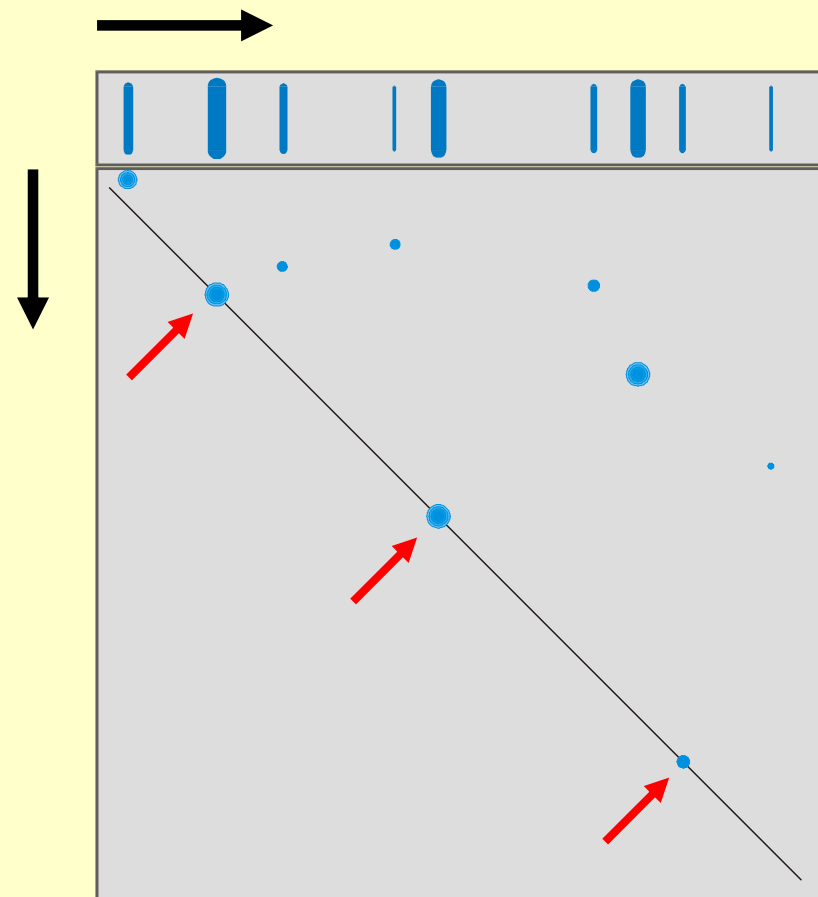
2. Native PAGE of supernatant in 1st D



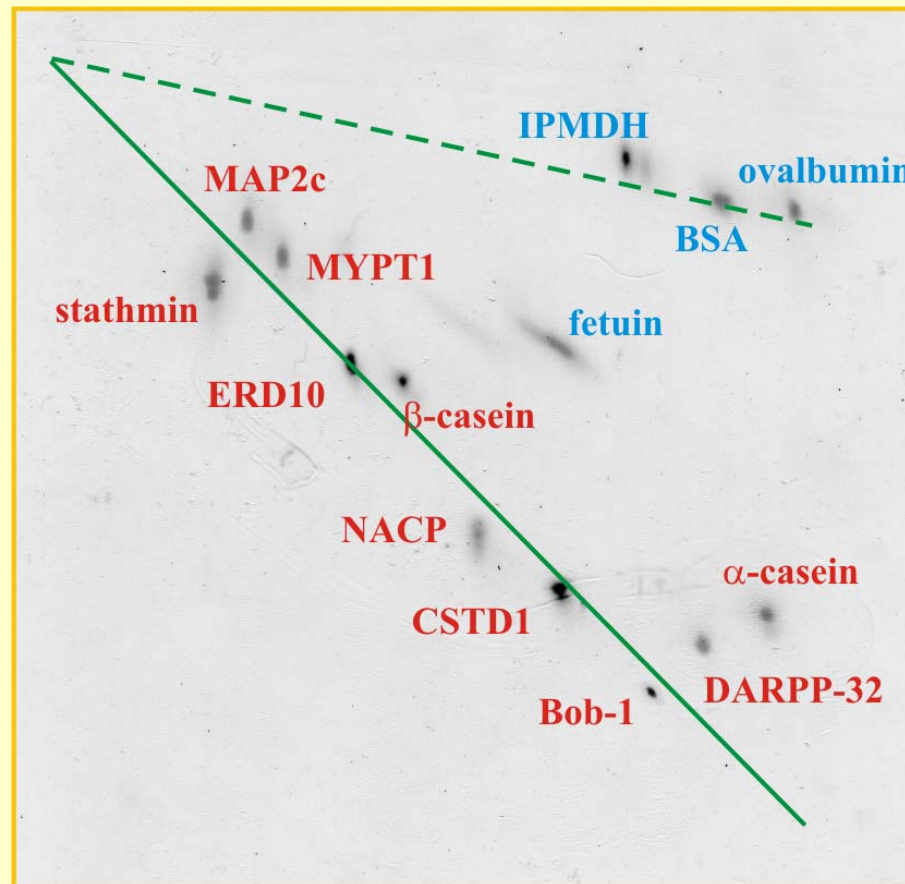
3. PAGE in 8M
UREA in 2nd D



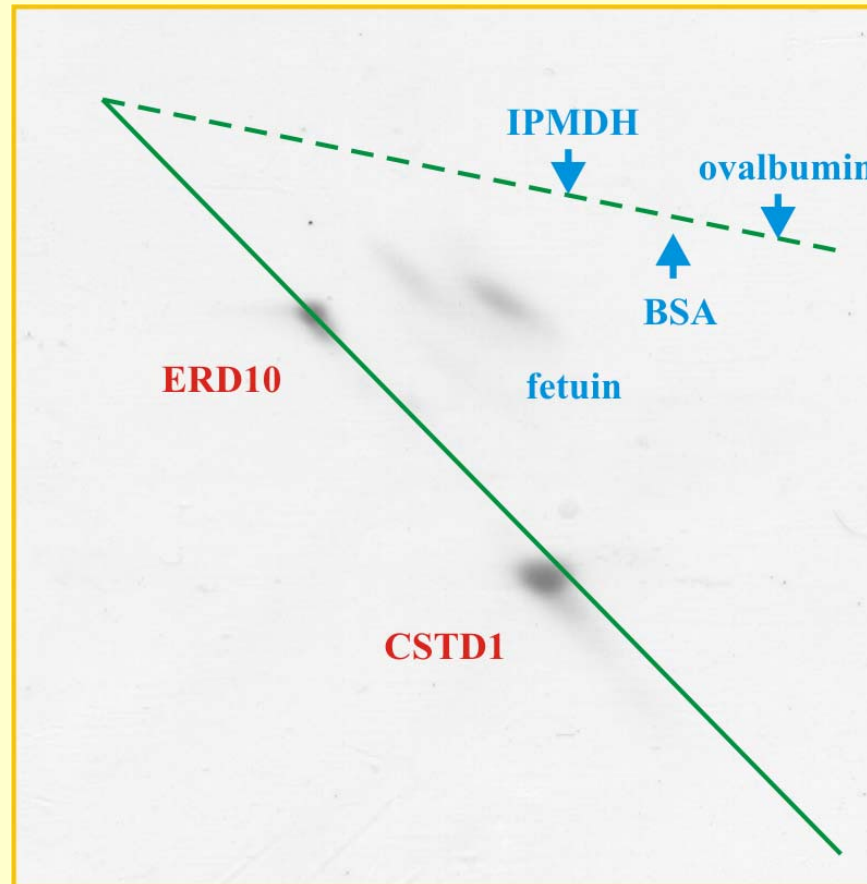
4. Identification of
selected spots by MS



Native/8M urea 2D electrophoresis separates disordered proteins and globular controls



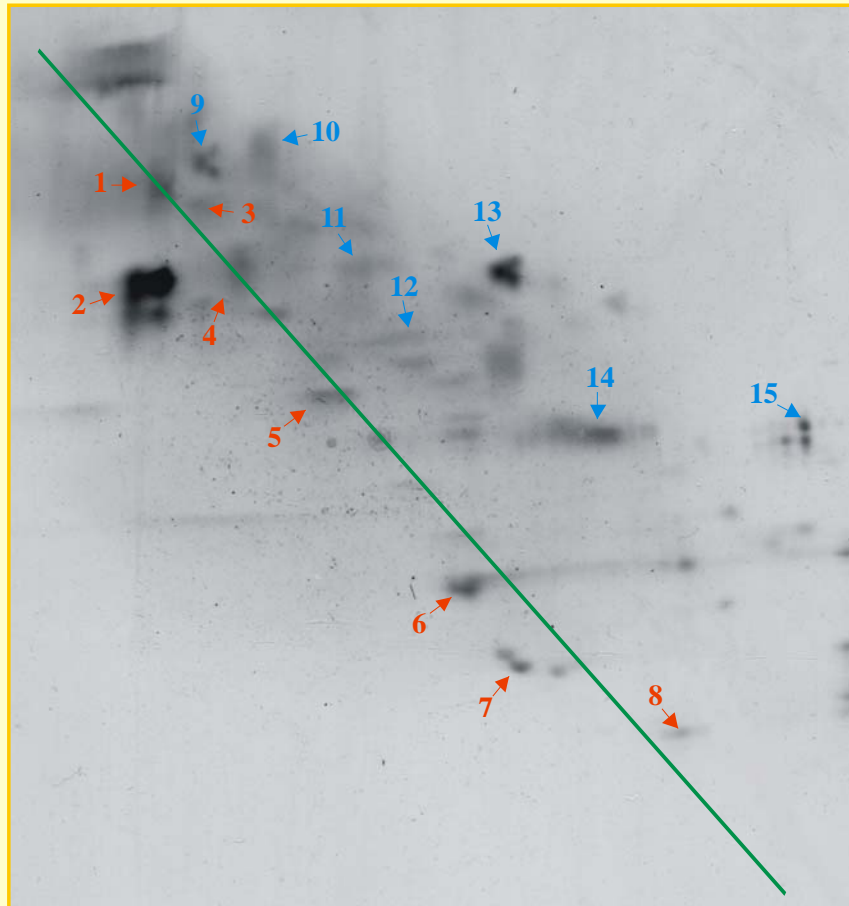
Probing the IUP nature of structurally uncharacterized proteins



Limited quality, microgram quantity

Separation and identification of IUPs from cell extracts

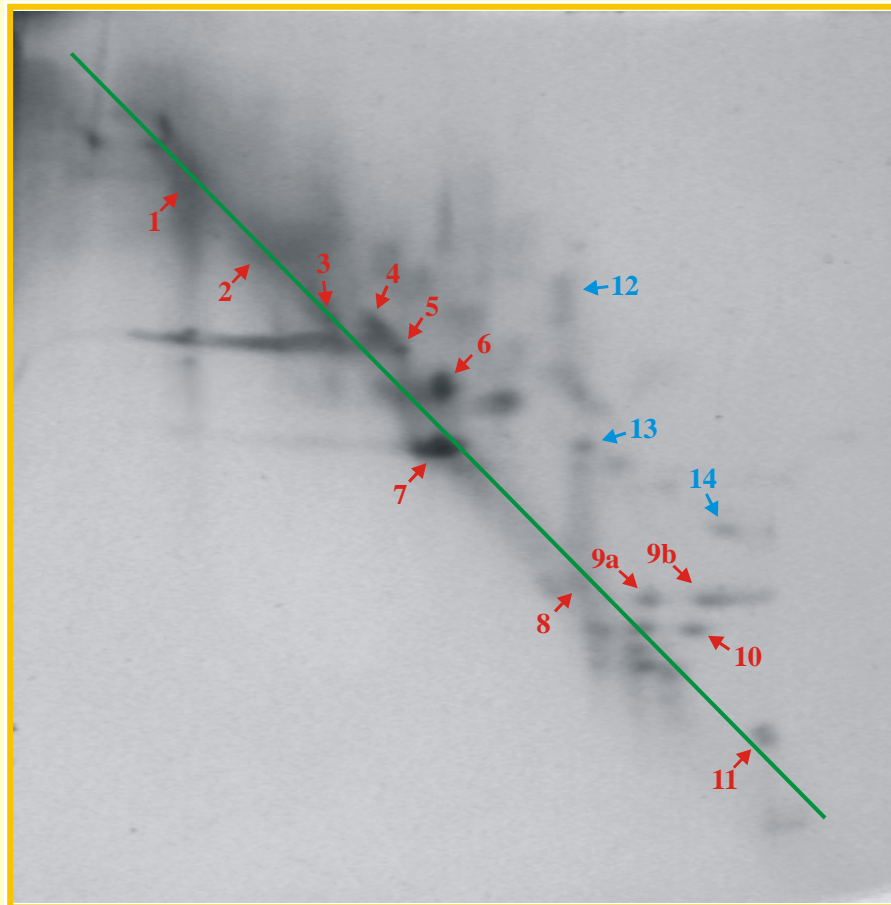
E. coli



Spot	protein/function
1	GroES (10 kDa chaperonin)
2	Ribosomal L7/L12
3	dnaK supressor
4	AcetylCoA carboxylase, BCCP subunit
5	Hypothetical protein yhgI
6	Hypothetical protein (ORF1)
7	Glycine cleavage complex H
8	Acyl carrier protein
9	Superoxide dismutase
10	Aspartate 1-decarboxylase
11	Hypothetical protein (ORF2)
12	Thioredoxin
13	PTS system, IIA component
14	FKBP-type peptidyl-prolyl cis-trans isomerase, His-rich
15	Flavodoxin 1
	NACP
	casein

Separation and identification of IUPs from cell extracts

S. cerevisiae



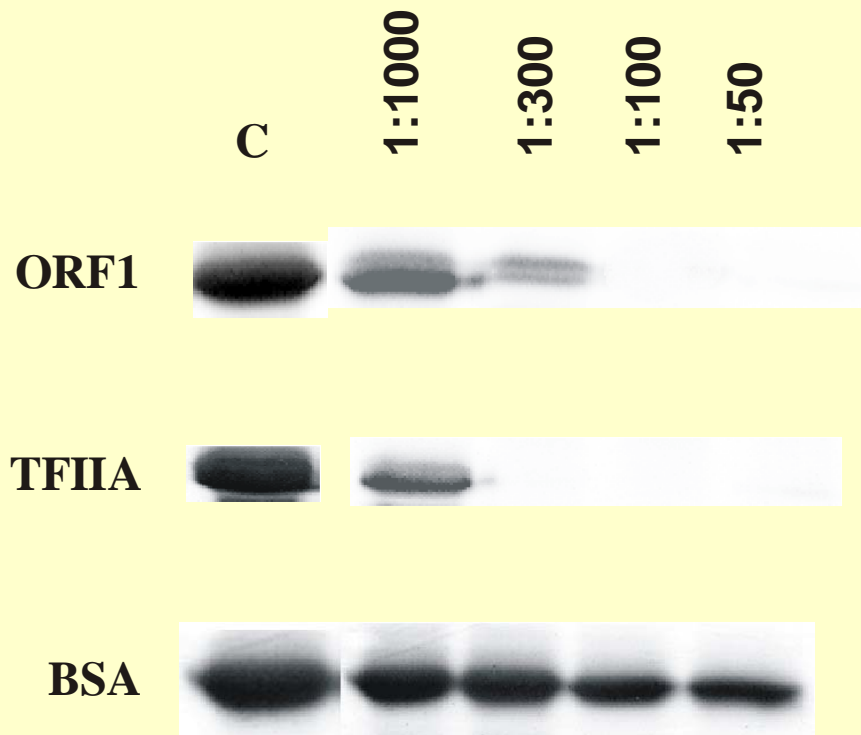
Spot	protein/function
1	Protein component of the small ribosomal subunit
2	Chain A, Yeast Superoxide dismutase
3	Transcription factor IIA large chain
4	Actin-binding main tropomyosin
5	Clathrin light chain
6	Translation elongation factor eEF-1 beta
7	Ribosomal protein L44'
8	Centromere DNA-binding complex subunit D
9	60S acidic ribosomal protein P2-beta (L45)
10	60S acidic ribosomal protein P2-alpha (L44)
11	Ubiquinol-cytochrome c oxidoreductase subunit 6
12	Myosin-2 light chain
13	Homologous to SUMO-1
14	Yeast calmodulin

Verifying the IUP nature of ORF1 and TFIIA

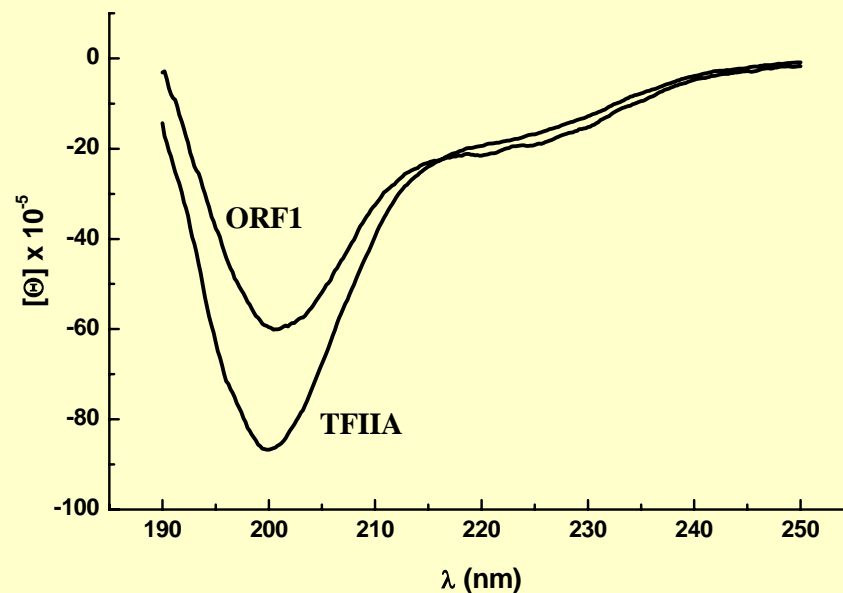
Gel filtration

	$M_{W, app} / M_W$
ORF1	4.0
TFIIA	4.4

Proteolysis



CD spectroscopy



Summary / 3

- A new 2D gelelectrophoresis technique, which separates IUPs from globular proteins
- There are two main applications of this technique
 - Probing IUP nature of a structurally uncharacterized protein
 - Identifying new IUPs from different cellular extracts
- This technique characterizes the global tendency for order/disorder

Acknowledgments

István Simon

Péter Tompa

Márk Sándor

Veronika Csizmók

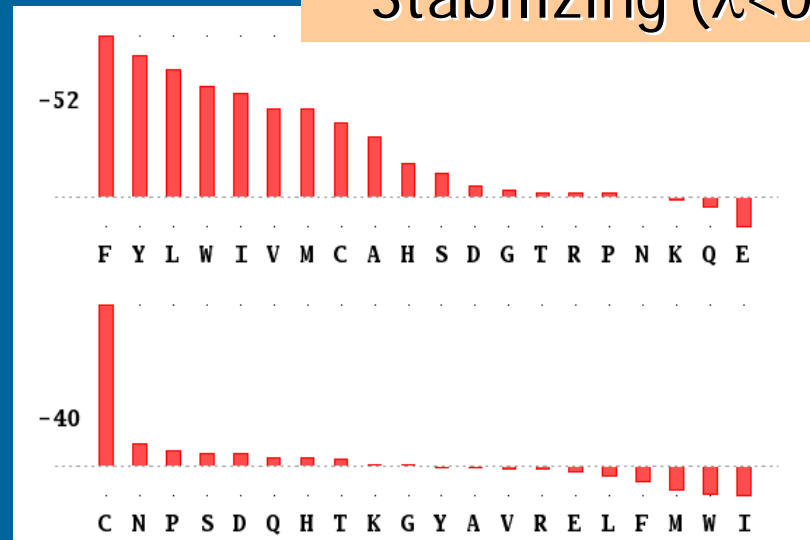
The end

Decomposition

$$E(\text{estimated}) / L = \sum_{ij} n_i P_{ij} n_j =$$

$$\lambda_1 (\mathbf{v}_1 \mathbf{n})^2 + \lambda_2 (\mathbf{v}_2 \mathbf{n})^2 + \dots + \lambda_{20} (\mathbf{v}_{20} \mathbf{n})^2$$

Stabilizing ($\lambda < 0$)



Destabilizing ($\lambda > 0$)

