



Evolution teaches predicting protein function



Burkhard Rost

CUBIC, NESG, C2B2

**Dept. Biochemistry & Mol. Biophysics
Columbia University**

<http://www.rostlab.org/>

Google "rost"



I. Introduction: protein function evolution

Protein function

Intuitive but not well-defined:

 chemical

how atom bound?

 biochemical

transferase

 cellular (kinase)

cell cycle

 developmental

time, regulatory

 physiological

related to disease

 genetic

dominant/recessive

Protein function as action:


Function =

anything that happens to or through a protein


Our goals

Predict protein function from sequence + structure

Where?

 nuclear/cytoplasmic/extra-cellular/mitochondrial/other, membrane/not/which, nuclear matrix, ER/Golgi/vesicle?

What?

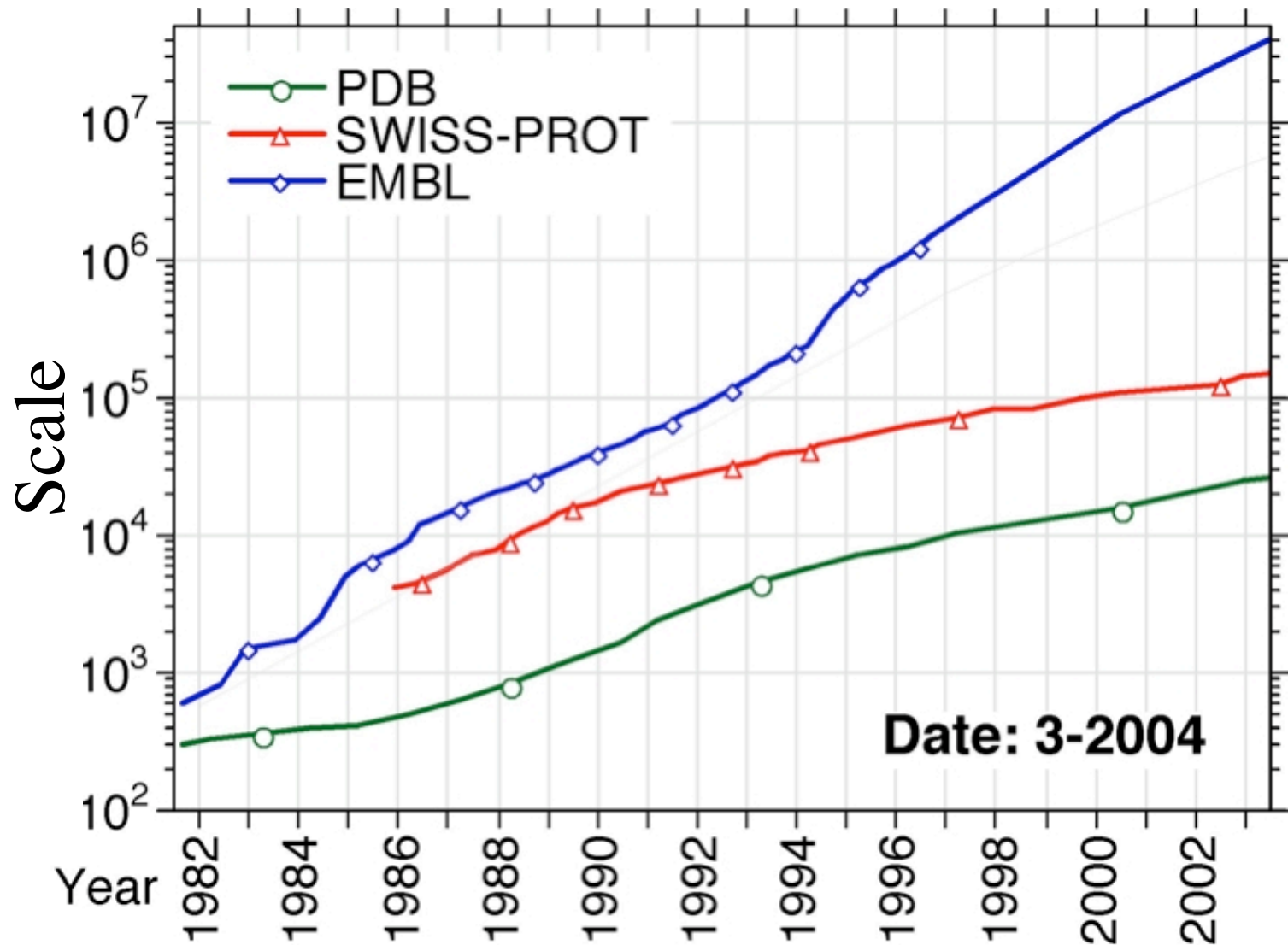
 protein-protein, protein-DNA, protein-small substrate, “is enzyme”, “is cell-cycle control protein”, “SNP deleterious?”

When?

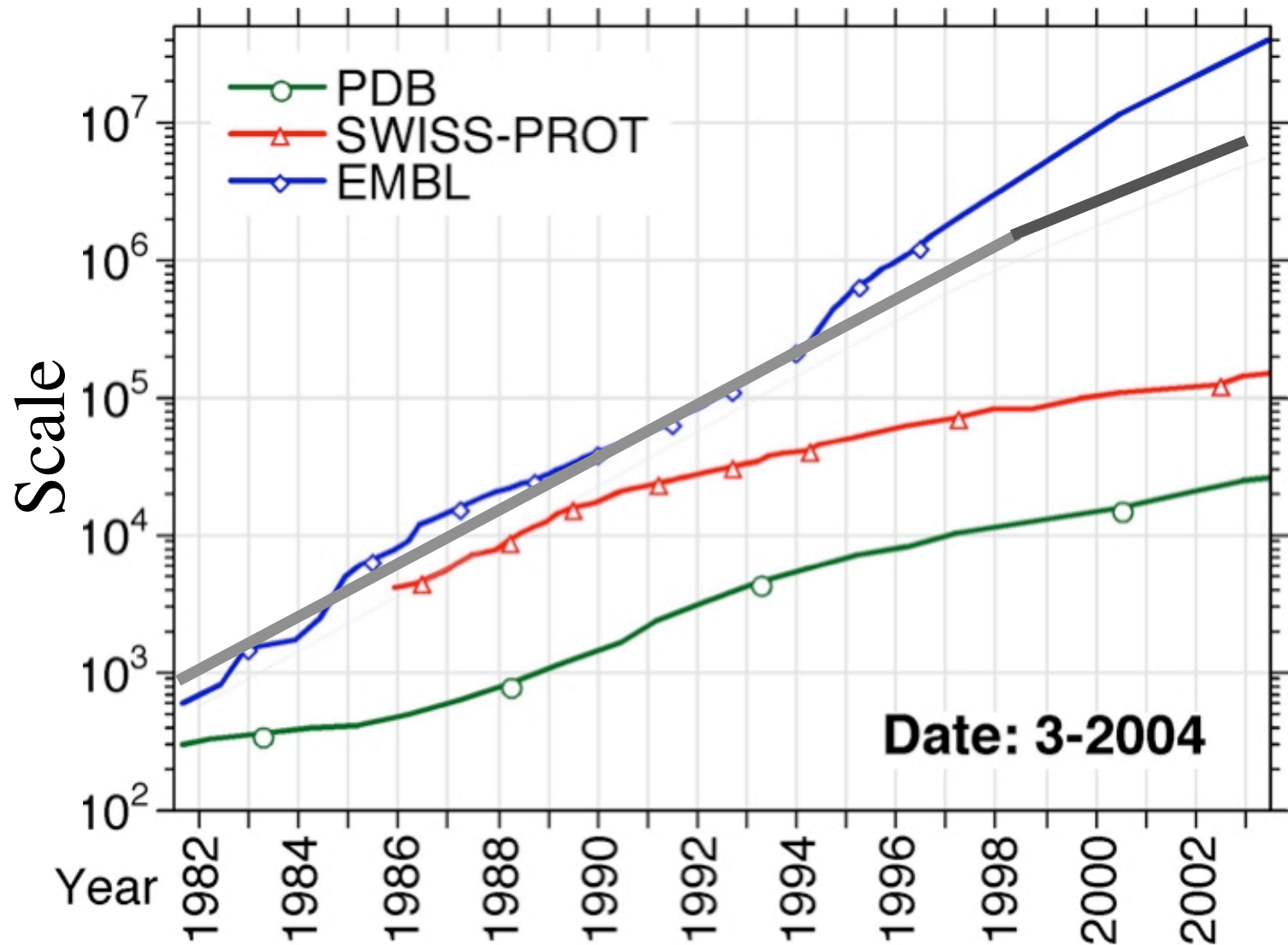
 pathways

Predict protein structure: focus on aspects relevant for function

Increasing wealth of experimental data!



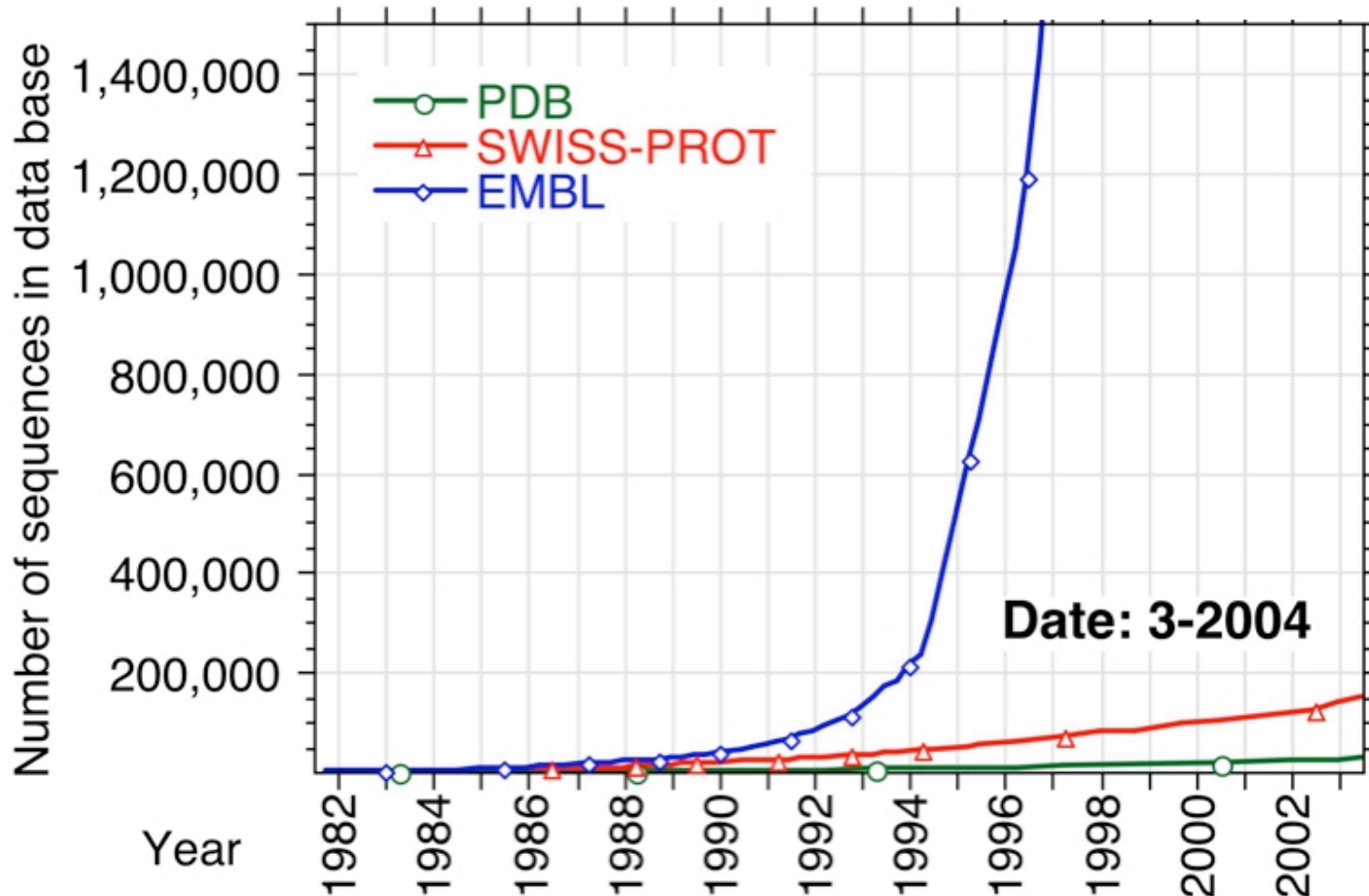
Increasing wealth of experimental data!



Gap sequence/annotation grows!

1.5 million protein sequences known today

>40 million
“gene”
sequences



Homology transfer accurate for very similar proteins

methyltransferase

identity	protein
100%	guanidinoacetate N-methyltransferase
99%	magnesium protoporphyrin IX synthase
70%	phosphoribosylglycinamide formyltransferase
65%	inositol 3-methyltransferase
65%	phosphoribosylglycinamide formyltransferase
63%	aspartate carbamoyltransferase
62%	glycine amidinotransferase
61%	inositol 3-methyltransferase

1				50	
fyn_human	VTLFVALYDY	EARTEDDLSF	HKGEKFQILN	SSEGDWWEAR	SLTTGETGYI
yrk_chick	VTLFIALYDY	EARTEDDLSF	QKGEKFHIIN	NTEGDWWEAR	SLSSGATGYI
fgr_human	VTLFIALYDY	EARTEDDLTF	TKGEKFHILN	NTEGDWWEAR	SLSSGKTGYI
yes_chick	VTVFVALYDY	EARTTDDLSF	KKGERFQIIN	NTEGDWWEAR	SIATGKTGYI
src_avis2	VTTFVALYDY	ESRTETDLSF	KKGERLQIVN	NTEGDWWEAR	SLTTGQTGYI
src_avis	VTTFVALYDY	ESRTETDLSF	KKGERLQIVN	NTEGDWWEAR	SLTTGQTGYI
src_avisr	VTTFVALYDY	ESRTETDLSF	KKGERLQIVN	NTEGDWWEAR	SLTTGQTGYI
src_chick	VTTFVALYDY	ESRTETDLSF	KKGERLQIVN	NTEGDWWEAR	SLTTGQTGYI
stk_hydat	VTIFVALYDY	EARISEDLSF	KKGERLQIIN	TADGDWWEAR	SLITNSEGYI
src_rsvpa	ESRIETDLSF	KKRERLQIVN	NTEGTWWEAR	SLTTGQTGYI
hck_human	..IVVALYDY	EAIHHEDLSF	QKGDQMVVLE	ES.GEWKAR	SLATRKEGYI
blk_mouse	..FVVALFDY	AAVNDRLQV	LKGEKIQVLR	.STGDWWEAR	SLVTGREGYV
hck_mouse	.TIVVALYDY	EAIHREDLSF	QKGDQMVVLE	.EAGEWWEAR	SLATKKEGYI
lyn_human	..IVVALYPY	DGIHPDDLFS	KKGEKMKVLE	.EHGEWWEAK	SLLTKKEGYI
lck_human	..LVIALHSY	EPSHDGDLGF	EKGEQIRILE	QS.GEWKAO	SLTTGQEGFI
ss81_yeastALYPY	DADDDeISF	EQNEILQVSD	.IEGRWWEAR	R.ANGETGYI
abl_mouse	..LFVALYDF	VASGDNTLSI	TKGEKIRVLG	YnnGEWCEAO	..TKNGQGWV
abl1_human	..LFVALYDF	VASGDNTLSI	TKGEKIRVLG	YnnGEWCEAO	..TKNGQGWV
src1_drome	..VVVLYDY	KSRDESLSF	MKGDRMEVID	DTESDWRVW	NLTTROEGLI
mysd_dicdiALYDF	DAESSMELSF	KEGDILTVDL	QSSGDWDAE	L..KGRRKV
yfj4_yeastVALYSF	AGEESGDLPF	RKGDVITILK	ksQNDWWTGR	V..NGREGF
abl2_human	..LFVALYDF	VASGDNTLSI	TKGEKIRVLG	YnnGEWWEAR	RSKNG.QGWV
tec_human	.EIVVAMYDF	QAAEGHDLRL	ERGQEYLILE	KNDVHWWRAR	D.KYNGEYI
abl1_caeel	..LFVALYDF	HGVGEEQLSL	RKGDQVRILG	YNKNNEWCEA	RlrLGEIGWV
txk_humanALYDF	LPREPCNLAL	RRAEYLILE	KYNPHWWEAR	D.RLNGEGLI
yha2_yeast	VRRVVALYDL	TTNEPDELSF	RKGDVITVLE	QVYRDWWEAR	L..RGNMGF
abp1_sacexAEYDY	EAGEDNELTF	AENDKIINIE	FVDDDWLGE	LETTGQKGLF

Homology transfer accurate for very similar proteins

methyltransferase

TRUE

FALSE

identity protein

100% **guanidinoacetate N-methyltransferase**

99% **magnesium protoporphyrin IX methyltransferase**

70% **phosphoribosylglycinamide formyltransferase**

65% **inositol 3-methyltransferase**

65% **phosphoribosylglycinamide formyltransferase**

63% **aspartate carbamoyltransferase**

62% **glycine amidinotransferase**

61% **inositol 3-methyltransferase**

Homology transfer accurate for very similar proteins

methyltransferase

TRUE

FALSE

identity protein

100% guanidinoacetate N-methyltransferase

99% magnesium protoporphyrin IX methyltransferase

~~70% phosphoribosylglycinamide formyltransferase~~

2/3 accuracy ; 2/4 coverage

65% inositol 3-methyltransferase

65% phosphoribosylglycinamide formyltransferase

63% aspartate carbamoyltransferase

62% glycine amidinotransferase

61% inositol 3-methyltransferase

Homology transfer accurate for very similar proteins

methyltransferase

TRUE

FALSE

identity protein

100% guanidinoacetate N-methyltransferase

99% magnesium protoporphyrin IX methyltransferase

70% **phosphoribosylglycinamide formyltransferase**

2/3 accuracy ; 2/4 coverage

65% inositol 3-methyltransferase

65% **phosphoribosylglycinamide formyltransferase**

63% **aspartate carbamoyltransferase**

62% **glycine amidinotransferase**

61% inositol 3-methyltransferase

3/8 accuracy ; 4/4 coverage

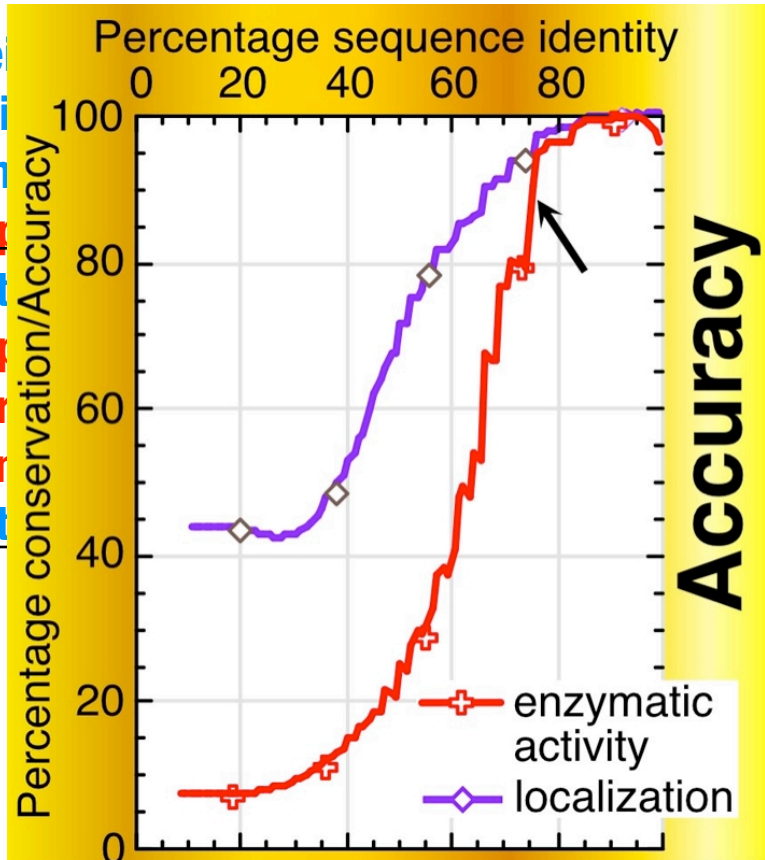
Homology transfer accurate for very similar proteins

methyltransferase

TRUE

FALSE

- identity prote
- 100% guan
- 99% magn
- 70% phosph
- 65% inosit
- 65% phosph
- 63% aspar
- 62% glycin
- 61% inosit



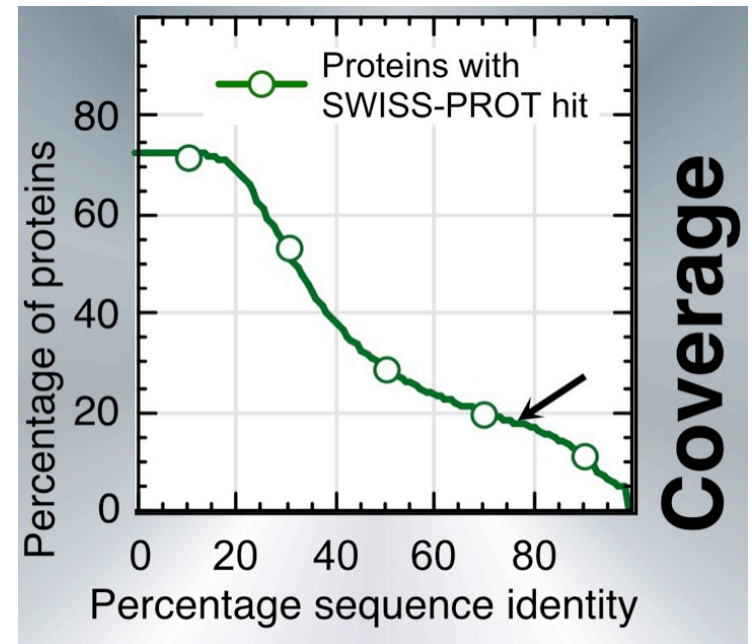
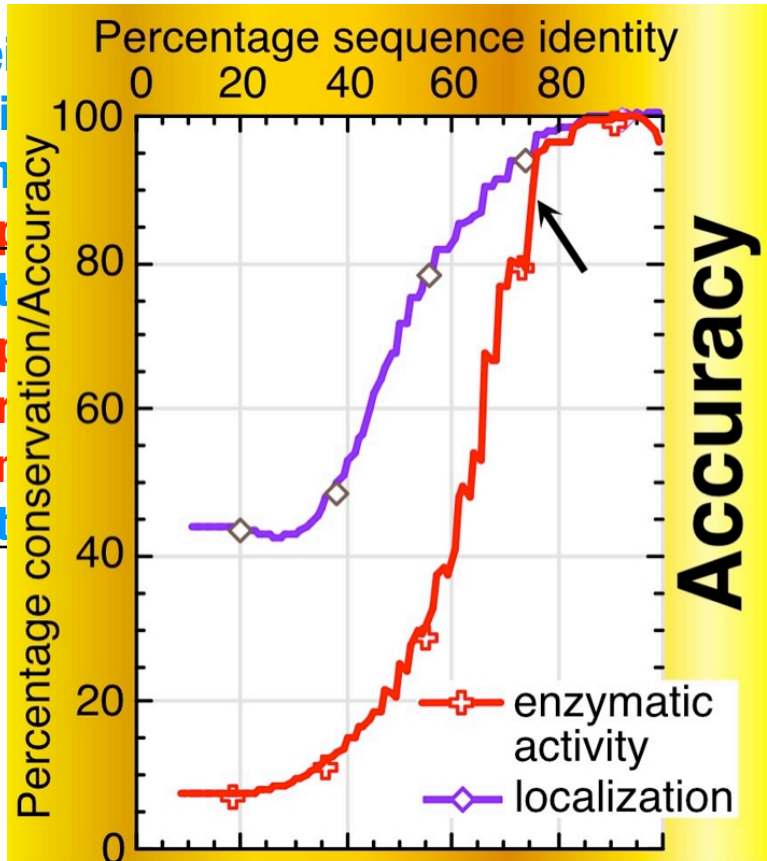
Homology transfer accurate for very similar proteins

methyltransferase

TRUE

FALSE

- identity prote
- 100% guan
- 99% magn
- 70% phosph
- 65% inosit
- 65% phosph
- 63% aspar
- 62% glycin
- 61% inosit



Some problems of homology transfer

not all annotations as informative as “methyltransferase”

ID 1433_TRIHA STANDARD; PRT; 262 AA.

DE 14-3-3 PROTEIN HOMOLOG (TH1433).

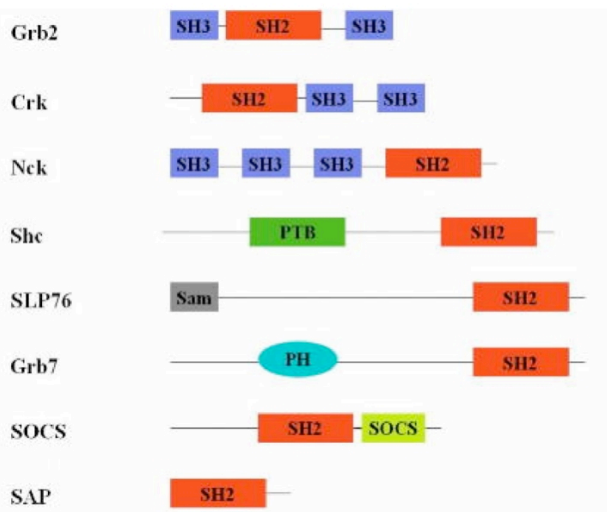
CC -!- DEVELOPMENTAL STAGE: HIGHEST EXPRESSION DURING THE ACTIVE GROWTH

CC PERIOD 10-12 HOURS AFTER GERMINATION.

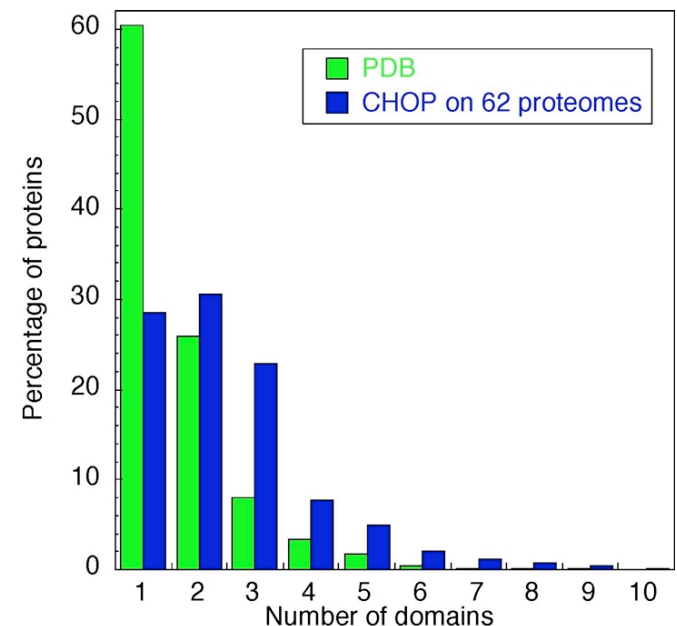
CC -!- SIMILARITY: BELONGS TO THE 14-3-3 FAMILY.

70% multi-domain proteins

adaptors and regulatory proteins



Schlessinger unpublished



Liu & Rost 2004 Proteins 55:678-686

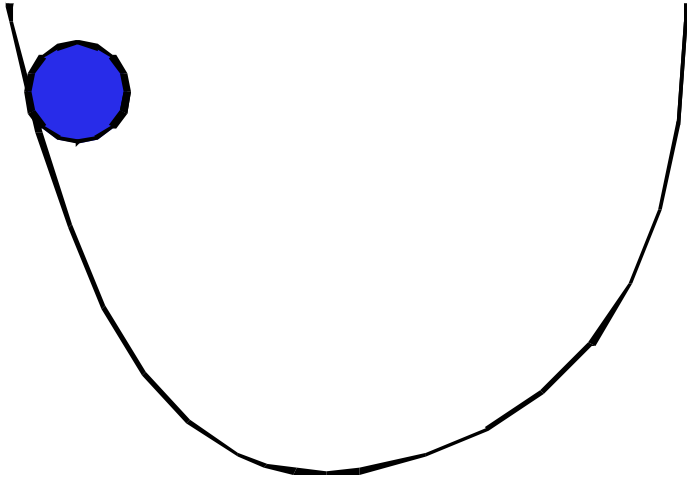
Less than 25% have *some* annotation

coverage of homology transfer

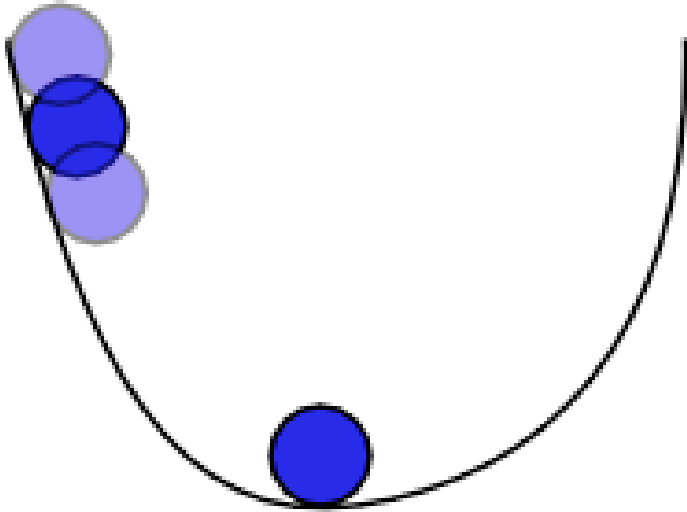
< 10-25%

we clearly need something more!

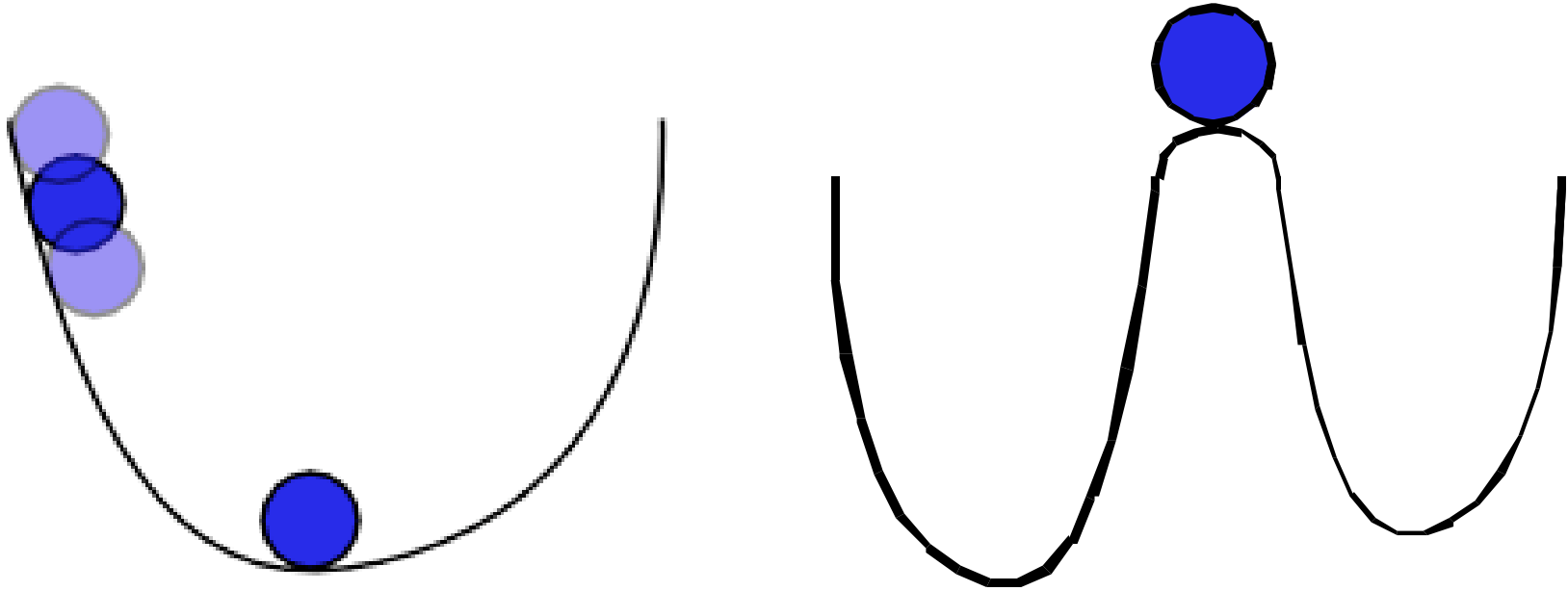
Prediction in terms of energy landscapes



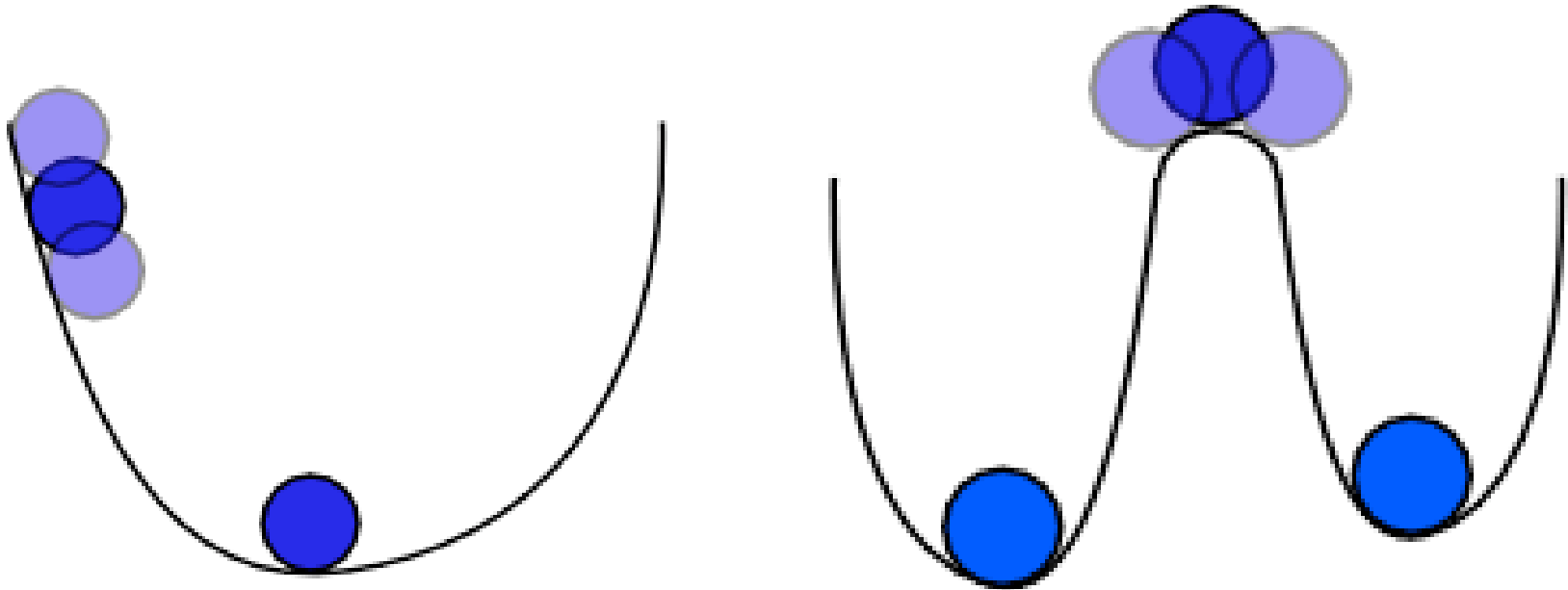
Prediction in terms of energy landscapes



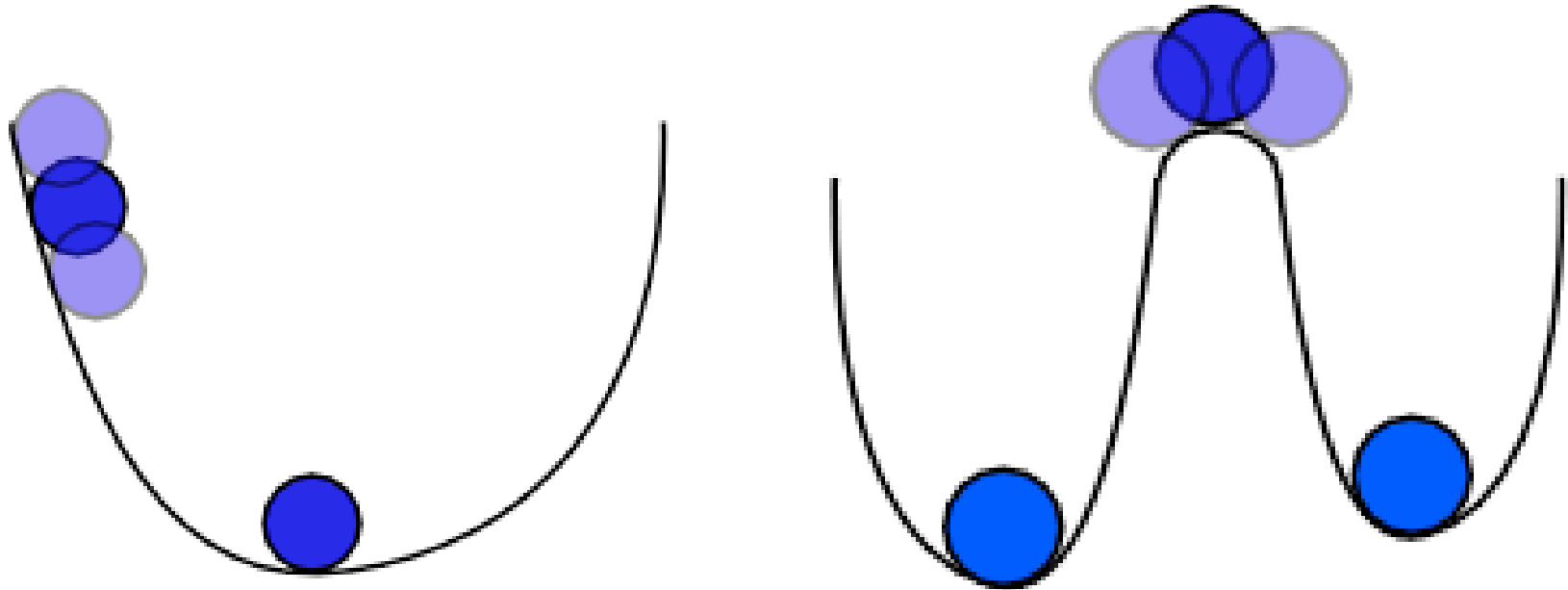
Prediction in terms of energy landscapes



Prediction in terms of energy landscapes



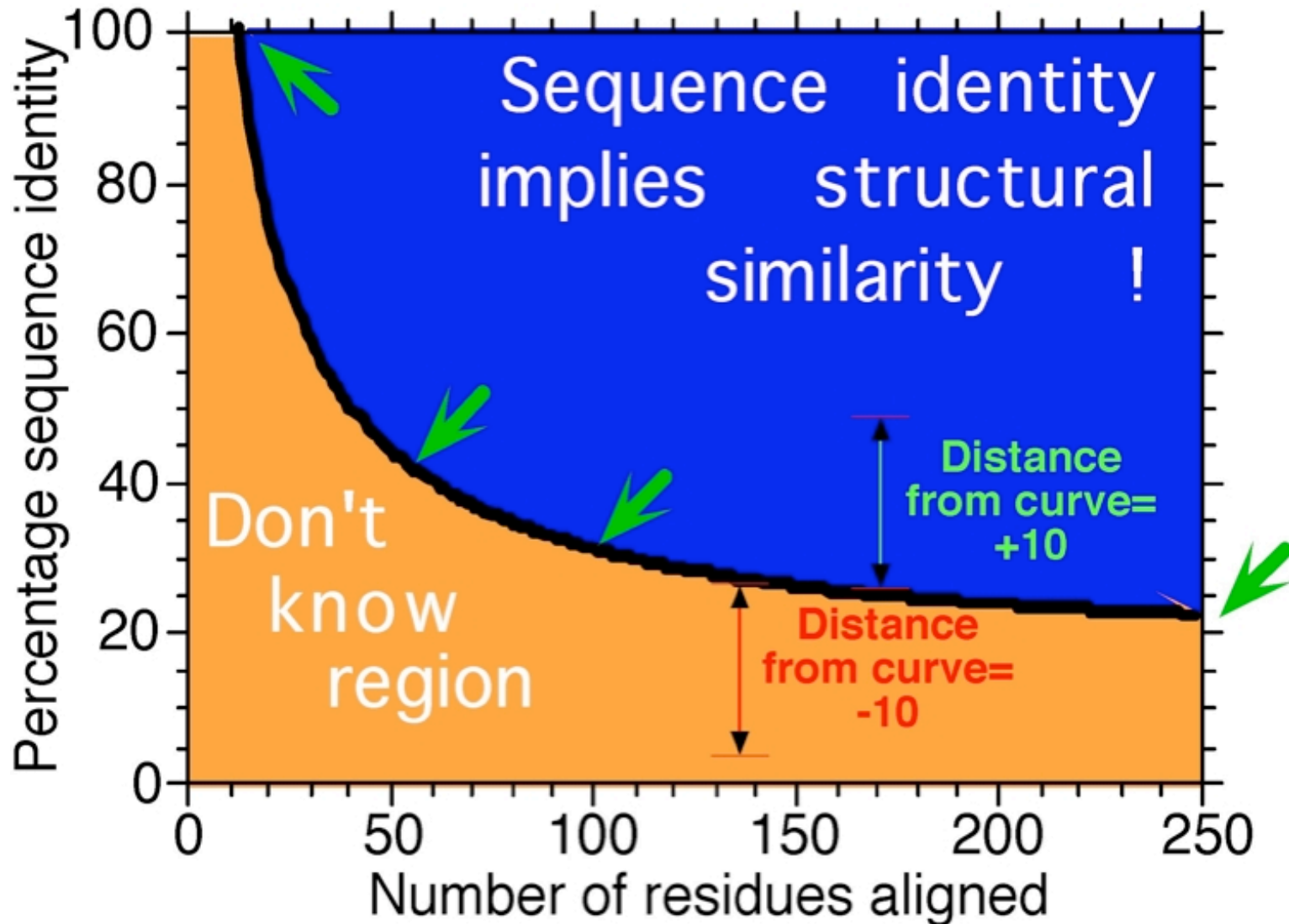
Prediction in terms of energy landscapes



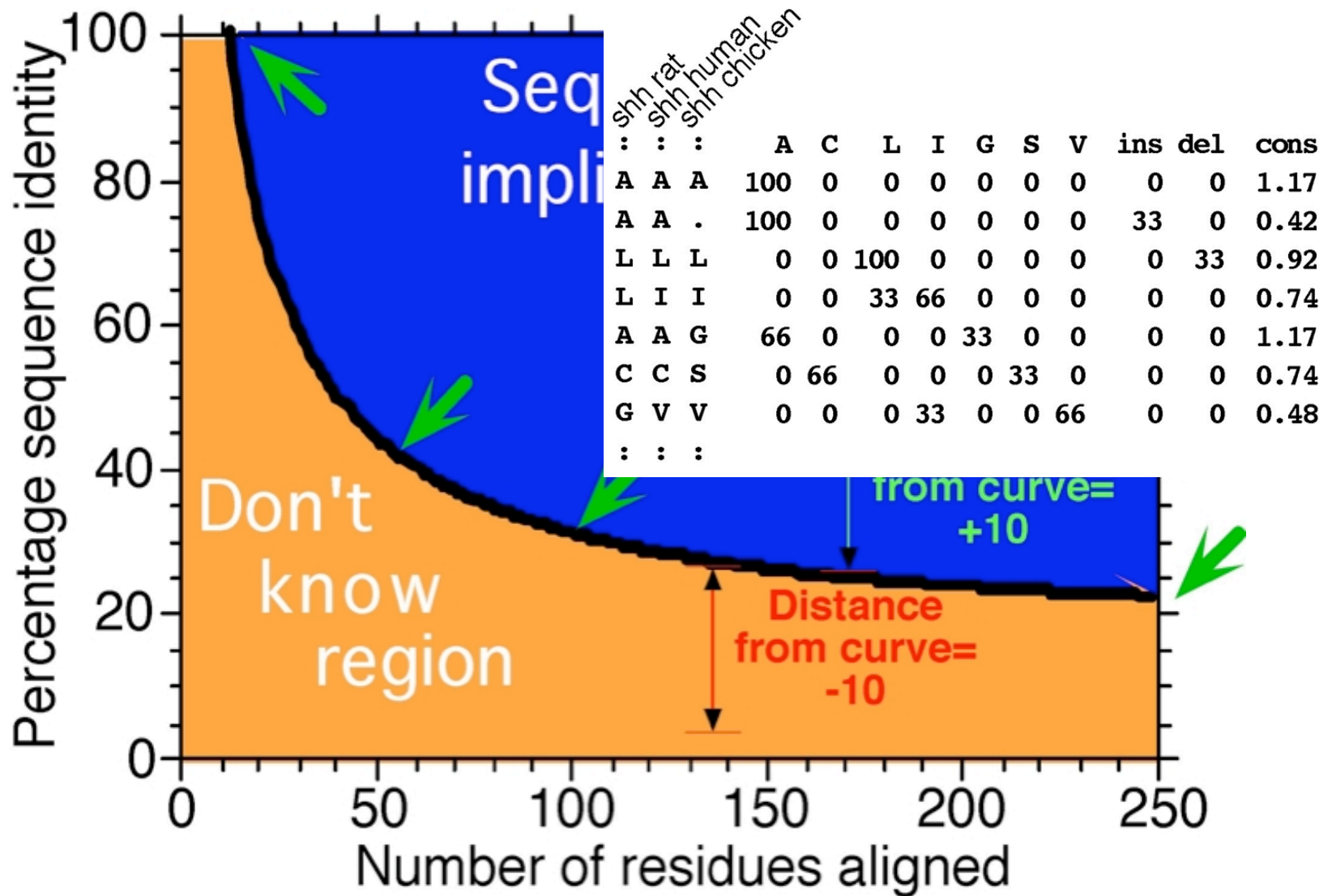
- Point mutation
- Binding (Substrate/Protein)
- Environmental change (DNA close/pH)

Need to know history to predict!

Evolution is history!



Evolution is history!

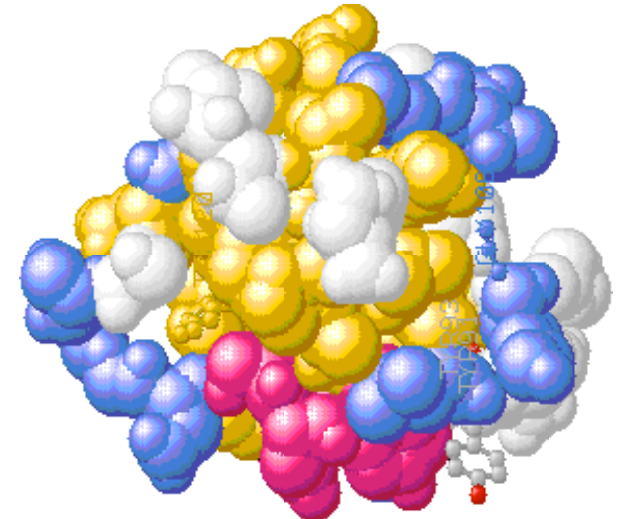
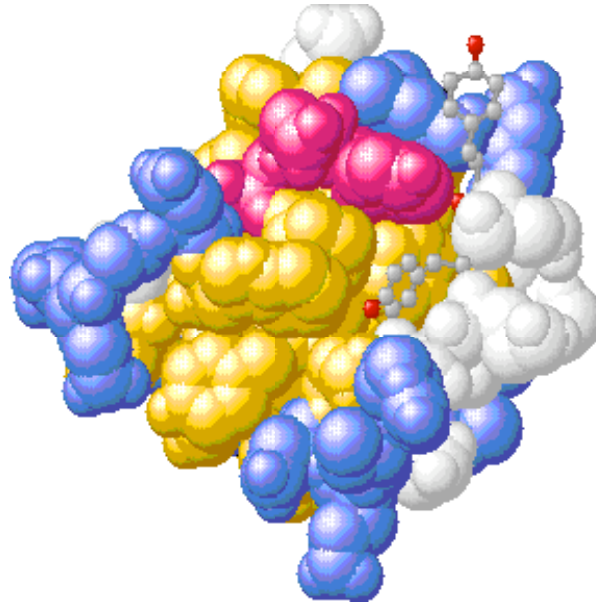


SH3

Src-homology 3 domain

one domain of proteins such as
Src tyrosine kinase (STK)

```
1 50  
fyn_human VTLFVALYDY EARTEDDL SF HKGEKFI LN SSEGDWWEAR SLTTGETGI  
yrk_chick VTLFIALYDY EARTEDDL SF QKGEKFI LN NTEGDWWEAR SLSSGATGI  
fgr_human VTLFIALYDY EARTEDDL TF TKGEKFI LN NTEGDWWEAR SLSSGKTGI  
yes_chick VTVFVALYDY EARTD DDL SF KKGERFI LN NTEGDWWEAR SIATGKTGI  
src_avis2 VTTFVALYDY ESRTETDL SF KKGERLI QVN NTEGDWLAH SLTTGQTGI  
src_avis VTTFVALYDY ESRTETDL SF KKGERLI QVN NTEGDWLAH SLTTGQTGI  
src_avisr VTTFVALYDY ESRTETDL SF KKGERLI QVN NTEGDWLAH SLTTGQTGI  
src_chick VTTFVALYDY ESRTETDL SF KKGERLI QVN NTEGDWLAH SLTTGQTGI  
stk_hydat VTFIVALYDY EARISEDLSF KKGERLI QIN TADGDWWEAR SLITNSEGI  
src_rsvpa ..... ESRIETDL SF KKRERLI QVN NTEGTWLAH SLTTGQTGI  
hck_human .IVVALYDY EAIHHEDLSF QKGDQMVLE ES.GEWKAR SLATRKEGI  
blk_mouse .FVVALFDY AAVNDRDLQV LKGEKLI QVLR .STGDWWEAR SLVTGREGV  
hck_mouse .TIVVALYDY EAIHREDLSF QKGDQMVLE .EAGEWWEAR SLATKKEGI  
lyn_human .IVVALYDY DGIHPDDL SF KKGEKMI VLE .EHGEWWEAR SLLTKKEGI  
lck_human .LVIALHSY EPSHDGDLGF EKGEQRI LE QS.GEWKAO SLTTGQEGFI  
ss81_yeast...ALYPY DADDDeISF EQNEILQVSD .IEGRWWEAR R.ANGETGI  
abl_mouse .LFVALYDF VASGDNTLSI TKGEKIRVLG YnnGEWCEAQ ..TKNGQGV  
abl_human..LFVALYDF VASGDNTLSI TKGEKIRVLG YnnGEWCEAQ ..TKNGQGV  
src1_drome..VVWLYDY KSRDESDFS MKGDRMEVID DTESDWRVV NLTRQEGFI  
mysd_dicdi...ALYDF DAESSMELSF KEGDILTVDL QSSGDWDAE L..KGRRKVI  
yjf4_yeast...VLYSF AGEESGDLFP RKGDVITILK ksQNDWWTGR V..NGREGF  
abl_human..LFVALYDF VASGDNTLSI TKGEKIRVLG YNNGEWSEV RSKNG.QGV  
tec_human .EIVVAMYDF QAAEGHDLRL ERGQYLLILE KNDVHWWRAR D.KYNGEGI  
abl1_caee1..LFVALYDF HGVGEEQLSL RKGDVIRILG YNKNEWCEA RlrLGEIGV  
txk_human ....ALYDF LPREPCNLAL RRAEYLLILE KYNPHWWEAR D.RLNGEGLI  
yha2_yeastVRRVVALYDL TTNEPDELSF RKGDVITVLE QVYRDWKGGA L..RGNMGF  
abp1_sacex...AEYDY EAGEDNELTF AENDKIINIE FVDDDWLGE LETTGQKGLF
```

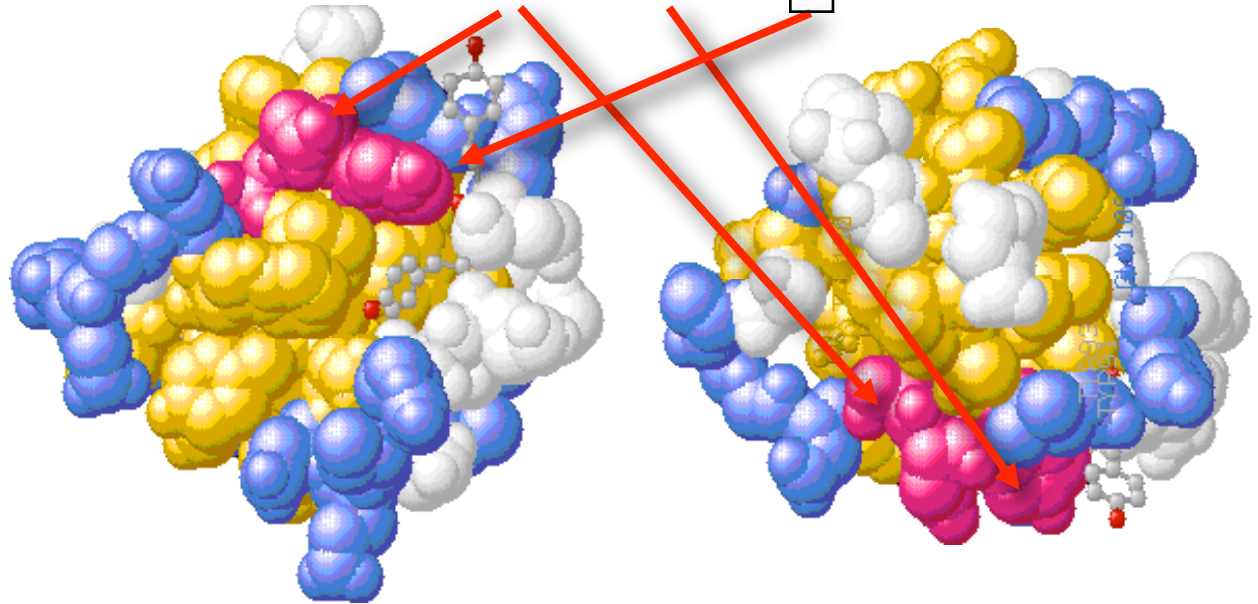


SH3

Src-homology 3 domain

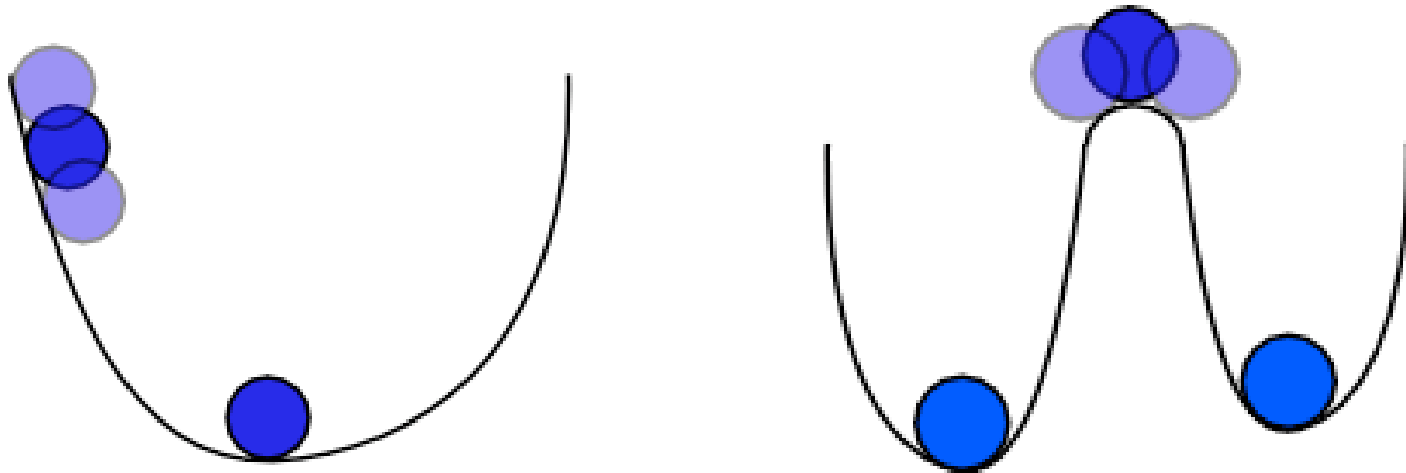
one domain of proteins such as
Src tyrosine kinase (STK)

```
1 50  
fyn_human VTLFVALYDY EARTEDDLSF HKGEKFOILN SSEGDWWEAR SLTTGTEGFI  
yrk_chick VTLFIALYDY EARTEDDLSF QKGEKFHIIN NTEGDWWEAR SLSSGATGFI  
fgr_human VTLFIALYDY EARTEDDLTF TKGEKFHILN NTEGDWWEAR SLSSGKTCFI  
yes_chick VTVFVALYDY EARTTDDLSF KKGERFOIIN NTEGDWWEAR SIATGKTGFI  
src_avis2 VTTFVALYDY ESRTETDLSF KKGERLIQVN NTEGDWLAH SLTTGQTGFI  
src_avis3 VTTFVALYDY ESRTETDLSF KKGERLIQVN NTEGDWLAH SLTTGQTGFI  
src_avisr VTTFVALYDY ESRTETDLSF KKGERLIQVN NTEGDWLAH SLTTGQTGFI  
src_chick VTTFVALYDY ESRTETDLSF KKGERLIQVN NTEGDWLAH SLTTGQTGFI  
stk_hydat VTFIVALYDY EARISEDLSF KKGERLIQIN TADGDWWEAR SLITNSEGFI  
src_rsvpa ..... ESRIETDLSF KKRERLIQVN NTEGTWLAH SLTTGQTGFI  
hck_human .IVVALYDY EAIHHEDLSF QKGDQMVLE .ESGEWKKAR SLATRKEGFI  
blk_mouse .FVVALFDY AAVNDRDLQV LKGEKIQVLR .STGDWWEAR SLVTGREGVFI  
hck_mouse .TIVVALYDY EAIHREDLSF QKGDQMVLE .EAGEWKKAR SLATKKEGFI  
lyn_human .IVVALYDY DGIHPDDLFSF KKGEKMKVLE .EHGEWKKAK SLLTKKEGFI  
lck_human .LVIALHSY EPSHDGDLGF EKGEQIRILE QS.GEWWKAO SLTTGQEGFI  
ss81_yeast...ALYPY DADDDeISF EQNEILQVSD .IEGRWKKAR R.ANGETGFI  
abl_mouse .LFVALYDF VASGDNTLSI TKGEKIRVLG YnnGEWCEAQ ..TKNGQGVFI  
abl1_human..LFVALYDF VASGDNTLSI TKGEKIRVLG YnnGEWCEAQ ..TKNGQGVFI  
src1_drome..VVWLYDY KSRDESDFSF MKGDRMEVID DTEGDWRVV NLTRQEGFI  
mysd_dicdi...ALYDF DAESSMELSF KEGDILTFLD QSSGDWDAE L..KGRRGVFI  
yjf4_yeast...VLYSF AGEESGDLFP RKGDVITILK ksQNDWWTGR V..NGREGVFI  
abl2_human..LFVALYDF VASGDNTLSI TKGEKIRVLG YNQGEWSEV RSKNG.QGVFI  
tec_human .EIVVAMYDF QAAEGHDLRL ERGQEYLILE KNDVHWKAR D.KYNGEGFI  
abl1_caee1..LFVALYDF HGVGEEQLSL RKGQVRIILG YNKVNEWCEA RlrLGEIGVFI  
txk_human ....ALYDF LPREPCNLAL RRAEYLLILE KYNPHWKKAR D.RLNGEGFI  
yha2_yeastVRRVVALYDL TTNEPDELSF RKGDVITVLE QVYEDWKKGA L..RGNMGVFI  
abp1_sacex...AEYDY EAGEDNELTF AENDKIINIE FVDDWMLGE LETTGQKGLFI
```



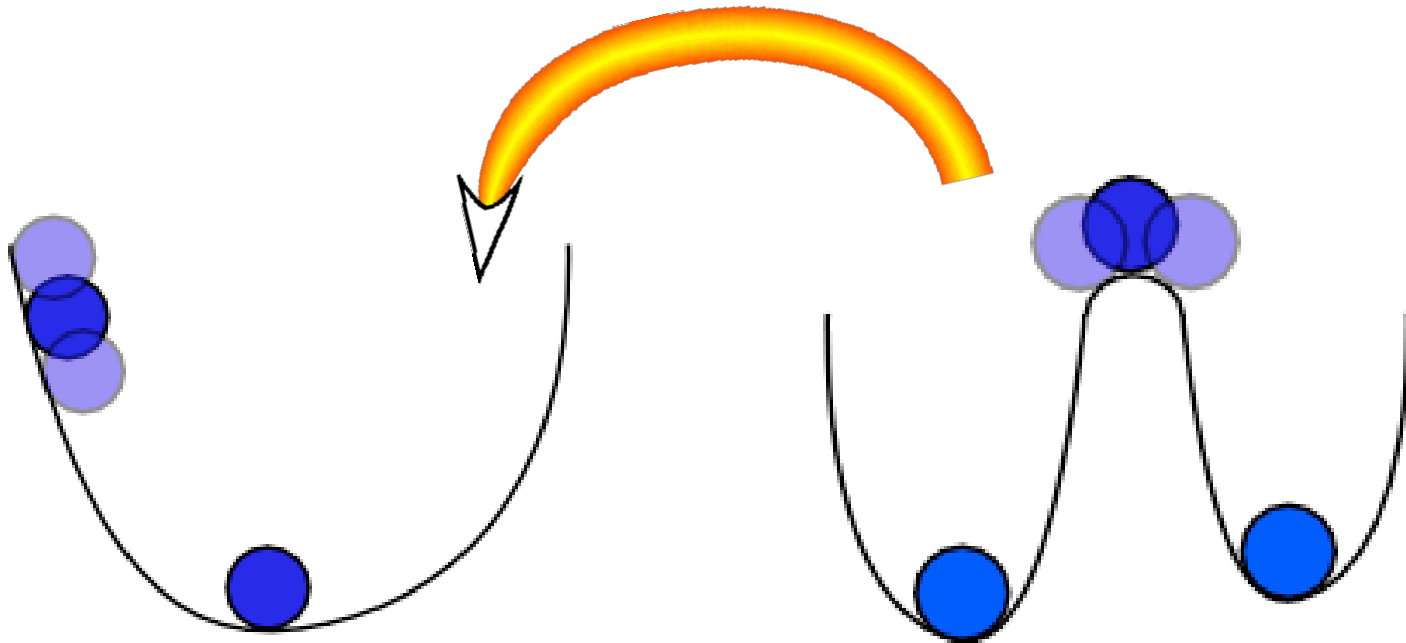
Evolution improves prediction

Evolutionary profile implicitly captures history of and individual protein!



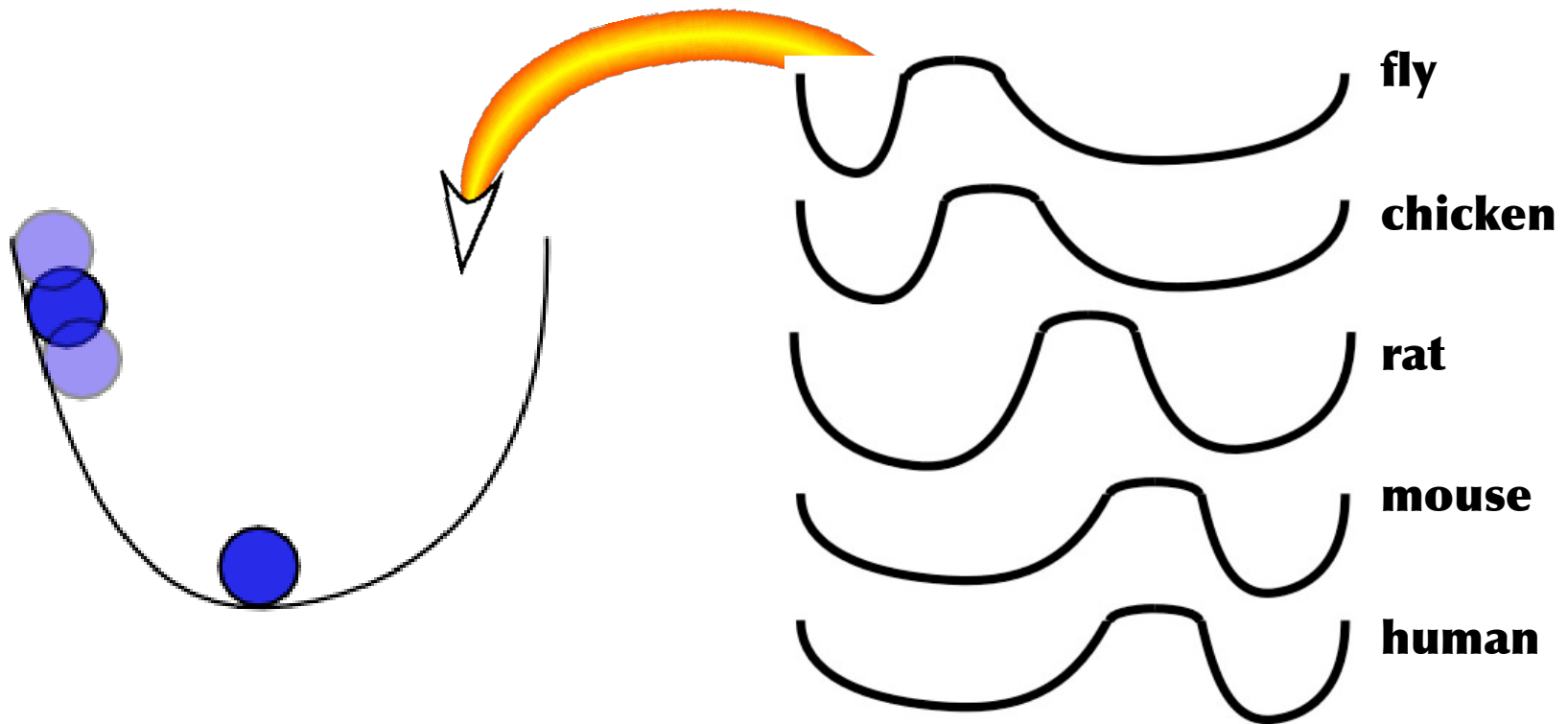
Evolution improves prediction

Evolutionary profile implicitly captures history of and individual protein!



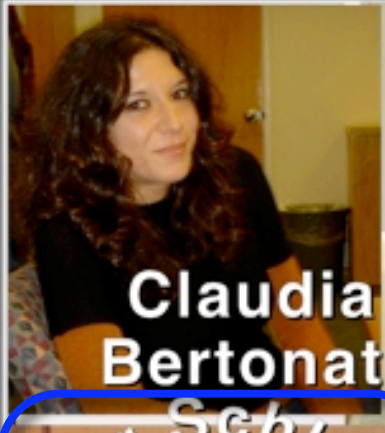
Evolution improves prediction

Evolutionary profile implicitly captures history of and individual protein!



II. Focus:

Predict physical
protein-protein
interactions



Claudia Bertoni



Henry Bigelow

Kaz Wrzeszczynski



Yanay Ofran

Ta-Tsen Soong

Yanay



Avner Schlessinger

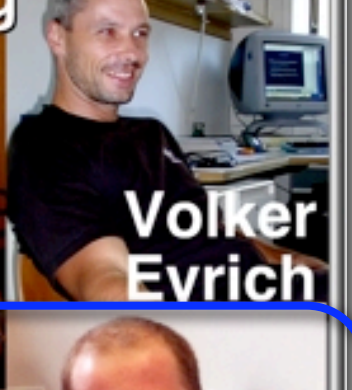


Jinfeng Liu

Ingrid Kohl Bromberg



Sara Gilman



Volker Eyrich



Marco Punta

Eyal Mozes



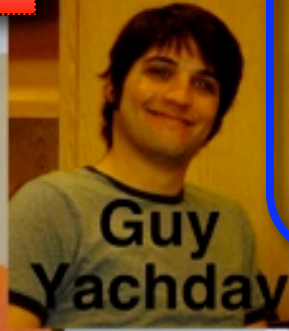
Darek Przybylski



Raj Nair



Andrew Kernytsky



Guy Yachday



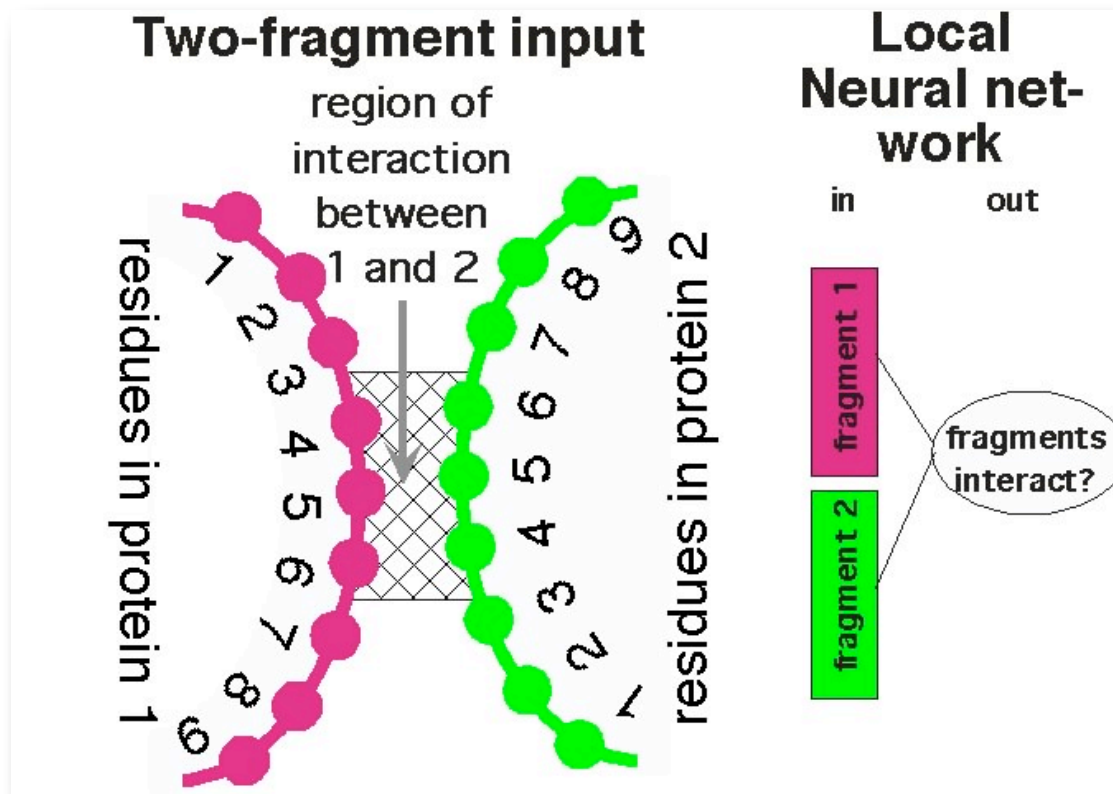
Sven Mika



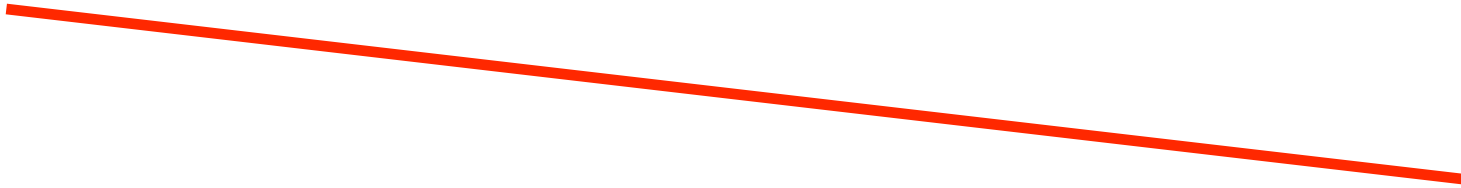
Phil Carter

1999: Want to predict protein-protein partners

- Implement simple method to do this
failed entirely: too many false positives



1999: Want to predict protein-protein partners




1999: Want to predict protein-protein partners

- ~~Implement simple method to do this
failed entirely: too many false positives~~

1999: Want to predict protein-protein partners

- ~~● Implement simple method to do this
failed entirely: too many false positives~~
- Reduce false positives:

1999: Want to predict protein-protein partners

 ~~Implement simple method to do this
failed entirely: too many false positives~~

 Reduce false positives:

 predict surface residues (PROFacc, 1999)
note: 1/2 of residues -> 1/4 of false positives!

1999: Want to predict protein-protein partners

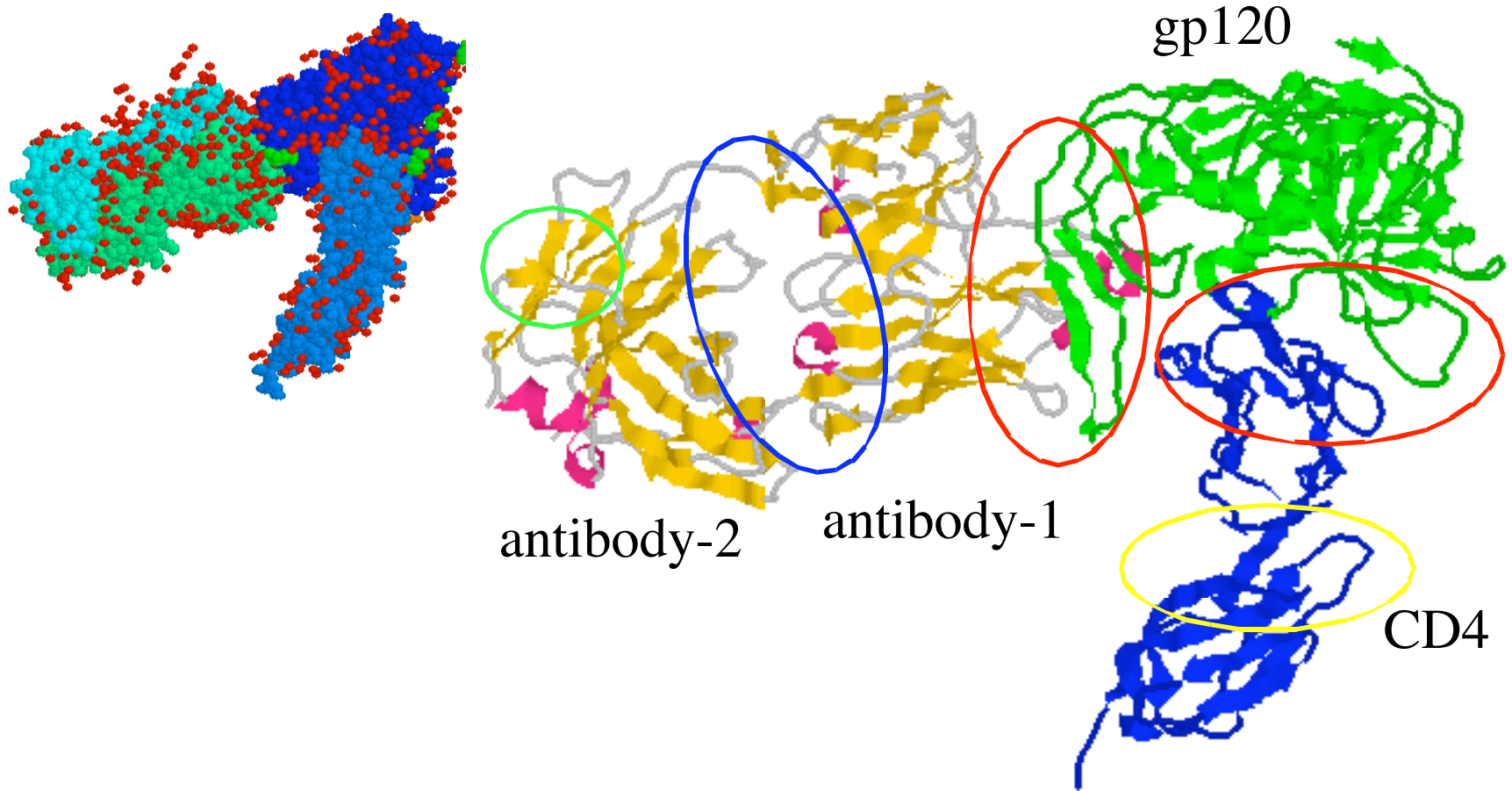
 ~~Implement simple method to do this
failed entirely: too many false positives~~

 Reduce false positives:

predict surface residues (PROFacc, 1999)
note: 1/2 of residues -> 1/4 of false positives!

predict residues in external interfaces (ISIS, 2004)

Different interfaces = different physics?

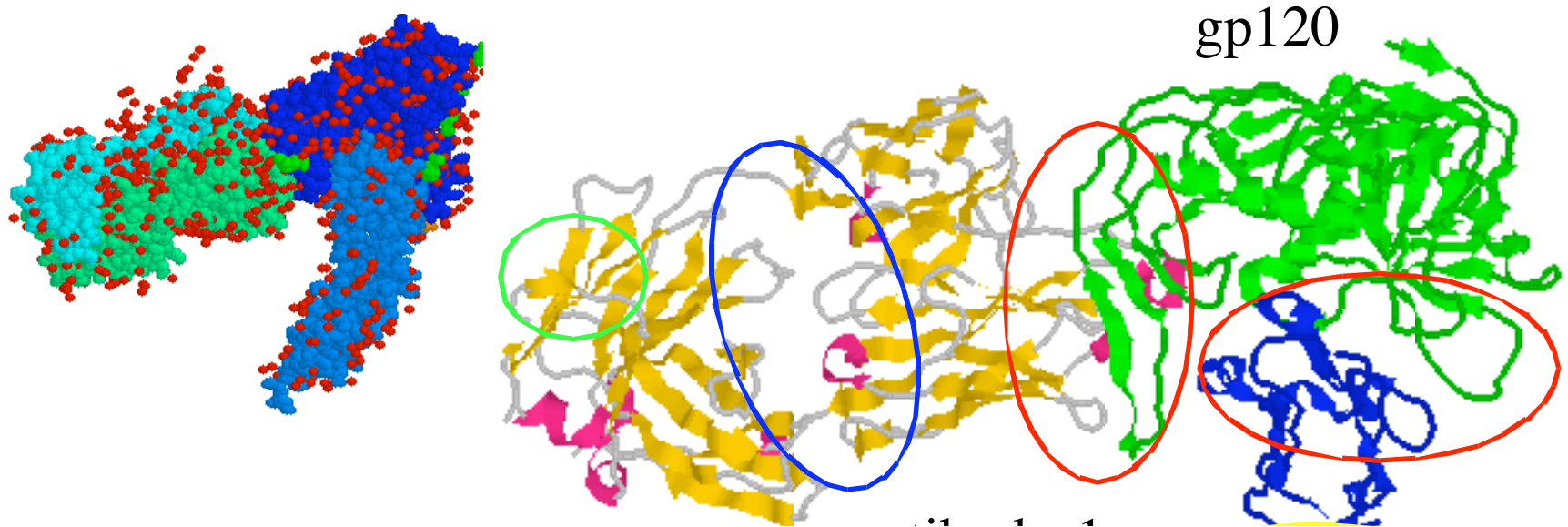


HIV gp120 / CD4 / FAB

PD Kwong, R Wyatt, J Robinson, RW Sweet, J Sodroski & WA Hendrickson (1998) *Nature* **393**, 648-659.

PD Kwong, R Wyatt, S Majeed, J Robinson, RW Sweet, J Sodroski & WA Hendrickson (2000) *Structure* **8**, 1329-1339.

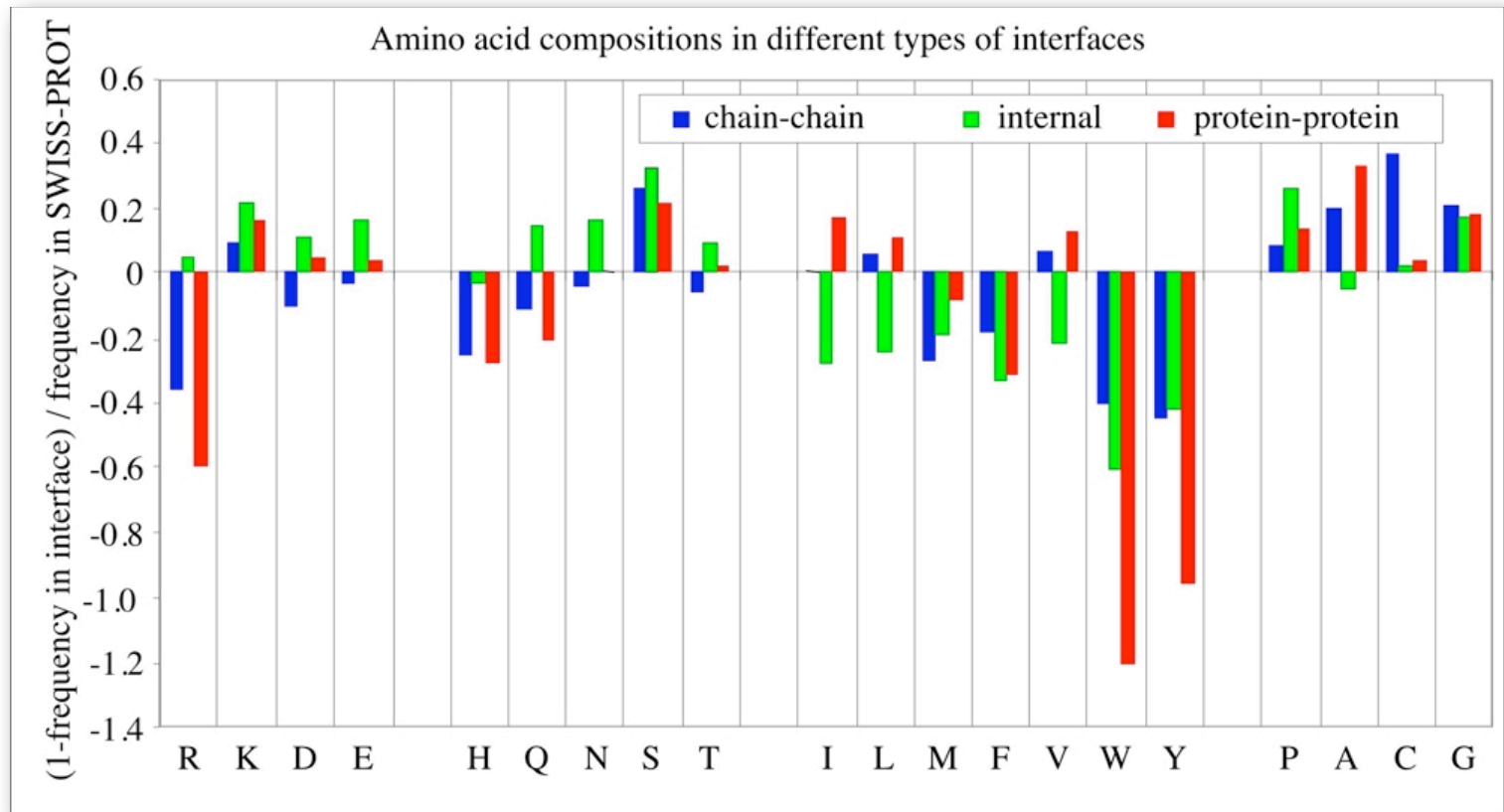
Different interfaces = different physics?



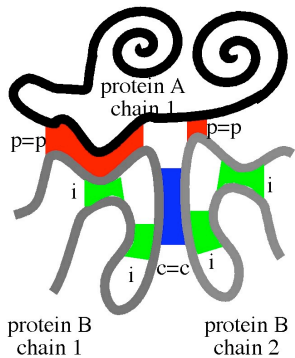
At least 6 types of interfaces differ in sequence!

Internal (inter-domain and intra-domain)
External homomers (permanent/transient)
External heteromers (permanent/transient)

Interface types differ in composition



3 interface types:
■ internal
■ chain-chain
■ protein-protein



Are these differences statistically significant?

Are these differences statistically significant?

Chi-square test:

-  known problem: small data sets
-  here millions of points

Are these differences statistically significant?

Chi-square test:

-  known problem: small data sets
-  here millions of points


 all differences $< 10^{-300}$
-> SIGNIFICANT

Are these differences statistically significant?

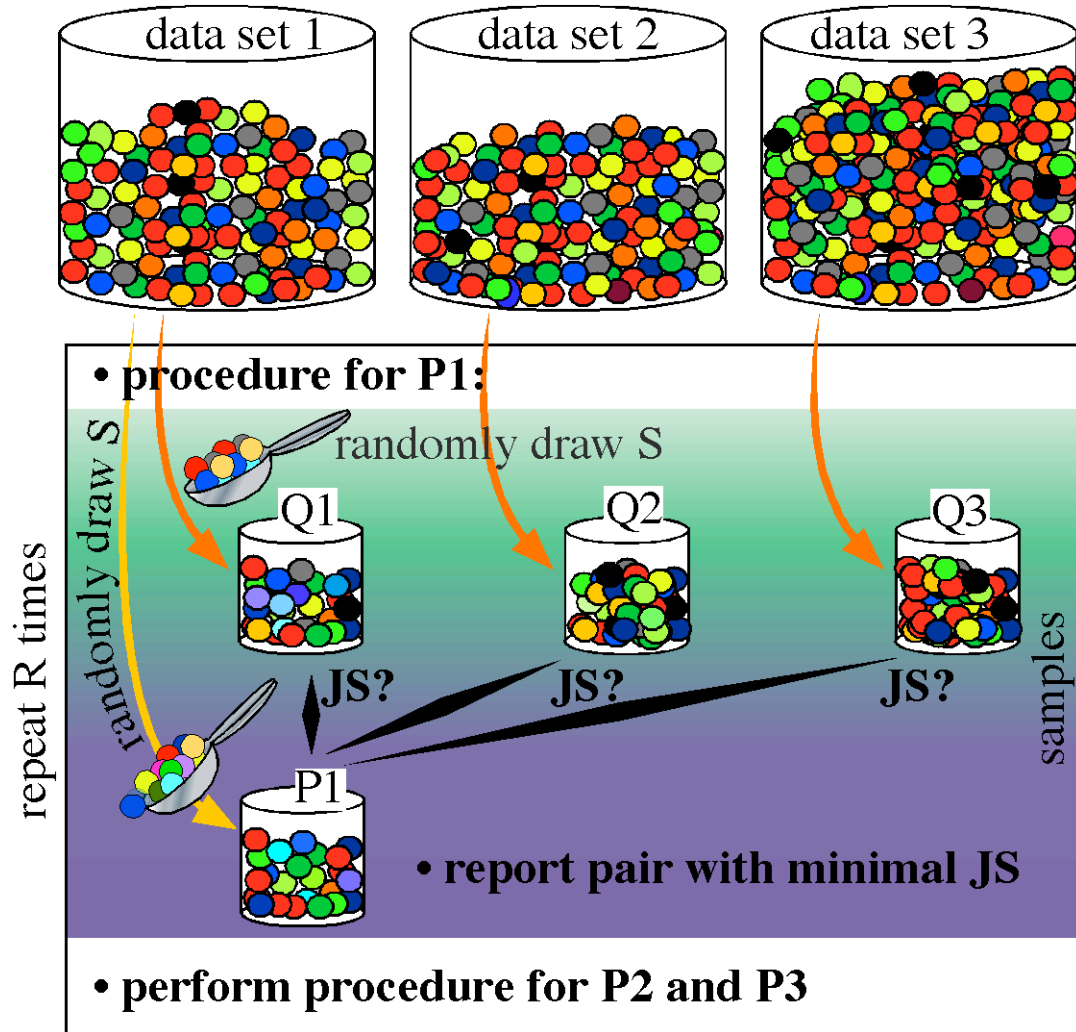
Chi-square test:

-  known problem: small data sets
-  here millions of points













 all differences $< 10^{-300}$
-> SIGNIFICANT

 ... unfortunately also:
proteins [a-b] vs [c-d]
1 vs 2 authors
random subsets ...

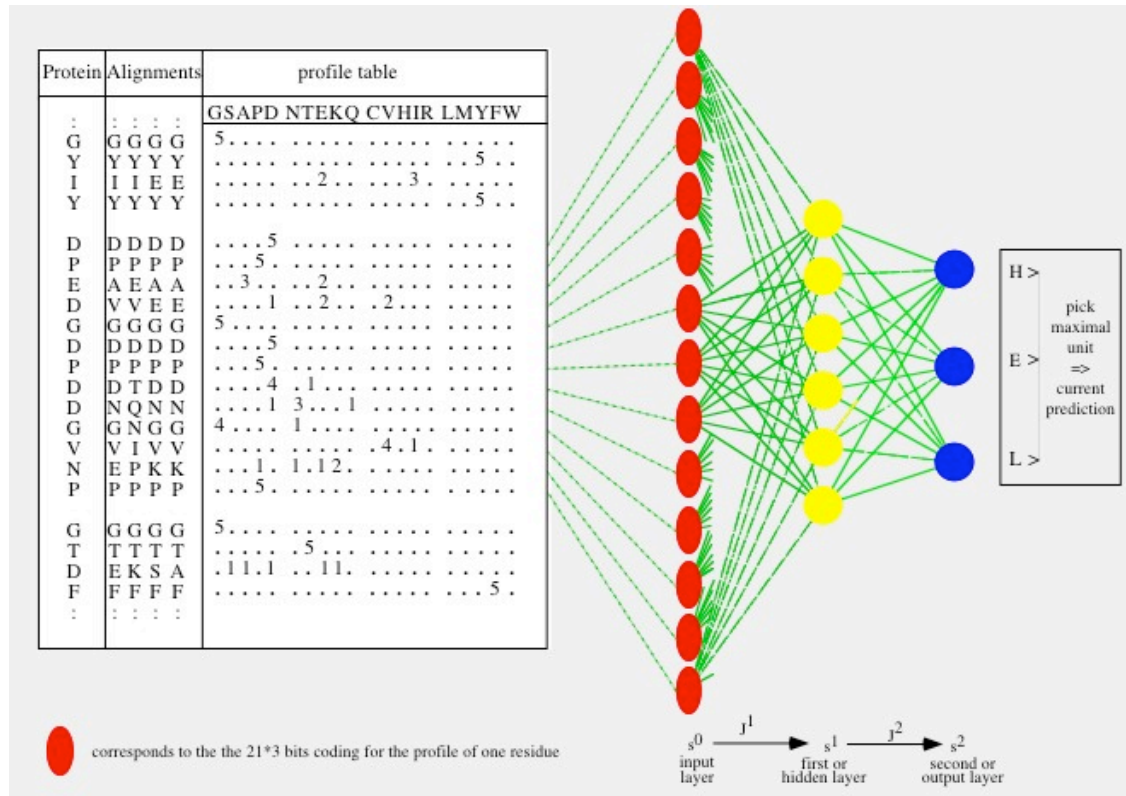
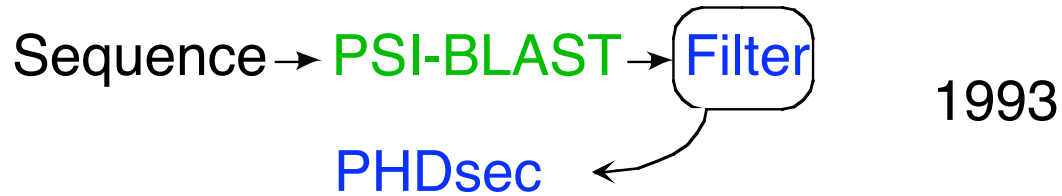
Find-self test (statistical significance)



Find-self test on six types of interfaces

							
internal		909	65	3	-	21	2
domain-domain		53	812	2	-	128	5
homo-obligomer		-	2	925	-	12	61
homo-oligomer		-	-	-	1000	-	-
hetero-obligomer		18	130	7	-	811	34
hetero-oligomer		-	8	58	-	38	896

Using evolution to predict structure



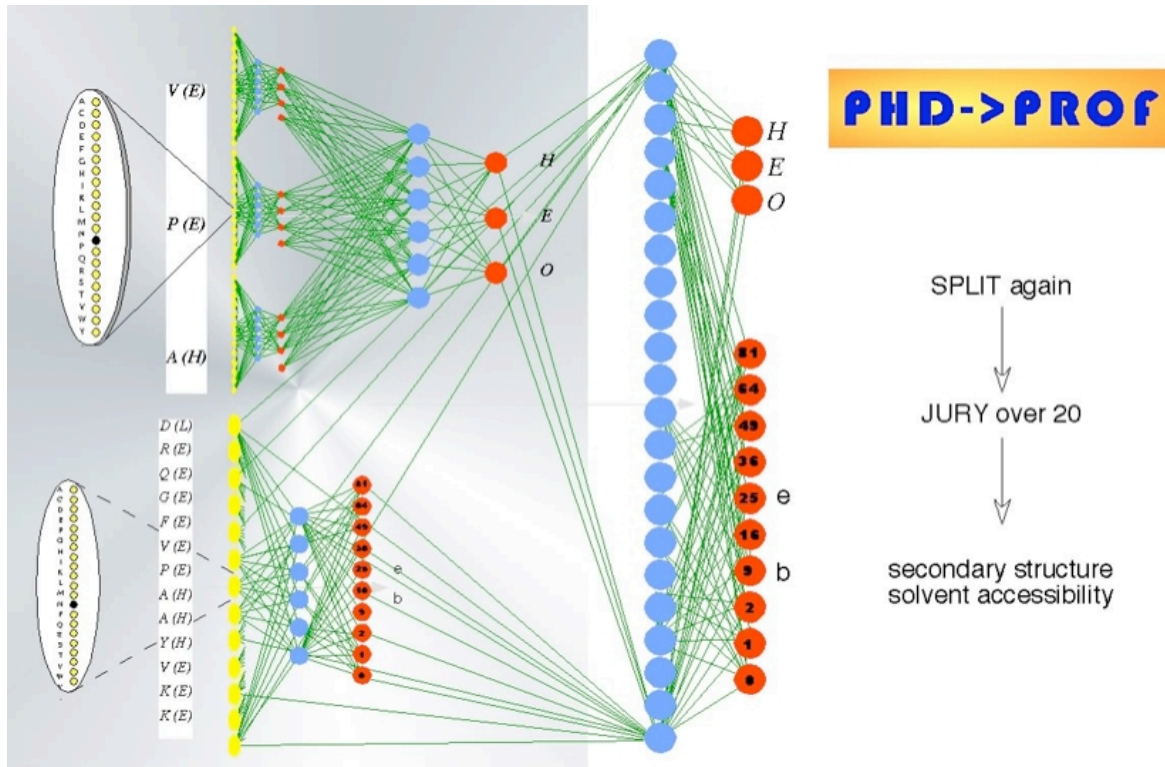
60%
->
72%

More complex system to predict structure

Sequence → PSI-BLAST → Filter

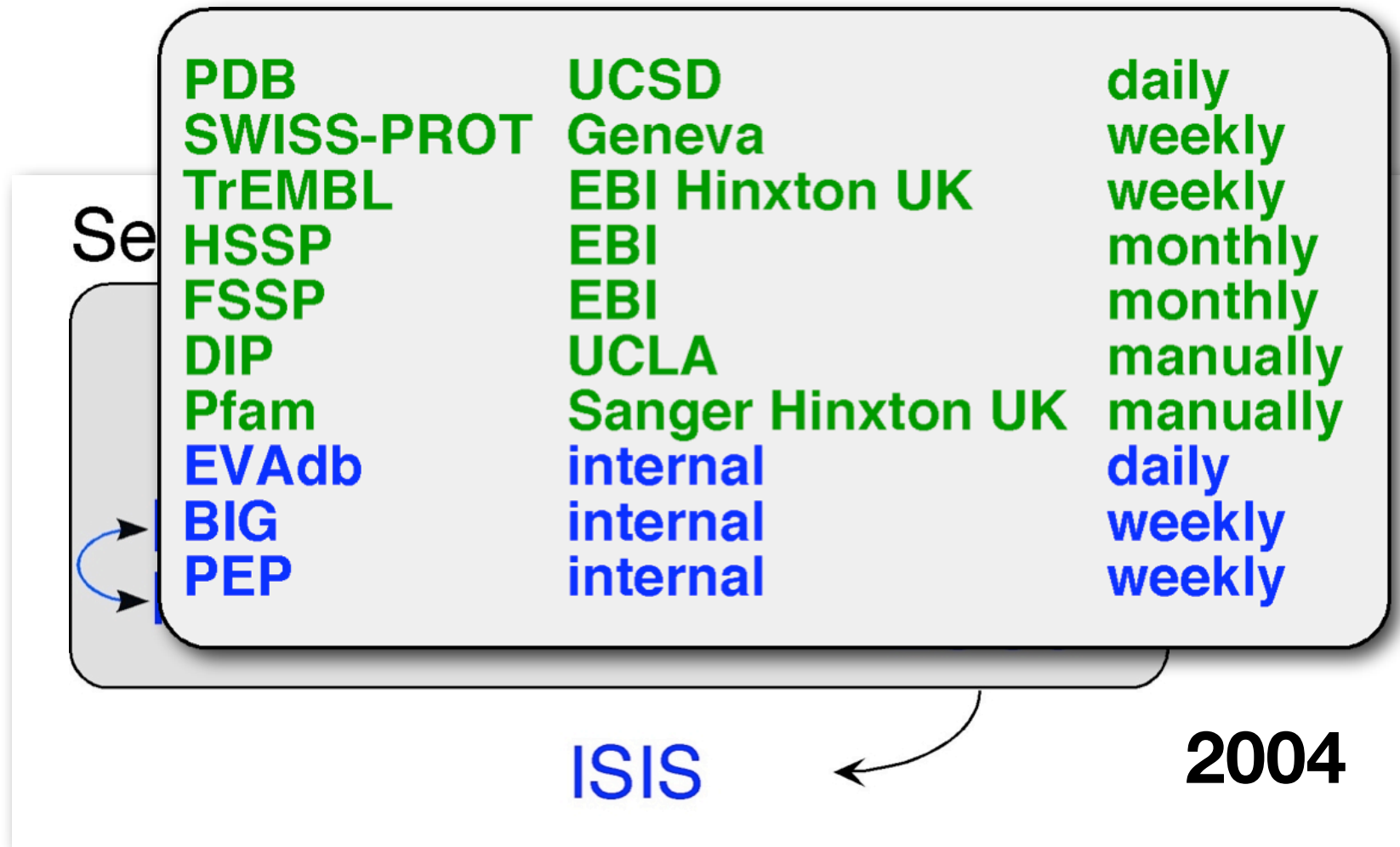
PROFsec
PROFacc

1999



Much more complex system for function

Much more complex system for function



What makes it work?

● Evolutionary information:

- Optimally choosing profile
- Explicitly using conserved residues

● (Predicted) 1D Structure

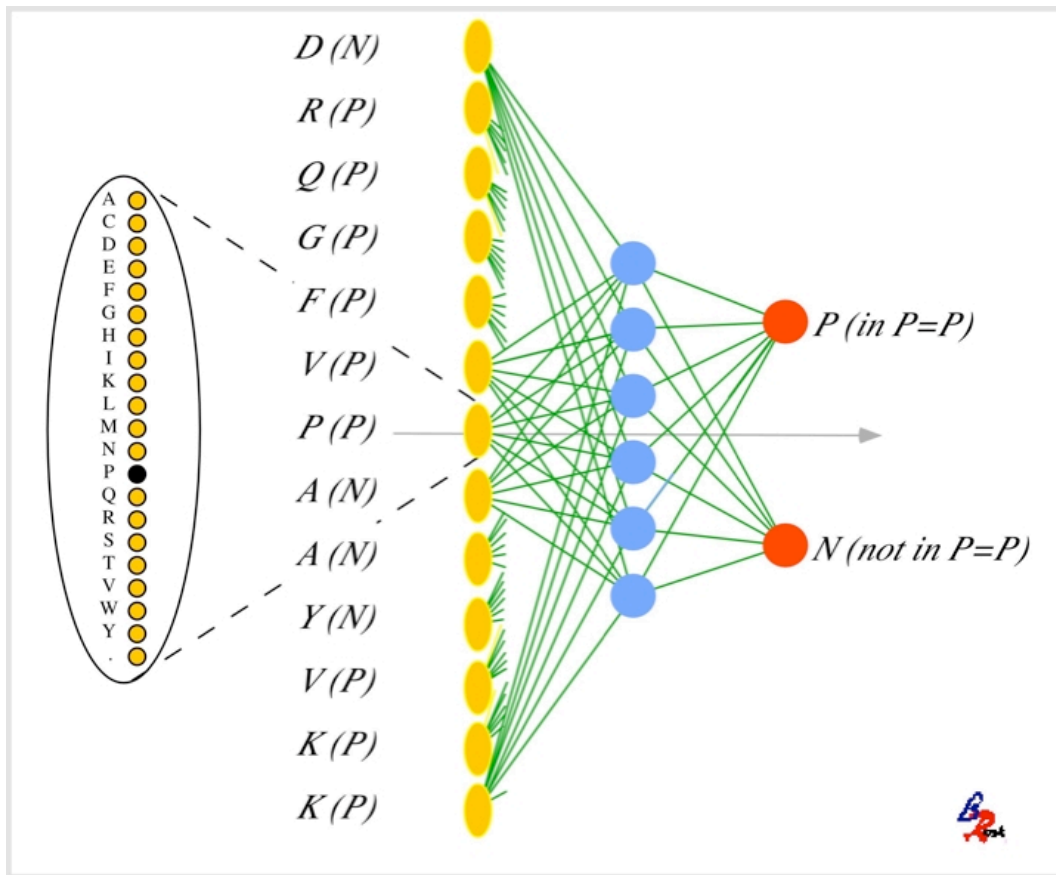
important: good prediction + used correctly

- Surface residues
- Secondary structure

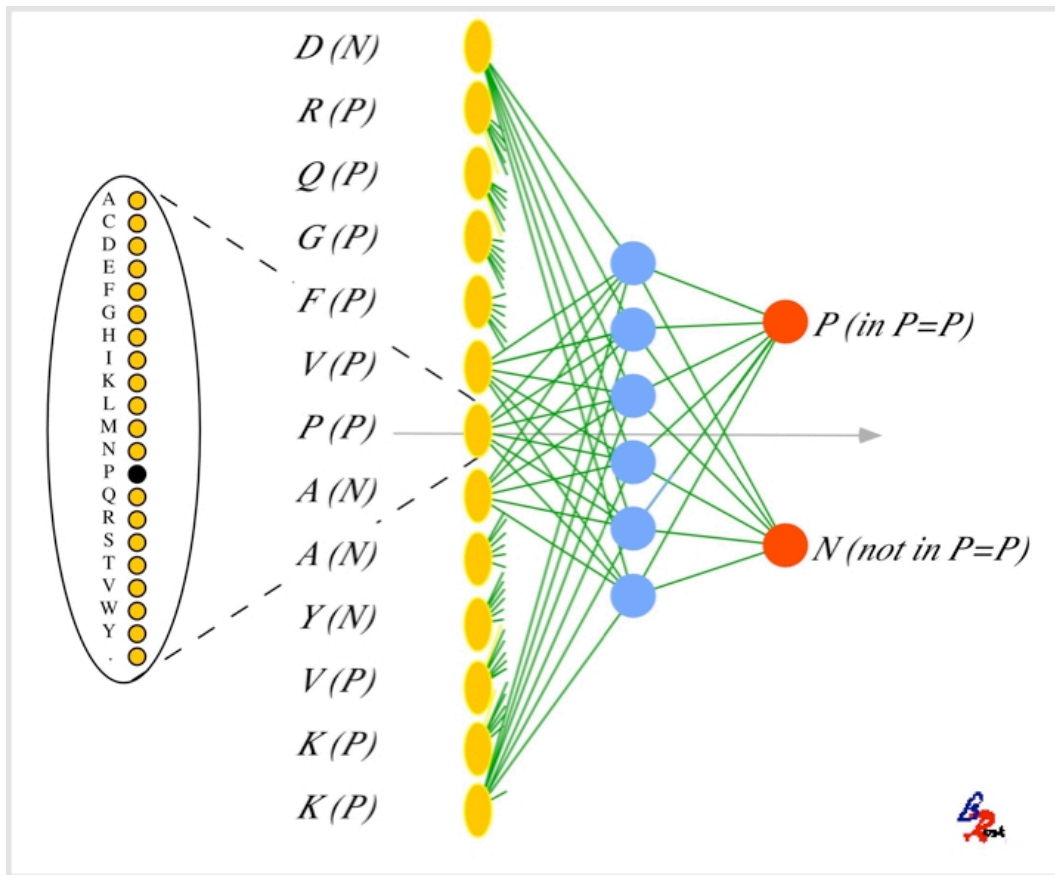
● Mark low-complexity and *sticky*

● Filtering “isolated predictions”

Strength of prediction reflects reliability?



Strength of prediction reflects reliability?



strong

0.9

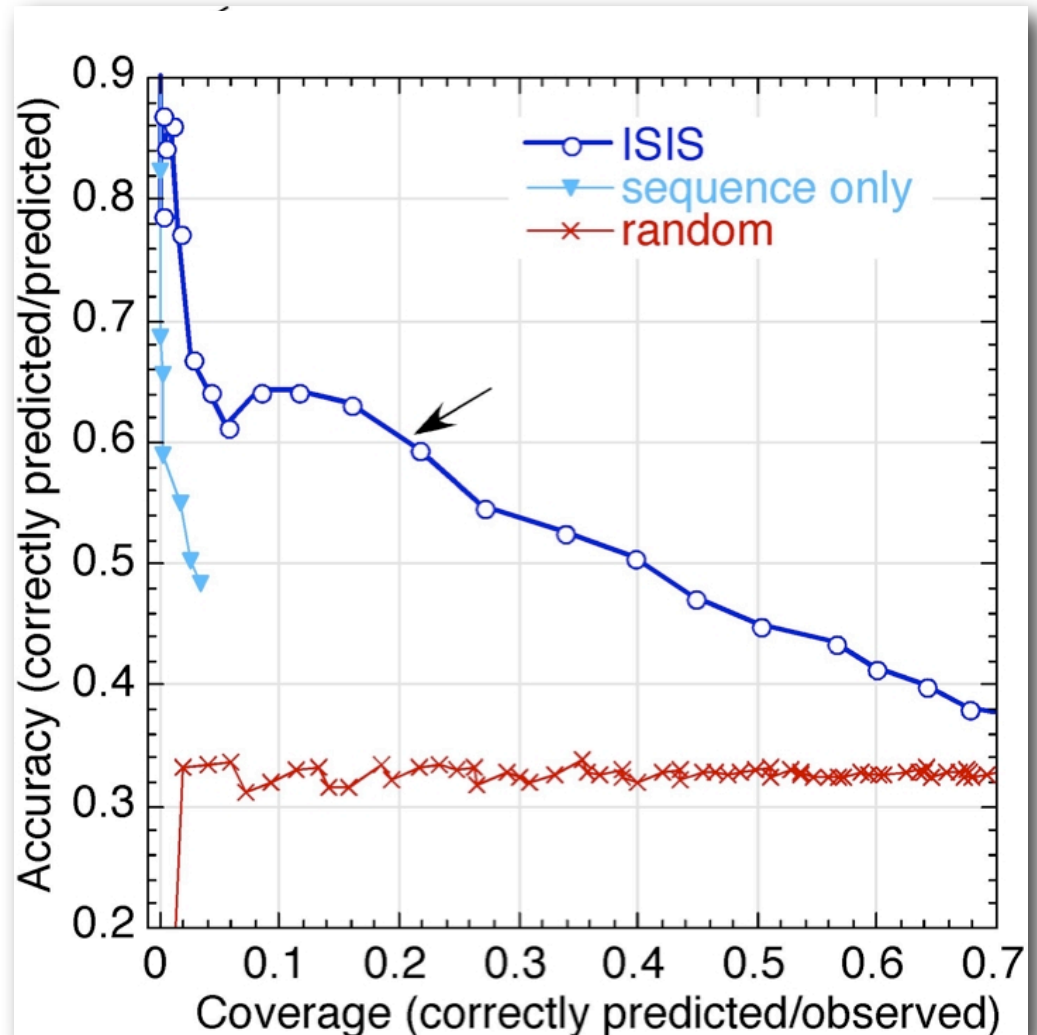
0.1

weak

0.6

0.4

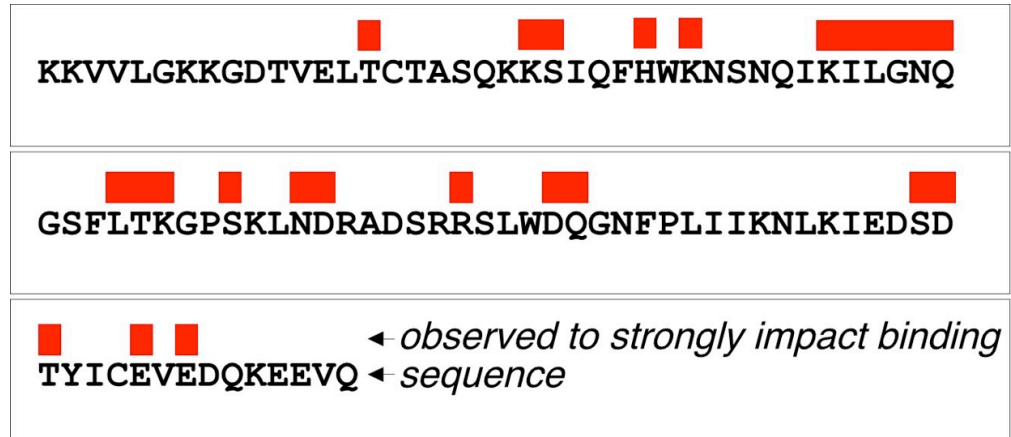
PP interfaces predicted from sequence



Prediction of *hot spots* for CD4

- alanine scan for V1 domain of CD4 (bound to gp120) (A Ashkenazi et al. & DJ Capon (1990) *PNAS* **87**, 7150)

red: observed

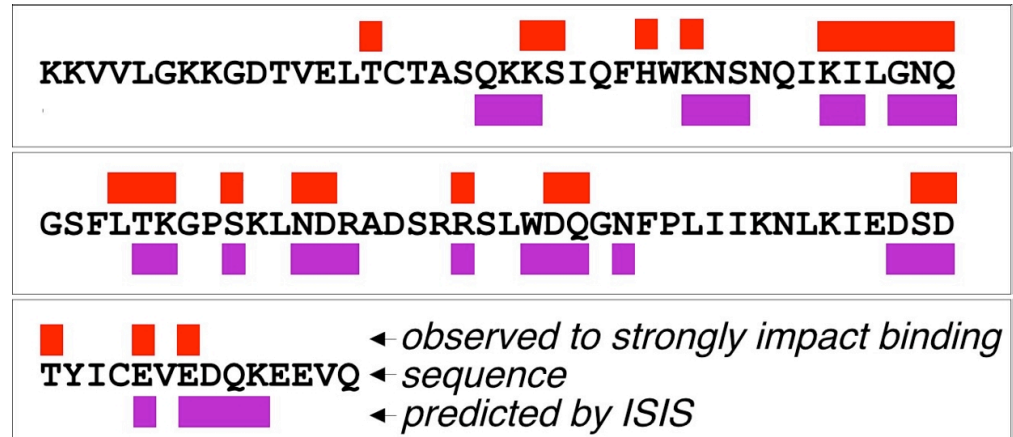


Prediction of *hot spots* for CD4

- alanine scan for V1 domain of CD4 (bound to gp120) (A Ashkenazi et al. & DJ Capon (1990) *PNAS* **87**, 7150)

red: observed
purple: predicted

(Y Ofran & B Rost (2006) *ISIS submitted*)



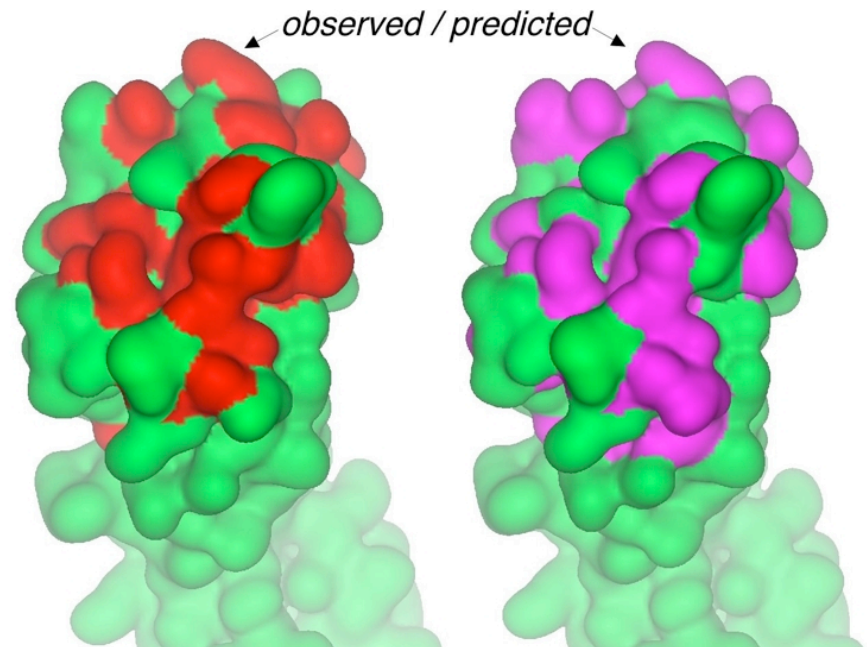
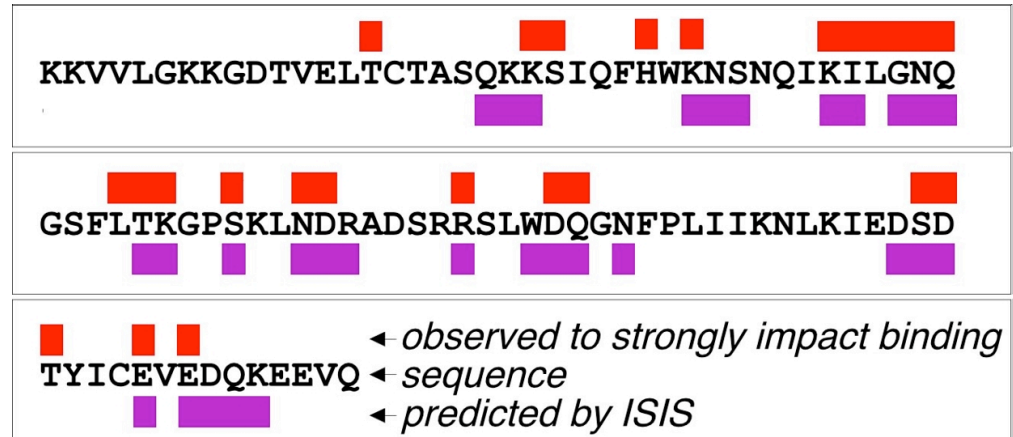
Prediction of *hot spots* for CD4

- alanine scan for V1 domain of CD4 (bound to gp120) (A Ashkenazi et al. & DJ Capon (1990) *PNAS* **87**, 7150)

red: observed
purple: predicted

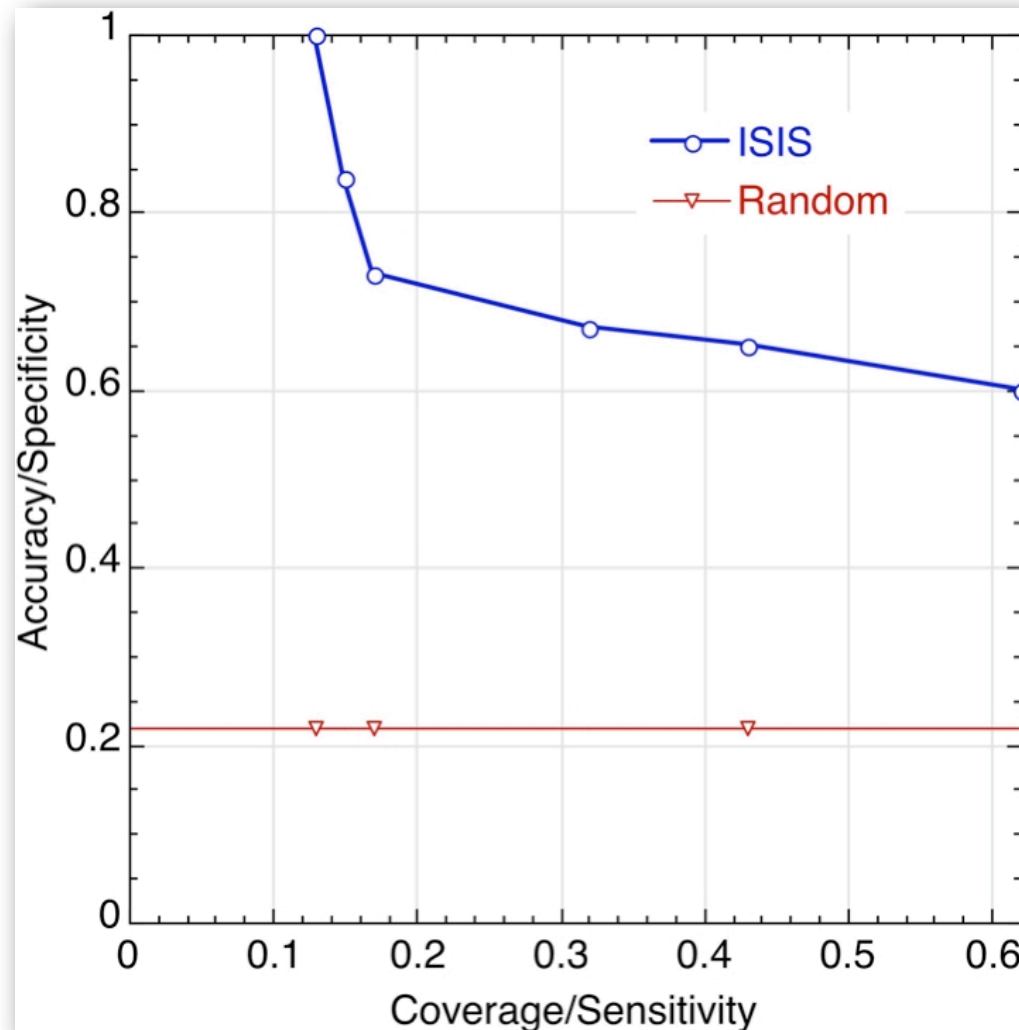
(Y Ofran & B Rost (2006) *ISIS submitted*)

- structure: PD Kwong et al. & WA Hendrickson (2000) *Structure* **8**, 1329-1339.



Hot spots reliably predicted from sequence!

hottest of hot = no error!





worst:
>60%
accuracy

Predict protein-protein binding partners

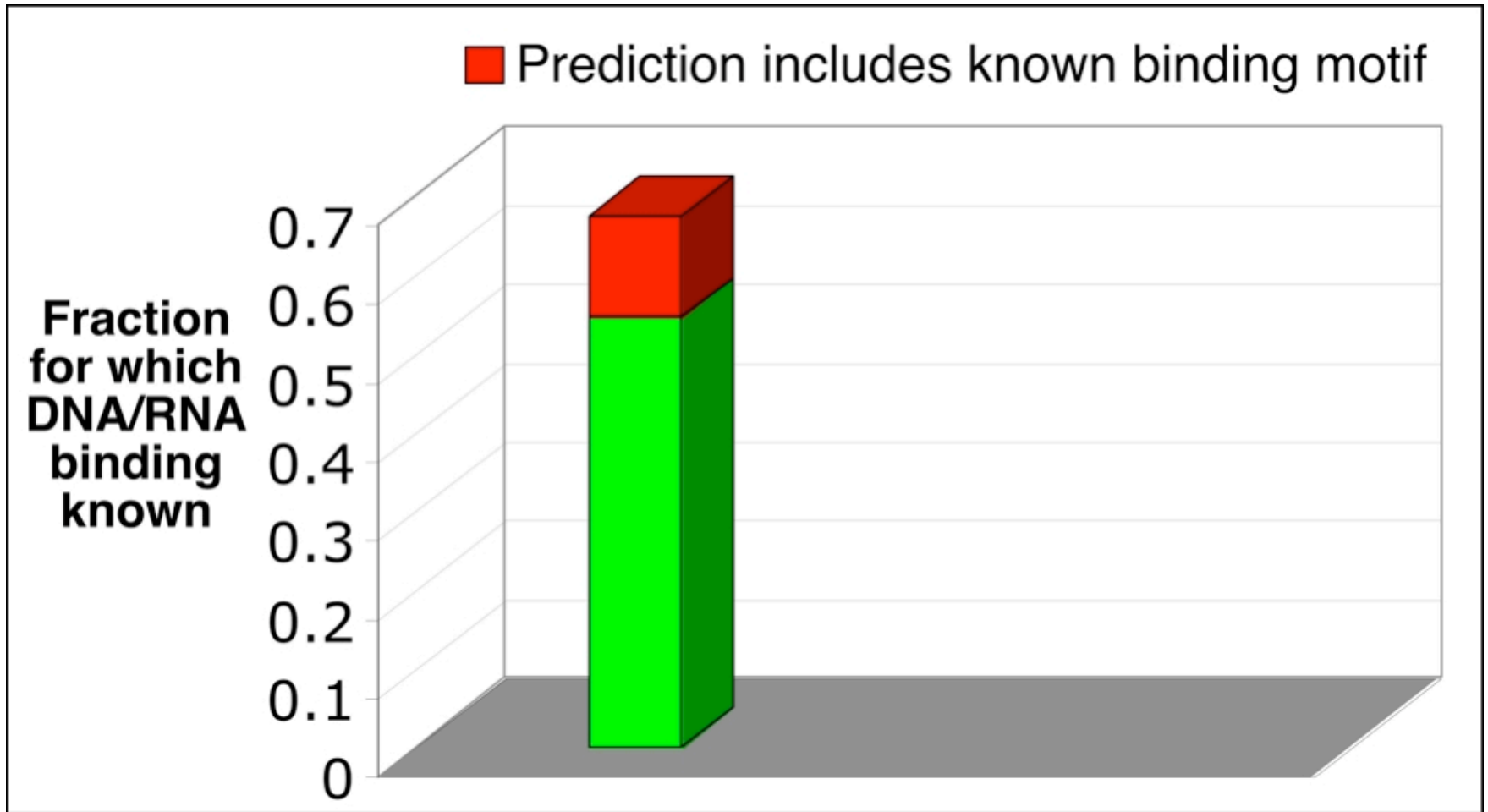
Reducing false positives:

- predict surface residues (PROFacc, 1999)
- predict residues in external interfaces (ISIS, 2004)
- predict residues saturated internally (PROFcon, 2004)
- localization (e.g. only all nuclear, LOctree, 2004)

Predict protein-protein binding partners

-  **Reducing false positives:**
 - predict surface residues (PROFacc, 1999)**
 - predict residues in external interfaces (ISIS, 2004)**
 - predict residues saturated internally (PROFcon, 2004)**
 - localization (e.g. only all nuclear, LOCtree, 2004)**
 -  **predict residues in protein-substrate interfaces (active)**







Most predictions are discoveries!



Predict protein-protein binding partners

Predict protein-protein binding partners

Reducing false positives:

-  predict surface residues (PROFacc, 1999)
-  predict residues in external interfaces (ISIS, 2004)
-  localization (e.g. only all nuclear, LOCo, 2004)
-  predict residues in protein-substrate interfaces (active)
-  predict residues saturated internally (PROFcon, 2004)
-  predict protein domains/improve alignments (2003/2004)

 **Put all together and predict binding partners!**

III. In passing:

Predict subcellular
localization

Predict sub-cellular localization

Homology

Alignment

Text analysis

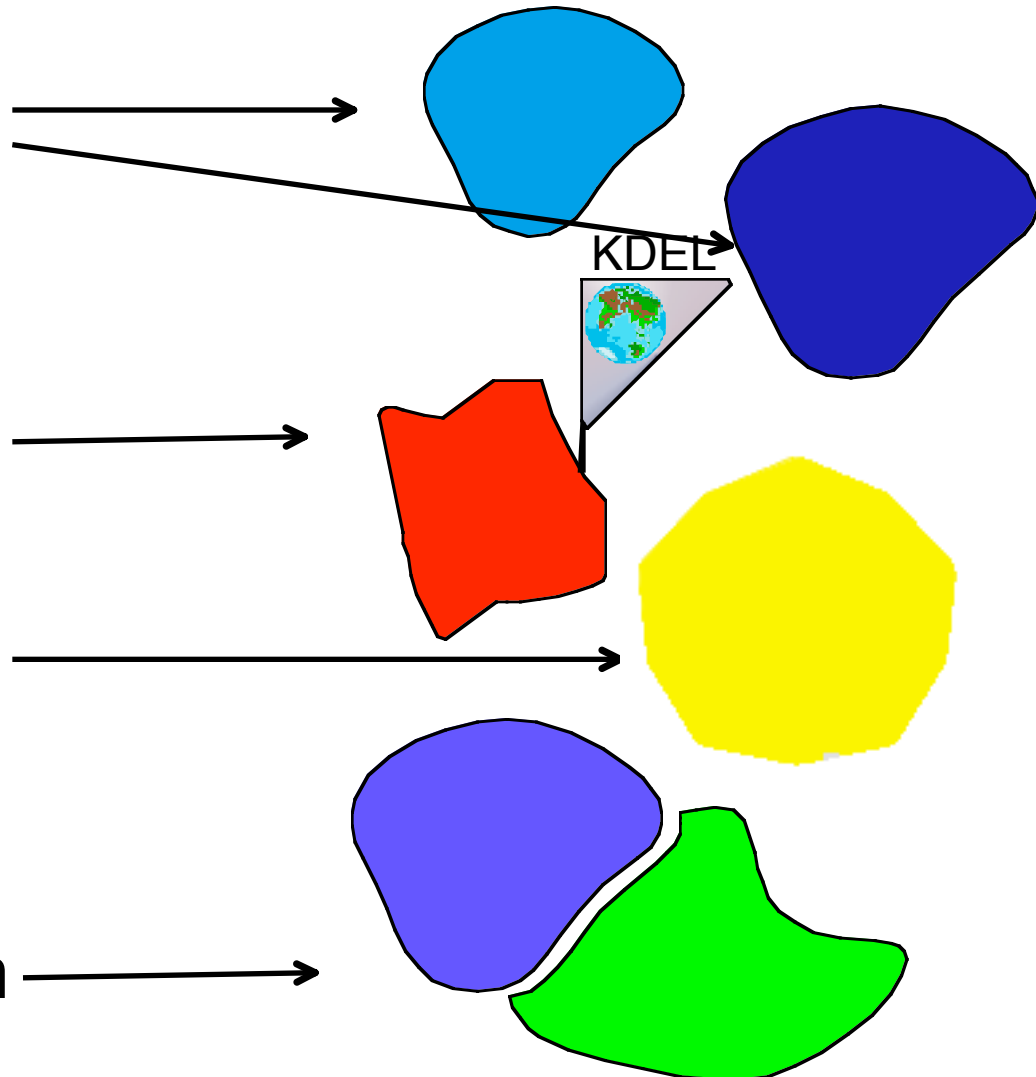
Motifs

De novo

structure

sequence

Protein-protein





Claudia Bertoni



Henry Bigelow



Kaz Wrzeszczynski



Yanay Ofran

Ta-Tsen Soong

Yana



Avner Schlessinger



Jinfeng Liu



Ingrid Kohl



Volker Eyrich



Marco Punta

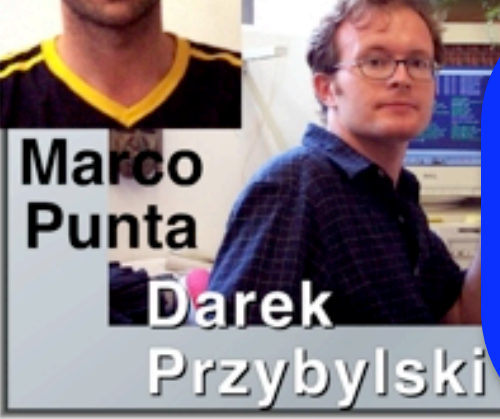


Eyal Mozes

Sara Gilman



Sven Mika



Darek Przybylski



Raj Nair



Andrew Kernytsky

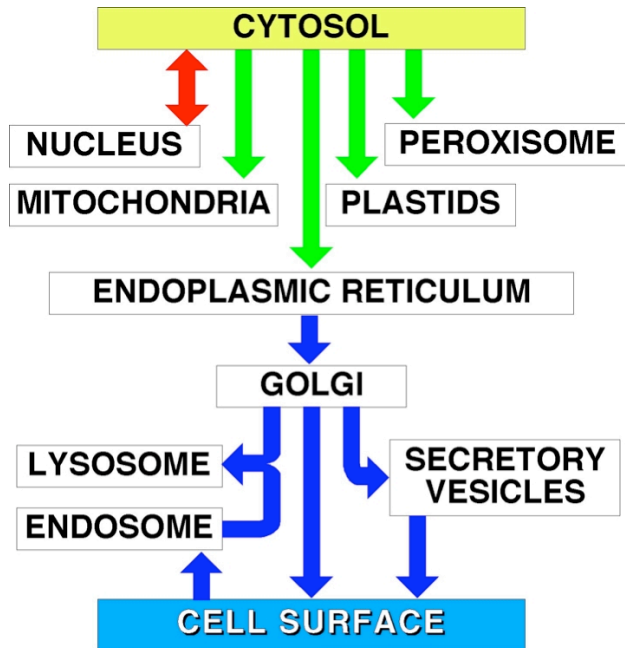


Guy Yachday



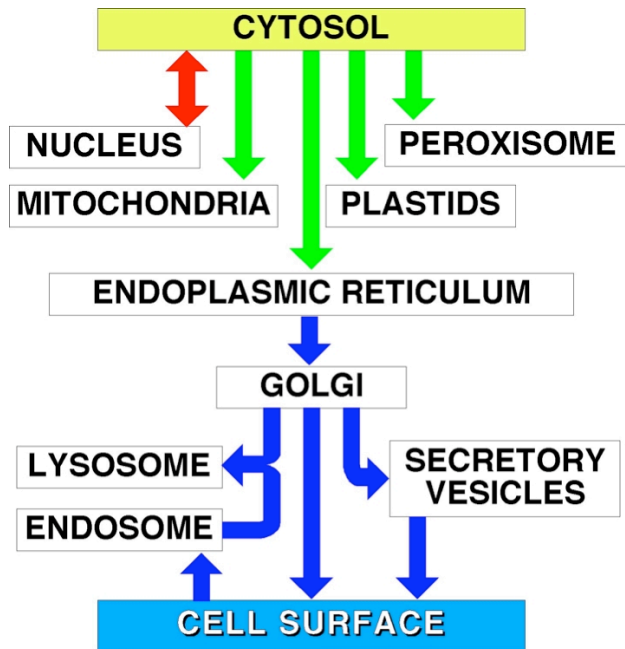
Phil Carter

Hierarchical prediction system

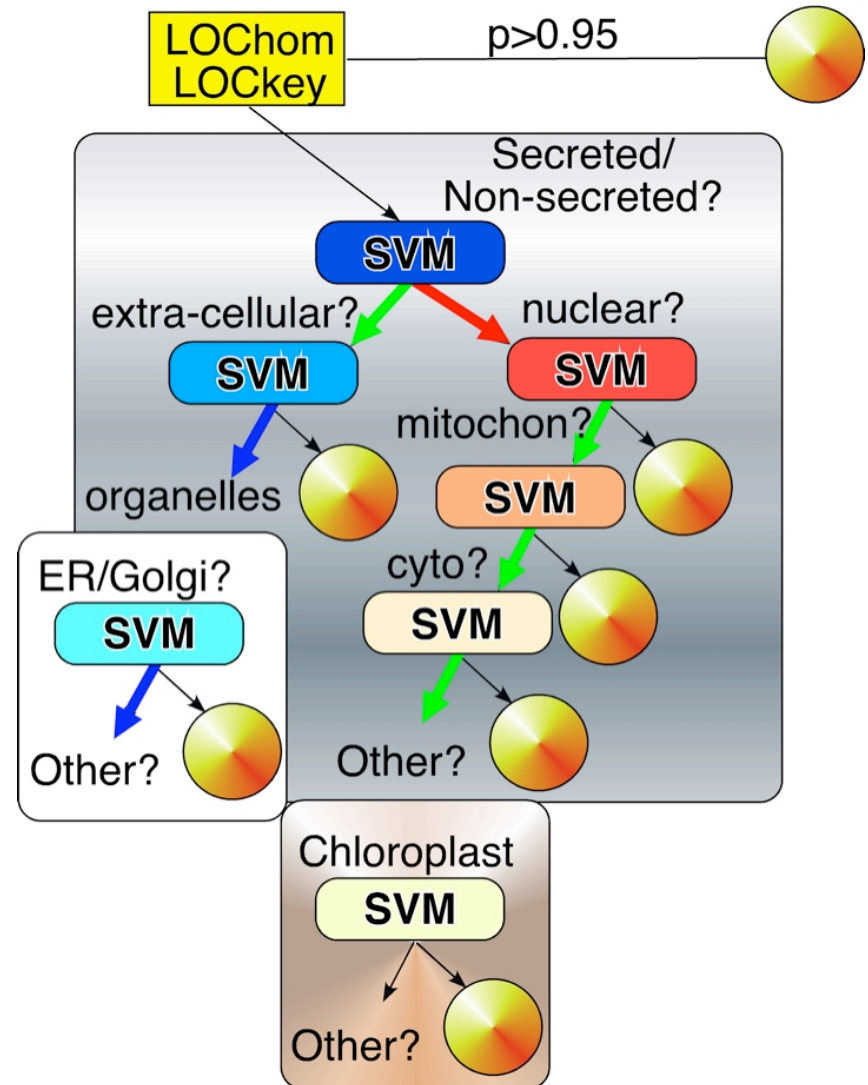


KEY: █ gated transport
█ transmembrane transport
█ vesicular transport

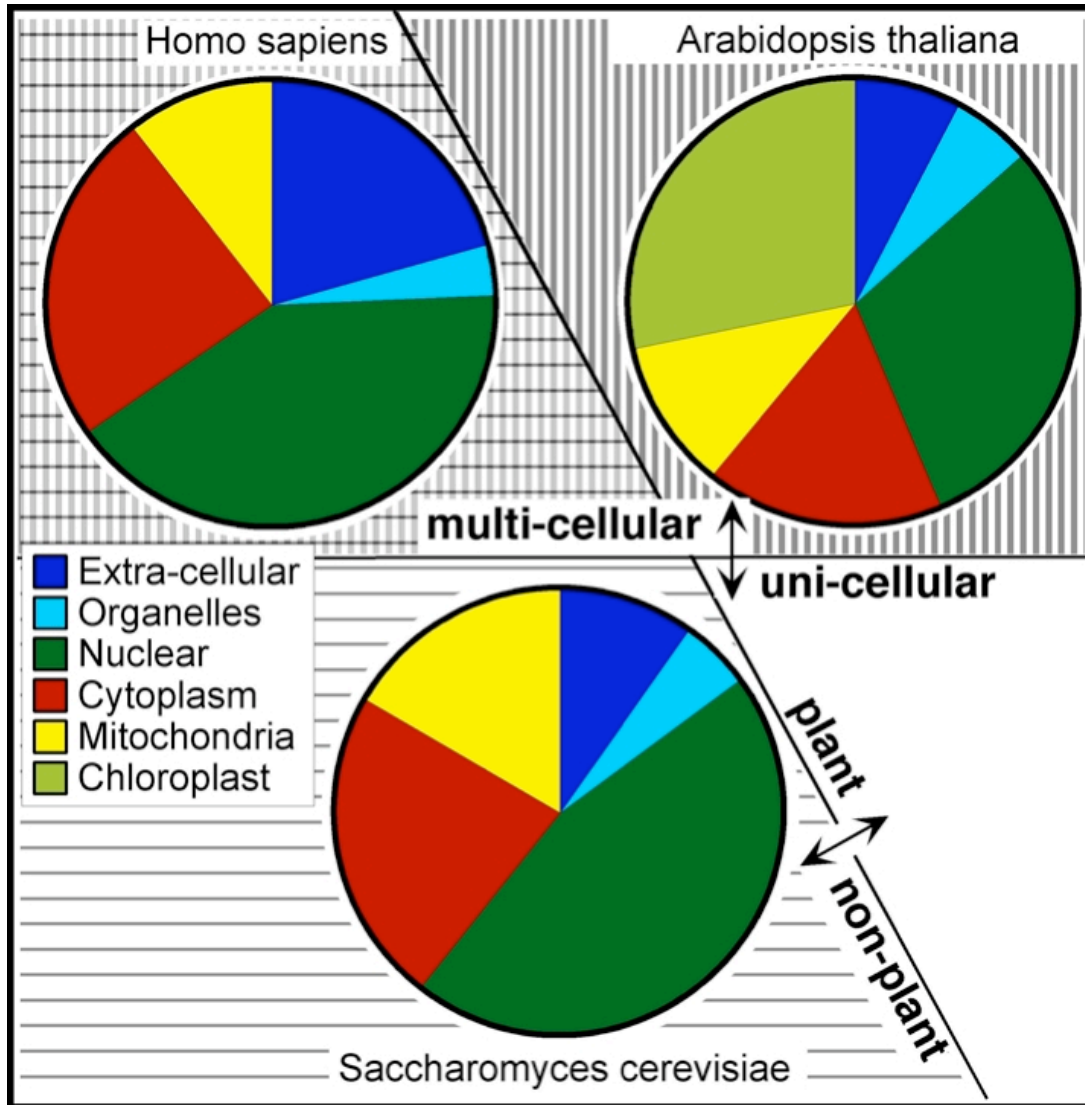
Hierarchical prediction system



KEY:
█ gated transport
█ transmembrane transport
█ vesicular transport



Complete map of localization



SWISS-PROT: transcription factor E2F-1

Description and origin of the Protein	
Description	Transcription factor E2F1 (E2F-1) (Retinoblastoma binding protein 3) (RBBP-3) (PRB-binding protein E2F-1) (PBR3) (Retinoblastoma-associated protein 1) (RBAP-1).
Gene name(s)	E2F1 OR RBBP3.
Organism source	Homo sapiens (Human).
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
Comments	
FUNCTION	TRANSCRIPTION ACTIVATOR THAT BINDS DNA COOPERATIVELY WITH DP PROTEINS THROUGH THE E2 RECOGNITION SITE, TTTCC/GCGC, FOUND IN THE PROMOTER REGION OF A NUMBER OF GENES WHOSE PRODUCTS ARE INVOLVED IN CELL CYCLE REGULATION OR IN DNA REPLICATION. THE DRTF1/E2F COMPLEX FUNCTIONS IN THE CONTROL OF CELL-CYCLE PROGRESSION FROM G1 TO S PHASE. E2F-1 BINDS PREFERENTIALLY RB1 PROTEIN, IN A CELL-CYCLE DEPENDENT MANNER. IT CAN MEDIATE BOTH CELL PROLIFERATION AND P53-DEPENDENT APOPTOSIS.
SUBUNIT	COMPONENT OF THE DRTF1/E2F TRANSCRIPTION FACTOR COMPLEX. FORMS HETERODIMERS WITH DP FAMILY MEMBERS. THE E2F-1 COMPLEX BINDS SPECIFICALLY HYPOPHOSPHORYLATED RETINOBLASTOMA PROTEIN RB1. DURING THE CELL CYCLE, RB1 BECOMES PHOSPHORYLATED IN MID-TO-LATE G1 PHASE, DETACHES FROM THE DRTF1/E2F COMPLEX, RENDERING E2F TRANSCRIPTIONALLY ACTIVE. VIRAL ONCOPROTEINS, NOTABLY E1A, T- ANTIGEN AND HPV E7, ARE CAPABLE OF SEQUESTERING RB PROTEIN, THUS RELEASING THE ACTIVE COMPLEX.
SUBCELLULAR LOCATION	NUCLEAR.
Keywords	

Transcription regulation; Activator; DNA-binding; Nuclear protein; Phosphorylation; Cell cycle; Apoptosis; Polymorphism;

SWISS-PROT: transcription factor E2F-1

Description and origin of the Protein

Description	Transcription factor E2F1 (E2F-1) (Retinoblastoma binding protein 3) (RBBP-3) (PRB-binding protein E2F-1) (PBR3) (Retinoblastoma-associated protein 1) (RBAP-1).
Gene name(s)	E2F1 OR RBBP3.
Organism source	Homo sapiens (Human).
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

Comments

FUNCTION	TRANSCRIPTION ACTIVATOR THAT BINDS DNA COOPERATIVELY WITH DP PROTEINS THROUGH THE E2 RECOGNITION SITE, TTTCC/GCGC, FOUND IN THE PROMOTER REGION OF A NUMBER OF GENES WHOSE PRODUCTS ARE INVOLVED IN CELL CYCLE REGULATION OR IN DNA REPLICATION. THE DRTF1/E2F COMPLEX FUNCTIONS IN THE CONTROL OF CELL-CYCLE PROGRESSION FROM G1 TO S PHASE. E2F-1 BINDS PREFERENTIALLY RB1 PROTEIN, IN A CELL-CYCLE DEPENDENT MANNER. IT CAN MEDIATE BOTH CELL PROLIFERATION AND P53-DEPENDENT APOPTOSIS.
SUBUNIT	COMPONENT OF THE DRTF1/E2F TRANSCRIPTION FACTOR COMPLEX. FORMS HETERODIMERS WITH DP FAMILY MEMBERS. THE E2F-1 COMPLEX BINDS SPECIFICALLY HYPOPHOSPHORYLATED RETINOBLASTOMA PROTEIN RB1. DURING THE CELL CYCLE, RB1 BECOMES PHOSPHORYLATED IN MID-TO-LATE G1 PHASE, DETACHES FROM THE DRTF1/E2F COMPLEX, RENDERING E2F TRANSCRIPTIONALLY ACTIVE. VIRAL ONCOPROTEINS, NOTABLY E1A, T- ANTIGEN AND HPV E7, ARE CAPABLE OF SEQUESTERING RB PROTEIN, THUS RELEASING THE ACTIVE COMPLEX.
SUBCELLULAR LOCATION	

Keywords

Transcription regulation; Activator; DNA-binding; Nuclear protein; Phosphorylation; Cell cycle; Apoptosis; Polymorphism;

Localization: better and more detail

<http://www.rostlab.org/services/nlprot/>

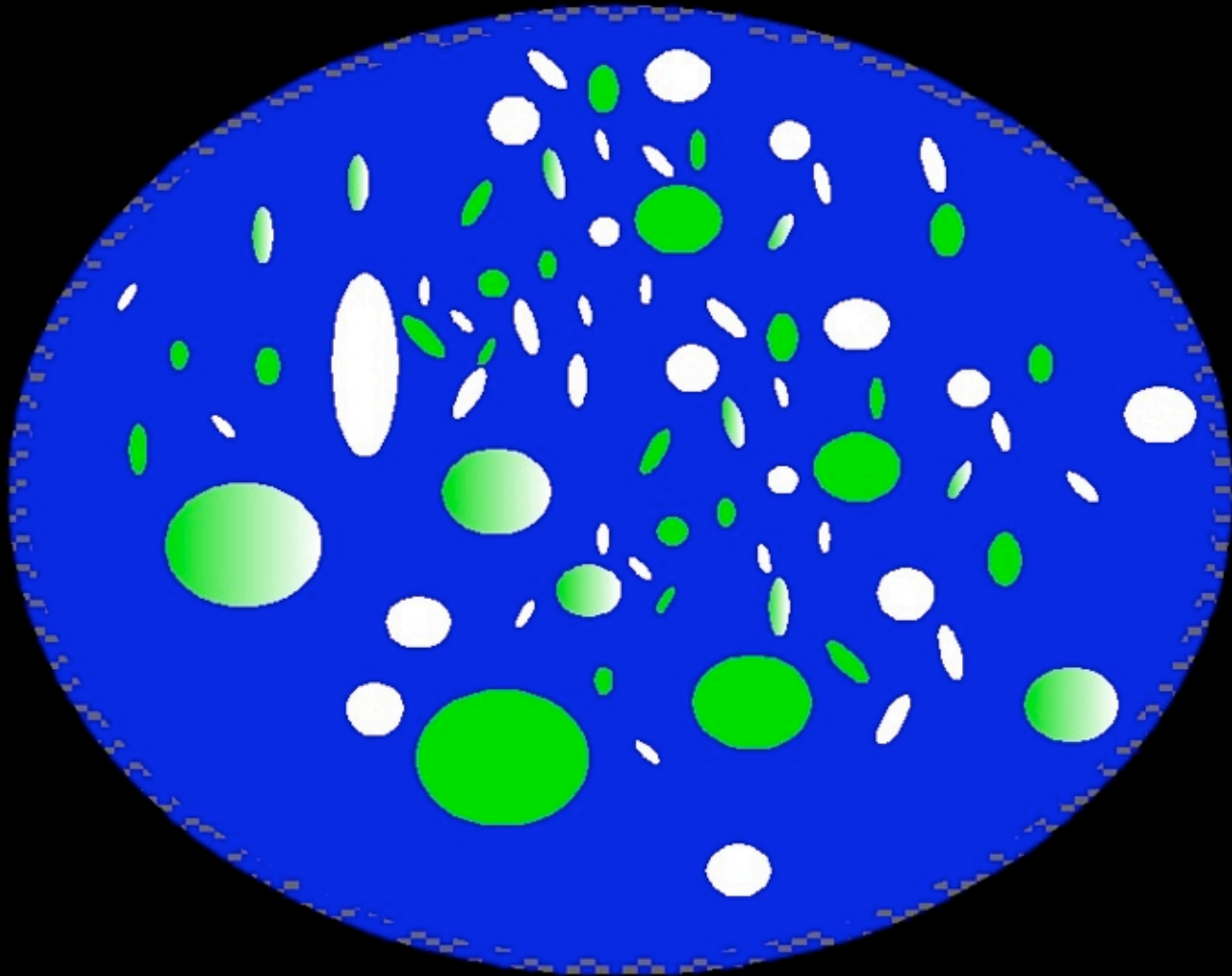
The image shows a PubMed search result for a paper on PAI-1 promoter polymorphism. Red circles and arrows highlight key terms: '4G/5G PAI-1 Promoter Polymorphism', 'PAI-1 activity and antigen concentrations', 'inflammatory reaction', and 'acute-phase PAI-1 levels'. A black box on the right displays the 'Prot View of Swiss-Prot: P43244' for MAT3_RAT. This box is divided into 'References' and 'Comments' sections. The 'References' section lists a paper by Noelling et al. (2001) with its MEDLINE and PubMed IDs. The 'Comments' section contains three bullet points: 'FUNCTION: Converts holo-ACP to apo-ACP', 'CATALYTIC ACTIVITY: Holo-[alpha]-caseinase', and 'SIMILARITY: Belongs to the acpD family of proteins'. Below the comments is a table with protein information: MAT3_RAT, accession number P43244, O35833, release date November 1995, and last modified in October 2001 and February 2003. The 'Name and origin of the protein' section lists 'Matrin 3' as the protein name, with synonyms 'None' and 'MAT3'. The gene name is 'MAT3' from 'Rattus norvegicus (Rat)'. The 'Comments' section at the bottom of the table lists: 'FUNCTION: May play a role in transcription or may interact with other nuclear matrix proteins to form the internal fibrogranular network.', 'SUBCELLULAR LOCATION: Nuclear matrix.', 'SIMILARITY: Contains 1 matrin-type zinc finger.', and 'SIMILARITY: Contains 2 RNA recognition motif (RRM) domains.'

NLProt first step toward:

- 🕒 machine-reading literature and
- 🕒 building databases from extracted information

IV. In passing:
Function from 3D-
Structural Genomics

Structural genomics: 1 structure / family for all



Speeding up structure determination

- **today: more structures in 27 days than in first 27 years**
- **8% 'new sequence-structure family'**

Speeding up structure determination

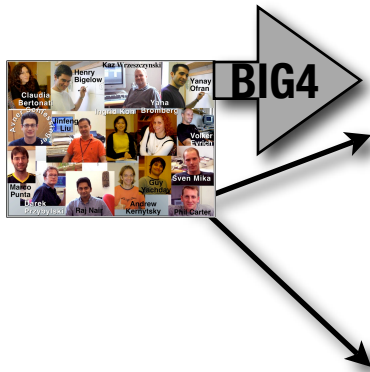
- today: more structures in 27 days than in first 27 years
- 8% 'new sequence-structure family'

Acronym	Name	Country
JCSG	The Joint Center for Structural Genomics	USA
MCSG	The Midwest Center for Structural Genomics	USA
NYSGRC	New York Structural Genomics Research Consortium	USA
NESG	Northeast Structural Genomics Consortium	USA
Gene3D	Accelerated Technologies Center for Gene to 3D Structure	USA
CESG	Center for Eukaryotic Structural Genomics	USA
CHTSB	Center for High-Throughput Structural Biology	USA
CSMP	Center for Structures of Membrane Proteins	USA
ICSFI	Integrated Center for Structure and Function Innovation	USA
NYCOMPS	New York Consortium on Membrane Protein Structure	USA
BSGC	Berkeley Structural Genomics Center	USA
SECSG	The Southeast Collaboratory for Structural Genomics	USA
SGPP	Structural Genomics of Pathogenic Protozoa Consortium	USA
S2F	Structure to function	USA
SGC	Structural Genomics Consortium	Canada
PSF	Protein Structure Factory	Germany
PSB	Partnership for Structural Biology	France
SGM	Structural Genomics of Micobacteria	France
YSG	Yeast Structural genomics	France
SPINE	Structural Proteomics in Europe	Europe
RSGI	RIKEN Structural Genomics Initiative	Japan

>\$150M/year

Speeding up structure determination

- today: more structures in 27 days than in first 27 years
- 8% 'new sequence-structure family'



Acronym	Name	Country
JCSG	The Joint Center for Structural Genomics	USA
MCSG	The Midwest Center for Structural Genomics	USA
NYSGRC	New York Structural Genomics Research Consortium	USA
NESG	Northeast Structural Genomics Consortium	USA
Gene3D	Accelerated Technologies Center for Gene to 3D Structure	USA
CESG	Center for Eukaryotic Structural Genomics	USA
CHTSB	Center for High-Throughput Structural Biology	USA
CSMP	Center for Structures of Membrane Proteins	USA
ICSFI	Integrated Center for Structure and Function Innovation	USA
NYCOMPS	New York Consortium on Membrane Protein Structure	USA
BSGC	Berkeley Structural Genomics Center	USA
SECSG	The Southeast Collaboratory for Structural Genomics	USA
SGPP	Structural Genomics of Pathogenic Protozoa Consortium	USA
S2F	Structure to function	USA
SGC	Structural Genomics Consortium	Canada
PSF	Protein Structure Factory	Germany
PSB	Partnership for Structural Biology	France
SGM	Structural Genomics of Micobacteria	France
YSG	Yeast Structural genomics	France
SPINE	Structural Proteomics in Europe	Europe
RSGI	RIKEN Structural Genomics Initiative	Japan

>\$150M/year



Claudia Bertoni



Henry Bigelow



Kaz Wrzeszczynski



Yanay Ofran

Ta-Tsen Soong

Yana



Avner Givon



Jinfeng Liu



Ingrid Koh



Bromberg



Volker Eyrich

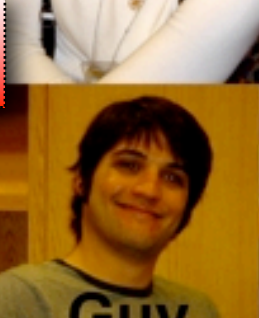


Marco Punta



Eyal Mozes

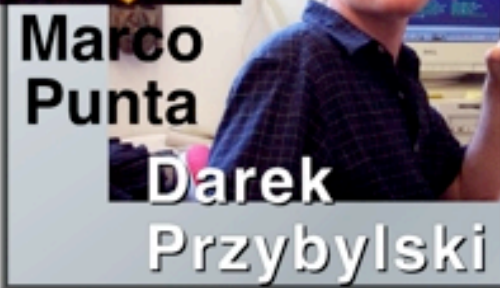
Sara Gilman



Guy Yachday



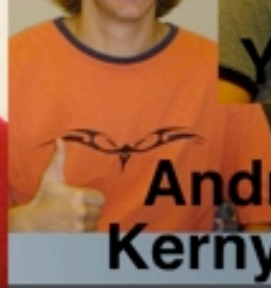
Sven Mika



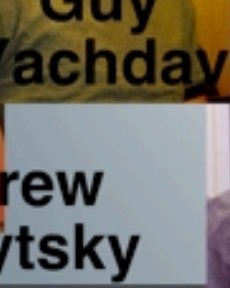
Darek Przybylski



Raj Nair



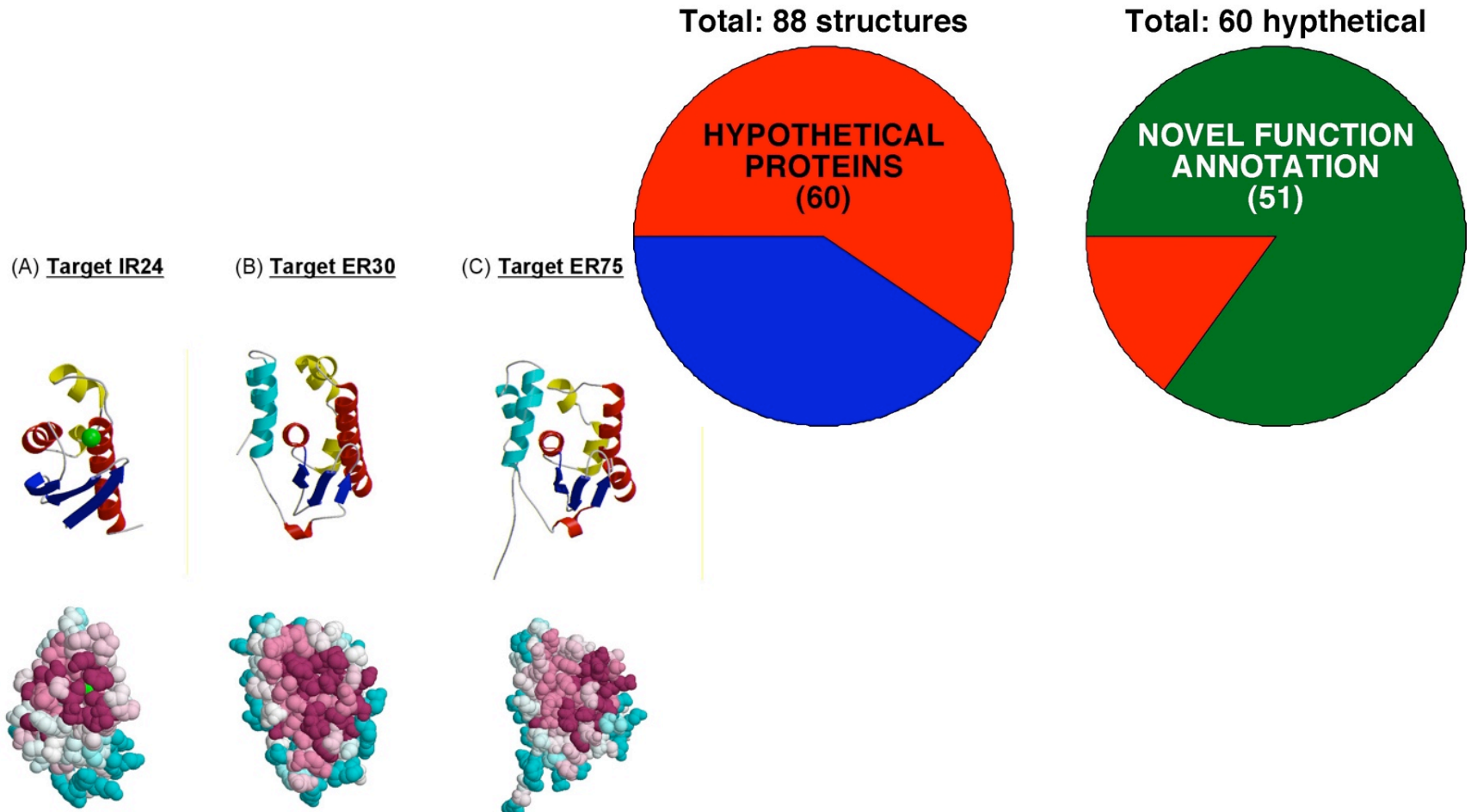
Andrew Kernytsky



Phil Carter

Structure reveals function

Claudia **Bertonati**, Sharon Goldsmith-Fischman & Barry **Honig**, unpublished





Claudia Bertoni



Henry Bigelow



Kaz Wrzeszczynski



Yanay Ofran

Ta-Tsen Soong

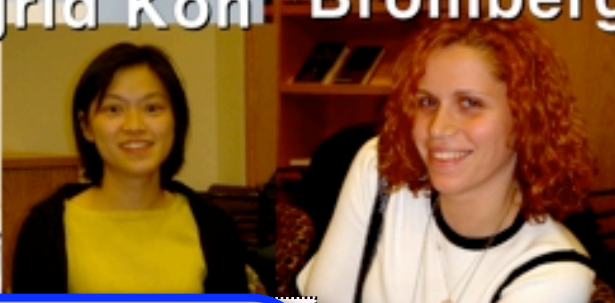
Yana



Avner Schlessinger



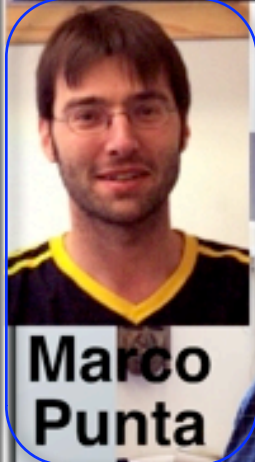
Jinfeng Liu



Ingrid Koh



Bromberg



Marco Punta



Eyal Mozes

Sara Gilman



Guy Yachday



Sven Mika



Darek Przybylski



Raj Nair



Andrew Kernytsky



Phil Carter

Automatic annotation of function

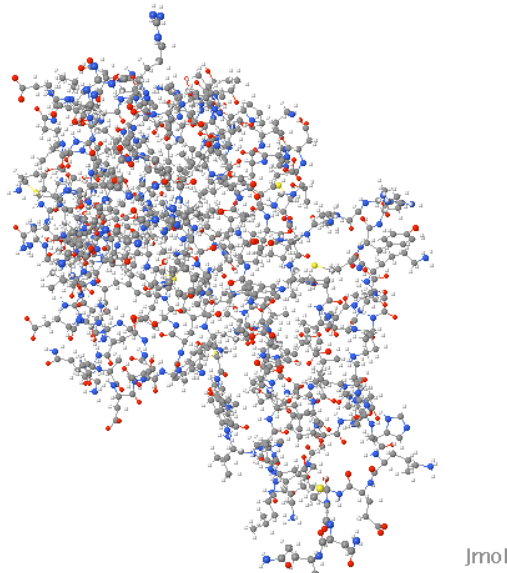
MODEL/PDBid: [XXX](#)

Protein Information

Protein name Protein CGI-126 (HSPC155).
Organism Homo sapiens
Gene Name Name=Ufc1;

Sequence

Chain: A Length: 175 AA
MADEATRRVVSEIPVLKTNAGPRDRELWVQRLKE
EYQSLIRYVENNKADNDWFRLESNKEGTRWFGK
CWYIHDLKYEFDIEFDIPITYPTAPEIAVPEL
DGKTAKMYRGGKICLTDHFKPLWARNVPKFGLAH
LMALGLPWLAVEIPDLIQKGVIOHKEKCNQLEH
HHHHH



Analysis

Databases

Sequence Annotation
Genomic Context
Structure Classification
Available Literature

Sequence

Sequence Similarity
Sequence Motifs

Structure

Structure Validation
Structure Homologues
Structure Motifs
Conservation Map
Electrostatic Potential
Cavities

Predictions-Predictions-Predictions-Prediction

Structure

Secondary Structure
TM, coiled coil, low complexity
Disorder Region Predictions
B-factors
Metal Binding Sites
Protein-Protein interaction

Function

Fully Automated Servers
Subcellular Localization
Posttranslational Modifications

Automatic annotation of function

GeneTegrate

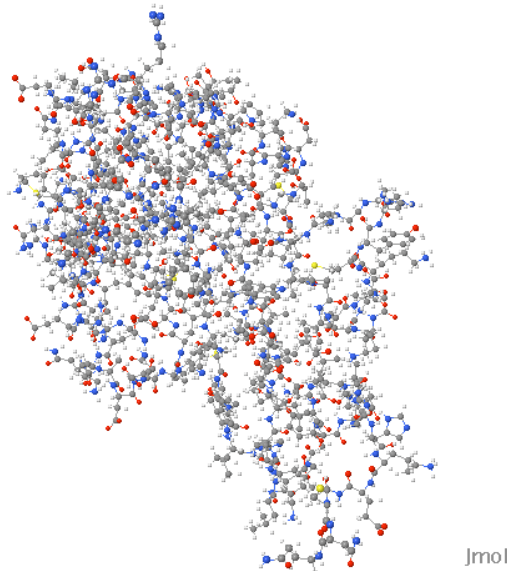
Protein Information

Protein name Protein CGI-126 (HSPC155).
Organism Homo sapiens
Gene Name Name=Ufc1;

Sequence

Chain: A Length: 175 AA
MADEATRRVVSEIPVLKTNAGPRDRELWVQRLKE
EYQSLIRYVENNKNADNDWFRLESNKEGTRWFGK
CWYIHDLKYEFDIEFDIPITYPTTAPEIAVPEL
DGKTAKMYRGGKICLTDHFKPLWARNVPKFLAH
LMALGLPWLAVEIPDLIQKGVIOHKEKCNQLEH
HHHHH

MODEL/PDBid: [XXX](#)



Analysis

Databases

Sequence Annotation
Genomic Context
Structure Classification
Available Literature

Sequence

Sequence Similarity
Sequence Motifs

Structure

Structure Validation
Structure Homologues
Structure Motifs
Conservation Map
Electrostatic Potential
Cavities

Predictions-Predictions-Predictions-Prediction

Structure

Secondary Structure
TM, coiled coil, low complexity
Disorder Region Predictions
B-factors
Metal Binding Sites
Protein-Protein interaction

Function

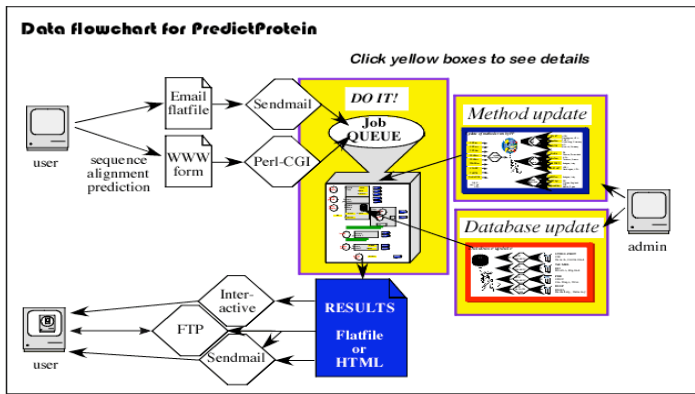
Fully Automated Servers
Subcellular Localization
Posttranslational Modifications

GeneTegrate: ontology for comp bio

PredictProtein

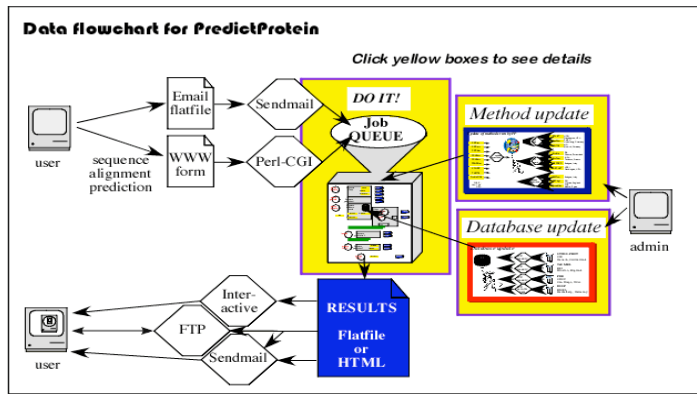
- growing since 1992
- >50,000 users
- from 102 countries

www.predictprotein.org/doc/flowchart/syn.html



GeneTegrate: ontology for comp bio

PredictProtein



- growing since 1992
- >50,000 users
- from 102 countries

www.predictprotein.org/doc/flowchart/syn.html

GeneTegrate

Yechiam Yemini (CU)

Yoav Freund (UCSD), Gal Kaiser (CU), Ken Ross (CU)

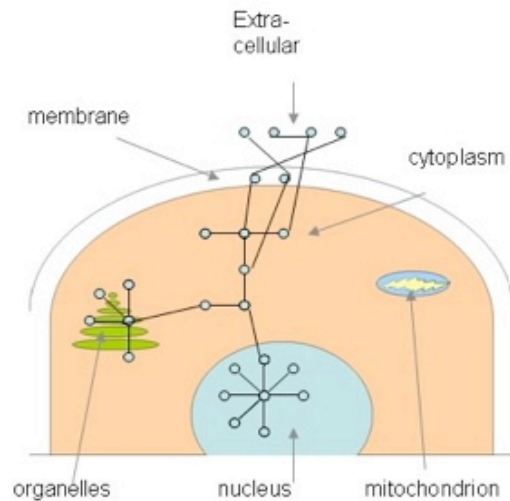
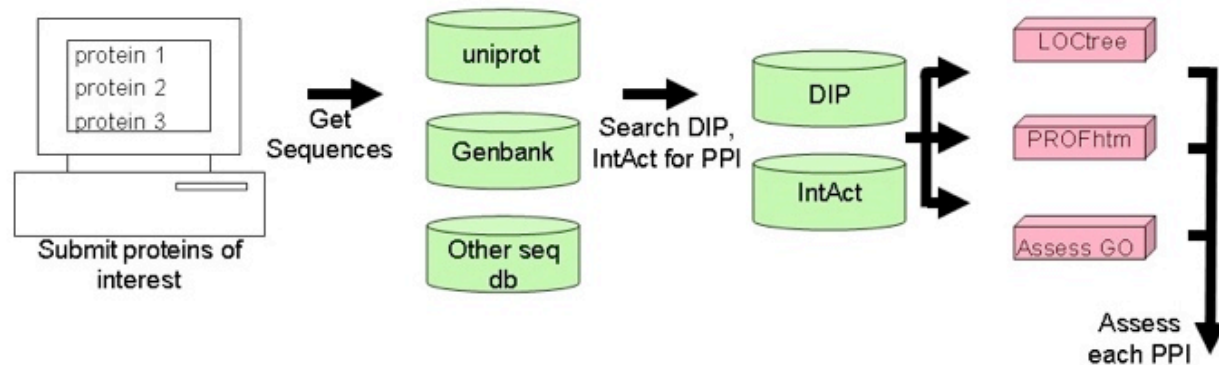
5 challenges:

- Diversity, Confidence, Scaling, Complexity, Reuse

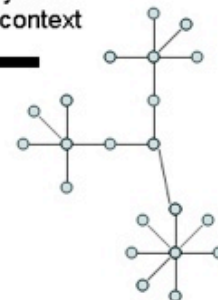
Solution:

- ontology for computational biology
- unified abstractions of enriched object-relationship semantic layer
- classifier-based indexing, look-ahead caching, generalized object-relationship spreadsheet

PiNat (Protein Interaction Network analysis tool)



Display in a cellular context



Generate network

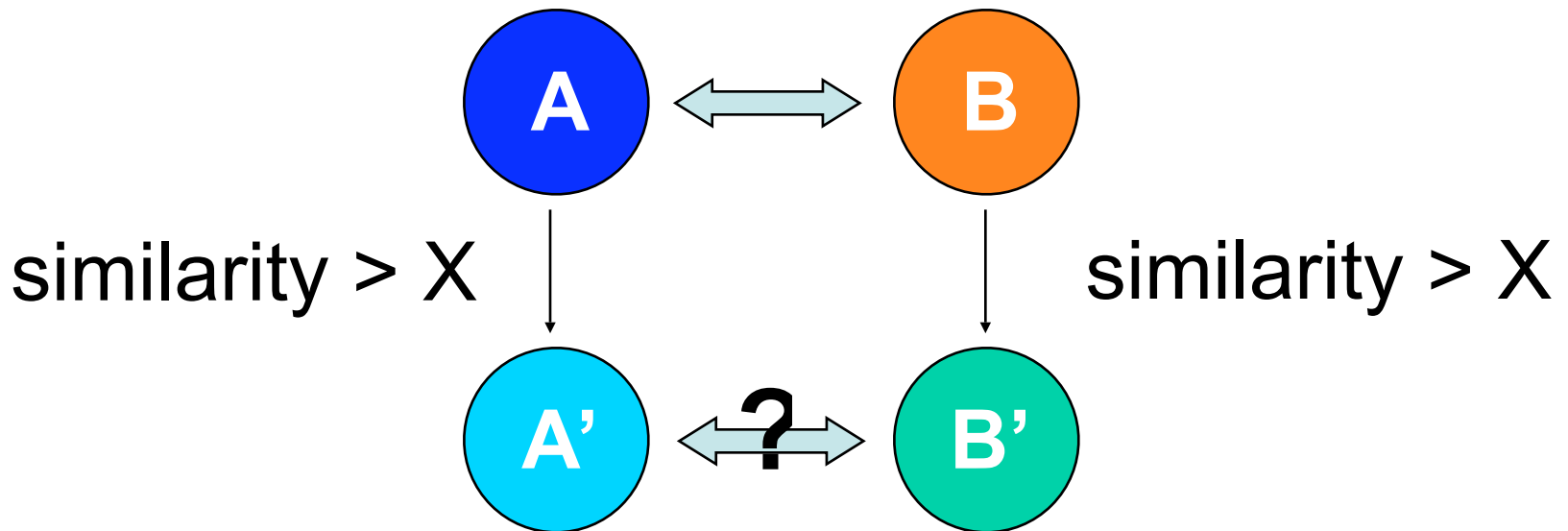
protein 1	protein 2	score
P1	P4	LOW
P1	P90	Neutral
P5	P6	Neutral
P10	P9	High
P12	P4	Neutral
P11	P6	Neutral
P7	P24	Neutral
P5	P7	LOW
P3	P34	LOW
P80	P4	Neutral
P5	P6	High
P1	P2	Neutral
P34	P6	Neutral

V. In passing:

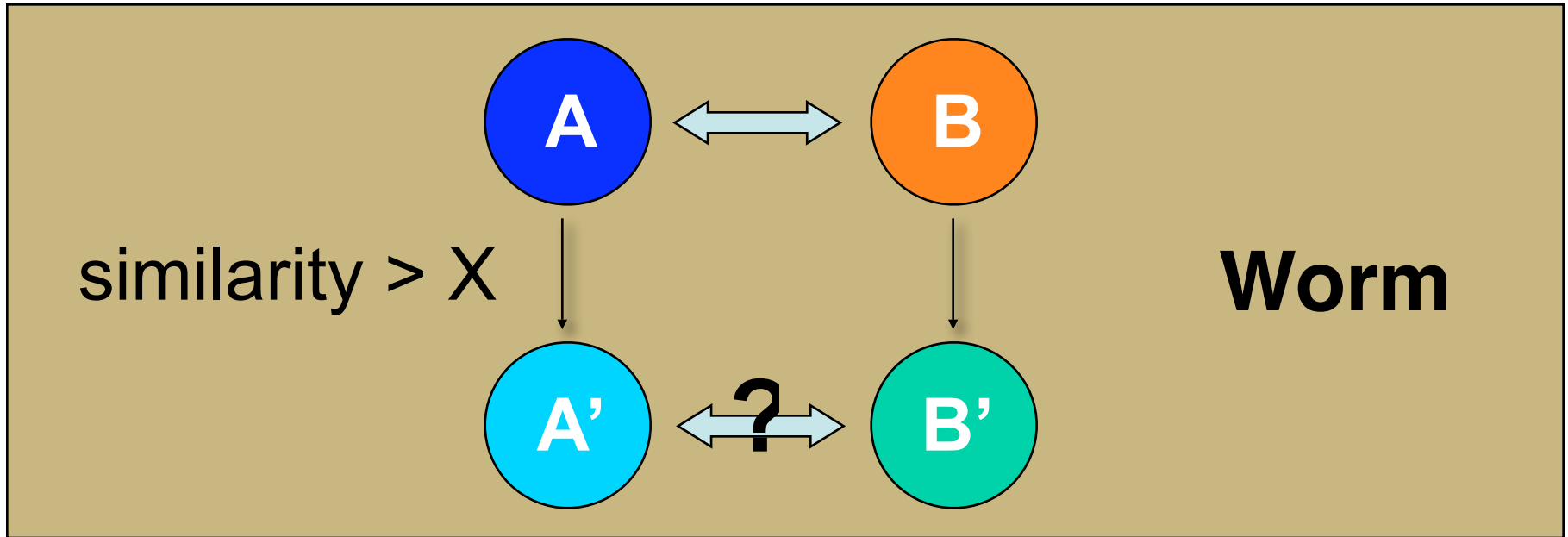
Model organisms pose problems for protein-protein interactions

Can we transfer binding through homology?

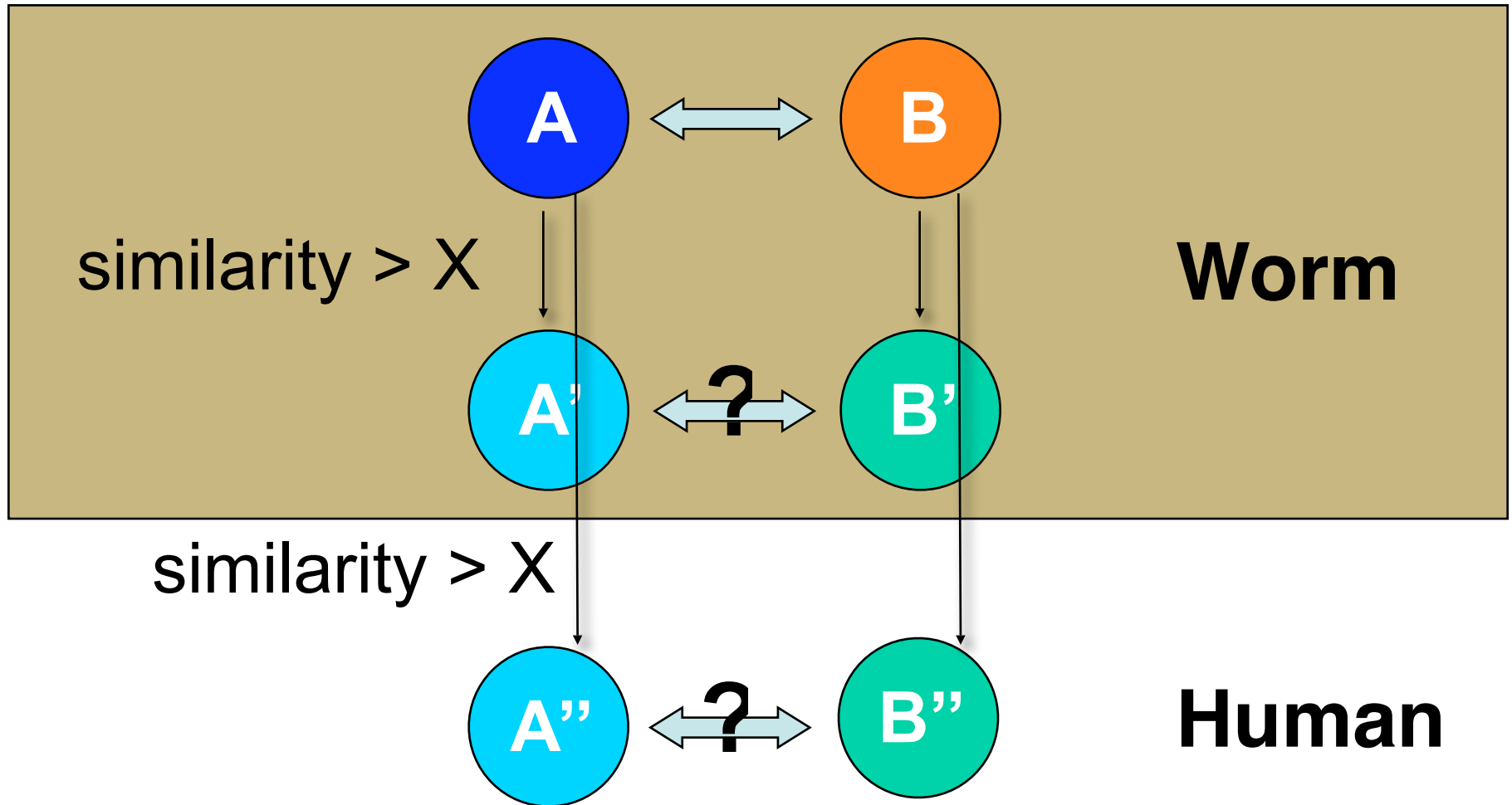
- Obviously, otherwise no value in model organisms ...



Inter and Intra-species the same?

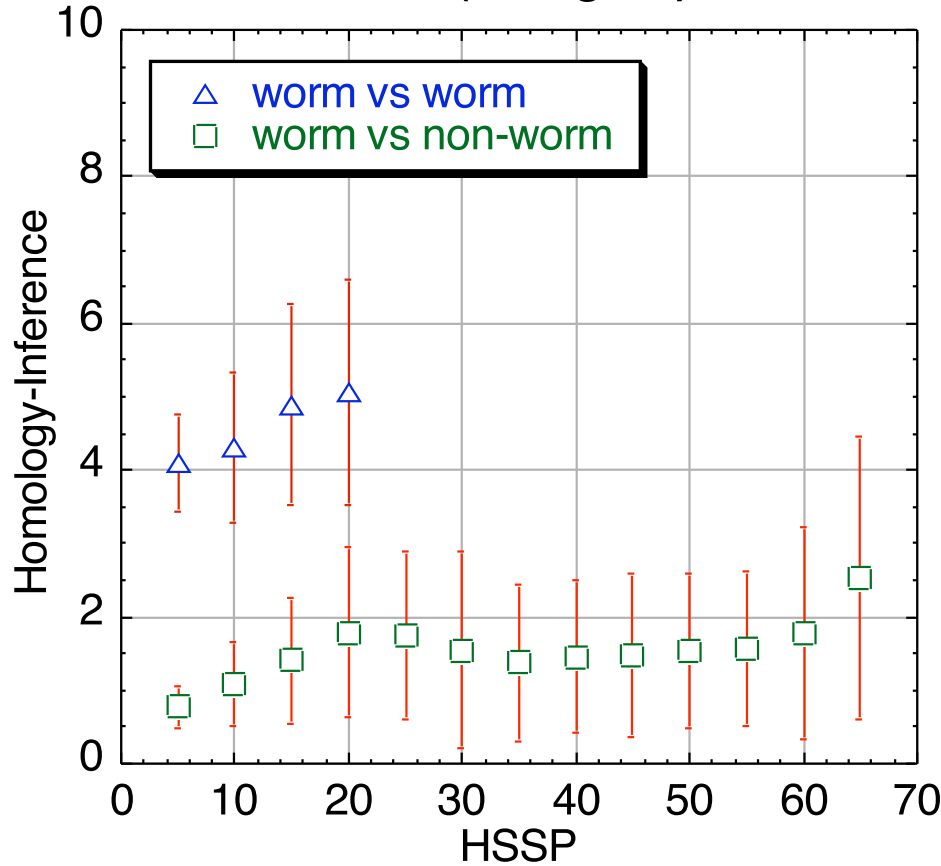


Inter and Intra-species the same?

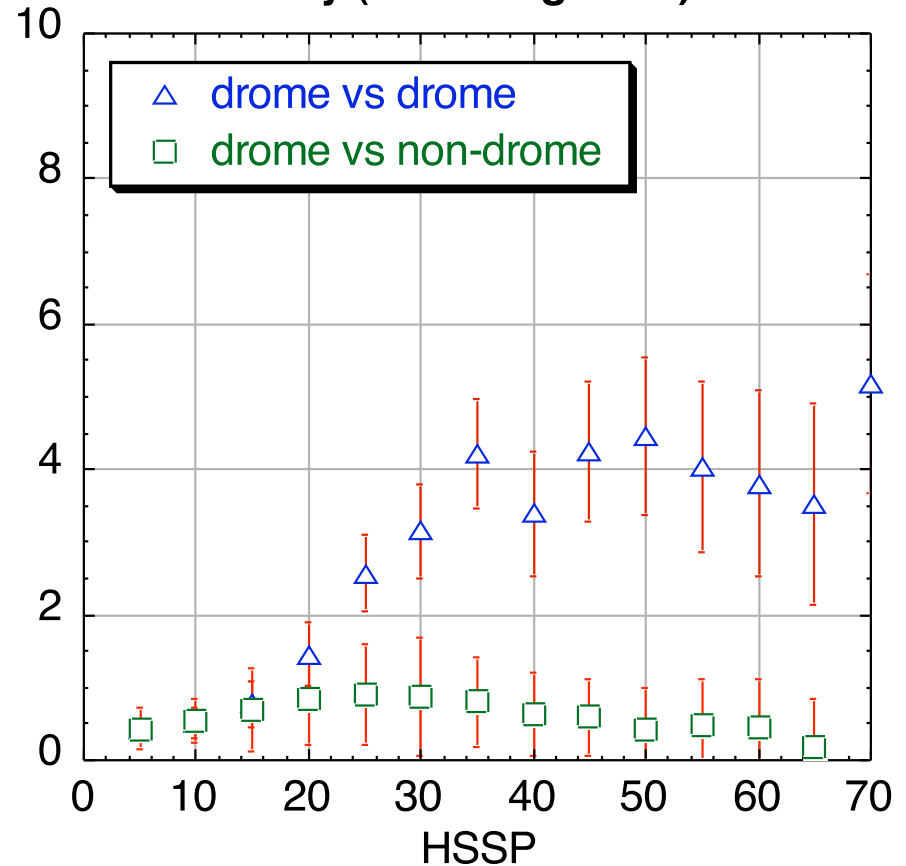


Much better intra-species

worm (C.Elegans)



fruit fly (Drosophila melanogaster)



Excerpt of work papers 2002-2006

1. CAF Andersen *et al.* 2002 *Structure* **10**:175-184
2. CP Chen & B Rost 2002 *Appl Bioinf* **1**:21-35
3. CP Chen *et al.* 2002 *Prot Sci* **11**:2774-2791
4. CP Chen & B Rost 2002 *Prot Sci* **27**:66-2773
5. J Glasgow & B Rost 2002 *Bioinformatics* **18**:S1
6. J Liu & B Rost 2002 *Bioinformatics* **18**:922-933
7. J Liu *et al.* 2002 *J Mol Biol* **322**:53-64
8. MA Marti-Renom *et al.* 2002 *Structure* **10**:435-440
9. R Nair & B Rost 2002 *Bioinformatics* **18**:S78-S86
10. R Nair & B Rost 2002 *Prot Sci* **11**:2836-2847
11. G Pollastri *et al.* 2002 *Proteins* **47**:228-235
12. D Przybylski & B Rost 2002 *Proteins* **46**:195-205
13. B Rost 2002 *J Mol Biol* **318**:595-608
14. B Rost 2002 *Curr Opin Str Biol* **12**:409-416
15. B Rost *et al.* 2002 *Bioinformatics* **18**:897
16. B Rost 2003 *Methods Biochem Anal* **44**:559-587
17. CAF Andersen & B Rost 2003 *Methods Biochem Anal* **44**:341-363
18. B Rost 2003 In *Artificial intelligence and heuristic methods in bioinformatics* (P Frasconi & R Shamir) IOS Press:34-50
19. B Rost *et al.* 2003 In *Handbook of Chemoinformatics - from data to knowledge* (J Gasteiger & T Engel) Wiley-VCH:1789-1811
20. Y Ofran & B Rost 2003 *FEBS Letters* **544**:236-239
21. R Nair *et al.* 2003 *NAR* **31**:397-399
22. P Carter *et al.* 2003 *NAR* **31**:410-413
23. KO Wrzeszczynski & B Rost 2003 In *Cell cycle checkpoint control protocols* (H Lieberman) Humana Press:219-233
24. J Liu & B Rost 2003 *Cur Opinion Chem Biol* **7**:5-11
25. R Zidovetzki *et al.* 2003 *JBC* **15**:555-575
26. Y Ofran & B Rost 2003 *JMB* **325**:377-387
27. IYY Koh *et al.* 2003 *NAR* **31**:3311-3315
28. R Nair & B Rost 2003 *NAR* **31**:3337-3340
29. VA Eyrich & B Rost 2003 *NAR* **31**:3308-3310
30. S Mika & B Rost 2003 *NAR* **31**:3789-3791
31. J Liu & B Rost 2003 *NAR* **31**:3833-3835
32. B Rost & J Liu 2003 *NAR* **31**:3300-3304
33. A Kernytsky & B Rost 2003 *NAR* **31**:3642-3644
34. P Carter *et al.* 2003 *NAR* **31**:3293-3295
35. R Nair & B Rost 2003 *Proteins* **53**:917-930
36. VA Eyrich *et al.* 2003 *Proteins* **53 Suppl 6**:548-560
37. B Rost *et al.* 2003 *CMLS* **60**:2637-2650
38. B Rost 2003 In *Protein structure determination, analysis, and modeling for drug discovery* (D. Chasman) Dekker:207-249
39. JM Aramini *et al.* 2003 *Prot Sci* **12**:2823-2830
40. D Przybylski & B Rost 2004 *JMB* **341**:255-269
41. J Liu & B Rost 2004 *NAR* **32**:3522-3530
42. J Liu *et al.* 2004 *Proteins* **56**:188-200
43. Z Wunderlich *et al.* 2004 *Proteins* **56**:181-187
44. S Mika & B Rost 2004 *Bioinformatics* **20**:1241-7
45. KO Wrzeszczynski & B Rost 2004 *CMLS* **61**:1341-1353
46. R Nair & B Rost 2004 *AI Magazine* **25**:45-56
47. J Liu & B Rost 2004 *Proteins* **55**:678-688
48. B Rost *et al.* 2004 *NAR* **32**:W321-W326
49. S Mika & B Rost 2004 *NAR* **32**:W634-W637
50. H Bigelow *et al.* 2004 *NAR* **32**:2566-2577
51. R Nair & B Rost 2004 *NAR* **32**:W517-W521
52. J Liu & B Rost 2004 *NAR* **32**:W569-W571
53. J Glasgow *et al.* 2004 *AI Magazine* **25**:7-8
54. KO Wrzeszczynski & B Rost 2004 *Meth Mol Biol* **241**:219-233
55. Z Wunderlich *et al.* 2004 *Proteins* **56**:181-7
56. R Powers *et al.* 2004 *J Biomolecular NMR* **30**:107-108
57. B Rost 2005 In *The Proteomics Protocols Handbook* (J. Walker) 875-901
58. A Schlessinger & B Rost 2005 *Proteins* **61**:115-126
59. Y Ofran & B Rost 2005 In *Bioinformatics* (A. D. Baxevanis and B. F. Ouellette) Wiley:197-222
60. S Mika & B Rost 2005 *NAR* **33**:D160-163
61. R Nair & B Rost 2005 *JMB* **348**:85-100
62. M Punta & B Rost 2005 *Bioinformatics* **21**:2960-2968
63. M Punta & B Rost 2005 *JMB* **348**:507-512
64. J Benach *et al.* 2005 *Acta Crystallogr D Biol Crystallogr* **61**:589-98
65. Grana *et al.* 2005 *Nucleic Acids Res* **33**:W347-51
66. HV Jagadish *et al.* 2005 *Bioinformatics* **21 Suppl 1**:i1-i2
67. A Schlessinger & B Rost 2005 *Proteins* **61**:115-26
68. The FANTOM Consortium 2005 *Science* **309**:1559-1563
69. Y Ofran, M Punta, R Schneider & B Rost 2005 *Drug Disc Today* **10**:1475-1482
70. R Powers *et al.* 2005 *Protein Science* **14**:2849-61
71. DA Snyder *et al.* 2005 *J Am Chem Soc* **127**:16505-16511
72. J Moulton *et al.* 2005 *Proteins* **61**:3-7
73. O Grana *et al.* 2006 *Proteins* **61**:214-224
74. A Schlessinger, Y Ofran, G Yachdav & B Rost 2006 *NAR* **34**:D777-D780
75. J Liu, J Gough & B Rost 2005 *PLoS Genetics* in press
76. R Nair & B Rost 2006 *In silico technology in drug target identification and validation* (Eds. D Leon & S Markel) Boca Raton, FL: CRC Press, in press.
77. D Przybylski & B Rost 2006 In *Bioinformatics - From Genomes to Therapies* (T Lengauer) Weinheim: Wiley-VCH, in press
78. Y Ofran & B Rost 2006 submitted 2004
79. S Mika & B Rost 2005 *PLoS Comp Biol* submitted
80. Y Ofran & B Rost 2006 in preparation

Excerpt of work papers 2002-2006

1. CAF Andersen *et al.* 2002 *Structure* **10**:175-184
2. CP Chen & B Rost 2002 *Appl Bioinf* **1**:21-35
3. CP Chen *et al.* 2002 *Prot Sci* **11**:2774-2791
4. CP Chen & B Rost 2002 *Prot Sci* **27**:66-2773
5. J Glasgow & B Rost 2002 *Bioinformatics* **18**:S1
6. J Liu & B Rost 2002 *Bioinformatics* **18**:922-933
7. J Liu *et al.* 2002 *J Mol Biol* **322**:53-64
8. MA Marti-Renom *et al.* 2002 *Structure* **10**:435-440
9. R Nair & B Rost 2002 *Bioinformatics* **18**:S78-S86
10. R Nair & B Rost 2002 *Prot Sci* **11**:2836-2847
11. G Pollastri *et al.* 2002 *Proteins* **47**:228-235
12. D Przybylski & B Rost 2002 *Proteins* **46**:195-205
13. B Rost 2002 *J Mol Biol* **318**:595-608
14. B Rost 2002 *Curr Opin Str Biol* **12**:409-416
15. B Rost *et al.* 2002 *Bioinformatics* **18**:897
16. B Rost 2003 *Methods Biochem Anal* **44**:559-587
17. CAF Andersen & B Rost 2003 *Methods Biochem Anal* **44**:341-363
18. B Rost 2003 In *Artificial intelligence and heuristic methods in bioinformatics* (P Frasconi & R Shamir) IOS Press:34-50
19. B Rost *et al.* 2003 In *Handbook of Chemoinformatics - from data to knowledge* (J Gasteiger & T Engel) Wiley-VCH:1789-1811
20. Y Ofran & B Rost 2003 *FEBS Letters* **544**:236-239
21. R Nair *et al.* 2003 *NAR* **31**:397-399
22. P Carter *et al.* 2003 *NAR* **31**:410-413
23. KO Wrzeszczynski & B Rost 2003 In *Cell cycle checkpoint control protocols* (H Lieberman) Humana Press:219-233
24. J Liu & B Rost 2003 *Cur Opinion Chem Biol* **7**:5-11
25. R Zidovetzki *et al.* 2003 *JBC* **15**:555-575
26. Y Ofran & B Rost 2003 *JMB* **325**:377-387
27. IYY Koh *et al.* 2003 *NAR* **31**:3311-3315
28. R Nair & B Rost 2003 *NAR* **31**:3337-3340
29. VA Eyrich & B Rost 2003 *NAR* **31**:3308-3310
30. S Mika & B Rost 2003 *NAR* **31**:3789-3791
31. J Liu & B Rost 2003 *NAR* **31**:3833-3835
32. B Rost & J Liu 2003 *NAR* **31**:3300-3304
33. A Kernytsky & B Rost 2003 *NAR* **31**:3642-3644
34. P Carter *et al.* 2003 *NAR* **31**:3293-3295
35. R Nair & B Rost 2003 *Proteins* **53**:917-930
36. VA Eyrich *et al.* 2003 *Proteins* **53** Suppl 6:560
37. B Rost *et al.* 2003 *CMLS* **60**:2637-2650
38. B Rost 2003 In *Protein structure determination, analysis, and modeling for drug discovery* (Chasman) Dekker:207-249
39. JM Aramini *et al.* 2003 *Prot Sci* **12**:2823-2830
40. D Przybylski & B Rost 2004 *JMB* **341**:255-260
41. J Liu & B Rost 2004 *NAR* **32**:3522-3530
42. J Liu *et al.* 2004 *Proteins* **56**:188-200
43. Z Wunderlich *et al.* 2004 *Proteins* **56**:181-187
44. S Mika & B Rost 2004 *Bioinformatics* **20**:1241-7
45. KO Wrzeszczynski & B Rost 2004 *CMLS* **73**:1341-1353
46. R Nair & B Rost 2004 *AI Magazine* **25**:45-56
47. J Liu & B Rost 2004 *Proteins* **55**:678-688
48. B Rost *et al.* 2004 *NAR* **32**:W321-W326
49. S Mika & B Rost 2004 *NAR* **32**:W634-W637
50. H Bigelow *et al.* 2004 *NAR* **32**:2566-2577
51. R Nair & B Rost 2004 *NAR* **32**:W517-W521
52. J Liu & B Rost 2004 *NAR* **32**:W569-W571
53. J Glasgow *et al.* 2004 *AI Magazine* **25**:7-8
54. KO Wrzeszczynski & B Rost 2004 *Meth Mol Biol* **241**:219-233
55. Z Wunderlich *et al.* 2004 *Proteins* **56**:181-7
56. R Powers *et al.* 2004 *J Biomolecular NMR* **30**:107-108
57. B Rost 2005 In *The Proteomics Protocols Handbook* (J. Walker) 875-901
58. A Schlessinger & B Rost 2005 *Proteins* **61**:115-126
59. Y Ofran & B Rost 2005 In *Bioinformatics* (A. D. Baxevanis and B. F. Ouellette) Wiley:197-222
60. S Mika & B Rost 2005 *NAR* **33**:D160-163
61. R Nair & B Rost 2005 *JMB* **348**:85-100
62. M P... 2005 *Bioinformatics* **21**:2960-...
63. ... 2005:507-512
64. ... *Crystallogr D Biol* ...
65. ... 2005:W347-51
66. ... *Proteins* **21** Suppl ...
67. ... 2005:115-26
68. **The FANTOM Consortium 2005**
Science
309:1559-1563
69. ... Rost 2005
70. ... 2005:14:2849-61
71. ... *Chem Soc* ...
72. ... 2005:3-7
73. ... 2005:61:214-224
74. A Schlessinger, Y Ofran, G Yachdav & B Rost 2006 *NAR* **34**:D777-D780
75. J Liu, J Gough & B Rost 2005 *PLoS Genetics* in press
76. R Nair & B Rost 2006 *In silico technology in drug target identification and validation* (Eds. D Leon & S Markel) Boca Raton, FL: CRC Press, in press.
77. D Przybylski & B Rost 2006 In *Bioinformatics - From Genomes to Therapies* (T Lengauer) Weinheim: Wiley-VCH, in press
78. Y Ofran & B Rost 2006 submitted 2004
79. S Mika & B Rost 2005 *PLoS Comp Biol* submitted
80. Y Ofran & B Rost 2006 in preparation

Excerpt of work papers 2002-2006

1. CAF Andersen *et al.* 2002 *Structure* **10**:175-184
2. CP Chen & B Rost 2002 *Appl Bioinf* **1**:21-35
3. CP Chen *et al.* 2002 *Prot Sci* **11**:2774-2791
4. CP Chen & B Rost 2002 *Prot Sci* **27**:66-2773
5. J Glasgow & B Rost 2002 *Bioinformatics* **18**:S1
6. J Liu & B Rost 2002 *Bioinformatics* **18**:922-933
7. J Liu *et al.* 2002 *J Mol Biol* **322**:53-64
8. MA Marti-Renom *et al.* 2002 *Structure* **10**:435-440
9. R Nair & B Rost 2002 *Bioinformatics* **18**:S78-S86
10. R Nair & B Rost 2002 *Prot Sci* **11**:2836-2847
11. G Pollastri *et al.* 2002 *Proteins* **47**:228-235
12. D Przybylski & B Rost 2002 *Proteins* **46**:19
13. B Rost 2002 *J Mol Biol* **318**:595-608
14. B Rost 2002 *Curr Opin Str Biol* **12**:409-41
15. B Rost *et al.* 2002 *Bioinformatics* **18**:897
16. B Rost 2003 *Methods Biochem Anal* **44**:5
17. CAF Andersen & B Rost 2003 *Methods Anal* **44**:341-363
18. B Rost 2003 In *Artificial intelligence and methods in bioinformatics* (P Frasconi & R S IOS Press:34-50
19. B Rost *et al.* 2003 In *Handbook of Chemoinformatics - from data to knowledge* (J Gasteiger & T Engel) Wiley-VCH:1789-1811
20. Y Ofran & B Rost 2003 *FEBS Letters* **544**:236-239
21. R Nair *et al.* 2003 *NAR* **31**:397-399
22. P Carter *et al.* 2003 *NAR* **31**:410-413
23. KO Wrzeszczynski & B Rost 2003 In *Cell cycle checkpoint control protocols* (H Lieberman) Humana Press:219-233
24. J Liu & B Rost 2003 *Cur Opinion Chem Biol* **7**:5-11
25. R Zidovetzki *et al.* 2003 *JBC* **15**:555-575
26. Y Ofran & B Rost 2003 *JMB* **325**:377-387
27. IYY Koh *et al.* 2003 *NAR* **31**:3311-3315
28. R Nair & B Rost 2003 *NAR* **31**:3337-3340
29. VA Eyrich & B Rost 2003 *NAR* **31**:3308-3310
30. S Mika & B Rost 2003 *NAR* **31**:3789-3791
31. J Liu & B Rost 2003 *NAR* **31**:3833-3835
32. B Rost & J Liu 2003 *NAR* **31**:3300-3304
33. A Kernytsky & B Rost 2003 *NAR* **31**:3642-3644
34. P Carter *et al.* 2003 *NAR* **31**:3293-3295
35. R Nair & B Rost 2003 *Proteins* **53**:917-930
36. VA Eyrich *et al.* 2003 *Proteins* **53 Suppl 6**:548-560
37. B Rost *et al.* 2003 *J Mol Biol* **30**:2637-2650
38. B Rost *et al.* 2003 *Structure determination, drug discovery* (D. S. 2823-2830
39. B Rost *et al.* 2003 *Proteins* **47**:255-269
40. D Przybylski & B Rost 2004 *JMB* **341**:255-269
41. B Rost *et al.* 2004 *J Mol Biol* **330**
42. B Rost *et al.* 2004 *J Mol Biol* **331**:187
43. B Rost *et al.* 2004 *J Mol Biol* **330**:1241-7
44. B Rost *et al.* 2004 *CML S*
45. B Rost *et al.* 2004 *Proteins* **52**:45-56
46. B Rost *et al.* 2004 *Proteins* **52**:678-688
47. B Rost *et al.* 2004 *Proteins* **52**:W326
48. B Rost *et al.* 2004 *Proteins* **52**:W634-W637
49. B Rost *et al.* 2004 *Proteins* **52**:2566-2577
50. B Rost *et al.* 2004 *NAR* **32**:W517-W521
51. R Nair *et al.* 2004 *NAR* **32**:W569-W571
52. J Liu & B Rost 2004 *NAR* **32**:W569-W571
53. J Glasgow *et al.* 2004 *AI Magazine* **25**:7-8
54. KO Wrzeszczynski & B Rost 2004 *Meth Mol Biol* **241**:219-233
55. Z Wunderlich *et al.* 2004 *Proteins* **56**:181-7
56. R Powers *et al.* 2004 *J Biomolecular NMR* **30**:107-108
57. B Rost 2005 In *The Proteomics Protocols Handbook* (J. Walker) 875-901
58. A Schlessinger & B Rost 2005 *Proteins* **61**:115-126
59. Y Ofran & B Rost 2005 In *Bioinformatics* (A. D. Baxevanis and B. F. Ouellette) Wiley:197-222
60. S Mika & B Rost 2005 *NAR* **33**:D160-163
61. R Nair & B Rost 2005 *JMB* **348**:85-100
62. M Punta & B Rost 2005 *Bioinformatics* **21**:2960-2968
63. M Punta & B Rost 2005 *JMB* **348**:507-512
64. J Benach *et al.* 2005 *Acta Crystallogr D Biol Crystallogr* **61**:589-98
65. Grana *et al.* 2005 *Nucleic Acids Res* **33**:W347-51
66. HV Jagadish *et al.* 2005 *Bioinformatics* **21 Suppl 1**:i1-i2
67. A Schlessinger & B Rost 2005 *Proteins* **61**:115-26
68. The FANTOM Consortium 2005 *Science* **309**:1559-1563
69. Y Ofran, M Punta, R Schneider & B Rost 2005 *Drug Disc Today* **10**:1475-1482
70. R Powers *et al.* 2005 *Protein Science* **14**:2849-61
71. DA Snyder *et al.* 2005 *J Am Chem Soc* **127**:16505-16511
72. J Moulton *et al.* 2005 *Proteins* **61**:3-7
73. O Grana *et al.* 2006 *Proteins* **61**:214-224
74. A Schlessinger, Y Ofran, G Yachdav & B Rost 2006 *NAR* **34**:D777-D780
75. J Liu, J Gough & B Rost 2005 *PLoS Genetics* in press
76. R Nair & B Rost 2006 *In silico technology in drug target identification and validation* (Eds. D Leon & S Markel) Boca Raton, FL: CRC Press, in press.
77. D Przybylski & B Rost 2006 In *Bioinformatics - From Genomes to Therapies* (T Lengauer) Weinheim: Wiley-VCH, in press
78. Y Ofran & B Rost 2006 submitted 2004
79. S Mika & B Rost 2005 *PLoS Comp Biol* submitted
80. Y Ofran & B Rost 2006 in preparation

Conclusions

- **Transient protein-protein interfaces specific
-> specific prediction very accurate**
- **Localization predicted at levels of accuracy
similar to high-throughput experiments**
- **Structural genomics is increasingly impacting
biology; it builds on computational biology**
- **Evolution provides the key for *de
novo* prediction of (protein) function**

Predict function from sequence+structure

● Molecular level

● Localization

● Protein-X interactions

● System level

● Networks

● Pathways

*THIS
is the
beginning*



Claudia Bertoni



Henry Bigelow



Kaz Wrzeszczynski



Yanay Ofran

Ta-Tsen Soong

Yana



Avner Schlessinger



Jinfeng Liu



Ingrid Koh



Volker Eyrich

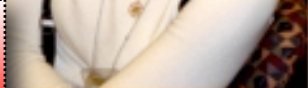


Marco Punta



Eyal Mozes

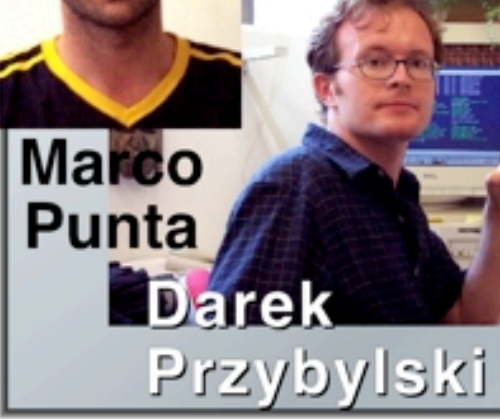
Sara Gilman



Yana Bromberg



Sven Mika



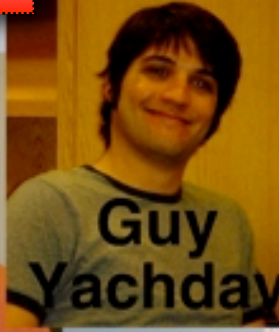
Darek Przybylski



Raj Nair



Andrew Kernytsky



Guy Yachday



Phil Carter

Thanksgiving



Group (left): Claus AF Andersen, Hepan Tang, Murat Cokol, Trevor Siggers, Chen Peter Chien, Shoshanna Posy, Venkatesh Mysore

STRX: Guy Montelione (Rutgers), Diana Murray (Cornell, NYC), Tom Acton (Rutgers), Liang Tong & John Hunt (Columbia), George DeTitta (Buffalo), Cheryl Arrowsmith (Toronto), Wayne Hendrickson (Columbia)
















General CU: Barry Honig, Ann McDermott, Art Palmer, David Hirsh, Yoav Freund, Yechiam Yemini, Dimitris Thanos, Richard Mann, Richard Axel, Eric Kandel, Max Gottesmann, Oliver Hobert, Iva Greenwald, Marty Chalfie, Larry Shapiro, Christine Leslie, Dimitris Anastassiou

EVA: Andrej Sali (UCSF), Alfonso Valencia (Madrid)

ASF: Anna Tramantano (Rome), Terry Gaasterland (UCSD), Reinhard Schneider (EMBL), Chris Sander (Sloan), Debbie Marks (Harvard)

Karima Djabali, Lena Rezkia Inge Rost

X=<http://www.rostlab.org>

 PredictProtein PP	X/predictprotein/
 META-PP	X/meta/submit_meta.html
 EVA	X/eva/
 services:	X/services/
 LOctree	X/services/loctree/
 PredictNLS	X/predictNLS/
 NORSp	X/services/norsp/
 DSSPcont	X/services/dsspcont/
 NLProt	X/services/nlprot/
 CHOP/CHOPnet	X/services/chop/
 ISIS	X/services/isis/
 databases:	X/db/
 PEP	X/db/PEP/
 CellCycleDB	X/db/cellcycledb/
 NMPdb	X/db/nmpdb/

