On Applications of CI in Life Sciences:

Stories from the Field of Protein Structure and Function Prediction

Jarek Meller

Departments of Environmental Health and Biomedical Engineering, University of Cincinnati & Division of Biomedical Informatics, Cincinnati Children's Hospital Research Foundation & Department of Informatics, Nicholas Copernicus University, Toruń

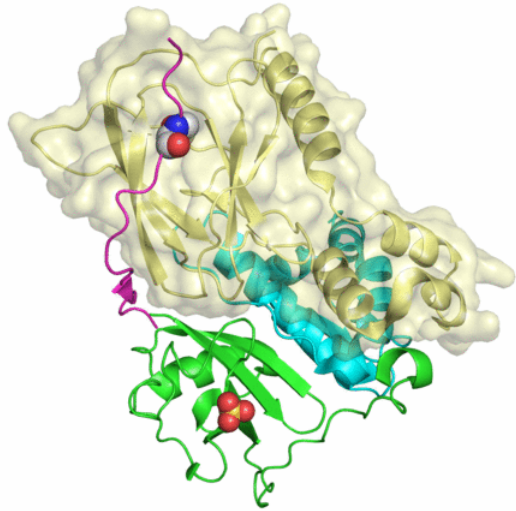
Joint work with Rafal Adamczak, Aleksey Porollo, Baoqiang Cao, Mukta Phatak, and Michael Wagner

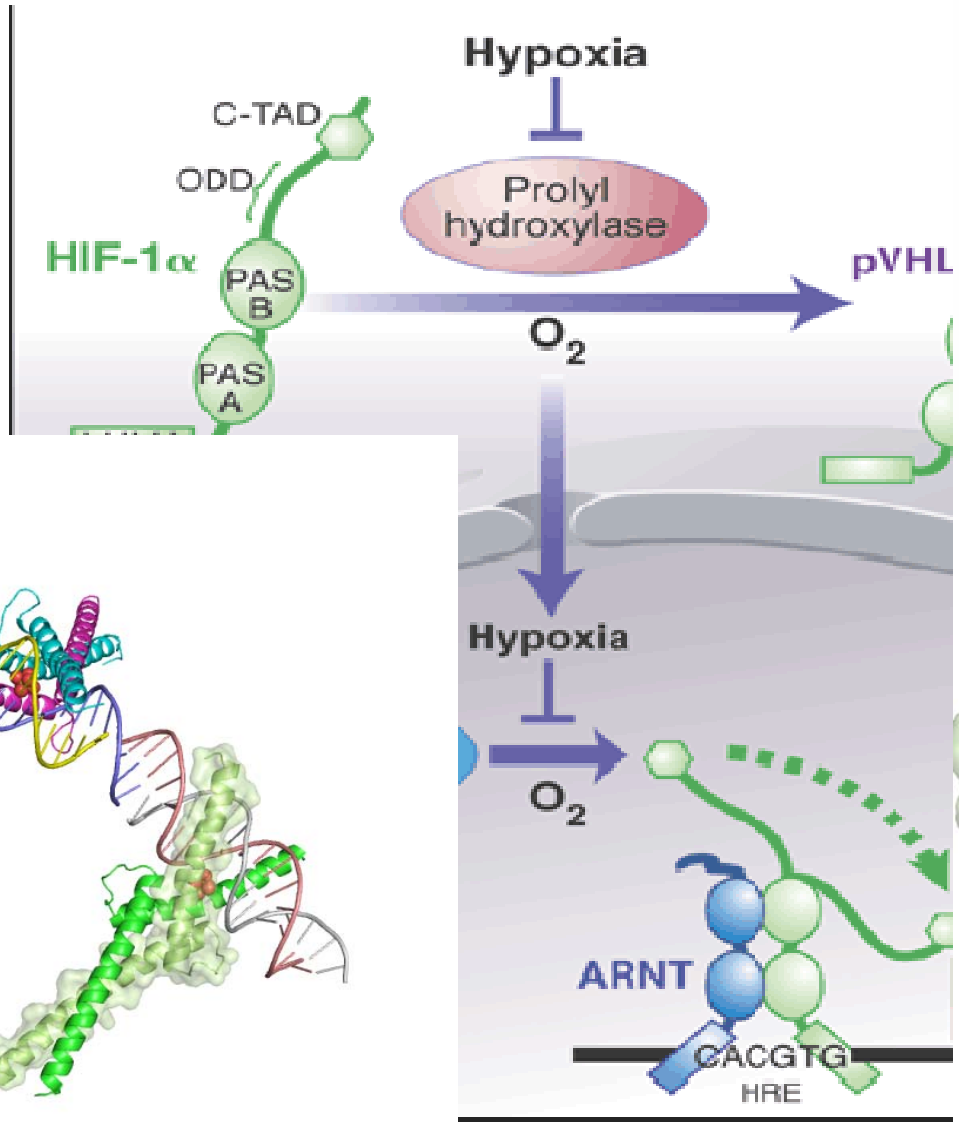
Outline: Some Lessons from Our Attempts to Improve Protein Structure and Function Prediction

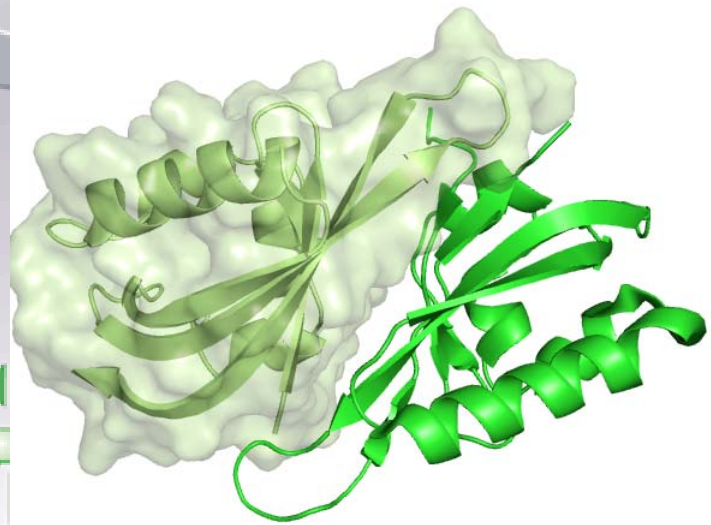
- Our general (knowledge-based) approach: from sequence to structure to function using Artificial Intelligence and Machine Learning as short cuts
- The importance of intermediate attributes such as solvent accessibility: functional predictions in the absence of the overall 3D structure
- Some lessons from our attempts to improve solvent accessibility prediction
- Generalizations for membrane proteins: limited data to extrapolate from
- Accurate recognition of transmembrane segments using prediction of (aqueous) solvent accessibility: an example of a non-trivial initial transformation and dimensionality reduction of input data
- Some more lessons from our attempts to learn from limited data: prediction of lipid accessibility in membrane proteins
- The other story (to be covered some other time) on genome-wide association studies: correlating genotypes and phenotypes using machine and statistical learning, and dealing with even bigger problems: millions of variables (genetic markers) with limited number of data points (patients/genotypes) and fuzzy phenotypes

Hypoxia-induced stabilization of Hif-

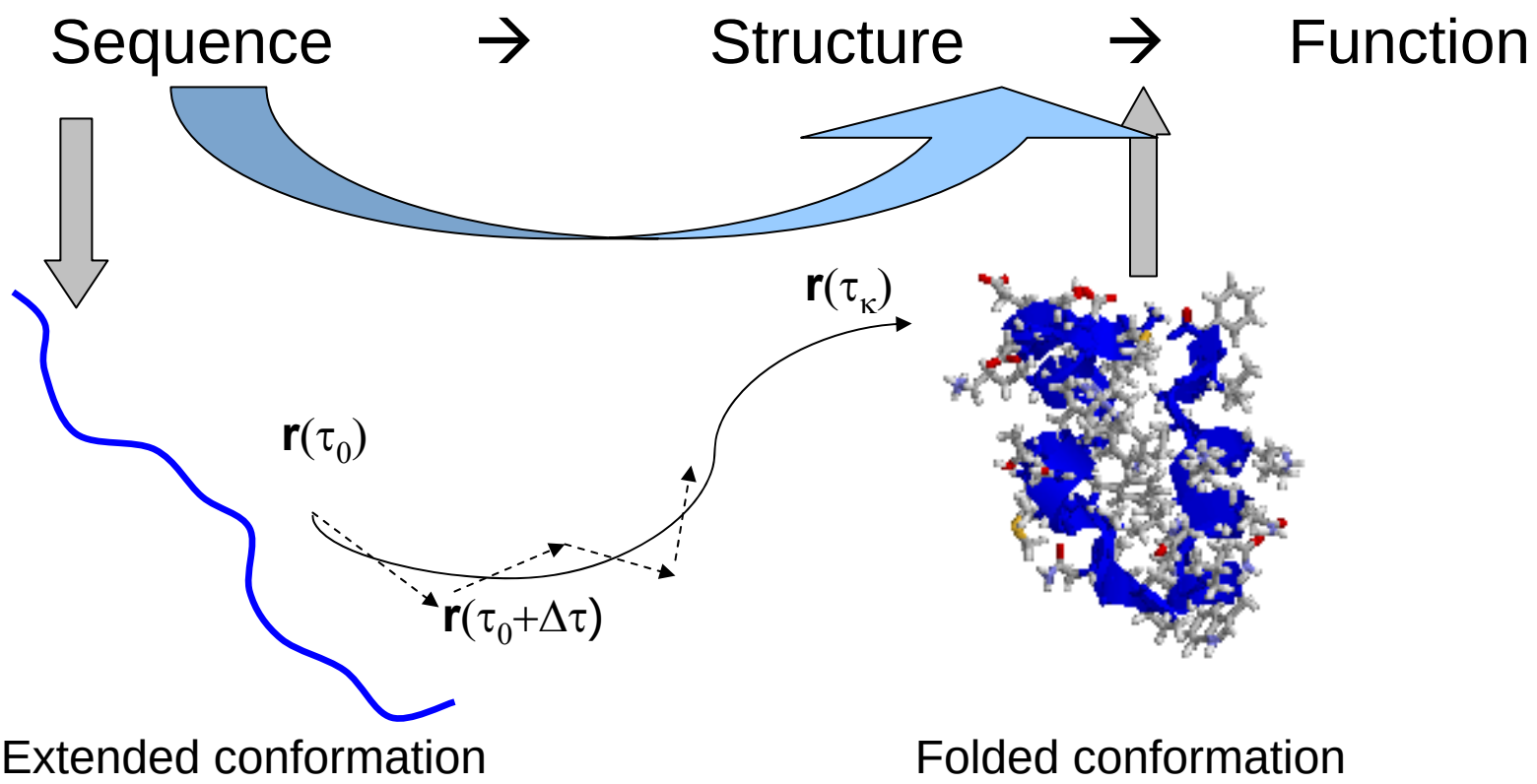
Graphics from R.K. Bruick and S.L.McKnight, Science 295





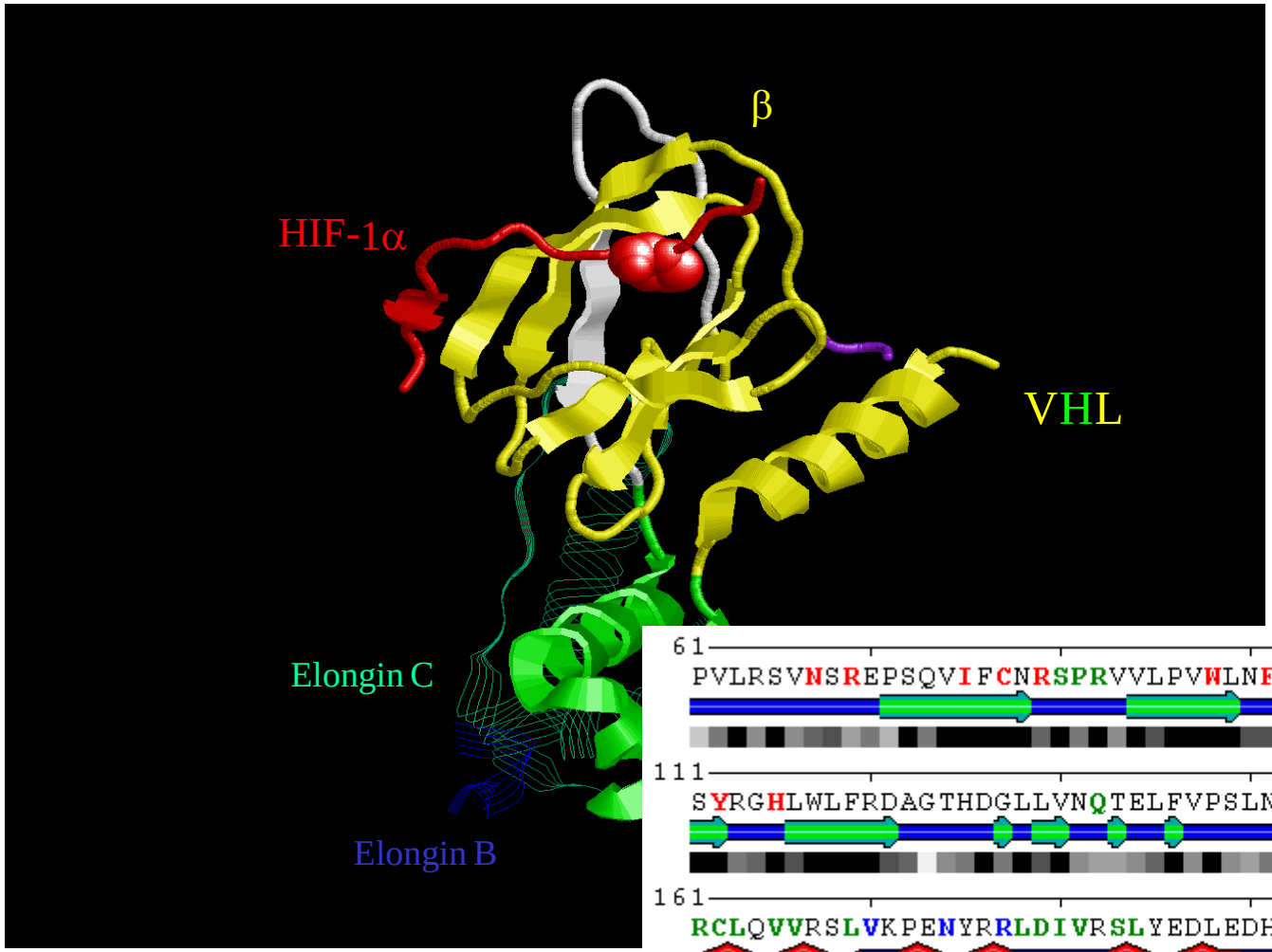


From Sequence to Structure to Function: Protein Folding Problem and CI/AI/ML Short Cuts



Machine learning to the rescue: correlating complex patterns in sequence with structural/functional outcomes using known examples.

Von Hippel-Lindau (VHL) Tumor Suppressor



Folding, stability and functional hot spots (interaction sites) …

Important Example: Predicting Protein Secondary Structures from Amino Acid Sequence

Successful applications of machine learning techniques for secondary structure prediction involve:

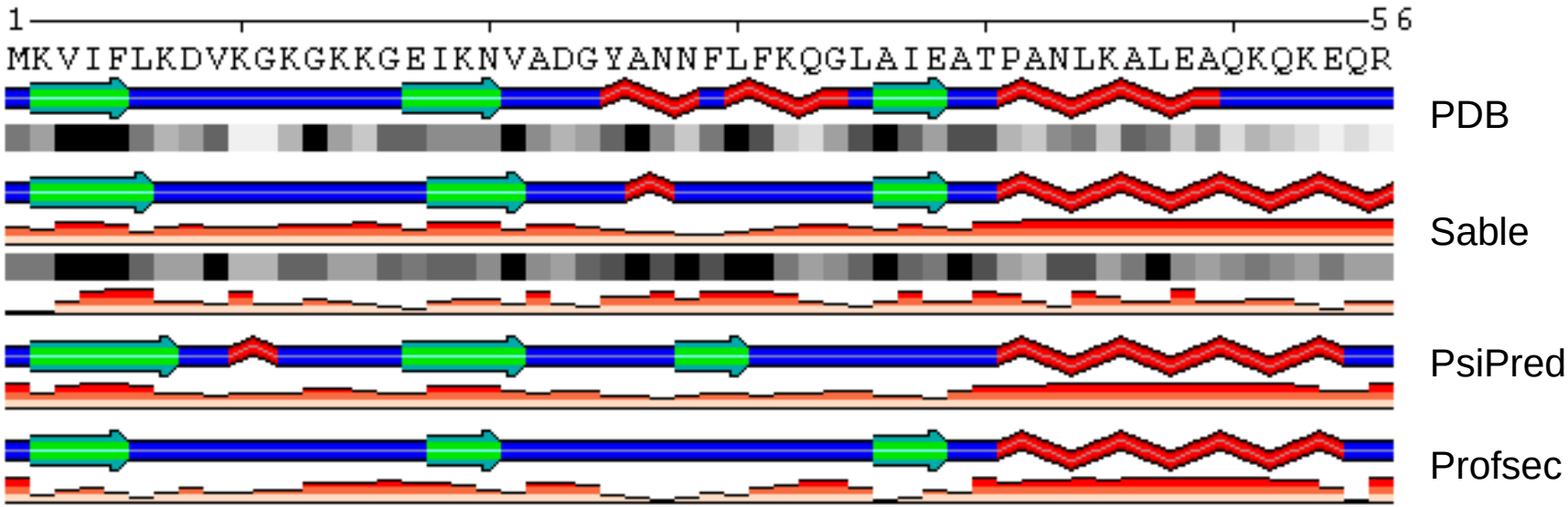
i) multiple alignment and family profile-based representation of local environment and structural propensities (Rost and Sander, Jones);

ii) the use of advanced machine learning techniques, such as Neural Networks (Rost and Sander, Qian and Sejnowski)

However, the latter is far less critical, and, in fact, NN, HMM or SVM-based methods all reach comparable accuracy if trained properly.

State-of-the-art secondary structure prediction methods yield classification accuracies of up to 80% for three state (H, E, C) problem.

Predicting Secondary Structures from Sequence



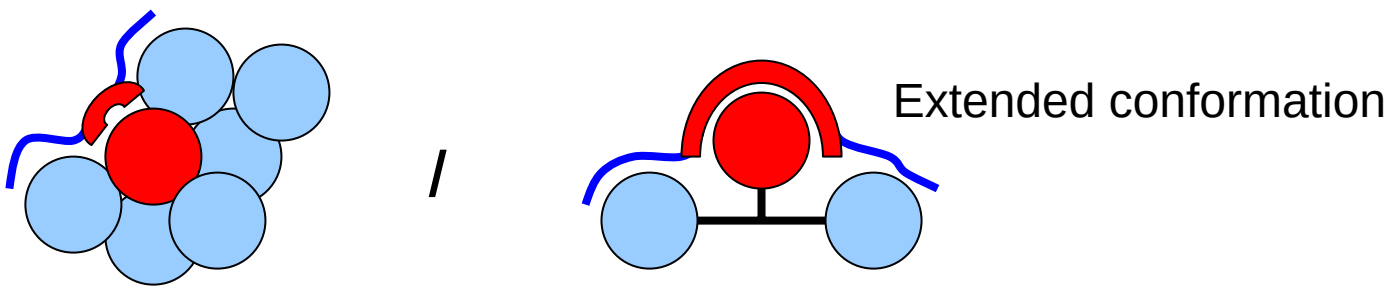
Tailored measures of accuracy, e.g., Segment Overlap Measure

Another Intermediate Problem: Which Residues Are Accessible to Solvent and Interaction Partners?

Relative **S**olvent **A**ccessibility of an amino acid residue in a protein quantifies the degree of exposure (surface exposed area, SEA) to solvent molecules in relative terms:

$$RSA = SEA / MAX\_SEA ; \qquad 0 <= SEA <= MAX\_SEA$$

Thus, RSA is a real valued number in the interval [0,1], which for convenience may be scaled to take the values between 0% and 100%, where 0% corresponds to fully buried and 100% to fully exposed residues, respectively. In membrane domains, lipid replaces water as the solvent, and **R**elative **L**ipid **A**ccessibility can be defined as above.



RSA Prediction: Regression vs. Classification

Classification approaches: relative solvent accessibility prediction is cast as a classification problem, i.e., the real valued RSA is **discretized** with two classes of residues (buried vs. exposed) distinguished by an arbitrary threshold, e.g., 25% RSA, classification accuracy above 70% (PHDacc (Rost and Sander), ACCpro (Pollastri et al.), Jnet (Cuff and Burton)).

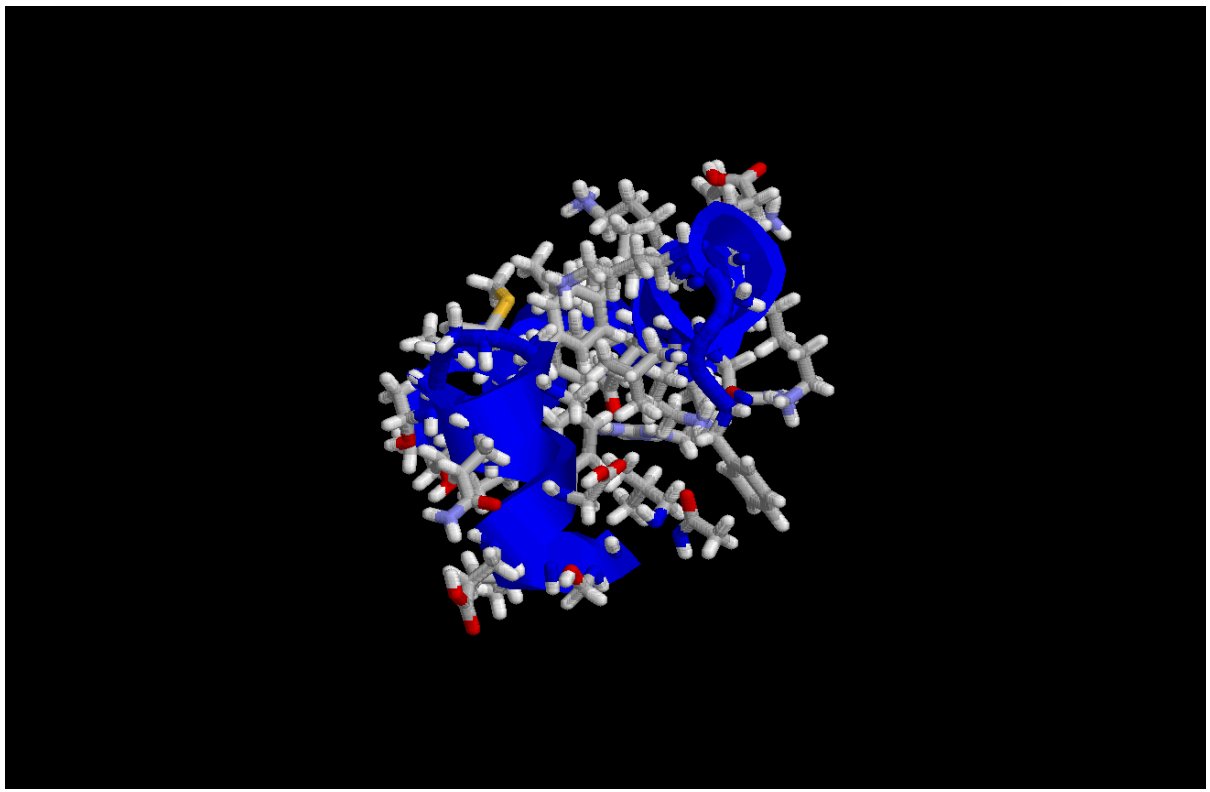
Classification approach to RSA prediction is not only somewhat clumsy but also inconsistent with the level of thermal fluctuations, conformational flexibility and resulting variations in observed RSA, e.g., in protein families.

Regression approaches:

PROF: B. Rost, unpublished
RVPNet: S. Ahmad, M. M. Gromiha, and A. Sarai, *Proteins* 50 (2003)
SABLE: R. Adamczak, A. Porollo, and J. Meller, Proteins 56 (2004)

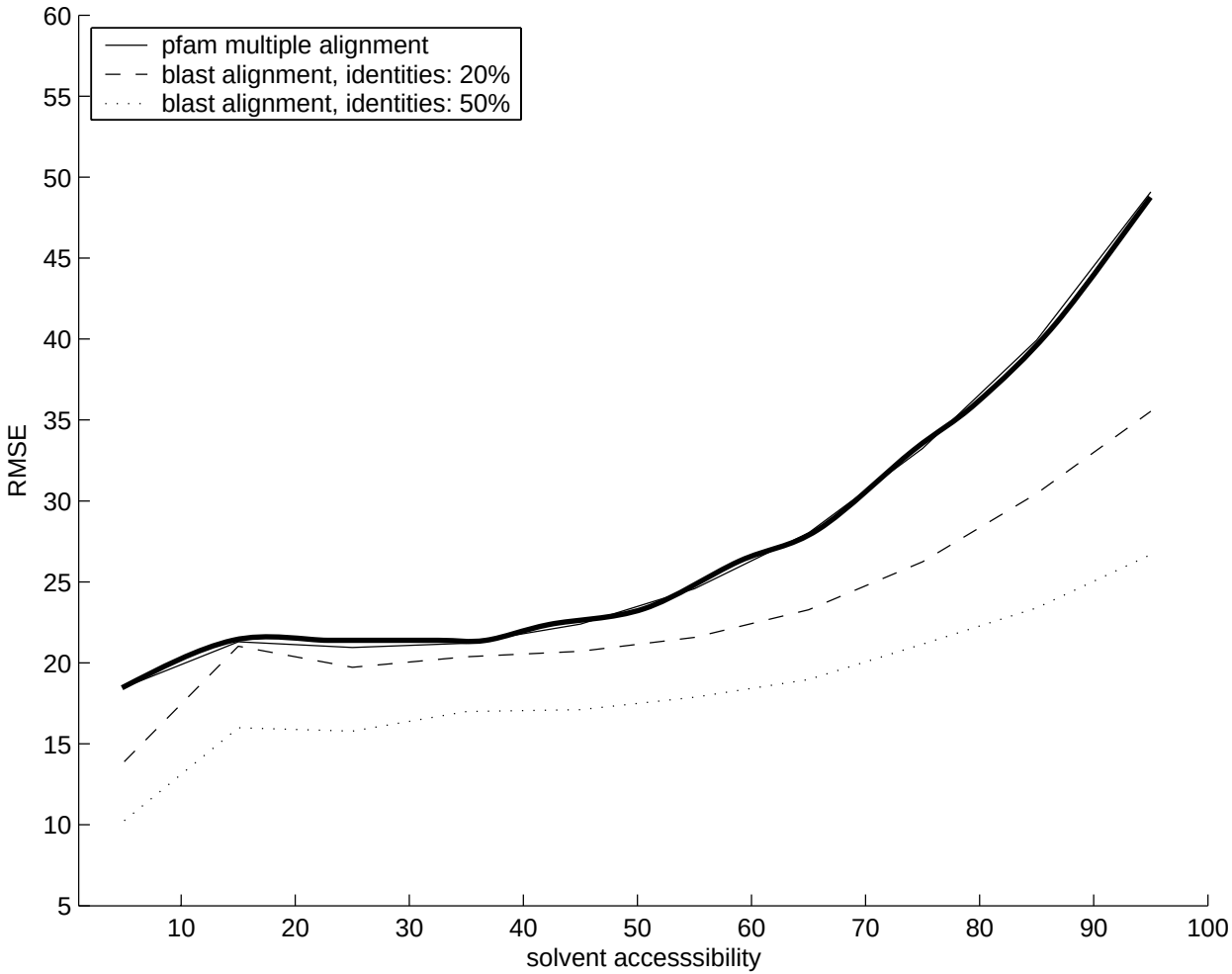
Ensemble of conformations in solution:

top NMR models for the villin headpiece domain and Hif PAS dimer

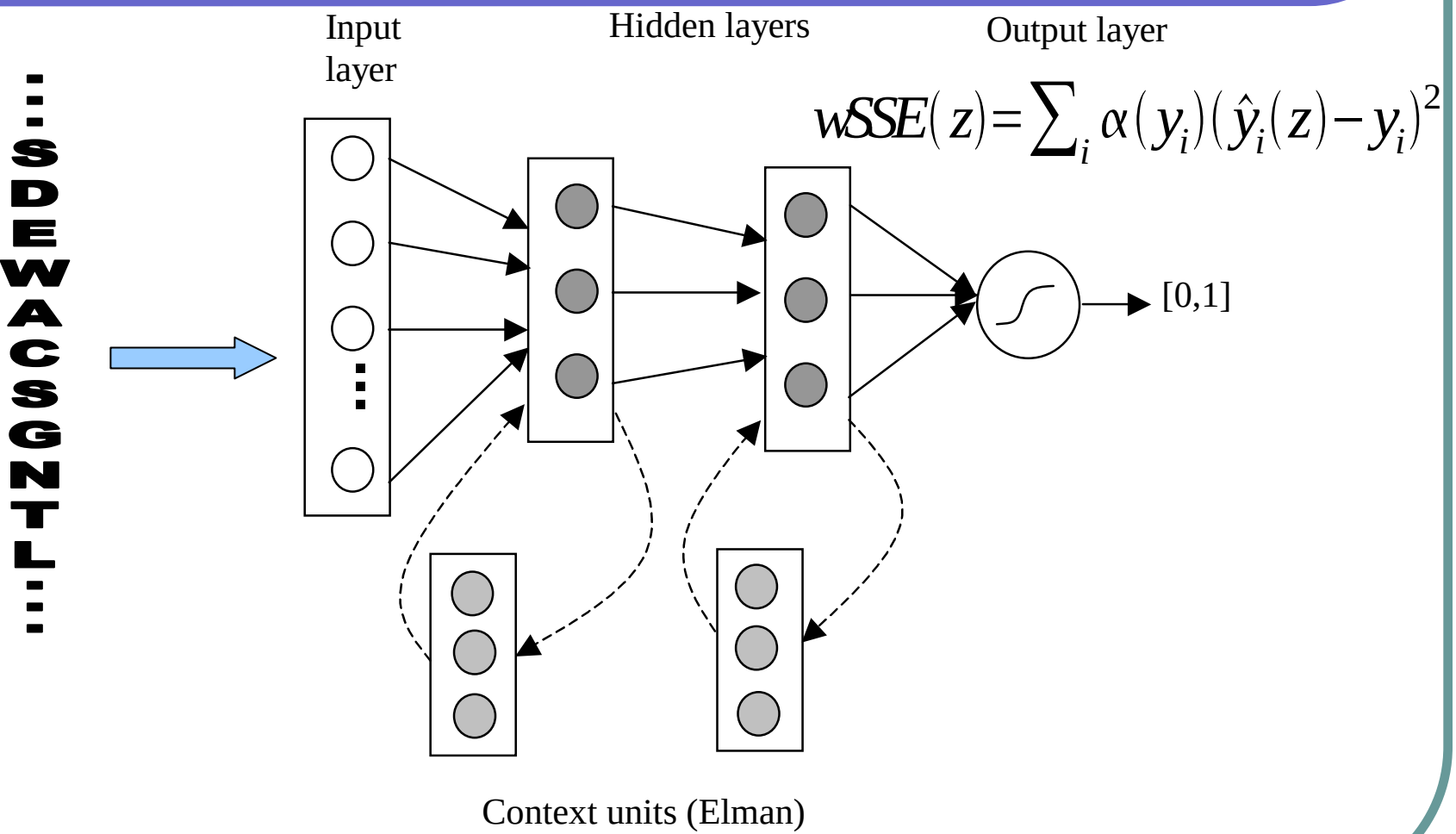


PDB codes: 2a24, 1unc

Variability in Surface Exposure for Structurally Equivalent Residues: “Soft” Approximation Problem



Neural Network-based Regression for RSA Prediction



R. Adamczak, A. Porollo, and J. Meller, Proteins 56 (2004)

Support Vector Regression for RSA Prediction

ε- insensitive SVR regression model:

$$min \left\| w \right\|_p + C \left\| \xi \right\|_1$$

$$s.t. \left| a_i^T \, w + \beta - y_i \right| - \xi_i \leq \varepsilon \qquad \text{for each } i$$

Here, $\left\| w \right\|_p \equiv \left(\sum_i \left| w_i \right|^p \right)^{\frac{1}{p}}$ and a_i is the vector that represents residue i.

Make the error bars dependent on the observed RSA, y_i :

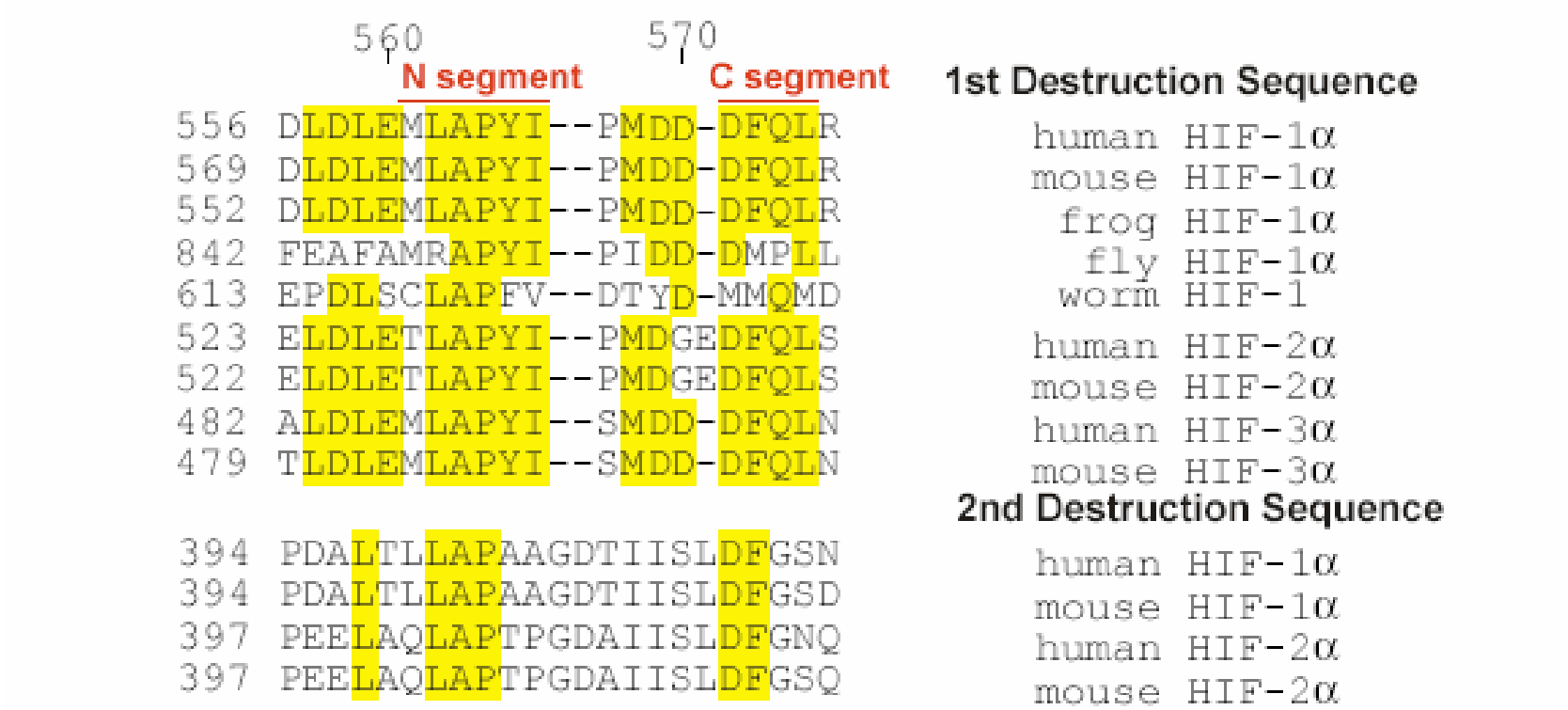
$$\varepsilon_i \equiv \varepsilon \left(y_i \right)$$

M. Wagner, R. Adamczak, A. Porollo and J. Meller; Journal of Computational Biology, Vol. 12 (2005)

Training Sets and Protocols

- To build training set we used 860 protein families derived from the PFAM database
- Input (sliding) window of length 11
- Multiple alignment (PSSM columns) plus additional features used to represent each residue
- Training set consisting of almost 200,000 vectors
- All features were standardized (mean=0, standard deviation=1)
- All networks/SVRs have been trained on different subsets of 90% randomly chosen vectors
- Control sets derived from new submissions to PDB

Multiple alignment and Psi-BLAST



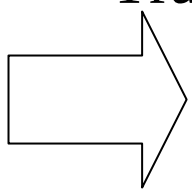
Iteratively redefining similarity measure (scoring matrix): PSSMs

Multiple Alignment and PSSM-based Representation

QUERY: **VDVRKVDISEISSALHVDVPFYVSATALCKLGNPLE**

Class: **BBBBEBEBEEEEEEEBBBEBEBBBEEEEBBBEBEBE**

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
V	0	-3	-3	-3	-1	-2	-3	-3	-3	3	1	-3	1	-1	-3	-2	0	-3	-1	4
D	-2	-2	1	6	-4	0	1	-2	-1	-3	-4	-1	-3	-4	-2	0	-1	-5	-3	-3
I	-2	-3	-4	-3	-1	-3	-4	-4	-4	5	1	-3	1	0	-3	-3	-1	-3	-1	3
S	1	-1	1	0	-1	0	0	0	-1	-3	-3	0	-2	-3	-1	4	1	-3	-2	-2
E	-4	2	0	-5	-3	1	2	-2	-1	-6	-2	1	-1	-3	-1	0	-1	2	-4	-2
I	2	-3	1	0	-1	1	1	0	-1	4	-3	0	-3	-2	-3	-1	1	-3	-2	-2
S	1	-2	1	1	-3	-2	0	1	1	-2	-3	0	-2	-3	-1	4	1	-3	-1	5
S	3	4	-2	2	-5	1	3	-1	-1	-3	4	3	2	1	–2	4	1	1	-2	-2
A	4	-1	-2	-1	1	3	4	2	5	3	-3	-4	-2	-3	-1	-1	-2	-2	0	0
L	1	-1	0	1	0	-1	3	0	1	-1	2	0	2	-5	0	1	1	-3	-2	-2



Multiple alignments and PSSMs obtained using the PsiBLAST program by S. Altschul et. al.: 3 iterations without pre-filtering (following in the footsteps of Rost, Jones and others).

Overall Accuracy of Different Regression Models for RSA Prediction on Independent Control Sets

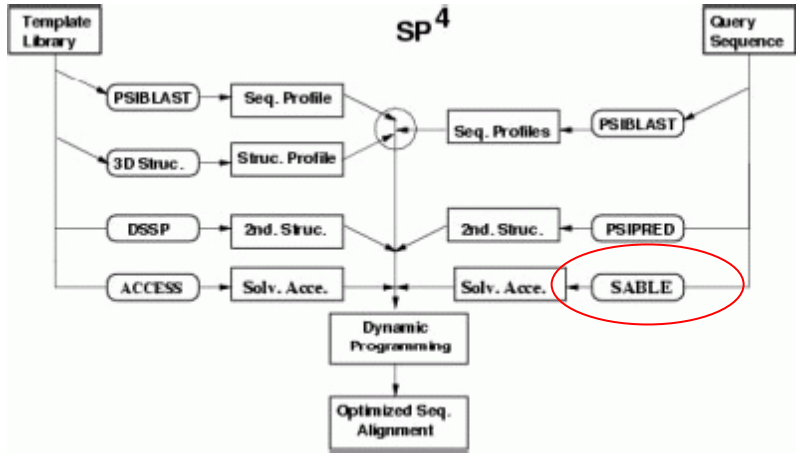
	S163 *cc / MAE / RMSE*	*S156* *cc / MAE / RMSE*	*S135* *cc / MAE / RMSE*	*S149* *cc / MAE / RMSE*
SABLE-a	**0.65 / 15.6 / 20.8**	**0.64 / 15.9 / 21.0**	**0.66 / 15.3 / 20.5**	**0.64 / 16.0 / 21.0**
SABLE-wa	**0.66 / 15.5 / 21.2**	**0.64 / 15.7 / 21.3**	**0.67 / 15.3 / 20.9**	**0.65 / 15.8 / 21.4**
LS	**0.63 / 16.3 / 21.0**	**0.62 / 16.5 / 21.1**	**0.65 / 15.9 / 20.5**	**0.62 / 16.5 / 21.2**
SVR1	**0.62 / 15.9 / 21.3**	**0.61 / 16.1 / 21.4**	**0.64 / 15.6 / 20.8**	**0.62 / 16.2 / 21.5**
SVR2	**0.62 / 16.6 / 22.8**	**0.61 / 16.7 / 22.7**	**0.64 / 16.4 / 22.5**	**0.61 / 16.9 / 23.0**

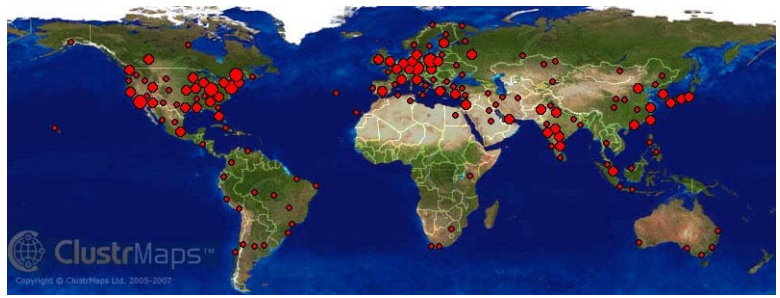
A total of 163+156+135+149=603 non-redundant chains without homology to our training set of 860 representative protein chains derived from PDB, Adamczak et al., Proteins

SABLE is a state-of-the-art RSA predictor

	MCC	Q2
Adamczak et al., Proteins 2004	0.52-0.54	76.5-77.3%
Chen & Zhou, Proteins 2005	0.54	77.2%
Garg et al., Proteins 2005	0.56	78.3%
Liu et al., Proteins 2007	0.53-0.55	74.3-77.9%

47 CASP6 proteins; Garg, Kaur and Raghava, Proteins 61 (2005)
16 FR/NF CASP6 proteins; Chen and Zhou, Nucl. Acids Res. 33 (2005)



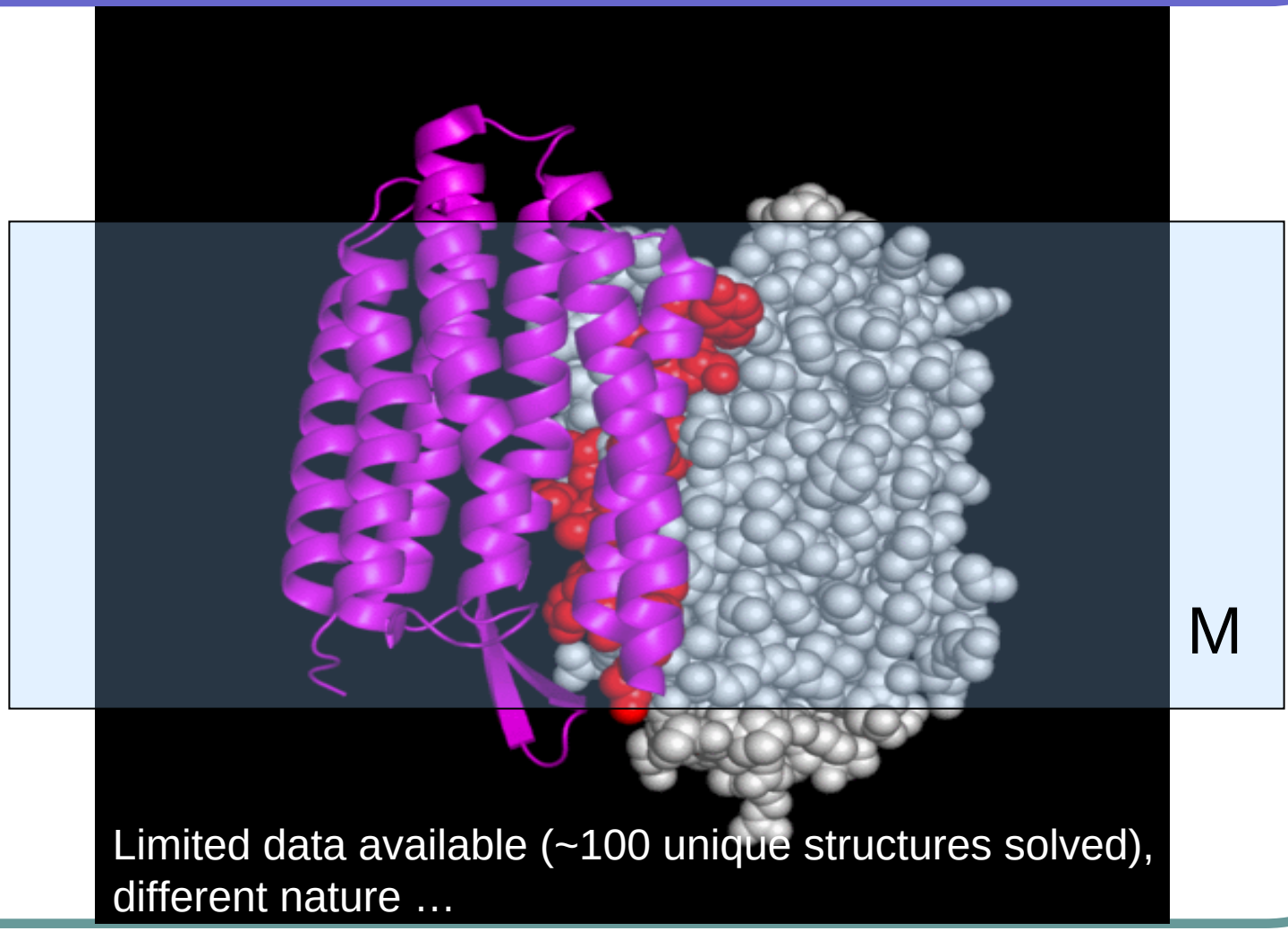


"The two-state accuracy by **SABLE** is 77.3% in the ProSup benchmark, 77.9% in the SALIGN benchmark, 74.3% in the Lindahl benchmark and, 75.3% in the LiveBench 8 benchmark. This accuracy is **consistent with the published performance** of this and other state-of-the-art predictors." Liu, Zhang, Liang and Zhou, Proteins 68 (2007)

Some take home messages …

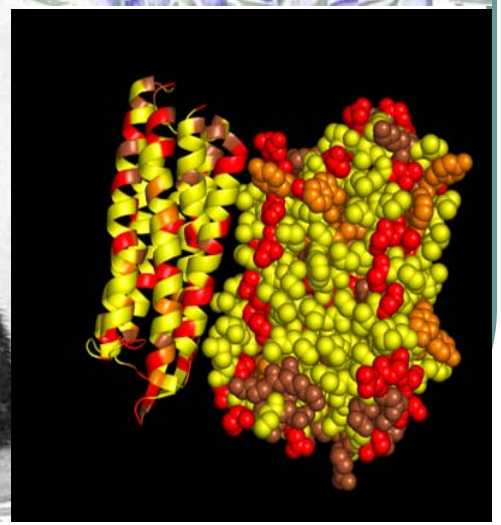
- Importance of domain knowledge and finding the right model for the problem (representation, learning approach etc.)
- In particular, RSA as “fuzzy” (variable) quantity implies specific error models and tailored regression approaches (understanding limits of what can be achieved)
- Importance of non-redundant and representative training and validation sets (identify and control potential biases)
- Importance of using different accuracy measures, including those that are important from the point of view of future applications (some of them somewhat qualitative)
- Importance of confidence measures: an additional meta-classifier trained to provide error estimates
- Cross-validation useful, but final validation on independent control sets necessary to obtain more realistic accuracy estimates (however painful)
- Room for meta-learning, although human brain hard to replace at this point …

Structural and Functional Predictions for Membrane Proteins



Limited data available (~100 unique structures solved), different nature …

Cats vs. Membrane Proteins

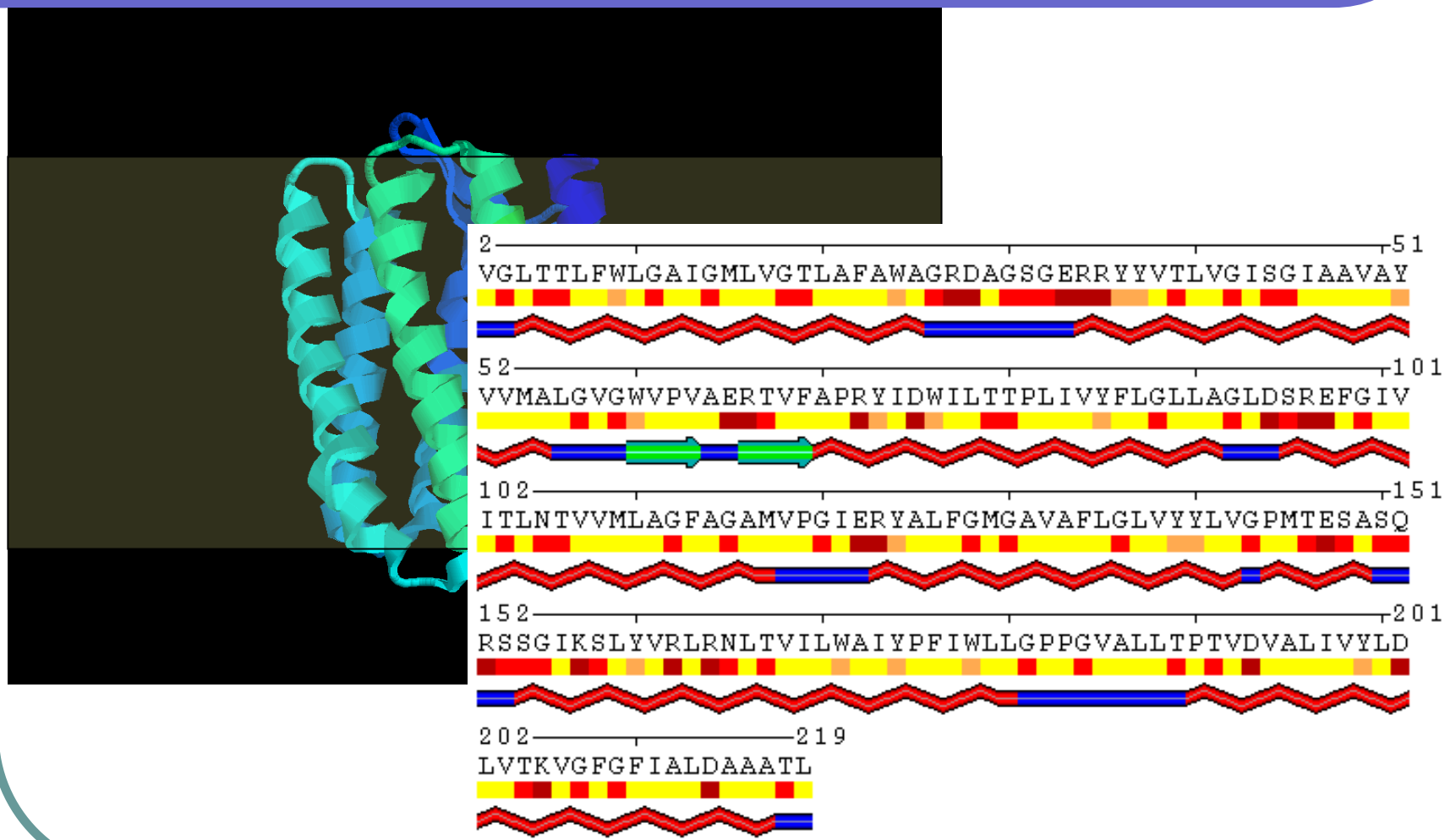


A bit of irony …

"While it's true that most cats find the bathing experience less than savory, professional cat breeders acclimate their pets to the process through **regular repetition**."

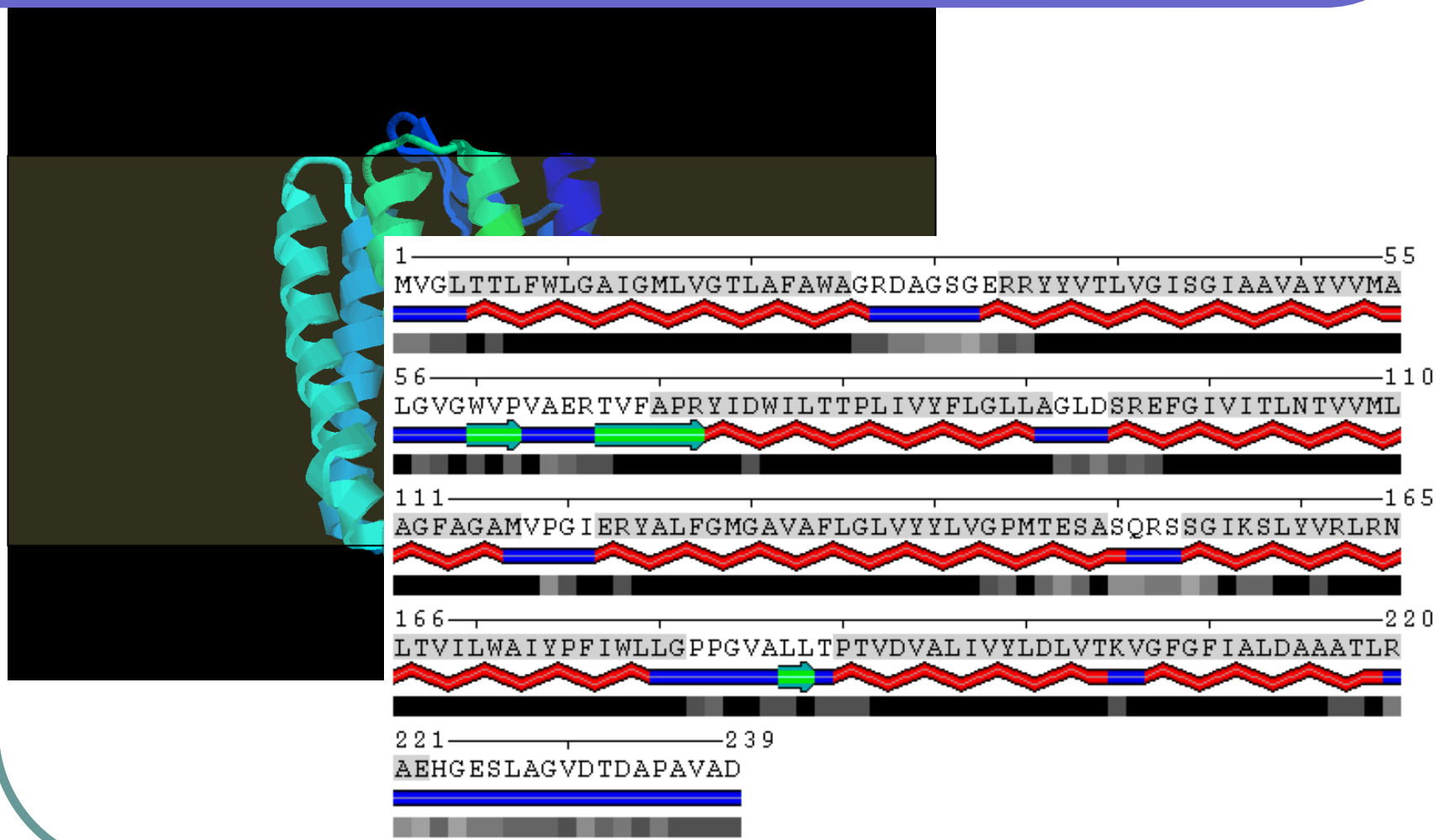
Courtesy of the **Society for the Prevention of Cruelty to Animals** (www.spca.com), as well as www.geckoandfly.com and http://courses.umass.edu/phys120/images/cat-and-mouse.jpg

Learning from Limited Data: Recognition of Transmembrane Domains

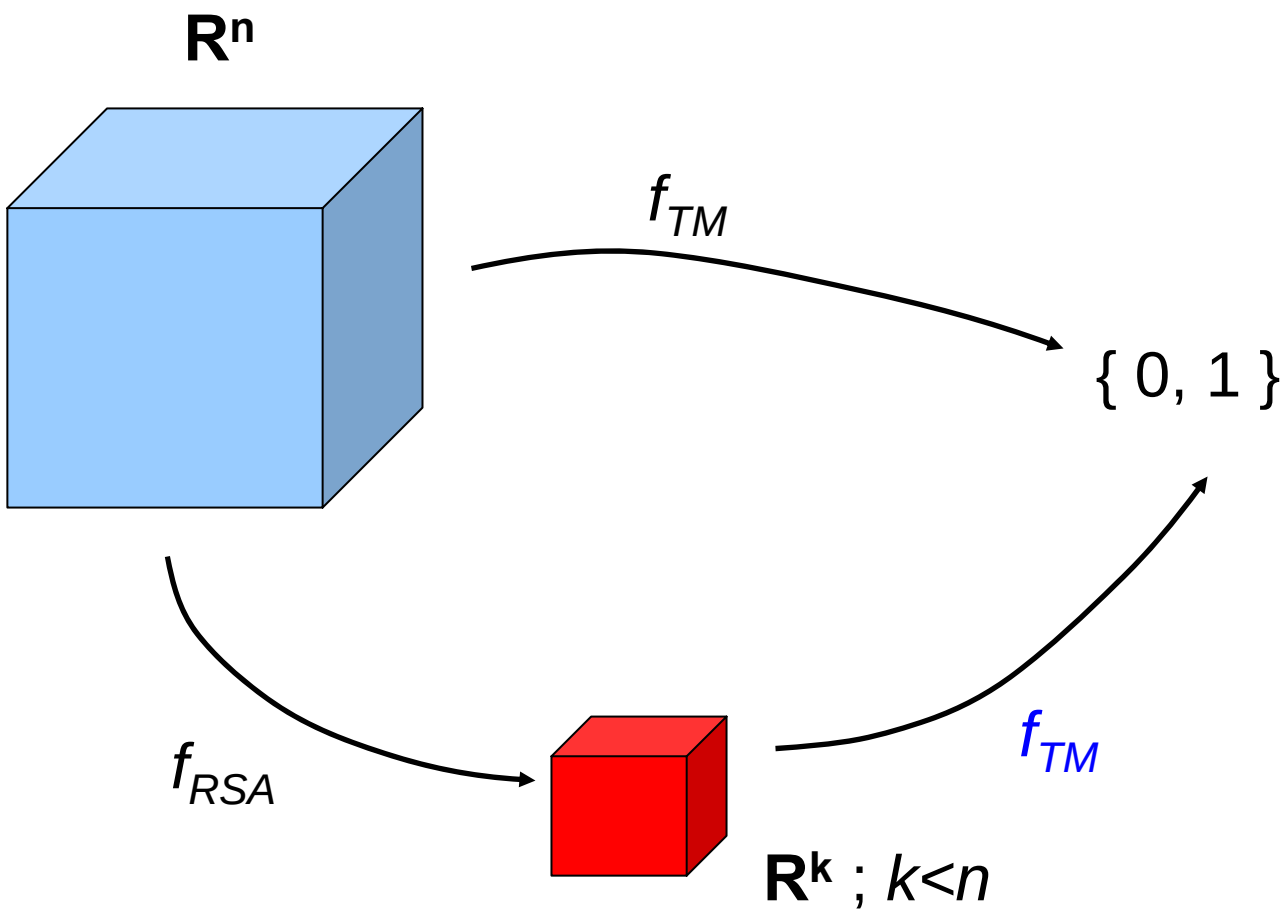


Recognition of TM Domains with Predicted RSA:

Compact Representation of an Amino Acid and Its Environment without Explicit Use of Multiple Alignments



Transformation of Evolutionary Profiles (MAs) into a Compact Representation by Using the RSA Prediction



Recognition of TM Domains Using RSA Prediction

RSA predictions are used in order to indicate residues unlikely to be exposed to aqueous environment, i.e., residues that are either buried in the hydrophobic core of a protein, or alternatively, “buried” in a membrane …

	Alpha-helical		Beta-barrel	
Features	Q2 %	MCC	Q2 %	MCC
RSA+SS (11)	87.9±0.8	0.74±0.02	77.9±3.3	0.50±0.09
RSA+SS (21)	**88.0±0.6**	**0.73±0.02**	**78.7±3.3**	**0.53±0.08**
RSA+SS (31)	87.4±0.7	0.73±0.020	77.9±3.6	0.53±0.08
MSA (11)	85.0±1.3	0.67±0.03	71.6±2.9	0.37±0.07
MSA (21)	**86.0±1.4**	**0.69±0.03**	**73.3±3.4**	**0.41±0.08**
MSA (31)	86.5±1.4	0.70±0.03	73.6±3.6	0.42±0.09

Cross-validated classification accuracy for transmembrane helices prediction with different feature spaces and a set of 72 TM chains.

Results in the TMH Benchmark Server (Chen and Rost, 2003): Reassessing the Overall Expectations

Method	Q_2	Q_{OK}	cGP	cSP
MINNOU	**89**	**80**	**1**	**8**
PHDhtm	80	84	2	23
HMMTOP2	80	83	6	48
TMHMM1	80	71	1	34
DAS	72	79	16	97
TopPred2	77	75	10	82
SOSUI	75	71	1	61

B. Cao, A. Porollo, R. Adamczak, M. Jarrell and J. Meller; *Enhanced Recognition of Protein Transmembrane Domains with Prediction-based Structural Profiles*, **Bioinformatics**, vol. 22 (3): 303-309 (2006)

Some take home messages …

- Early and optimistic estimates of accuracy of TMH prediction methods suggested that this problem had essentially been solved (claims of over 95% accuracy etc.)
- These estimates, however, were based on cross-validation studies using small and biased samples of TM proteins
- Recent reassessment from Rost group and others – still ways to go
- Importance of compact representations, low complexity models – risk of overfitting and overestimating the accuracy still significant
- New TM proteins being resolved structurally (e.g. ion channels) reveal novel, more complex architectures/folds
- New RSA-based representation provides a unique transformation of multiple alignment input data, using a predictor trained exclusively on soluble proteins, and thus minimizing the risk of biasing and overfitting

Support Vector Regression for RLA prediction

ε- insensitive SVR regression model:

$$min \left\| w \right\|_p + C \left\| \xi \right\|_1$$

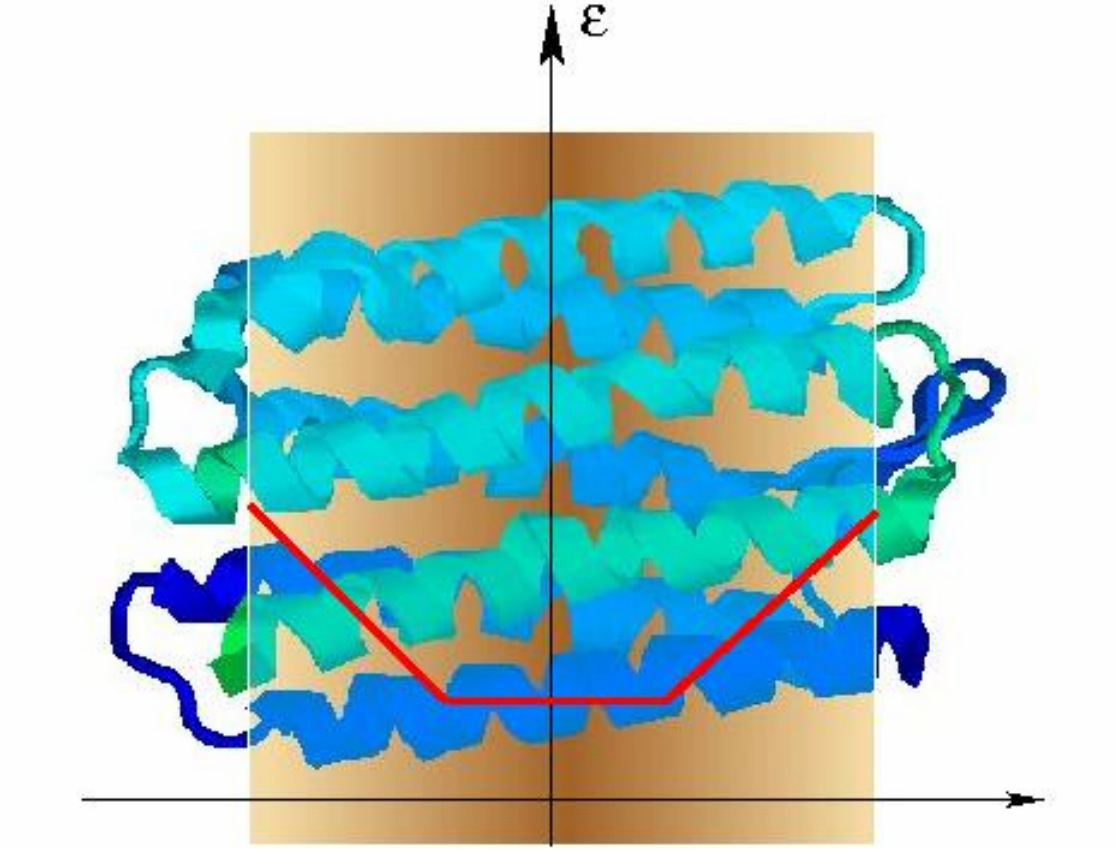
$$s.t. \left| a_i^T \, w + \beta - y_i \right| - \xi_i \leq \varepsilon \qquad \text{for each } i$$

Here, $\left\| w \right\|_p \equiv \left(\sum_i \left| w_i \right|^p \right)^{\frac{1}{p}}$ and a_i is the vector that represents residue *i*.

Make the error bars dependent on the observed RSA, y_i :

$$\varepsilon_i \equiv \varepsilon \left(y_i \right)$$

Using Flexible SVRs for Lipid Accessibility Prediction



RLA Prediction: Need for Low Complexity Models

Representation	NN	SVR
RSA	0.34±0.02	0.36±0.04
MSA	0.32±0.02	0.45±0.02
MSA+WW	0.33±0.02	0.45±0.02
MSA+TMLIP2H	0.32±0.02	0.46±0.02
MSA+SABLE	0.35±0.02	0.47±0.02
MSA+SABLE+WW	0.33±0.03	0.47±0.02
MSA+SABLE+TMLIP2H	0.36±0.02	0.47±0.02

Cross-validated accuracies in terms of correlation coefficients on a non-redundant set of 72 alpha-helical TM proteins (about 7 thousand TM residues).

Performance of our new RLA predictor on an independent control set: robust predictions with good generalization ☺

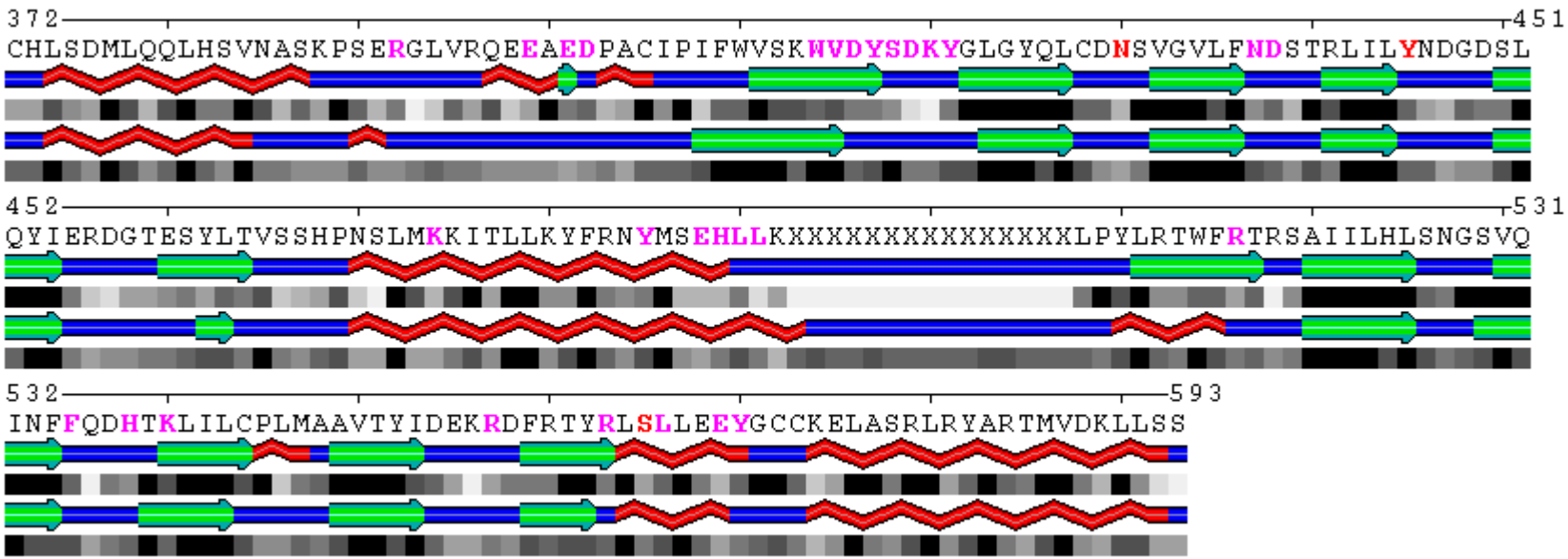
PDB Chain ID	CC	RMSE [%]	MAE [%]
1xfh_C	0.50	21.5	16.4
1vry_A	0.35	34.4	31.1
1xqf_A	0.62	15.6	12.7
1yq3_D	0.57	17.5	13.8
1s5l_z	0.53	18.5	14.7
1s5l_x	0.40	14.4	10.9
1w5c_F	0.56	23.7	20.6
2axt_h	0.56	19.7	17.6
1yew_K	0.59	16.7	14.4
1yew_J	0.33	25.6	22.1
1yew_A	0.80	19.2	17.1
1q90_M	0.60	20.8	16.3
2bbj_E	0.51	13.9	12.3
1zcd_A	0.64	16.3	13.6
1c17_M	0.40	24.2	18.8
2a65_A	0.51	18.9	16.0
Average	***0.53±0.03***	***19.9±1.3***	***16.6±1.2***

Summary

- Improved, regression-based real-valued RSA prediction (correlation coefficients between observed and predicted RSA ~ 0.67)
- Enhanced trans-membrane domain prediction with a compact representation obtained using RSA predictions
- RLA prediction using compact SVR models: good generalization, correlation coefficients between observed and predicted RSA ~ 0.5
- Applications in both *de novo* simulations and fold recognition (filtering out incorrect models)
- Enhanced recognition of protein-protein interaction sites based on the difference between predicted and experimentally observed RSA (RLA)
- Other applications of RSA/RLA predictions: post-translational modifications sites, recognition of pore interfaces in ion channels, recognition of binding sites for ligands, analysis of functional consequences of point mutations etc.
- The other story (to be covered some other time) on genome-wide association studies: correlating genotypes and phenotypes using machine and statistical learning, and dealing with even bigger problems: millions of variables (genetic markers) with limited number of data points (patients/genotypes) and fuzzy phenotypes

Biases in RSA Predictions for Residues within Interaction Interfaces

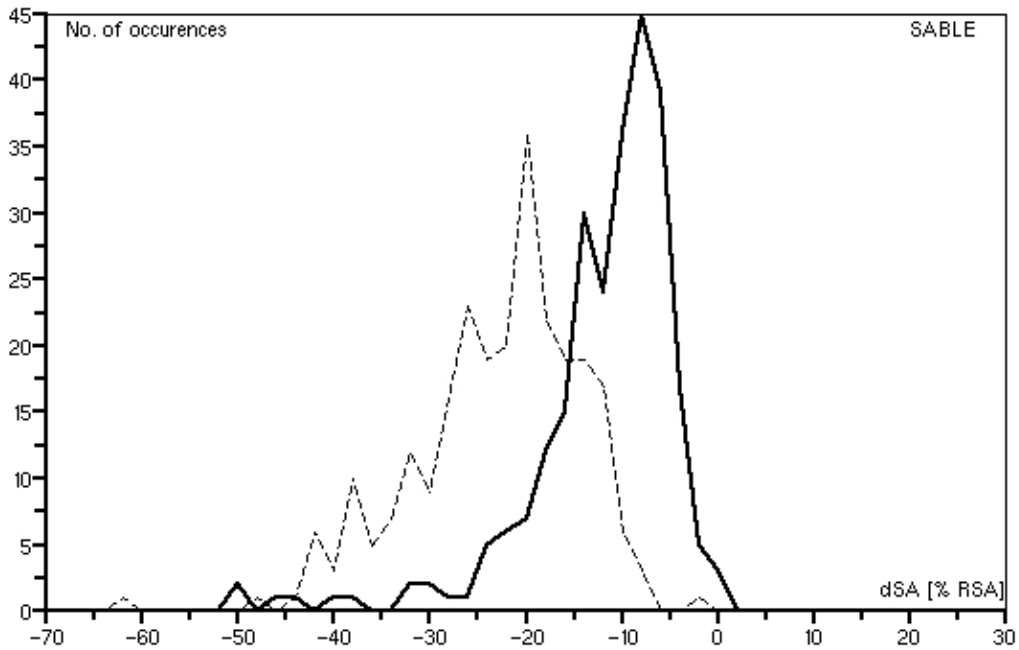
Prediction “errors” at interaction interfaces: differences between predicted and actual (observed in an unbound structure) RSA values.



Predictions obtained using SABLE; picture generated using the POLYVIEW server (A. Porollo) – also used to generate most animations and other pictures used in this presentation.

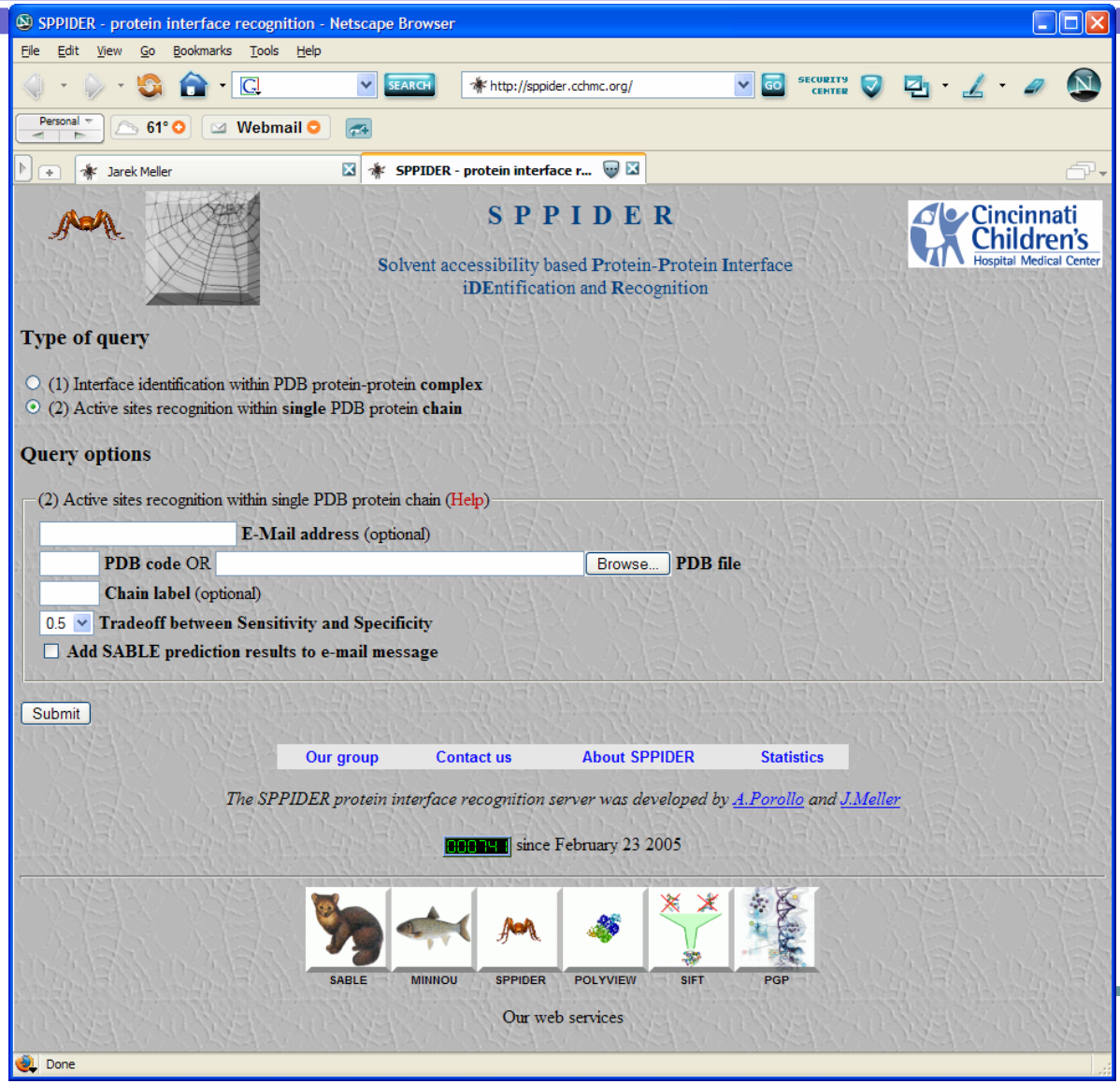
A Novel Fingerprint of Interaction Sites Obtained by Contrasting Predicted and Observed Solvent Accessibilities

Distributions of prediction “errors” (dSA) for interacting vs. non-interacting sites



A. Porollo and J. Meller; *Prediction-based Fingerprints of Protein Interactions*, **Proteins: Structure, Function and Bioinformatics**, 66 (2007)

You are welcome to visit our zoo:



Our servers:

http://sable.cchmc.org

http://sppider.cchmc.org

http://minnou.cchmc.org

http://sift.cchmc.org

Visualization:

http://polyview.chmcc.org