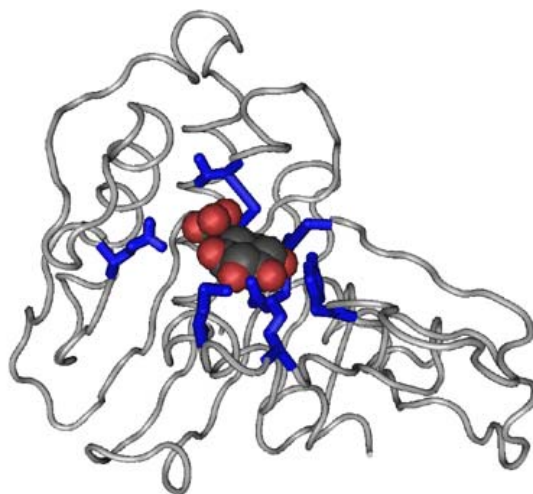


# *Predicting Protein Domains and Small Molecule Binding Sites*



**Michel Dumontier\***, Howard Feldman% and Christopher Hogue\$

*\* Departments of Biology, Biochemistry and Computer Science,  
Carleton University, Ottawa*

*% Chemical Computing Group, Montreal, Quebec, Canada*

*\$ Department of Biochemistry, University of Toronto*

June 29, 2006

# Outline

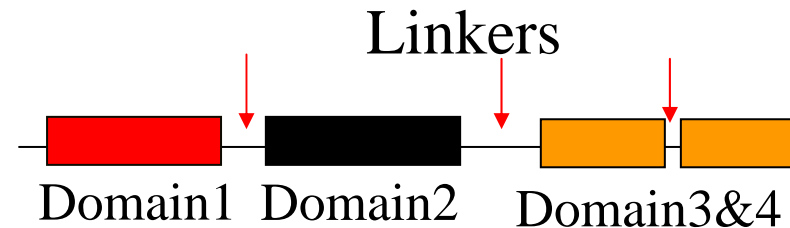
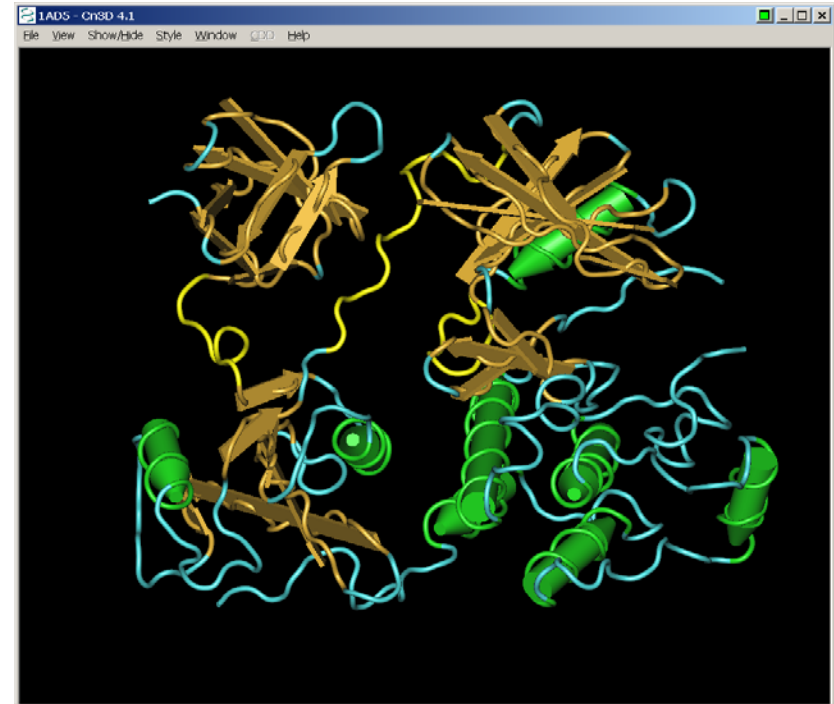
- Predicting Domain Boundaries by Sequence Alone
- Predicting Protein Small Molecule Interactions using Conserved Domains

# Domains and Linkers

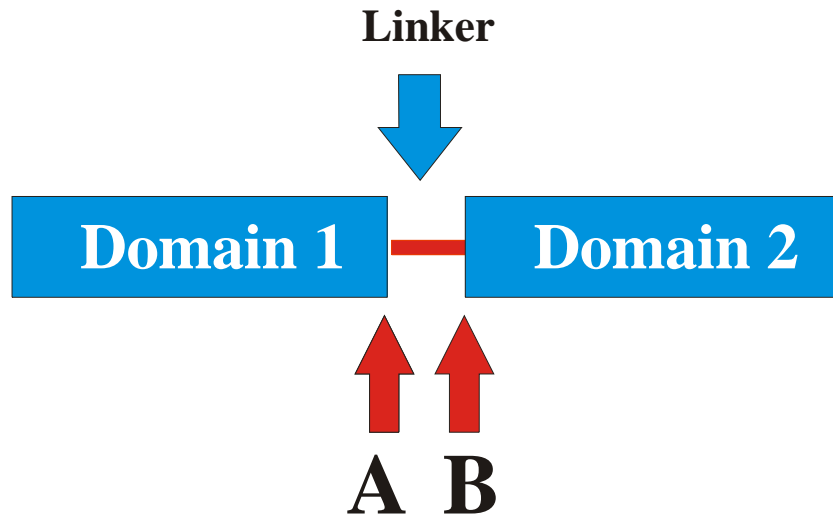
Choice and position of amino acids are important factors in protein folding.

So,

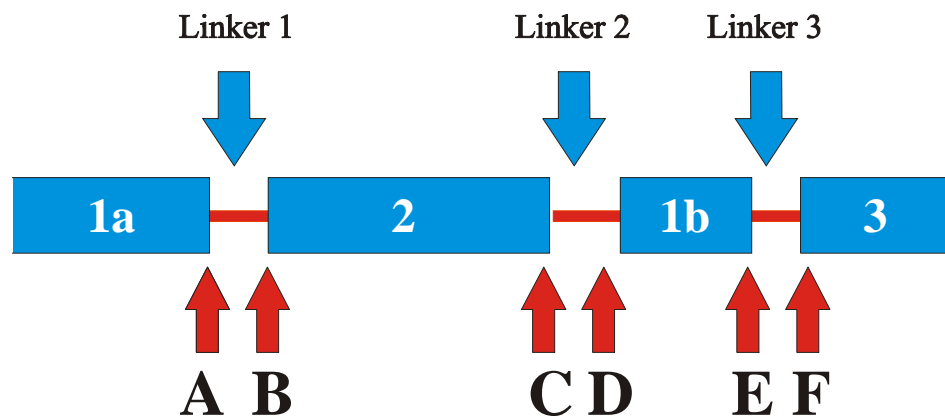
- What rules exist for the choice amino acid in domains and the intervening sequences that link them.
- Can this information be used to predict linkers?



# Delineating Domains & Linkers



**Simple Case:**  
Contiguous Domains with  
well defined linker



**Complicated Case:**  
Segmented Domains with  
multiple linkers

**JMB**

Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

SCIENCE @ DIRECT®



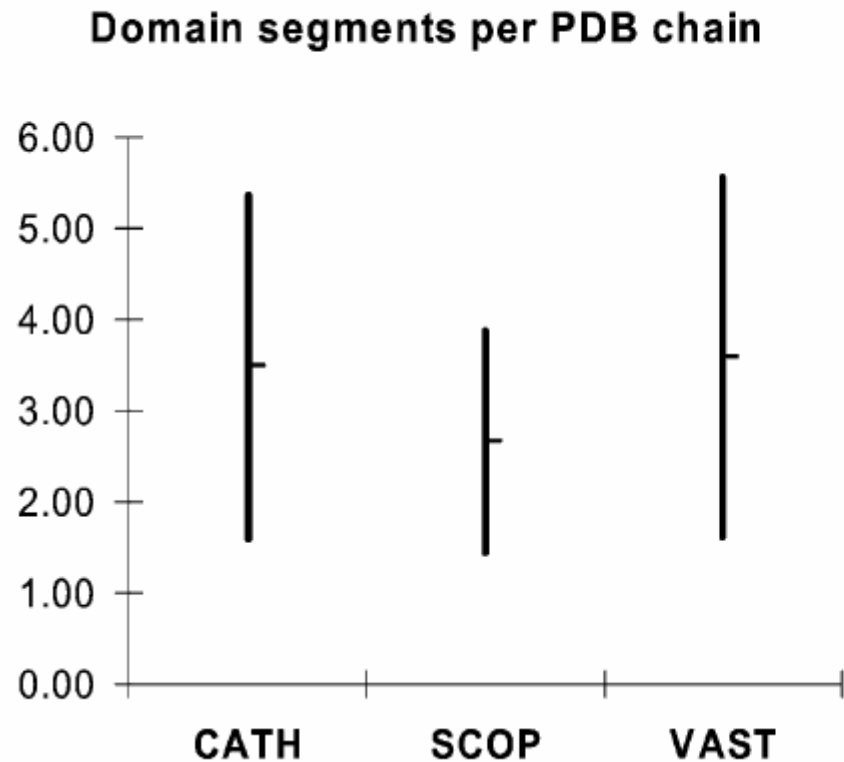
## **Armadillo: Domain Boundary Prediction by Amino Acid Composition**

Michel Dumontier<sup>1,2</sup>, Rong Yao<sup>2</sup>, Howard J. Feldman<sup>2</sup> and Christopher W. V. Hogue<sup>1,2\*</sup>

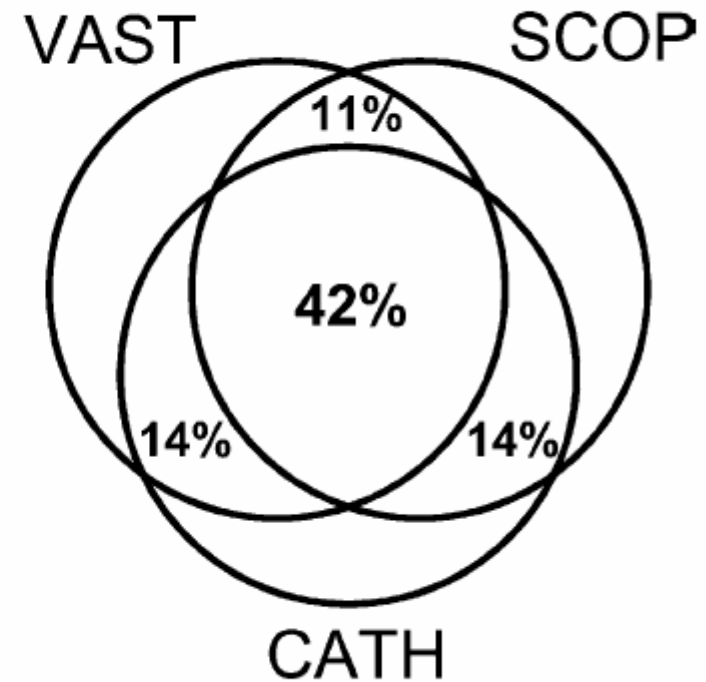
- **GOAL:**
  - Determine the amino acid composition bias in domain linker regions
  - Use this empirical knowledge to make a sequence based prediction (no MSA).

# Comparing Domain Definitions

(a)

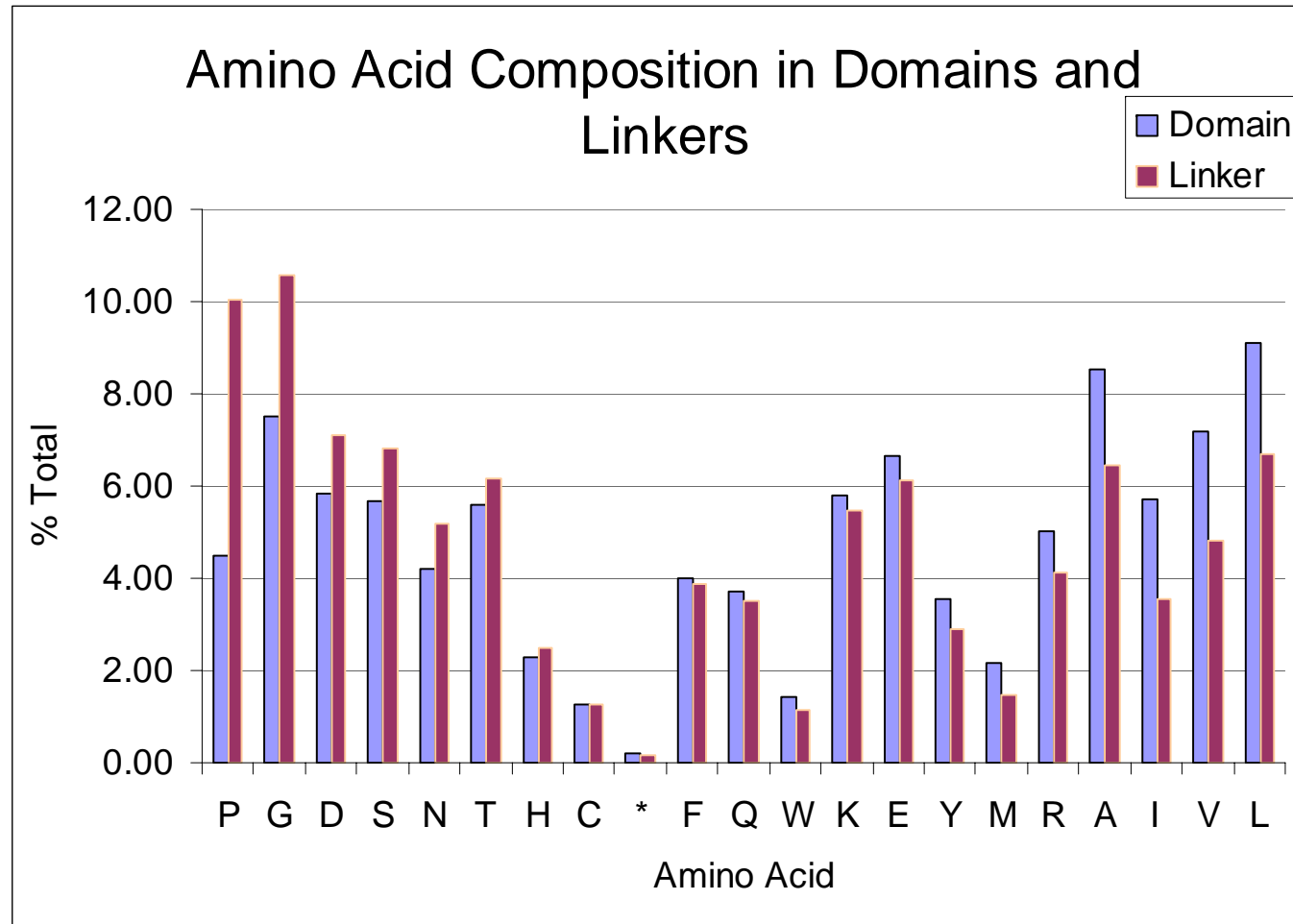


(b)



- Non-redundant set of 655 segmented, multi-domain proteins.

# Hydrophobic residues decreased in domain linkers



# Domain Linker Index

$$DLI_{aa} = -\log \left( \frac{\sum_{j=0}^{nprot} \frac{n_{i,l}}{n_{i,l,t}}}{n_{i,d} / n_{i,d,t}} \right)$$

where  $DLI_{aa}$  is the negative log probability of the propensity for amino acid  $i$  in the linker region (l) and in the full protein set (t), where  $n_{i,l}$  and  $n_{i,d}$  are the number of amino acid type  $i$ , respectively.

# Log Likelihood Indices

Table 2. Domain linker propensity indices

AA	DLI	REI	GHL	KDH
A	0.806	-1.721	0.273	0.767
R	0.508	1.637	-1.287	-1.342
N	-0.823	-0.042	0.447	-1.008
D	-0.773	-0.042	0.691	-1.008
C	-0.161	-0.042	1.894	1.001
Q	0.052	0.798	-0.450	-1.008
E	0.127	0.798	-0.485	-1.008
G	-1.264	-0.882	1.397	0.030
H	-0.380	-0.042	-0.163	-0.907
I	1.441	-0.042	0.639	1.671
L	0.893	-0.042	-0.781	1.436
K	0.072	1.637	0.447	-1.142
M	1.205	0.798	-0.319	0.800
F	-0.049	-0.042	-1.078	1.101
P	-2.799	-2.561	-2.646	-0.372
S	-0.738	-0.042	0.421	-0.104
T	-0.465	-0.042	-0.189	-0.070
W	0.560	-0.042	0.874	-0.137
Y	0.590	0.798	-0.041	-0.271
V	1.199	-0.882	0.352	1.570

Amino acid indices used for identifying domain linkers. The domain linker propensity index (DLI) was derived in this study from analysis of the residues in linkers as compared to those in domains. The DLI and other indices, REI,<sup>32</sup> GHL<sup>33</sup> and KDH<sup>39</sup> were normalized to a zero mean, unit standard deviation for use by Armadillo in domain linker predictions.

**REI** – Residue Entropy Index – derived from Galzitskaya & Melnik<sup>1</sup> is based on the number of degrees of freedom for each amino acid residue.

**GHL** – derived from George-Heringa Amino Acid propensity of all linkers in their study<sup>2</sup>

**KDH** – Kyte-Doolittle hydrophathy index - control

<sup>1</sup>Galzitskaya, O. V. & Melnik, B. S. (2003). *Protein Sci.* 12, 696–701.

<sup>2</sup>George, R. A. & Heringa, J. (2002). *Protein Eng.* 15, 871–879.

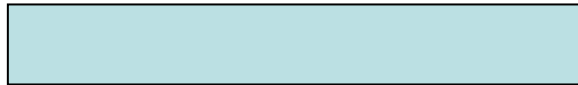
# Applying the index

sequence

Replace AA with index values

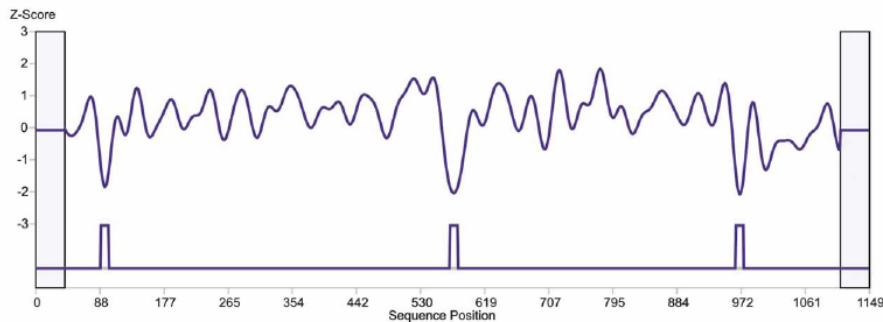
1 2 3 3 1 -1 -2 -3 1 5 3 4 2 1 -1 -2 -5 -4 -5 -2 0 1

Smoothing window (15 aa in length)

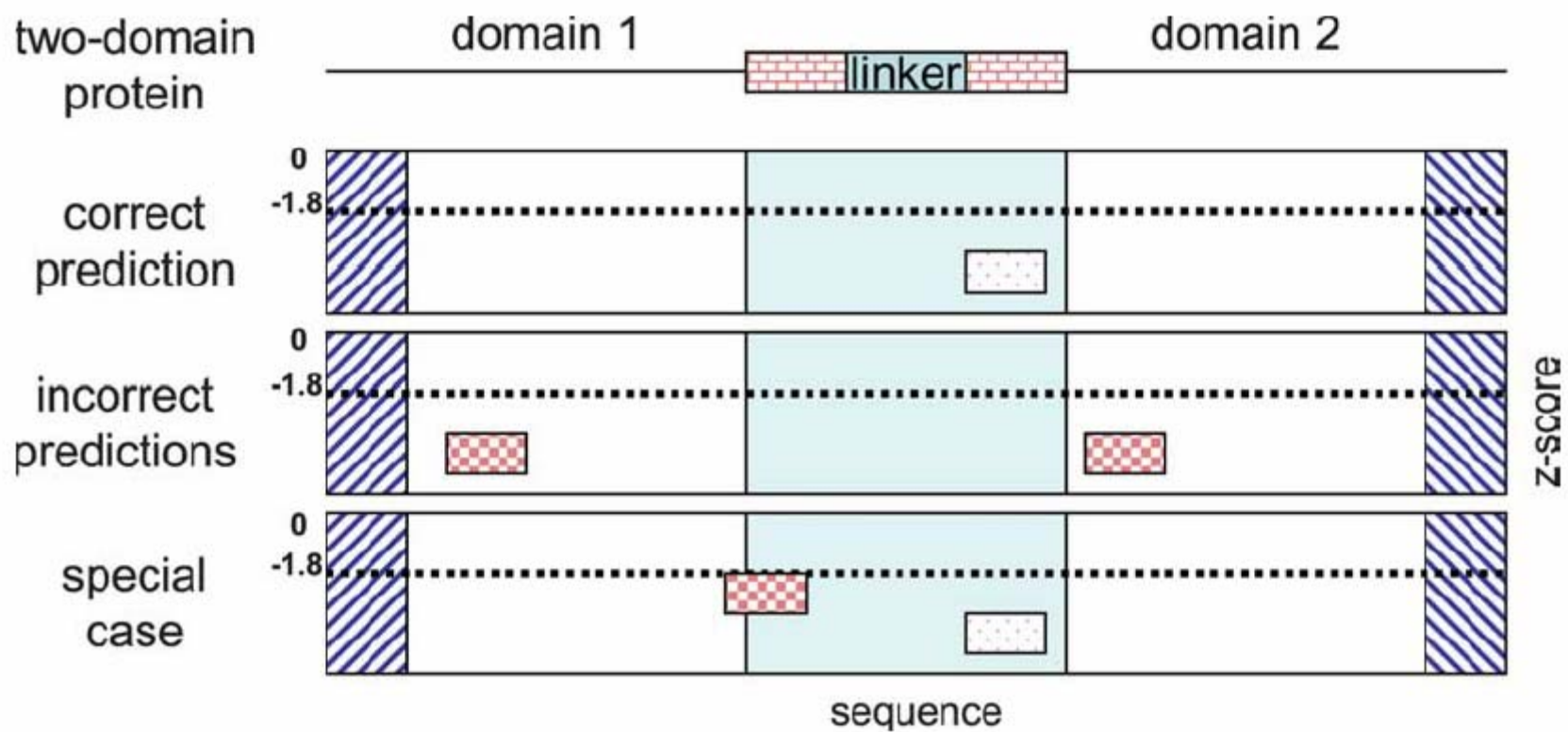


FFT & Low-Pass Filter to further reduce noise and smooth curve

Normalization & Z-score estimation (standardize across different indices)



# Evaluation



# Comparison to other methods

**Domain Guess by Size** (Wheelan 2003) : Based on statistical distribution of domain lengths.

- 28% sensitivity (3x better than random)
- Armadillo 2x more sensitive

**Neural Network** (Miyazaki 2002): Uses linker amino acid propensities

- 74 single linker, multi-domain (continuous) proteins
- NN: 59%:36% (sensitivity:specificity)
- Armadillo : 54%:49%

**DomSSEA** (Marsden 2002) uses secondary structure prediction and alignment

- ~200 two-domain (single linker, continuous)
- DomSSEA: 49%, 53% (consensus)
- Armadillo : 63.9% sensitivity

**SnapDRAGON** (George 2002): multiple sequence alignment & 3D models

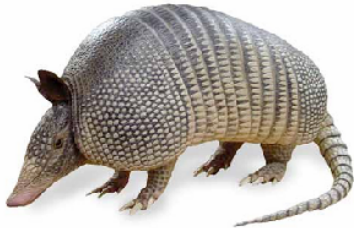
- SnapDRAGON <400 residues requires 1 hour on 100 linux nodes.
- Armadillo : <1 second
- Continuous domains: SnapDRAGON 42%:40% -- Armadillo 44%:33%
- Segmented domains: SnapDRAGON 33%:40% -- Armadillo 34%:44%<sup>12</sup>

Armadillo Domain Prediction - Microsoft Internet Explorer

File Edit View Favorites Tools Help Google Search Web Search Site

Address http://armadillo.mshri.on.ca

## Armadillo: Domain Linker Prediction



Proteins are often composed of multiple structural/functional domains. Domain linkers link these domains together and have been found to contain an amino acid signature that is distinct from the structurally compact domains. Armadillo predicts the linker regions of proteins from their sequence using amino acid indices that reflect the propensity of amino acids in those linker regions.

Important details on using and interpreting Armadillo can be found [here](#).

Armadillo accepts any of the following identifiers:

- NCBI GenInfo Identifier (GI)
- Accession (Accession)
- MMDB Identifier (MMDB)
- PDB Identifier (PDB) - 4 letter code + 1 letter chain (optional)
- FASTA formatted sequence (FASTA)

If you select a structure identifier (MMDB or PDB), then you can compare the prediction to the VAST or SCOP domain definitions.  
 VAST  SCOP

Please choose one of the identifiers in the drop down list, and enter the identifier in the text box. Click [here](#) for a test entry, then press the Predict! button.

OR

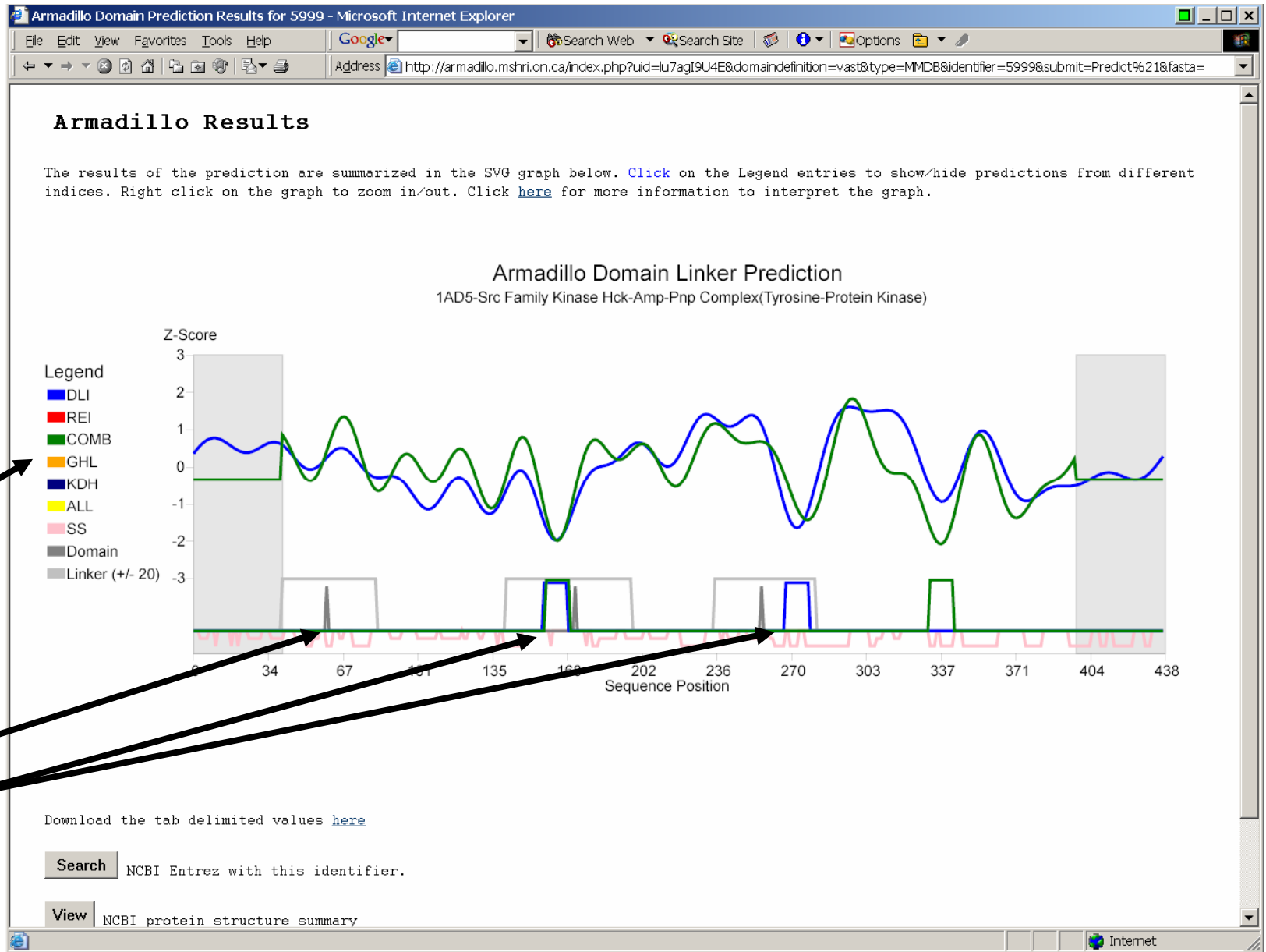
Enter a FASTA formatted sequence.

[armadillo.blueprint.org](http://armadillo.blueprint.org)

HELP!

Example:

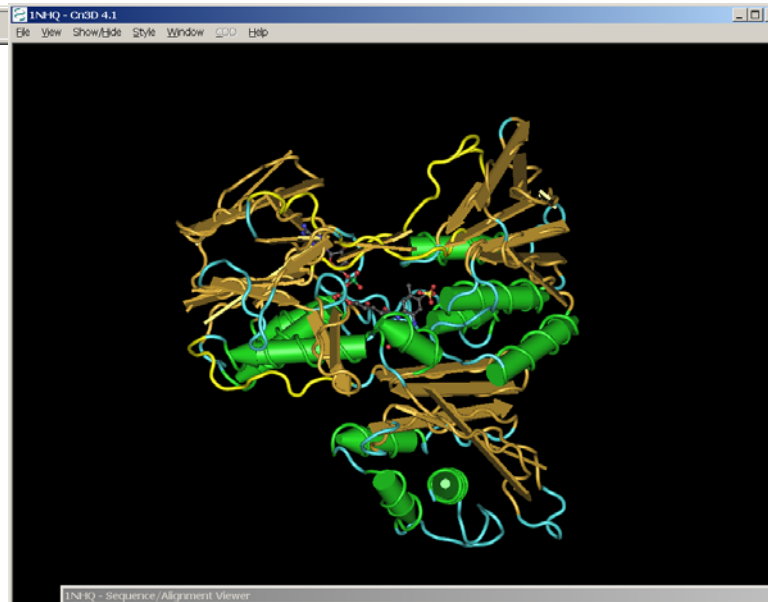
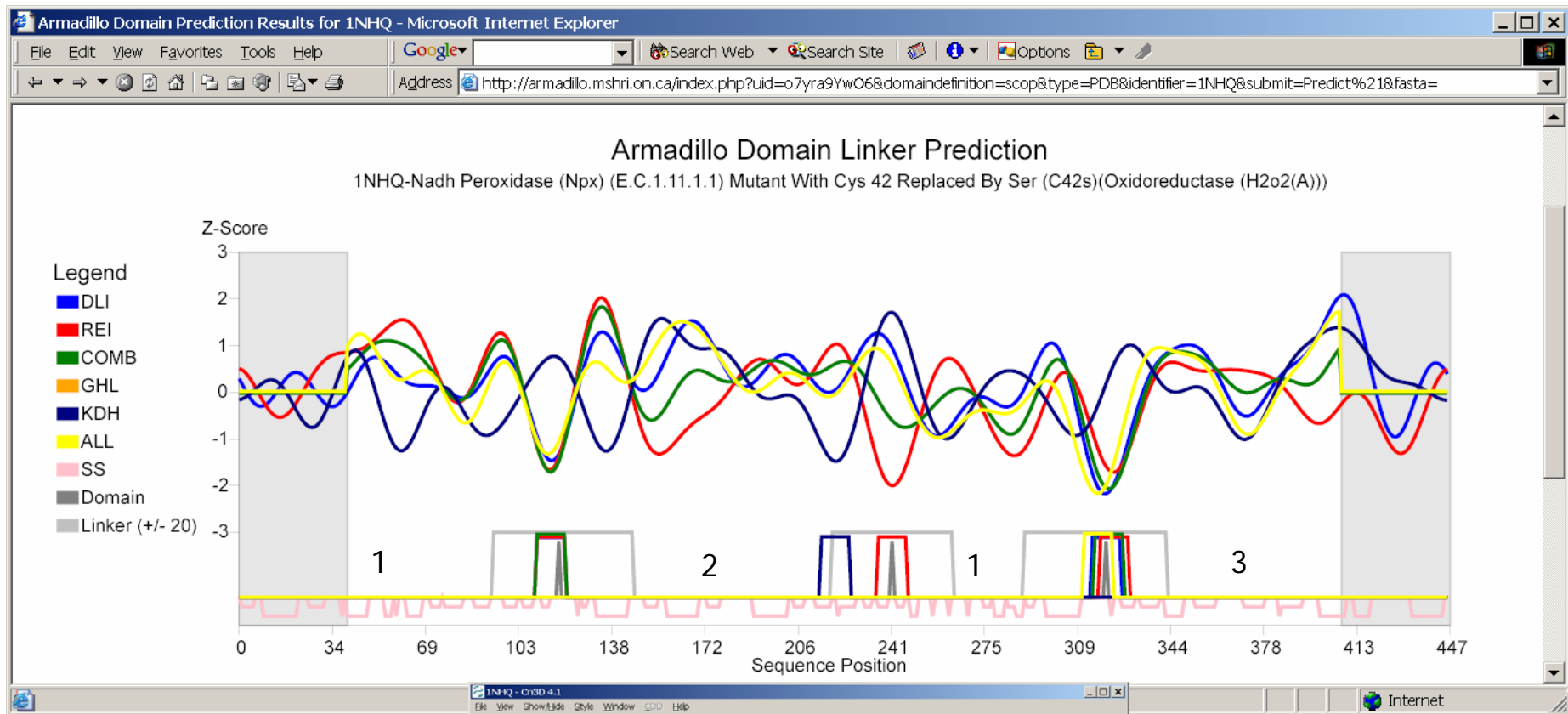
Src Family Kinase



Show/Hide by clicking on legend

Linker

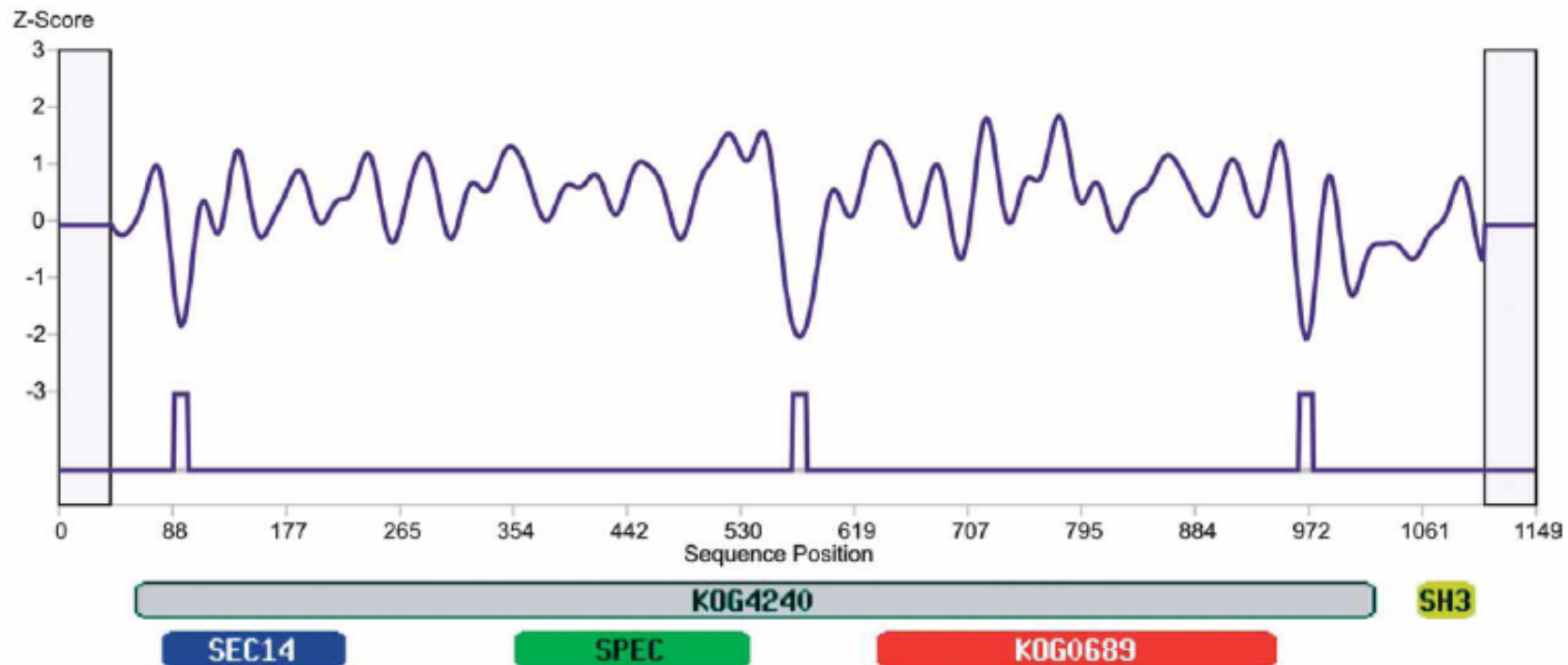
Predictions



# Predicting Linkers for Proteins with Conserved Domains

## Armadillo Domain Linker Prediction

gi|6014925|sp|Q64096|DBS\_MOUSE Guanine nucleotide exchange factor DBS (DBL's big sister) (MCF2 transforming sequence-like protein)

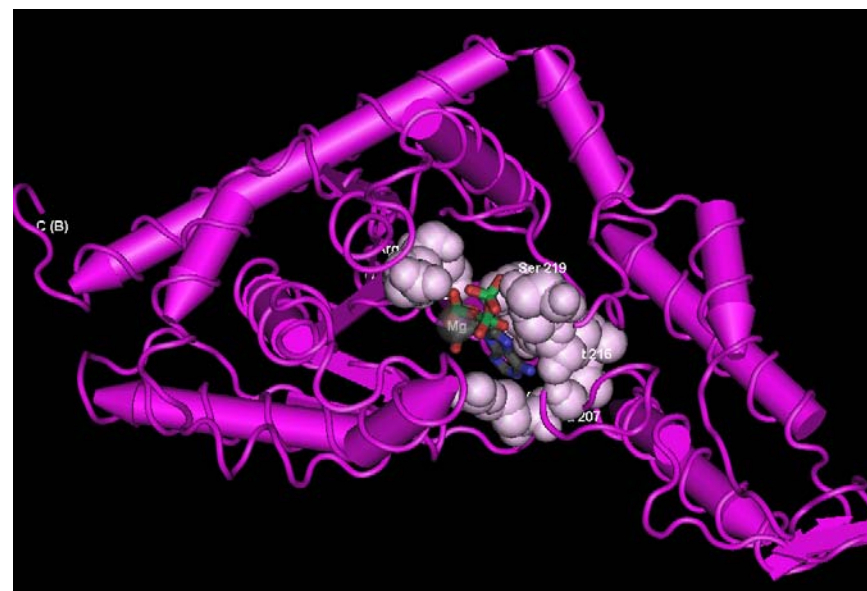


# Outline

- Predicting Domain Boundaries by Sequence Alone
- **Predicting Protein Small Molecule Interactions using Conserved Domains**

# Protein Small-Molecule Interaction Database

- Derived from MMDB (PDB) structure database
- Captured in the 3DSM division of the Biomolecular Interaction Network Database (BIND) – <http://bind.ca>
- Filtered for crystallographic symmetry, buffer agents, non-biologically interesting small molecules
- 23,000+ non-redundant small molecule interactions.



Tryptophanyl-tRNA Synthetase

PDB: 1YID  
BIND Id:330151

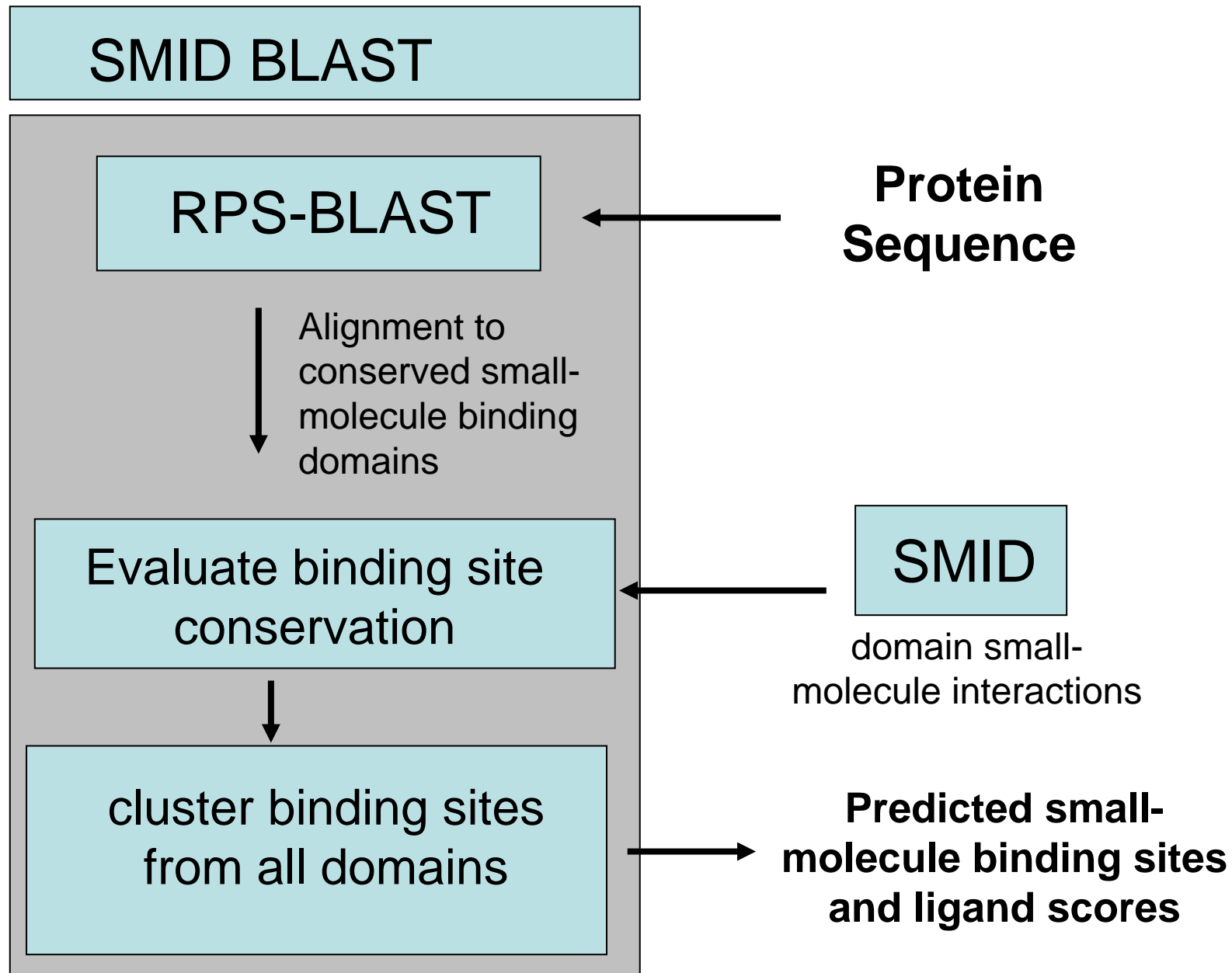
# SMID: A **Domain** Small-Molecule Interaction Database

- Map binding sites from the protein structure to the conserved domain.
- Identify conserved domains using RPS-BLAST
  - Includes SMART, PFAM, CD domain alignments
- ~50,000 domain small-molecule binding sites.



# SMID-BLAST

- Enables users to identify putative small-molecule binding sites in proteins for which a crystal-structure has not yet been determined.
- Requires that protein has a conserved small molecule binding domain.
- Annotates binding sites on query protein
  - Based on PDB structural interactions
- Freely available
  - Web interface
    - <http://smid.blueprint.org>
  - Standalone tool
    - <ftp://ftp.blueprint.org/pub/SMID/tool/>

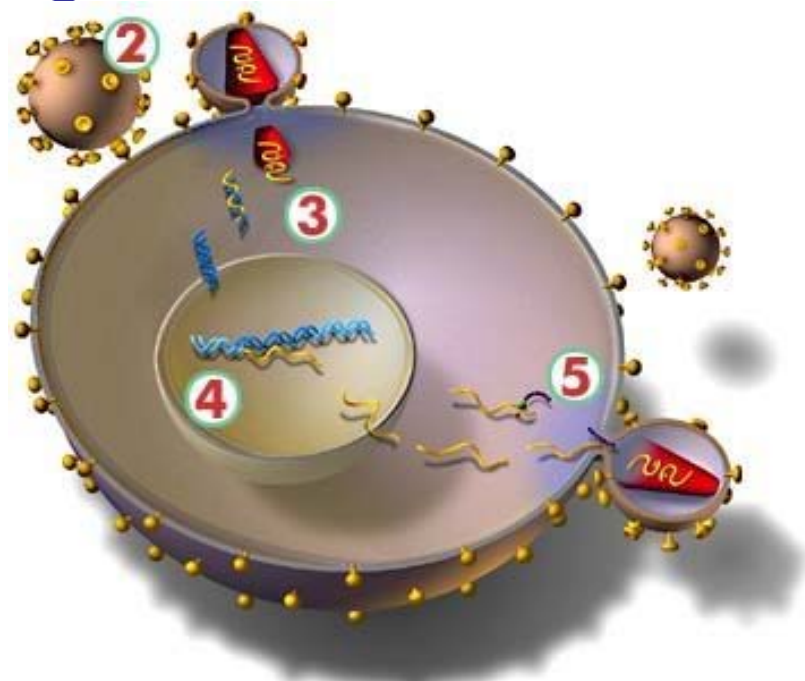


# SMID-BLAST Validation

- Can we **predict** the same **ligand**?
  - 600 PDB chains having 1652 small molecule interactions
  - **62% exact ligand predicted**
    - 25% with best ligand score
- How well do we **predict** the **binding sites**?
  - Over 70% predictions had >80% correct binding residues

# SMID-BLAST Example: HIV Integrase

- Mediates integration of the viral genome into the host DNA.
- Has no mammalian counterpart
- Zn binding domain, a catalytic core and DNA-binding domain.



Use **SMID-BLAST** to make short list of small molecules that may interact with the integrase

- Basis for pharmacological studies to determine inhibition.

# Small molecules predicted to bind to HIV Integrase

- Y3
  - known interactor with Avian Sarcoma Virus integrase (24% sequence identity)
  - Used as a basis for finding integrase inhibitors with *in silico* search and validated with experimental assays<sup>1</sup>

Small Molecule Summary: <a href="#">Help</a>			
Site #	Molecule	Binding Site(s) on query	Final Ligand Score
1	<a href="#">Zn2+</a>	64, 116, 152	911.719
1	<a href="#">100</a>	64, 66, 151..152, 155..156, 159	674.050
1	<a href="#">Mn2+</a>	64, 116	554.275
1	<a href="#">Mg2+</a>	64..65, 116	462.967
2	<a href="#">Zn2+</a>	12, 16, 40, 43	448.878
2	<a href="#">K+</a>	37..40, 43..44	434.232
3	<a href="#">TTA</a>	167..169, 174	438.099
3	<a href="#">TTO</a>	168	109.525
4	<a href="#">Y3</a>	60, 62, 114, 149..150, 153	224.743

<sup>1</sup>Chen et al Bioorg Med Chem. 2000 Oct;8(10):2385-98.

# SMID Genomes

- **Bridges** the gap between structural proteomics and genomics
- Small-molecule binding site predictions for proteins of **1616** completely sequenced genomes
- Allows for **comparative analysis** of small-molecule binding profiles



## SMID Genomes

**SMID Genomes** offers a simple interface to browse, search or compare predicted small molecule interactions in an *organism-specific* or cross-genomes manner. SMID Genomes is built by running SMID-BLAST over the NCBI non-redundant (NR) sequence database and genome information is obtained from the NCBI's RefSeq database.

Use the search box to search organisms, small molecules or domains or browse one of the taxonomic collections.

### ***Compare Genome Small-Molecule Binding Profiles***

Use this feature to compare the overlap of small molecule hits for up to 5 genomes simultaneously. Compare Genomes.

#### **SEARCH**

Protein GI  
Domain Identifier  
Small Molecule Name  
Taxonomy name or identifier



#### **BROWSE**

23	Archae	
234	Bacteria	
19	Eukaryote	
271	Phage	
1069	Virus	

#### **STATISTICS**

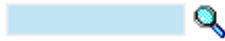
1,616 Genomes  
481,605 Proteins With SM Hits  
46.9% Protein Coverage  
13,010,584 SM Interactions

# Browsing SMID-Genomes

<u>TaxID</u>	<u>Organism</u>	<u>%P</u>	<u>Hits</u>	<u>Prot</u>	<u>SM</u>	<u>D</u>	<u>SMID</u>	<u>E</u>
180454	<u>Anopheles gambiae</u> PEST	48%	7426	15328	2715	2111	236353	
3702	<u>Arabidopsis thaliana</u>	47%	13625	29095	2589	2201	364280	
6239	<u>Caenorhabditis elegans</u>	40%	9044	22729	2587	2119	273387	
284593	<u>Candida glabrata</u> CBS138	47%	2455	5181	1802	1710	64185	
214684	<u>Cryptococcus neoformans</u> JEC21	47%	3122	6594	1979	1832	78969	
284592	<u>Debaryomyces hansenii</u> CBS767	45%	2853	6318	1944	1772	67952	
7227	<u>Drosophila melanogaster</u>	49%	9514	19386	2611	2096	341206	
284813	<u>Encephalitozoon cuniculi</u> GB-M1	42%	829	1996	779	1035	23230	
33169	<u>Eremothecium gossypii</u>	48%	2270	4718	1744	1690	58731	
55529	<u>Guillardia theta</u> nucleomorph	42%	263	632	283	430	8379	
9606	<u>Homo sapiens</u>	52%	15347	29511	3069	2301	548894	
284590	<u>Kluyveromyces lactis</u> NRRL Y-1140	46%	2432	5327	1832	1702	61434	
10090	<u>Mus musculus</u>	52%	14090	27071	3079	2278	466530	
39947	<u>Oryza sativa (japonica</u> cultivar-group)	32%	11671	36946	2493	2160	285898	
36329	<u>Plasmodium falciparum</u> 3D7	37%	1952	5267	1466	1482	53025	
10116	<u>Rattus norvegicus</u>	55%	13327	24061	3092	2266	468303	
4932	<u>Saccharomyces cerevisiae</u>	47%	2751	5867	1859	1757	72111	E
284812	<u>Schizosaccharomyces</u> pombe 972h-	49%	2455	5035	1831	1691	66968	
7955	<u>Zebrafish</u>	56%	17265	30602	3088	2272	553067	

Essential  
Genes





**Eukaryote:**Homo sapiens [taxid: 9606]

3 **views** to browse protein domain small-molecule hits for the genome.



Views: [\[Protein\]](#) [\[Small Molecule\]](#) [\[Domain\]](#)

Export Results

15347 of 29511 (52%) proteins have distinct conserved domain hits that bind small molecules.

Help

**SMID** [Proteins](#) [SM](#) [Domain](#) [Description](#) [Links](#)

<a href="#">[SMID]</a> <a href="#">865</a> (6%)	<a href="#">1</a>	zf-C2H2	pfam00096: Zinc finger, C2H2 type. The C2H2 zinc finger is the classical zinc finger domain. The two conserved cysteines and ... <a href="#">more</a>	<a href="#">[G]</a>
<a href="#">[SMID]</a> <a href="#">798</a> (5%)	<a href="#">13</a>	7tm_1	pfam00001: 7 transmembrane receptor (rhodopsin family). <a href="#">more</a>	<a href="#">[G]</a>
<a href="#">[SMID]</a> <a href="#">760</a> (5%)	<a href="#">158</a>	S_TKc	smart00220: Serine/Threonine protein kinases, catalytic domain; Phosphotransferases. Serine or threonine-specific kinase subf... <a href="#">more</a>	<a href="#">[G]</a>
<a href="#">[SMID]</a> <a href="#">759</a> (5%)	<a href="#">158</a>	S_TKc	cd00180: Serine/Threonine protein kinases, catalytic domain. Phosphotransferases of the serine or threonine-specific kinase s... <a href="#">more</a>	<a href="#">[G]</a>
<a href="#">[SMID]</a> <a href="#">741</a> (5%)	<a href="#">156</a>	Pkinase	pfam00069: Protein kinase domain. <a href="#">more</a>	<a href="#">[G]</a>
<a href="#">[SMID]</a> <a href="#">728</a> (5%)	<a href="#">154</a>	TyrKc	smart00219: Tyrosine kinase, catalytic domain; Phosphotransferases. Tyrosine-specific kinase subfamily.... <a href="#">more</a>	<a href="#">[G]</a>

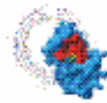
View small-molecule binding sites on genomic proteins

Click on numbers to narrow the scope of the search

# Compare Binding Profiles

## Malaria

- a disease that directly impacts 300-500 million people worldwide and is a prominent economic and social problem in the developing world
- Compare binding profiles & identify small molecules that target proteins of *Plasmodium falciparum* exclusively.



SMID Genomes

### Small Molecule Comparison Across Genomes

3386 small molecules were found across the genomes. In the table below, **number of small molecules** exclusively for each or combination of genome are marked by a '+'. To view the small molecules, click on the number.

A Homo sapiens

B Anopheles gambiae PEST

C Plasmodium falciparum 3D7



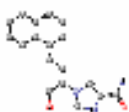
	A	B	C
<u>540</u>	+	-	-
<u>1133</u>	+	+	-
<u>1319</u>	+	+	+
<u>77</u>	+	-	+
<u>247</u>	-	+	-
<u>16</u>	-	+	+
<u>54</u>	-	-	+

#### Select Organisms:

NCBI taxonomy identifier  
or any part of a scientific name.

1	<input type="text" value="homo sapiens"/>
	<input checked="" type="checkbox"/> Homo sapiens
2	<input type="text" value="anopheles"/>
	<input checked="" type="checkbox"/> Anopheles gambiae PEST
3	<input type="text" value="plasmodium"/>
	<input checked="" type="checkbox"/> Plasmodium falciparum 3D7
4	<input type="text"/>
5	<input type="text"/>
	<input type="button" value="Search"/> <input type="button" value="Compare"/>

# DOXP & Fosmidomycin

Links	SM	Proteins
[G]	 <a href="#">FG1</a>	<a href="#">A</a> [SMID] hypothetical protein [GI: <a href="#">23509747</a> ]
[G]	 <a href="#">fosmidomycin</a>	<a href="#">A</a> [SMID] 1-deoxy-D-xylulose 5-phosphate reductoisomerase [GI: <a href="#">23509863</a> ]
[G]	 <a href="#">FR3</a>	<a href="#">A</a> [SMID] AMP deaminase, putative [GI: <a href="#">23619215</a> ]

- **Fosmidomycin**
  - known antibiotic
  - potent inhibitor of **DOXP reductoisomerase**, a key enzyme of the alternative pathway of **isoprenoid synthesis**.

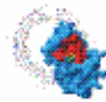
- *Plasmodium* is dependent on this pathway, because it lacks the primary isoprenoid synthesis pathway.

- Effective treatment
- In clinical trials

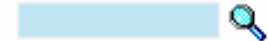
Parasitol Res. 2003 Jun;90 Suppl 2:S71-6.

# Ligand Spectrum

- Hits 73% of bacteria
- For eukaryotes, hit to malaria parasite, rat, but only plants (rice, thale cress) have true ortholog of reductoisomerase but they also have both synthesis pathways (KEGG)



**SMID Genomes**



## Small Molecule Search: fosmidomycin

176 of 1616 organisms found to have proteins with hits to the **fosmidomycin** small molecule.

<b>Archae</b>	1/23	4.3%
<b>Bacteria</b>	171/234	73.1%
<b>Eukaryote</b>	4/19	21.1%

Export Results

Help

<u>TaxID</u>	<u>Class</u>	<u>Organism</u>	<u>Hits</u>	<u>%P</u>	<u>Links</u>
3702	Eukaryote	<a href="#">Arabidopsis thaliana</a>	<a href="#">1</a>	< 0.1%	<a href="#">[SMID]</a>
39947	Eukaryote	<a href="#">Oryza sativa (japonica cultivar-group)</a>	<a href="#">1</a>	< 0.1%	<a href="#">[SMID]</a>
36329	Eukaryote	<a href="#">Plasmodium falciparum 3D7</a>	<a href="#">1</a>	< 0.1%	<a href="#">[SMID]</a>
10116	Eukaryote	<a href="#">Rattus norvegicus</a>	<a href="#">1</a>	< 0.1%	<a href="#">[SMID]</a>

# Essential Genes Make Better Targets

- Database of Essential Genes (DEG)
  - 9 organisms
- *Mycoplasma genitalium*
  - nonchlamydial nongonococcal urethritis in men
- *M. penetrans*

## Select Organisms:

NCBI taxonomy identifier  
or any part of a scientific name.

1 homo sapiens  
 Homo sapiens

2 mycoplasma  
 Mycoplasma gallisepticum R  
 Mycoplasma genitalium G37  
 Mycoplasma hyopneumoniae 232  
 Mycoplasma hyopneumoniae 7448  
 Mycoplasma hyopneumoniae J  
 Mycoplasma mobile 163K  
 Mycoplasma mycoides subsp. mycoides SC str. PG1  
 Mycoplasma penetrans HF-2  
 Mycoplasma pneumoniae M129  
 Mycoplasma pulmonis UAB CTIP  
 Mycoplasma synoviae 53  
 Mycoplasma arthritis bacteriophage MAV1  
 Mycoplasma virus P1

3  
4  
5

Search Compare

# Possible Urethritis Targets

**A** Homo sapiens

**B** Mycoplasma genitalium G37

**C** Mycoplasma penetrans HF-2

- FM2 (formycin A der.) nucleoside inhibitor acts on essential gene
- hexameric form in lower orgs.

**A B C**

**2351** + - -

**55** + + -

**432** + + +

**231** + - +

**8** - + -

**15** - + +

**29** - - +

## Legend

**A** Mycoplasma genitalium G37

**B** Mycoplasma penetrans HF-2

## Links

[G]



## SM

CDI

## Proteins

A [SMID] hypothetical protein [GI:[12045318](#)]  
B [SMID] putative enzyme of deoxy-xylulose pathway YgbB [GI:[26554474](#)]

[G]



FM1

A E [SMID] purine-nucleoside phosphorylase (deoD) [GI:[12044899](#)]  
B [SMID] purine nucleoside phosphorylase [GI:[26553558](#)]

[G]



FM2

A E [SMID] purine-nucleoside phosphorylase (deoD) [GI:[12044899](#)]  
B [SMID] purine nucleoside phosphorylase [GI:[26553558](#)]

[G]



MDR

A E [SMID] purine-nucleoside phosphorylase (deoD) [GI:[12044899](#)]  
B [SMID] 5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase [GI:[26553973](#)]

[G]

SB9

SB9

A E [SMID] polypeptide deformylase (def) [GI:[12044958](#)]  
B [SMID] polypeptide deformylase [GI:[26554017](#)]

[G]



Spermidine

A E [SMID] lipoprotein, putative [GI:[12044895](#)]  
B [SMID] putative lipoprotein [GI:[26554309](#)]  
B [SMID] putative lipoprotein [GI:[26554314](#)]

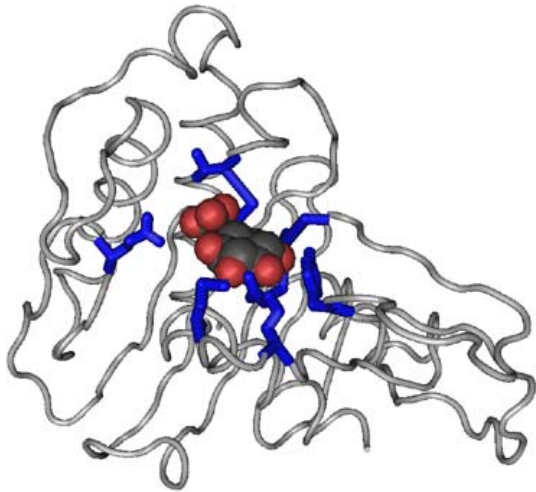
# Conclusions

- *Simple* domain linker prediction based on amino acid composition
  - Good start for sequences that have no similarity to anything else
  - Prove useful in combining with more sophisticated methods
- *Annotation* of small molecule binding sites based on conserved domains
  - Interesting drug targets can be identified by comparing binding profiles between sequenced genomes.

# Acknowledgements

- Dumontier Lab

- Jose Cruz
- Zhen Liu
- Daniel Oropeza
- Salim Quadri



- Blueprint

- Christopher Hogue
- Howard Feldman
- Kevin Snyder
- Susan Ling
- John Salama
- Marc Dumontier

- Funding Agencies

- Genome Canada
- Genome Ontario
- CIHR
- ORDCF

Michel Dumontier

michel\_dumontier@carleton.ca

dumontierlab.com

# Extra Slides

# Analyzing Complete Genomes

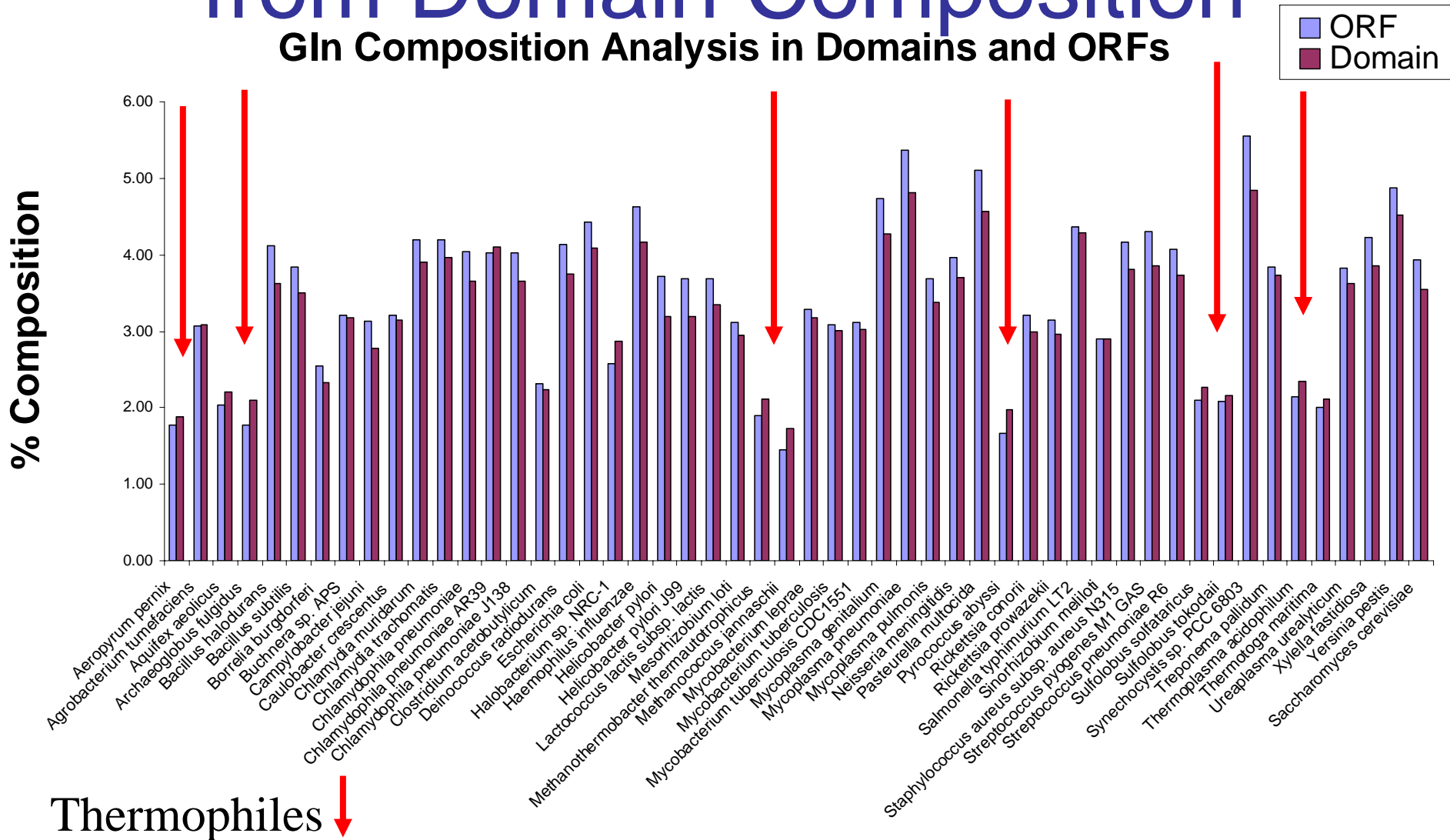
T-test suggests that the distribution of amino acids significantly changes from conserved domains to rest of the full open-reading frame

- Preference for smaller hydrophobic residues in domains.
- Preferences for negatively charged residues over their amide derivatives
- Universal preferences across all genomes

	no-filt	filt
ASP	-41.5	-41.1
GLU	-26.1	-39.1
VAL	-38.0	-57.7
GLY	-28.8	-47.7
HIS	-22.8	-27.2
ALA	-14.0	-34.4
ARG	-2.5	-3.6
ILE	-0.8	-21.1
CYS	0.0	-2.2
LYS	-0.6	-2.5
THR	-0.6	-14.6
PRO	-0.4	-2.6
MET	-0.3	-11.1
ASN	-6.4	-3.5
GLN	-11.0	-5.4
TYR	-7.0	-2.7
PHE	-21.7	-12.1
TRP	-27.8	-20.7
SER	-35.4	-30.4
LEU	-45.3	-13.7

# Species-Specific Signatures from Domain Composition

## Gln Composition Analysis in Domains and ORFs



# Compositional Bias Sufficient to Identify Domains From Different Species

- The log likelihood of a domain having the amino acid composition expected from domains of a particular organism
- Cross-validation:
  - Attempt to discriminate between the model and template domain
  - Average of  $85\pm 8\%$  success in identifying species-specific domains using amino acid composition alone

# Ligand Score

- Initial Ligand Score =  $(1 - \log_{10} E)^{1/2} * Id / S^2$ 
  - E is the RPS-BLAST E-value
  - Id is the % identity of binding site residues
  - S is the relative entropy score
- Final Ligand Score incorporates the binding site occupancy from clustering of all domain hits.