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
- **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

- Non-redundant set of 655 segmented, multi-domain proteins.
Hydrophobic residues decreased in domain linkers

Amino Acid Composition in Domains and Linkers

Hydrophobic residues decreased in domain linkers
Domain Linker Index

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

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

<table>
<thead>
<tr>
<th>AA</th>
<th>DLI</th>
<th>REI</th>
<th>GHL</th>
<th>KDH</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.806</td>
<td>−1.721</td>
<td>0.273</td>
<td>0.767</td>
</tr>
<tr>
<td>R</td>
<td>0.508</td>
<td>1.637</td>
<td>−1.287</td>
<td>−1.342</td>
</tr>
<tr>
<td>N</td>
<td>−0.823</td>
<td>−0.042</td>
<td>0.447</td>
<td>−1.088</td>
</tr>
<tr>
<td>D</td>
<td>−0.773</td>
<td>−0.042</td>
<td>0.691</td>
<td>−1.088</td>
</tr>
<tr>
<td>C</td>
<td>−0.161</td>
<td>−0.042</td>
<td>1.894</td>
<td>1.011</td>
</tr>
<tr>
<td>Q</td>
<td>0.052</td>
<td>0.798</td>
<td>−0.450</td>
<td>−1.008</td>
</tr>
<tr>
<td>F</td>
<td>0.127</td>
<td>0.798</td>
<td>−0.485</td>
<td>−1.008</td>
</tr>
<tr>
<td>G</td>
<td>−1.264</td>
<td>−0.882</td>
<td>1.397</td>
<td>0.030</td>
</tr>
<tr>
<td>H</td>
<td>−0.380</td>
<td>−0.042</td>
<td>−0.163</td>
<td>−0.907</td>
</tr>
<tr>
<td>I</td>
<td>1.441</td>
<td>−0.042</td>
<td>0.639</td>
<td>1.671</td>
</tr>
<tr>
<td>L</td>
<td>0.893</td>
<td>−0.042</td>
<td>−0.781</td>
<td>1.436</td>
</tr>
<tr>
<td>K</td>
<td>0.072</td>
<td>1.637</td>
<td>0.447</td>
<td>−1.142</td>
</tr>
<tr>
<td>M</td>
<td>1.205</td>
<td>0.798</td>
<td>−0.319</td>
<td>0.800</td>
</tr>
<tr>
<td>F</td>
<td>−0.049</td>
<td>−0.042</td>
<td>−1.078</td>
<td>1.101</td>
</tr>
<tr>
<td>P</td>
<td>−2.799</td>
<td>−2.561</td>
<td>−2.646</td>
<td>−0.372</td>
</tr>
<tr>
<td>S</td>
<td>−0.738</td>
<td>−0.042</td>
<td>0.421</td>
<td>−0.104</td>
</tr>
<tr>
<td>T</td>
<td>−0.465</td>
<td>−0.042</td>
<td>−0.189</td>
<td>−0.070</td>
</tr>
<tr>
<td>W</td>
<td>0.560</td>
<td>−0.042</td>
<td>0.874</td>
<td>−0.137</td>
</tr>
<tr>
<td>Y</td>
<td>0.590</td>
<td>0.798</td>
<td>−0.041</td>
<td>−0.271</td>
</tr>
<tr>
<td>V</td>
<td>1.199</td>
<td>−0.882</td>
<td>0.352</td>
<td>1.570</td>
</tr>
</tbody>
</table>

REI – Residue Entropy Index – derived from Galzitskaya & Melnik\(^1\) 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\(^2\)

KDH – Kyte-Doolittle hydropathy index - control

---


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%
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 GeneInfo Identifier (GI)
- Accession (Accession)
- PDBS Identifier (PDBS)
- PDB Identifier (PDB) - 4 letter code + 1 letter chain (optional)
- FASTA formatted sequence (FASTA)

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

© VAST © SCOP

Example:
Src Family Kinase

Help!
Armadillo Results

The results of the prediction are summarized in the SVO graph below. Click on the Legend entries to show/hide predictions from different indices. Right click on the graph to zoom in/out. Click here for more information to interpret the graph.

Download the tab delimited values here.
Armadoillo Domain Linker Prediction

1NHQ-Nadh Peroxidase (Npx) (E.C.11.1.1) Mutant With Cys 42 Replaced By Ser (C42s)(Oxidoreductase (H2o2/A))
Predicting Linkers for Proteins with Conserved Domains
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.
Interaction Between Domain: TrpRS_core and ATP

**SMD id:** 424831

**TrpRS_core** (c000060) Trypsophyl tRNA synthetase (TrpRS) catalytic core domain. TrpRS is a homodimer which attaches Trp to the appropriate tRNA. The domain is responsible for the ATP-dependent formation of the enzyme-bound aminoacyl-adenylate. It contains a characteristic HIGN and HMBKS motifs, which are involved in ATP binding.

**Small Molecule:** ATP

**MNID id:** 34221_B  **PDB id:** 1YID_B

**BIND id:** 330151

**Domain Family Multiple Alignment**

**Options:** Launch Viewer

**Viewer:** Cons (ASN1)

**Complexity:** All-Atom Model

---

The locations where
\[
- \text{O-(hydroxy)(hydroxy)phospho-phospho-phospho-phospho-phospho-hydroxy-5-deoxy-D-ribose-9-phosphate-6-amino (ATP) bonds the domain for this record are highlighted, red for conserved residues, blue for similar residues, and yellow for non-conserved residues. Lowercase residues do not align with the consensus and represent insertions or deletions relative to the consensus. Conservation cannot be measured when the binding site aligns only to gaps in the consensus, and in this case, no coloured residues are displayed below.}
\]

The PDB code highlighted in red, if any, is the structure from which this interaction was derived. The binding site residues in this PDB file will be highlighted, but please note some additional residues may be highlighted as well, if there are redundant interactions from other PDB files with a similar binding site.
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

RPS-BLAST

Alignment to conserved small-molecule binding domains

Evaluate binding site conservation

cluster binding sites from all domains

SMID

domain small-molecule interactions

Predicted small-molecule binding sites and ligand scores

Protein Sequence
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 assays¹

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** 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 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](http://smid.blueprint.org).

Currently available at [http://smid.blueprint.org](http://smid.blueprint.org)
### Browsing SMID-Genomes

<table>
<thead>
<tr>
<th>TaxID</th>
<th>Organism</th>
<th>%P Hits</th>
<th>Prot</th>
<th>SN</th>
<th>D</th>
<th>SMID E</th>
</tr>
</thead>
<tbody>
<tr>
<td>160454</td>
<td><em>Anopheles gambiae PEST</em></td>
<td>48%</td>
<td>7426</td>
<td>15328 2715 2111 236353</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3702</td>
<td><em>Arabidopsis thaliana</em></td>
<td>47%</td>
<td>13625 29095 2569 2201 364280</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6239</td>
<td><em>Caenorhabditis elegans</em></td>
<td>40%</td>
<td>9044 22729 2587 2119 273387</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>284593</td>
<td><em>Candida glabrata CBS138</em></td>
<td>47%</td>
<td>2455 5181 1802 1710 64185</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>214684</td>
<td><em>Cryptococcus neoformans JEC21</em></td>
<td>47%</td>
<td>3122 6594 1979 1832 78969</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>284592</td>
<td><em>Debaryomyces hansenii CBS767</em></td>
<td>45%</td>
<td>2853 6318 1944 1772 67952</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7227</td>
<td><em>Drosophila melanogaster</em></td>
<td>49%</td>
<td>9514 19386 2611 2096 341206</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>284813</td>
<td><em>Encephalitozoon cuniculi GB-M1</em></td>
<td>42%</td>
<td>829 1996 779 1035 23230</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33169</td>
<td><em>Eremothecium gossypii</em></td>
<td>48%</td>
<td>2270 4718 1744 1690 58731</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55529</td>
<td><em>Guillardia theta nucleornorph</em></td>
<td>42%</td>
<td>263 632 283 430 8379</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9606</td>
<td><em>Homo sapiens</em></td>
<td>52%</td>
<td>15347 29511 3069 2301 548894</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>284590</td>
<td><em>Kluveromyces lactis NRRL Y-1140</em></td>
<td>46%</td>
<td>2432 5327 1832 1702 61434</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10090</td>
<td><em>Mus musculus</em></td>
<td>52%</td>
<td>14090 27071 3079 2278 466530</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39947</td>
<td><em>Oryza sativa (japonica cultivar-group)</em></td>
<td>32%</td>
<td>11671 36946 2493 2160 285898</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36329</td>
<td><em>Plasmodium falciparum 3D7</em></td>
<td>37%</td>
<td>1952 5267 1466 1482 53025</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10116</td>
<td><em>Rattus norvegicus</em></td>
<td>55%</td>
<td>13327 24061 3092 2266 468303</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4932</td>
<td><em>Saccharomyces cerevisiae</em></td>
<td>47%</td>
<td>2751 5867 1699 1757 7211</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>284812</td>
<td><em>Schizosaccharomyces pombe 972h</em></td>
<td>49%</td>
<td>2455 5035 1831 1691 66968</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7955</td>
<td><em>Zebrafish</em></td>
<td>56%</td>
<td>17265 30602 3088 2272 553067</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3 views to browse protein domain small-molecule hits for the genome.

**Eukaryote:** Homo sapiens [taxid: 9605]

Views: [Protein] [Small Molecule] [Domain]

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

### Table of Small-Molecule Binding Sites on Genomic Proteins

<table>
<thead>
<tr>
<th>SMID</th>
<th>Proteins</th>
<th>SM</th>
<th>Domain</th>
<th>Description</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>[SMID] 666 (5%)</td>
<td>1</td>
<td>zfC2H2</td>
<td>pfam00066: Zinc finger, C2H2 type. The C2H2 zinc finger is the classical zinc finger domain. The two conserved cysteines and...</td>
<td>more</td>
<td></td>
</tr>
<tr>
<td>[SMID] 796 (5%)</td>
<td>13</td>
<td>7tm_1</td>
<td>pfam00001: 7 transmembrane receptor (rhodopsin family)</td>
<td>more</td>
<td></td>
</tr>
<tr>
<td>[SMID] 790 (5%)</td>
<td>156</td>
<td>S_TKc</td>
<td>smart0220: Serine/Threonine protein kinases, catalytic domain; Phosphotransferases. Serine or threonine-specific kinase subf...</td>
<td>more</td>
<td></td>
</tr>
<tr>
<td>[SMID] 759 (5%)</td>
<td>156</td>
<td>S_TKc</td>
<td>cd00130: Serine/Threonine protein kinases, catalytic domain. Phosphotransferases of the serine or threonine-specific kinase subf...</td>
<td>more</td>
<td></td>
</tr>
<tr>
<td>[SMID] 741 (5%)</td>
<td>156</td>
<td>Pkinase</td>
<td>pfam00069: Protein kinase domain.</td>
<td>more</td>
<td></td>
</tr>
<tr>
<td>[SMID] 726 (5%)</td>
<td>154</td>
<td>TyrKc</td>
<td>smart00219: Tyrosine kinase, catalytic domain; Phosphotransferases. Tyrosine-specific kinase subfamily...</td>
<td>more</td>
<td></td>
</tr>
</tbody>
</table>

View small-molecule binding sites on genomic proteins

Click on numbers to narrow the scope of the search
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.

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 genoi are marked by a ‘+’. To view the small molecules, click on the number.

<table>
<thead>
<tr>
<th></th>
<th>A Homo sapiens</th>
<th>B Anopheles gambiae PEST</th>
<th>C Plasmodium falciparum 3D7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+ - -</td>
<td>+ + -</td>
<td>+ + +</td>
</tr>
<tr>
<td></td>
<td>540</td>
<td>1133</td>
<td>1319</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+ - +</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td></td>
<td>247</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>54</td>
</tr>
</tbody>
</table>

Select Organisms:

- homo sapiens
- anopheles
- Anopheles gambiae PEST
- plasmodium
- Plasmodium falciparum 3D7

Search  Compare
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

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)
Essential Genes Make Better Targets

- Database of Essential Genes (DEG)
  - 9 organisms

- **Mycoplasma genitalium**
  - nonchlamydial nongonococcal urethritis in men

- **M. penetrans**
Possible Urethritis Targets

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

Legend:
- A Homo sapiens
- B Mycoplasma genitalium G37
- C Mycoplasma penetrans HF-2

Links | SM | Proteins
--- | --- | ---
A | CD1 | A [SMID] hypothetical protein [GI:15045216]
B |  | B [SMID] putative enzyme of deoxy-xylulose pathway YgbD [GI:26554474]

| 2351 | + + - | A [SMID] purine-nucleoside phosphorylase (deoD) [GI:12044899]
| 55 | + + - | B [SMID] purine nucleoside phosphorylase [GI:26555586]

| 432 | + + + | A [SMID] purine-nucleoside phosphorylase (deoD) [GI:12044899]
| 231 | + + - | B [SMID] purine nucleoside phosphorylase [GI:26555586]

| 15 | + + + | A [SMID] purine-nucleoside phosphorylase (deoD) [GI:12044899]
| 29 | + + - | B [SMID] 5'-methylthioadenosine/5'-adenosylhomocysteine nucleosidase [GI:26555375]

| SB6 | SB9 | A [SMID] polypeptide deformylase (def) [GI:12044950]
|  |  | B [SMID] polypeptide deformylase [GI:26555017]

| Spermidine | | A [SMID] lipoprotein, putative [GI:12044955]
|  |  | B [SMID] putative lipoprotein [GI:26554820]
|  |  | C [SMID] putative lipoprotein [GI:26554514]
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

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

Dumontier et al. 2002. BMC Bioinformatics. 3: 39
Species-Specific Signatures from Domain Composition

Gln Composition Analysis in Domains and ORFs

% Composition

Thermophiles
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±8% success in identifying species-specific domains using amino acid composition alone
Ligand Score

• Initial Ligand Score = \( (1 - \log_{10} E)^{\frac{1}{2}} \times \frac{Id}{S^2} \)
  – 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.