Characterizing and Predicting Hexose-Binding Sites

Houssam Nassif
http://pages.cs.wisc.edu/~hous21/

CIBM Seminar
25 January 2011
Outline

1. Background
   - Motivation
   - Hexoses
   - Atomic Interactions

2. Problem Representation

3. Glucose Binding Modeling
   - Classification Approach
   - Results

   - Rule Inference
   - Results
Hexoses Pathways

- 6-carbon sugar molecules
- Key role in several biochemical pathways
  - cellular energy release
  - signaling
  - carbohydrate synthesis
  - regulation of gene expression...
Tasks

- Galactose, glucose, mannose
- High specificity to diverse protein families
- Lack of glucose model
- No data-driven comparison to biochemical findings

Tasks

- Glucose-binding model
- Empirical comparison to wet-lab findings
Tasks

- Galactose, glucose, mannose
- High specificity to diverse protein families
- Lack of glucose model
- No data-driven comparison to biochemical findings

 Tasks

- Glucose-binding model
- Empirical comparison to wet-lab findings
Outline

1 Background
   - Motivation
   - Hexoses
     - Atomic Interactions

2 Problem Representation

3 Glucose Binding Modeling
   - Classification Approach
   - Results

4 Hexose Binding Rules Empirical Generation
   - Rule Inference
   - Results
Hexose Stereoisomers

Figure: (a) D-Galactose; (b) D-Glucose; (c) D-Mannose
Hexose Structure

Figure: Glucose

- Contains two functional groups
- Both groups can interact together

(a) Carbonyl
(b) Hydroxyl
Hexose Cyclization

- The molecule folds on itself and forms a **pyranose** ring.
- In two different ways. Watch the star!

(a) $\alpha$-pyranose

(b) Open chain

(c) $\beta$-pyranose
Outline

1 Background
   - Motivation
   - Hexoses
   - Atomic Interactions

2 Problem Representation

3 Glucose Binding Modeling
   - Classification Approach
   - Results

4 Hexose Binding Rules Empirical Generation
   - Rule Inference
   - Results
Covalent Bonds

- Close and strong interaction
- Forms a molecule
- Atoms share electrons
- Electronegativity:
  - Equal $\Rightarrow$ nonpolar
  - Different $\Rightarrow$ polar
- Partial charges

**Figure:** Covalent bond

**Definition**

**Electronegativity:** Measure of atom’s attraction for electrons
Covalent Bonds

- Close and strong interaction
- Forms a molecule
- Atoms share electrons
- Electronegativity:
  - Equal $\Rightarrow$ nonpolar
  - Different $\Rightarrow$ polar
- Partial charges

**Figure:** Covalent bond

**Definition**

**Electronegativity:** Measure of atom’s attraction for electrons
Covalent Bonds

- Close and strong interaction
- Forms a molecule
- Atoms share electrons
- Electronegativity:
  - Equal $\Rightarrow$ nonpolar
  - Different $\Rightarrow$ polar
- Partial charges

**Figure:** Covalent polar bond

**Definition**

**Electronegativity:** Measure of atom’s attraction for electrons
Hydrogen Bonds

- Attraction between a positively charged H and a negatively charged atom
- Hexose attaches to the protein using hydrogen bonds

Figure: Hydrogen bond
Van der Waals and Hydrophobicity

**Definition**

**Van der Waals Forces:** Weak electrostatic attraction and repulsion forces

**Definition (Hydrophobicity)**


- Dual nature:
  - Pyranose ring is hydrophobic
  - Hydroxyl group is hydrophilic
Van der Waals and Hydrophobicity

**Definition**

**Van der Waals Forces:** Weak electrostatic attraction and repulsion forces

**Definition (Hydrophobicity)**

**Hydrophobic:** water hating. **Hydrophilic:** water loving. Hydrophobic/Hydrophilic atoms tend to gather together.

- Dual nature:
  - Pyranose ring is hydrophobic
  - Hydroxyl group is hydrophilic
Outline

1. Background
   - Motivation
   - Hexoses
   - Atomic Interactions

2. Problem Representation

3. Glucose Binding Modeling
   - Classification Approach
   - Results

   - Rule Inference
   - Results
Binding-Site Representation
Binding-Site Feature Extraction

1: procedure EXTRACT FEATURES(binding site center)
2: for all concentric layers do
3: for all PDB atoms do
4: get coordinates
5: get charge
6: get hydrophobicity
7: get hydrogen-bonding
8: get residue
9: end for
10: end for
11: end procedure
## Binding-Site Features

<table>
<thead>
<tr>
<th>Atomic Feature</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charge</td>
<td>Negative, Neutral, Positive</td>
</tr>
<tr>
<td>Hydrogen-bonding</td>
<td>Non-hydrogen bonding, Hydrogen-bonding</td>
</tr>
<tr>
<td>Hydrophobicity</td>
<td>Hydrophilic, Hydroneutral, Hydrophobic</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Residue Grouping</th>
<th>Amino Acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aromatic</td>
<td>HIS, PHE, TRP, TYR</td>
</tr>
<tr>
<td>Aliphatic</td>
<td>ALA, ILE, LEU, MET, VAL</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASN, CYS, GLN, GLY, PRO, SER, THR</td>
</tr>
<tr>
<td>Acidic</td>
<td>ASP, GLU</td>
</tr>
<tr>
<td>Basic</td>
<td>ARG, LYS</td>
</tr>
</tbody>
</table>
Empirical evidence suggests that hexose docking is not accompanied by protein conformational changes (galactose)

- **Hexose dataset**
  - Mine PDB for glucose/hexoses
  - Discard theoretical structures and redundancies
  - Discard covalently bound and floating in medium
  - Impose 30% cut-off overall sequence identity
  - Discard if other ligands bind or are present

- **Non-hexose dataset**
  - Non-sugar binding sites
  - Glucose/hexose-like binding sites
  - Random non-binding sites
Classifier Outline

a) Known glucose binding sites
b) Known non-glucose sites
c) Unknown site
d) Site feature vector
e) Non-Site feature vector
f) Unknown site feature vector
g) Classifier (training phase)
h) classifier (testing phase)
i) Glucose binding site
j) Not a glucose binding site
Outline

1. Background
   - Motivation
   - Hexoses
   - Atomic Interactions

2. Problem Representation

3. Glucose Binding Modeling
   - Classification Approach
   - Results

   - Rule Inference
   - Results
Support Vector Machines (SVM)

- Construct the *optimal separating hyperplane* (usually in a higher feature space)
  - Maximize *margins*: minimal distance from the hyperplane
  - Only *Support Vectors (SV)* specify the margins/hyperplane
  - Small number of SV ⇔ good generalization
Support Vector Machines (SVM)

- Construct the *optimal separating hyperplane* (usually in a higher feature space)
- Maximize *margins*: minimal distance from the hyperplane
- Only *Support Vectors (SV)* specify the margins/hyperplane
- Small number of SV $\leftrightarrow$ good generalization
Support Vector Machines (SVM)

- Construct the *optimal separating hyperplane* (usually in a higher feature space)
- Maximize *margins*: minimal distance from the hyperplane
- Only *Support Vectors (SV)* specify the margins/hyperplane
- Small number of SV $\iff$ good generalization
Support Vector Machines (SVM)

- Construct the *optimal separating hyperplane* (usually in a higher feature space)
- Maximize *margins*: minimal distance from the hyperplane
- Only *Support Vectors (SV)* specify the margins/hyperplane
- Small number of SV $\Leftrightarrow$ good generalization
Random Forest (RF)

- High features/examples ratio $\Rightarrow$ curse of dimensionality
- **Feature selection**: select the best feature subset

Random Forest feature selection:
- Based on multiple classification trees
- Provides direct feature importance measure
- Can be used when feature number $\gg$ samples
- Robust to noise
- Low bias and low variance
Random Forest (RF)

- High features/examples ratio $\Rightarrow$ *curse of dimensionality*
- *Feature selection*: select the best feature subset

Random Forest feature selection:
- Based on multiple classification trees
- Provides direct feature importance measure
- Can be used when feature number $\gg$ samples
- Robust to noise
- Low bias and low variance
Experimental Setting

<table>
<thead>
<tr>
<th>Ligand</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glucose</td>
<td>43</td>
</tr>
<tr>
<td>Non-sugar</td>
<td>36</td>
</tr>
<tr>
<td>Other sugars</td>
<td>15</td>
</tr>
<tr>
<td>Non-binding</td>
<td>17</td>
</tr>
</tbody>
</table>

- 8 concentric layers
  - Inner layer width: 3 Å
  - Other layers width: 1 Å
- Non-linear RBF SVM
- Tune gamma and cost parameters
- Leave-one-out cross-validation
Outline

1. Background
   - Motivation
   - Hexoses
   - Atomic Interactions

2. Problem Representation

3. Glucose Binding Modeling
   - Classification Approach
   - Results

   - Rule Inference
   - Results
**Importance of Water and Ions**

- Ordered water molecules and ions affect ligand specificity

<table>
<thead>
<tr>
<th>Properties</th>
<th>Whole set error</th>
<th>Subset error*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Include water and ions</td>
<td>18.92%</td>
<td>7.81%</td>
</tr>
<tr>
<td>Discard water</td>
<td>18.92%</td>
<td>10.94%</td>
</tr>
<tr>
<td>Discard ions</td>
<td>20.27%</td>
<td>7.81%</td>
</tr>
<tr>
<td>Discard water and ions</td>
<td>20.27%</td>
<td>12.5%</td>
</tr>
</tbody>
</table>

* Lacks the other sugars binding sites negatives
## Properties Feature Selection

<table>
<thead>
<tr>
<th>Property</th>
<th>RF</th>
<th>Feature Number</th>
<th>Error (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>SV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charge</td>
<td>false</td>
<td>24</td>
<td>24.32</td>
<td>79.31</td>
<td>73.33</td>
<td>77.03</td>
</tr>
<tr>
<td></td>
<td>true</td>
<td>5</td>
<td>14.86</td>
<td>86.21</td>
<td>84.44</td>
<td>44.59</td>
</tr>
<tr>
<td>H-Bonding</td>
<td>false</td>
<td>16</td>
<td>17.57</td>
<td>82.76</td>
<td>82.22</td>
<td>41.89</td>
</tr>
<tr>
<td></td>
<td>true</td>
<td>3</td>
<td>14.86</td>
<td>82.76</td>
<td>86.67</td>
<td>47.30</td>
</tr>
<tr>
<td>Hydro</td>
<td>false</td>
<td>24</td>
<td>16.22</td>
<td>72.41</td>
<td>91.11</td>
<td>65.57</td>
</tr>
<tr>
<td></td>
<td>true</td>
<td>15</td>
<td>12.16</td>
<td>82.76</td>
<td>91.11</td>
<td>40.54</td>
</tr>
<tr>
<td>Residues</td>
<td>false</td>
<td>48</td>
<td>21.62</td>
<td>48.28</td>
<td>97.78</td>
<td>100.0</td>
</tr>
<tr>
<td></td>
<td>true</td>
<td>19</td>
<td>09.46</td>
<td>93.10</td>
<td>88.89</td>
<td>41.89</td>
</tr>
<tr>
<td>Combined</td>
<td>false</td>
<td>112</td>
<td>18.92</td>
<td>75.86</td>
<td>84.44</td>
<td>79.73</td>
</tr>
<tr>
<td></td>
<td>true</td>
<td>24</td>
<td>08.11</td>
<td>89.66</td>
<td>93.33</td>
<td>40.54</td>
</tr>
</tbody>
</table>
Charge Features

- Negatively charged
- Layer 1: Steric hindrance, non-binding sites
- Layer 2: Small moiety non-sugar binding sites
Hydrogen Bond Features

- Importance of layer 3: Hydrogen-bonding atoms at the protein-glucose interface
Hydrophobicity Features

- Mostly hydrophilic
- Notice layer 7 hydrophobic feature
- Dual nature
Residue Features

- Prominence of negatively charged carboxylate residues
- Aromatic residue plays a role in glucose docking
### Glucose Binding-Site Classifier

<table>
<thead>
<tr>
<th>Features</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>L4</th>
<th>L5</th>
<th>L6</th>
<th>L7</th>
<th>L8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative Charge</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Neutral Charge</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non H-Bonding</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H-Bonding</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Hydrophilic</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Hydroneutral</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydrophobic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Neutral Residue</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acidic Residue</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Glucose Binding Modeling Summary

- First glucose binding model
- Requires specification of binding-site
- Model sensitive to negative dataset
- Findings in accordance with biochemical knowledge

Outline

1. Background
   - Motivation
   - Hexoses
   - Atomic Interactions

2. Problem Representation

3. Glucose Binding Modeling
   - Classification Approach
   - Results

   - Rule Inference
   - Results
Inductive Logic Programming (ILP)

Definition

Inductive Logic Programming (ILP): Machine learning approach that learns a set of first-order logic rules that explain the data

1. Generates easy to interpret if-then rules
2. Allows user interaction through background knowledge
3. Operates on relational datasets
Inductive Logic Programming (ILP)

**Definition**

Inductive Logic Programming (ILP): Machine learning approach that learns a set of first-order logic rules that explain the data

1. Generates easy to interpret if-then rules
2. Allows user interaction through background knowledge
3. Operates on relational datasets
ILP Example

Example

\[ P(A), \text{red}(A), \text{big}(A), \text{round}(A), \text{sibling}(A, B) \]
ILP Example

Example

\( P(A), \text{red}(A), \text{big}(A), \text{round}(A) \)

\( \text{sibling}(A, B) \)
ILP Example

Example

\( P(A), \, red(A), \, big(A), \, round(A) \), sibling(A, B)
ILP Example

Example

\[ P(A), \text{red}(A), \text{big}(A), \text{round}(A), \text{sibling}(A, B) \]
ILP Example

Example

P(A), red(A), big(A), round(A)
sibling(A, B)

- P(X) if square(X)
- P(X) if red(X) ∧ big(x)
  - 1 false positive
- P(X) if sibling(X, Y) ∧ square(Y)
  - 1 false negative
- Form theory
ILP Example

Example

\[ P(A), \text{red}(A), \text{big}(A), \text{round}(A), \text{sibling}(A, B) \]
**ILP Example**

- $P(X)$ if $\text{square}(X)$
- $P(X)$ if $\text{red}(X) \land \text{big}(x)$
  - 1 false positive
- $P(X)$ if $\text{sibling}(X, Y) \land \text{square}(Y)$
  - 1 false negative
- Form *theory*

**Example**

$P(A), \text{red}(A), \text{big}(A), \text{round}(A), \text{sibling}(A, B)$
ILP Example

Example
\[ P(A), \text{red}(A), \text{big}(A), \text{round}(A), \text{sibling}(A, B) \]

- \( P(X) \) if \( \text{square}(X) \)
- \( P(X) \) if \( \text{red}(X) \land \text{big}(x) \)
  - 1 false positive
- \( P(X) \) if \( \text{sibling}(X, Y) \land \text{square}(Y) \)
  - 1 false negative
- Form \textit{theory}
ILP Example

Example:

\[ P(A), \text{red}(A), \text{big}(A), \text{round}(A) \]

\[ \text{sibling}(A, B) \]

- \( P(X) \) if \( \text{square}(X) \)
- \( P(X) \) if \( \text{red}(X) \wedge \text{big}(x) \)
  - 1 false positive
- \( P(X) \) if \( \text{sibling}(X, Y) \wedge \text{square}(Y) \)
  - 1 false negative
- Form \textit{theory}
**ILP Search**

- Pick a positive instance
- Construct the Bottom Clause, most specific clause
- Top-down search: Start with most general rule, add bottom clause predicates
- Bottom-up search: Start with bottom clause, remove predicates

**Example (Bottom Clause (A))**

- $\text{red}(A)$, $\text{big}(A)$, $\text{round}(A)$,
- $\text{siblings}(A, B)$,
- $\text{red}(B)$, $\text{big}(B)$, $\text{round}(B)$
ILP Search

- Pick a positive instance
- Construct the **Bottom Clause**, most specific clause
  - **Top-down search**: Start with most general rule, add bottom clause predicates
  - **Bottom-up search**: Start with bottom clause, remove predicates

Example (Bottom Clause (A))

\[
\text{red}(A), \text{big}(A), \text{round}(A),
\text{siblings}(A, B),
\text{red}(B), \text{big}(B), \text{round}(B)
\]
ILP Search

- Pick a positive instance
- Construct the *Bottom Clause*, most specific clause
- *Top-down search*: Start with most general rule, add bottom clause predicates
- *Bottom-up search*: Start with bottom clause, remove predicates

**Example (Bottom Clause (A))**

- red(A), big(A), round(A), sibling(A, B), red(B), big(B), round(B)
ILP Search

- Pick a positive instance
- Construct the *Bottom Clause*, most specific clause
- *Top-down search:* Start with most general rule, add bottom clause predicates
- *Bottom-up search:* Start with bottom clause, remove predicates

**Example (Bottom Clause (A))**

```
red(A), big(A), round(A),
sibling(A, B),
red(B), big(B), round(B)
```
### Experimental Setting

<table>
<thead>
<tr>
<th>Ligand</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Galactose</td>
<td>33</td>
</tr>
<tr>
<td>Glucose</td>
<td>35</td>
</tr>
<tr>
<td>Mannose</td>
<td>12</td>
</tr>
<tr>
<td>Non-sugar</td>
<td>27</td>
</tr>
<tr>
<td>Hexose-like</td>
<td>22</td>
</tr>
<tr>
<td>Non-binding</td>
<td>31</td>
</tr>
</tbody>
</table>

- One layer
- Compute distances between atoms and center
- 10-folds cross-validation
- Try both search techniques
- Compare empirical generated rules to known biochemical ones
Outline

1. Background
   - Motivation
   - Hexoses
   - Atomic Interactions

2. Problem Representation

3. Glucose Binding Modeling
   - Classification Approach
   - Results

   - Rule Inference
   - Results
Known Biochemical Rules

1. Hexose pyranose hydrophobically stacks on aromatic residues ring (Trp, Tyr, Phe, His)
2. May be sandwiched between two or more aromatics
3. Planar polar residues establish network of hydrogen-bonds with hexose (Asn, Asp, Gln, Glu, Arg)
4. Hydrogen-bonding atoms interface with hexose
5. Frequency of hydrogen-bonding: (Asp, Asn) > Glu > (Arg, His, Trp, Lys) > (Tyr, Gln) > (Ser, Thr)
6. Hydrophobic-hydrophilic dual nature
Known Biochemical Rules (cont.)

7. Partial negative charge
8. Ordered water molecules and ions affect ligand specificity
9. High sugar interface propensity (Trp, Tyr, Phe, His, Asn, Asp, Gln, Glu, Arg, Met)
10. Val/Ile presence (galactin, ricin, lectin)
11. A co-occurrence between Phe/Tyr and Asn/Asp (lectin)
12. Conserved positions for Asn, Asp, Gly and Phe/Tyr (lectin)
13. Spatial disposition is not conserved \textit{per se}, but is conserved with respect to the docking position (galactose)
Top-Down Rules Using Aleph

1. It contains a TRP residue and a GLU with an OE1 atom that is 8.53 Å away from an Oxygen atom with a negative partial charge (GLU, ASP, Sulfate, Phosphate, C-terminus Oxygen).
   [Pos cover = 22, Neg cover = 4]

2. It contains a TRP, PHE or TYR residue, an ASP and an ASN. ASP and an ASN’s OD1 atoms are 5.24 Å apart.
   [Pos cover = 21, Neg cover = 3]

3. It contains a VAL or ILE residue, an ASP and an ASN. ASP and ASN’s OD1 atoms are 3.41 Å apart.
   [Pos cover = 15, Neg cover = 0]
Top-Down Rules Using Aleph (cont.)

4. It contains a hydrophilic non-hydrogen bonding Nitrogen atom (PRO, ARG) with a distance of 7.95 Å away from a HIS’s ND1 atom, and 9.60 Å away from a VAL or ILE’s CG1 atom.
[Pos cover = 10, Neg cover = 0]

5. It has a hydrophobic CD2 atom (LEU, PHE, TYR, TRP, HIS), a PRO, and two hydrophilic OE1 atoms (GLU, GLN) 11.89 Å apart.
[Pos cover = 11, Neg cover = 2]

6. It contains an ASP residue B, two identical atoms Q and X, and a hydrophilic hydrogen-bonding atom K 8.29 Å apart from X. Atoms K, Q and X have the same charge. B’s OD1 atom share the same Y-coordinate with K and the same Z-coordinate with Q.
[Pos cover = 8, Neg cover = 0]
7 It contains a SER residue, and two NE2 atoms (GLN, HIS) 3.88 Å apart.
[Pos cover = 8, Neg cover = 2]

8 It contains an ASN residue and a PHE, TYR or HIS residue, whose CE1 atom is 7.07 Å away from a Calcium ion.
[Pos cover = 5, Neg cover = 0]

9 It contains a LYS or ARG, a PHE, TYR or ARG, a TRP, and a Sulfate or a Phosphate ion.
[Pos cover = 3, Neg cover = 0]
Top-Down Rules Insight

- Aromatics (Trp, Tyr, Phe): 1, 2, 5, 8, 9
- Histidine: 4, 5, 7, 8
- Planar-polar (Asn, Asp, Gln, Glu, Arg): 1 – 9
- High propensity residues: 1 – 9
- Negatively charged atoms/residues: 1, 2, 3, 5, 6
- Dual hydrophobic/hydrophilic: 5
- Presence of ions: 1, 8, 9
- Val/Ile presence: 3
- Phe/Tyr and Asn/Asp co-occurrence: 2, 8
- Trp and Glu co-occurrence: 1
1. It contains an ASP residue whose CG atom is 5.4 Å away from the binding center, and two different ASN residues. [Pos cover = 37, Neg cover = 4]

2. It contains an ASN residue whose N atom is 8.2 Å away from the binding center, and an ASN residue whose N and ND2 atoms are 4.1 Å apart and whose N and O atoms are 3.6 Å apart. [Pos cover = 30, Neg cover = 0]

3. It contains an ASN whose N and C atoms are 2.4 Å apart, and a GLU whose CB and CG atoms are 8.0 Å and 6.9 Å away from the binding center, respectively. [Pos cover = 24, Neg cover = 0]
It contains CYS and LEU residues, and an ASP whose N and OD2 atoms are 4.6 Å apart, and whose C atom is 7.6 Å away from the binding center.
[Pos cover = 18, Neg cover = 0]

It contains a TRP whose CB atom is 7.1 Å away from the binding center, and whose N and CD1 atoms are 4.0 Å apart.
[Pos cover = 14, Neg cover = 0]

It contains a TYR whose CB and OH atoms are 5.6 Å apart, a HIS whose ND1 atom is 8.9 Å away from the binding center, and a TYR whose O atom is 9.8 Å away from the binding center.
[Pos cover = 6, Neg cover = 0]
Bottom-Up Rules Insight

- Aromatics (Trp, Tyr, Phe): 5, 6
- Histidine: 6
- Aromatic sandwich: 6
- Negatively charged atoms/residues: 1, 3, 4
- Planar-polar (Asn, Asp, Gln, Glu, Arg): 1, 2, 3, 4
- Hydrogen-bonding atoms interface: 1
- Conserved positions for Asn, Asp, Tyr: 1, 2, 4, 6
- Conformation conserved with respect to the ligand: 1 – 6
- Dependency over Leu and Cys: 4
Detecting Stereochemical Dispositions
### Sugar Binding Site Classifiers Error Rates

<table>
<thead>
<tr>
<th>Program</th>
<th>Error (%)</th>
<th>Method and Data set</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General sugar binding sites classifiers</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aleph hexose predictor</td>
<td>32.50</td>
<td>10-folds cross-validation, 80 hexose and 80 non-hexose or non-binding sites</td>
</tr>
<tr>
<td>ProGolem hexose predictor</td>
<td>16.70</td>
<td>10-folds cross-validation, 80 hexose and 80 non-hexose or non-binding sites</td>
</tr>
<tr>
<td>Shionyu-Mitsuyama et al.</td>
<td>31.00</td>
<td>Test set, 61 polysaccharide binding sites</td>
</tr>
<tr>
<td>Taroni et al.</td>
<td>35.00</td>
<td>Test set, 40 carbohydrate binding sites</td>
</tr>
<tr>
<td>Malik and Ahmad</td>
<td>39.00</td>
<td>Leave-one-out, 40 carbohydrate and 116 non-carbohydrate binding sites</td>
</tr>
<tr>
<td><strong>Specific sugar binding sites classifiers</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COTRAN</td>
<td>5.09</td>
<td>Overall performance over 6-folds, totaling 106 galactose and 660 non-galactose binding sites</td>
</tr>
<tr>
<td>SVM Nassif et al.</td>
<td>8.11</td>
<td>Leave-one-out, 29 glucose and 35 non-glucose or non-binding sites</td>
</tr>
</tbody>
</table>
Hexose Binding Rules Summary

- First hexose binding rules empirical generation and validation
- Recovered most of known rules, potential for discovery


Create \( j \) bootstrap datasets (select \( n \) with replacement)

Out-of-bag (OOB): \( \approx \frac{1}{3} \) of items not included

Grow a decision tree over each dataset
  - At each tree node, select \( q \) features randomly
  - Split node according to best split among the \( q \) features
  - Each tree remains unpruned (low-bias)

Let the tree classify its own OOB data

Compute the number of correctly classified samples

Permute the values of feature \( k \) in the OOB

Classify modified OOB, compute classification difference

**Feature Importance Score:** Resulting accuracy decrease
Hexose dataset:
- 160 instances
- 152 unique proteins
- 122 CATH superfamilies

Definition

**Sensitivity**: Ability to detect true positives ($TP/P$)

**Specificity**: Ability to reject true negatives ($TN/N$)
# Atomic Chemical Properties

<table>
<thead>
<tr>
<th>PDB atom symbol</th>
<th>Residues</th>
<th>Partial Charge</th>
<th>Hydrophobicity</th>
<th>Hydrogen Bonding</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Amino acid oxygen atoms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>All amino acids</td>
<td>0</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>OXT</td>
<td>All amino acids</td>
<td>-ve</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>OE1, OE2, OD1, OD2</td>
<td>GLU, ASP</td>
<td>-ve</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>OE1, OD1, OG, OG1, OH</td>
<td>GLN, ASN, SER, THR, TYR</td>
<td>0</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td><strong>Amino acid carbon atoms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>All amino acids</td>
<td>0</td>
<td>HNEUT</td>
<td>NHB</td>
</tr>
<tr>
<td>CA</td>
<td>All amino acids</td>
<td>0</td>
<td>HNEUT</td>
<td>NHB</td>
</tr>
<tr>
<td>CB, CG, CD, CE, CG2, CZ</td>
<td>ALA, SER, THR, CYS, ASP, ASN, GLU, GLN, ARG, LYS, PRO</td>
<td>0</td>
<td>HNEUT</td>
<td>NHB</td>
</tr>
<tr>
<td>CB, CD1, CD2, CE1, CE2, CE3, CG, CG1, CG2, CE, CH2, CZ, CZ2, CZ3</td>
<td>LEU, VAL, ILE, MET, PHE, TYR, TRP, HIS</td>
<td>0</td>
<td>HPHOB</td>
<td>NHB</td>
</tr>
</tbody>
</table>
### Appendix

#### Atomic Chemical Properties (cont.)

<table>
<thead>
<tr>
<th>PDB atom symbol</th>
<th>Residues</th>
<th>Partial Charge</th>
<th>Hydrophobicity</th>
<th>Hydrogen Bonding</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Amino acid nitrogen atoms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>All amino acids except PRO</td>
<td>0</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>N</td>
<td>PRO</td>
<td>0</td>
<td>HPHIL</td>
<td>NHB</td>
</tr>
<tr>
<td>NE2, ND1, ND2</td>
<td>GLN, ASN, HIS</td>
<td>0</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>NZ, NE, NH1, NH2</td>
<td>LYS, ARG</td>
<td>+ve</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>NE1</td>
<td>TRP</td>
<td>0</td>
<td>HNEUT</td>
<td>HB</td>
</tr>
<tr>
<td><strong>Amino acid sulfur atoms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SG</td>
<td>CYS</td>
<td>0</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>SD</td>
<td>MET</td>
<td>0</td>
<td>HNEUT</td>
<td>HB</td>
</tr>
<tr>
<td><strong>Water and ions atoms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>HOH</td>
<td>0</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>O1, O2, O3, O4</td>
<td>SO4, 2HP</td>
<td>-ve</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>CA, MG, ZN, MN, FE</td>
<td>CA, MG, ZN, MN, FE</td>
<td>+ve</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
</tbody>
</table>
SVM trained using an exclusively nonbinding sites negative set

<table>
<thead>
<tr>
<th>Property</th>
<th>SVM error</th>
<th>Support Vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charge</td>
<td>5.26%</td>
<td>73.68%</td>
</tr>
<tr>
<td>Hydrogen Bonding</td>
<td>3.51%</td>
<td>61.40%</td>
</tr>
<tr>
<td>Hydrophobicity</td>
<td>5.26%</td>
<td>68.42%</td>
</tr>
</tbody>
</table>
## Baseline Algorithms

<table>
<thead>
<tr>
<th>Fold</th>
<th>$k$NN</th>
<th>BS$k$NN</th>
<th>NB</th>
<th>DT</th>
<th>Pr DT</th>
<th>Per</th>
<th>SC</th>
<th>Aleph</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>25.0</td>
<td>25.0</td>
<td>43.75</td>
<td>31.25</td>
<td>37.5</td>
<td>43.75</td>
<td>31.25</td>
<td>25.0</td>
</tr>
<tr>
<td>1</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>31.25</td>
<td>25.0</td>
<td>43.75</td>
<td>31.25</td>
<td>37.5</td>
</tr>
<tr>
<td>2</td>
<td>18.75</td>
<td>18.75</td>
<td>25.0</td>
<td>12.5</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
</tr>
<tr>
<td>3</td>
<td>18.75</td>
<td>18.75</td>
<td>37.5</td>
<td>6.25</td>
<td>12.5</td>
<td>31.25</td>
<td>12.5</td>
<td>50.0</td>
</tr>
<tr>
<td>4</td>
<td>25.0</td>
<td>37.5</td>
<td>37.5</td>
<td>25.0</td>
<td>37.5</td>
<td>25.0</td>
<td>12.5</td>
<td>31.25</td>
</tr>
<tr>
<td>5</td>
<td>31.25</td>
<td>31.25</td>
<td>37.5</td>
<td>31.25</td>
<td>18.75</td>
<td>37.5</td>
<td>31.25</td>
<td>18.75</td>
</tr>
<tr>
<td>6</td>
<td>31.25</td>
<td>18.75</td>
<td>25.0</td>
<td>37.5</td>
<td>31.25</td>
<td>37.5</td>
<td>25.0</td>
<td>25.0</td>
</tr>
<tr>
<td>7</td>
<td>31.25</td>
<td>25.0</td>
<td>37.5</td>
<td>25.0</td>
<td>31.25</td>
<td>31.25</td>
<td>37.5</td>
<td>43.75</td>
</tr>
<tr>
<td>8</td>
<td>18.75</td>
<td>18.75</td>
<td>31.25</td>
<td>25.0</td>
<td>12.5</td>
<td>31.25</td>
<td>31.25</td>
<td>25.0</td>
</tr>
<tr>
<td>9</td>
<td>31.25</td>
<td>31.25</td>
<td>50.0</td>
<td>50.0</td>
<td>31.25</td>
<td>43.75</td>
<td>25.0</td>
<td>43.75</td>
</tr>
</tbody>
</table>

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>25.63</td>
<td>25.0</td>
<td>35.0</td>
<td>27.5</td>
<td>26.25</td>
<td>35.0</td>
<td>26.25</td>
<td>32.5</td>
</tr>
<tr>
<td>std dev</td>
<td>5.47</td>
<td>6.59</td>
<td>8.44</td>
<td>12.22</td>
<td>9.22</td>
<td>7.34</td>
<td>8.23</td>
<td>10.54</td>
</tr>
<tr>
<td>upper bound</td>
<td>29.54</td>
<td>29.71</td>
<td>41.03</td>
<td>36.23</td>
<td>32.84</td>
<td>40.24</td>
<td>32.13</td>
<td>40.03</td>
</tr>
</tbody>
</table>