Characterizing and Predicting Hexose-Binding Sites

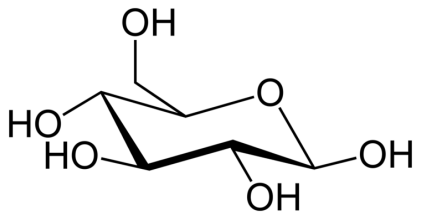
Houssam Nassif
http://pages.cs.wisc.edu/~hous21/

CIBM Seminar
25 January 2011

Outline

Outline

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

Hexose Stereoisomers

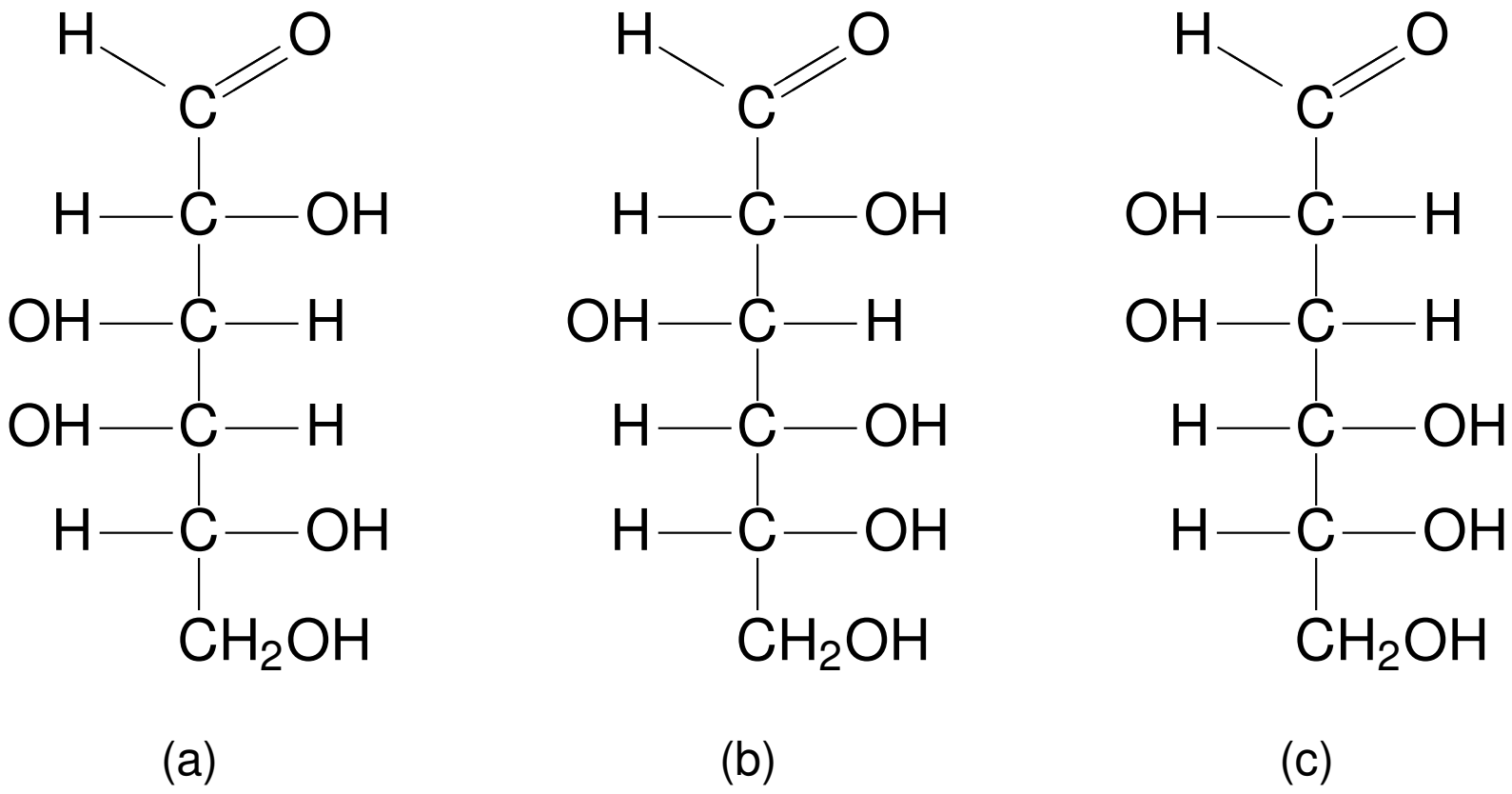


Figure: (a) D-Galactose; (b) D-Glucose; (c) D-Mannose

Hexose Structure

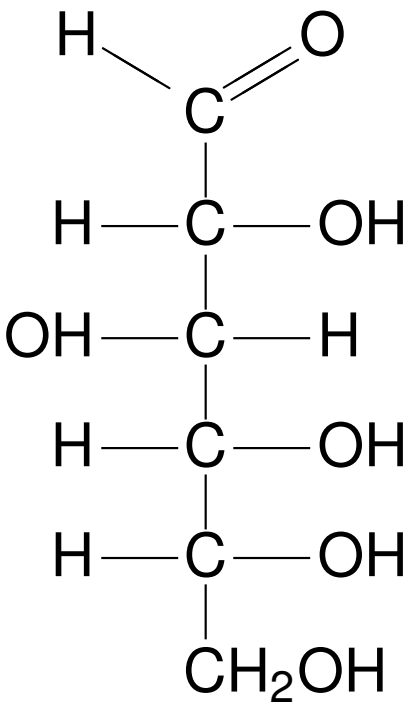
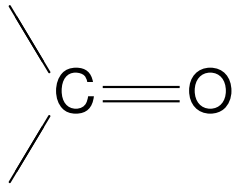


Figure: Glucose

- Contains two functional groups
- Both groups can interact together

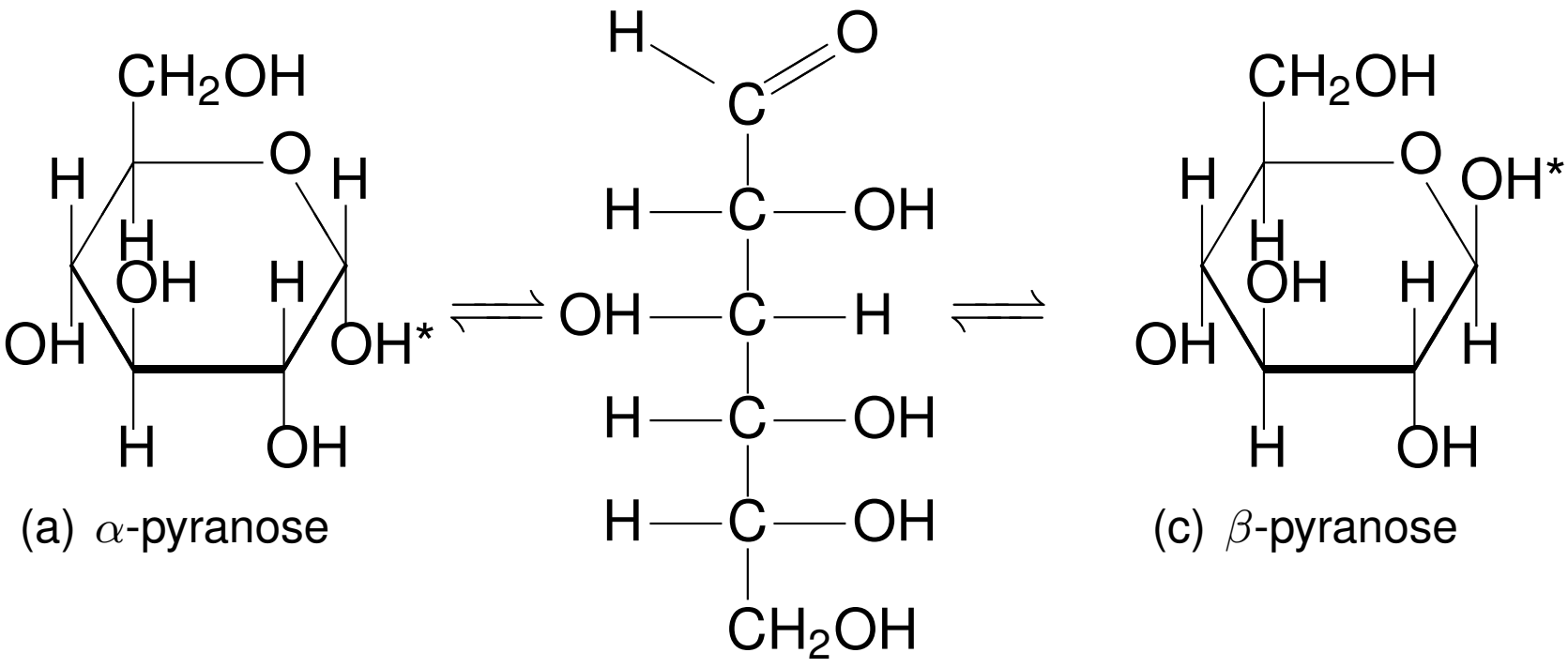


(a) Carbonyl

(b) Hydroxyl

Hexose Cyclization

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



Outline

Covalent Bonds

Figure: Covalent bond

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

Definition

Electronegativity: Measure of atom's attraction for electrons

Covalent Bonds

Figure: Covalent bond

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

Definition

Electronegativity: Measure of atom's attraction for electrons

Covalent Bonds

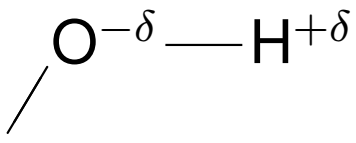


Figure: Covalent polar bond

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

Definition

Electronegativity: Measure of atom's attraction for electrons

Hydrogen Bonds

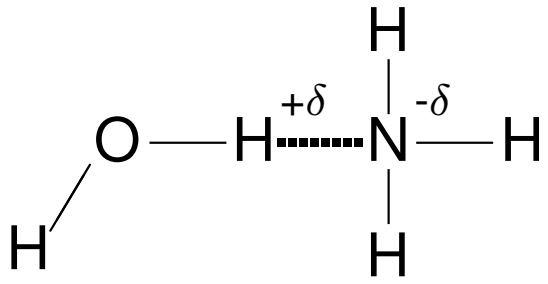


Figure: Hydrogen bond

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

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

Van der Waals and Hydrophobicity

Definition

Van der Waals Forces: Weak electrostatic attraction and repulsion forces

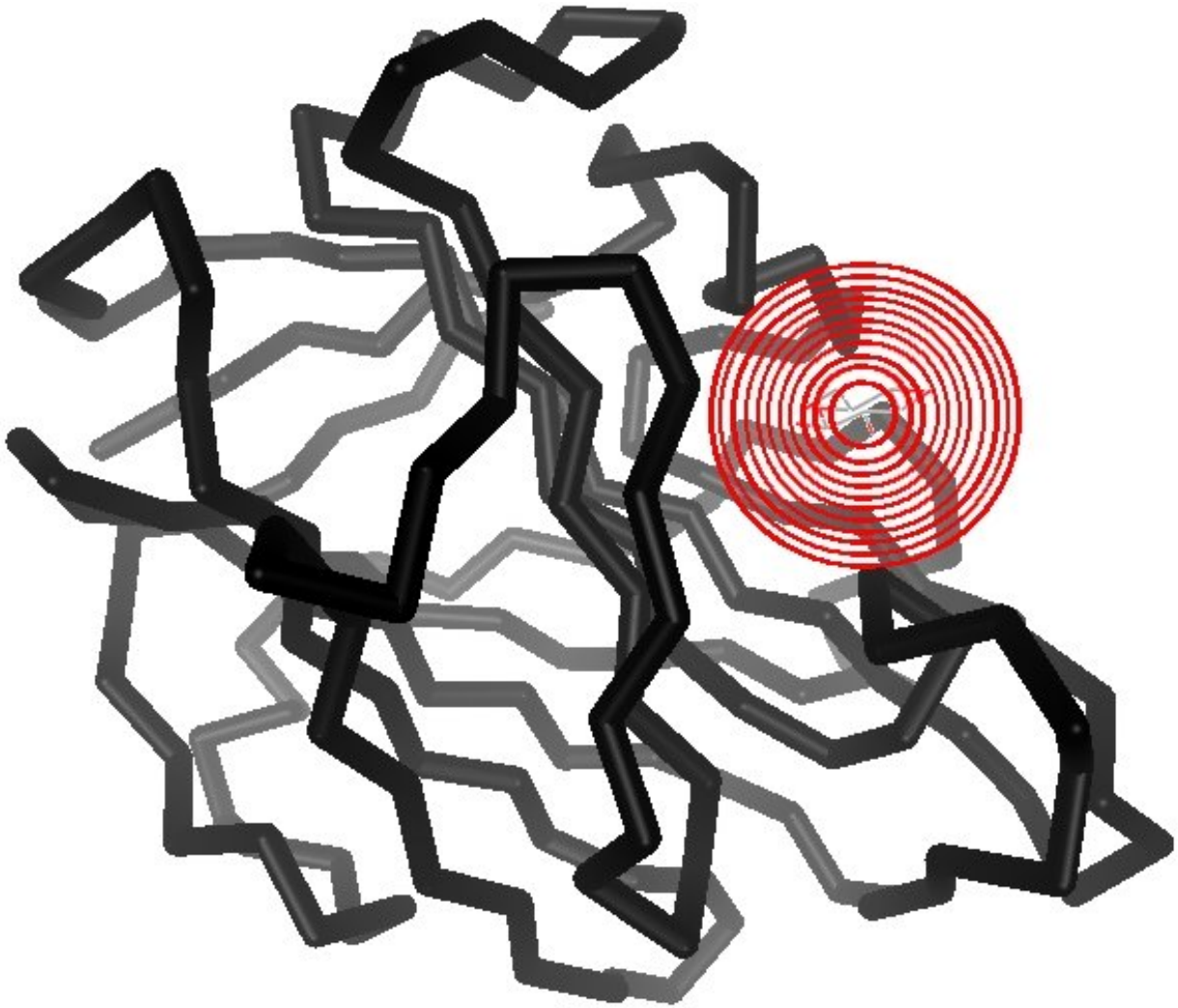
Definition (Hydrophobicity)

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

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

Outline

Binding-Site Representation



Binding-Site Feature Extraction

```
1:  procedure ExtractFeatures(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

Atomic Feature	Values
Charge	Negative, Neutral, Positive
Hydrogen-bonding	Non-hydrogen bonding, Hydrogen-bonding
Hydrophobicity	Hydrophilic, Hydroneutral, Hydrophobic

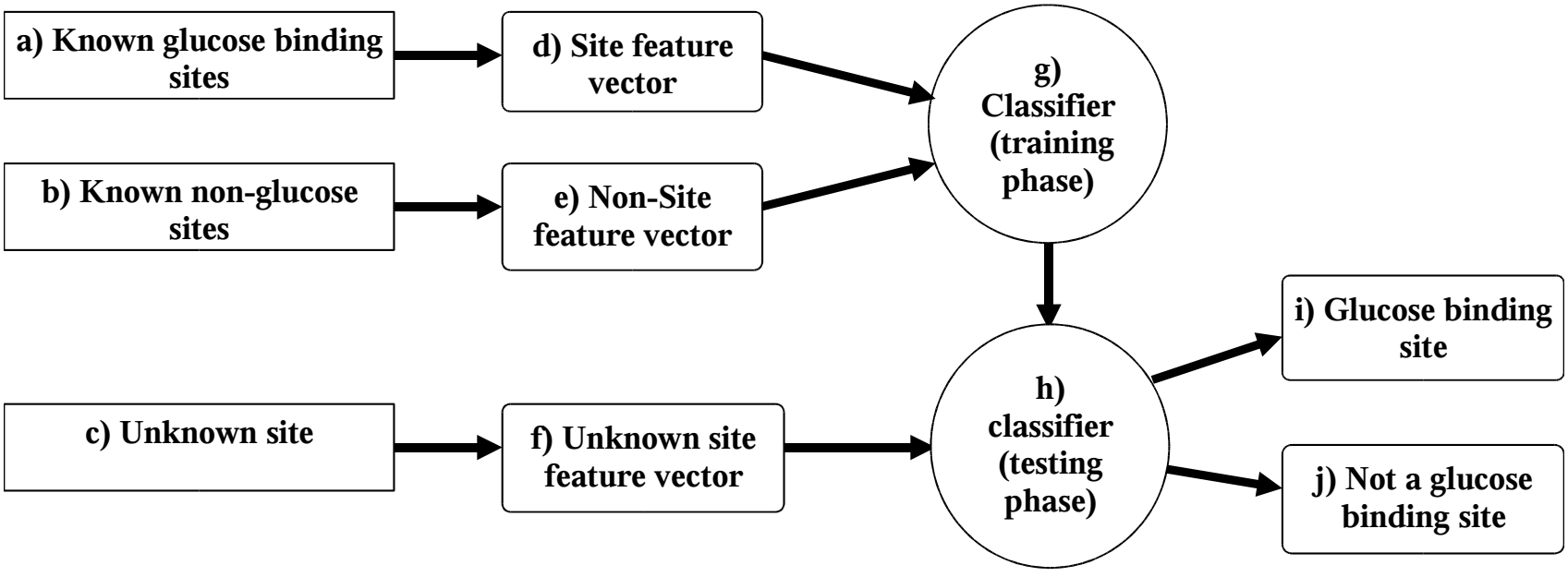
Residue Grouping	Amino Acids
Aromatic	HIS, PHE, TRP, TYR
Aliphatic	ALA, ILE, LEU, MET, VAL
Neutral	ASN, CYS, GLN, GLY, PRO, SER, THR
Acidic	ASP, GLU
Basic	ARG, LYS

Data Mining

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

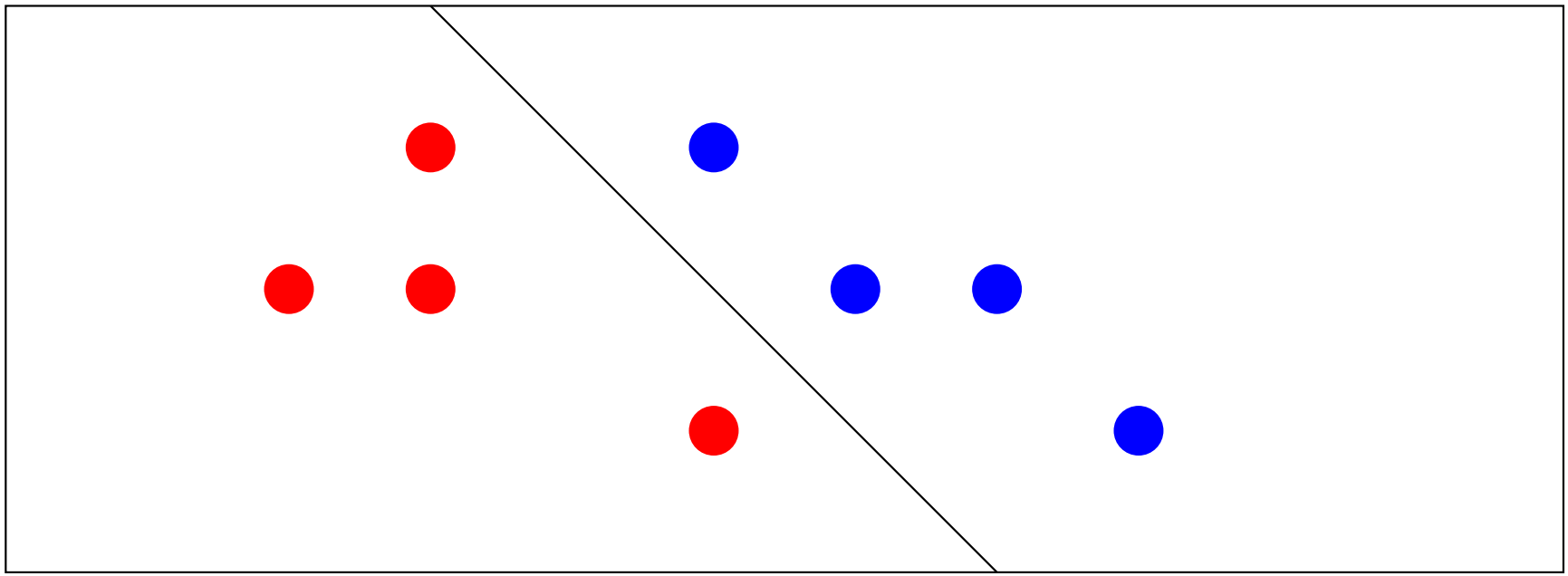
Classifier Outline



Outline

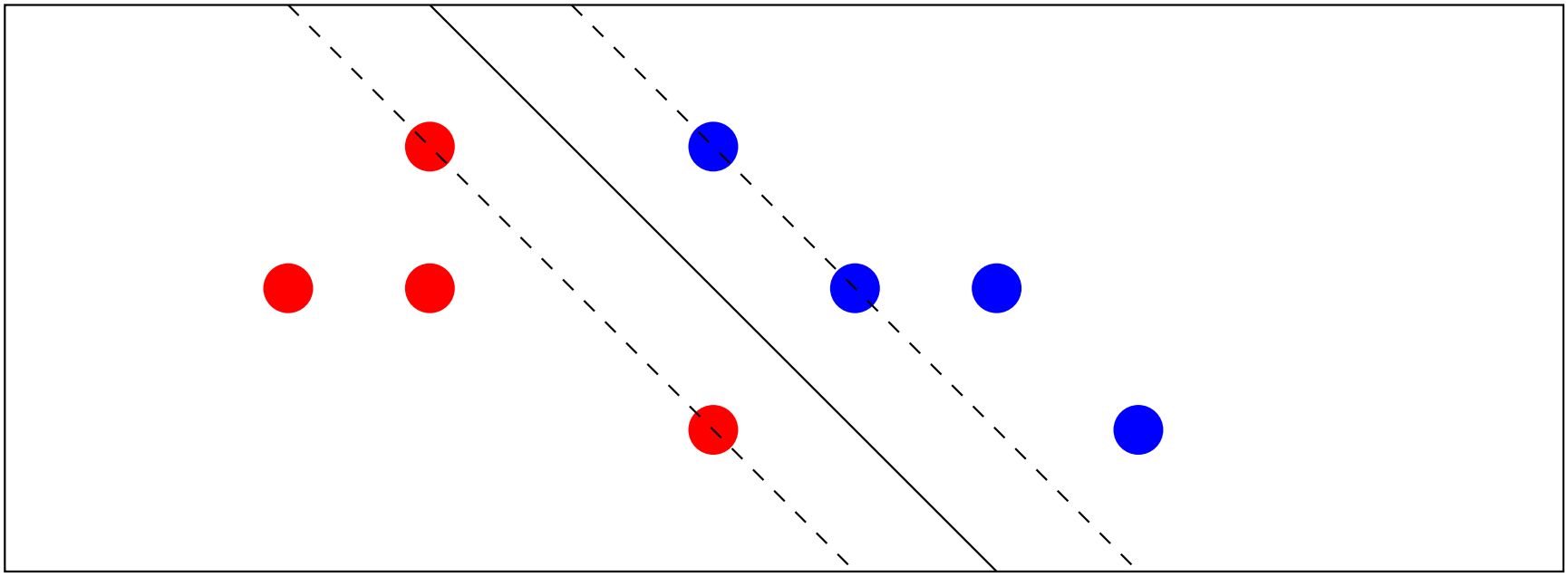
Support Vector Machines (SVM)

- 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



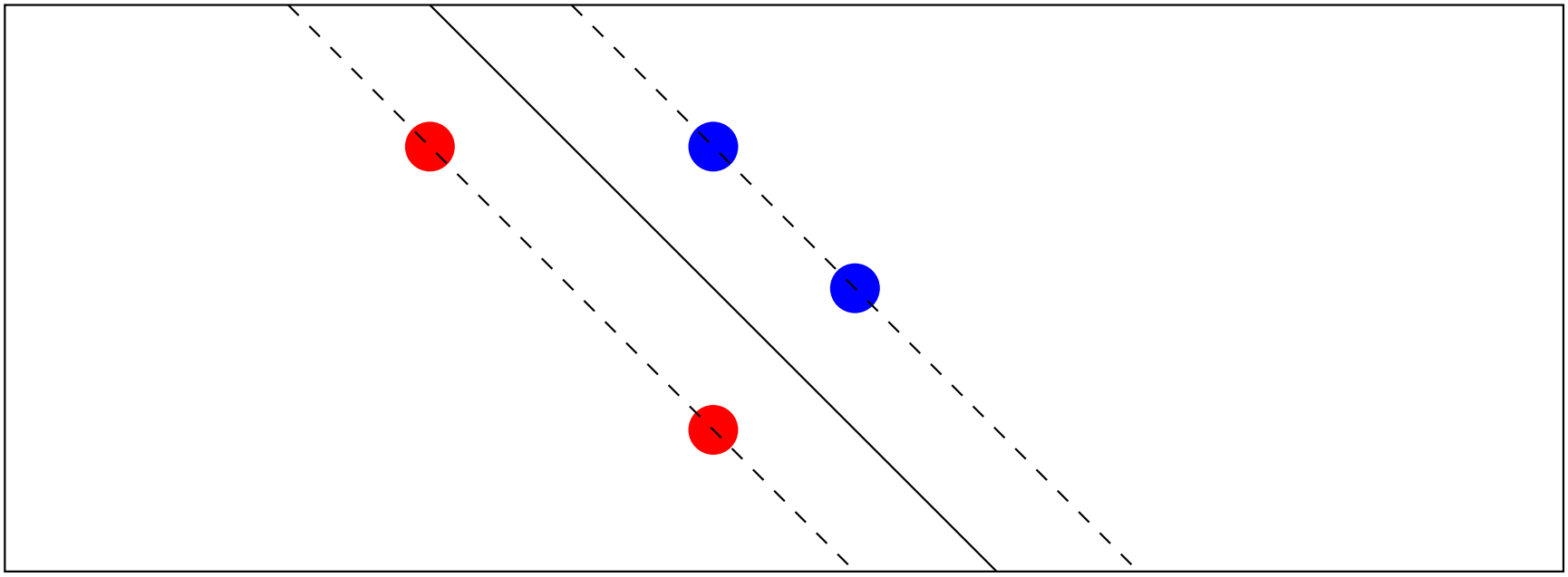
Support Vector Machines (SVM)

- 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 $\Leftrightarrow$ good generalization



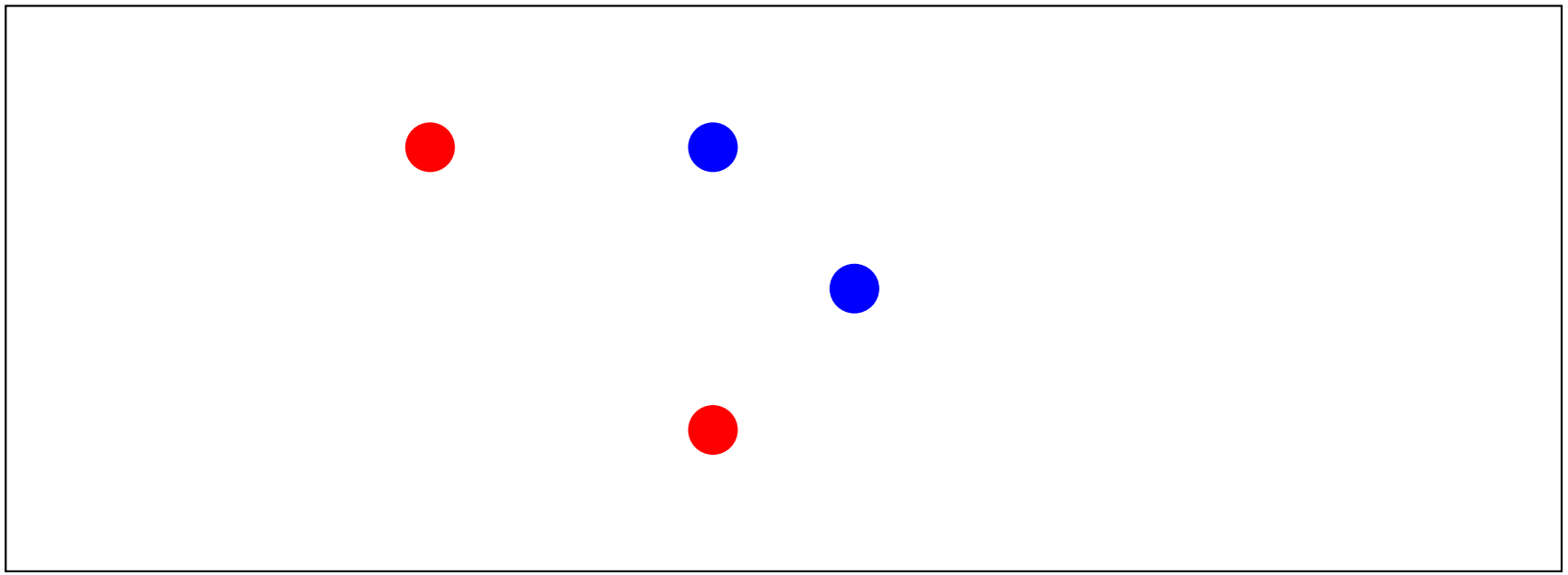
Support Vector Machines (SVM)

- 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 $\Leftrightarrow$ 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 $\Leftrightarrow$ 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

Ligand	Number
Glucose	43
Non-sugar	36
Other sugars	15
Non-binding	17

- 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

Importance of Water and Ions

- Ordered water molecules and ions affect ligand specificity

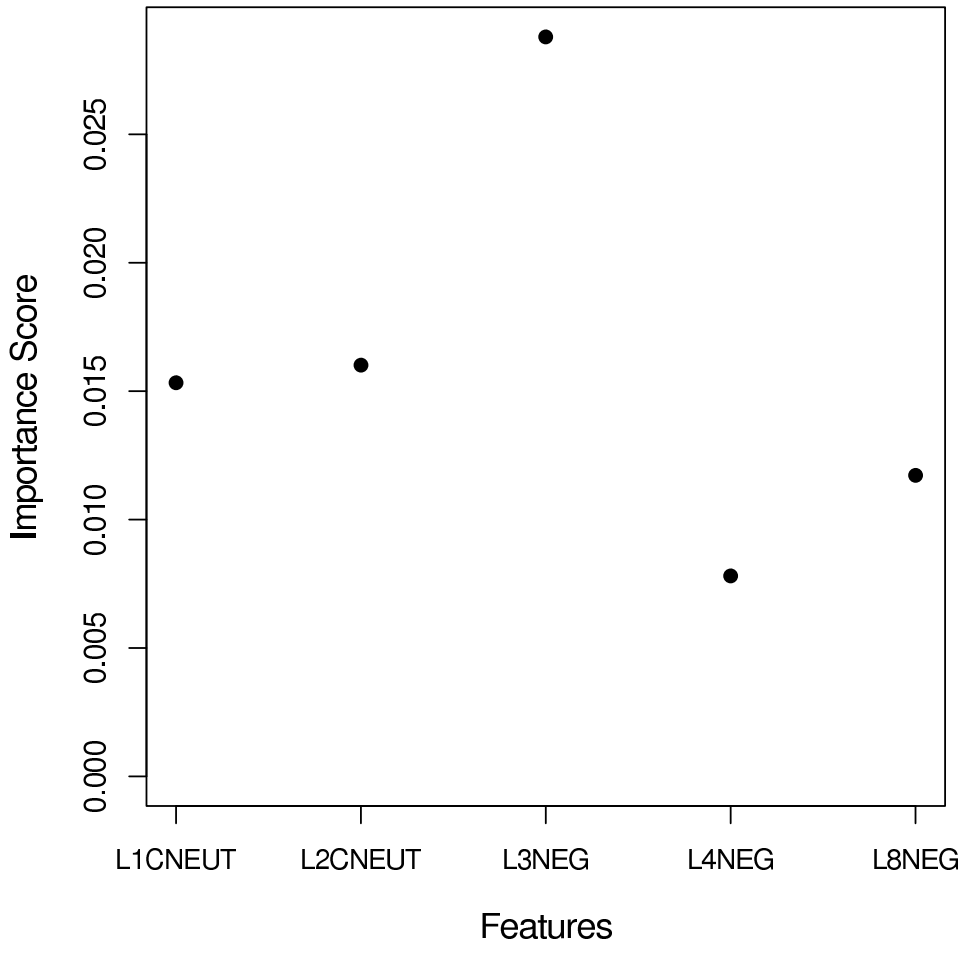
Properties	Whole set error	Subset error*
Include water and ions	18.92%	7.81%
Discard water	18.92%	10.94%
Discard ions	20.27%	7.81%
Discard water and ions	20.27%	12.5%

* Lacks the other sugars binding sites negatives

Properties Feature Selection

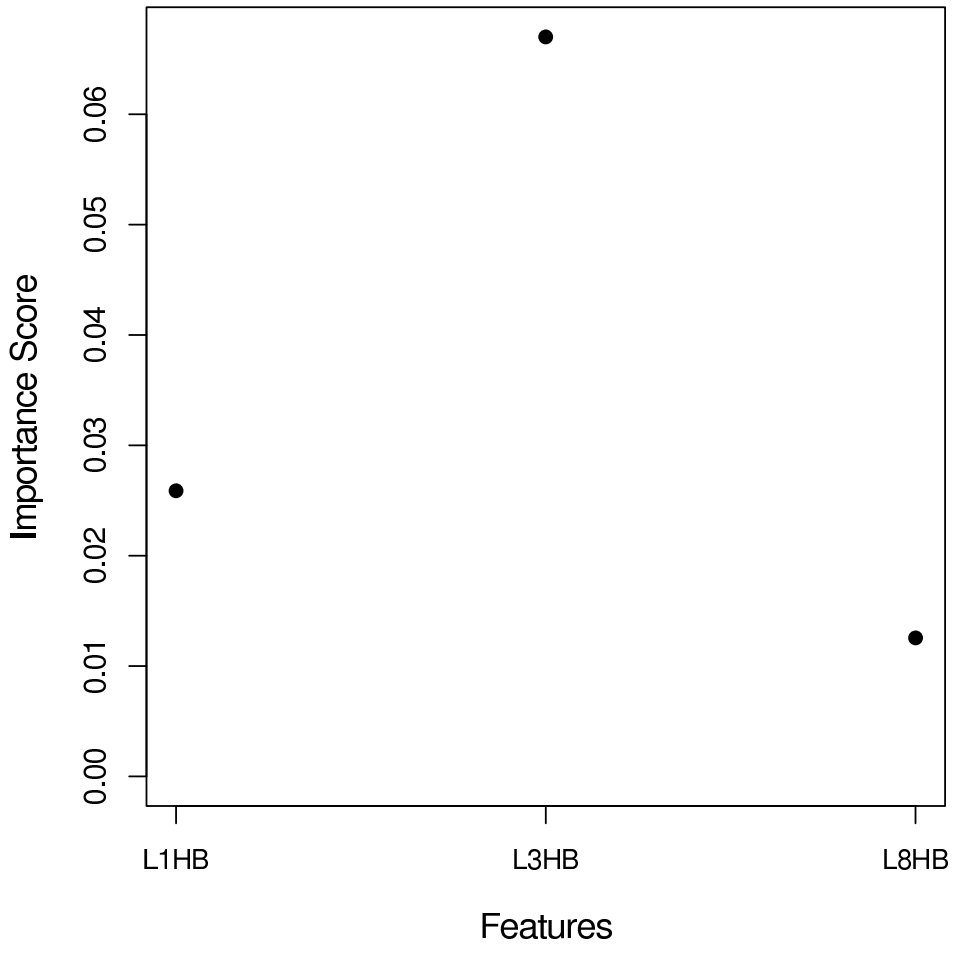
Property	RF	Feature Number	Error (%)	Sensitivity (%)	Specificity (%)	SV (%)
Charge	false	24	24.32	79.31	73.33	77.03
	true	5	14.86	86.21	84.44	44.59
H-Bonding	false	16	17.57	82.76	82.22	41.89
	true	3	14.86	82.76	86.67	47.30
Hydro	false	24	16.22	72.41	91.11	65.57
	true	15	12.16	82.76	91.11	40.54
Residues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

Charge Features



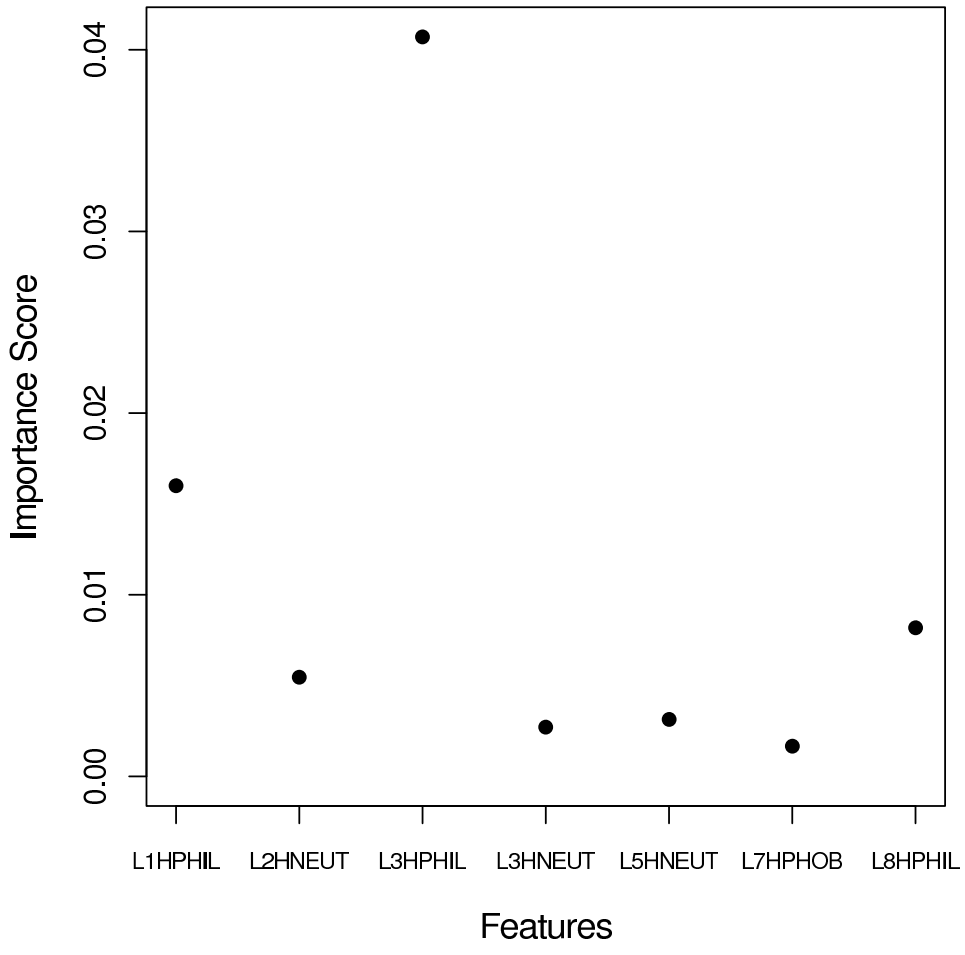
- Negatively charged
- Layer 1: Steric hindrance, non-binding sites
- Layer 2: Small moiety non-sugar binding sites

Hydrogen Bond Features



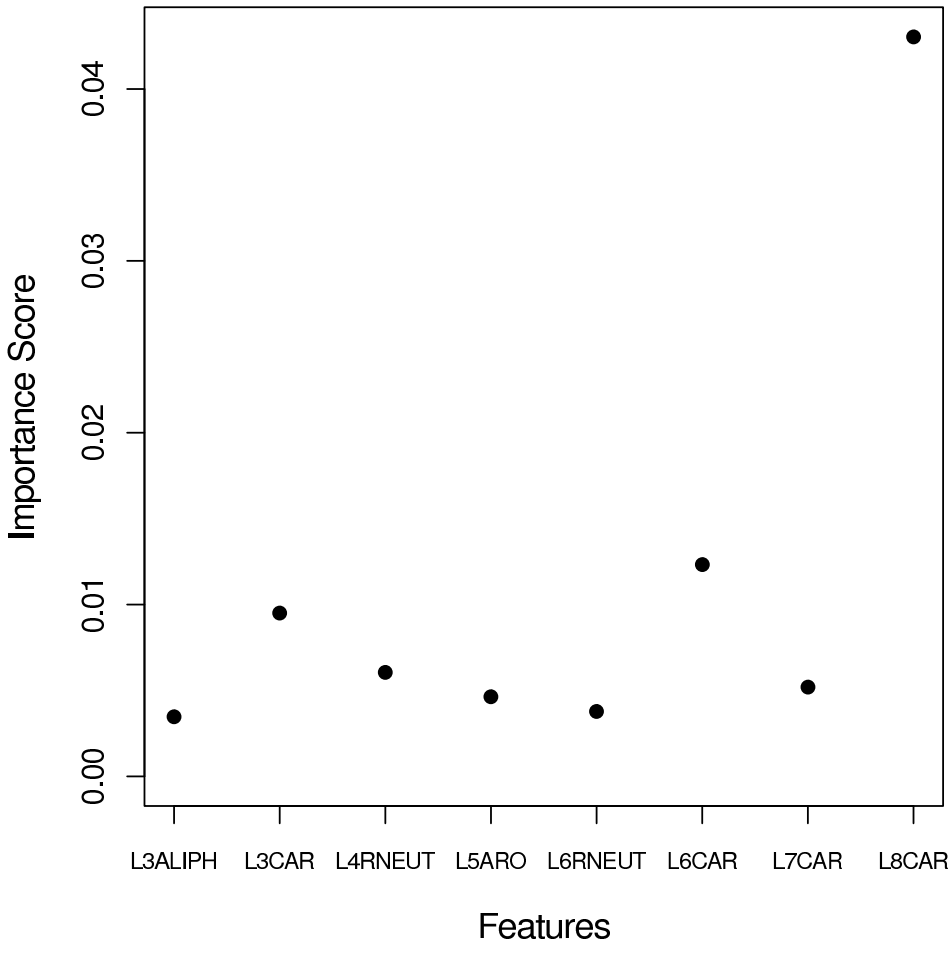
- Importance of layer 3: Hydrogen-bonding atoms at the protein-glucose interface

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

Features	L1	L2	L3	L4	L5	L6	L7	L8
Negative Charge			X				X	X
Neutral Charge	X	X						
Non H-Bonding	X							
H-Bonding	X		X					X
Hydrophilic	X		X					X
Hydroneutral		X	X					
Hydrophobic					X		X	
Neutral Residue				X	X		X	
Acidic Residue			X		X	X	X	X

Glucose Binding Modeling Summary

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

H. Nassif, H. Al-Ali, S. Khuri, and W. Keyrouz.
Prediction of Protein-Glucose Binding Sites Using SVMs.
Proteins, 77(1):121-132, 2009.

Outline

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

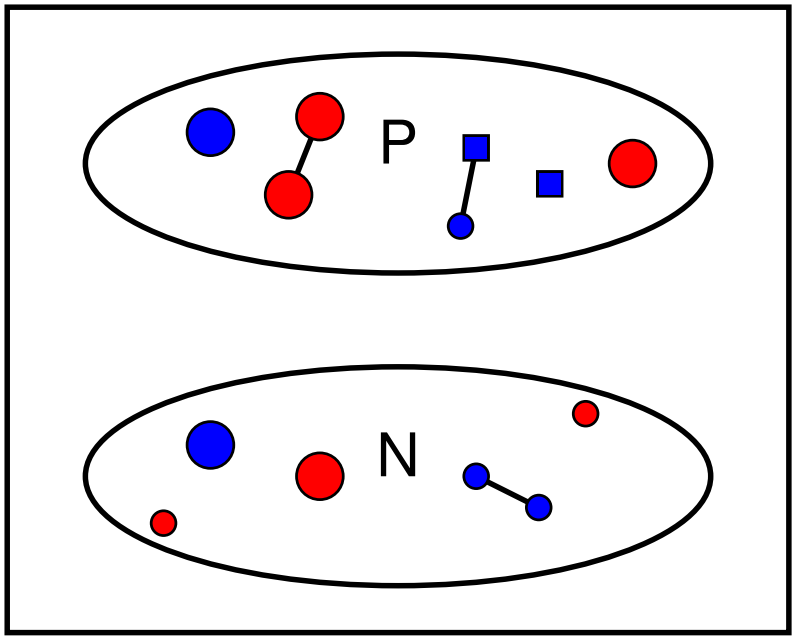
Inductive Logic Programming (ILP)

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

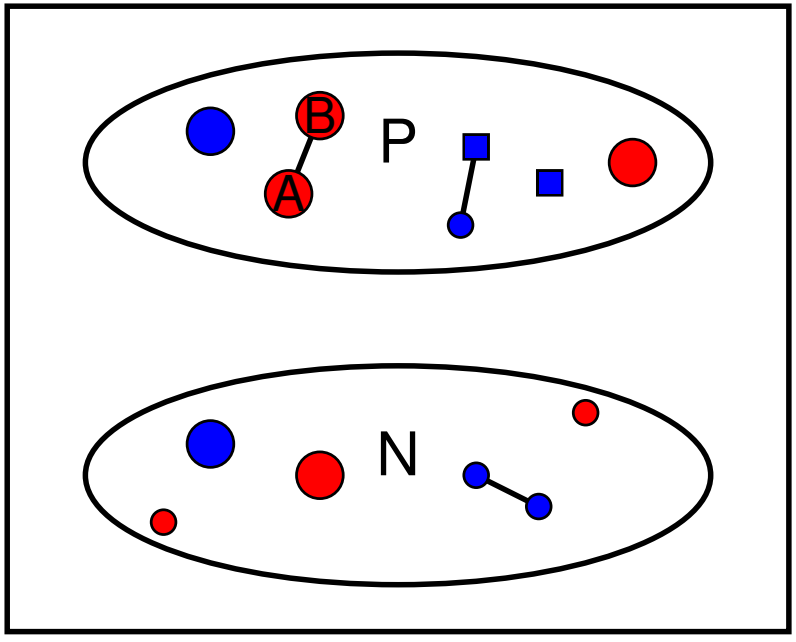


- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(x)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form *theory*

Example

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

ILP Example

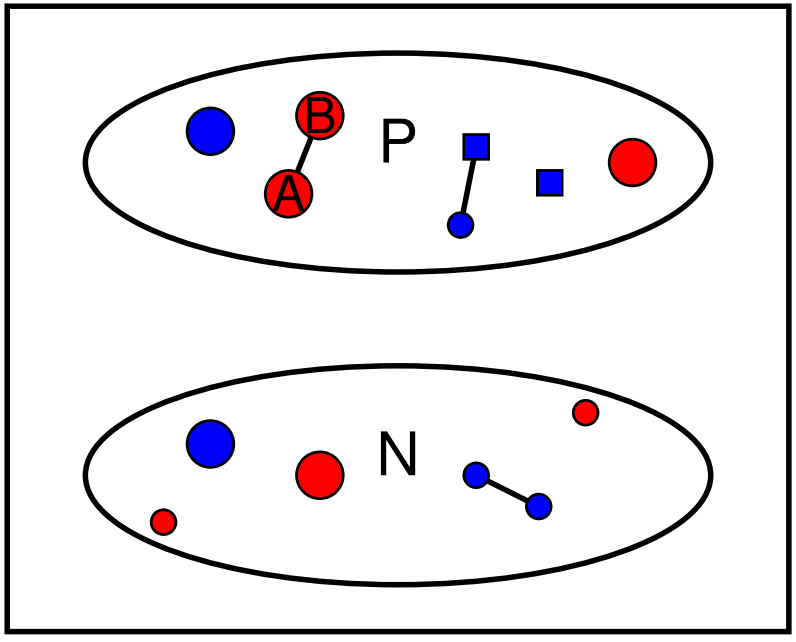


Example

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

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

ILP Example

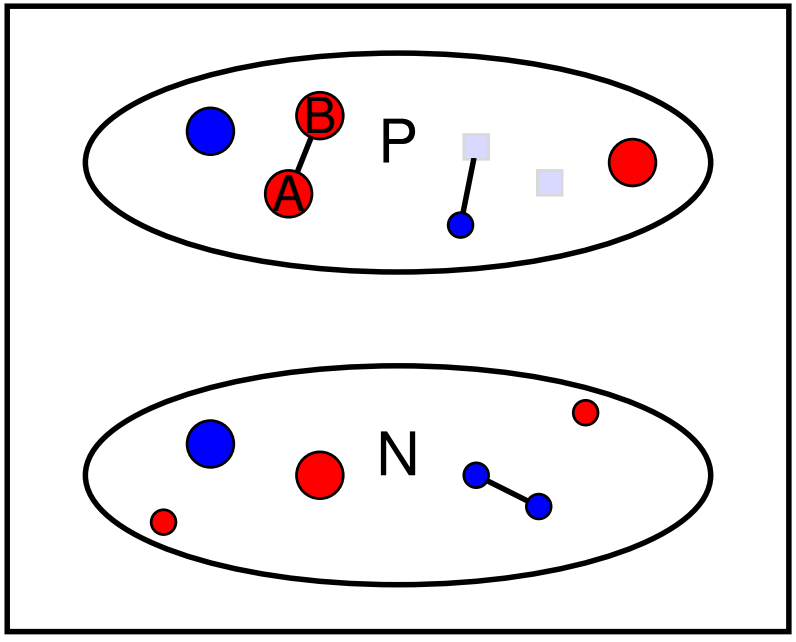


Example

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

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

ILP Example

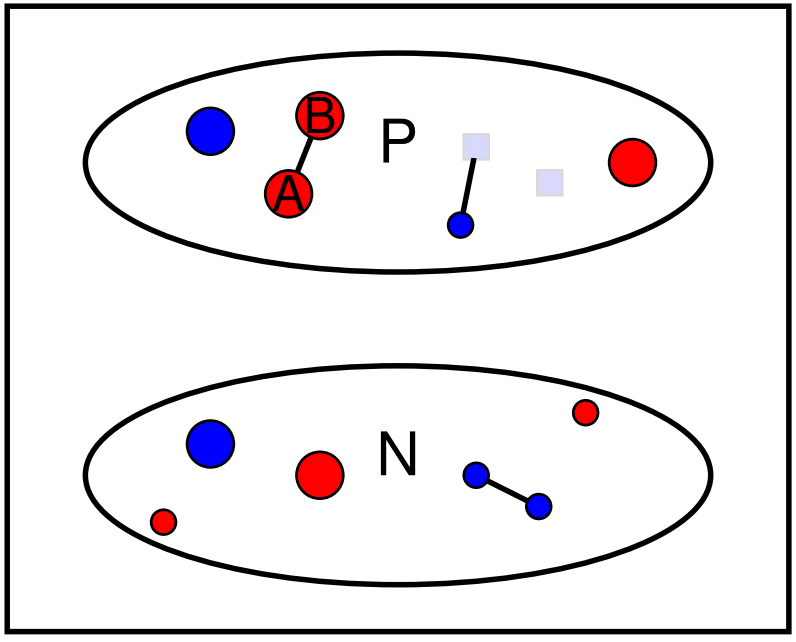


Example

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

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

ILP Example

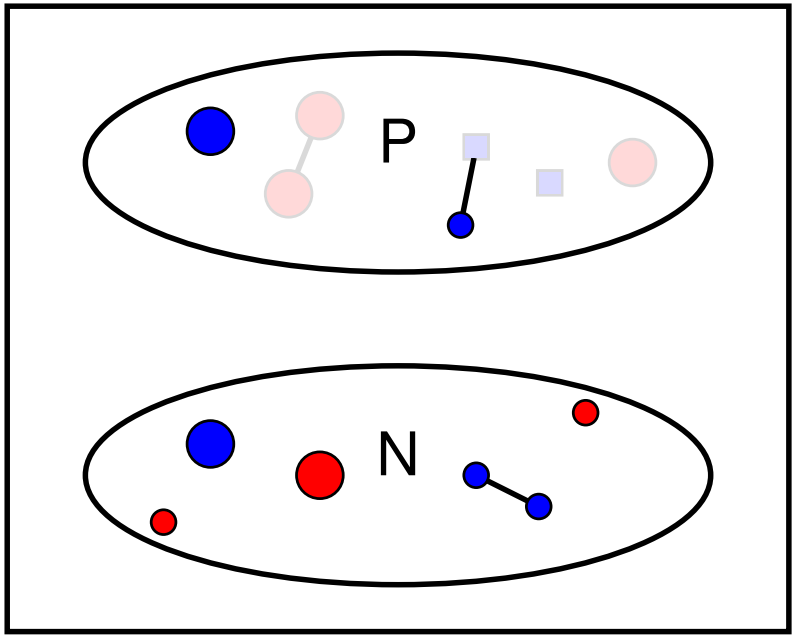


Example

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

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

ILP Example

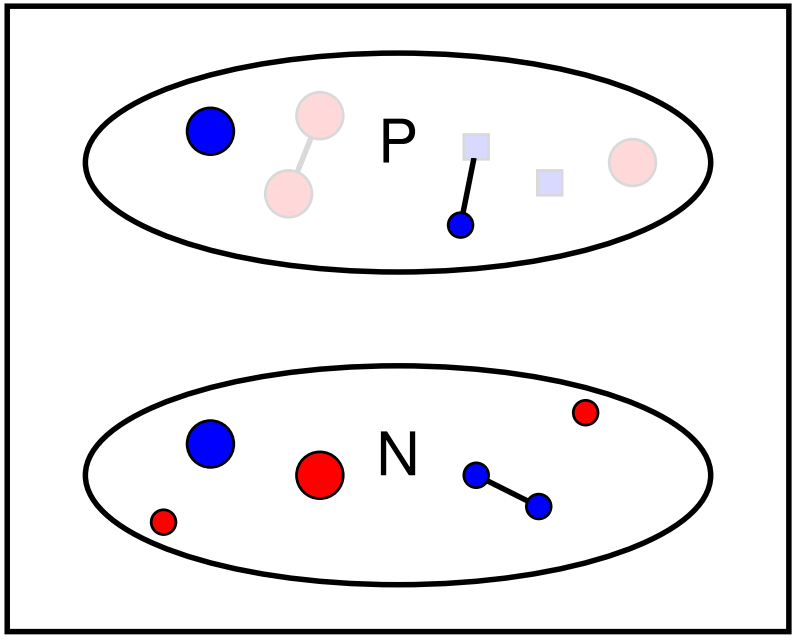


Example

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

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

ILP Example

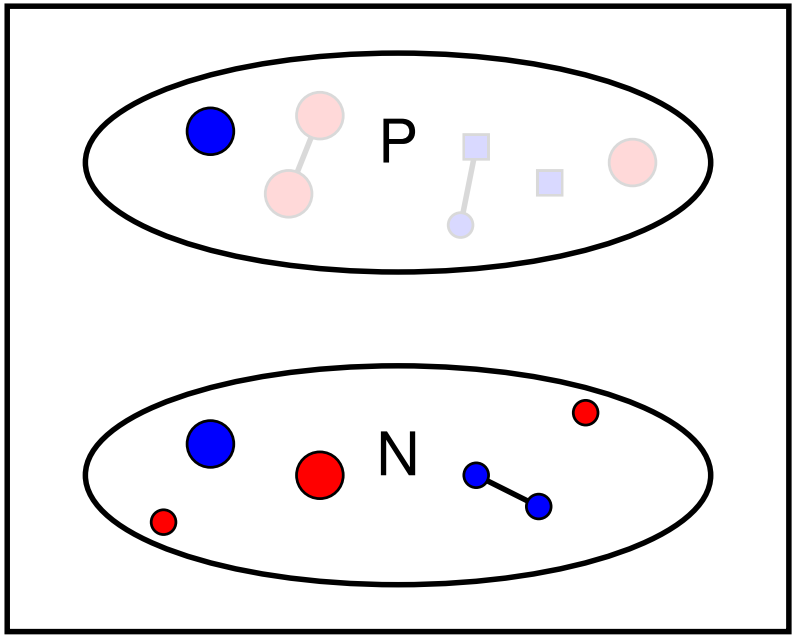


Example

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

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

ILP Example

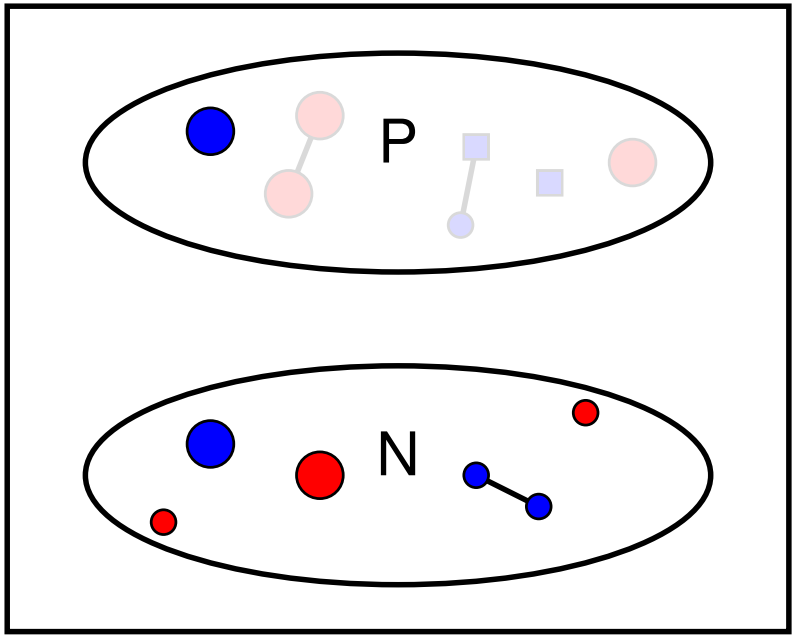


Example

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

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

ILP Example

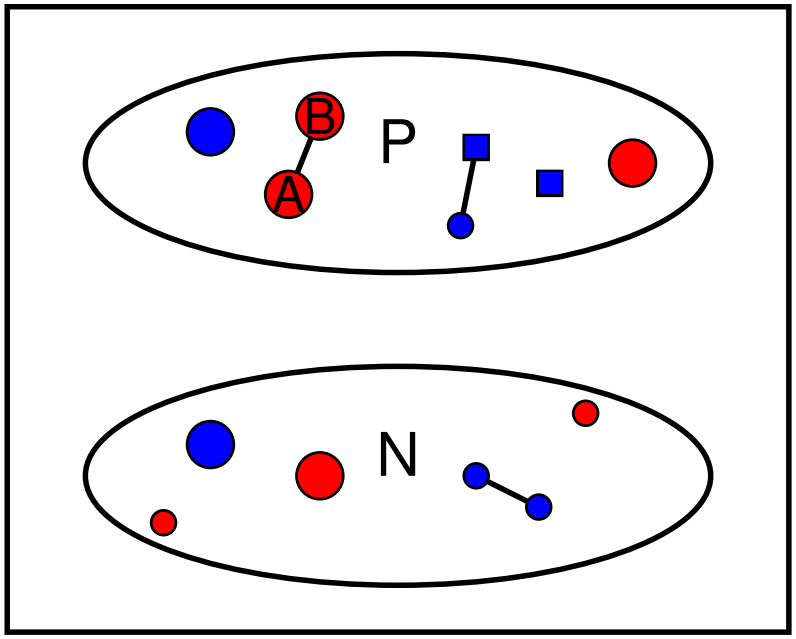


Example

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

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

ILP Search

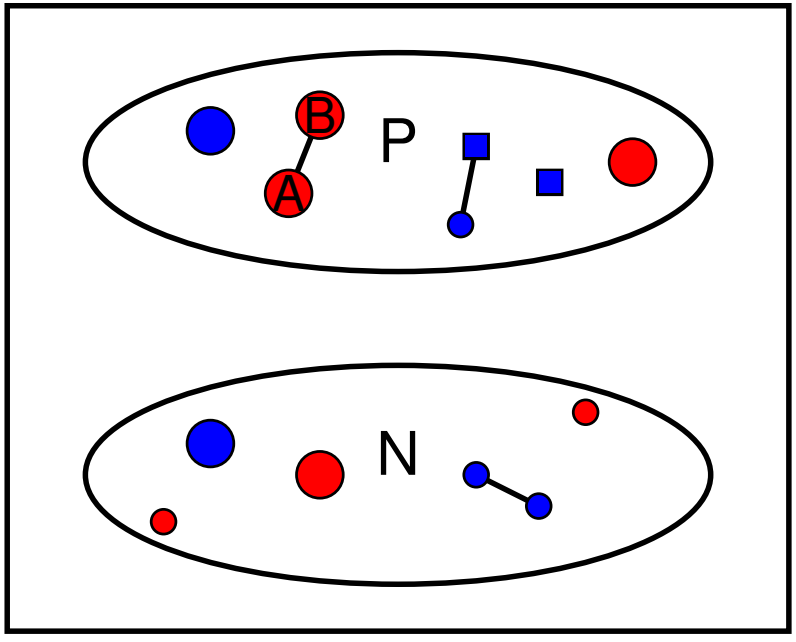


- Pick a positive instance
- Construct the *Bottom Clause*, most specific clause
- *Top-down search*: Start with most general rule, add bottom clause predicates
- *Bottom-up search*: Start with bottom clause, remove predicates

Example (Bottom Clause (A))

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

ILP Search

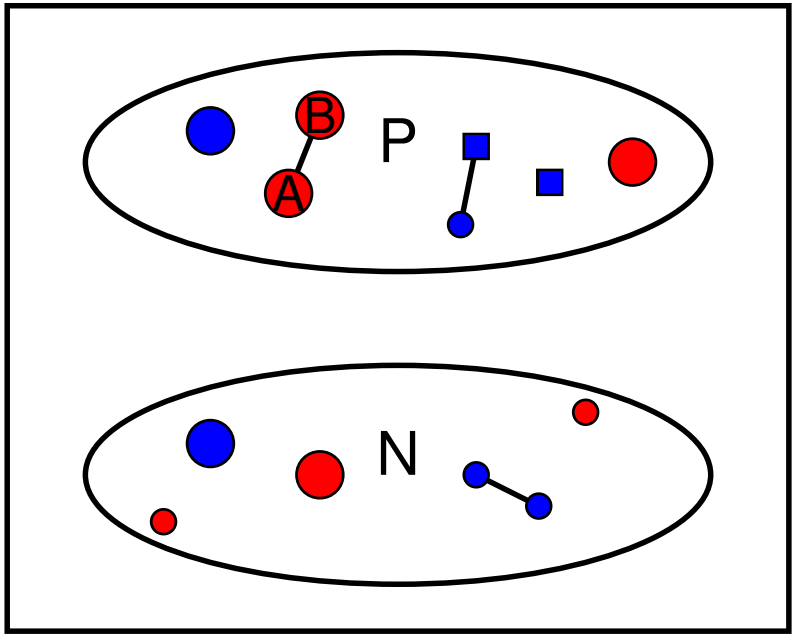


Example (Bottom Clause (A))

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

- 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

ILP Search

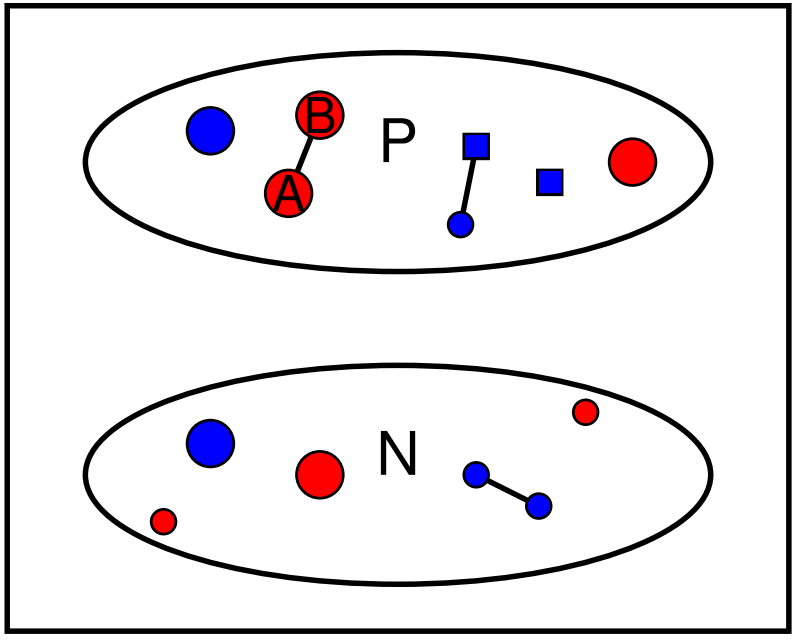


Example (Bottom Clause (A))

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

- 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

ILP Search



Example (Bottom Clause (A))

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

- 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

Experimental Setting

Ligand	Number
Galactose	33
Glucose	35
Mannose	12
Non-sugar	27
Hexose-like	22
Non-binding	31

- 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

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 *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]

Top-Down Rules Using Aleph (cont.)

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

Bottom-Up Rules Using ProGolem

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]

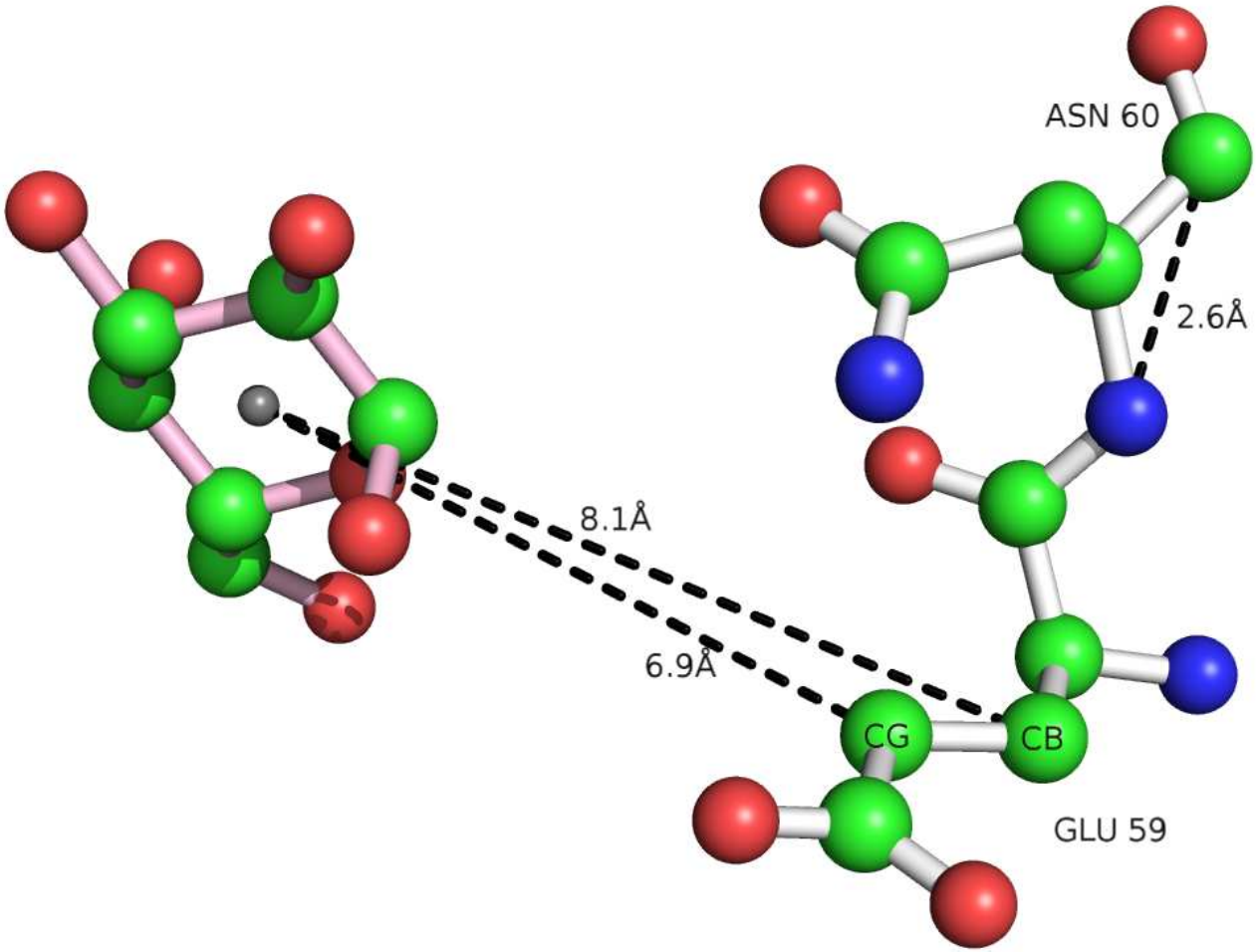
Bottom-Up Rules Using ProGolem (cont.)

4. 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]
5. 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]
6. 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

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

Hexose Binding Rules Summary

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


H. Nassif, H. Al-Ali, S. Khuri, W. Keyrouz and D. Page.
An ILP Approach to Validate Hexose Binding Biochemical Knowledge.
ILP'09, Leuven, Belgium, pp. 149-165, 2009.


J. Santos, H. Nassif, D. Page, S. Muggleton and M. Sternberg.
Automated identification of features of protein-ligand interactions using ILP: Application to hexose binding.
Submitted.

RF Feature Importance Score

- Create j bootstrap datasets (select n with replacement)
- Out-of-bag (OOB): $\approx 1/3$ 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

Data

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

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

Atomic Chemical Properties (cont.)

PDB atom symbol	Residues	Partial Charge	Hydro-phobicity	Hydrogen Bonding
Amino acid nitrogen atoms				
N	All amino acids except PRO	0	HPHIL	HB
N	PRO	0	HPHIL	NHB
NE2, ND1, ND2	GLN, ASN, HIS	0	HPHIL	HB
NZ, NE, NH1, NH2	LYS, ARG	+ve	HPHIL	HB
NE1	TRP	0	HNEUT	HB
Amino acid sulfur atoms				
SG	CYS	0	HPHIL	HB
SD	MET	0	HNEUT	HB
Water and ions atoms				
O	HOH	0	HPHIL	HB
O1, O2, O3, O4	SO4, 2HP	-ve	HPHIL	HB
CA, MG, ZN, MN, FE	CA, MG, ZN, MN, FE	+ve	HPHIL	HB

Nonbinding Sites Negative Set

SVM trained using an exclusively nonbinding sites negative set

Property	SVM error	Support Vectors
Charge	5.26%	73.68%
Hydrogen Bonding	3.51%	61.40%
Hydrophobicity	5.26%	68.42%

Baseline Algorithms

Fold	*k*NN	BS*k*NN	NB	DT	Pr DT	Per	SC	Aleph
0	25.0	25.0	43.75	31.25	37.5	43.75	31.25	25.0
1	25.0	25.0	25.0	31.25	25.0	43.75	31.25	37.5
2	18.75	18.75	25.0	12.5	25.0	25.0	25.0	25.0
3	18.75	18.75	37.5	6.25	12.5	31.25	12.5	50.0
4	25.0	37.5	37.5	25.0	37.5	25.0	12.5	31.25
5	31.25	31.25	37.5	31.25	18.75	37.5	31.25	18.75
6	31.25	18.75	25.0	37.5	31.25	37.5	25.0	25.0
7	31.25	25.0	37.5	25.0	31.25	31.25	37.5	43.75
8	18.75	18.75	31.25	25.0	12.5	31.25	31.25	25.0
9	31.25	31.25	50.0	50.0	31.25	43.75	25.0	43.75
mean	25.63	25.0	35.0	27.5	26.25	35.0	26.25	32.5
std dev	5.47	6.59	8.44	12.22	9.22	7.34	8.23	10.54
lower bound	21.71	20.29	28.97	18.77	19.66	29.76	20.37	24.97
upper bound	29.54	29.71	41.03	36.23	32.84	40.24	32.13	40.03