Application of Decision Tree

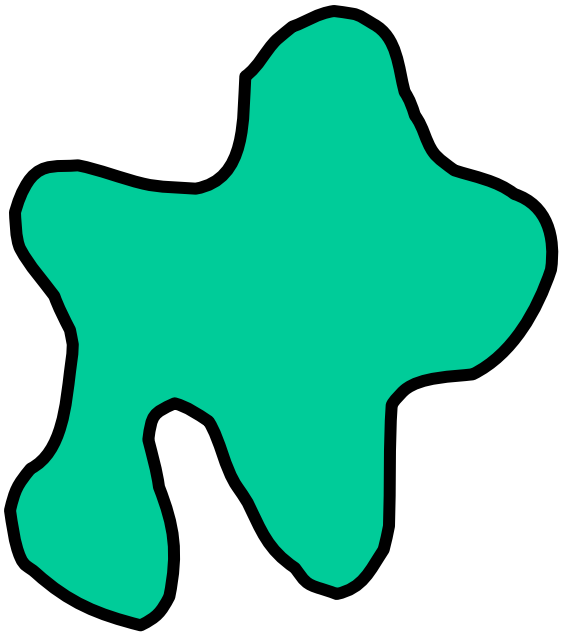
Protein Solvent Accessibility Prediction

What is a molecular surface?

A molecular surface is a closed 3D "manifold".

What's a "manifold"?
Here is a 2D manifold.

A cell, for example, is a 3D manifold. It is continuous, closed, non-intersecting. It has an inside and an outside.

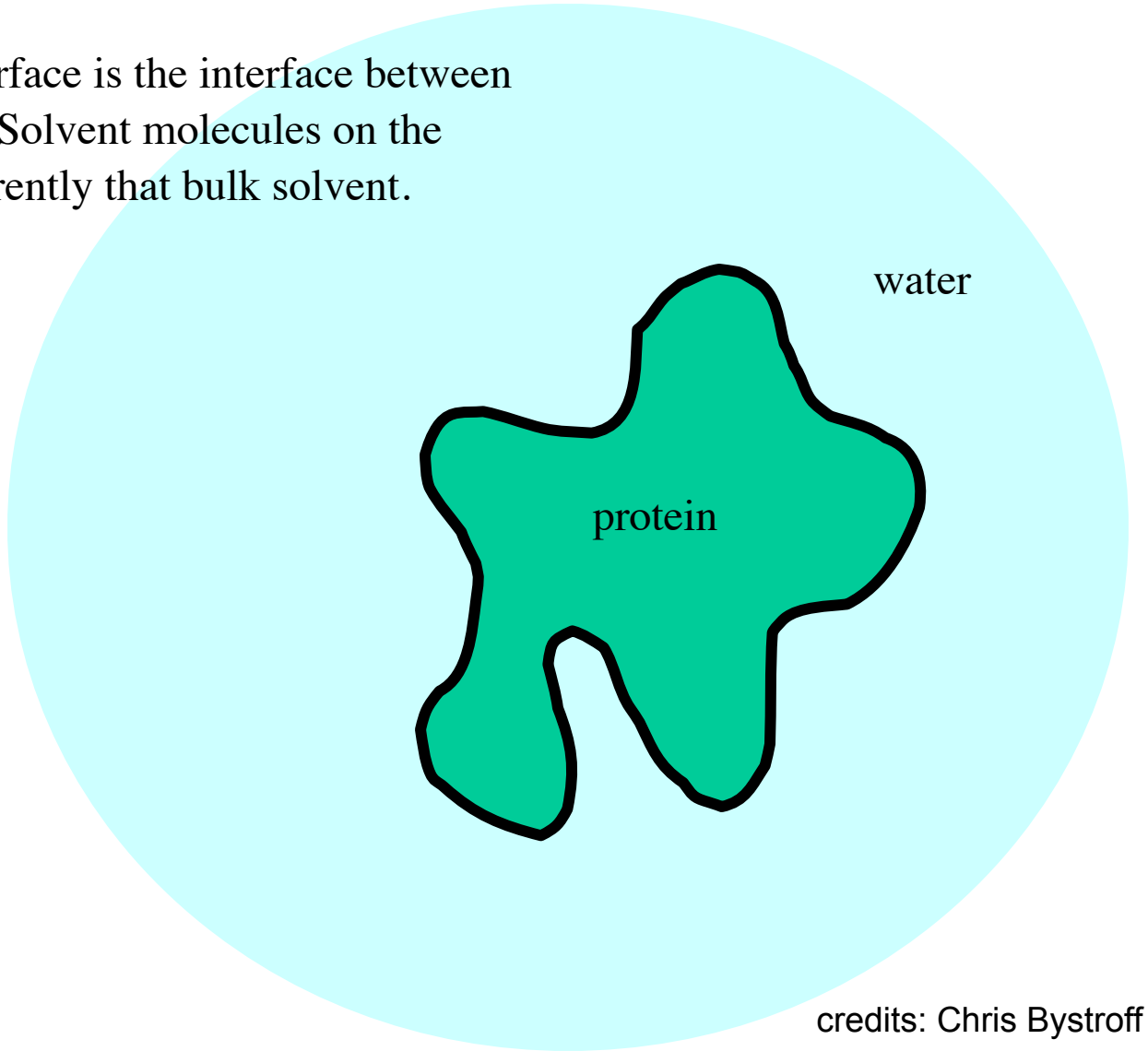


credits: Chris Bystroff

Solvent accessible surface

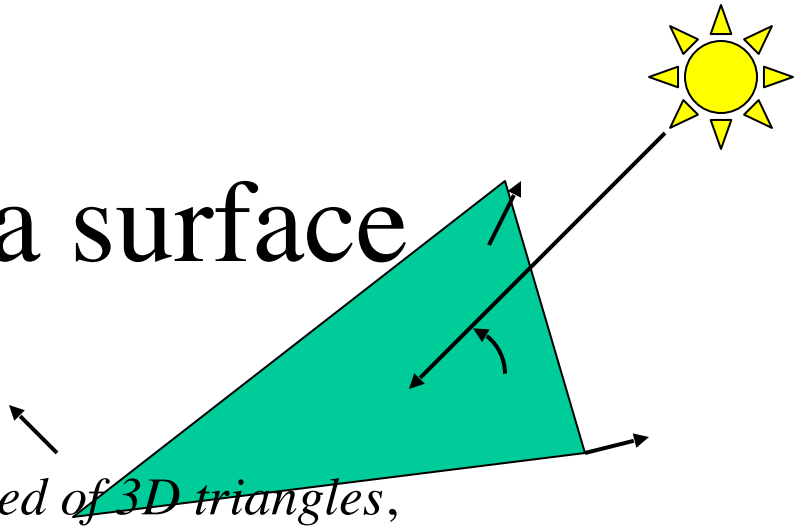
The solvent accessible surface is the interface between molecule and its solvent. Solvent molecules on the surface may behave differently that bulk solvent.

Surfaces have:
- size/area
- electrostatic properties
- shape properties



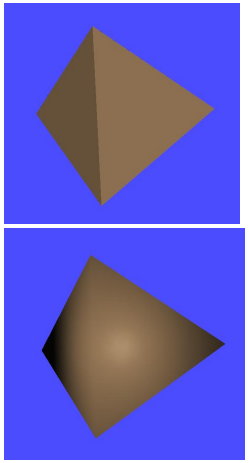
credits: Chris Bystroff

Rendering a surface



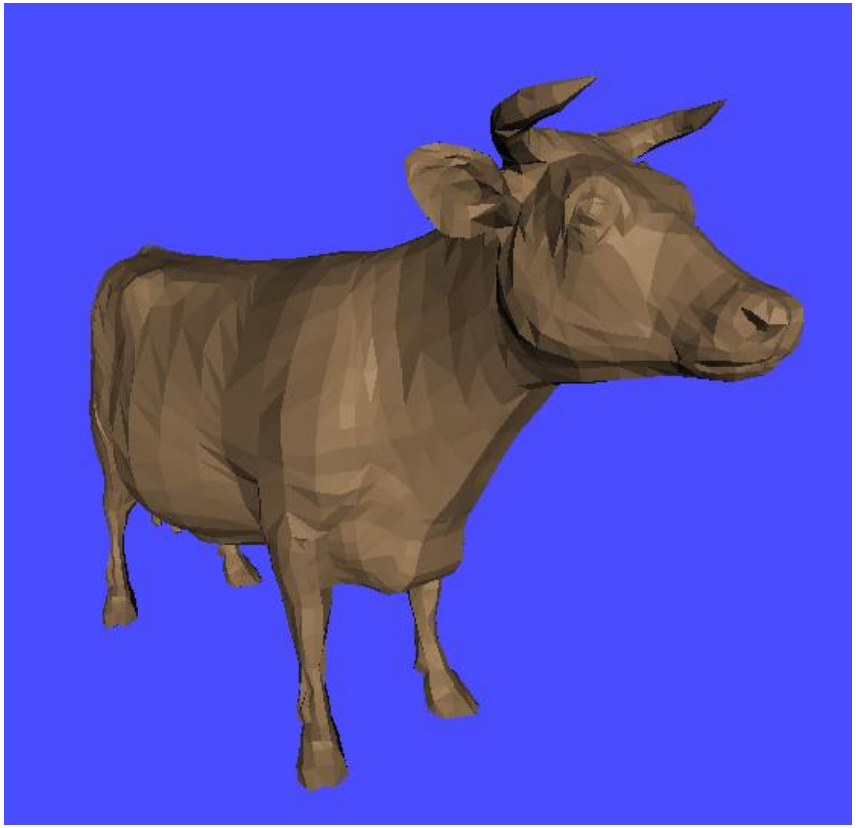
- *A surface of any shape may be composed of 3D triangles*, that is, 3 sets of xyz coordinates, one for each vertex.
- To display the surface on the screen, each triangle is rotated and translated according to the current **frame of reference**.
- Continuous triangles make a continuous surface.
- Then, each pixel is assigned a **brightness** according to the angle between the triangle and the light source.

- Phong shading may be applied to simulate *curvature*. In this case, each pixel in the triangle has a different brightness, depending on where it is.

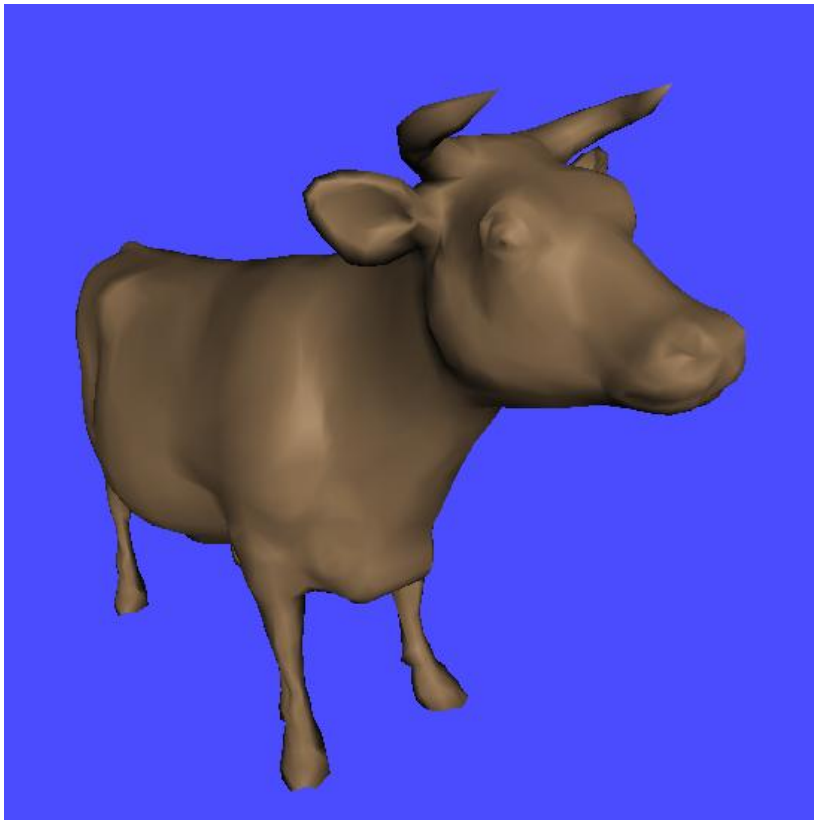


credits: Chris Bystroff

Surfaces maybe described as a set of connected triangles



A cow-shaped manifold made of triangles.



Phong-shaded cow. Shading give the illusion of higher resolution.

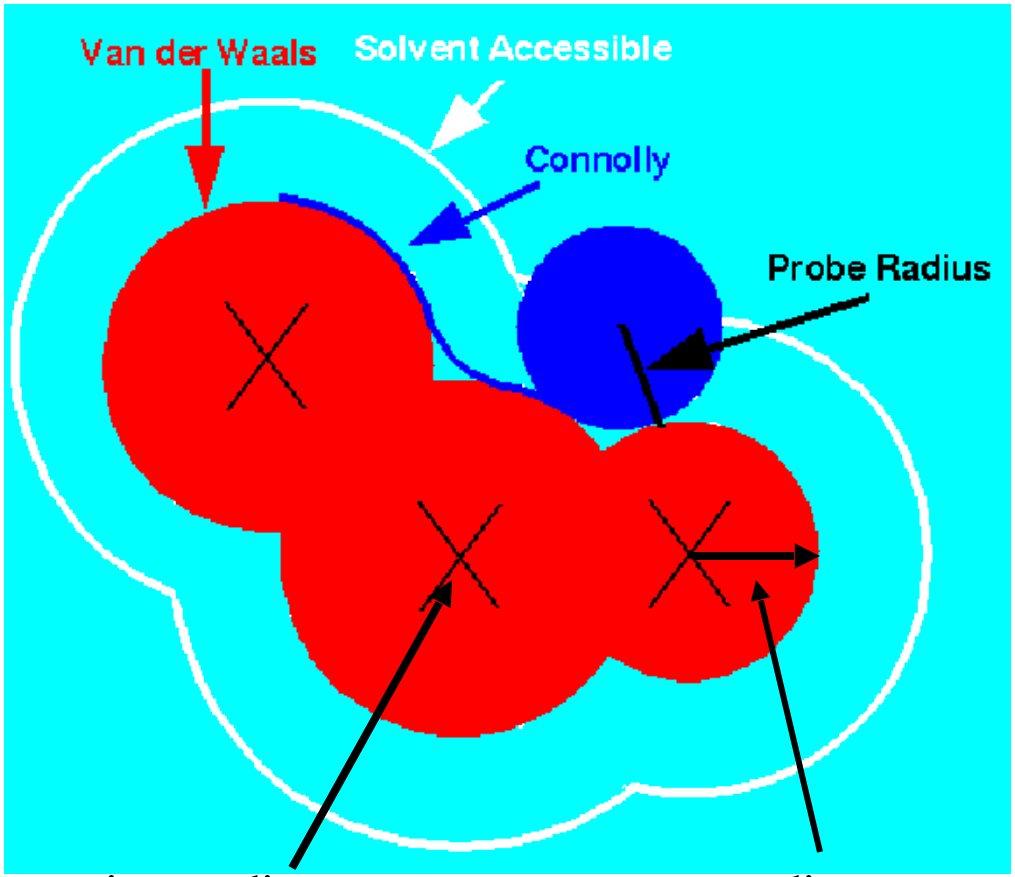
credits: Chris Bystroff

The Connolly surface

Conceptually, roll a probe sphere over the molecule...

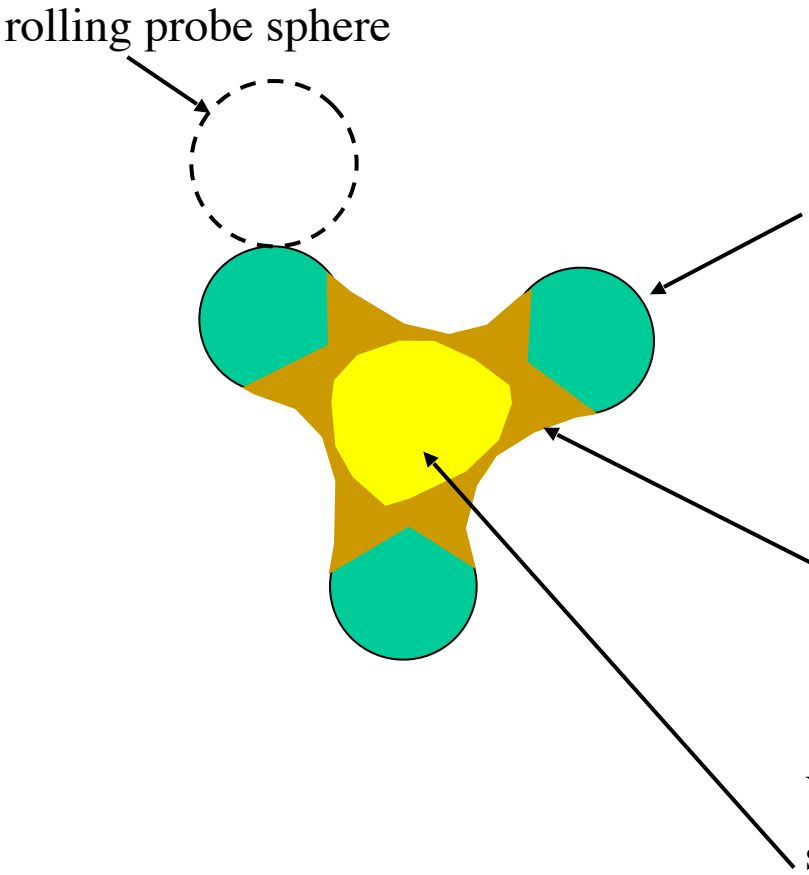
- Everywhere the center of the sphere goes is the Solvent Accessible Surface (SAS).

- Everywhere the sphere touches (including empty space) is the Solvent Excluded (or "Connolly") Surface (SES).



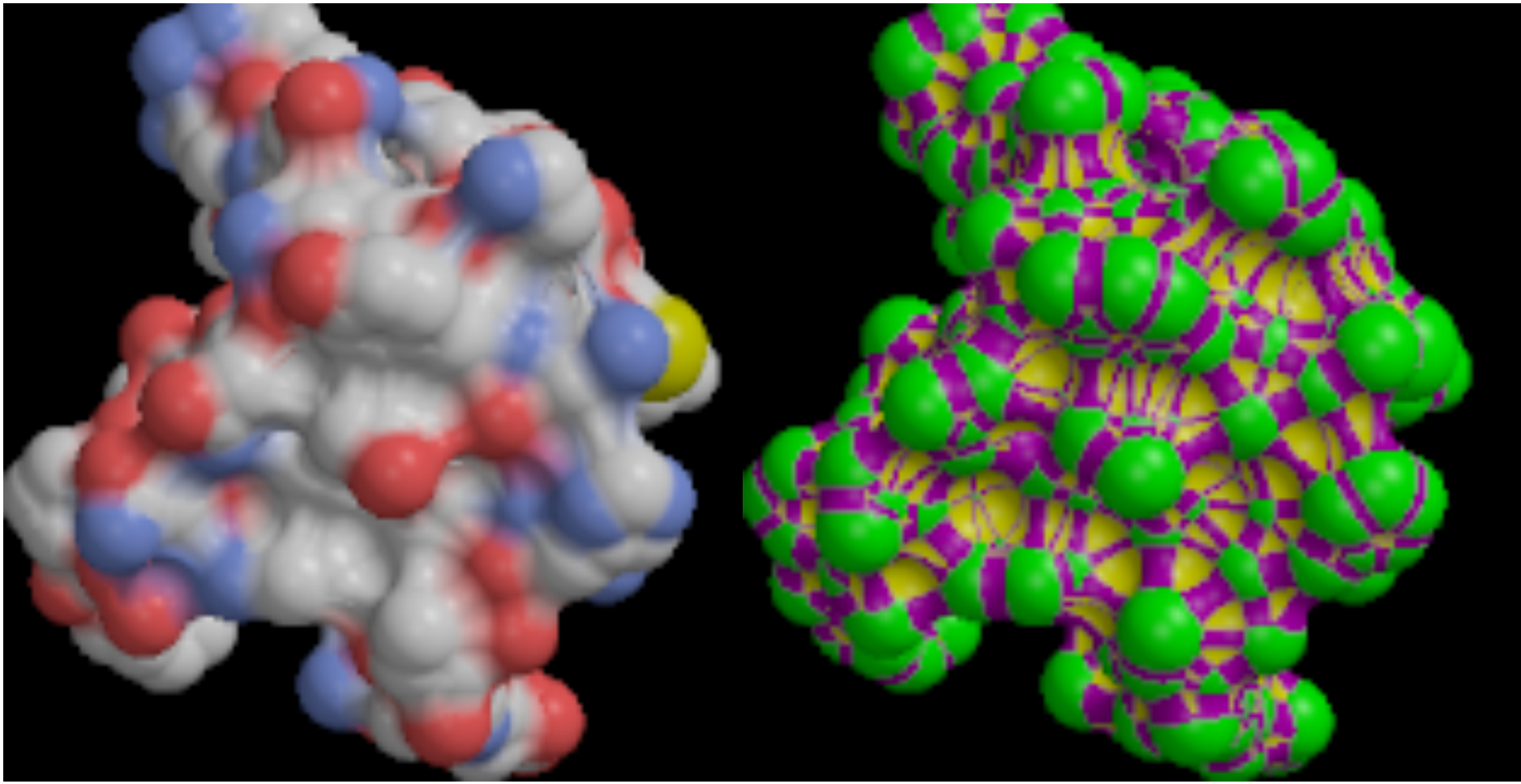
credits: Chris Bystroff

Surface shapes



credits: Chris Bystroff

Coloring by atom, by shape



Surfaces maybe shaded by partial charge.

or by shape. Yellow parts are ‘reentrant’.

credits: Chris Bystroff

Chimera demo
1BYI

Exposure of amino acid to solvent

is quantified by solvent accessibility

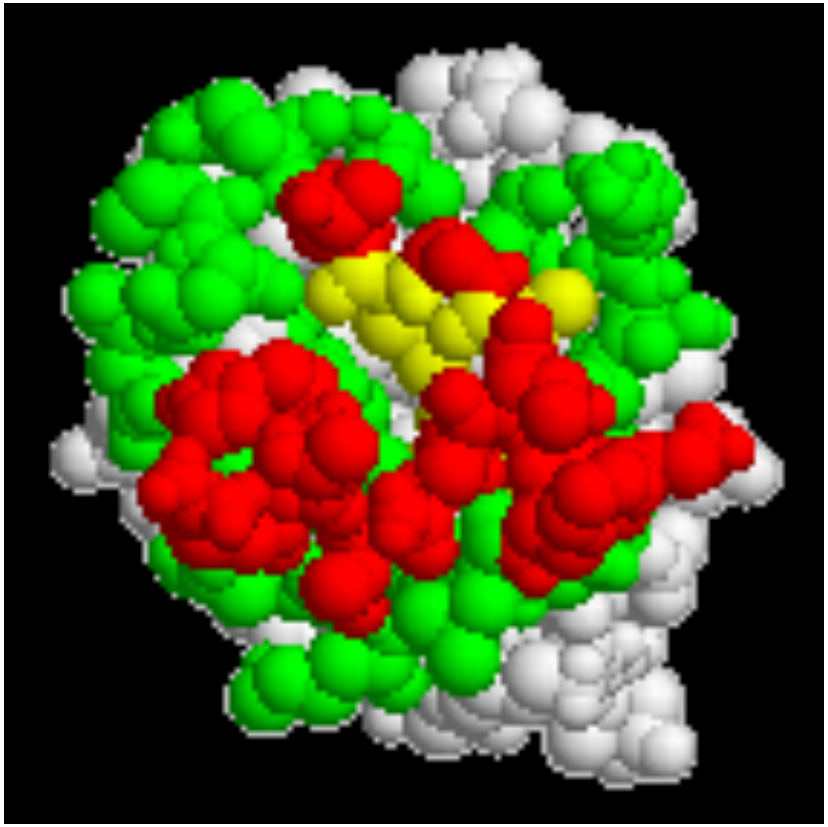
Solvent Accessibility

Size of the area of an amino acid that is exposed to solvent (water).

•Maximum solvent accessible area for each amino acid is its whole surface area.

•Hydrophobic residues like to be Buried inside (interior).

•Hydrophilic residues like to be exposed on the surface.



Google image

Structure Analysis

•If we know the 3D protein structure, we can calculate solvent accessible surface area (SASA) for each amino acids from 3D structure

•Then we can calculate the relative solvent accessibility (RSA) by calculating what percentage is exposed to solvent

•If the RSA is < 25 %, then buried (**b**), else exposed (**e**)

•Most widely used tool: **DSSP** (Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. Kabsch and Sander, 1983)

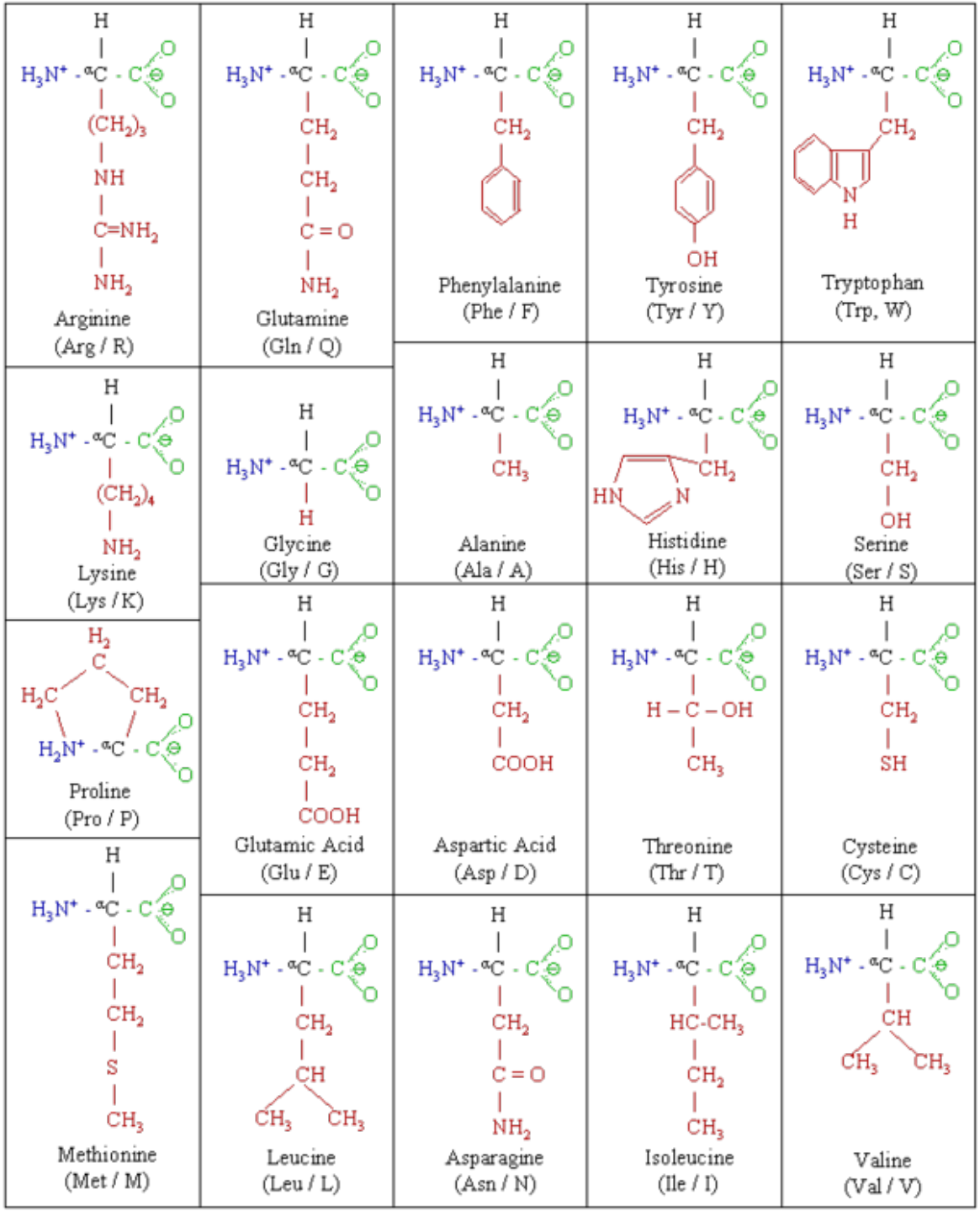
But what this has to do
with decision tree?

Protein Relative Solvent Accessibility Prediction

- We have two classes: buried (**b**), else exposed (**e**)
  - Binary labels (Y)

- We know that amino acids has some chemical properties (X)
  - Like some are charrged,

  - Instance space, X
  - Sample of labeled training data { $\langle x^{(i)}, y^{(i)} \rangle$}
  - Hypothesis space, H = { f: X→Y }



20 naturally occurring amino acid residues

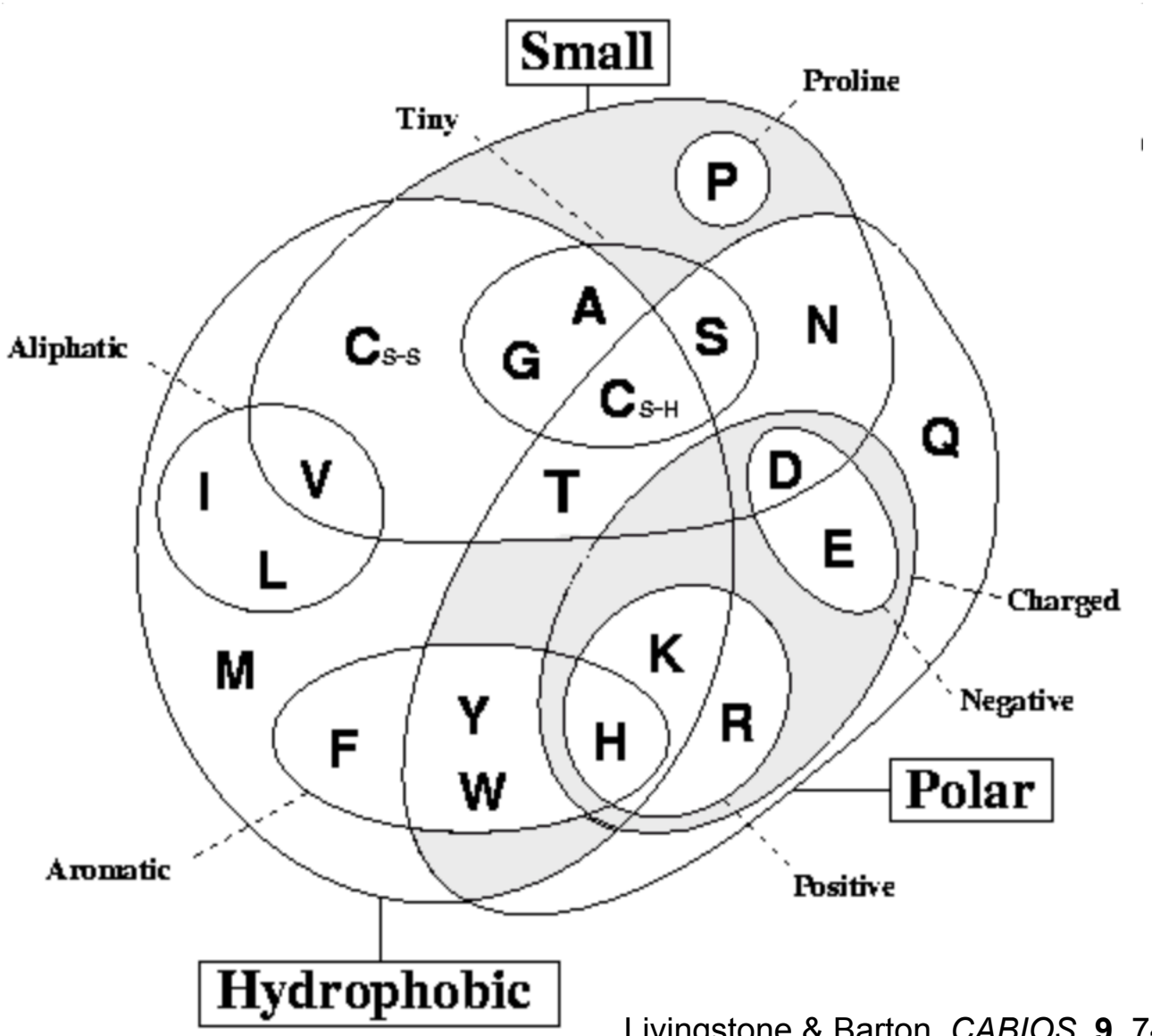
Google image

Properties of Amino Acids

Amino acid	Abbrev.	Side chain	Hydro-phobic	Polar	Charged	Small	Tiny	Aromatic or Aliphatic	van der Waals volume	Codon	Occurrence in proteins (%)
Alanine	Ala, A	$-CH_3$	X	-	-	X	X	-	67	GCU, GCC, GCA, GCG	7.8
Cysteine	Cys, C	$-CH_2SH$	X	-	-	X	-	-	86	UGU, UGC	1.9
Aspartate	Asp, D	$-CH_2COOH$	-	X	negative	X	-	-	91	GAU, GAC	5.3
Glutamate	Glu, E	$-CH_2CH_2COOH$	-	X	negative	-	-	-	109	GAA, GAG	6.3
Phenylalanine	Phe, F	$-CH_2C_6H_5$	X	-	-	-	-	Aromatic	135	UUU, UUC	3.9
Glycine	Gly, G	-H	X	-	-	X	X	-	48	GGU, GGC, GGA, GGG	7.2
Histidine	His, H	$-CH_2-C_3H_3N_2$	-	X	positive	-	-	Aromatic	118	CAU, CAC	2.3
Isoleucine	Ile, I	$-CH(CH_3)CH_2CH_3$	X	-	-	-	-	Aliphatic	124	AUU, AUC, AUA	5.3
Lysine	Lys, K	$-(CH_2)_4NH_2$	-	X	positive	-	-	-	135	AAA, AAG	5.9
Leucine	Leu, L	$-CH_2CH(CH_3)_2$	X	-	-	-	-	Aliphatic	124	UUA, UUG, CUU, CUC, CUA, CUG	9.1
Methionine	Met, M	$-CH_2CH_2SCH_3$	X	-	-	-	-	-	124	AUG	2.3
Asparagine	Asn, N	$-CH_2CONH_2$	-	X	-	X	-	-	96	AAU, AAC	4.3
Proline	Pro, P	$-CH_2CH_2CH_2-$	X	-	-	X	-	-	90	CCU, CCC, CCA, CCG	5.2
Glutamine	Gln, Q	$-CH_2CH_2CONH_2$	-	X	-	-	-	-	114	CAA, CAG	4.2
Arginine	Arg, R	$-(CH_2)_3NH-C(NH)NH_2$	-	X	positive	-	-	-	148	CGU, CGC, CGA, CGG, AGA, AGG	5.1
Serine	Ser, S	$-CH_2OH$	-	X	-	X	X	-	73	UCU, UCC, UCA, UCG, AGU,AGC	6.8
Threonine	Thr, T	$-CH(OH)CH_3$	X	X	-	X	-	-	93	ACU, ACC, ACA, ACG	5.9
Valine	Val, V	$-CH(CH_3)_2$	X	-	-	X	-	Aliphatic	105	GUU, GUC, GUA, GUG	6.6
Tryptophan	Trp, W	$-CH_2C_8H_6N$	X	-	-	-	-	Aromatic	163	UGG	1.4
Tyrosine	Tyr, Y	$-CH_2-C_6H_4OH$	X	X	-	-	-	Aromatic	141	UAU, UAC	3.2

Livingstone & Barton, *CABIOS*, **9**, 745-756, 1993

Venn diagram of Amino Acid Properties



Livingstone & Barton, *CABIOS*, **9**, 745-756, 1993

Attributes

Hydrophobic
Polar
Small
Proline
Tiny
Aliphatic
Aromatic
Positive
Negative
Charged

Labels

Buried
Exposed

Livingstone & Barton, *CABIOS*, **9**, 745-756, 1993

Protein RSA Prediction using Decision Trees

- Well posed function approximation problems:
  - Instance space, X
  - Sample of labeled training data { $\langle x^{(i)}, y^{(i)} \rangle$ }
  - Hypothesis space, H = { f: X→Y }

- Learning is a search/optimization problem over H
  - Various objective functions
    - minimize training error (0-1 loss)
    - among hypotheses that minimize training error, select smallest (?)

- Decision tree learning
  - Greedy top-down learning of decision trees (ID3, C4.5, ...)
  - Overfitting and tree/rule post-pruning
  - Extensions…