

Digitally Assessing Protein Properties

Biochemistry Boot Camp
Session #2
Nick Fitzkee
nfitzkee@chemistry.msstate.edu

Protein as Chemicals

- Molecular weight
- Chemical formula (e.g. $C_{274}H_{427}N_{69}O_{93}S_1$)
- Isoelectric point
- Sequence & Residue composition
- Solubility
- Structure
- Concentration/extinction coefficient

→ How do we access this information?

Sequence of GB3

- Primary Structure:

NT-Met-Gln-Tyr-Lys-...-Thr-Glu-**CT**

- More convenient:

```
MQYKLVINGK TLKGETTTKA VDAETAEKAF  
KQYANDNGVD GVWTYDDATK TFTVTE
```

- Can we search this (think Google)?

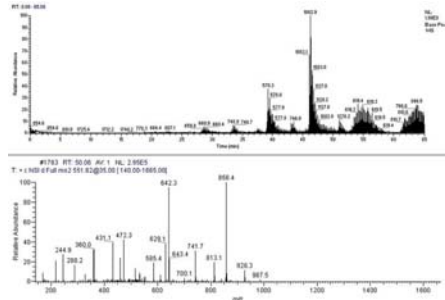
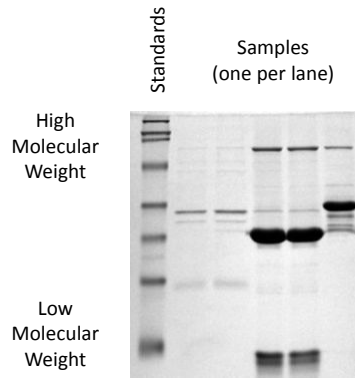
Website #1: Protparam

- <http://web.expasy.org/protparam/>
- **Input:** Protein sequence (one-letter codes)
- **Output:** Basic chemical properties
 - Molecular weight
 - Isoelectric point (pI)
 - Extinction coefficient

Molecular Weight

Polyacrylamide Gel Electrophoresis (SDS-PAGE)

Mass Spectrometry (ESI-MS, LC-MS)

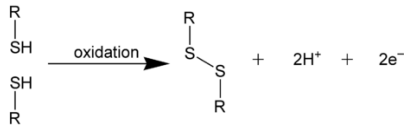


Sources: en.wikipedia.org/wiki/SDS-PAGE, en.wikipedia.org/wiki/Protein_mass_spectrometry

Residue Composition

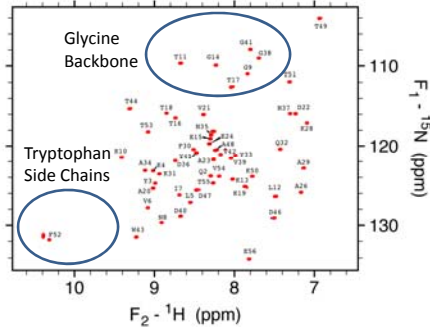
Disulfide Formation (Cysteine Content)

Protein ¹⁵N HSQC (NMR)

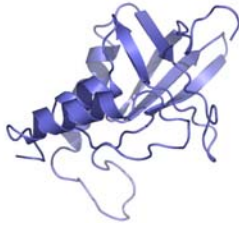


Reducing Agents:

- 2-Mercaptoethanol (BME, 5-10 mM)
- Dithiothreitol (DTT, 1-5 mM)
- Tris-(2 carboxyethyl) phosphine (TCEP, < 1 mM)

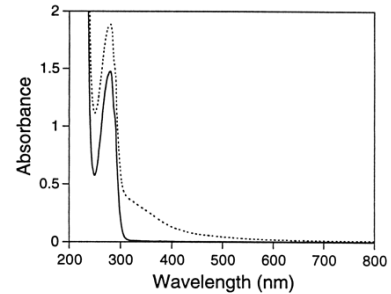


Extinction Coefficient



Tryptophan side chain absorbs light at 280 nm

More absorbance → More protein



If we know the extinction coefficient, we can *estimate* the concentration.

Calculating Protein Concentration

(Beer's Law)

- **UV-Vis:** Absorbance at 280 nm is 0.348 in a 0.3 cm quartz cuvette
 - Most cuvettes are 1 cm
- **Protparam:** Extinction coefficient at 280 nm is $9970 \text{ M}^{-1} \text{ cm}^{-1}$
- **Beer's Law:** $A = \epsilon Cl$



Source: www.malvernstore.com

Caveats: Extinction Coefficient

- Uncertainty can be as much as 10%
 - Can be worse if your technique is poor!
- No Trp means low absorbance at 280 nm
 - See Anthis, N.J. and Clore, G.M. (2013) *Protein Science*. <http://www.ncbi.nlm.nih.gov/pubmed/?term=23526461>
- DNA, other impurities or compounds may artificially increase absorbance at 280 nm

Think and Discuss

The extinction coefficient can be calculated from primary structure alone. Why is this important?

Website #2: Entrez

- <http://www.ncbi.nlm.nih.gov/sites/gquery>
- **Input:** Gene names, organisms, authors, etc.
- **Output:** Curated summary of research
 - Accepted DNA and protein sequences
 - Summaries of associated diseases
 - Recent research papers

Entrez Tricks

- **Database restriction**

| | |
|---------------------|---------------------------------|
| srcdb refseq [prop] | Only search reference sequences |
| srcdb pdb [prop] | Only search the PDB |
- **Journal restriction**

| | |
|----------------------|-----------------------------------------------------|
| 1998:2003 [dp] | Dates from 1998-2003 |
| fitzkee_nc [auth] | Author name is Fitzkee, N. C. |
| j am chem soc [jour] | Journal name is JACS (need to know abbreviation) |

Entrez Tricks

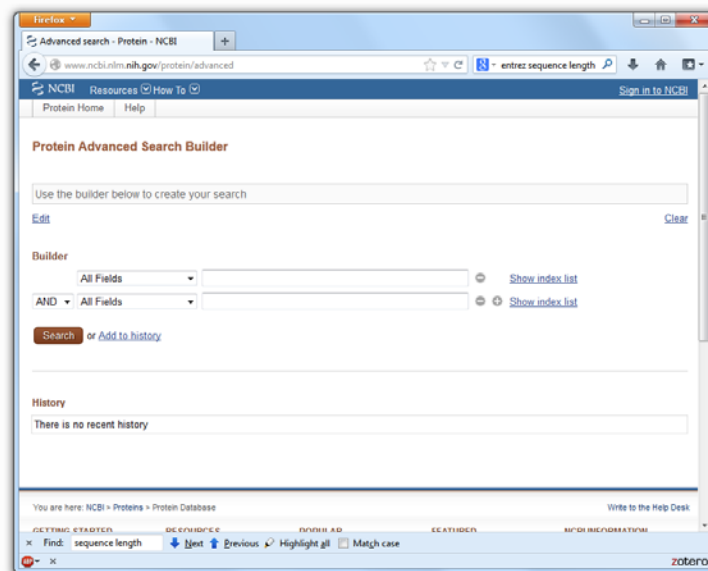
- Combining Terms

| | |
|-------------------|---------------------------|
| xx AND yy | Must have xx and yy |
| xx OR yy | Must have either xx or yy |
| NOT zz | Without term zz |
| xx AND (yy OR zz) | Complex example |

- Chemical Properties

75:100 [sequence length]
3500:6000 [molecular weight]

Advanced Searches



Practice

- What's the sequence of your favorite protein?
- What's the extinction coefficient of human heart fatty acid binding protein?
- What human disease is associated with phenylalanine hydroxylase?

Website #3: Protein Data Bank

- <http://rcsb.org/>
- **Input:** Protein name, PDB ID, authors, etc.
- **Output:** 3D coordinates of protein structures
 - Author information on methods
 - Cofactors and other information

What is a PDB file?

- Example: Ricin (2AAI)
- Text file contains a summary of information used in structure determination
- Most important: ATOM records contain X, Y, Z in *Ångströms* (1×10^{-10} m)
 - Most atoms have a radius of 0.5-2 Å

Properties of PDB Files

- Experimental methodology:
 - X-Ray: Typically more precise
 - NMR: Need lots of “restraints;” sometimes hard to assess quality
- “Good” Structures (for X-Ray)
 - Low resolution ($< 2\text{Å}$)
 - Low R-value ($< 20\%$)
 - Low R_{free} -value ($< 25\%$)

Searching the PDB

Query Refinements: Select an item or pie chart Hide

Organism

- Homo sapiens (458)
- Mus musculus (15)
- Haemophilus influenzae (14)
- Methanosarcina thermophila (12)
- Drosophila melanogaster (6)
- Thalassiosira weissflogii (6)
- Coccomyxa sp. PA (5)
- Other (36)

Taxonomy

- Eukaryota (502)
- Bacteria (31)
- Archaea (18)
- Unassigned (5)
- Viruses (1)
- Other (1)

Experimental Method

- X-ray (554)
- Neutron Diffraction (2)
- Hybrid (1)

X-ray Resolution

- less than 1.5 Å (86)
- 1.5 - 2.0 Å (273)
- 2.0 - 2.5 Å (165)
- 2.5 - 3.0 Å (30)
- 3.0 and more Å (1)
- more choices...

Release Date

- before 2000 (128)
- 2000 - 2005 (68)
- 2005 - 2010 (168)
- 2010 - today (193)
- this year (11)
- more choices...

Polymer Type

- Protein (557)

Enzyme Classification

- 4: Lyases (538)
- 3: Hydrolases (9)

SCOP Classification

- All beta proteins (261)
- Alpha and beta proteins (a/b) (12)

Protein Symmetry

- Cyclic (528)
- Dihedral (29)

Protein Stoichiometry

- Monomer (480)
- Homomer (71)
- Heteromer (6)

Refine Query with Advanced Search
Remove Similar: Select Percent Similarity ▼

Advanced Searching

Advanced Search Interface

Text Search ?

Text Result Count

557
PDB Entries
(Structures)

AND

Choose a Query Type: ?

Result Count

[Add Search Criteria](#)

Remove Similar Sequences at: 90% ? Identity ?

Match all ▼ of the above conditions. Clear All Parameters Submit Query

Think and Discuss

What are the advantages to large, public databases of scientific information? Are there any disadvantages?

Summary

- Protein properties depend on their primary, secondary, tertiary, and quaternary structure
- Computer databases can organize huge amounts of data on biomolecular systems
- Entrez and the PDB are curated from published research worldwide