Digitally Assessing Protein Properties

Biochemistry Boot Camp 2017
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

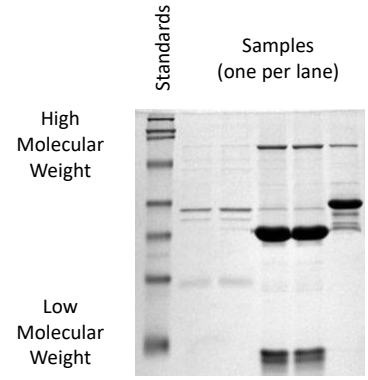
- Can we search this (think Google)?

Website #1: Protparam

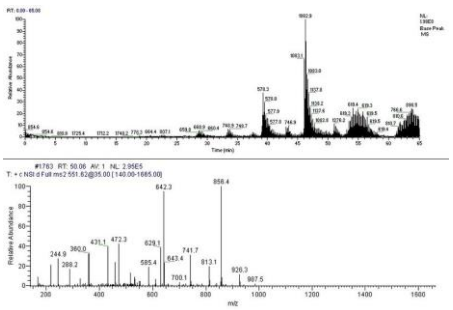
- http://web.expasy.org/protparam/
- **Input:** Protein sequence (one-letter codes)
- **Output:** Basic chemical properties
  - Molecular weight
  - Isoelectric point (pI)
  - Extinction coefficient

Molecular Weight

Polyacrylamide Gel Electrophoresis (SDS-PAGE)



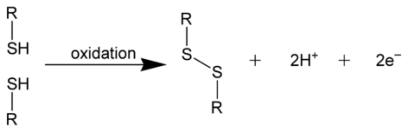
Mass Spectrometry (ESI-MS, LC-MS)



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

Residue Composition

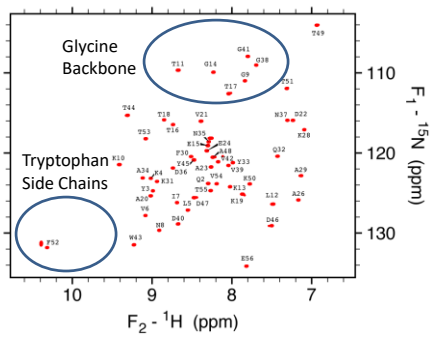
Disulfide Formation (Cysteine Content)



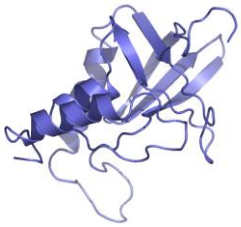
Reducing Agents:

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

Protein ^{15}N HSQC (NMR)

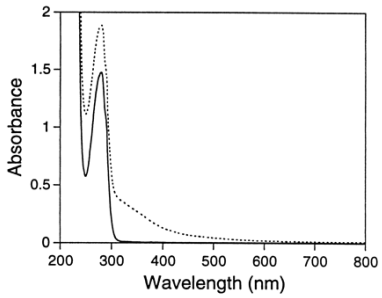


Extinction Coefficient



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 M^{-1} cm^{-1}
- **Beer's Law:** $A = \epsilon C l$

Source: www.malvernstore.com

What If My Protein Doesn't Have Trp?

- No Trp means low (no) absorbance at 280 nm
- Protein backbone has intrinsic absorbance at 205 nm
  - See Anthis, N.J. and Clore, G.M. (2013) *Protein Science.* http://www.ncbi.nlm.nih.gov/pubmed/?term=23526461
  - Website: http://nickanthis.com/tools/a205.html
- Complications:
  - Protein concentration will need to be quite low, which may introduce dilution errors
  - Many buffers absorb at 205 nm, these can overwhelm the protein signal (even when using a blank)
  - **Solution:** Careful dilution, use water as a blank if possible

Caveats: Extinction Coefficient

- Uncertainty can be as much as 10%
  - Can be worse if your technique is poor!
- Absorbance values need to be between 0.1-1.0 for highest accuracy
  - Estimate your expected A_{280} and dilute if necessary
- **Scattering of aggregates:** If the baseline is not zero at 600 nm, you are probably not getting an accurate value!
- DNA, other impurities or other 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: NCBI Databases

- 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

NCBI Tricks #1

- 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)

NCBI Tricks #2

- Combining Terms

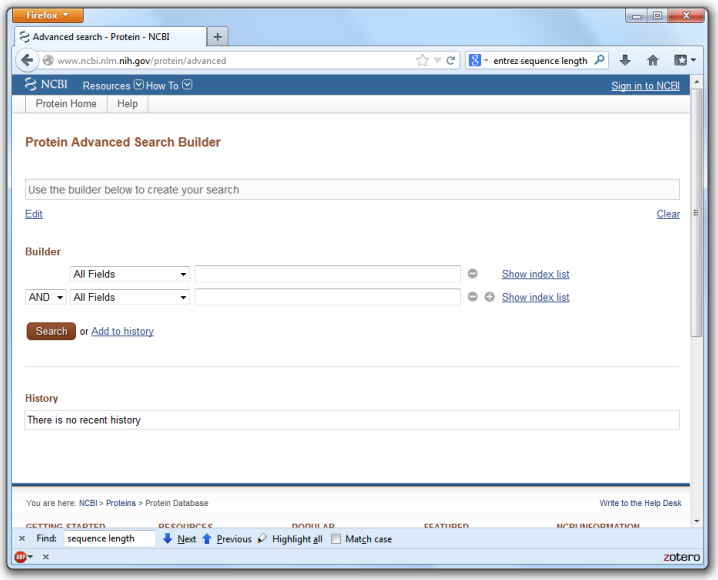
xx AND yy	Must have xx and yy
xx OR yy	Must have either xx or yy
NOT zz	Without term zz
xx AND (yy OR zz)	Complex example

- 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

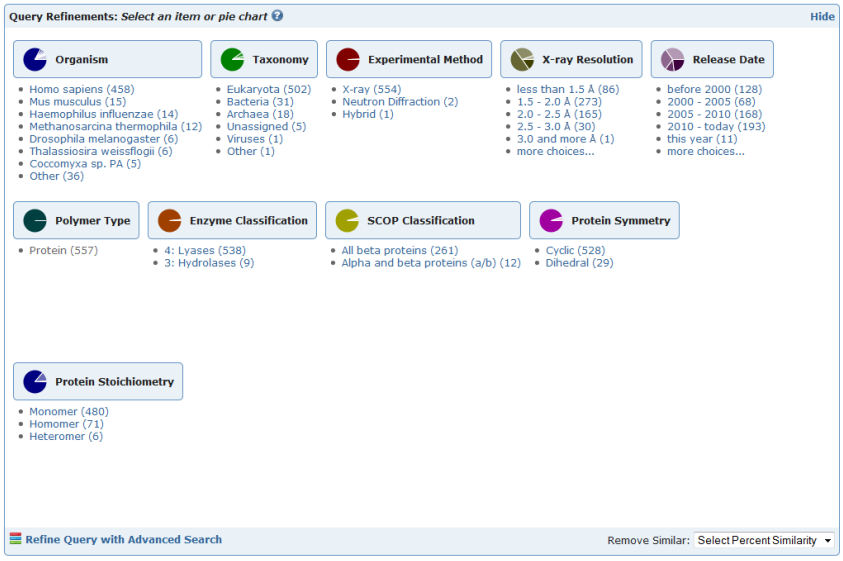
What is a PDB file?

- 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

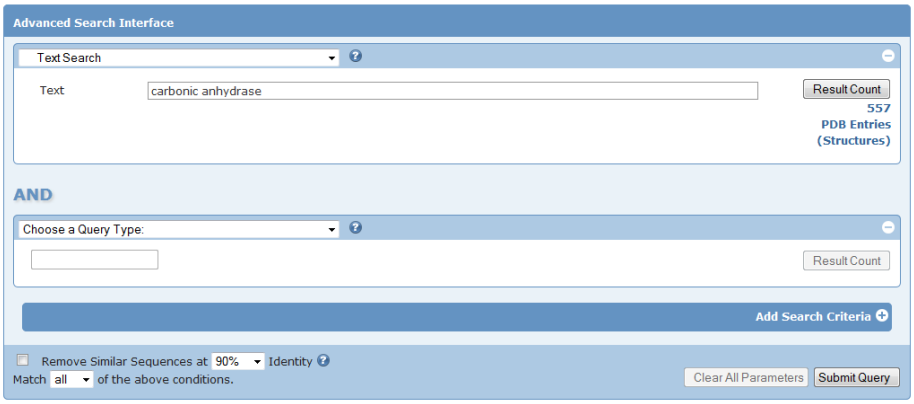
Properties of PDB Files

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

Searching the PDB



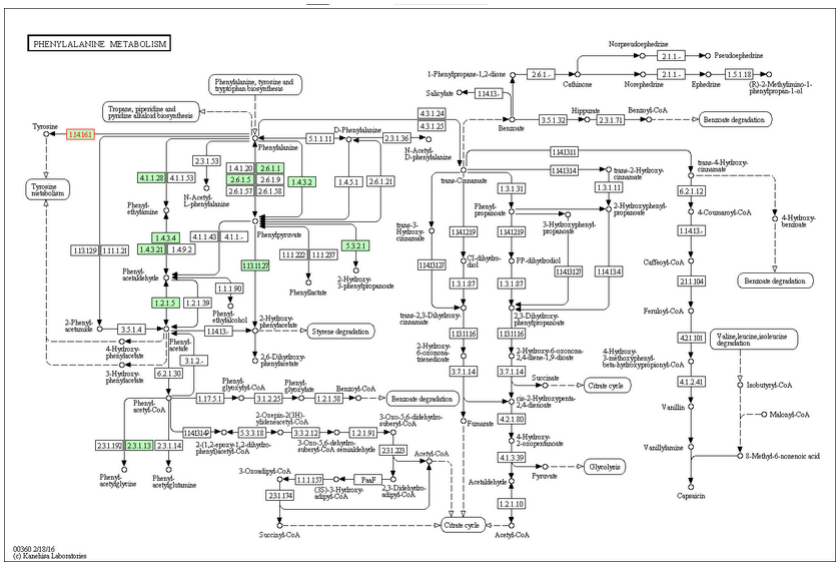
Advanced Searching



Website #4: KEGG

- http://www.genome.jp/kegg/
  (Kyoto Encyclopedia of Genes and Genomes)
- **Input:** Protein name, PDB ID, authors, etc.
- **Output:** What reactions does an enzyme catalyze?
  - Metabolic pathways
  - The "big picture"

Pathway for Phenylalanine Hydroxylase



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