

# Digitally Assessing Protein Properties

Biochemistry Boot Camp 2023

Session #2

Nick Fitzkee

[nfitzkee@chemistry.msstate.edu](mailto: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)?

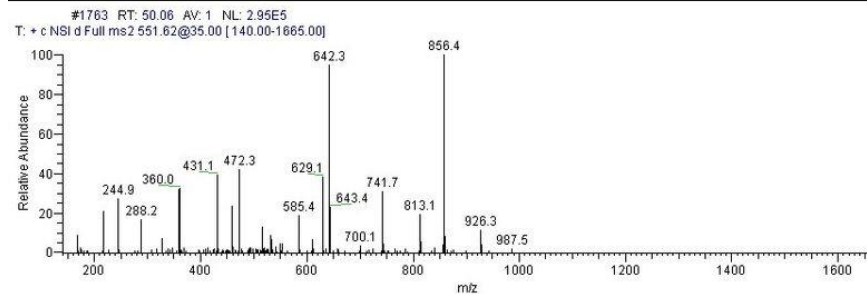
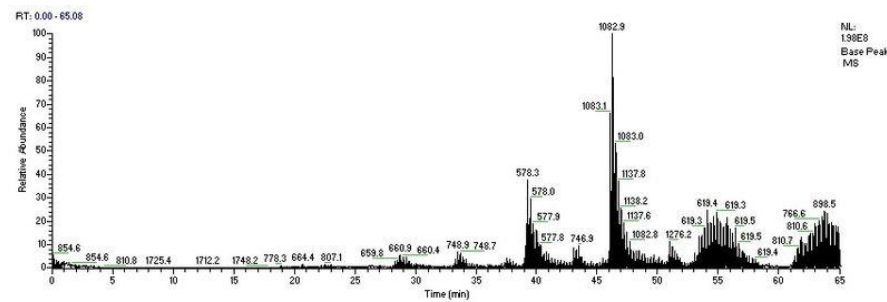
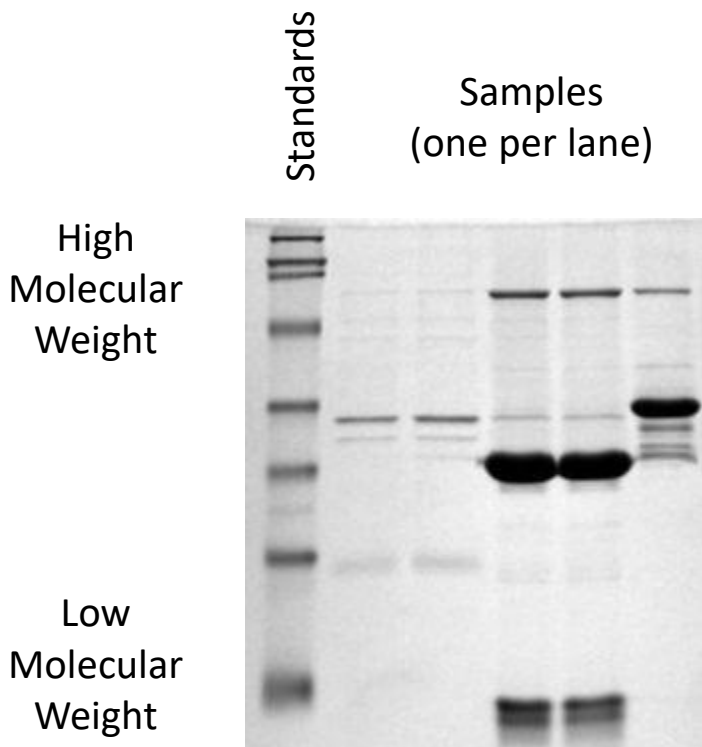
# Website #1: Protparam

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

# Molecular Weight

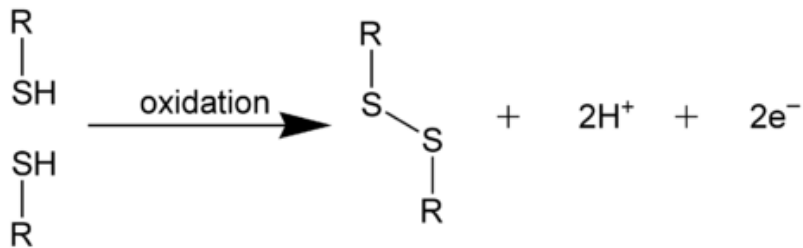
Polyacrylamide Gel Electrophoresis  
(SDS-PAGE)

Mass Spectrometry  
(ESI-MS, LC-MS)



# Residue Composition

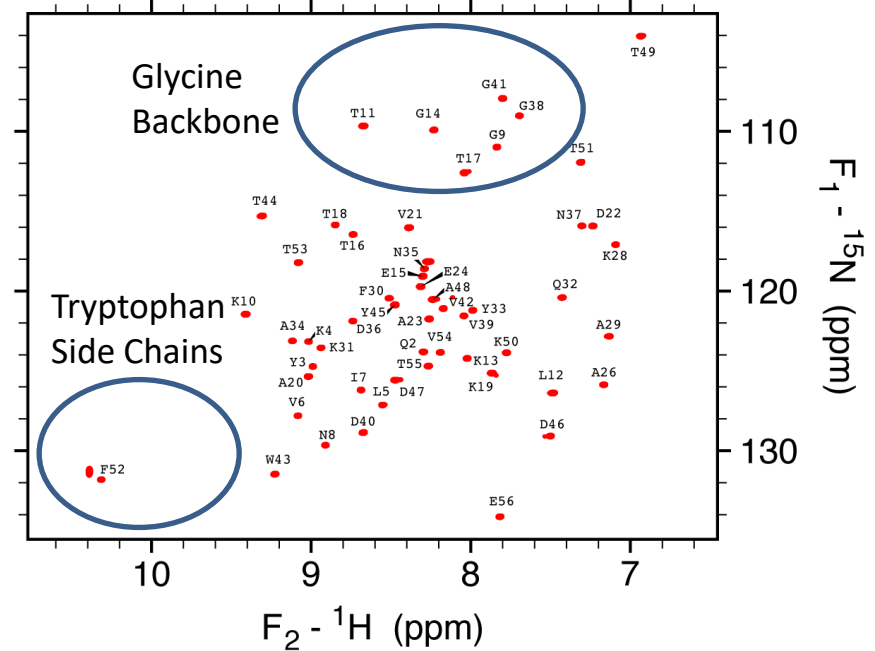
## Disulfide Formation (Cysteine Content)



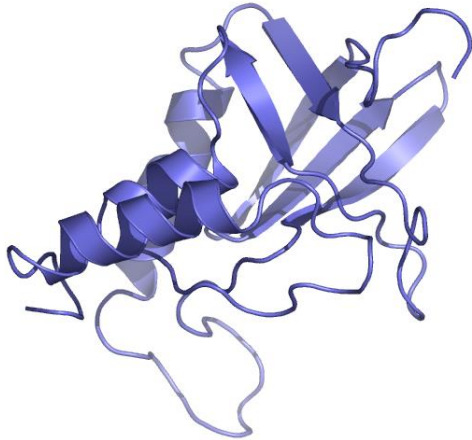
### Reducing Agents:

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

## Protein $^{15}\text{N}$ HSQC (NMR)

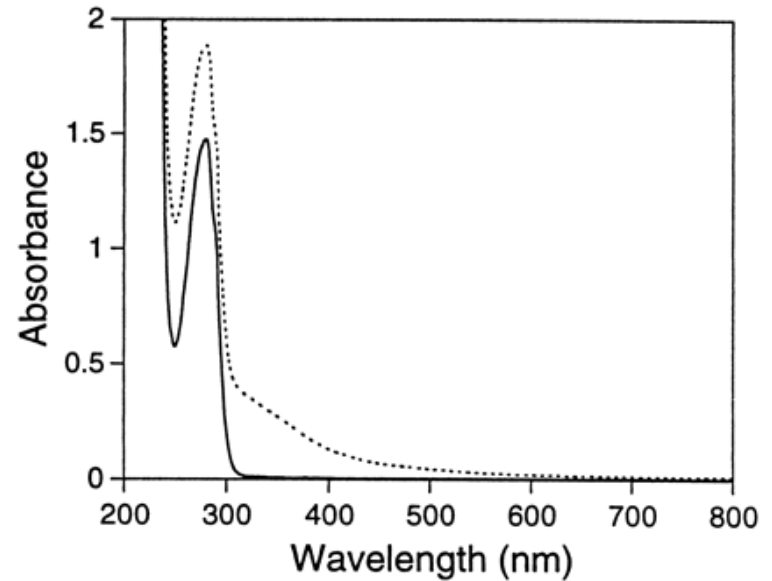


# Extinction Coefficient



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  $9970 \text{ M}^{-1} \text{ cm}^{-1}$
- **Beer's Law:**  $A = \epsilon Cl$



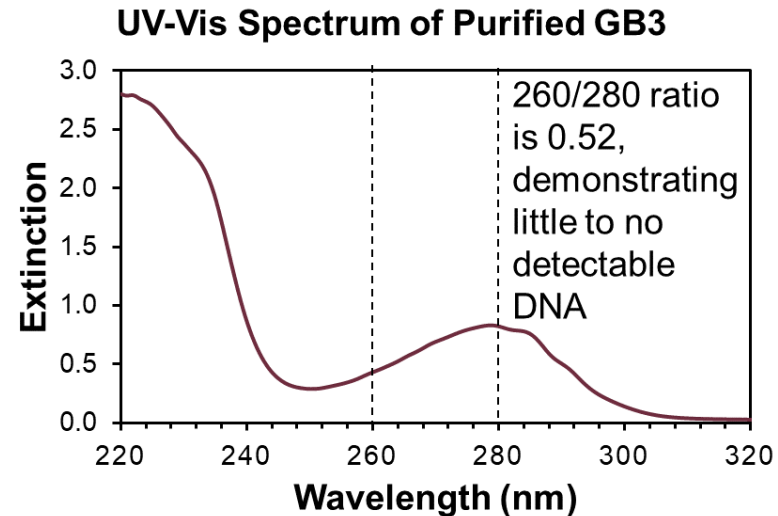


# What If My Protein Doesn't Have Trp?

- 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 ( $260/280$  ratio  $< 0.6$ )



## *Think and Discuss*

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

# Website #2: NCBI Databases

- <https://www.ncbi.nlm.nih.gov/>
- **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

The screenshot shows a web browser window with the following elements:

- Browser Tab:** "Advanced search - Protein - NCBI"
- Address Bar:** "www.ncbi.nlm.nih.gov/protein/advanced" with a search engine icon and the text "entrez sequence length".
- Page Header:** "NCBI Resources How To" and "Sign in to NCBI".
- Navigation:** "Protein Home" and "Help" links.
- Section Header:** "Protein Advanced Search Builder".
- Instruction:** "Use the builder below to create your search".
- Builder Section:**
  - Field selection: "All Fields" dropdown.
  - Logic: "AND" dropdown.
  - Field selection: "All Fields" dropdown.
  - Buttons: "Search" and "Add to history".
  - Links: "Show index list" (twice).
- History Section:** "History" with the text "There is no recent history".
- Footer:** "You are here: NCBI > Proteins > Protein Database" and "Write to the Help Desk".
- Page Navigation:** "GETTING STARTED", "RESOURCES", "POPULAR", "FEATURED", "MORE INFORMATION".
- Search Bar:** "Find: sequence length" with "Next", "Previous", "Highlight all", and "Match case" options.
- Browser Status Bar:** "zotero" icon.

# *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* ( $1 \times 10^{-10}$  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 ( $< 2\text{\AA}$ )
  - Low R-value ( $< 20\%$ )
  - Low  $R_{\text{free}}$ -value ( $< 25\%$ )

# Searching the PDB

The screenshot shows the RCSB PDB search interface. The search query is "Full Text = 'fatty acid binding protein'". The search summary indicates 205,446 structures. The left sidebar shows refinement options, with a red circle highlighting the "Scientific Name of Source Organism" section. The main content area displays two search results: 2HMB (THREE-DIMENSIONAL STRUCTURE OF RECOMBINANT HUMAN MUSCLE FATTY ACID-BINDING PROTEIN) and 1A57 (THE THREE-DIMENSIONAL STRUCTURE OF A HELIX-LESS VARIANT OF INTESTINAL FATTY ACID BINDING PROTEIN, NMR, 20 STRUCTURES).

Search Summary: This query matches 205,446 Structures.

Refinements

- Structure Determination Methodology
  - experimental (205,446)
- Scientific Name of Source Organism
  - Homo sapiens (62,681)
  - Mus musculus (8,680)
  - synthetic construct (7,949)
  - Escherichia coli (7,197)
  - Escherichia coli K-12 (4,188)
  - Rattus norvegicus (3,778)
  - Bos taurus (3,589)
  - Saccharomyces cerevisiae (3,264)
  - Severe acute respiratory syndrome coronavirus 2 (3,105)
  - Saccharomyces cerevisiae S288C (2,516)
  - [More...](#)
- Taxonomy
  - Eukaryota (113,126)
  - Bacteria (69,678)
  - Riboviria (12,867)

1 to 25 of 205,446 Structures | Page 1 of 8,218 | Sort by Score

**2HMB**  
THREE-DIMENSIONAL STRUCTURE OF RECOMBINANT HUMAN MUSCLE FATTY ACID-BINDING PROTEIN  
Zanotti, G., Scapin, G., Spadon, P., Veerkamp, J.H., Sacchettini, J.C.  
(1992) J Biol Chem **267**: 18541-18550  
**Released** 1994-01-31  
**Method** X-RAY DIFFRACTION 2.1 Å  
**Organisms** Homo sapiens  
**Macromolecule** MUSCLE FATTY ACID BINDING PROTEIN (protein)  
**Unique Ligands** PLM

**1A57**  
THE THREE-DIMENSIONAL STRUCTURE OF A HELIX-LESS VARIANT OF INTESTINAL FATTY ACID BINDING PROTEIN, NMR, 20 STRUCTURES  
Steele, R.A., Emmert, D.A., Kao, J., Hodsdon, M.E., Frieden, C., Cistola, D.P.  
(1998) Protein Sci **7**: 1332-1339  
**Released** 1998-05-27

Note refinements!

# Advanced Searching

The screenshot displays the RCSB PDB search interface. The browser address bar shows the URL: `https://www.rcsb.org/search?request={"query":"%3A("type":"%3A"group"%2C"nodes"%3A[{"typ`. The search query entered is "fatty acid binding protein".

The **Advanced Search Query Builder** section is active, showing the following query components:

- Full Text:** fatty acid binding protein
- Structure Attributes:**
  - Refinement R Factors (All) < 0.28
  - Refinement Resolution < 2.0 Å
- Chemical Attributes:** (Collapsed)
- Sequence Similarity:** (Collapsed)
- Sequence Motif:** (Collapsed)
- Structure Similarity:** (Collapsed)
- Structure Motif:** (Collapsed)
- Chemical Similarity:** (Collapsed)

At the bottom, the search is configured to return **Structures** grouped by **No Grouping**. The search results are displayed as **Search Summary**: This query matches 205,446 Structures.

# 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”

# Search Result: Intestinal FABP

**KEGG** ORTHOLOGY: K08751 [Help](#)

<b>Entry</b>	K08751 KO
<b>Name</b>	FABP2
<b>Definition</b>	fatty acid-binding protein 2, intestinal
<b>Pathway</b>	ko03320 PPAR signaling pathway ko04975 Fat digestion and absorption
<b>Brite</b>	KEGG Orthology (KO) [BR:ko00001] Organismal Systems Endocrine system 03320 PPAR signaling pathway K08751 FABP2; fatty acid-binding protein 2, intestinal Digestive system 04975 Fat digestion and absorption K08751 FABP2; fatty acid-binding protein 2, intestinal <a href="#">BRITE hierarchy</a>
<b>Genes</b>	HSA: 2169(FABP2) PTR: 740421(FABP2) PPS: 100991717(FABP2) GGO: 101151281(FABP2) PON: 100445937(FABP2) NLE: 100581617(FABP2) MCC: 705475(FABP2) MCF: 102140395(FABP2) CSAB: 103236178(FABP2) RRO: 104663589(FABP2) » show all <a href="#">Taxonomy</a> <a href="#">KOALA</a> <a href="#">UniProt</a>
<b>Reference</b>	PMID:20716527
<b>Authors</b>	Storch J, Thumser AE
<b>Title</b>	Tissue-specific functions in the fatty acid-binding protein family.
<b>Journal</b>	J Biol Chem 285:32679-83 (2010) DOI:10.1074/jbc.R110.135210

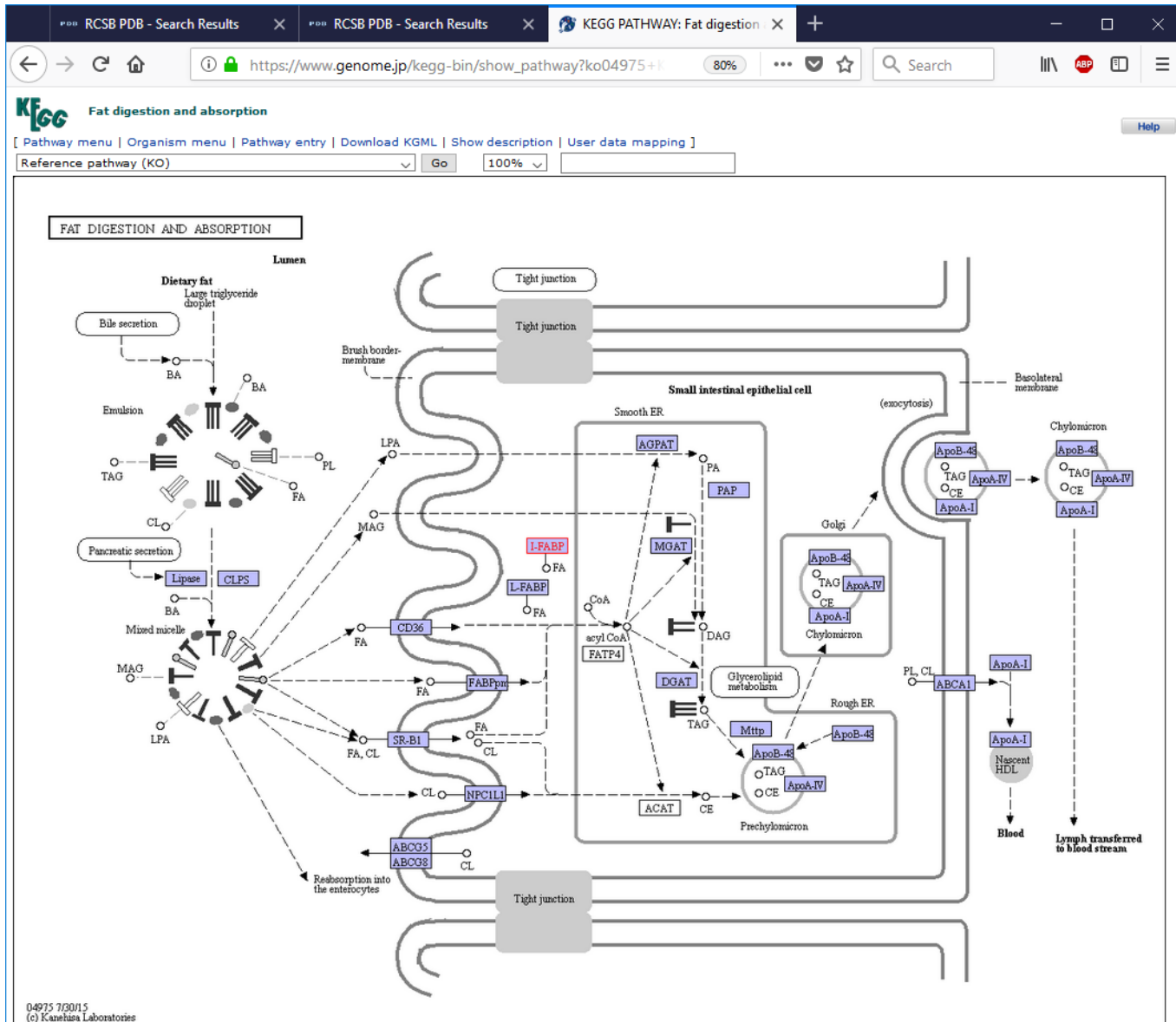
**All links**

- Ontology (2)
  - KEGG BRITE (2)
- Pathway (4)
  - KEGG PATHWAY (4)**
- Gene (663)
  - KEGG GENES (138)
  - KEGG MGENES (14)
  - RefGene (488)
  - EGENES (21)
  - OC (4)
- Protein sequence (68)
  - UniProt (62)
  - SWISS-PROT (6)
- Literature (1)
  - PubMed (1)
- All databases (740)

[Download RDF](#)

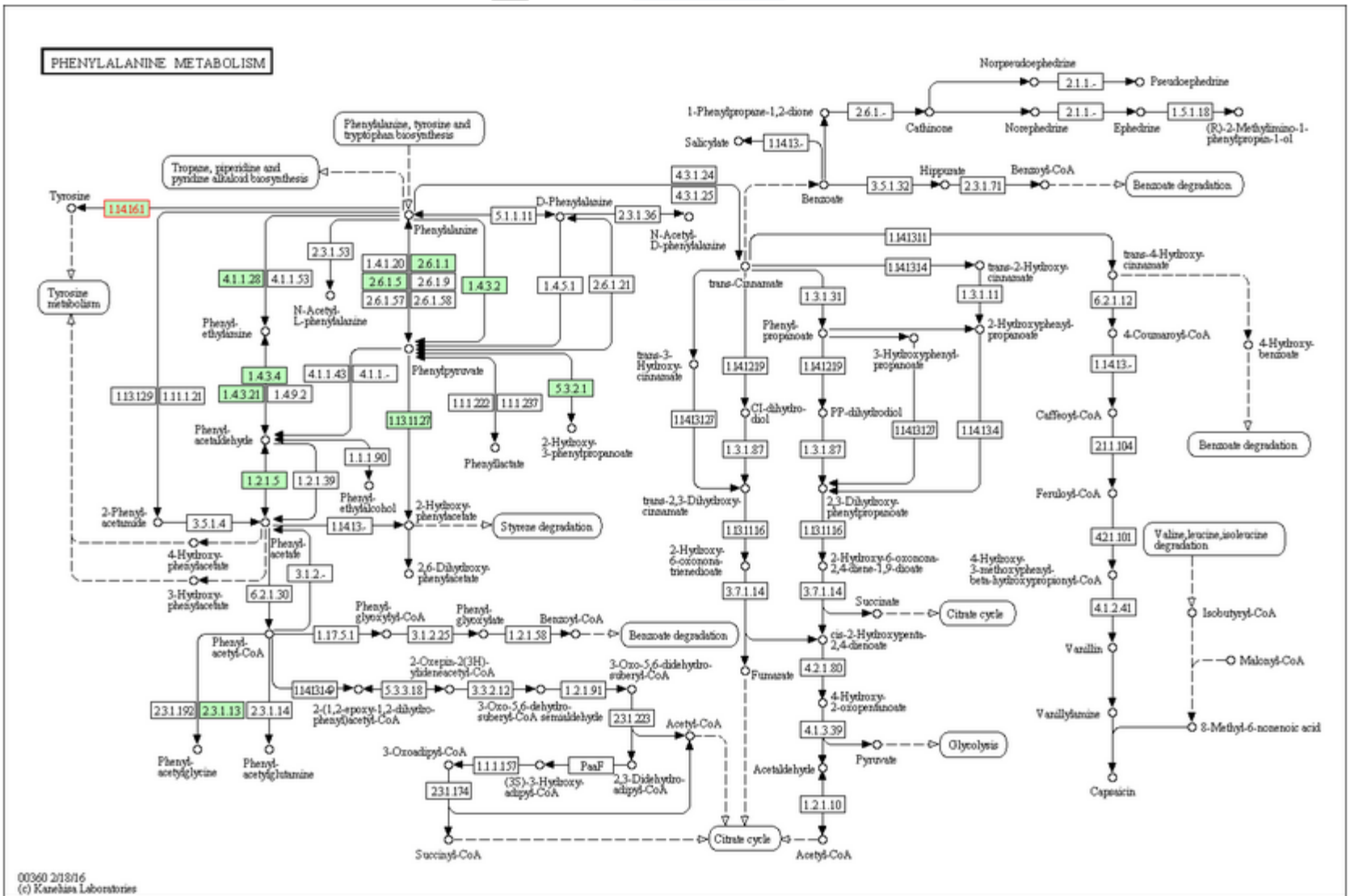
DBGET integrated database retrieval system

# Search Result: Fat Digestion and Absorption





# 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