Pet Project:

IL-6, or your favorite gene

What can we know about this gene?

- **Search for “curated” databases.**
- **To prepare for future analysis, save annotated sequence files as genename.html (in a target folder).**
- **For downstream sequence analysis, save pure sequence as FASTA format file.**

Representation of sequence – sequence file format

1.) FASTA – simple and clean

```
> gene_name, (other info)
MASASASKJHKLJLKJLDSDFSF
SSDSASFSFD…
```

Practice / DIY: retrieve sequence in Fasta format and save the file in the local computer.

How to store sequence files

- Pure text format is clean and allows down stream sequence analysis.
- .doc or .rtf allows formatting during annotation – however, extra information are inserted thus NOT suitable for computational analysis.
- !! No space in file or folder name!!- Or trouble will find you.

Practice – file types

- Using file Finder (Mac) or Explorer (PC) to view downloaded files.
- Change the “Options” so that the file **extensions (.txt) are revealed**.
- Edit the downloaded sequence file in MS Word, highlight a section of the sequence with Bold font or color and save as .doc
- Open the .doc file in a TextEditor – observe the inserted characters.

Practice – file types (Cont.)

- Load the “Mysequence.doc” file to Webcutter using “Choose file” and then “Upload sequence file”.
  - -Notice that the “sequence” in the sequence box are nonsense characters.
- Clear input; Browse and then load the .txt file. Run an analysis.

Keep you sequences as .seq or .fasta file for downstream analysis.

Sequence Representation - nucleotide

N	G	R	C	W	T	G	Y	C	Y
A	G	A	C	A	T	G	C	C	C
C		G		T			T		T
G									
T									

For complete list of IUB/IUPAC nucleic acid codes, see
https://people.bath.ac.uk/jm2219/biology/degenerate.htm

Sequence Representation - amino acids

Q:

What's the common property of these amino acids ?

1. D, E
2. I, L, V, M, F

Sequence Representation - amino acids

Example:

W D L L A Q I L C Y A L R I Y

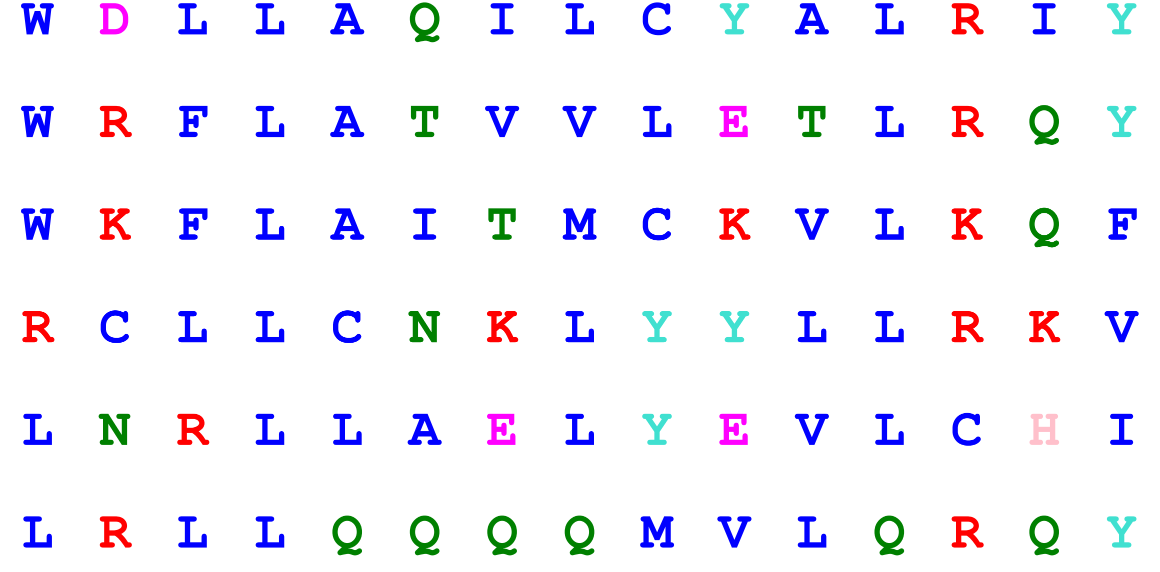
W R F L A T V V L E T L R Q Y

W K F L A I T M C K V L K Q F

R C L L C N K L Y Y L

L N R L L A E L Y E V

L R L L Q Q Q Q M V L



W D L L A Q I L C Y A L R I Y

W R F L A T V V L E T L R Q Y

W K F L A I T M C K V L K Q F

R C L L C N K L Y Y L L R K V

L N R L L A E L Y E V L C H I

L R L L Q Q Q Q M V L Q R Q Y

Coloring based on aa property.

Public Resources for Bioinformatics

- Databases
- Analysis Tools

Observe: List of databases and service at NCBI, EBI, KEGG, and Ensembl.

Public Resources (I) – Databases and data sources

- Over 1,000 in the sea of databases.
- Content-specific, such as DNA, Protein, Structure, etc.
- Species-specific, such as flybase, wormbase, OMIM, etc.
- System-specific, such as MetaCyc, AFCS, etc.

Pet Project:

IL-6, or your favorite gene

What can we know about this gene?

- **Search for “curated” databases.**
- **To prepare for future analysis, save annotated sequence files as genename.html (in a target folder).**
- **For downstream sequence analysis, save pure sequence as FASTA format file.**

Observe/Practice

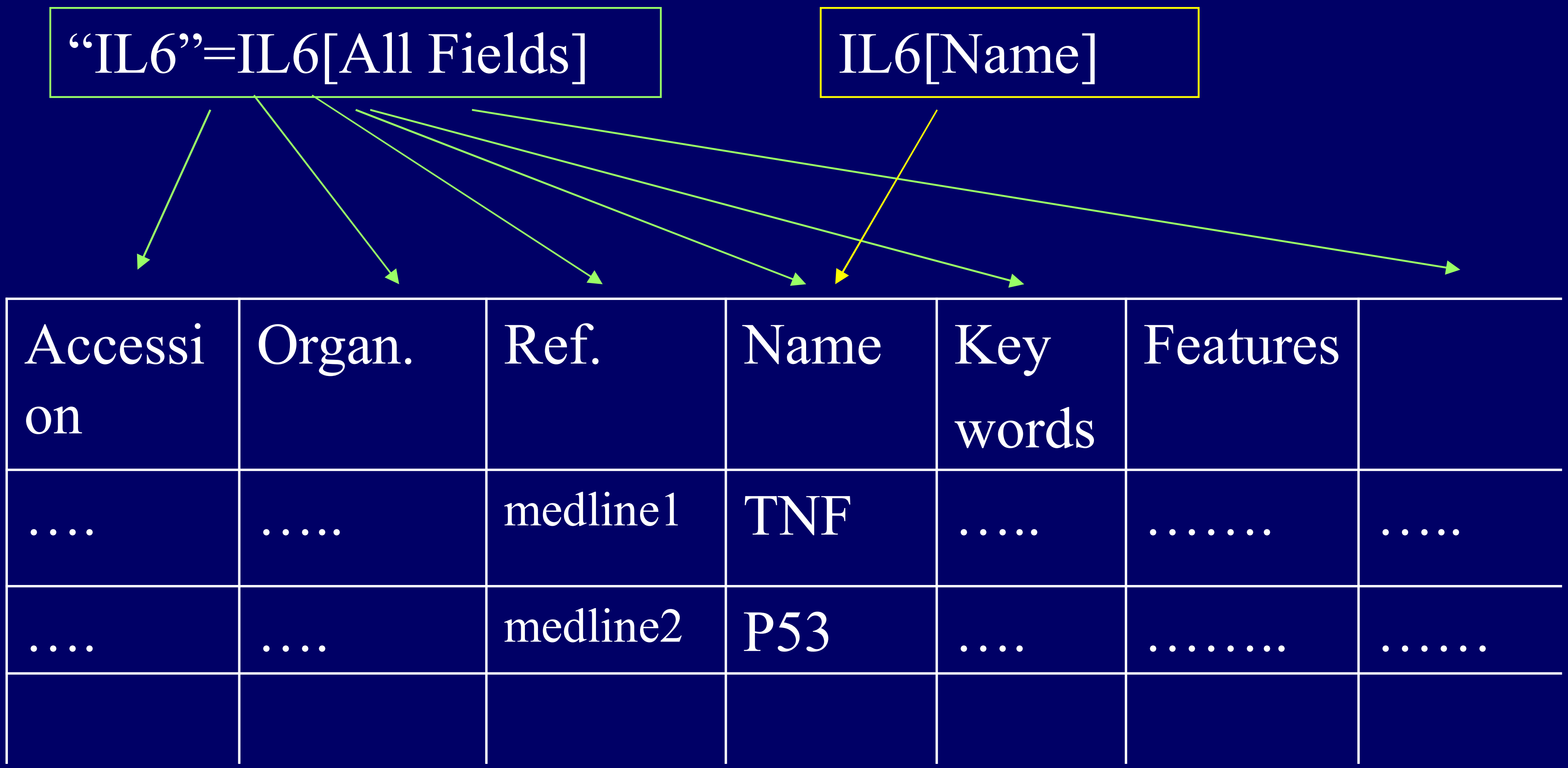
Search for IL6 (or your favorite gene) in the Gene database and the Proteins databases.

- why do we get so many hits?

Search for IL6 in the default “All Text” vs. search in the [Gene Name] field only in the Gene database.

Compare results.

Database concept – tables in relational databases



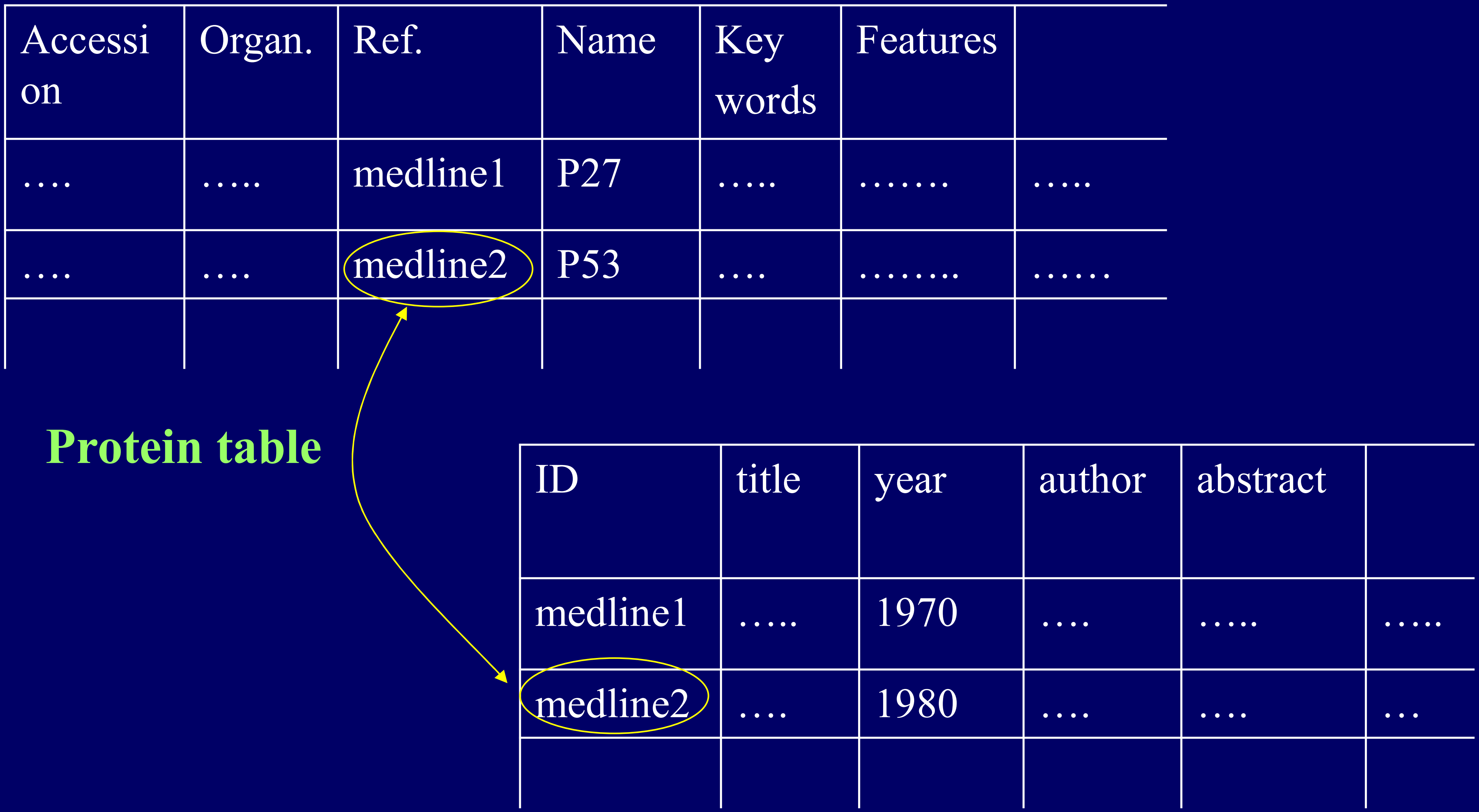
“IL6”=IL6[All Fields]

IL6[Name]

Accession	Organ.	Ref.	Name	Key words	Features	
….	…..	medline1	TNF	…..	…….	…..
….	….	medline2	P53	….	……..	……

Gene table

Database concept – relationship between tables allows linkage



Accession	Organ.	Ref.	Name	Key words	Features	
….	…..	medline1	P27	…..	…….	…..
….	….	medline2	P53	….	……..	……

Protein table

ID	title	year	author	abstract	
medline1	…..	1970	….	…..	…..
medline2	….	1980	….	….	…

Reference table

Observe/Practice

Observe the links from the IL6 **Gene** page:

- RefSeq
- OMIM
- SNP
- GEO
- Etc.

The "Gene" entry is the pivotal point for many NCBI resources.

Representation of genes and related information

The need to represent associated info with sequence

- Different aspects of the gene (such as protein, nucleotide, structure (PDB), OMIM etc.)
- Specialized databases (such GEO, SNP)
- Complex / customized data structure

  - Object-oriented data representation

Observe

Observe entries involving IL6 (or your gene) in Reactome.

Where and how much information are available for my gene?

Observe: The information contents and presentation format for the same gene in SwissProt, NCBI protein, NCBI Genes, etc..

Public Resources (II) – Analysis tools

- Web-based analysis tools – easy to use, but often with less customization options.
- Stand-alone analysis tools – requires installation and configuration, but provides more customizatio0n options.
- Commercial analysis tools
- Scripting for bioinformatics projects

web-based tools

- Identification of web-based bioinformatics resources.
  - Portals, lists,
  - Google search
- Organization
  - Book mark.
  - html page.

web-based tools

Practice –design QPCR primers for IL-6 or your gene.