Bioinformatics I
Biological Databases

Dr Manaf A Guma

University Of Anbar- college of Applied sciences-Hit

Department of applied chemistry

1

What are the secondary Databases?

- **1- Prosite database https://prosite.expasy.org**
- It is an important secondary biological database is Prosite [prosite].
- It resides at the SIB [expasy].
- Classification of proteins in Prosite is determined using single conserved motifs, i.e., short sequence regions (10–20 amino acids) that are conserved in related proteins and usually have a key role in the protein's function.
- The search for such sequence motifs in unknown proteins can provide a first hint of an affiliation to a protein family or function.
- A typical regular expression in Prosite would have the following form: [GSTNE]-[GSTQCR]-[FYW]-{ANW}-x(2)-P.

2

What are the secondary Databases?

- **2- PRINTS** **http://130.88.97.239/PRINTS/index.php**
- The PRINTS database [prints] uses fingerprints to classify sequences. Fingerprints consist of several sequence motifs, represented in the PRINTS database by short, local, ungapped alignments (talk about it later).
- The PRINTS database takes advantage of the fact that proteins usually contain functional regions that result in several sequence motifs per protein.

3

What are the secondary Databases?

- **3- Pfam database** **https://pfam.xfam.org/**
- The Pfam database [pfam] classifies protein families according to profiles.
- A profile is a pattern that evaluates the probability of the appearance of a given amino acid, an insertion, or a deletion at every position in a protein sequence.
- Pfam is based on sequence alignments.
- High-quality, manually checked alignments serve as starting points for the automatic construction of hidden Markov models (HMMs).

4

What are the secondary Databases?

- **4- Interpro database https://www.ebi.ac.uk/interpro/**
- The Integrated Resource of Protein Families, Domains and Sites (Interpro) [interpro] integrates important secondary databases into a comprehensive signature database.
- Interpro merges the databases Swissprot, TrEMBL, Prosite, Pfam, PRINTS, ProDom, Smart, and TIGRFAMs [tigr] and thereby allows a simple and simultaneous query of these databases.
- The result page combines the output of the individual queries.
- The Interpro Web server offers a few intuitive query facilities for text and sequence searches.

5

What are the secondary Databases?

- **5-Genotype-Phenotype Databases https://www.omim.org and https://www.ncbi.nlm.nih.gov/gap/phegeni**
- A number of genotype-phenotype databases have been established that record relationships between genes and the biological properties of organisms.
- The Online Mendelian Inheritance in Man (OMIM) https://www.omim.org database of the NCBI [omim] is perhaps the best-known genotype- phenotype database.
- For diseases to emerge and progress, several genes or their products are frequently required. The identification of genes relevant to disease is, therefore, of vital impor- tance in a target-based approach to rational drug development.
- The Online Mendelian Inheritance in Animals (OMIA) database [omia] at the NCBI also contains genotype-phenotype relationships of various animals, except mice and humans.
- For mice, the relevant database is in the Mouse Genome Database (MGD) [mgd].

6

What are the secondary Databases?

- **6- PhenomicDB** https://academic.oup.com/bioinformatics/article/21/3/418/237882
- The PhenomicDB database is a multiorganism genotype-phenotype database contain- ing data from humans and other important organisms such as the mouse, zebra fish (*Danio rerio*), fruit fly (*D. melanogaster*), nematode (*C. elegans*), baker’s yeast (*S. cere- visiae*), and cress plant (*Arabidopsis thaliana*).
- PhenomicDB integrates data from the aforementioned and other primary genotype-phenotype databases. A complete listing of all underlying data sources can be found on the home page [phenomicdb]

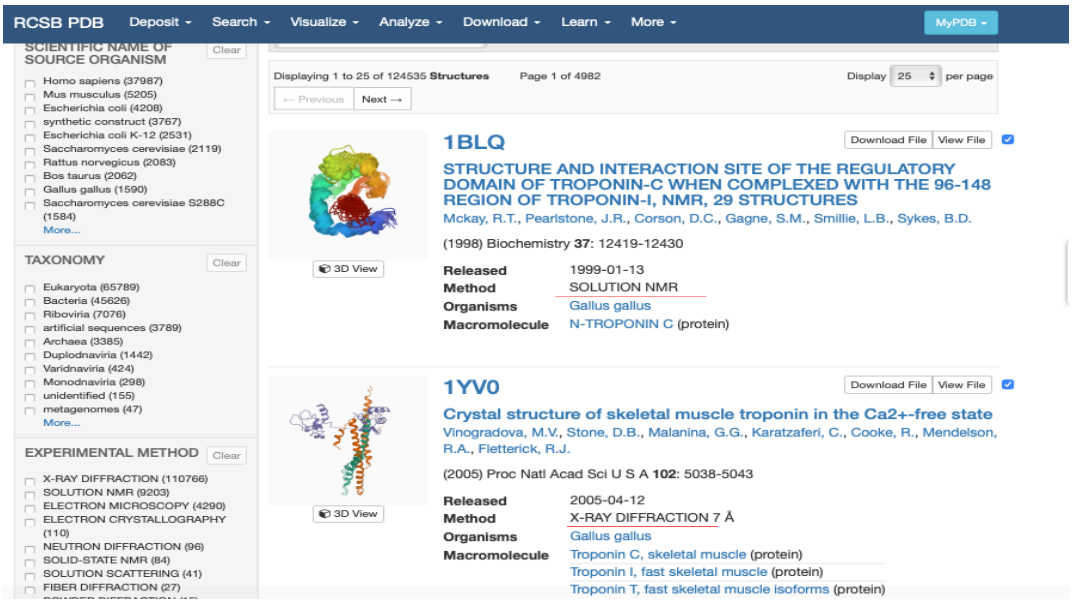
7

Molecular Structure Databases

- **Secondly: bank databases for proteins structures.**
- **1- Protein Data Bank** https://www.rcsb.org
- The PDB is a database of experimentally determined crystal structures of biological macromolecules and is coordinated by a consortium located in the USA, Europe, and Japan [wwpdb] (Berman et al. 2000).
- The PDB was founded at the Brookhaven National Laboratory in 1971, reflected in the frequent use of the name Brookhaven Protein Data Bank.
- These are predominantly proteins, but also include DNA and RNA structures and protein–nucleic acid complexes. (that were solved by X-ray, NMR and Cryo EM techniques)

8

What can you see in (RCSB) web? PDB



SCIENTIFIC NAME OF SOURCE ORGANISM

□ Homo sapiens (37987)
□ Mus musculus (5205)
□ Escherichia coli (4208)
□ synthetic construct (3767)
□ Escherichia coli K-12 (2531)
□ Saccharomyces cerevisiae (2119)
□ Rattus norvegicus (2083)
□ Bos taurus (2062)
□ Gallus gallus (1590)
□ Saccharomyces cerevisiae S288C (1584)

More...

TAXONOMY

□ Eukaryota (65789)
□ Bacteria (45626)
□ Riboviria (7076)
□ artificial sequences (3789)
□ Archaea (3385)
□ Duplodnaviria (1442)
□ Varidnaviria (424)
□ Monodnaviria (298)
□ unidentified (155)
□ metagenomes (47)

More...

EXPERIMENTAL METHOD

□ X-RAY DIFFRACTION (110766)
□ SOLUTION NMR (9203)
□ ELECTRON MICROSCOPY (4290)
□ ELECTRON CRYSTALLOGRAPHY (110)
□ NEUTRON DIFFRACTION (96)
□ SOLID-STATE NMR (84)
□ SOLUTION SCATTERING (41)
□ FIBER DIFFRACTION (27)

Displaying 1 to 25 of 124535 **Structures** Page 1 of 4982 Display 25 per page

1BLQ

STRUCTURE AND INTERACTION SITE OF THE REGULATORY DOMAIN OF TROPONIN-C WHEN COMPLEXED WITH THE 96-148 REGION OF TROPONIN-I, NMR, 29 STRUCTURES

Mckay, R.T., Pearlstone, J.R., Corson, D.C., Gagne, S.M., Smillie, L.B., Sykes, B.D.

(1998) Biochemistry **37**: 12419-12430

Released	1999-01-13
Method	SOLUTION NMR
Organisms	Gallus gallus
Macromolecule	N-TROPONIN C (protein)

1YV0

Crystal structure of skeletal muscle troponin in the Ca2+-free state

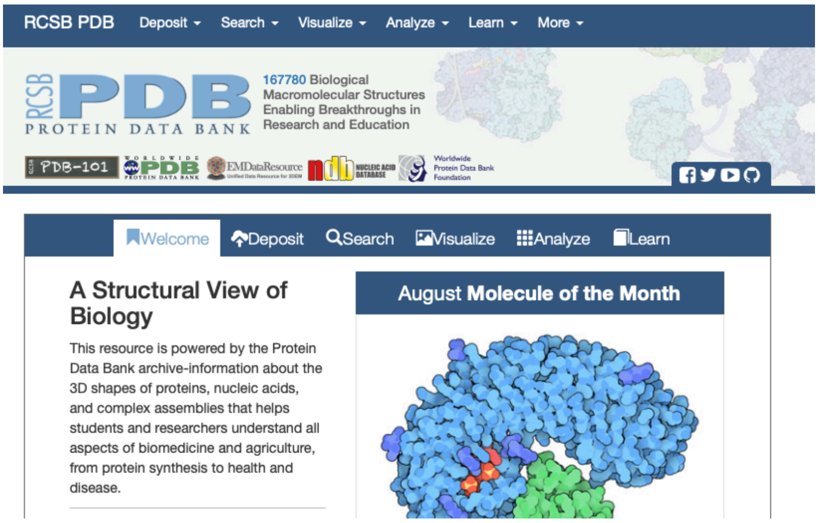
Vinogradova, M.V., Stone, D.B., Malanina, G.G., Karatzaferi, C., Cooke, R., Mendelson, R.A., Fletterick, R.J.

(2005) Proc Natl Acad Sci U S A **102**: 5038-5043

Released	2005-04-12
Method	X-RAY DIFFRACTION 7 Å
Organisms	Gallus gallus
Macromolecule	Troponin C, skeletal muscle (protein) Troponin I, fast skeletal muscle (protein) Troponin T, fast skeletal muscle isoforms (protein)

9

PDB web search



167780 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

August Molecule of the Month

10

Molecular Structure Databases

- **2-SCOP database http://scop.mrc-lmb.cam.ac.uk**
- Proteins that perform a similar biological function and are evolutionary related must have a similar structural organization, at least in the region of their active centres.
- It should, therefore, be possible to predict the function of an unknown protein by comparison of its structural organization with that of known proteins.
- Two databases, SCOP and CATH, provide such predictions. SCOP (Structural Classification Of Proteins) [scop] classifies proteins of a known structure in a hierarchical manner.
- The three main classifications are families, superfamilies, and folds.
- Families describe proteins with a clear evolutionary relationship to each othe.

11

Molecular Structure Databases

- **3- CATH database https://www.cathdb.info**
- The CATH database [cath] (Greene et al. 2007) classifies protein structures hierarchically into four categories: Class (C), Architecture (A), Topology (T), and Homologous Superfamily (H).
- The classification of proteins into the Class category is mainly automatic, but it can be complemented by manual intervention when required. In the Class category, the proportion of secondary structural elements is taken into account without consideration of their arrangement or connections.
- Four classes of proteins are distinguished: proteins composed mainly of helices (*mainly alpha*), sheets (*mainly beta*), both helices and sheets (*alpha-beta*), and, finally, proteins with very few secondary structural elements.

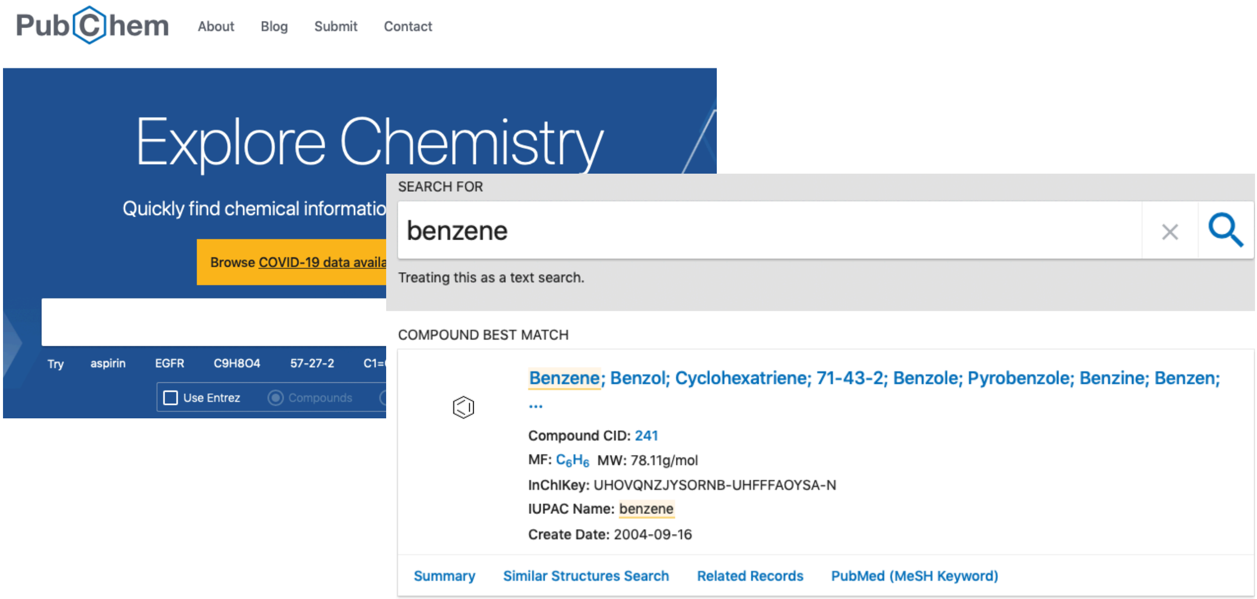
12

Molecular Structure Databases

- 4-**PubChem database** https://pubchem.ncbi.nlm.nih.gov
- The PubChem database at the NCBI [pubchem] stores small chemical molecules and information about their biological activities.
- It consists of three components, PubChem Compound, PubChem Substance, and PubChem BioAssay. PubChem Compound contains about 91 million molecules (July 2016) together with their two-dimensional (2D) molecular structures.
- The PubChem databases have multiple applications thanks to internal and external database linking, including to PubMed.
- For example, with a known enzyme inhibitor it is possible to find other similar potential inhibitors.

13

Molecular Structure Databases



Explore Chemistry

Quickly find chemical informatio

SEARCH FOR

benzene

Treating this as a text search.

COMPOUND BEST MATCH

Benzene; Benzol; Cyclohexatriene; 71-43-2; Benzole; Pyrobenzole; Benzine; Benzen; ...

Compound CID: 241

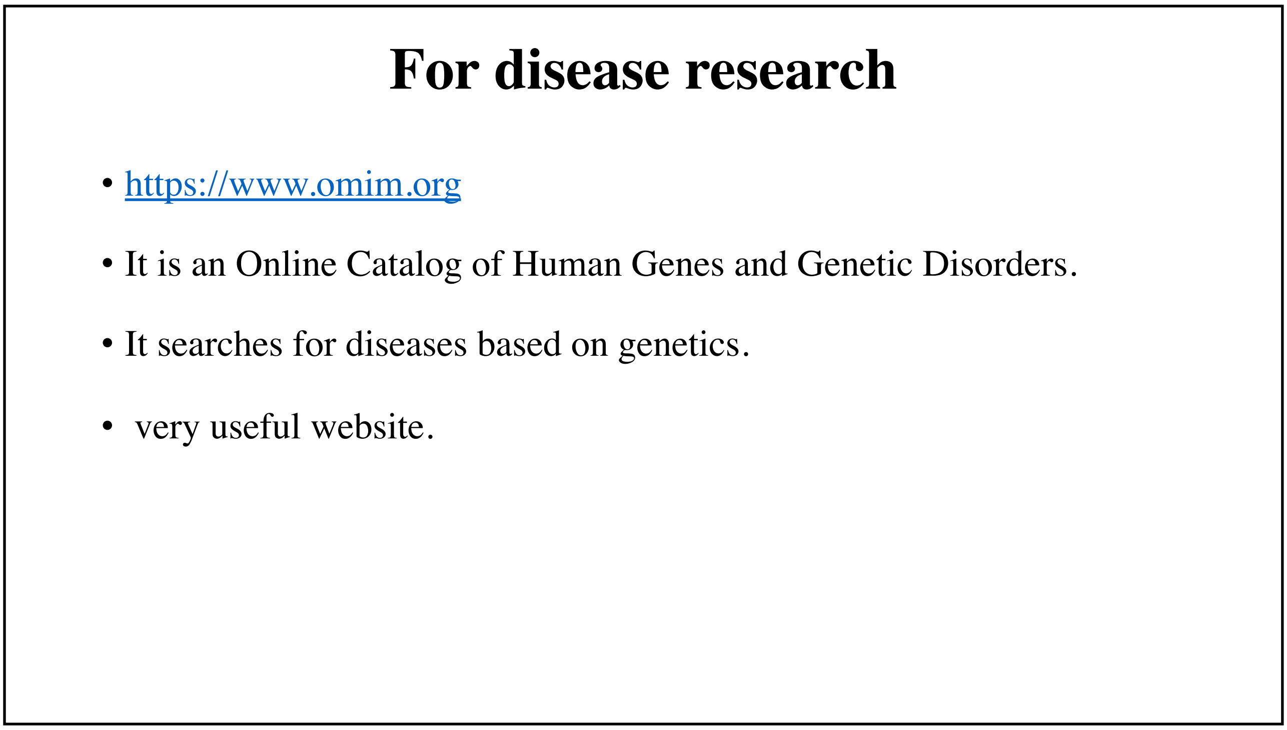
MF: C_6H_6 MW: 78.11g/mol

InChIKey: UHOVQNZJYSORNB-UHFFFAOYSA-N

IUPAC Name: benzene

Create Date: 2004-09-16

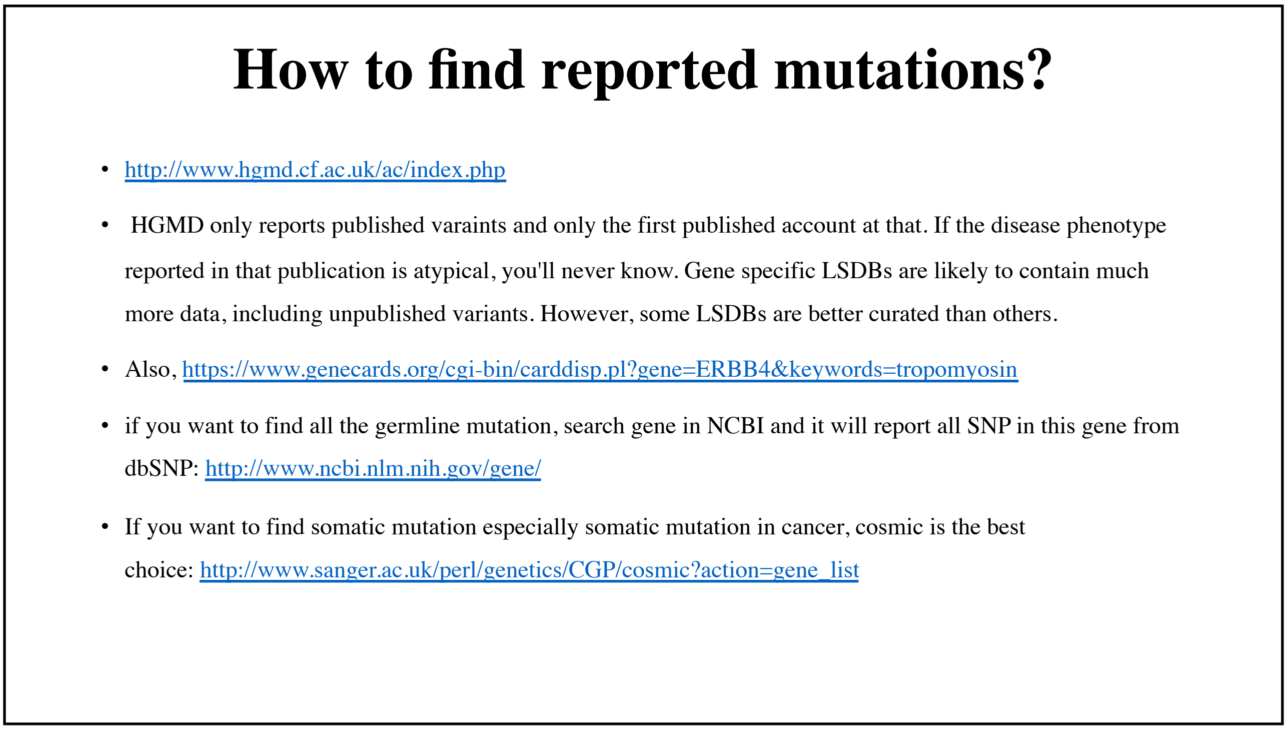
14



For disease research

- https://www.omim.org
- It is an Online Catalog of Human Genes and Genetic Disorders.
- It searches for diseases based on genetics.
- very useful website.

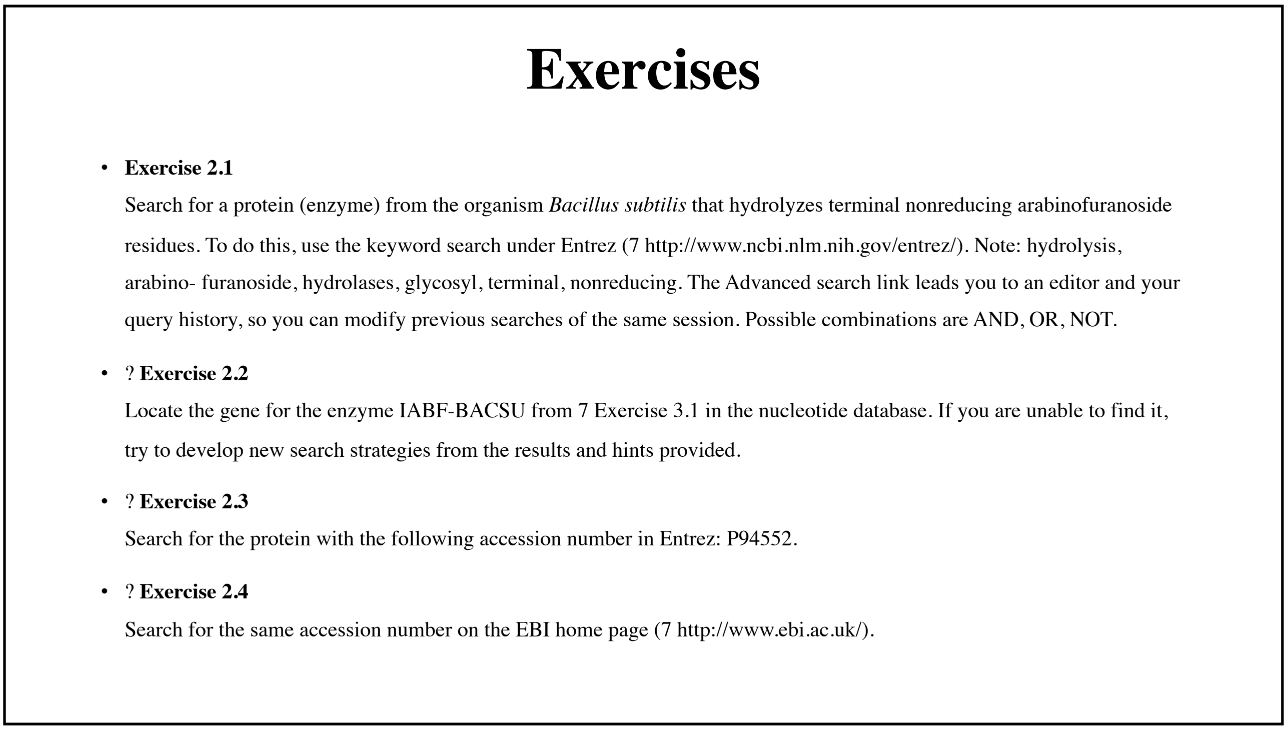
15



How to find reported mutations?

- http://www.hgmd.cf.ac.uk/ac/index.php
- HGMD only reports published varaints and only the first published account at that. If the disease phenotype reported in that publication is atypical, you'll never know. Gene specific LSDBs are likely to contain much more data, including unpublished variants. However, some LSDBs are better curated than others.
- Also, https://www.genecards.org/cgi-bin/carddisp.pl?gene=ERBB4&keywords=tropomyosin
- if you want to find all the germline mutation, search gene in NCBI and it will report all SNP in this gene from dbSNP: http://www.ncbi.nlm.nih.gov/gene/
- If you want to find somatic mutation especially somatic mutation in cancer, cosmic is the best choice: http://www.sanger.ac.uk/perl/genetics/CGP/cosmic?action=gene_list

16



Exercises

- **Exercise 2.1**

  Search for a protein (enzyme) from the organism *Bacillus subtilis* that hydrolyzes terminal nonreducing arabinofuranoside residues. To do this, use the keyword search under Entrez (7 http://www.ncbi.nlm.nih.gov/entrez/). Note: hydrolysis, arabino- furanoside, hydrolases, glycosyl, terminal, nonreducing. The Advanced search link leads you to an editor and your query history, so you can modify previous searches of the same session. Possible combinations are AND, OR, NOT.

- ? **Exercise 2.2**

  Locate the gene for the enzyme IABF-BACSU from 7 Exercise 3.1 in the nucleotide database. If you are unable to find it, try to develop new search strategies from the results and hints provided.

- ? **Exercise 2.3**

  Search for the protein with the following accession number in Entrez: P94552.

- ? **Exercise 2.4**

  Search for the same accession number on the EBI home page (7 http://www.ebi.ac.uk/).