EE550
Computational Biology

Week 1 Course Notes

Instructor: Bilge Karaçalı, PhD

Syllabus

Schedule **:** Tuesday 13:30, 14:30, 15:30

Text **:** Paul G. Higgs, Teresa K. Attwood, “Bioinformatics and Molecular Evolution,” Wiley-Blackwell, 2005

Instructor **:** Bilge Karaçalı, PhD

Office **:** EEE Building Room K1-32

E-mail **:** bilge@iyte.edu.tr

Summary **:**

This course will begin with a broad perspective of quantitative and high throughput biology. Computational methods for pattern detection and clustering will be introduced in the analysis of amino acid sequences of proteins. Probabilistic models of genetic evolution will be developed along with sequence alignment and motif detection algorithms. RNA and DNA analysis with microarrays will be discussed. Dynamic modelling of gene transcription networks will be introduced.

Grading **:**

Midterm	20%	Final	30%
Homework	20%	Project	30%

Course Outline:

Week 1: Introduction to computational biology

Week 2: Nucleic acid and protein structure

Week 3: Evolution mechanism through mutations

Week 4: Probabilistic amino acid sequence evolution models

Week 5: Gene and protein databases

Week 6: Sequence alignment

Week 7: Searching sequence databases

Week 8: Inter-species evolutionary relationships via phylogenetic trees

Week 9: Optimality criteria in phylogenetic tree construction

Week 10: Pattern searching in functional protein groups: Sequence motifs

Week 11: Bioinformatics

Week 12: Microarray data analysis

Week 13: Systems biology – Gene transcription networks

Week 14: Regulation of gene transcription

Topics

- Introduction to computational biology
  - Computation in life sciences
  - Quantitative and high throughput biology
  - Molecular biology on the internet
  - Bioinformatics

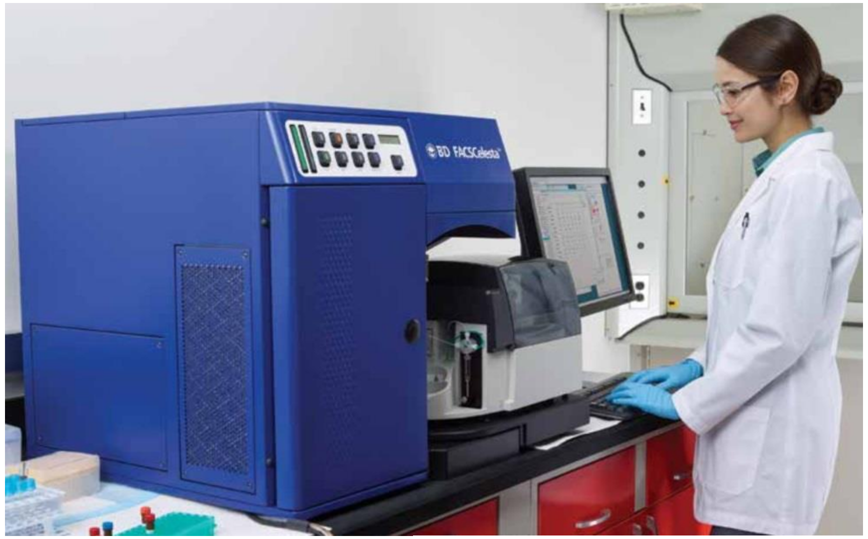
Computation and Life Sciences

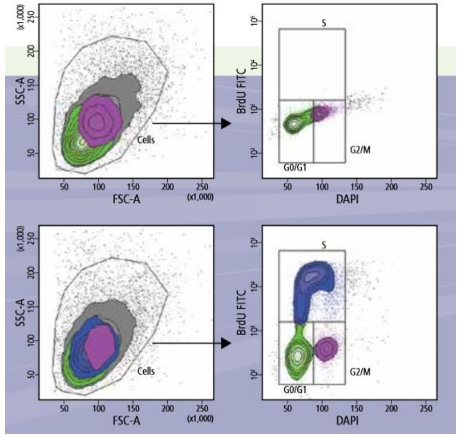
- Conventional life sciences research
  - Focus on a well-defined problem
  - Accumulate evidence using lab experiments
  - Construct a theory that makes predictions
  - Validate the predictions by additional lab experiments
- Accumulation of knowledge
  - over many years of empirical research
  - by contributions from many individuals working independently
  - on related areas of a unified framework

Source: https://www.jllrealviews.com/industries/life-sciences/flexible-space-major-breakthrough-lab-design/

Computation and Life Sciences

- Technological innovations in life sciences research
  - Development of biosensor technologies
    - Medical imaging systems
    - Immunohistochemical staining
    - Electrophysiological monitoring
    - …
  - Introduction of high-throughput techniques
    - DNA microarrays
    - Flow cytometry
    - …





Source: http://static.bdbiosciences.com/documents/BD-FACSCelesta-Brochure.pdf

Computation and Life Sciences

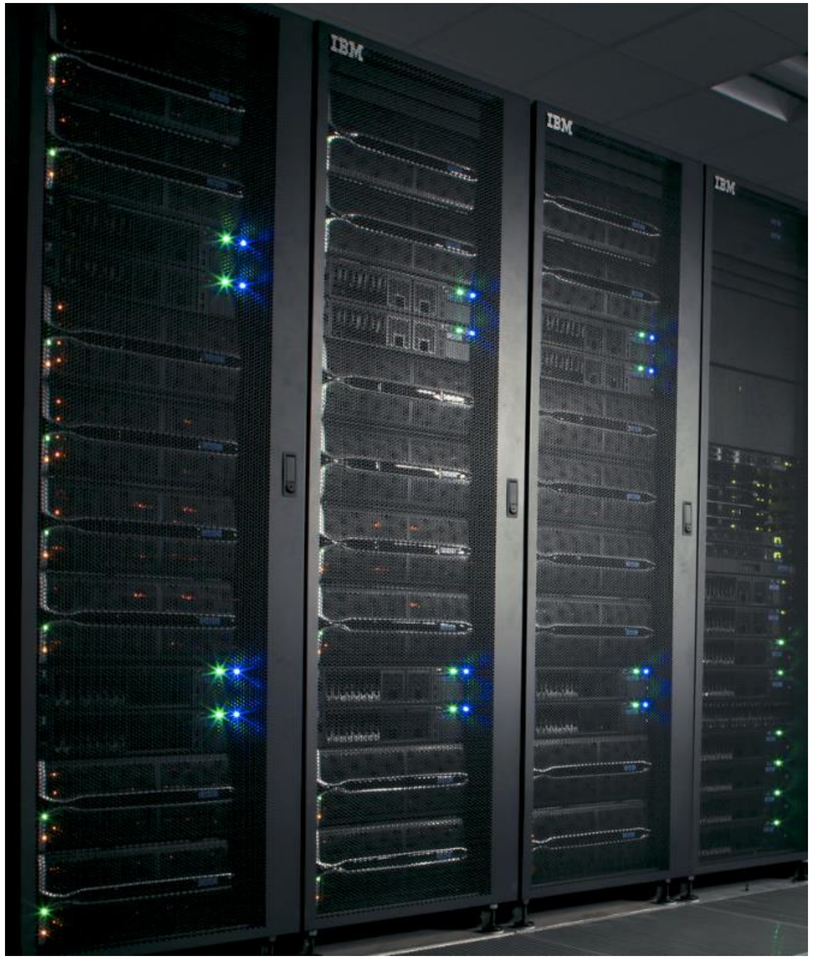
- The result:

  Increasing difficulty in analyzing rapidly accumulating biological data using conventional manual techniques

  - Time
  - Quantitation
  - Standardization
  - Dimensionality

Computation and Life Sciences

- Computational methods in engineering sciences
  - Run on computer hardware
  - Operate on high-dimensional numeric data
  - Make statistically viable inferences on the problem at hand
- Detailed **quantitative** analysis of biological data using computational methods
  - Minimal operating cost
  - Adequate scaling with increasing data and data dimensions
  - Exhaustive joint evaluation of all available evidence
  - Statistically significant deductions and predictions



Source: https://www-03.ibm.com/press/us/en/photos.wss?topic=1

Computation and Life Sciences

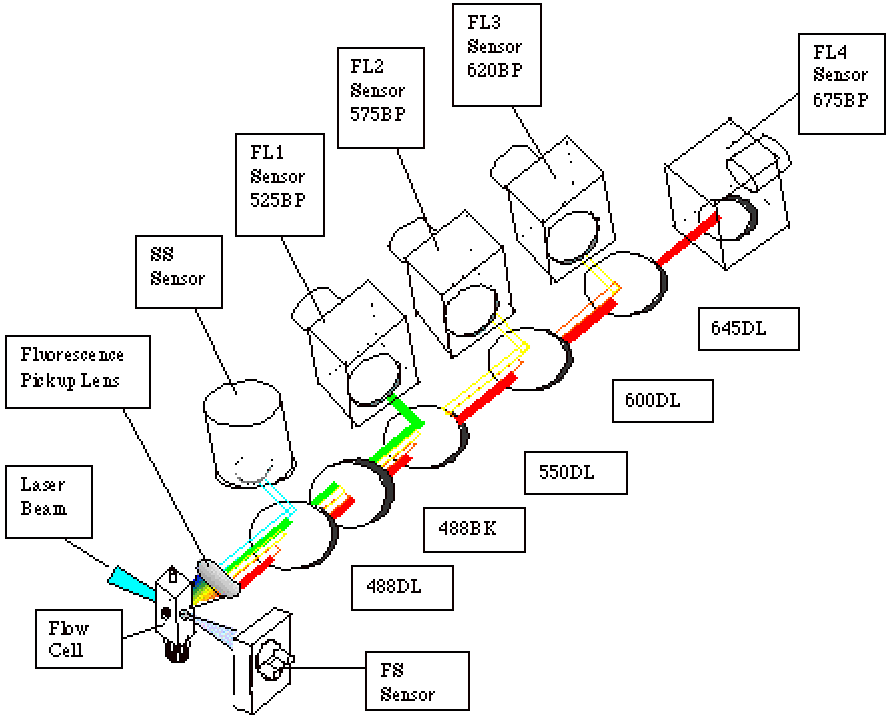
- Two main avenues of contribution
  - Replicating human expert decisions
    - cost of personnel training to carry out the required analysis
      - money, time and effort
    - cost of the trained expert personnel as they carry out the required analysis
      - money, time and effort
  - Performing analyses that are impossible by conventional manual analysis
    - joint analysis of multiple parameters

Quantitative and High-Throughput Biology

- Nomenclature
  - Quantitative: represented by numeric measurements
  - High-throughput: many measurements at once
  - Biology: knowledge of life processes
- Quantitative and high-throughput data in biology
  - Gene and protein sequence analysis
  - Gene expressions in DNA microarrays
  - Cell expressions in multi-color flow cytometry experiments
  - …

Case in point: Multi-Color Flow Cytometry

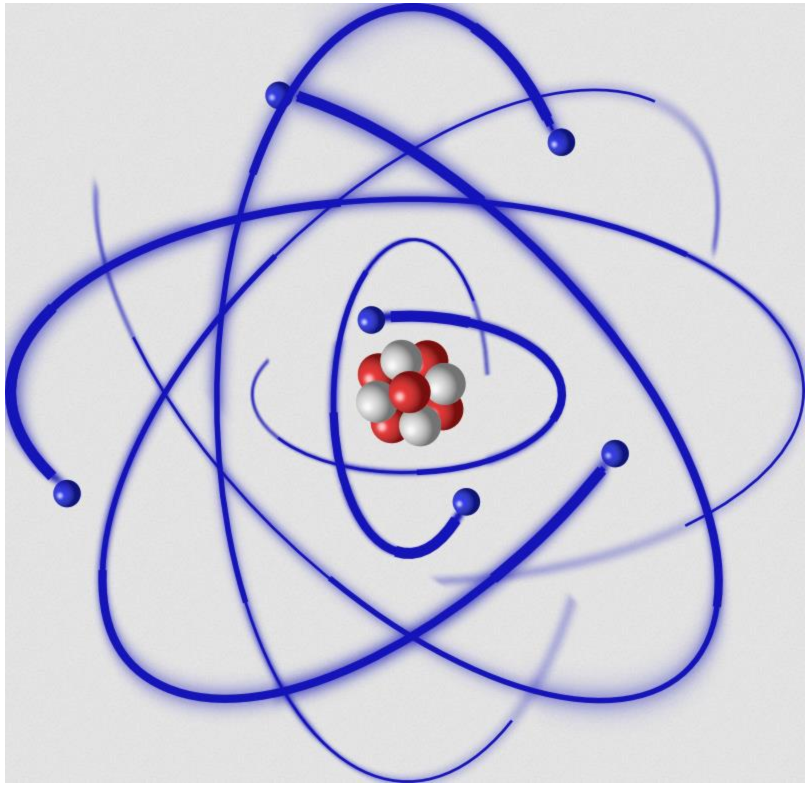
- Provides a multivariate profile for each cell in an emulsion
  - Morphological (structural) parameters:
    - Forward scatter – related to cell size
    - Side scatter – related to granularity and surface curvature
  - Fluorescence (functional) parameters:
    - Multiple intensity parameters indicating amounts of molecular markers in cell cytoplasm or membrane
- Evaluates several hundred cells per second
  - Thousands and thousands of high-dimensional feature vectors to be analyzed
- Computational data analysis required for
  - Computing the percentages of known cell subsets
  - Identifying new cell subsets
  - Comparing cell distributions across individuals and populations



Source: http://biology.berkeley.edu/crl/flow_cytometry_basic.html

Gene and Protein Sequence Data

- Molecular biology is based on the chemistry of carbon
  - All molecules of biological significance are made of carbon atoms
    - DNA and RNA
    - Proteins
    - Carbohydrates
    - Lipids
  - The versatility and variability of carbon allows it to form complex molecules of all shapes and sizes



Source: http://myriverside.sd43.bc.ca/annie-rosep-2013/2014/10/20/atom-stories/

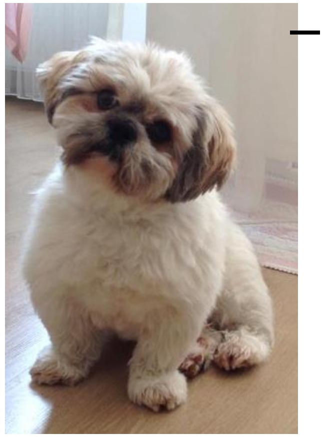
Gene and Protein Sequence Data

- Genes and proteins are **sequential polymers of carbon-based molecules** with
  - specific building instructions and
  - properties
- Computational analysis of gene and protein sequence data aims at inferring
  - the relationships between **sequence** and **function**
  - the evolutionary history

Gene Sequence Data

- The genome in every organism is encoded by the DNA molecule
  - Each cell has a full copy of the organism’s genome
  - In eukaryotes, the DNA is tightly packed inside the nucleus
- The DNA is a very long polymer chain made up of 4 types of nucleotides
  - Adenine
  - Guanine
  - Cytosine
  - Thymine

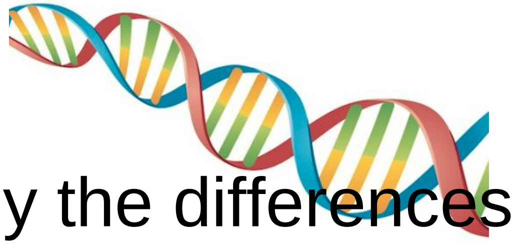
- The arrangement of these nucleotides in a sequence encodes the genetic information of the organism
  - Suppose each species comes with its own set of DNA (more or less)
  - For a DNA segment of 500 nucleotides, there are $4^{500} \approx 10^{301}$ different possibilities
  - This number is more than enough to account for all different species that ever were and will ever be on earth for the foreseeable future



Gene Sequence Data

- Genes are special segments of the DNA molecule
  - Not all regions on a DNA molecule are "transcribed" during the organism's life cycle

    "transcribed" → "used in the synthesis of proteins"
  - Genes are the regions that are transcribed



- Differences between organisms are determined by the differences in their genes
  - Genes are **polymer chains of nucleotides** just like the full DNA
    - A gene represents a specific sequence of nucleotides
  - Differences between the nucleotide sequences of two genes accounts for the differences in the function of the associated proteins as well as their distant common heritage
- The **genome** of an organism refers to the collection of all its genes
  - **Human genome** is about 1% of the total human DNA

Gene Sequence Data

- Computational analysis of gene sequence data deals with
  - their similarities across different species
  - their evolutionary relationships
  - their variations under mutations
  - their variations in different disease and risk conditions
  - …
- The analysis methods are based on
  - Probability and statistics
  - Discrete mathematics
  - Graph theory
  - …

```
TCACCTATGCCGGAGAGTCAGCGATTATG
AAGTCGGTGCCGGAGTAGCGAACTCGGGA
TTGATTAGGCAGACGGGGGTTAAACGCCC
GTGTGGTTTACCCGCAATTGGGTCGACCG
TCTTAGGCGACCGTTGATTTGCGACTGGT
TTGGATTGGGGCGCGTAACGTTTCCTCAC
CCTTGTTCGACGAAGAAAAAGAATGGGTG
GAGGGAGGGGGCCGTCATACCTGAGTGGT
GTAGTCTCCGAGCGAAATCCGAGTGCTTC
CAGACACAACAACGCTTGGGGGGGACTAG
TAAAGACGGTTACATATCGCGCGCTTTAG
CCTTGTGTATGCCGGGACGTGGAGGTACG
CGGTCATGAGCTGTGATTTGAATGGGCAT
TGCTCTGGGGCCAAATCATGGTTCTCACA
GAAGGTAATGTATAAAGCCGCAAACGTAG
ACCAACCTTTGGACGAGCTTTGGTCGCGC
AGCTACCAGAAGAATCAGCACTCACTTGT
GTTCCACGCGAACACTAGCTTCTATTGAA
CGTAAGTCTTAGTACGAACACCGGGGCGT
CTTCTTGTGAGTTTGCCGGCTTAGCCCAT
```

Protein Sequence Data

- Proteins are also **polymer chains**, but **of amino acids**
  - 20 naturally occurring **amino acids**
  - Amino acids joined together by peptide bonds form **polypeptides**
  - One or more polypeptides coming together as complexes form **proteins**; stable molecules with specific function
- Protein structure is dictated primarily by its **amino acid sequence**
  - Primary structure refers to the sequence itself
  - Secondary and higher structures refer to non-covalent interactions between non-neighboring amino acid residues
- Variations in the amino acid sequence is responsible for variations in the protein’s biochemical properties
  - Three-dimensional conformation
  - Molecular interactions and binding
  - …

Accumulation of Gene and Protein Data

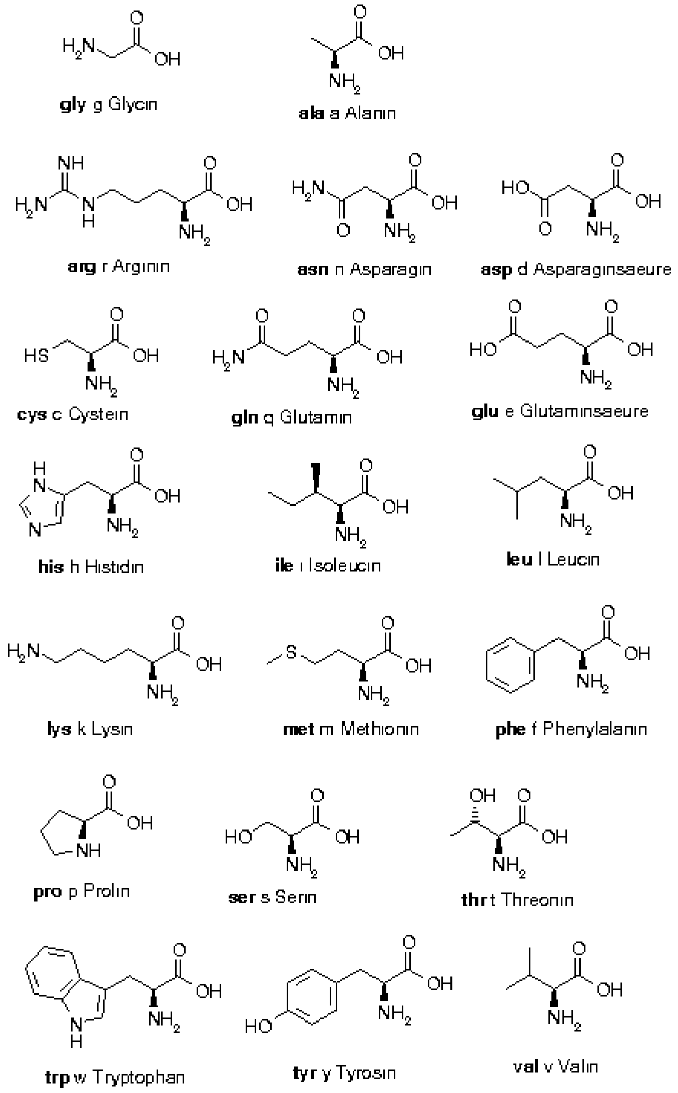
- Human Genome Project
  - An international project funded by the US Department of Energy and the US National Institutes of Health
  - Primary goal of sequencing 25,000 human genes
    - Some ongoing discussion on the total number
  - Project start in 1990, first working draft in 2000, completion in 2003
  - All sequences made publicly available on the internet
- A parallel project launched in 1998 by Celera Genomics
  - Used a different sequencing technique
  - Completed earlier and with less cost
  - Incorporated the Human Genome Project's findings into their research
  - But denied public access to their own findings
    - Tried to patent the gene sequences
    - The effort fell apart when genetic sequences were declared to be unpatentable
- The experience prompted universal agreement in publishing all genetic data on the internet for free
  - Created a utopian environment where researchers could make use of each other's published data with full agreement that others would be able to use theirs

Protein Sequence Data

- Amino acids

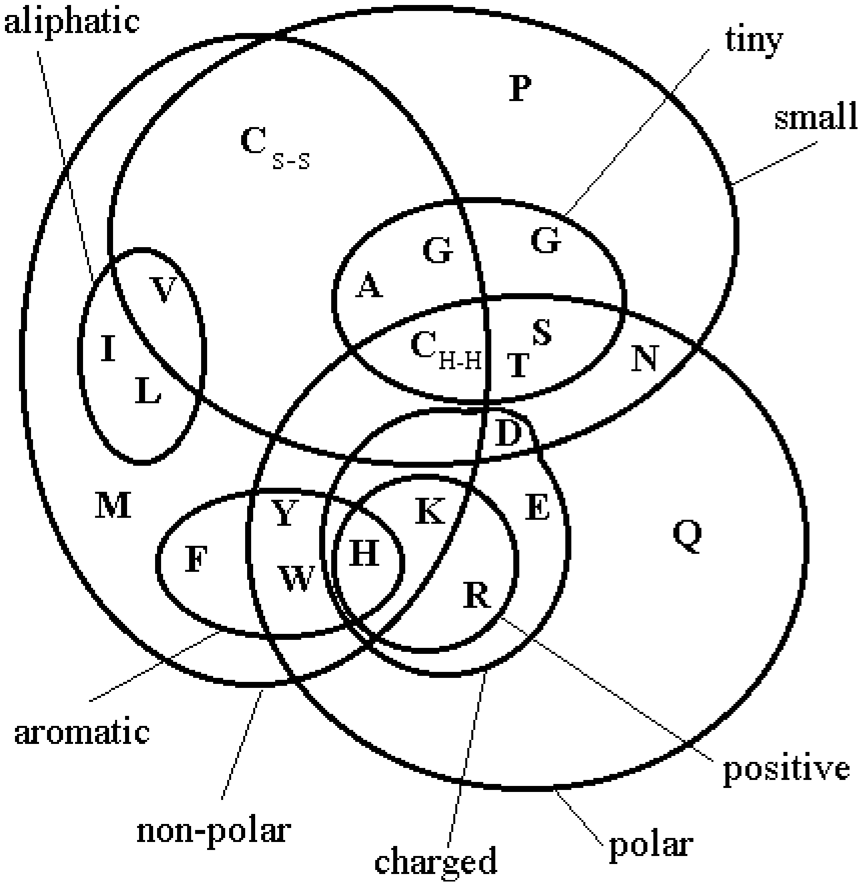
Alanine	ala a	CH3-CH(NH2)-COOH
Arginine	arg r	HN=C(NH2)-NH-(CH2)3-CH(NH2)-COOH
Asparagine	asn n	H2N-CO-CH2-CH(NH2)-COOH
Aspartic acid	asp d	HOOC-CH2-CH(NH2)-COOH
Cysteine	cys c	HS-CH2-CH(NH2)-COOH
Glutamine	gln q	H2N-CO-(CH2)2-CH(NH2)-COOH
Glutamic acid	glu e	HOOC-(CH2)2-CH(NH2)-COOH
Glycine	gly g	NH2-CH2-COOH
Histidine	his h	NH-CH=N-CH=C-CH2-CH(NH2)-COOH
Isoleucine	ile i	CH3-CH2-CH(CH3)-CH(NH2)-COOH
Leucine	leu l	(CH3)2-CH-CH2-CH(NH2)-COOH
Lysine	lys k	H2N-(CH2)4-CH(NH2)-COOH
Methionine	met m	CH3-S-(CH2)2-CH(NH2)-COOH
Phenylalanine	phe f	Ph-CH2-CH(NH2)-COOH
Proline	pro p	NH-(CH2)3-CH-COOH
Serine	ser s	HO-CH2-CH(NH2)-COOH
Threonine	thr t	CH3-CH(OH)-CH(NH2)-COOH
Tryptophan	trp w	Ph-NH-CH=C-CH2-CH(NH2)-COOH
Tyrosine	tyr y	HO-p-Ph-CH2-CH(NH2)-COOH
Valine	val v	(CH3)2-CH-CH(NH2)-COOH

Source: http://www.chemie.fu-berlin.de/chemistry/bio/amino-acids_en.html



Protein Sequence Data

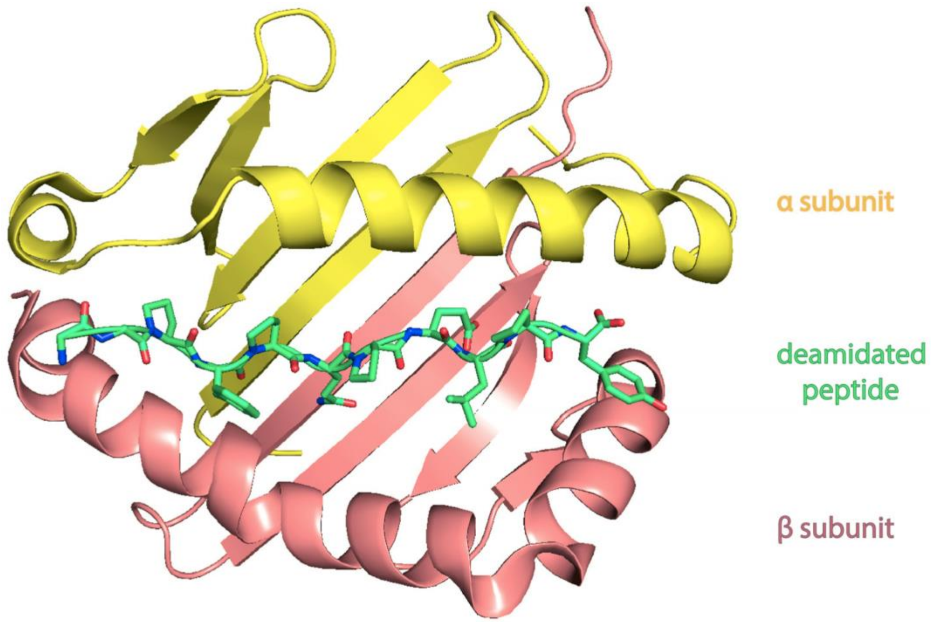
- Physicochemical properties of amino acids
  - Residue volume
  - Surface area
  - Side chain acidity
  - Solubility
  - Crystal density
  - Isoelectric point at 25°C
  - Hydrophobicity
  - Polarity
  - Aromaticity
  - Aliphaticity
  - Charge
  - …



Source: http://prowl.rockefeller.edu/aainfo/pchem.htm

Protein Sequence Data

- Computational analysis of protein sequence data deals with
  - Homologies across species
  - Sequence similarities across different proteins and protein groups
  - Conserved sequence motifs
  - Prediction of the protein function
  - Prediction of protein-protein interactions
  - …
- The analysis methods are based on
  - Probability and statistics
  - Graph theory
  - Discrete mathematics
  - …



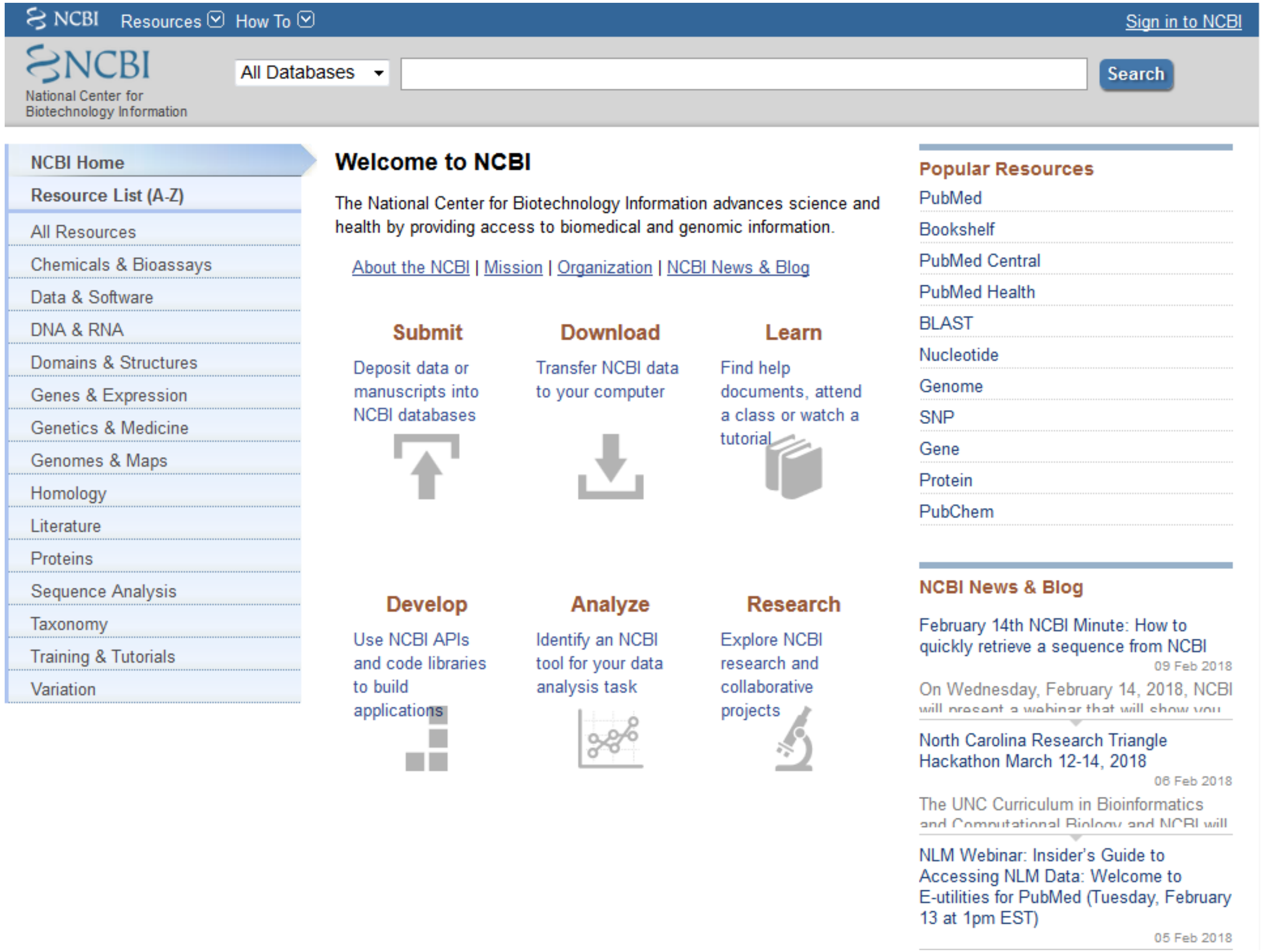
Gluten

Source: Anastasia V. Balakireva and Andrey A. Zamyatnin, "Properties of Gluten Intolerance: Gluten Structure, Evolution, Pathogenicity and Detoxification Capabilities," Nutrients, 8(10), 644, 2016

Gene and Protein Data on the Internet

- European Molecular Biology Laboratory nucleotide sequence database
  - URL http://www.ebi.ac.uk/embl/
  - Maintained by the European Bioinformatics Institute (UK)
- GenBank
  - URL http://www.ncbi.nlm.nih.gov/Database/
  - Maintained by the National Center of Biotechnology Information (USA)
- DNA Databank of Japan
  - URL http://www.ddbj.nig.ac.jp/
  - Maintained by the National Institute of Genetics (Japan)
- KEGG
  - URL http://www.genome.jp/kegg/
  - Maintained by the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo
- UniProt
  - URL http://www.uniprot.org/
  - Maintained by the European Bioinformatics Institute, the Swiss Institute of Bioinformatics, and the Protein Information Resource
- PDB
  - URL http://www.rcsb.org/pdb/home/home.do
  - Maintained by Rutgers, The State University of New Jersey and the University of California, San Diego
- …

Example: GenBank



Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

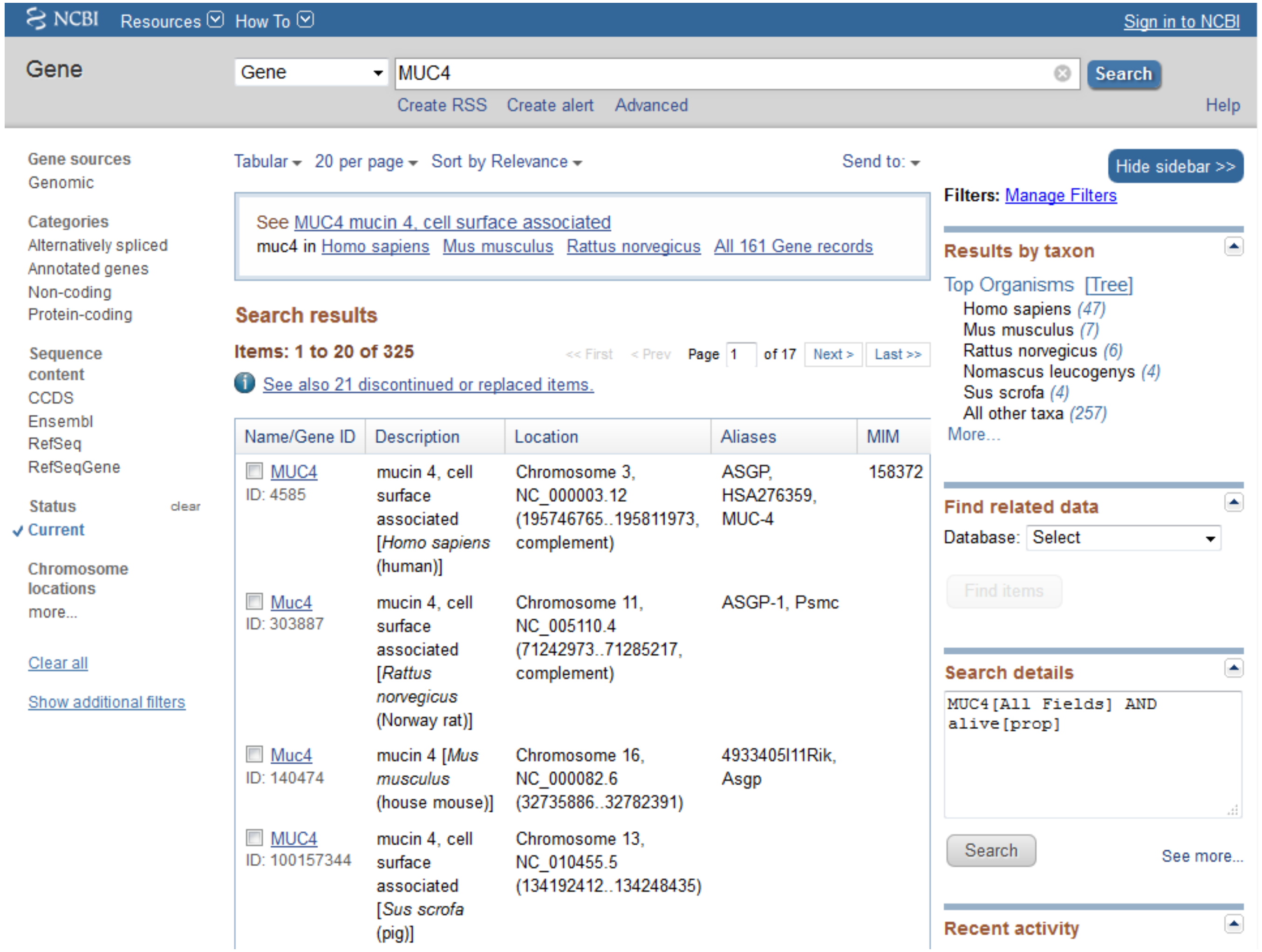
NCBI News & Blog

February 14th NCBI Minute: How to quickly retrieve a sequence from NCBI
09 Feb 2018
On Wednesday, February 14, 2018, NCBI will present a webinar that will show you

North Carolina Research Triangle Hackathon March 12-14, 2018
06 Feb 2018
The UNC Curriculum in Bioinformatics and Computational Biology and NCBI will

NLM Webinar: Insider's Guide to Accessing NLM Data: Welcome to E-utilities for PubMed (Tuesday, February 13 at 1pm EST)
05 Feb 2018

Example: GenBank



Gene sources
Genomic

Categories
Alternatively spliced
Annotated genes
Non-coding
Protein-coding

Sequence content
CCDS
Ensembl
RefSeq
RefSeqGene

Status clear
✓ Current

Chromosome locations
more...

Clear all

Show additional filters

See MUC4 mucin 4, cell surface associated
muc4 in Homo sapiens Mus musculus Rattus norvegicus All 161 Gene records

Search results

Items: 1 to 20 of 325

See also 21 discontinued or replaced items.

Name/Gene ID	Description	Location	Aliases	MIM
□ MUC4 ID: 4585	mucin 4, cell surface associated [*Homo sapiens* (human)]	Chromosome 3, NC_000003.12 (195746765..195811973, complement)	ASGP, HSA276359, MUC-4	158372
□ Muc4 ID: 303887	mucin 4, cell surface associated [*Rattus norvegicus* (Norway rat)]	Chromosome 11, NC_005110.4 (71242973..71285217, complement)	ASGP-1, Psmc	
□ Muc4 ID: 140474	mucin 4 [*Mus musculus* (house mouse)]	Chromosome 16, NC_000082.6 (32735886..32782391)	4933405I11Rik, Asgp	
□ MUC4 ID: 100157344	mucin 4, cell surface associated [*Sus scrofa* (pig)]	Chromosome 13, NC_010455.5 (134192412..134248435)		

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]

Homo sapiens	(47)
Mus musculus	(7)
Rattus norvegicus	(6)
Nomascus leucogenys	(4)
Sus scrofa	(4)
All other taxa	(257)

More...

Find related data

Database: Select

Find items

Search details

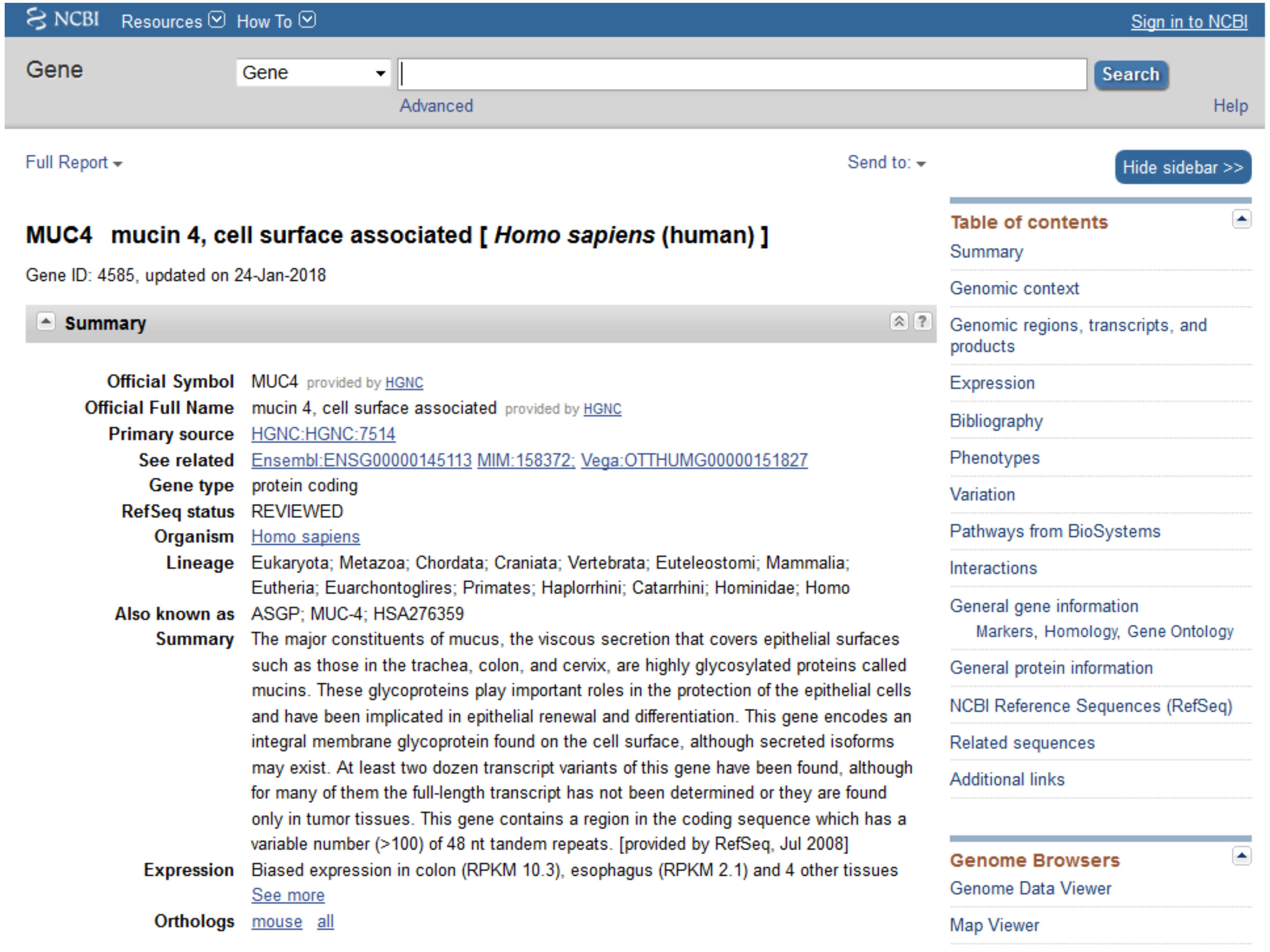
```
MUC4[All Fields] AND
alive[prop]
```

Search

See more...

Recent activity

Example: GenBank



MUC4 mucin 4, cell surface associated [*Homo sapiens* (human)]

Gene ID: 4585, updated on 24-Jan-2018

Summary

Official Symbol	MUC4 provided by HGNC
Official Full Name	mucin 4, cell surface associated provided by HGNC
Primary source	HGNC:HGNC:7514
See related	Ensembl:ENSG00000145113 MIM:158372; Vega:OTTHUMG00000151827
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	ASGP; MUC-4; HSA276359
Summary	The major constituents of mucus, the viscous secretion that covers epithelial surfaces such as those in the trachea, colon, and cervix, are highly glycosylated proteins called mucins. These glycoproteins play important roles in the protection of the epithelial cells and have been implicated in epithelial renewal and differentiation. This gene encodes an integral membrane glycoprotein found on the cell surface, although secreted isoforms may exist. At least two dozen transcript variants of this gene have been found, although for many of them the full-length transcript has not been determined or they are found only in tumor tissues. This gene contains a region in the coding sequence which has a variable number (>100) of 48 nt tandem repeats. [provided by RefSeq, Jul 2008]
Expression	Biased expression in colon (RPKM 10.3), esophagus (RPKM 2.1) and 4 other tissues See more
Orthologs	mouse all

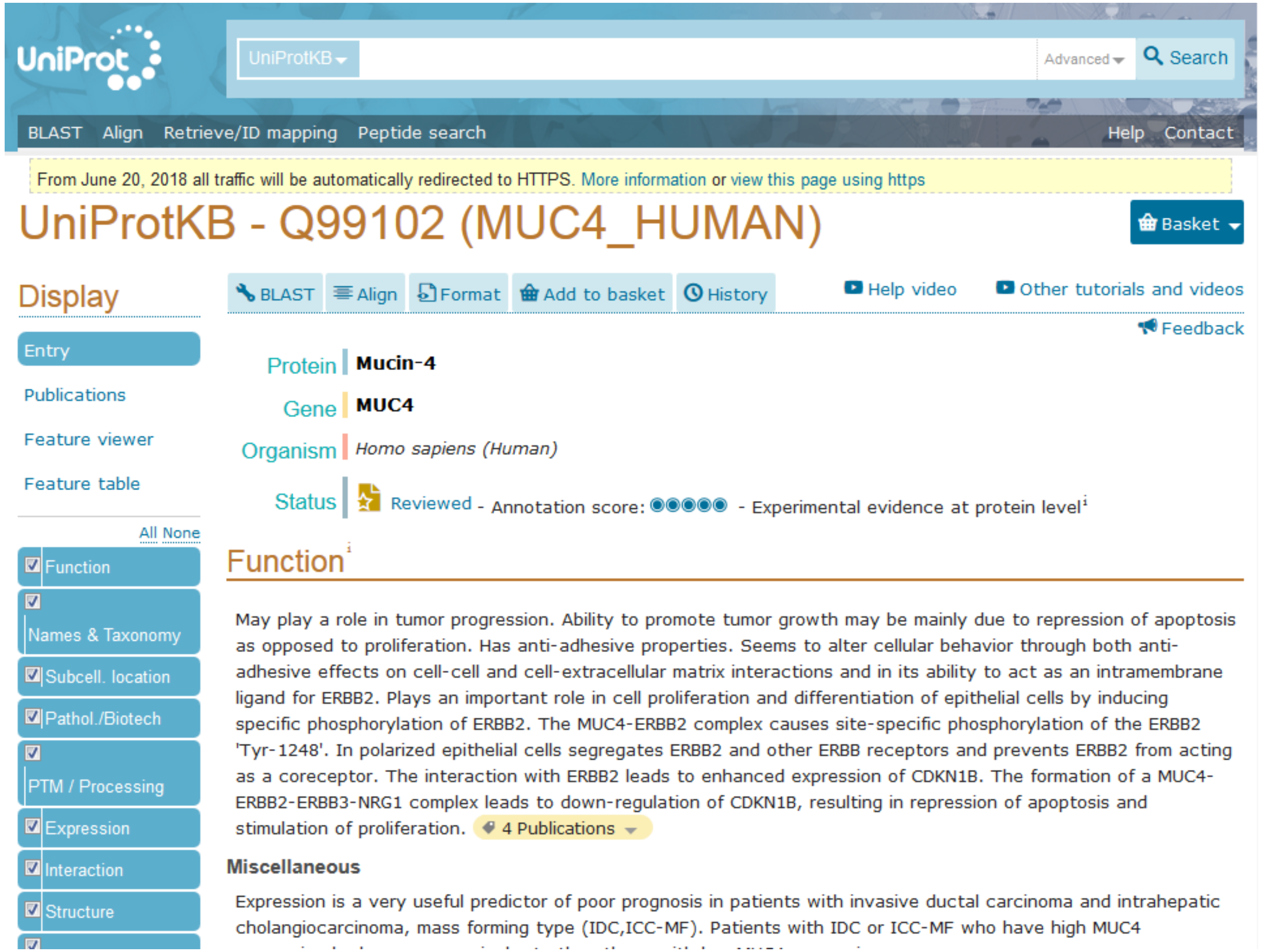
Table of contents

Genome Browsers

Genome Data Viewer

Map Viewer

Example: UniProt



From June 20, 2018 all traffic will be automatically redirected to HTTPS. More information or view this page using https

UniProtKB - Q99102 (MUC4_HUMAN)

Protein | **Mucin-4**

Gene | **MUC4**

Organism | *Homo sapiens (Human)*

Status | Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level[i]

Function[i]

May play a role in tumor progression. Ability to promote tumor growth may be mainly due to repression of apoptosis as opposed to proliferation. Has anti-adhesive properties. Seems to alter cellular behavior through both anti-adhesive effects on cell-cell and cell-extracellular matrix interactions and in its ability to act as an intramembrane ligand for ERBB2. Plays an important role in cell proliferation and differentiation of epithelial cells by inducing specific phosphorylation of ERBB2. The MUC4-ERBB2 complex causes site-specific phosphorylation of the ERBB2 'Tyr-1248'. In polarized epithelial cells segregates ERBB2 and other ERBB receptors and prevents ERBB2 from acting as a coreceptor. The interaction with ERBB2 leads to enhanced expression of CDKN1B. The formation of a MUC4-ERBB2-ERBB3-NRG1 complex leads to down-regulation of CDKN1B, resulting in repression of apoptosis and stimulation of proliferation. 4 Publications

Miscellaneous

Expression is a very useful predictor of poor prognosis in patients with invasive ductal carcinoma and intrahepatic cholangiocarcinoma, mass forming type (IDC,ICC-MF). Patients with IDC or ICC-MF who have high MUC4

Bioinformatics

- The analysis of quantitative biological data using computational and statistical methods
  - Data collection
  - Modeling
  - Method development
  - Validation
  - New hypothesis generation
- Requires computers
  - Experiments *in silico*
  - Fast and efficient implementation of mathematical algorithms
  - Statistical significance of results
- Molecular biological context indispensable
  - Evolutionary foundations
  - Pharmaceutical (smart drug) research and pharmacogenomics

Systems Biology

- Mathematical modelling of interactions between groups of genes and/or biomolecules
  - interaction between genes → gene transcription networks
  - interaction between biomolecules → signaling networks
  - interaction between genes and biomolecules → hybrid networks
- Requires computers
  - Experiments *in silico*
  - Fast and efficient implementation of mathematical algorithms
  - Statistical significance of results
- Molecular biological context indispensable
  - Evolutionary foundations
  - Pharmaceutical (smart drug) research and pharmacogenomics

Python Resources for Computational Biology

- Biopython
  - https://biopython.org/
- Biotite
  - https://github.com/biotite-dev/biotite
- General purpose python resources
  - https://guides.library.cmu.edu/bioinfo/r-and-python
- Google’s Colab environment
  - https://colab.research.google.com/

Summary

- Biomolecular data explosion necessitates analysis using computational methods
- These methods are developed using statistical and mathematical principles
- Online databases provide all the data in the world