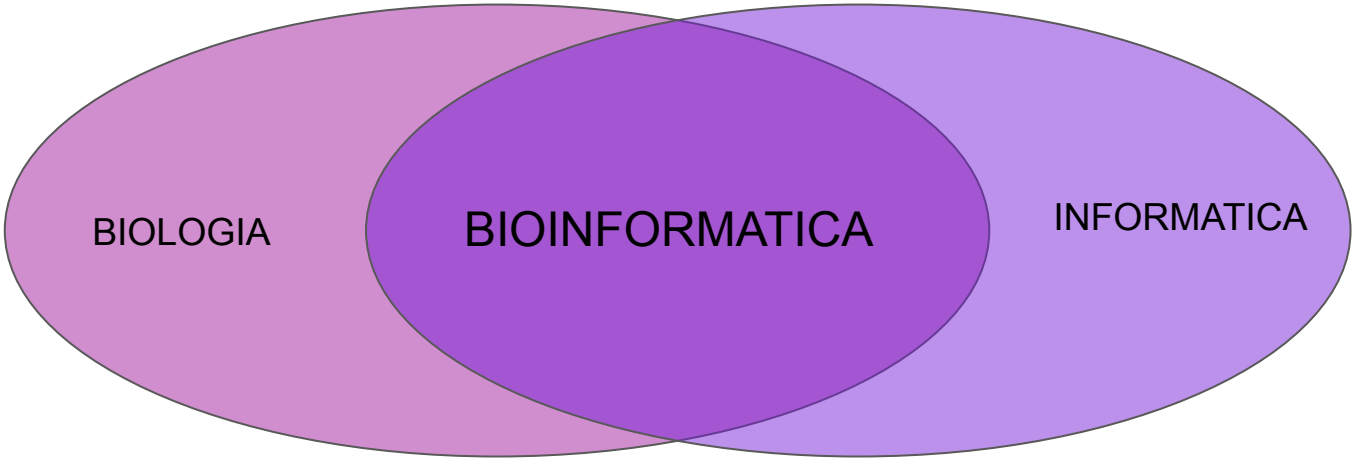
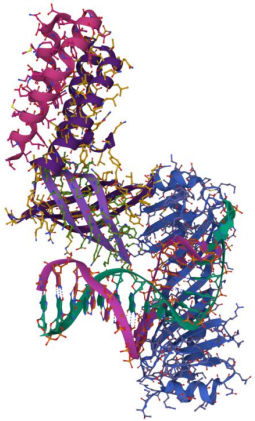
Annalaura Sabatucci



DEF: Approccio computazionale allo studio di sistemi biologici (a livello molecolare) => studio *in silico* (contrapposto allo studio *in vitro* o *in vivo*)

Cosa si può studiare mediante approccio bioinformatico



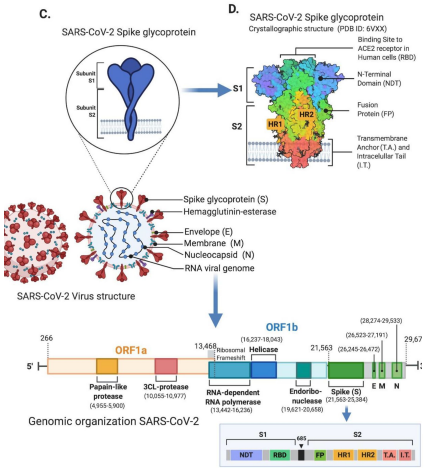
- Relazioni **evolutive** tra molecole o frammenti di molecole
- Relazioni **strutturali** tra molecole
- Relazione **tra sequenza e struttura**
- Interazione tra molecole
- **Network** di interazione (INTERATTOMA); di regolazione; di metabolica (METABOLOMA)
- STRUTTURA DEI GENOMI; RELAZIONE TRA GENOMI (Comparative Genomics)
- Modificazioni epigenetiche dei genomi (es metilazione: metiloma)

AREE DELLA BIOINFORMATICA

- **GENOMICA** (**Computational Genomics**) Study of entire genomes. Huge amount of data, fast algorithms, limited to sequence.

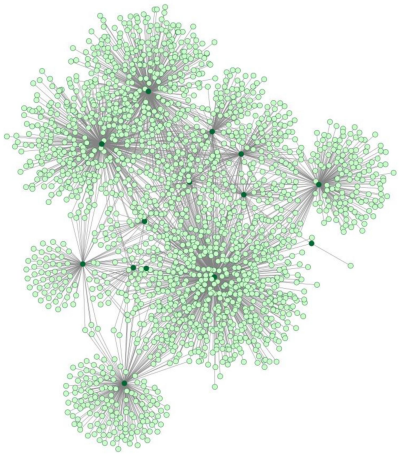
- **BIOINFORMATICA STRUTTURALE** Study of the folding process of bio-molecules. Less structural data than sequence data available, step toward function.

- **SYSTEMS BIOLOGY** Study of complex interactions in biological systems. High level of representation.



Schematic diagram of the genomic structure of the 29.3 kb(nCoV) gene and domain structure of the1273aa spike glycoprotein.

http://dx.doi.org/10.1136/jclinpath-2020-206658



The protein–protein interaction network obtained from IMEx. Dark green nodes represent SARS-CoV-2 proteins, whereas light green nodes are human proteins.

Biomolecules **2022**, *12*(2), 188; **https://doi.org/10.3390/biom12020188**

Computational genomics

La genomica computazionale si focalizza sulla comprensione dei principi secondo cui in DNA regola la biologia di qualsiasi specie a livello molecolare

Al giorno d'oggi abbiamo a disposizione una enorme quantità di dati biologici

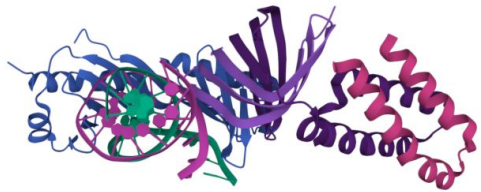
=> gli studi computazionali sono diventati uno strumento determinante nello studio della biologia

L'uso di algoritmi e tecniche di **machine learning** permettono di scoprire segnali biologici in grandi genomi, ricostruire network cellulari e scoprire meccanismi di evoluzione del genoma.

(da: MIT EECS dpt)

RISORSE DEL WEB

- BANCHE DATI (databases)
- GENOME BROWSERS

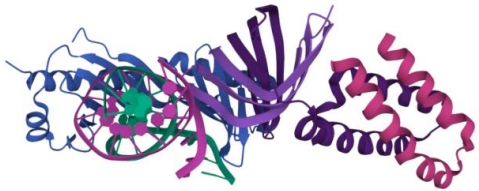


BANCHE DATI

preistoria

In 1955 English biochemist Frederick Sanger sequenced the amino acids of insulin, the first of any protein.

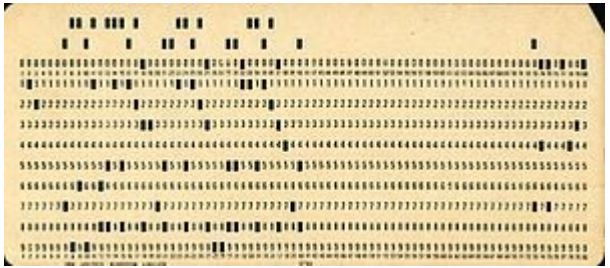
Sanger's work "revealed that a protein has a definite constant, genetically determined sequence—and yet a sequence with no general rule for its assembly. Therefore it had to have a code" (Judson, *Eighth Day of Creation*, 188).

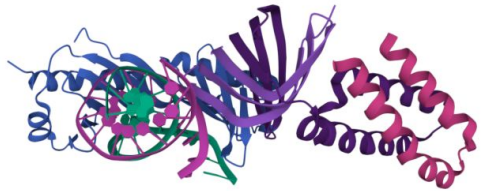


BANCHE DATI

Dr. Margaret Oakley Dayhoff, pioniera della **bioinformatica**, nel 1965 ha pubblicato 'Atlas of protein sequence and structure'. Da questo progetto ha preso vita il **database** Protein Information Resource (PIR) che nel 2002 diventa Uniprot (**Universal Protein Resource**)

Margaret Oakley Dayhoff in the computer room at the National Biomedical Research Foundation, in front of the punch card reader.





UniProt

UniProtKB | Advanced | Search

BLAST | Align | Retrieve/ID mapping | Peptide search | SPARQL | Help | Conta

UniProt BETA **The new UniProt website is here!** Take me to UniProt BETA

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB

UniProt Knowledgebase

Swiss-Prot (566,996)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL

UniRef

Sequence clusters

UniParc

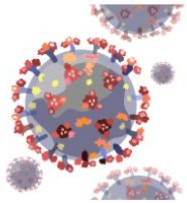
Sequence archive

Proteomes

Proteome sets

Supporting data

Literature citations | Taxonomy | Subcellular locations



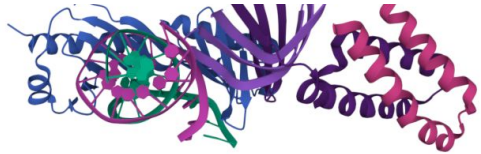
New UniProt portal for the latest SARS-CoV-2 coronavirus protein entries and receptors, updated independent of the general UniProt release cycle.

View SARS-CoV-2 Proteins and Receptors

News

Forthcoming changes
Planned changes for UniProt

UniProt release 2022_01
A phospholipase for clear vision | Cross-references to MANE-Select



GENOME BROWSERS

DEF: GENOME BROWSER: interfaccia grafica per dati genomici di un database biologico.

I genome browser permettono di visualizzare ed analizzare interi genomi.
I dati rappresentati sono 'annotati' ovvero portano annotazioni che permettono di predire le proteine espresse da un gene; la regolazione; la variazione; effettuare analisi comparative e molto altro

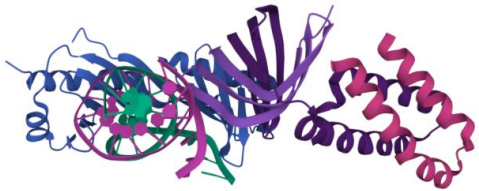
Si possono trovare molti genome browsers. Tra questi, FREE ed accessibili online e tra i più utilizzati possiamo citare:

UCSC Genome Browser,

NCBI's Genome Data Viewer.

Ensembl Genome Browser and

NB: Quelli elencati coprono molti genomi, ma per alcune specie esistono genome browsers dedicati.



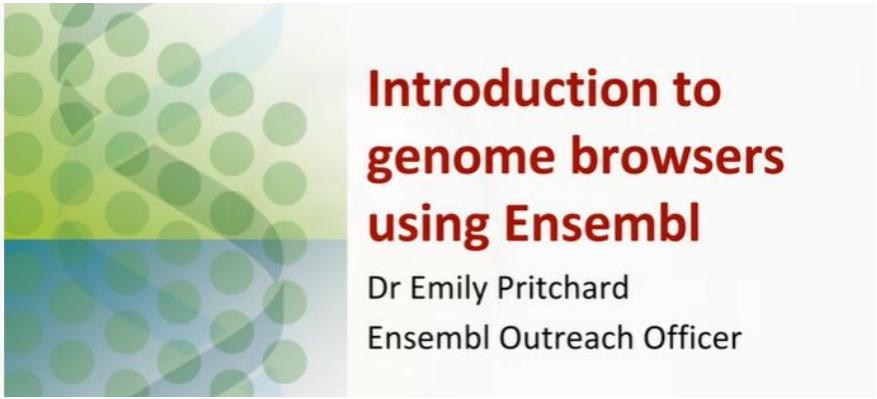
Ensembl hosts several genomic databases and resources for comparative genomics, variation, gene regulation, and epigenetics. Browse through over eighty genomes, from **alpacas** through **zebrafish**. Ensembl also provides tools for identifying phenotypic effects of genetic variants and comparing your data with known genomic variation.

The **UCSC Genome Browser** provides an interactive interface for navigating several sources of genomic information including sequence variation, epigenetic modifications, and transcription factor binding sites. In addition to the genomes of several model organisms, the UCSC Genome Browser also has annotated genomes of many other species, including those of naked mole-rats, the Ebola virus, and over ten species of fruit flies!

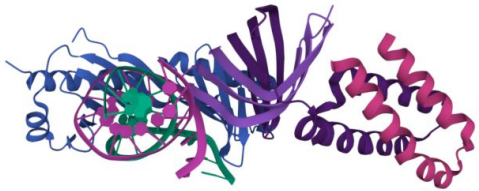
The **Variation Viewer** at NCBI allows you to search the human genome and visualize known genomic variants. The viewer integrates data from several genetic databases, including Gene, dbSNP, and dbVar. The viewer also includes information on medical genetics and gene expression. You can even use the Variation Viewer to compare your own data to what is already known about genetic variation and gene expression.

The EnsEMBL browser

https://youtu.be/42qZyXSH0Cc



Cos' è un genome browser
Cos'è un genome assembly e perché si aggiorna
Cosa sono le annotazioni (snp, deletion, insertion...)
Cos'è il regulatory build



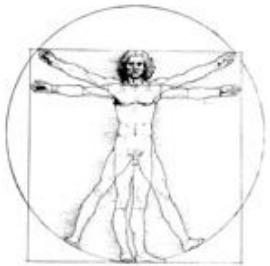
Genome Reference Consortium

GRC Home | Data | Help | Report an Issue | Contact Us | Credits | Curators Only

Human | Mouse | Zebrafish | Rat | Chicken | Paper Supplemental Data

Genome Assemblies

The GRC has built tools to facilitate the curation of genome assemblies based on the sequence overlaps of long, high quality sequences (clones and PCR products, not short sequence reads). The GRC currently supports production of assemblies for human, mouse or zebrafish. If your assembly data fits this model and you are interested in using these tools, please contact us. Subscribe to the grc-announce email list to receive email notification for all GRC assembly updates.



Human

The human genome assembly was produced as part of the Human Genome Project (HGP). The previous assembly (NCBI36) was the last one produced by the HGP and was described in 2004 (PMID: 15496913); this was the starting point for the GRC. The assembly is based largely on assembling overlapping clone sequences.

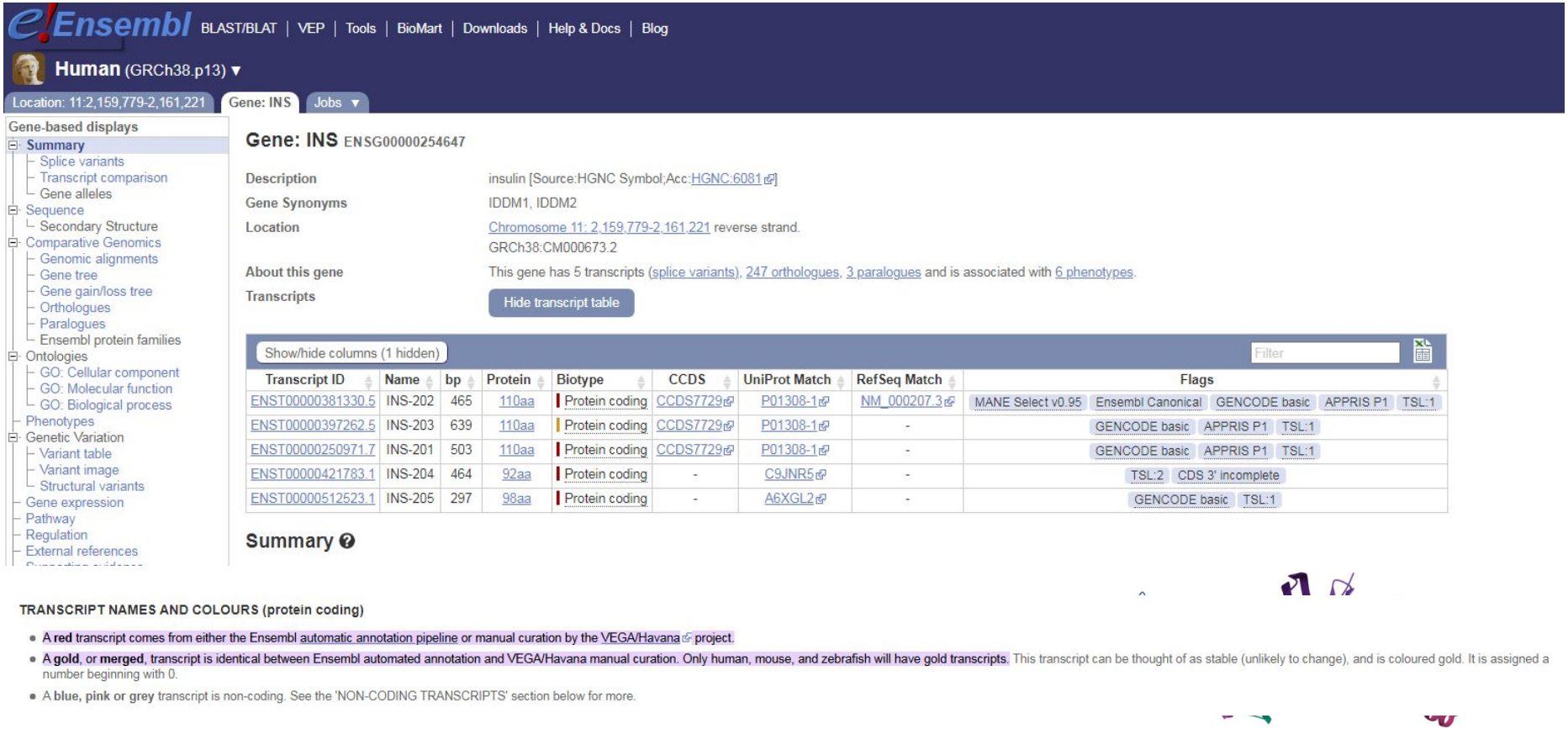
Human assembly information

Current major assembly	GRCh38
Regions with alternate loci	178
Assembly N50	67,794,873 bp
Remaining gaps	875
Patch release version	p14
Patches released	FIX: 164, NOVEL: 90

More human assembly statistics...

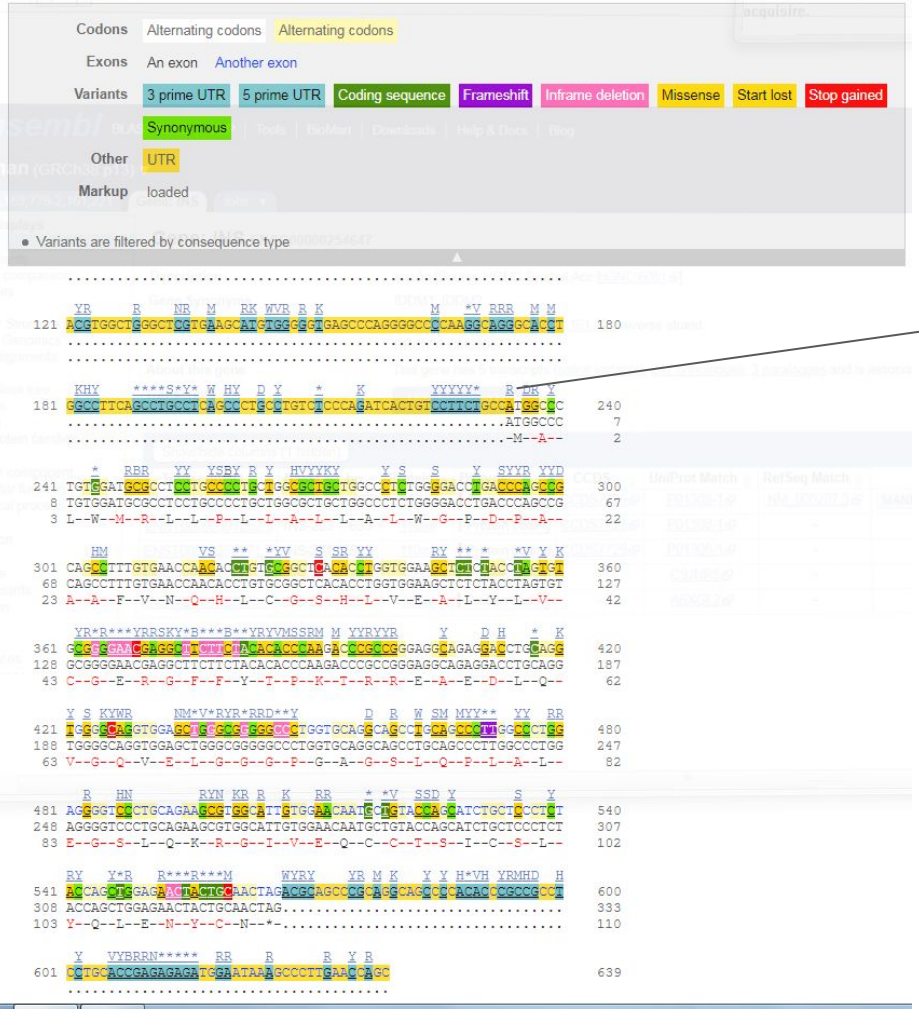
Marzo 2022

ANNOTAZIONE
Es. gene insulina umana

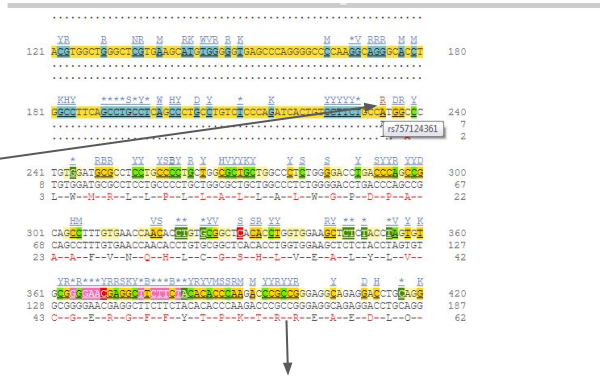


TRANSCRIPT NAMES AND COLOURS (protein coding)

- A **red** transcript comes from either the Ensembl automatic annotation pipeline or manual curation by the VEGA/Havana project.
- A **gold**, or **merged**, transcript is identical between Ensembl automated annotation and VEGA/Havana manual curation. Only human, mouse, and zebrafish will have gold transcripts. This transcript can be thought of as stable (unlikely to change), and is coloured gold. It is assigned a number beginning with 0.
- A **blue, pink or grey** transcript is non-coding. See the 'NON-CODING TRANSCRIPTS' section below for more.



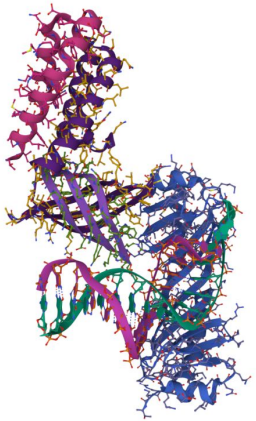
snp (single nucleotide polimorphism)



rs757124361 SNP

Most severe consequence	start lost | See all predicted consequences
Alleles	T/C | Ancestral: T | Highest population MAF: < 0.01
Change tolerance	CADD: C:23.5 | GERP: 1.70
Location	Chromosome 11:2160971 (forward strand) | VCF: 11 2160971 rs757124361 T C
Evidence status	
HGVS names	This variant has 29 HGVS names - Show
Synonyms	ClinGen Allele Registry CA5818208 (C)
Original source	Variants (including SNPs and indels) imported from dbSNP (release 154) | View in dbSNP
About this variant	This variant overlaps 18 transcripts.

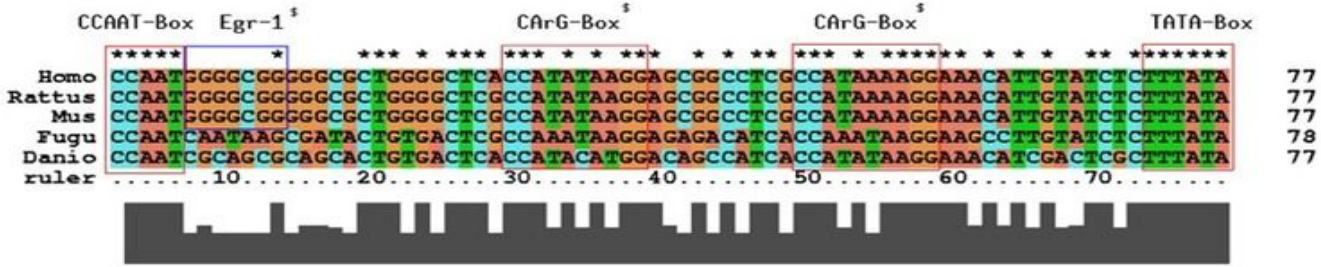
ALIGNMENT

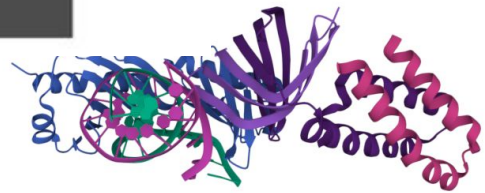


Similarity profiles

Organism	 CHIMP	 MOUSE	 CHICKEN	 FRUIT FLY
Gene Conservation with Humans (%)	99.5	88	75	60

Researchers can learn a great deal about the structure and function of human genes by examining their counterparts in **model organisms**.



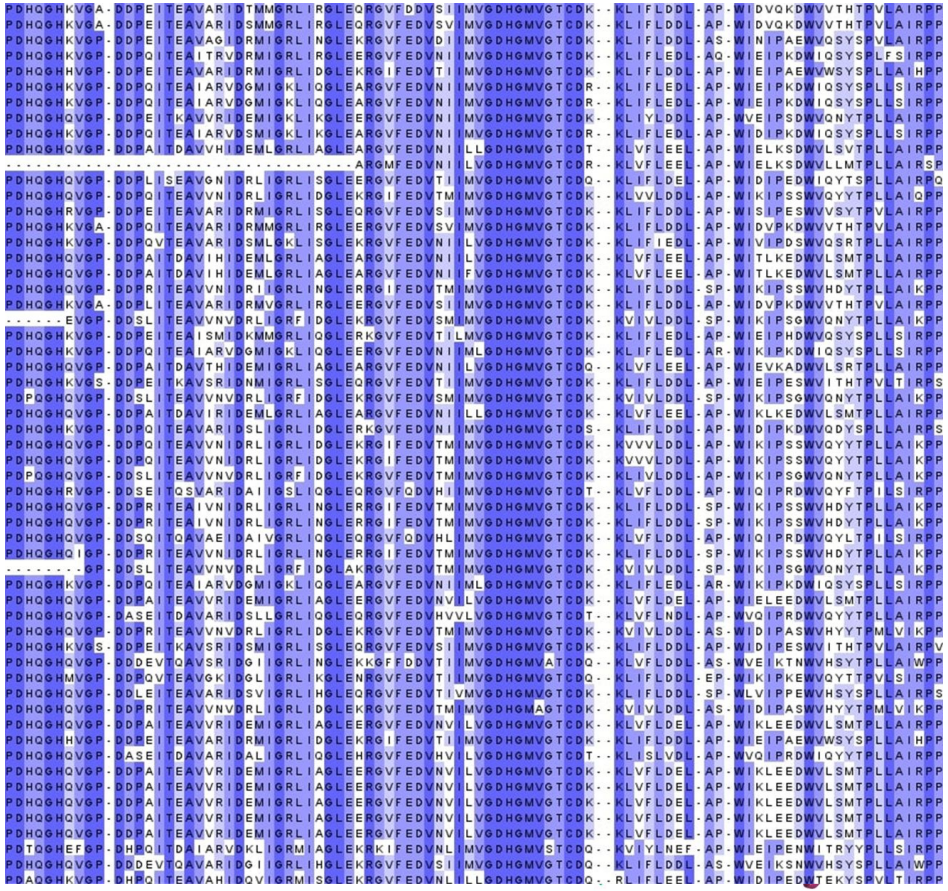


ALLINEAMENTO DI SEQUENZE

Un allineamento di più o 2 sequenze nucleotidiche o proteiche si basa sull'assunzione che esista una relazione di tipo evolutivo tra queste.

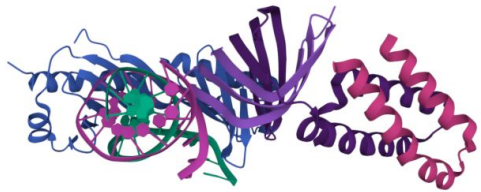
Le tecniche di confronto di sequenze mediante allineamento e gli algoritmi di ricerca nei database biologici sono tool fondamentali della bioinformatica

Il confronto tra sequenze aiuta nella comprensione dell'informazione contenuta in queste e della loro funzione



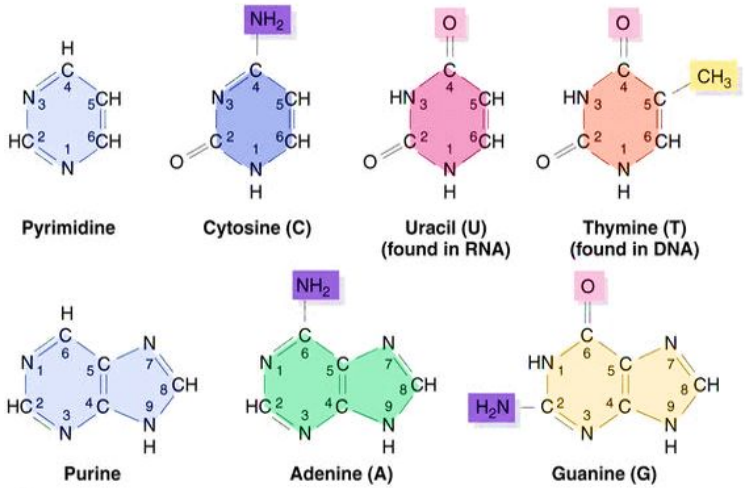
Gli allineamenti di sequenza vengono usati ad esempio per

- Determinare la funzione di una nuova sequenza genica
- Determinare la relazione evolutiva tra geni, proteine e specie intere
- Predire la struttura e la funzione di nuove proteine (step iniziale dell' Homology modeling)

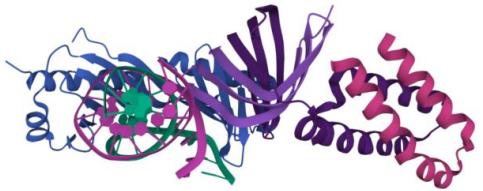


SEQUENZE

una sequenza nucleotidica o proteica è una successione di caratteri di tipo testo
DNA/RNA: tipicamente 4 caratteri (A-C-G-T/U)
proteine: 20 caratteri (uno per ogni aminoacido: si usa la notazione ad 1 lettera)



A	Alanine
R	Arginine
N	Asparagine
D	Aspartic acid
C	Cysteine
E	Glutamic acid
Q	Glutamine
G	Glycine
H	Histidine
I	Isoleucine
L	Leucine
K	Lysine
M	Methionine
F	Phenylalanine
P	Proline
S	Serine
T	Threonine
W	Tryptophan
Y	Tyrosine
V	Valine



FORMATO FASTA

Il formato più utilizzato per rappresentare una sequenza è il FASTA
(suffisso del file: .fas; .fasta, è un file di tipo testo, quindi leggibile da qualsiasi programma di elaborazione testi)
Una sequenza in formato fasta è formata da 2 parti:
una riga di intestazione
una serie di righe di dati di sequenza vera e propria

La riga di intestazione inizia sempre con il carattere > seguito da una descrizione
La sequenza vera e propria inizia nella riga successiva

NB IMPORTANTE IL FONT UTILIZZATO!!!! (vd Courier New)

CARATTERI AMMESSI

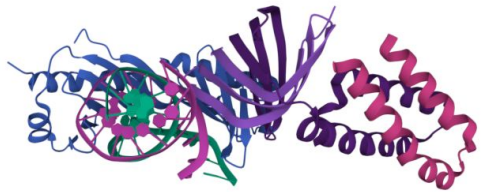
A	adenosine	C	cytidine	G	guanine
T	thymidine	N	A/G/C/T (any)	U	uridine
K	G/T (keto)	S	G/C (strong)	Y	T/C (pyrimidine)
M	A/C (amino)	W	A/T (weak)	R	G/A (purine)
B	G/T/C	D	G/A/T	H	A/C/T
V	G/C/A	-	gap of indeterminate length		

A	alanine	P	proline
B	aspartate/asparagine	Q	glutamine
C	cystine	R	arginine
D	aspartate	S	serine
E	glutamate	T	threonine
F	phenylalanine	U	selenocysteine
G	glycine	V	valine
H	histidine	W	tryptophan
I	isoleucine	Y	tyrosine
K	lysine	Z	glutamate/glutamine
L	leucine	X	any
M	methionine	*	translation stop
N	asparagine	-	gap of indeterminate length

NB
Quando si lavora con le sequenze è IMPORTANTE IL FONT UTILIZZATO!!!!

(ogni carattere testo deve occupare lo stesso spazio per poter lavorare)

Courier	A-B-C-D-E-F-G-H-I-L-*
Questrial	A-B-C-D-E-F-G-H-I-L-*
Arial	A-B-C-D-E-F-G-H-I-L-*



ESEMPI di sequenze

Sequences (2+)

Sequence status: Complete.
Sequence processing: The displayed sequence is further processed into a mature form.
This entry describes **2** isoforms produced by **alternative splicing**. Align | Add to basket
This entry has 2 described isoforms and 2 potential isoforms that are computationally mapped. Show all | Align All

Isoform 1 (identifier: **P01308-1**) [UniParc] FASTA | Add to basket
This isoform has been chosen as the canonical sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.
« Hide

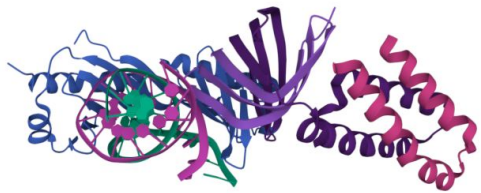
Length: 110
Mass (Da): 11,981
Last modified: July 21, 1986 - v1
Checksum: C2C3B23B85E520E5

BLAST GO

```
        10         20         30         40         50
MALWMRLLPL LALLALWGPD PAAAFVNQHL CGSHLVEALY LVCGERGFFY
        60         70         80         90        100
TPKTRREAED LQVGQVELGG GPGAGSLQPL ALEGSLQKRG IVEQCCTSIC
       110
SLYQLENYCN
```

```
>sp|P01308|INS_HUMAN Insulin OS=Homo sapiens OX=9606 GN=INS PE=1 SV=1
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
```

```
>
ATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGAC
CCAGCCGCAGCCTTTGTGAACCAACACCTGTGCGGCTCACACCTGGTGGAAGCTCTCTAC
CTAGTGTGCGGGGAACGAGGCTTCTTCTACACACCCAAGACCCGCCGGGAGGCAGAGGAC
CTGCAGG
```



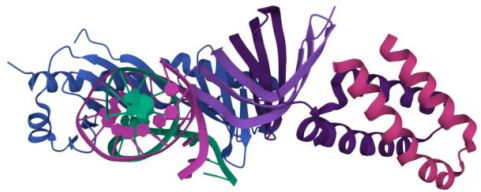
ALLINEAMENTO BINARIO

L’allineamento più semplice è l’allineamento binario:

confronto di 2 caratteri appartenenti a 2 sequenze diverse.

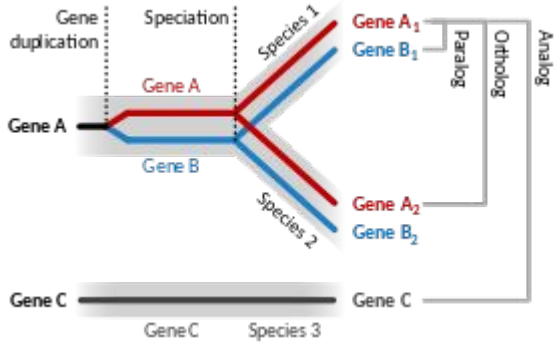
L’allineamento di nucelotidi o aminoacidi riflette la la loro relazione evolutiva, ovvero la presenza di un antenato comune.

La similarità viene quantificata mediante un valore razionale (percentuale di omologia)



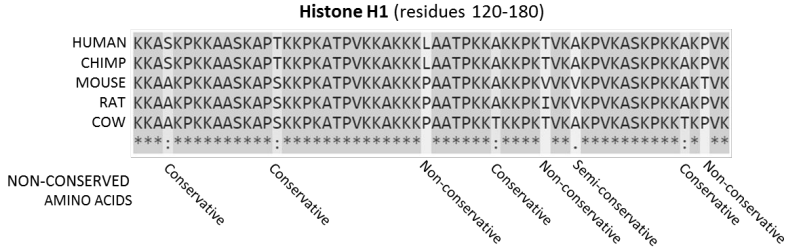
2 segmenti di DNA possono avere un antenato comune per

- SPECIAZIONE (ortologhi)
- DUPLICAZIONE (paraloghi)

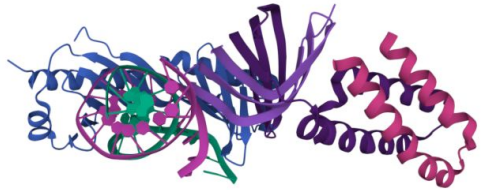


Il grado di similarità tra 2 sequenze viene misurato mediante la percentuale di 'omologia'

PROTEINE: distinguiamo
percentuale di residui identici (percentuale di identità)
percentuale di residui con proprietà fisicochimiche simili (percentuale di similarità)

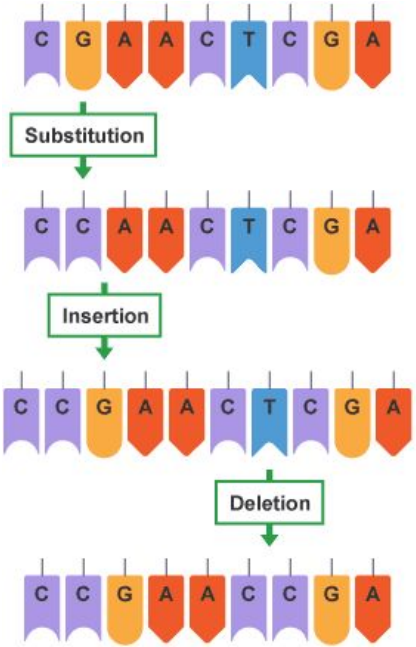


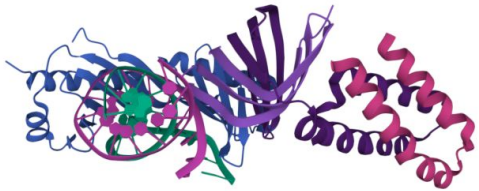
By Thomas Shafee - Own work, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=37188728



In una data posizione all'interno della sequenza possono trovarsi 3 tipi di variazioni:
una SOSTITUZIONE (di un carattere con un altro)
un'INSERZIONE (aggiunta di 1 o più caratteri)
una DELEZIONE (perdita di uno o più carateri)

In natura, inserzioni e delezioni sono significativamente meno frequenti delle sostituzioni!





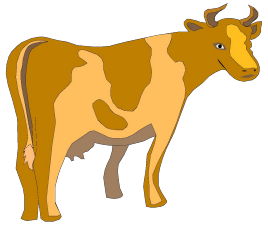
ALLINEAMENTO DI SEQUENZE:

ALIGN (es. in UNIPROT): allineamento di 2 o più sequenze proteiche
https://www.uniprot.org/align/

ES. allineamento tra:

Insulina bovina (insulin bos taurus): uniprot P01317

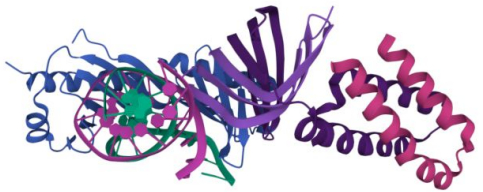
Insulina umana (insulin homo sapiens: uniprot P01308



Job status: COMPLETED

```
P01317 INS_BOVIN              1  MALWTRLRPLLALLALWPPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEG        60
P01308 INS_HUMAN              1  MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED        60
                                 **** ** ********* * ** ******************************:***.*.

P01317 INS_BOVIN             61  PQVGALELAGGPGAGGL-----EGPPQKRGIVEQCCASVCSLYQLENYCN                  105
P01308 INS_HUMAN             61  LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN                  110
                                  *** :**.******.*     **  **********:*:***********
```



ANNOTATION

Highlight

Annotation

- [ ] Peptide
- [ ] Beta strand
- [ ] Disulfide bond
- [x] Signal peptide
- [ ] Helix
- [ ] Turn
- [ ] Propeptide
- [ ] Natural variant

Amino acid properties

- [ ] Similarity
- [ ] Hydrophobic
- [ ] Negative

```
P01317 INS_BOVIN          1  MALWTRLRPLLALLALWPPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEG   60
P01308 INS_HUMAN          1  MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED   60
                             **** ** ********* * ** ****************************:***.*.

P01317 INS_BOVIN         61  PQVGALELAGGPGAGGL-----EGPPQKRGIVEQCCASVCSLYQLENYCN            105
P01308 INS_HUMAN         61  LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN            110
                              *** :**.******.*     **  **********:*:***********
```

You may add additional sequences to this alignment (in FASTA format)

Add sequence and align

PTM / Processing

Molecule processing

Feature key	Position(s)	Description
Signal peptide	1 – 24	1 Publication
Peptide (PRO_0000015819)	25 – 54	Insulin B chain
Propeptide (PRO_0000015820)	57 – 87	C peptide
Peptide (PRO_0000015821)	90 – 110	Insulin A chain