

Introduzione alla bioinformatica

Loriano Storchi

loriano@storchi.org

<http://www.storchi.org/>

Definizione

- **Bioinformatica:** scienza multi-disciplinare, al crocevia tra biologia, chimica, matematica, fisica ed informatica, che analizza l'informazione biologica con metodi computazionali al fine di formulare ipotesi sui processi della vita. (Anna Tramontano)
- Applicazione di tecniche computazionali nella comprensione ed organizzazione di tutta l'informazione associata alle strutture biologiche. La fisiologia di un organismo vivente e' per buona parte determinata dai suoi geni che possono essere visti e trattati come informazione digitale
- Esempio, possibile applicazione
 - Scoperta una nuova sequenza proteica, posso cercare di dedurre la sua funzionalita' , in modo approssimato, confrontandola con tutte le sequenze proteiche note al mondo.
 - Gestire questo tipo di ricerca e' impossibile per un'essere umano ma invece fattibile usando computers (algoritmi di ricerca, database, protocolli, etc etc) ed in generale strumenti informatici

Definizione

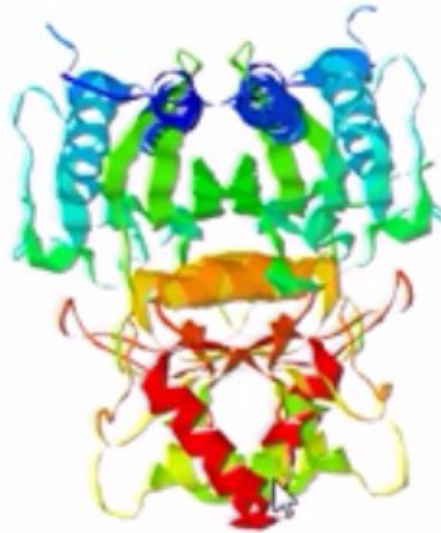
- E' necessario memorizzare analizzare ed interpretare una enorme quantita' di dati. L'intera sequenza del genoma umano, scritta in Times New Roman, dimensione 12, avrebbe una lunghezza di 5000 km!
- DNA (memoria) l'RNA (comunicazione) ed in fine le proteine (esecuzione)
- Quali parti del DNA sono importanti e controllano determinati processi ?
- Quale e' la funzione di certe proteine ?
- Come posso confrontare due sequenze proteiche o genomiche ?
- Molto di piu'.....

Struttura delle proteine altro esempio

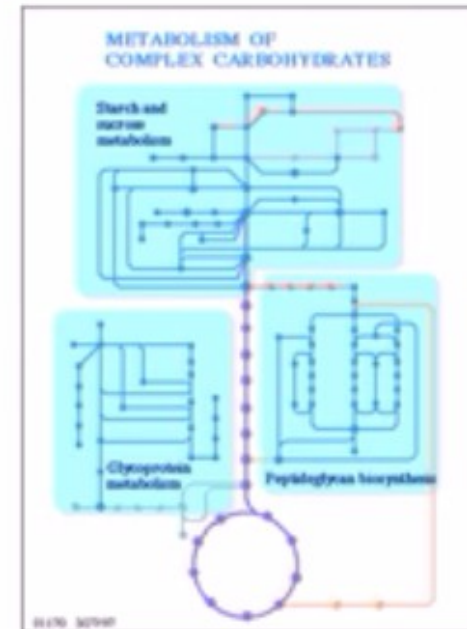
•Relationships between

```
TDQAAFDTNIVTLTRFVMEQGRKARGTGEM  
TQLLSLCTAVKAISTAVRKAGIAHLYGIA  
GSTNVTGDQVKKLDVLSNDLVINVLKSSFA  
TCVLVTEEDKNAIIVEPEKRGKYVVCFDPL  
DGSSNIDCLVSIPTIFGIYRKNSTDEPSEK  
DALQPGRNLVAAGYALYGSATMLV
```

sequence



3D structure

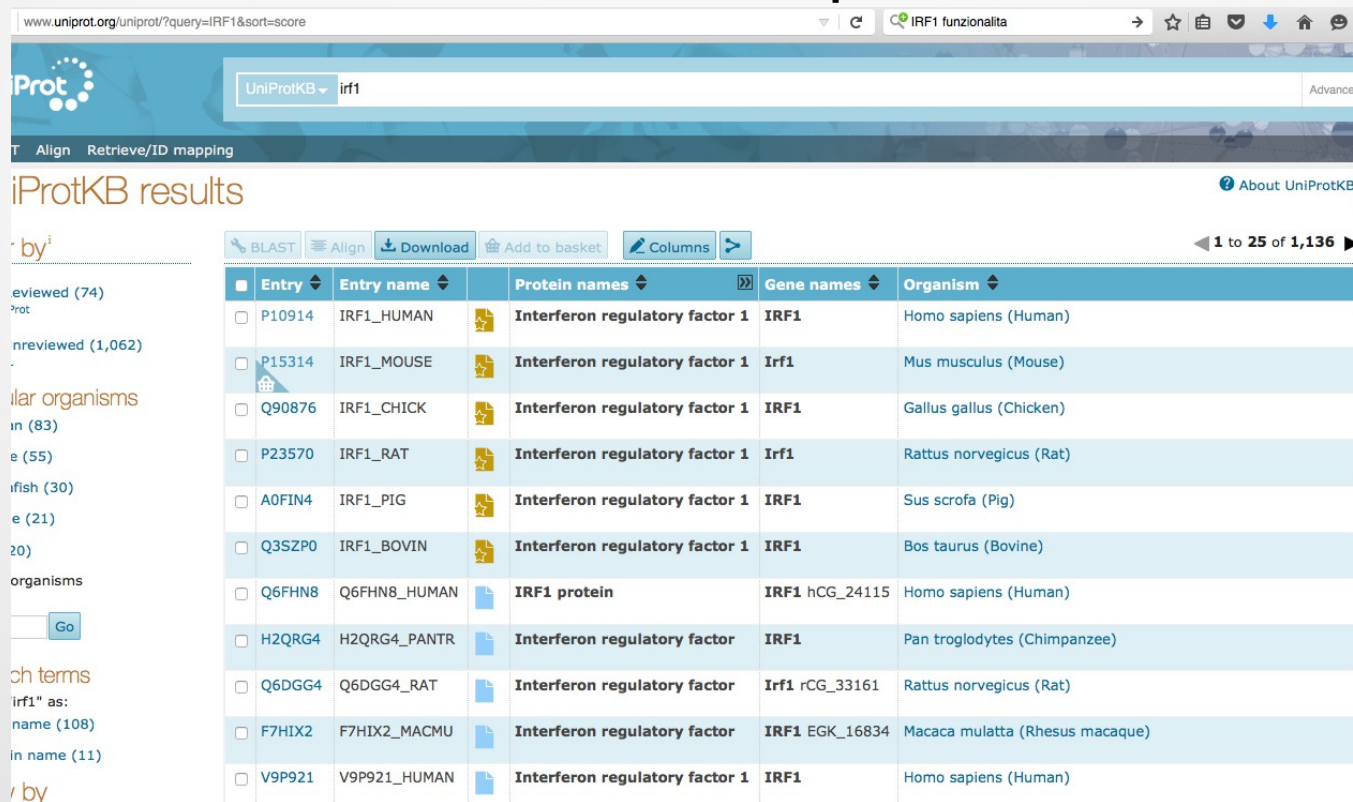


protein functions

Partendo dalla sequenza possiamo “ricostruire” la struttura 3D che ci permettera' di avere informazioni sulla funzionalita' della proteina

Esempio Banche dati, formati ed Omologia

- Interferon regulatory factor 1 (gene IRF1) vediamo un semplice flusso di operazioni molto preliminare
- Partiamo da una ricerca su Uniprot:



www.uniprot.org/uniprot/?query=IRF1&sort=score

UniProtKB irf1

Align Retrieve/ID mapping

iProtKB results

1 to 25 of 1,136

Entry	Entry name	Protein names	Gene names	Organism
P10914	IRF1_HUMAN	Interferon regulatory factor 1	IRF1	Homo sapiens (Human)
P15314	IRF1_MOUSE	Interferon regulatory factor 1	Irf1	Mus musculus (Mouse)
Q90876	IRF1_CHICK	Interferon regulatory factor 1	IRF1	Gallus gallus (Chicken)
P23570	IRF1_RAT	Interferon regulatory factor 1	Irf1	Rattus norvegicus (Rat)
A0FIN4	IRF1_PIG	Interferon regulatory factor 1	IRF1	Sus scrofa (Pig)
Q3SZP0	IRF1_BOVIN	Interferon regulatory factor 1	IRF1	Bos taurus (Bovine)
Q6FHN8	Q6FHN8_HUMAN	IRF1 protein	IRF1 hCG_24115	Homo sapiens (Human)
H2QRG4	H2QRG4_PANTR	Interferon regulatory factor	IRF1	Pan troglodytes (Chimpanzee)
Q6DGG4	Q6DGG4_RAT	Interferon regulatory factor	Irf1 rCG_33161	Rattus norvegicus (Rat)
F7HIX2	F7HIX2_MACMU	Interferon regulatory factor	IRF1 EGK_16834	Macaca mulatta (Rhesus macaque)
V9P921	V9P921_HUMAN	Interferon regulatory factor 1	IRF1	Homo sapiens (Human)

Standard IUB/IUPAC (FASTA)

Acidi Nucleici

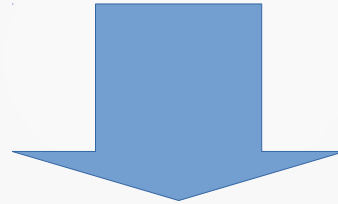
A	Adenina	R	G o A (Purine)	B	G T C
C	Citosina	Y	T o C (Pirimidine)	D	G A T
G	Guanina	K	G o T	H	A C T
T	Timina	M	A o C	V	G C A
U	Uracile	W	A o T	N	A C G T (Any)
				-	Gap

Aminoacidi

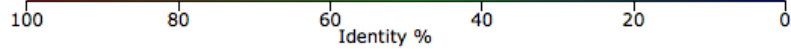
A	Alanina	B	Acido Aspartico o Asparagina	C	Cisteina
D	Acido Aspartico	E	Acido Glutammico	F	Fenilalanina
G	Glicina	H	Istidina	I	Isoleucina
K	Lisina	L	Leucina	M	Metionina
N	Asparagina	P	Prolina	Q	Glutammina
R	Arginina	S	Serina	T	Treonina
U	Selenocisteina	V	Valina	W	Triptofano
Y	Tirosina	Z	Acido Glutammico o Glutammina	X	Qualsiasi (Any)
*	Stop traduzione	-	Gap		

BLAST

```
>sp|P10914|IRF1_HUMAN Interferon regulatory factor 1 OS=Homo sapiens GN=IRF1 PE=1 SV=2
MPITRMRMRPWLEMQINSNQIPGLIWINKEEMIFQIPWKHAAKHGWDINKDACLFERSWAI
HTGRYKAGEKEPDPKTKANFRFCAMNSLPDIEEVKDQSRNKGSSAVRVYRMLPPLTKNQR
KERKSKSSRDAKSKAKRKSCGDSSPDTFSDGLSSSTLPDDHSSYTVPGYMQDLEVEQALT
PALSPCAVSSTLPDWHIPVEVVPDSTSDLYNFQVSPMPSTSEATDEDEEGKLPEDIMKL
LEQSEWQPTNVDGKGYLLNEPGVQPTSVYGDFSCKEEPEIDSPGGDIGLSLQRVFTDLKN
MDATWLDSLLTPVRLPSIQAIPCAP
```



BLAST (Basic Local Alignment Search Tool) questo tool puo' eseguire numerosi confronti nell'unita' di tempo, e quindi permette di fare una ricerca per similarita' nell'intero database



[Edit and resubmit](#) Order by: Score

Overview

Entry	Protein names	Match hit	Identity
P15314	Interferon regulatory factor 1 (Mus musculus)		84.5%
P23906	Interferon regulatory factor 2 (Mus musculus)		45.2%


Alignments

[BLAST](#) [Align](#) [Download](#) [Add to basket](#) [Columns](#)

1 to 2 of 2 Show 25

Entry	Alignment overview	Info	Status	Protein names	3D	Cross-reference (PDB)
<input type="checkbox"/>	Query: sp P10914 IRF1_HUMAN B20151120996H4OU3U6					
<input type="checkbox"/>	P15314 IRF1_MOUSE - Interferon regulatory factor 1 Mus musculus (Mouse) - View alignment	E-value: 0.0 Score: 1,495 Ident.: 84.5%		Interferon regulatory factor 1 (Mus musculus)	X-ray crystallography (1)	1IF1 X-ray 3.00 A/B 1-113 [↗]
<input type="checkbox"/>	P23906 IRF2_MOUSE - Interferon regulatory factor 2 Mus musculus (Mouse) - View alignment	E-value: 730E-66 Score: 552 Ident.: 45.2%		Interferon regulatory factor 2 (Mus musculus)	X-ray crystallography (1) NMR spectroscopy (2)	1IRF NMR - A 2-1: 1IRG NMR - A 2-1: 2IRF X-ray 2.20 G/H/I/J/K/L 1-1:

Allineamento



P10914	IRF1_HUMAN	1	MPITRMRMRPWLEMQINSNQIPGLIWINKEEMIFQIPWKHAAKHGWDINKDACLFRSWAI	60
P15314	IRF1_MOUSE	1	MPITRMRMRPWLEMQINSNQIPGLIWINKEEMIFQIPWKHAAKHGWDINKDACLFRSWAI	60
P10914	IRF1_HUMAN	61	HTGRYKAGEKEPDPKTWKFRCAMNSLPDIEEVKDQSRNKGSSAVRVYRMLPPLTKNQR	120
P15314	IRF1_MOUSE	61	HTGRYKAGEKEPDPKTWKFRCAMNSLPDIEEVKDQSRNKGSSAVRVYRMLPPLTRNQR	120
P10914	IRF1_HUMAN	121	KERKSKSSRDAKSKAKRKCSCGSSPDTFSDGLSSSTLPDDHSSYTVPGYM-QDLEVEQAL	179
P15314	IRF1_MOUSE	121	KERKSKSSRD KSK KRK CGD SPDTFSDGLSSSTLPDDHSSYT GY+ QDL++E+ +	180
P10914	IRF1_HUMAN	180	TPALSPCAVSSSTLPDWHIPVEVVPDSTSDLYNFQVSPMPSTSEATTDEDEEGKLPEDIMK	239
P15314	IRF1_MOUSE	181	TPALSPCVSSSLSEWHMQMDIIPDSTTDLYNLQVSPMPSTSEAATDEDEEGKIAEDLMK	240
P10914	IRF1_HUMAN	240	LLEQSEWQPTNVDGKGYLLNEPGVQPTS SVYGFDFSCKEEPEIDSPGGDIGLSLQRVFTDLK	299
P15314	IRF1_MOUSE	241	LFEQSEWQPTHIDGKGYLLNEPGTQLSSVYGFDFSCKEEPEIDSPRGDIGIGIQHVFTDK	300
P10914	IRF1_HUMAN	300	NMDA-TWLDSSL-TPVRL-PSIQAIPCAP	325
P15314	IRF1_MOUSE	301	NMD+ W+DSSL VRL PSIQAIPCAP	329

Ad esempio i segni + indicano similitudine chimica anche se non c'è una vera e propria coincidenza

Omologia

- Usiamo swiss-model

Start a New Modelling Project

Target Sequence:
*(Format must be Fasta,
Clustal, Promod,
plain string, or a valid
UniProtKB AC)*

Target `MPITRMRMRPWLEMQINSNQIPGLIWINKEEMIQIIPWKHAAKHGWDINKDACLFRSWAIHTGRYKAGEKEPDPKTWKANFRFCAMNSLPD` 90
Target `IEEVKDQSRNKGSSAVRVYRMLPPLTKNQRKERKSKSRDAKSKAKRRCGSDSSPDTFSDGLSSSTLPDDHSSYTPVGYMQDLEVEQAL` 180
Target `PALSPCAVSSSTLPDWHIPVEVVPDSTSDLYNFQVSPMPSTSEATTDEDEEGKLPEDIMKLLQSEWQPTNVDGKGYLLNEPGVQPTSVYG` 270
Target `DFSCKEEPEIDSPGGDIGLSLQRVFTDLKNMDATWLDSSLTPVRLPSIQAIPCA` 325

Reset Form

+ Upload Target Sequence File...

Project Title:

Untitled Project

Email:

Optional

Search For Templates

Build Model

By using the SWISS-MODEL server, you agree to comply with the following [terms of use](#) and to cite the corresponding [articles](#).
I have read the terms of use, and hereby state that I am (Please select)

You are currently not logged in - to take advantage of the workspace, please [log in](#) or [create an account](#).

(There is no requirement to create an account to use any part of SWISS-MODEL, however you will gain the benefit of seeing a list of your previous modelling projects here.)

Modello di omologia

Untitled Project Created: today at 07:47

Summary

Templates 15

Models 3



Model Results

Order by: GMQE



Model 03

Oligo-State

MONOMER (matching prediction)

1 x POTASSIUM ION

Ligand 1 in contact with: Chain A : M85, N86, L88, I91

Ligands

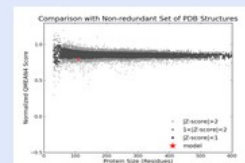
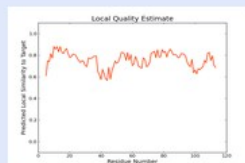
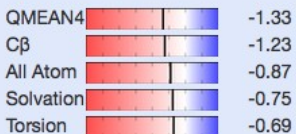
1 x K⁺

GMQE

0.25

QMEAN4

-1.33

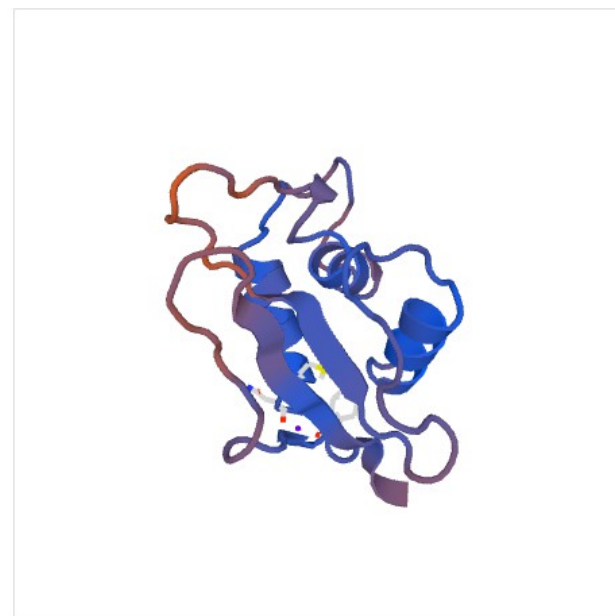


Template Seq Identity Coverage

2irf.1.C 76.11% INTERFERON REGULATORY FACTOR 2

Model-Template Alignment

```
Model_03 MPITRMRMRPWLEMQINSNQIPGLIWINKEEMIFQIPWKHAAKHGWDINKDACLFRSWAIHTGRYKAGEKEPDPK 75
2irf.1.C MPVERMRMRPWLEEQINSNTIPGLKWLNKEKKIFQIDWMMHAAARHGWDVEKDAPLFRNWAHTGKHQPGIDKPPK 75
Model_03 TWKANFRCAMNSLPDIIEVKDQSRNKGSSAVRVYRMLFPLTKNQRKERKSKSSRDAKSKAKRKCSDSSPDTFSD 150
2irf.1.C TWKANFRCAMNSLPDIIEVKDRSIXKGNNAFRVYRMLP----- 113
Model_03 GLSSSTLPDDHSSYVPGYMQDLEVEQALTPALSPCAVSSSTLPDWHIPVEVVPDSTSDLYNFQVSPMPSTSEATT 225
2irf.1.C -----
```



View



PDB formato

- E' costituito da numerosi record quello fondamentale e' il record ATOM:

```
      1         2         3         4         5         6         7         8
1234567890123456789012345678901234567890123456789012345678901234567890
ATOM   145  N   VAL A   25           32.433  16.336  57.540  1.00 11.92           A1  N
ATOM   146  CA  VAL A   25           31.132  16.439  58.160  1.00 11.85           A1  C
ATOM   147  C   VAL A   25           30.447  15.105  58.363  1.00 12.34           A1  C
ATOM   148  O   VAL A   25           29.520  15.059  59.174  1.00 15.65           A1  O
ATOM   149  CB  AVAL A   25           30.385  17.437  57.230  0.28 13.88           A1  C
ATOM   150  CB  BVAL A   25           30.166  17.399  57.373  0.72 15.41           A1  C
ATOM   151  CG1 AVAL A   25           28.870  17.401  57.336  0.28 12.64           A1  C
```

- A parte i primi campi ovvi c'e' poi occupancy, temp. Factor, (posizione dell'atomo), Segment id, simbolo dell'elemento e carica

Altro esempio bioinformatico

- Eyleless e' un gene dalla moscerino della frutta a seguito della sua rimozione le mosche nascono senza occhi. C'e' n gene umano (Aniridia) la cui mancanza piuttosto che eccessiva mutazione non fa sviluppare l'iride dell'occhio.
- Se si va a fare un confronto fra i due geni, ad esempio via BLAST in NCBI, si vede che i due geni si assomigliano molto.
- Questo tipo di operazione fatta la calcolatore e' immediata.