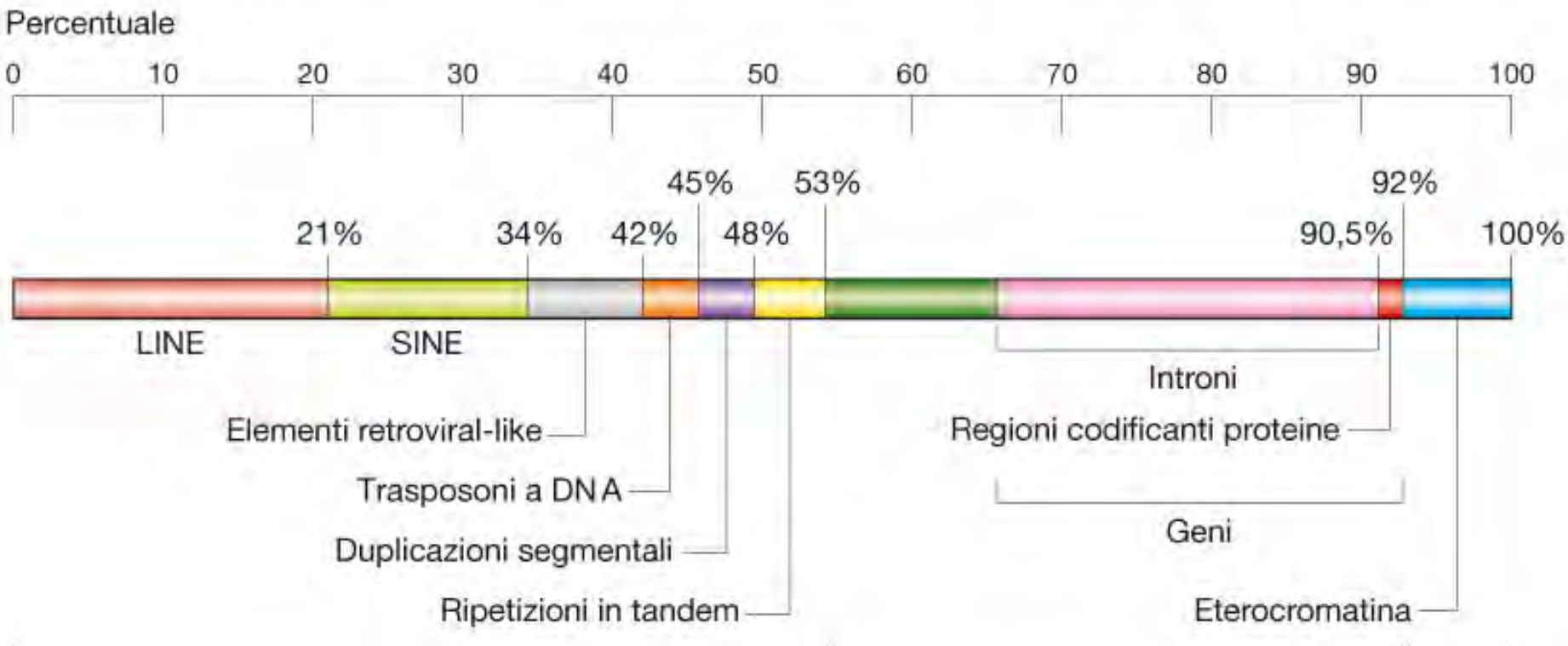
Bioinformatica ed applicazioni di bioinformatica strutturale

Bioinformatica

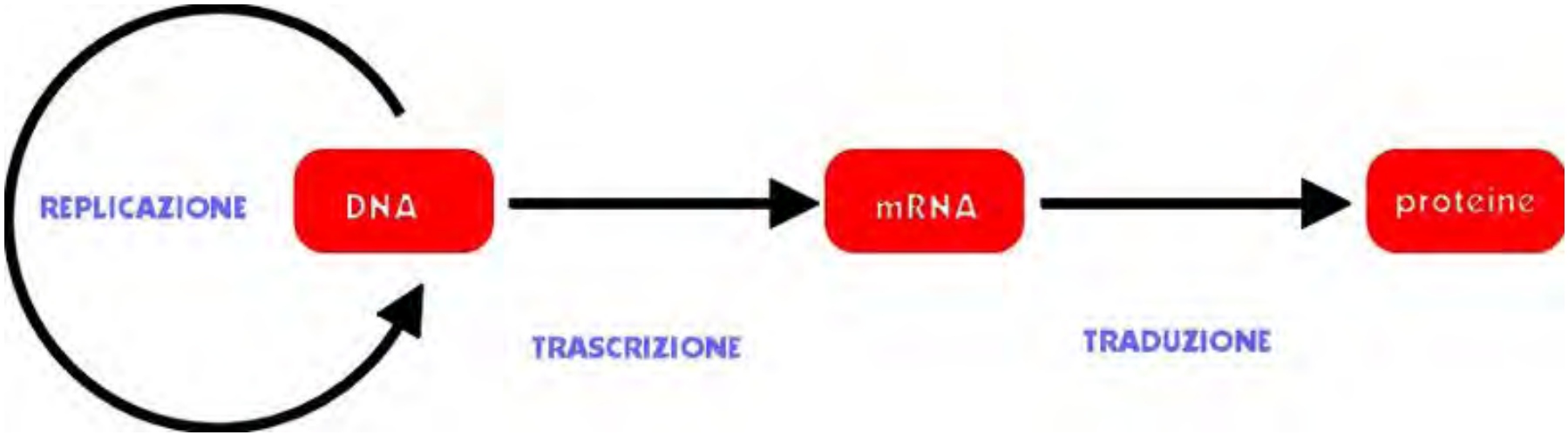
- Le banche dati
- Programmi per estrarre ed analizzare i dati

I numeri

- Cellule nell'uomo
- Genoma umano
- Geni nell'uomo



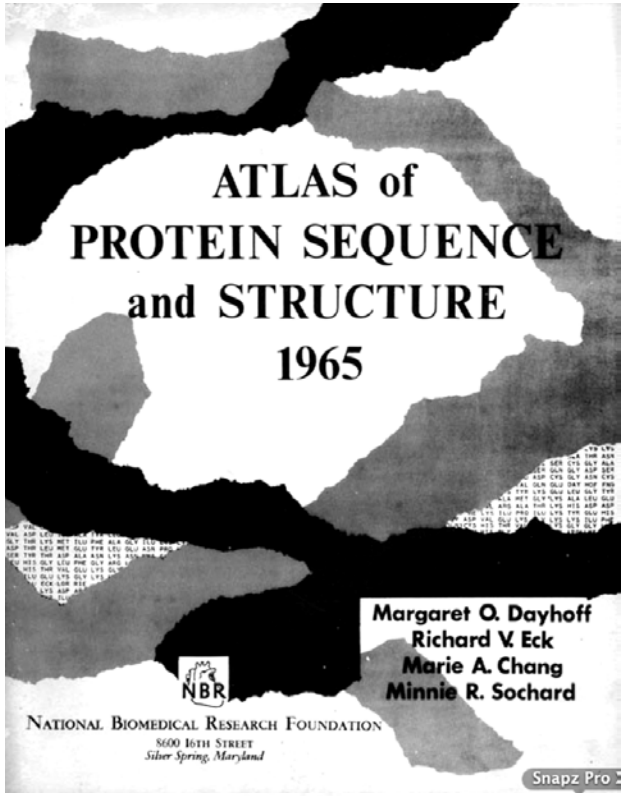
Il dogma centrale
livelli di informazione

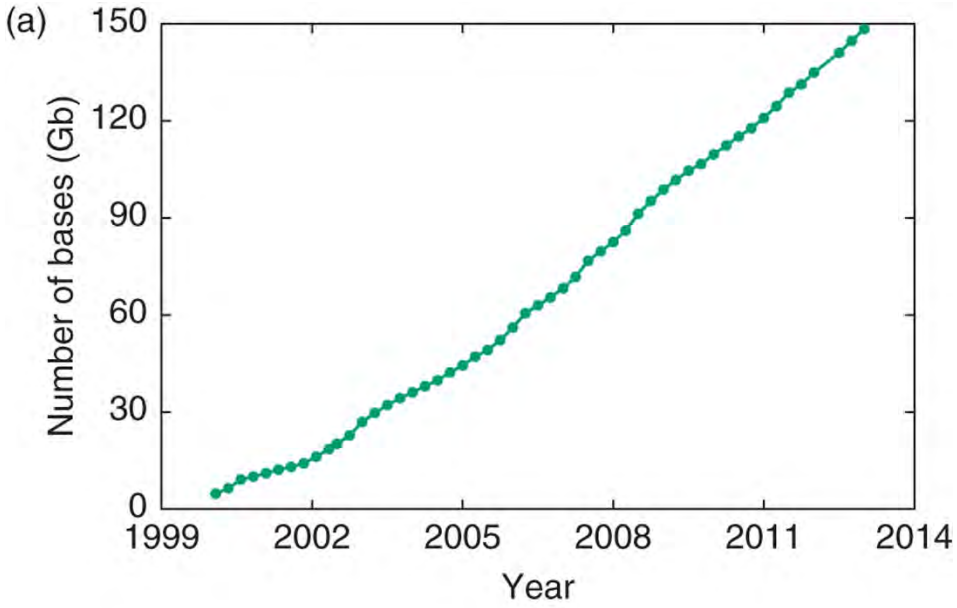


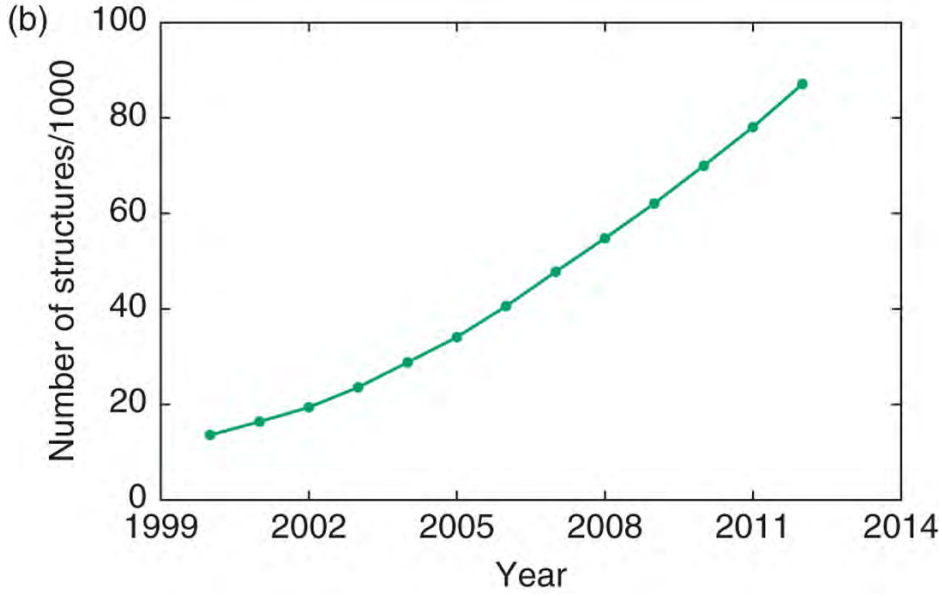
Le banche dati per sequenze di geni, DNA, cDNA, proteine e strutture 3D. Informazione della funzione, dell'espressione, del knockout o over-espressione, i partner di interazioni.

Banche dati

Banche dati primarie e secondarie







Archivi

Dati primari negli archivi

- Sequenze di acidi nucleici
- Sequenze di interi genomi
- Sequenze di amminoacidi delle proteine
- Strutture di proteine ed acidi nucleici
- Strutture di piccole molecole (metaboliti)
- Funzione delle proteine
- Pattern di espressione dei geni
- Networks (pathway metabolici, interazioni proteine e geni, regolazione)
- Pubblicazioni

Unannotated ➔ Preliminary ➔ Unreviewed ➔ Standard

Banche dati

- **Acidi Nucleici: NCBI, ENA (EBI), DDBJ.**
- **Genomi: ENSEMBL**
- **Proteine: PIR, SWISS-PROT, TrEMBL, PROSITE, UNIPROT**
- **Strutture: RCSB, PDB, PDBJ**
- **Pubblicazioni: PubMed**

Gateways per le banche dati

- **Entrez: 35 banche dati di NCBI**
- **Mutazioni di malattie: OMIM**
- **Analisi delle proteine: ExPASy**

Entrez Banche dati

Name	Contents
Taxonomy	Organisms in GenBank
SNP	Short genetic variations
dbVar	Genomic structural variation
Gene	Gene-centred information
SRA	Sequence Read Archive
BioSystems	Pathways and systems of interacting molecules
HomoloGene	Eukaryotic homology groups
OMIM	Online Mendelian Inheritance in Man
OMIA	Online Mendelian Inheritance in Animals
Probe	Sequence-specific reagents
BioProject	Aggregated biological research project data
dbGaP	Genotype and phenotype
UniGene	Gene-oriented clusters of transcript sequences
CDD	Conserved protein domain database
Clone	Integrated data for clone resources
UniSTS	Markers and mapping data
PopSet	Population study data sets
GEO Profiles	Expression and molecular abundance profiles
GEO DataSets	Experimental sets of Gene Expression Omnibus (GEO) data
Epigenomics	Epigenetic maps and data sets
PubChem BioAssay	Bioactivity screens of chemical substances
PubChem Compound	Unique small molecule chemical structures
PubChem Substance	Deposited chemical substance records
Protein Clusters	A collection of related protein sequences
BioSample	Biological material description
PubMed	Biomedical literature citations and abstracts
PubMed Central	Free, full-text journal articles
Site Search	NCBI web and ftp sites
Books	Online books

Homo sapiens insulin (INS), transcript variant 1, mRNA

NCBI Reference Sequence: NM_000207.2

FASTA Graphics

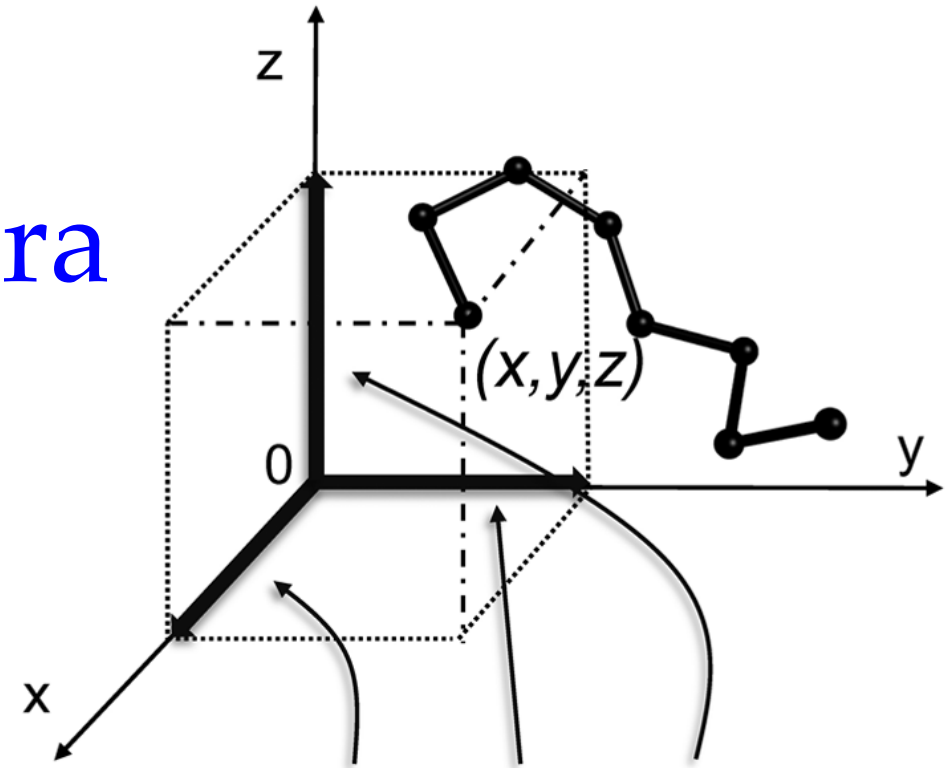
Go to:

```
LOCUS       NM_000207                469 bp    mRNA    linear   PRI 15-SEP-2016
DEFINITION  Homo sapiens insulin (INS), transcript variant 1, mRNA.
ACCESSION   NM_000207
VERSION     NM_000207.2
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 469)
  AUTHORS   Li S, Bouzar C, Cottet-Rousselle C, Zagotta I, Lamarche F, Wabitsch
            M, Tokarska-Schlattner M, Fischer-Posovszky P, Schlattner U and
            Rousseau D.
  TITLE     Resveratrol inhibits lipogenesis of 3T3-L1 and SGBS cells by
            inhibition of insulin signaling and mitochondrial mass increase
  JOURNAL   Biochim. Biophys. Acta 1857 (6), 643-652 (2016)
   PUBMED   26968895
```

...

```
     exon            247..465
                     /gene="INS"
                     /gene_synonym="IDDM; IDDM1; IDDM2; ILPR; IRDN; MODY10"
                     /inference="alignment:Splign:1.39.8"
     STS             247..465
                     /gene="INS"
                     /gene_synonym="IDDM; IDDM1; IDDM2; ILPR; IRDN; MODY10"
                     /standard_name="GDB:179433"
                     /db_xref="UniSTS:155046"
     regulatory      446..451
                     /regulatory_class="polyA_signal_sequence"
                     /gene="INS"
                     /gene_synonym="IDDM; IDDM1; IDDM2; ILPR; IRDN; MODY10"
     polyA_site      465
                     /gene="INS"
                     /gene_synonym="IDDM; IDDM1; IDDM2; ILPR; IRDN; MODY10"
ORIGIN
        1 agccctccag gacaggctgc atcagaagag gccatcaagc agatcactgt ccttctgcca
       61 tggccctgtg gatgcgcctc ctgcccctgc tggcgctgct ggccctctgg ggacctgacc
      121 cagccgcagc ctttgtgaac caacacctgt gcggctcaca cctggtggaa gctctctacc
      181 tagtgtgcgg ggaacgaggc ttcttctaca cacccaagac ccgccgggag gcagaggacc
      241 tgcaggtggg gcaggtggag ctgggcgggg gccctggtgc aggcagcctg cagcccttgg
      301 ccctggaggg gtccctgcag aagcgtggca ttgtggaaca atgctgtacc agcatctgct
      361 ccctctacca gctggagaac tactgcaact agacgcagcc cgcaggcagc cccacacccg
      421 ccgcctcctg caccgagaga gatggaataa agcccttgaa ccagcaaaa
//
```

La struttura PDB



ATOM	1	N	VAL	A	1	19.329	29.021	42.856	1.00	43.02	N
ATOM	2	CA	VAL	A	1	20.128	30.234	42.646	1.00	37.40	C
ATOM	3	C	VAL	A	1	21.619	29.864	42.729	1.00	29.27	C
ATOM	4	O	VAL	A	1	22.006	29.309	43.777	1.00	31.99	O
ATOM	5	CB	VAL	A	1	19.841	31.263	43.772	1.00	60.46	C
ATOM	6	CG1	VAL	A	1	20.254	32.668	43.348	1.00	64.75	C
ATOM	7	CG2	VAL	A	1	18.439	31.182	44.322	1.00	72.38	C
ATOM	8	N	LEU	A	2	22.373	30.245	41.718	1.00	24.14	N
ATOM	9	CA	LEU	A	2	23.818	29.940	41.779	1.00	20.91	C
ATOM	10	C	LEU	A	2	24.485	31.055	42.576	1.00	20.08	C
ATOM	11	O	LEU	A	2	24.354	32.219	42.170	1.00	27.23	O
ATOM	12	CB	LEU	A	2	24.354	29.824	40.346	1.00	19.33	C
ATOM	13	CG	LEU	A	2	23.889	28.647	39.501	1.00	23.06	C
ATOM	14	CD1	LEU	A	2	24.336	28.777	38.046	1.00	28.74	C
ATOM	15	CD2	LEU	A	2	24.422	27.336	40.080	1.00	23.33	C

Abbreviazioni

- NP: is for protein, Natural Protein
- NM: is for mRNA, Natural mRNA
- NR: is for RNA not codifing
- NT: contigs (DNA)
- XP or XM: these are referenced protein and mRNA seq, generated by insilico approach.
- CDs: coding sequence
- CON: Constructed
- EST: Expressed Sequence Tag from cDNA
- GSS: Genome Sequence Scan

manutenzione, annotazione, controllo di qualità

Il formato FASTA

```
>NM_000207.2:60-392 Homo sapiens insulin
ATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCC
CTCTGGGGACCTGACCCAGCCGCAGCCTTTGTGAACCAACACCTG
TGCGGCTCACACCTGGTGGAAGCTCTCTACCTAGTGTGCGGGGAA
CGAGGCTTCTTCTACACACCCAAGACCCGCCGGGAGGCAGAGGAC
CTGCAGGTGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGTGCAGGC
AGCCTGCAGCCCTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTGGC
ATTGTGGAACAATGCTGTACCAGCATCTGCTCCCTCTACCAGCTG
GAGAACTACTGCAACTAG
```

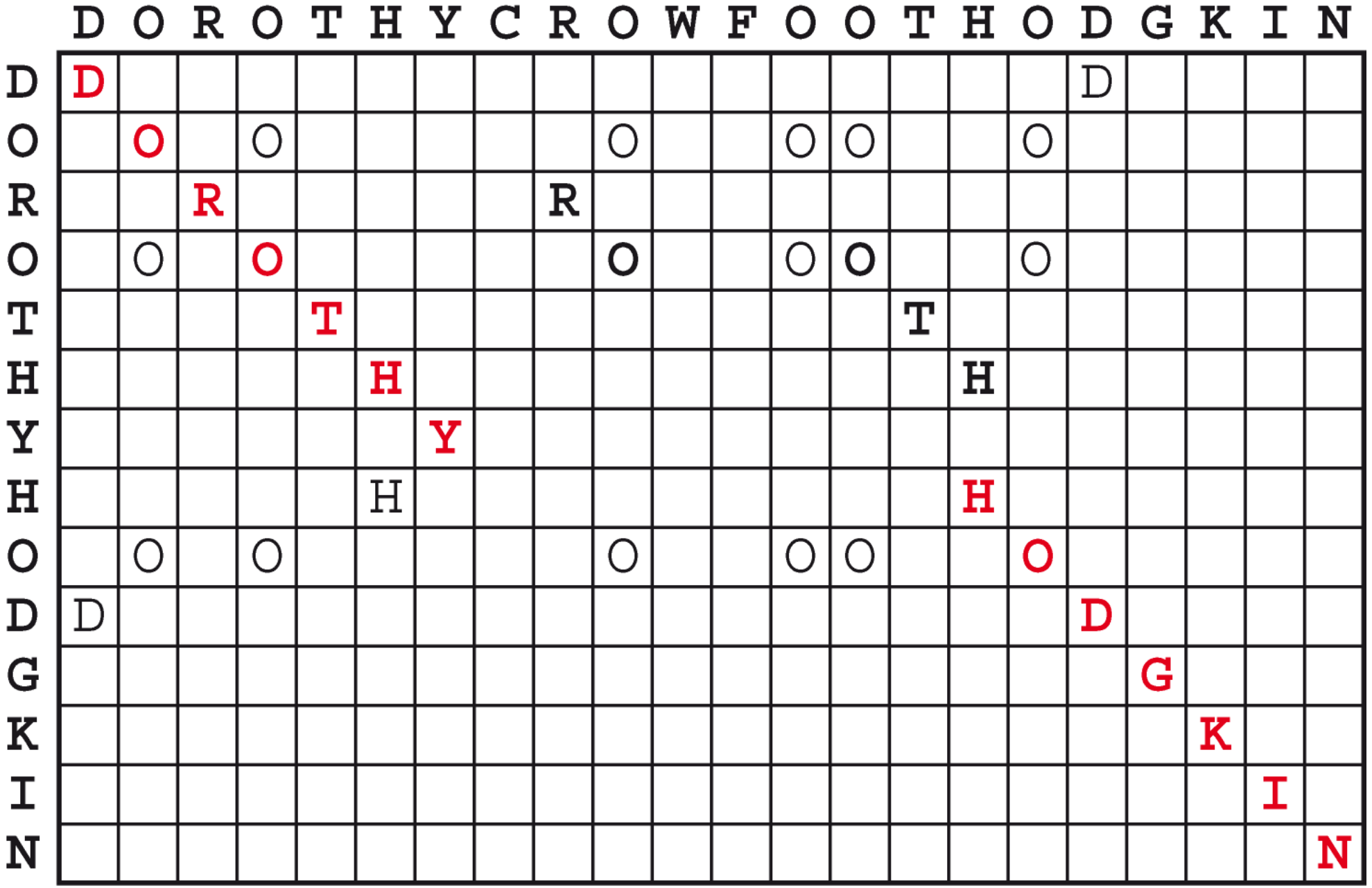
Bioinformatica

- Le banche dati.
- Programmi per estrarre ed analizzare i dati: estrarre sequenze o altri informazioni, allineamento delle sequenze, alberi filogenetici, analisi delle sequenze dei promotori, predizione dei domini, pattern nella sequenza, strutture ed interazioni.

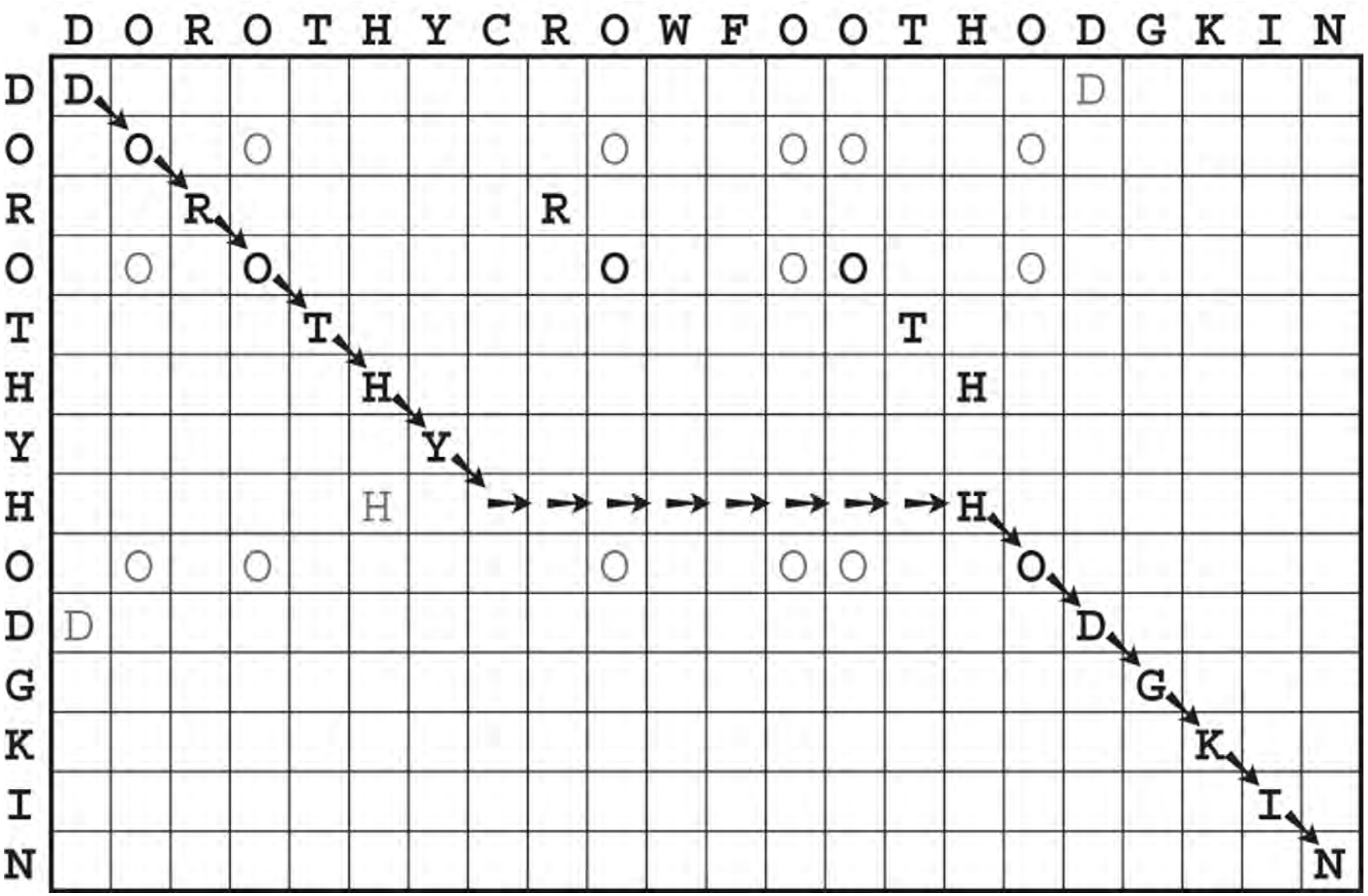
Allineamenti

- Global match
- Local match
- Motif match
- Allineamento di due sequenze
- Allineamento multiplo

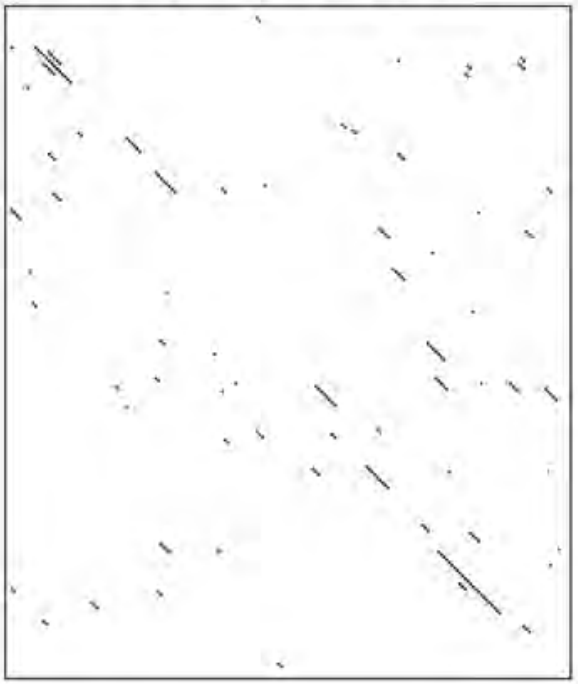
Dotplot



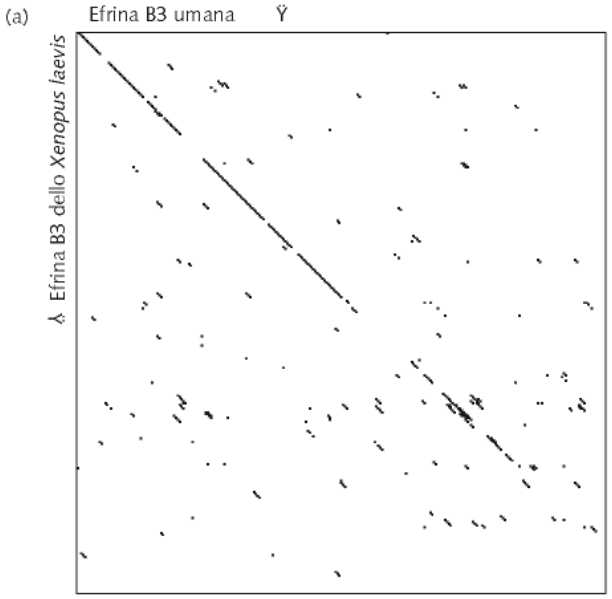
Dotplot



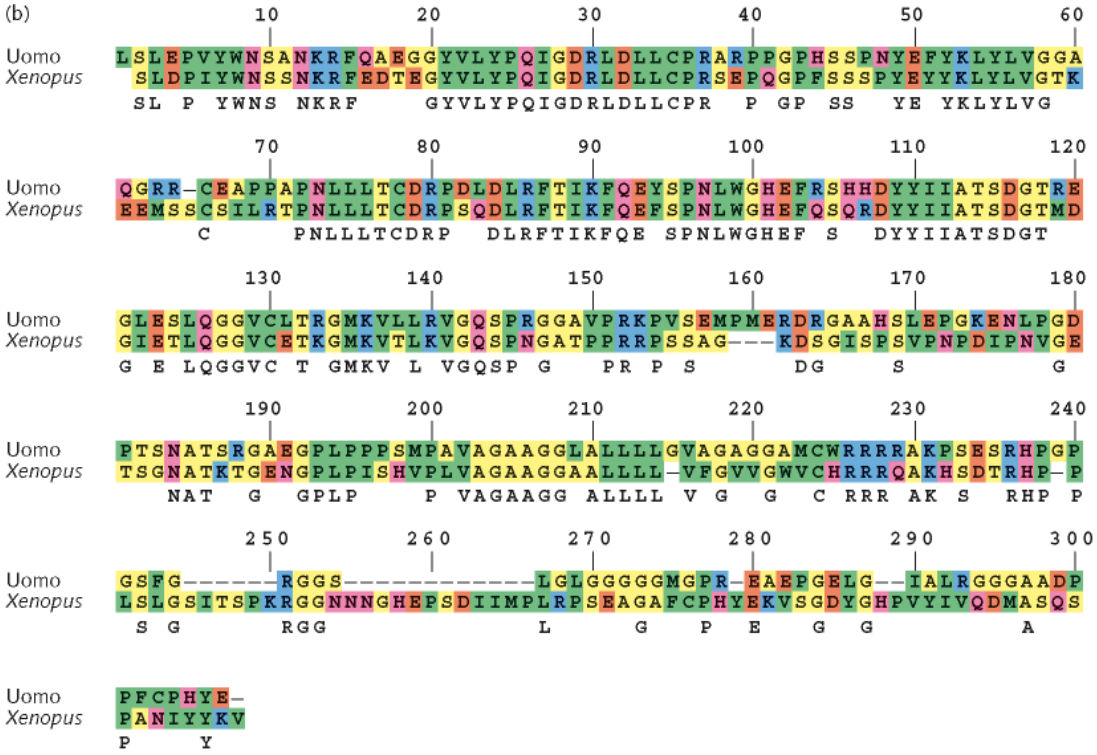
Dotplot




```
NPOS = 251   NIDENT  = 66   %IDENT = 26.29

IPEYVD-WRQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNQYSEQELLD-C-D
        | |          |||||||||| ||  |   | | |       |   ||  |
--DAREQWPQCPTIKEIRDQGSCGSCWAFGAVEAISDRICIHTNVSVEVSAEDLLTCCGS
+
RRSYGCNGGYP------WSALQLVAQYGI--HYRN-TY-----P--YEGVQRYCRSREKG
    |||||||      |    ||       |     |           |    |
MCGDGCNGGYPAEAWNFWTRKGLVSGGLYESHVGCRPYSIPPCEHHVNGSRPPCTGEGDT

PYAAK------TDGVRQVQPYNQGALLYSIANQPVSV-V-----LQ---AAGKDFQLYRG
|   |           |   |       |                          || ||
PKCSKICEPGYSPTYKQDKHYGYNSYSVSNSEKDIMAEIYKNGPVEGAFSVYSDFLLYKS

GIFVGPCGNKV-DHAVAAV--GY--GPNYILIKNSWGTGWGENGYIRIKRGTGNSYGVCG
|      |     ||      |   |  | |  ||| | || ||   | ||     |
GVYQHVTGEMMGGHAIRILGWGVENGTPYWLVANSWNTDWGDNGFFKILRGQ-DHCGIES

LYTSSFYPVKN
       |
EVVAGI-PRTD
```



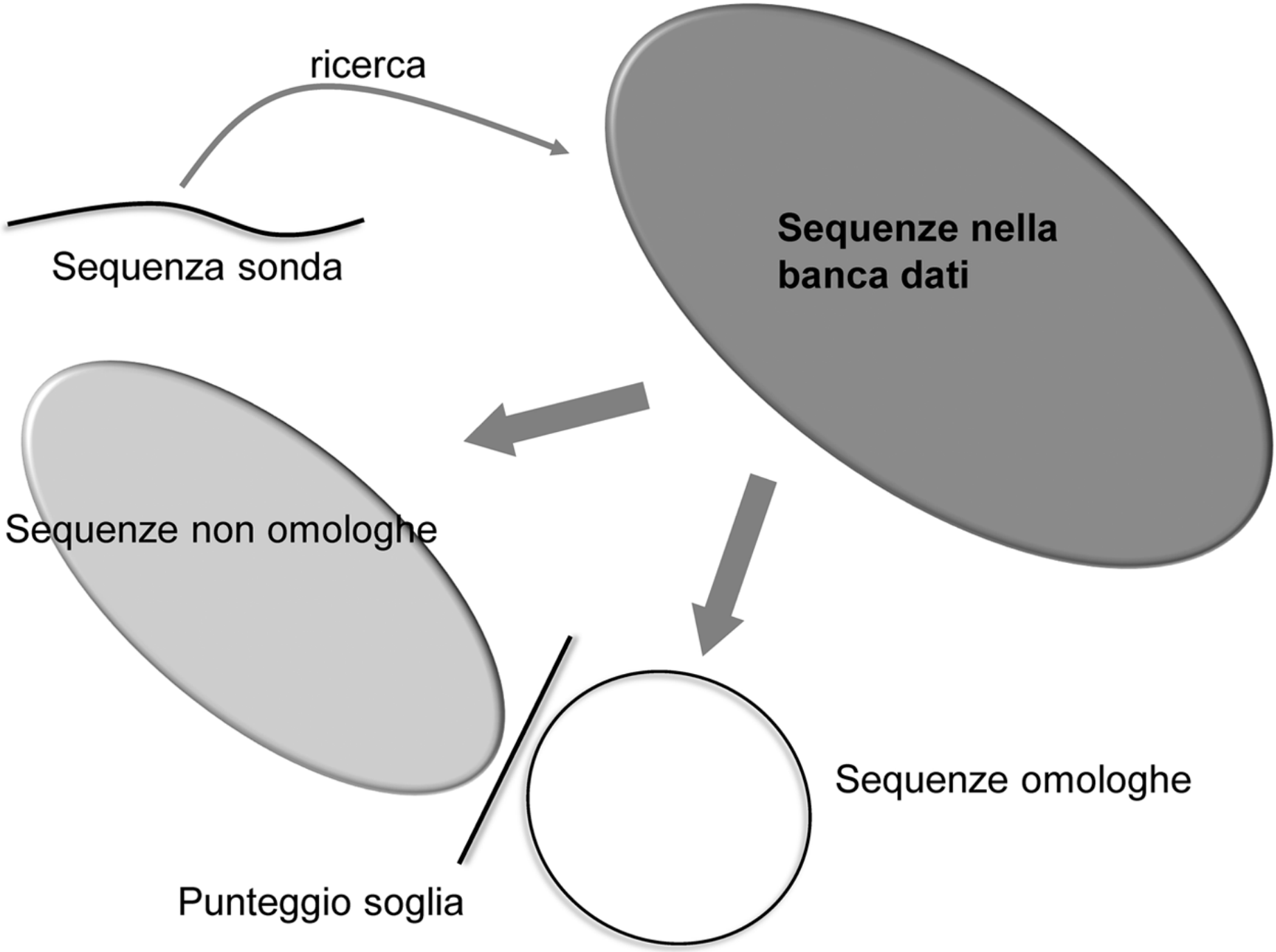
Allineamento di due sequenze

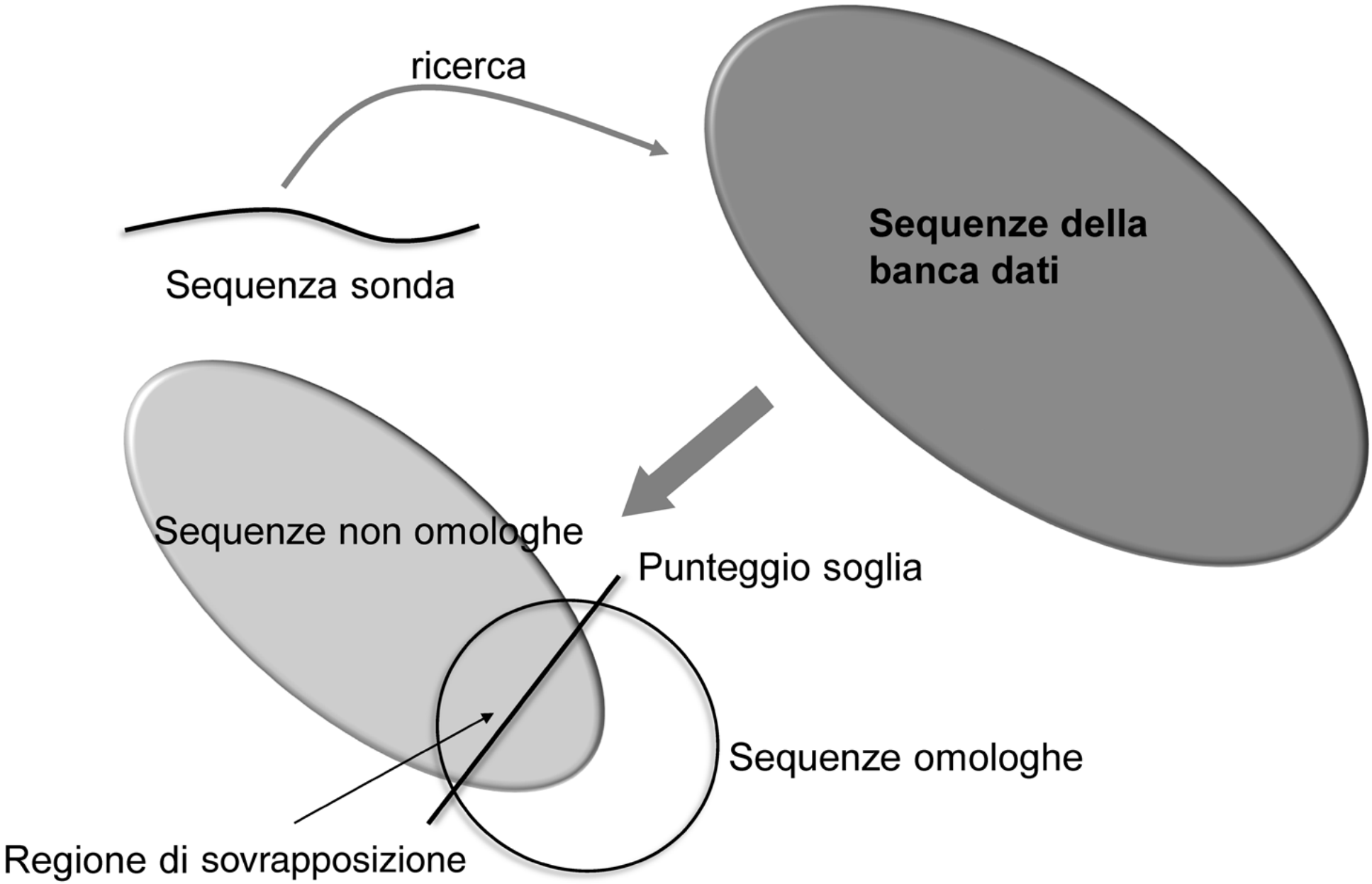


ID > 45% strutture e funzioni in commune o simili

ID> 25% ripiegamento simile

ID 18-25% twilight zone


ricerca
Sequenza sonda
Sequenze nella banca dati
Sequenze non omologhe
Sequenze omologhe
Punteggio soglia


ricerca
Sequenza sonda
Sequenze della banca dati
Sequenze non omologhe
Punteggio soglia
Sequenze omologhe
Regione di sovrapposizione

Misure della dissimilarità

- Distanza Hamming: 2

agtc

cgta

- Distanza Levenshtein (edit): 3

ag-tcc

cgctca

- Matrice per punteggio di similarità

PAM250 matrice di Dayhoff

Point Accepted Mutation Matrix

PAM	0	30	80	110	200	250
% ID	100	75	60	50	25	20

	C Cys	S Ser	T Thr	P Pro	A Ala	G Gly	N Asn	D Asp	E Glu	Q Gln	H His	R Arg	K Lys	M Met	I Ile	L Leu	V Val	F Phe	Y Tyr	W Trp
C Cys	12																			
S Ser	0	2																		
T Thr	−2	1	3																	
P Pro	−3	1	0	6																
A Ala	−2	1	1	1	2															
G Gly	−3	1	0	−1	1	5														
N Asn	−4	1	0	−1	0	0	2													
D Asp	−5	0	0	−1	0	1	2	4												
E Glu	−5	0	0	−1	0	0	1	3	4											
Q Gln	−5	−1	−1	0	0	−1	1	2	2	4										
H His	−3	−1	−1	0	−1	−2	2	1	1	3	6									
R Arg	−4	0	−1	0	−2	−3	0	−1	−1	1	2	6								
K Lys	−5	0	0	−1	−1	−2	1	0	0	1	0	3	5							
M Met	−5	−2	−1	−2	−1	−3	−2	−3	−2	−1	−2	0	0	6						
I Ile	−2	−1	0	−2	−1	−3	−2	−2	−2	−2	−2	−2	−2	2	5					
L Leu	−6	−3	−2	−3	−2	−4	−3	−4	−3	−2	−2	−3	−3	4	2	6				
V Val	−2	−1	0	−1	0	−1	−2	−2	−2	−2	−2	−2	−2	2	4	2	4			
F Phe	−4	−3	−3	−5	−4	−5	−4	−6	−5	−5	−2	−4	−5	0	1	2	−1	9		
Y Tyr	0	−3	−3	−5	−3	−5	−2	−4	−4	−4	0	−4	−4	−2	−1	−1	−2	7	10	
W Trp	−8	−2	−5	−6	−6	−7	−4	−7	−7	−5	−3	2	−3	−4	−5	−2	−6	0	0	17
	C Cys	S Ser	T Thr	P Pro	A Ala	G Gly	N Asn	D Asp	E Glu	Q Gln	H His	R Arg	K Lys	M Met	I Ile	L Leu	V Val	F Phe	Y Tyr	W Trp

The PAM-250 Log Odds Substitution Matrix

BLOSUM62 di Henikoff e Henikoff

BLOcks SUbstitution Matrix

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Inserzioni / delezioni

- Inizio alto "gap penalty"

```
aaagaaa
aaa-aaa
```

- Estensione "gap penalty" (1-10%)

```
aaaggggaaa
aaa----aaa
```

```
Per esempio:
ClustalW: gap=10, estenzione=0.1
BLOSUM62: gap=11, estenzione=1
```

Dynamic programming

- Metodo per trovare l'allineamento ottimale (globale) tra due sequenze
- + trova sempre l'allineamento con miglior punteggio (dato matrice di sostituzione e gap penalty)
- - trovare l'allineamento biologicamente corretto
- - il tempo di allineare le sequenze con n e m unità e proporzionale a n x m (non conveniente per ricerche in banche dati)

Screening delle banche dati

metodi approssimativi e con allineamenti locali

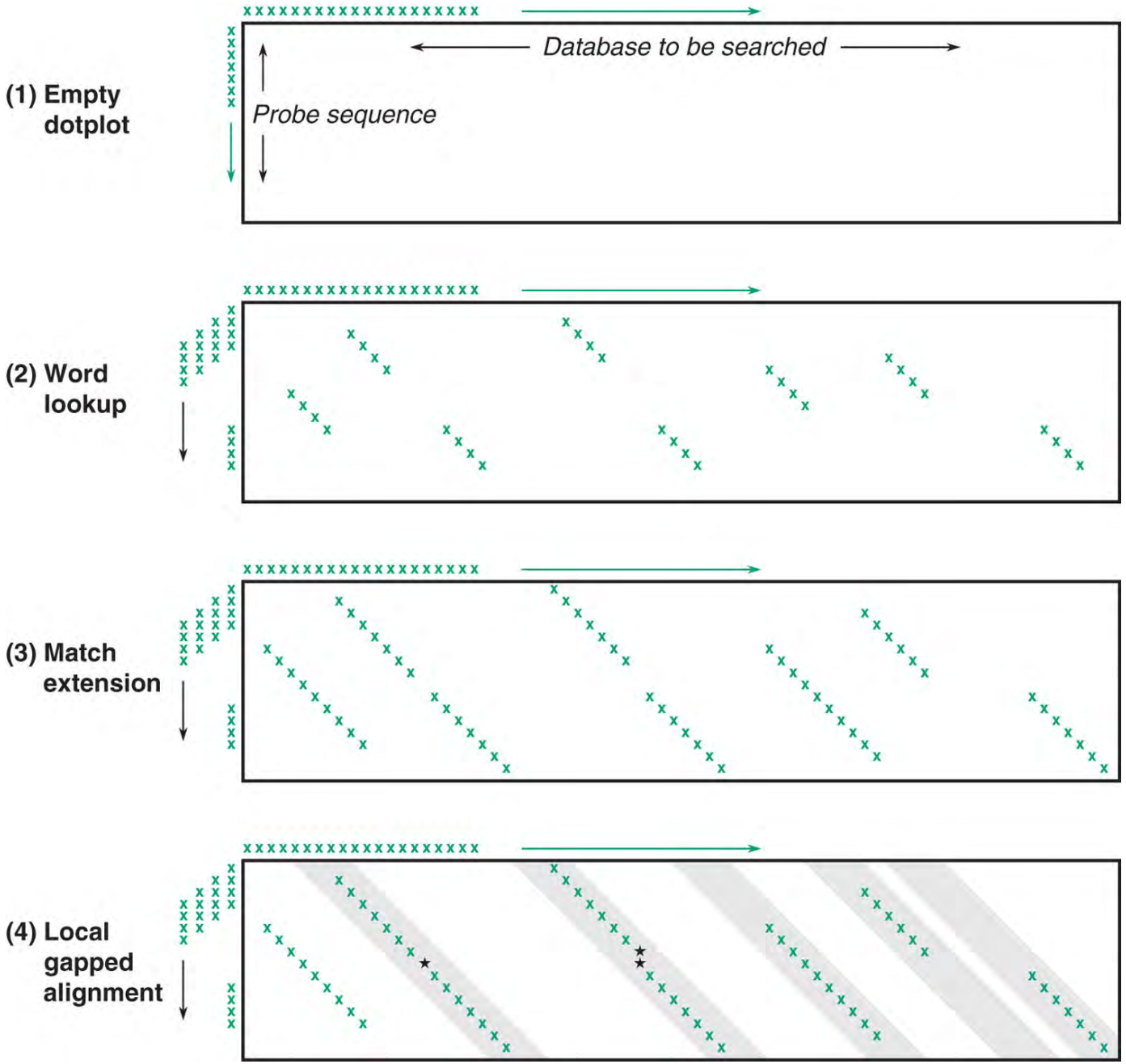
Allineamento con BLAST

- BLAST Basic Local Alignment Search Tool

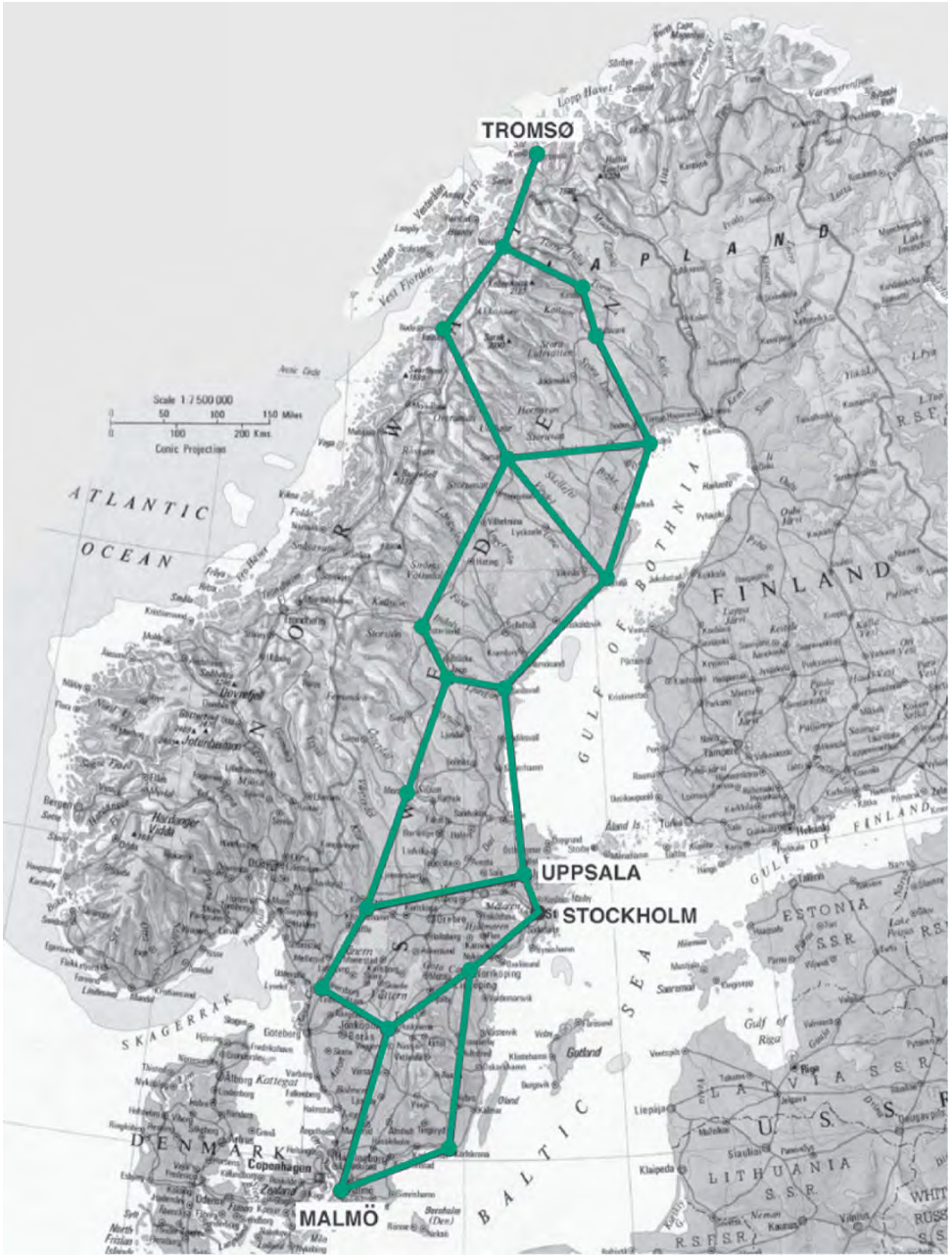
BLASTN, BLASTP

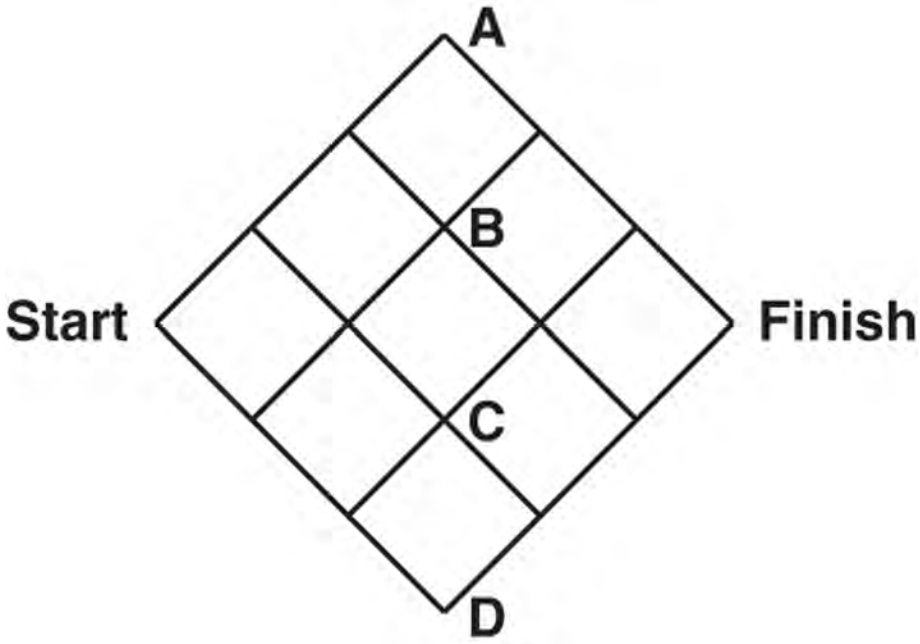
- Sensibilità, selettività e velocità

BLAST

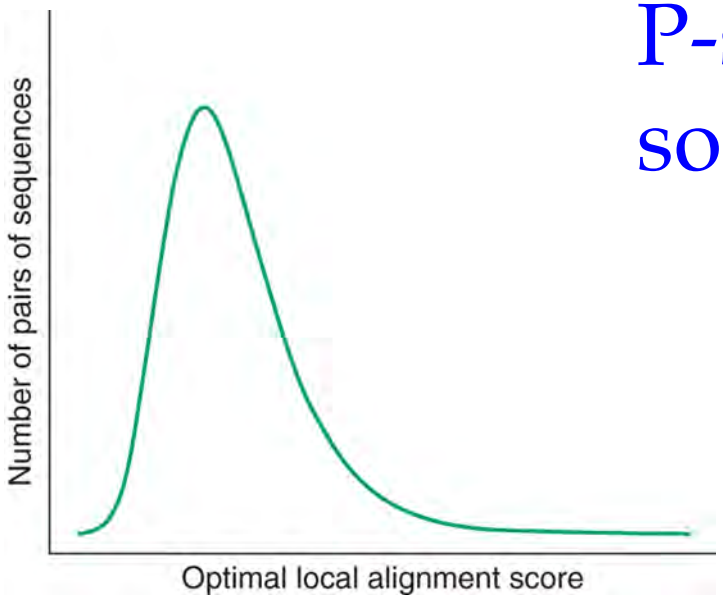

(1) Empty dotplot
Database to be searched
Probe sequence
(2) Word lookup
(3) Match extension
(4) Local gapped alignment

Programma dinamica





La significatività dell'allineamento



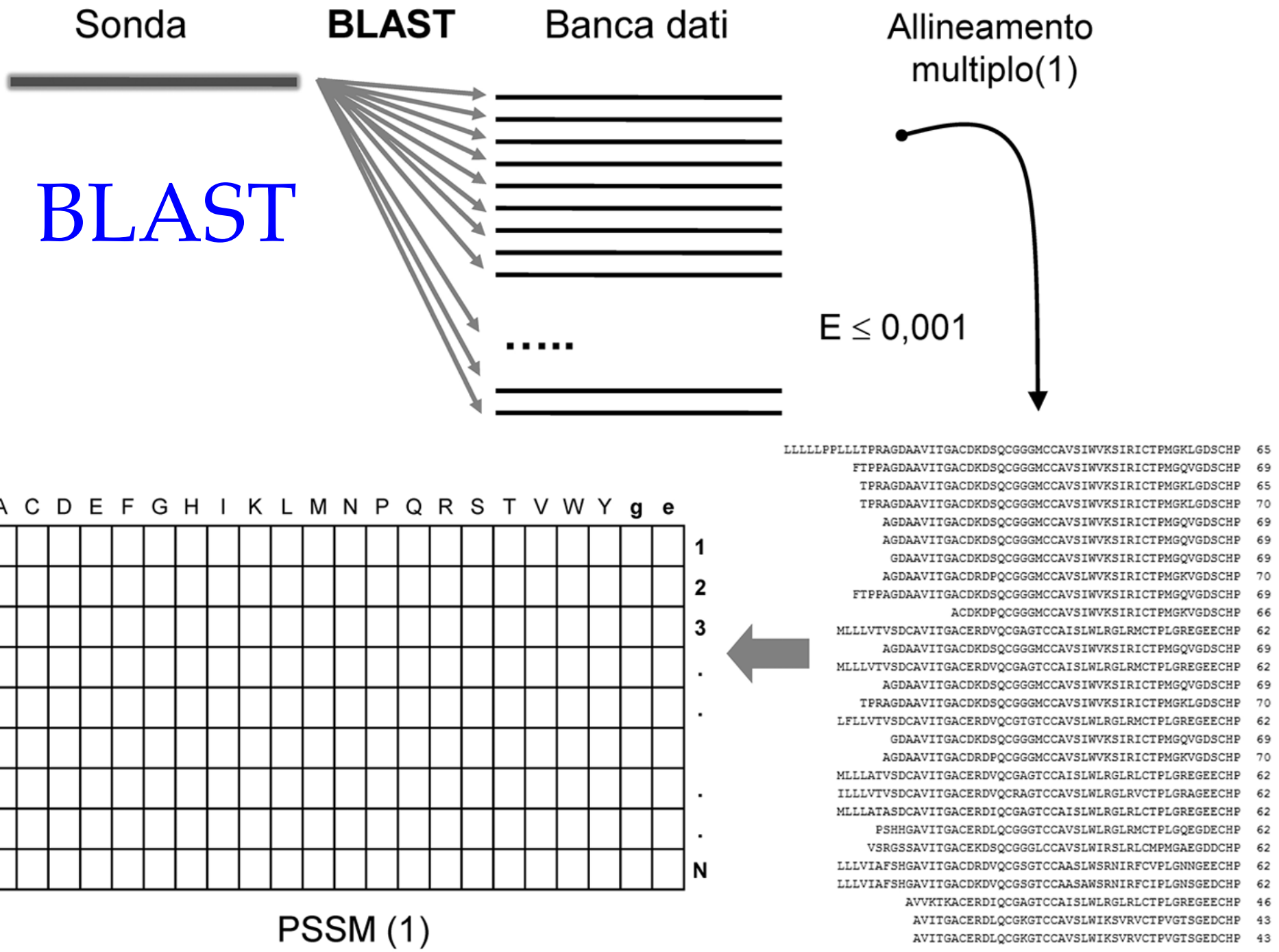
P-score: probabilità che la somiglianza è casuale

- P ≤ 10^{-100} match esatto
- P 10^{-100}-10^{-50} quasi ID (SNP o alleli)
- P 10^{-50}-10^{-10} omologia vicina
- P 10^{-5}-10^{-1} omologia distante
- P > 10^{-1} probabilmente insignificante

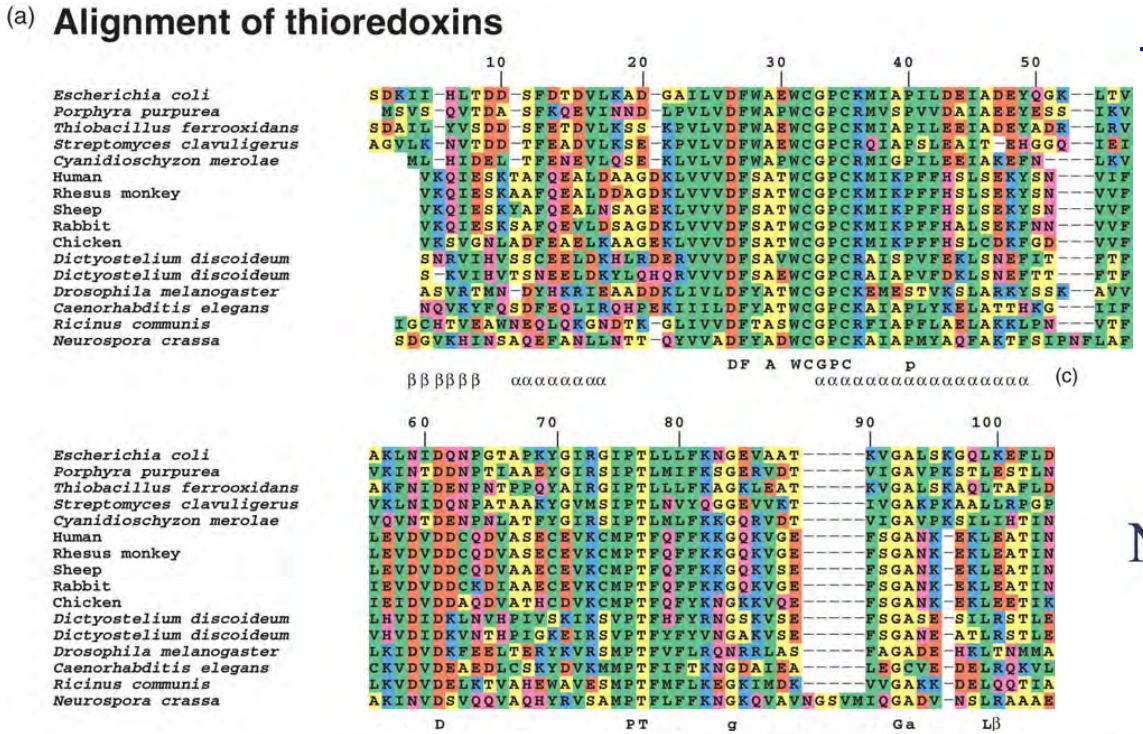
Z-score: ≥ 5 significativo

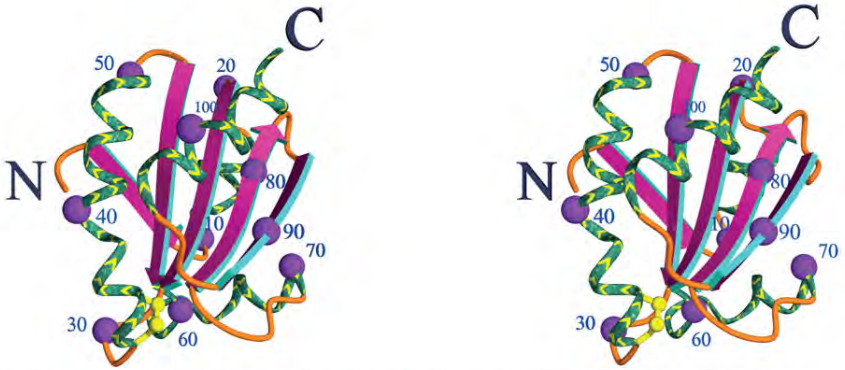
E value: probabilità di trovare match migliore

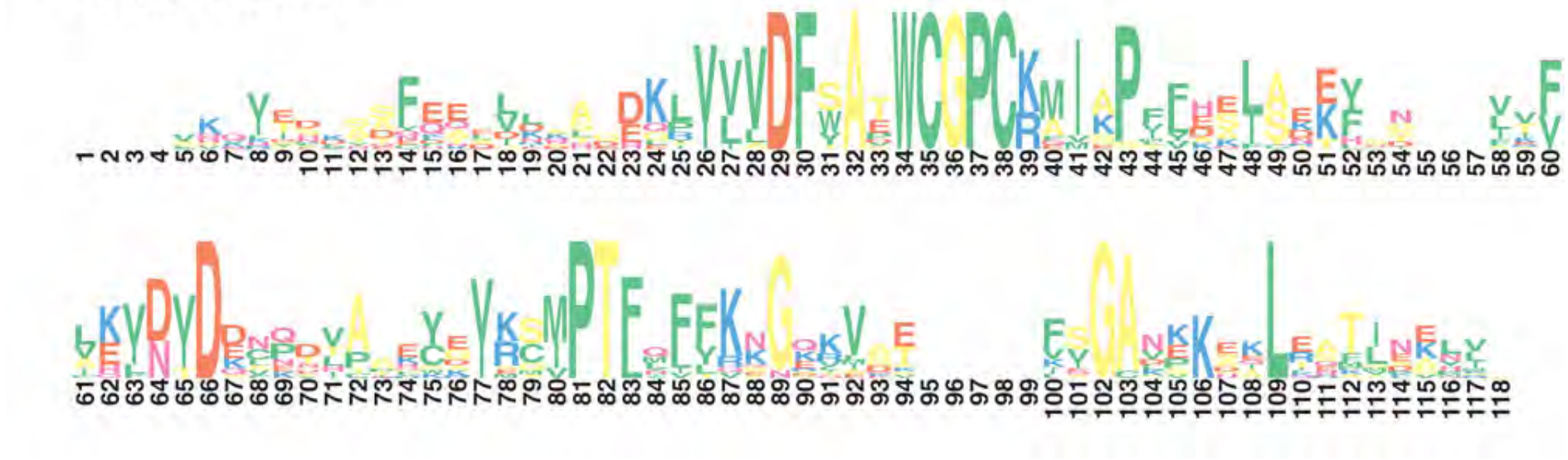
- E ≤ 0.02 sequenze probabilmente omologhe
- E 0.02-1 omologia non sicura
- E > 1 match casuale


Sonda
BLAST
Banca dati
Allineamento multiplo(1)
BLAST
E ≤ 0,001
LLLLLPPLLLTPRAGDAAVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGKLGDSCHP 65
FTPPAGDAAVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGQVGDSCHP 69
TPRAGDAAVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGKLGDSCHP 65
TPRAGDAAVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGKLGDSCHP 70
AGDAAVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGQVGDSCHP 69
AGDAAVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGQVGDSCHP 69
GDAAVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGQVGDSCHP 69
AGDAAVITGACDRDPQCGGGMCCAVSLWVKSIRICTPMGKVGDSCHP 70
FTPPAGDAAVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGQVGDSCHP 69
ACDKDPQCGGGMCCAVSIWVKSIRICTPMGKVGDSCHP 66
MLLLVTVSDCAVITGACERDVQCGAGTCCAISLWLRGLRMCTPLGREGEECHP 62
AGDAAVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGQVGDSCHP 69
MLLLVTVSDCAVITGACERDVQCGAGTCCAISLWLRGLRMCTPLGREGEECHP 62
AGDAAVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGQVGDSCHP 69
TPRAGDAAVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGKLGDSCHP 70
LFLLVTVSDCAVITGACERDVQCGTGTCCAVSLWLRGLRMCTPLGREGEECHP 62
GDAAVITGACDKDSQCGGGMCCAVSIWVKSIRICTPMGQVGDSCHP 69
AGDAAVITGACDRDPQCGGGMCCAVSLWVKSIRICTPMGKVGDSCHP 70
MLLLATVSDCAVITGACERDVQCGAGTCCAISLWLRGLRLCTPLGREGEECHP 62
ILLLVTVSDCAVITGACERDVQCRAGTCCAVSLWLRGLRVCTPLGRAGEECHP 62
MLLLATASDCAVITGACERDIQCGAGTCCAISLWLRGLRLCTPLGREGEECHP 62
PSHHGAVITGACERDLQCGGGTCCAVSLWLRGLRMCTPLGQEGDECHP 62
VSRGSSAVITGACEKDSQCGGGLCCAVSLWIRSLRLCMPMGAEGDDCHP 62
LLLVIAFSHGAVITGACDRDVQCGSGTCCAASLWSRNIRFCVPLGNNGEECHP 62
LLLVIAFSHGAVITGACDKDVQCGSGTCCAASAWSRNIRFCIPLGNSGEDCHP 62
AVVKTKACERDIQCGAGTCCAISLWLRGLRLCTPLGREGEECHP 46
AVITGACERDLQCGKGTCCAVSLWIKSVRVCTPVGTSGEDCHP 43
AVITGACERDLQCGKGTCCAVSLWIKSVRVCTPVGTSGEDCHP 43
A C D E F G H I K L M N P Q R S T V W Y g e
1
2
3
N
PSSM (1)

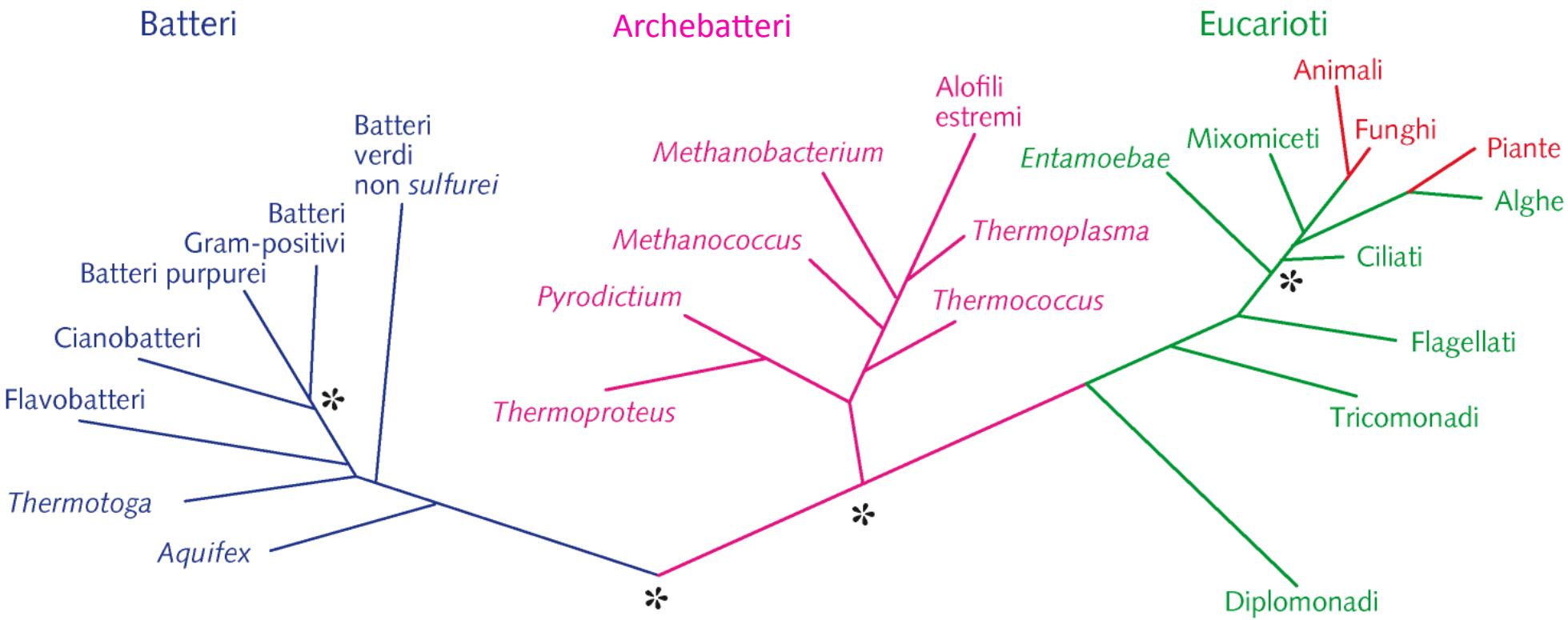
Allineamento multiplo di thioredoxins





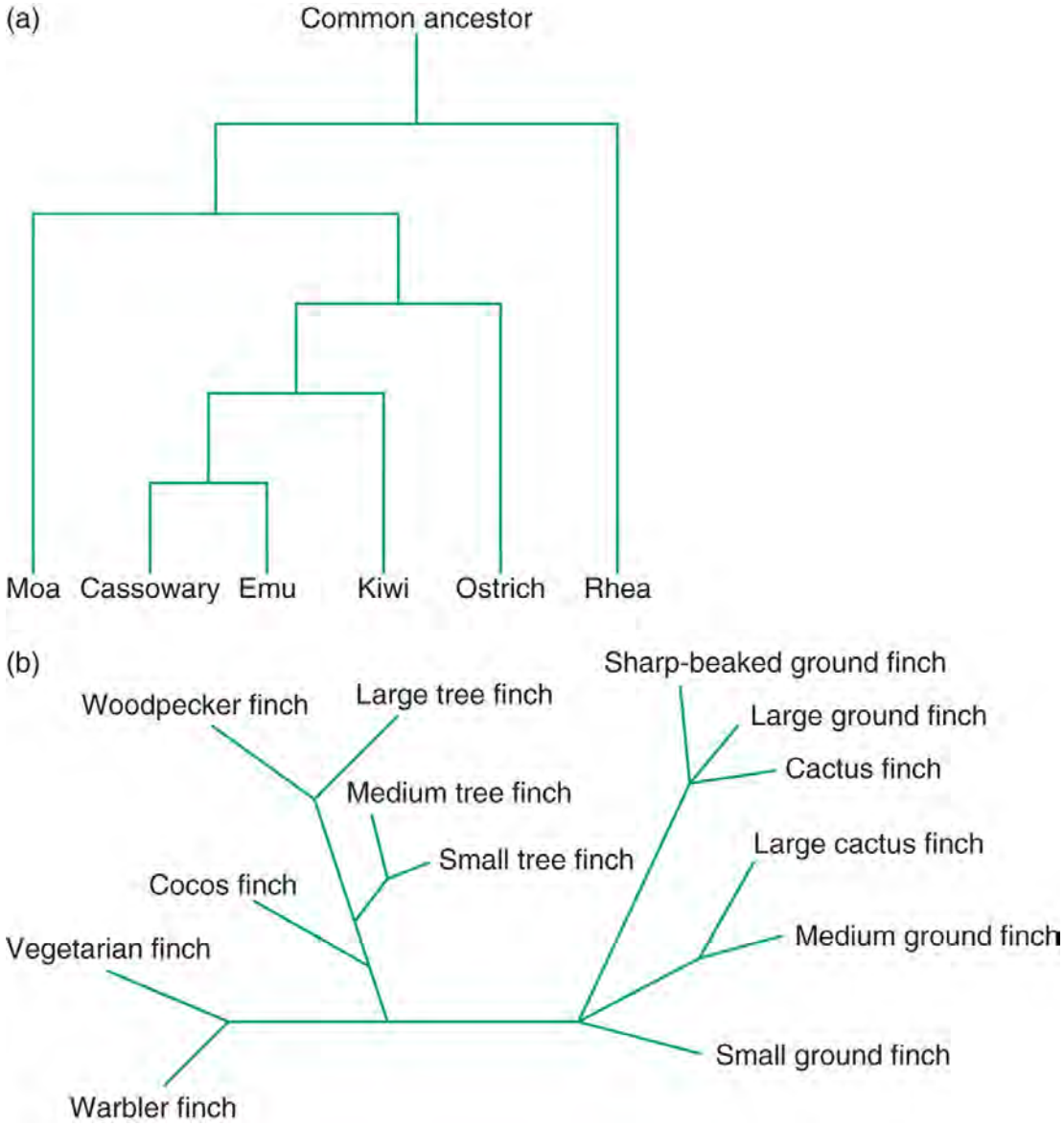


Albero filogenetico



Alberi filogenetici

- Omologia
- Identità
- Similarità
- Clustering
- Evoluzione divergente



Programmi per analizzare proteine su ExPASy

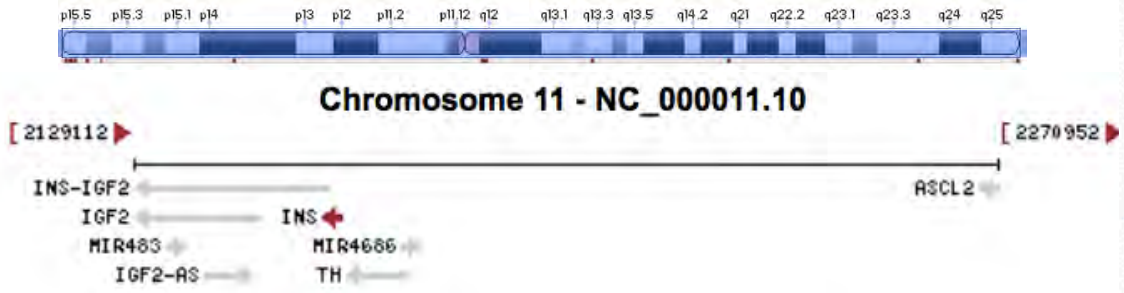
- **ProtParam**
- **Mascot**
- **ScanProsite**
- **NetNGlyc**
- **TargetP**
- **Pfam**
- **SMART**
- **SignalP**
- **TMHMM**

SDSC Biology Workbench
(http://workbench.sdsc.edu)

Scopo

- Ritrovare la sequenza del gene per una proteina (human insulin) e la sequenza proteica
- Analizzare la sequenza proteica dell'insulina umana
- Trovare omologhi dell'insulina umana con BLAST
- Fare un'allineamento multiplo
- Fare un albero filogenetico con gli omologhi dell'insulina
- Analizzare la struttura dell'insulina e le interazioni con il suo recettore

L'insulina



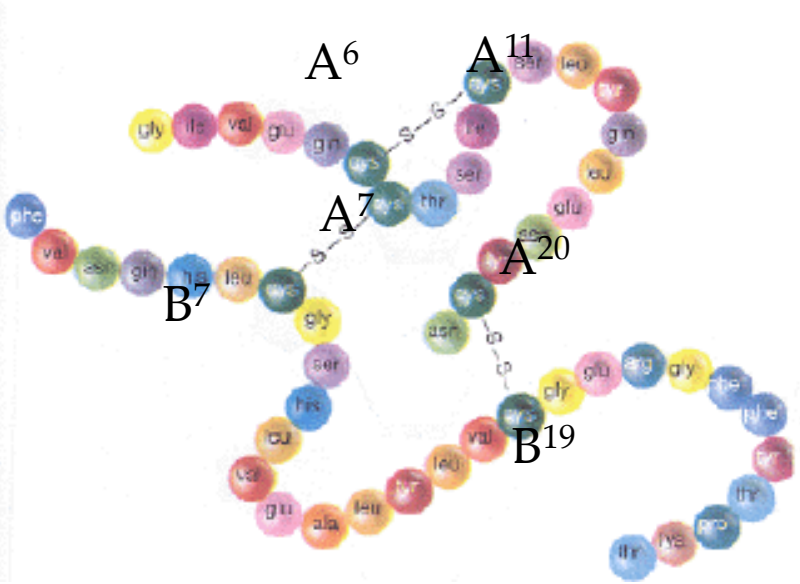
trascrizione

mRNA

traduzione

Pre-proinsulina

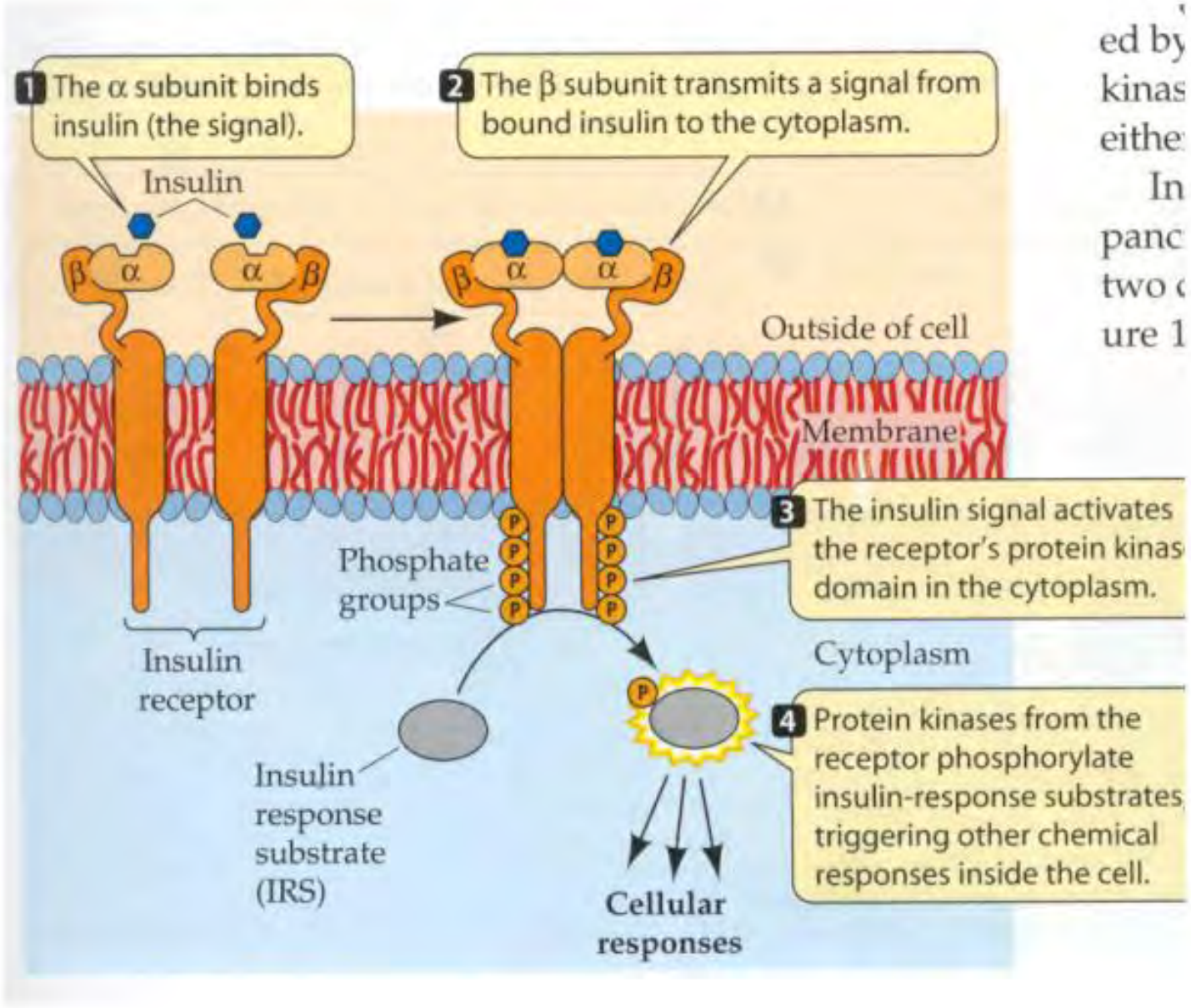
Catena A 21 amminoacidi



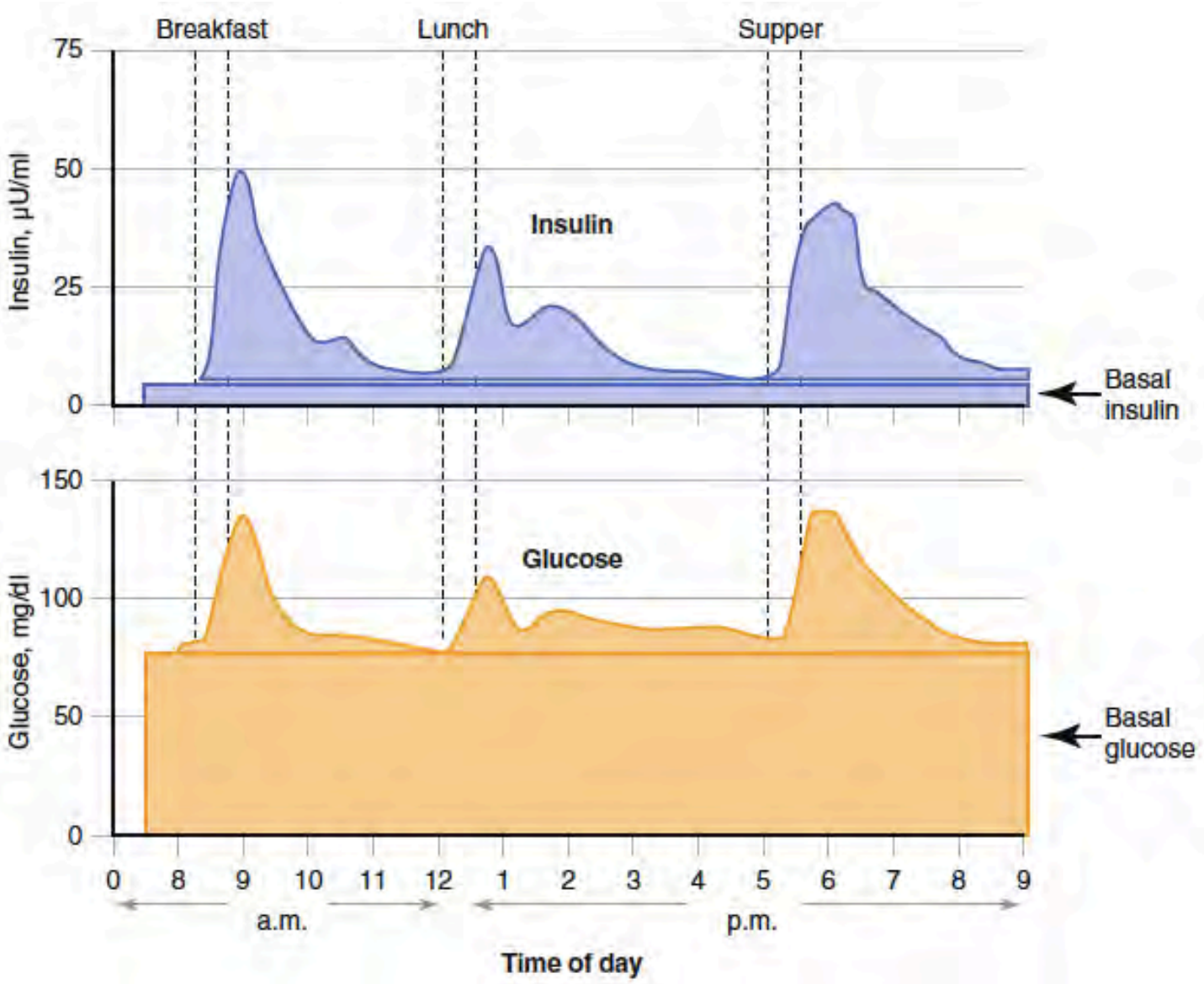
Catena B 30 amminoacidi

Specie	Catena A		Catena B
Umana	Thr8	Ile10	Thr30
Suina	Thr8	Ile10	Ala30
Bovina	Ala8	Val10	Ala30

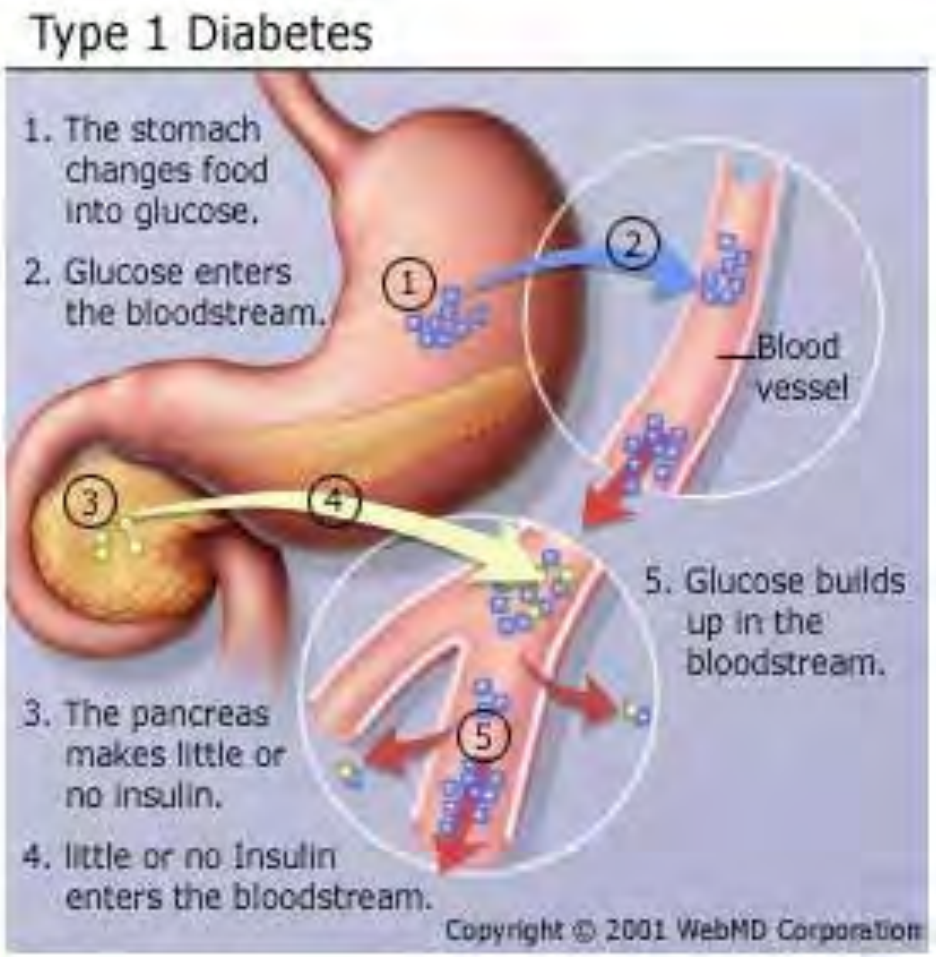
Il recettore dell'insulina

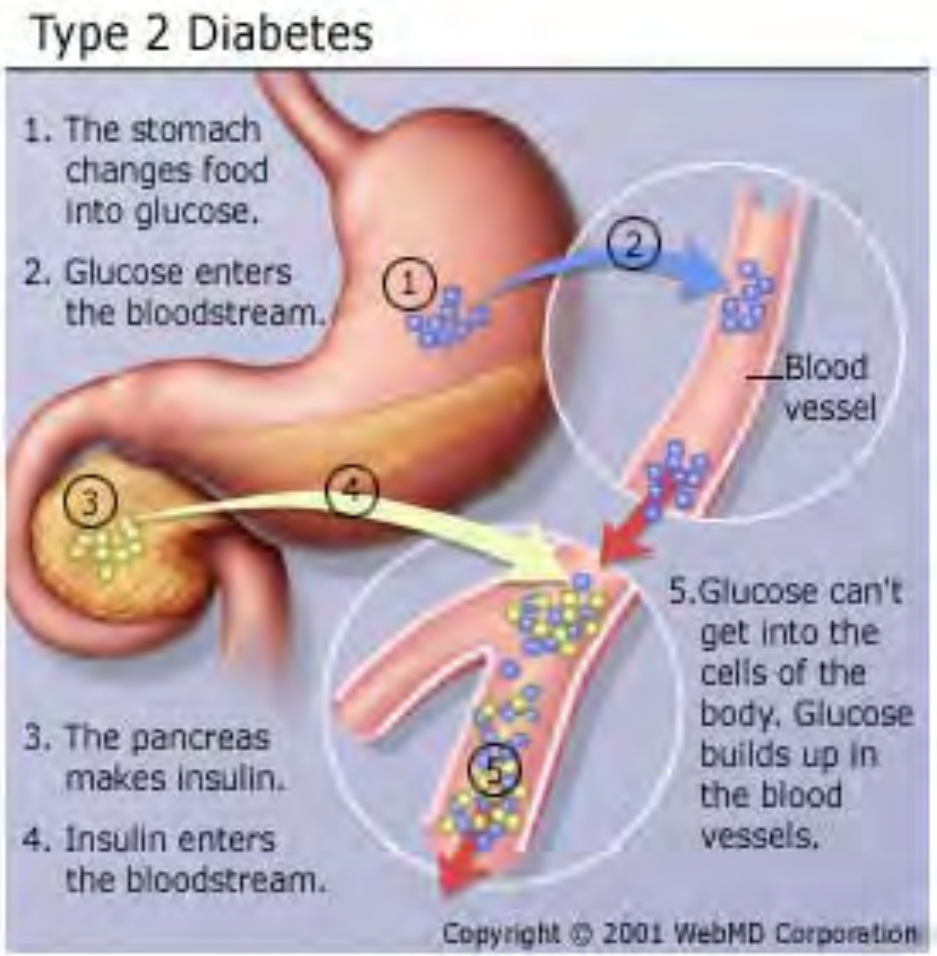


Effetti con insulina



Diabete tipo I e II





Ritrovare la sequenza dell'insulina umana ed analizzarla

Entrez NCBI http://www.ncbi.nlm.nih.gov/sites/gquery

-Cercare "human insulin", selezionare i hit di "Gene"

-Copiare la sequenza in formato "FASTA"

-Cliccare su "BLAST" (a destra) ed inserire la sequenza in "Protein BLAST"

-Selezionare un numero di sequenze omologhe e cliccare su multiple alignment

-Creare un albero filgenetico

Analizzare la sequenza proteica (domini, modificazioni post-traduzionale)

PFAM http://pfam.xfam.org

2. Come si chiama il dominio nella proteina?