Key-sites (1)

1) **National Center for Biotechnology Information** (NCBI): www.ncbi.nlm.nih.gov

- GQuery
- Pubmed
- Gene

2) **European Bioinformatics Institute** (EBI): www.ebi.ac.uk

- Services

Key-sites (2)

4) **Nucleic Acid Research** (NAR): nar.oxfordjournals.org

Database Issue → Compilation paper (Tables)

Web Server Issue → Editorial (bioinformatics.ca/links_directory)

5) **Wikipedia**: www.wikipedia.org

6) **Google**: www.google.com

Sequence-structure-function relationships

... ACG TCA GTA CAT CCG TAA ...

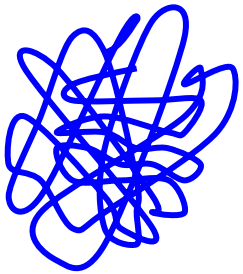
Sequenza gene

Codice genetico

... T S V H P K ...

Sequenza aminoacidica

Protein folding

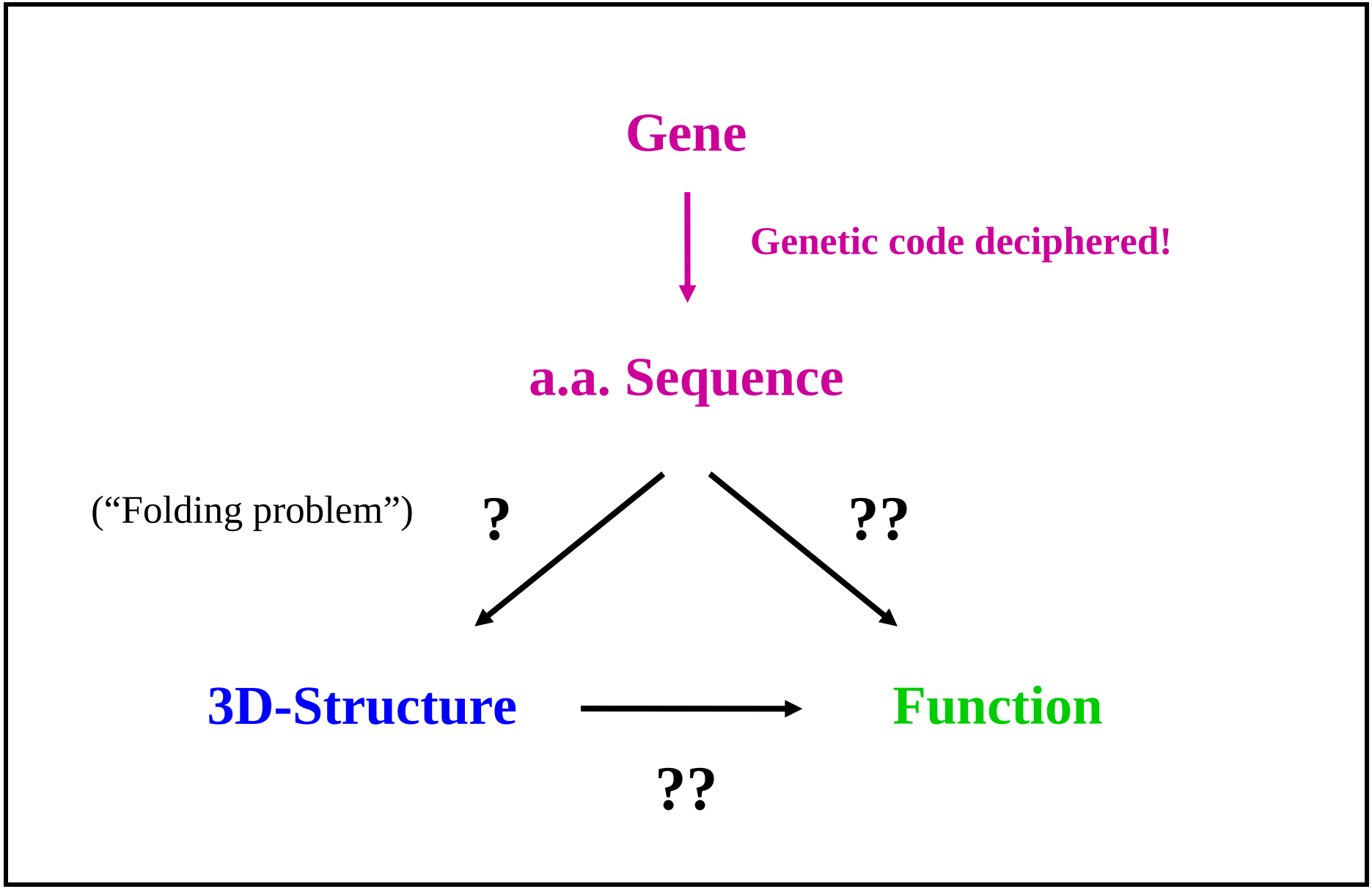


Struttura proteica

ATP → ADP+P

Funzione biologica

Sequence-structure-function relationships



Gene-protein sequence

Gene (nt sequence) → **Genetic code deciphered!** → **a.a. Sequence (primary structure)**

Algorithm: nt triplet – aa correspondence

...
ACG
TCA
GTA
CAT
CCG
TAA
...

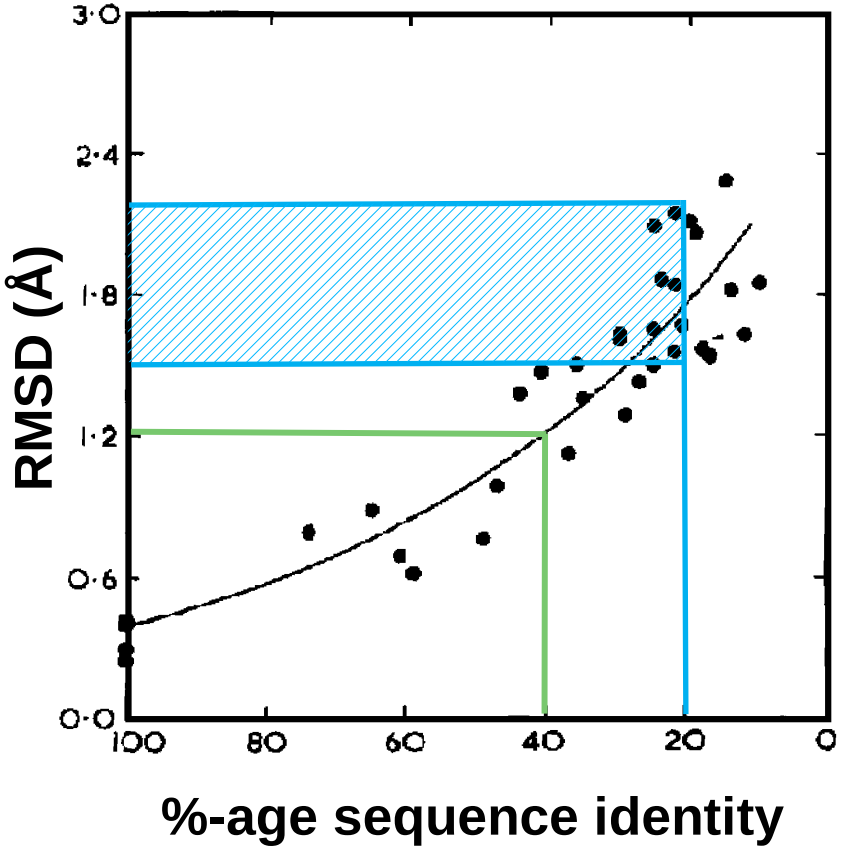
GGG GGA GGC GGT	CGG CGA CGC CGT	AGG AGA AGC AGT	TGG TGA TGC TGT
GCG GCA GCC GCT	CCG CCA CCC CCT	ACG ACA ACC ACT	TCG TCA TCC TCT
GAG GAA	CAG CAA	AAG AAA	TAG TAA
GAC GAT	CAC CAT	AAC AAT	TAC TAT
GTG GTA GTC GTT	CTG CTA CTC CTT	ATG ATA ATC ATT	TTG TTA TTC TTT

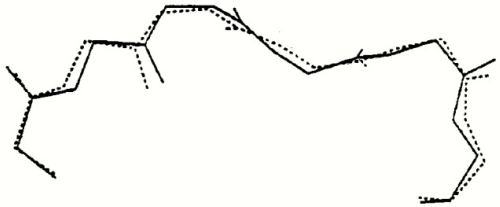
Gly	Arg	***	Trp
		Ser	Cys
Ala	Pro	Thr	Ser
Glu	Gln	Lys	***
Asp	His	Asn	Tyr
Val	Leu	Met	Leu
		Ile	Phe

...
T
S
V
H
P
K
...

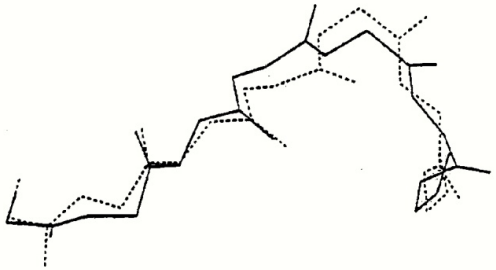
Sequence-structure relationships

Structure similarity

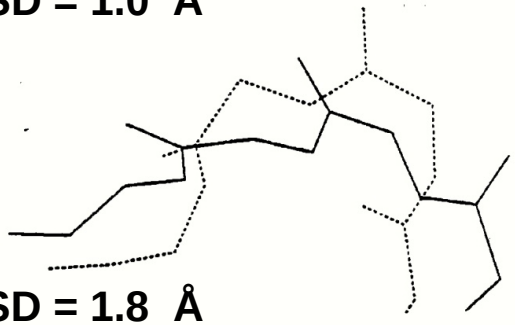




RMSD = 0.5 Å



RMSD = 1.0 Å



RMSD = 1.8 Å

Similar sequences → Similar structures

Sequence-structure relationships

a.a. Sequence $\xrightarrow{?}$ **3D-Structure**

"Folding problem"

i) Structure of protein A is known

ii) Structure of protein B is unknown

iii) Proteins A and B have similar sequences => similar structures

=> Structure of protein B can be modelled on the structure of protein A

(Homology / Comparative Modelling)

evolutionary	Relationships between sequences and structures in DBs	**+++**
ab initio	Physical-Chemical Principles	**-**

Homology

Homology = Evolutionary relationship = Common ancestor

'*High*' or '*Low*' Homology	**No!!!**
'**Close**' or '**Distant**' Homology	**Yes**
'*Measured*' Homology	**No**

Homology is inferred from measurable parameters:

- %-age of sequence identity
- structure similarity (RMSD)

Protein sequence analysis

- Name (full name or symbol)

- Sequence identifier (s)

- Name or identifier of homologue in different species

- ...

e.g., granulocyte colony-stimulating factor

National Center for Biotechnology Information (NCBI):
www.ncbi.nlm.nih.gov

- GQuery
- Pubmed
- Gene

Protein sequence analysis

Fasta format:

>1-line
AA sequence (1 or more lines)

```
>gi|4503079|ref|NP_000750.1| granulocyte colony-stimulating
factor isoform a precursor [Homo sapiens]
MAGPATQSPMKLMALQLLLWHSALWTVQEATPLGPASSLPQSFLLKCLEQVRKIQGDGAALQ
EKLVSECATYKLCHPEELVLLGHSLGIPWAPLSSCPSQALQLAGCLSQLHSGLFLYQGLLQA
LEGISPELGPTLDTLQLDVADFATTIWQQMEELGMAPALQPTQGAMPAFASAFQRRAGGVLV
ASHLQSFLEVSYRVLRHLAQP
```

3D structure? Biological function?

Residues essential for 3D structure / function?

<u>Save Fasta sequence</u>

Protein sequence analysis

1) NCBI gene annotation

2) UniProt annotation

3) BLAST

Sequence Comparison Methods (SCM)

Blast

1.) Provide the **Query** sequence
in Fasta format

```
>gi|4503079|ref|NP_000750.1| granulocyte
colony-stimulating factor isoform a precursor
[Homo sapiens]
MAGPATQSPMKLMALQLLLWHSALWTVQEATPLGPASSLPQSFLLKC
LEQVRKIQGDGAALQEKLVSECATYKLCHPEELVLLGHSLGIPWAPL
SSCPSQALQLAGCLSQLHSGLFLYQGLLQALEGISPELGPTLDTLQL
DVADFATTIWQQMEELGMAPALQPTQGAMPAFASAFQRRAGGVLVAS
HLQSFLEVSYRVLRHLAQP
```

2.) Select the sequence **Database**

NR; PDB; RefSeq; SwissProt; Organism; Exclude; Specialized

Sequence Comparison Methods (SCM)

BLAST

3.) Algorithm **parameters:**

- Max target sequences
- Expect threshold
- Filter Low complexity regions

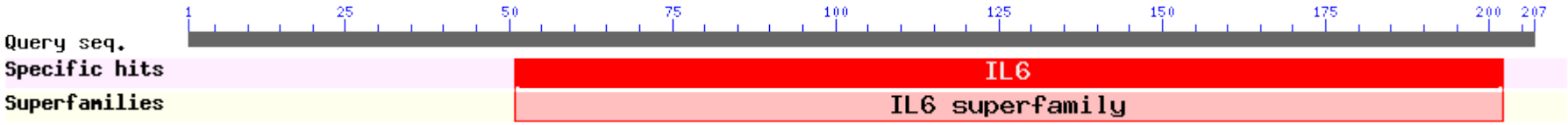
Sequence Comparison Methods (SCM)

Program output

```
>gi|4503079|ref|NP_000750.1| granulocyte colony-stimulating factor isoform a precursor
[Homo sapiens]
MAGPATQSPMKLMALQLLLWHSALWTVQEATPLGPASSLPQSFLLKCLEQVRKIQGDGAALQEKLVSECATYKLCHPEELVLLGHSL
GIPWAPLSSCPSQALQLAGCLSQLHSGLFLYQGLLQALEGISPELGPTLDTLQLDVADFATTIWQQMEELGMAPALQPTQGAMPAFA
SAFQRRAGGVLVASHLQSFLEVSYRVLRHLAQP
```

1.) Conserved domains (CDD)

Putative conserved domains have been detected, click on the image below for detailed results.



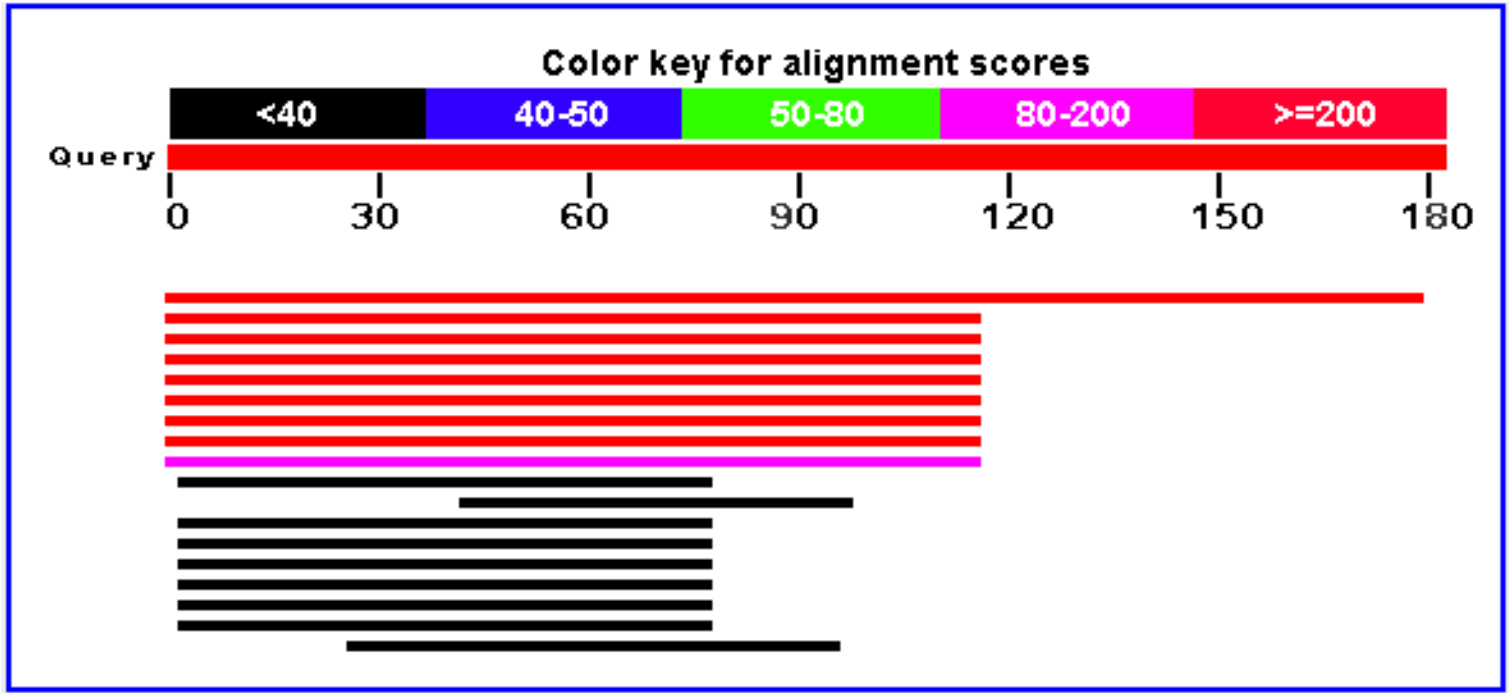
Domain-based DBs:

CDD; Pfam; SMART; Superfamily; etc.

Sequence Comparison Methods (SCM)

2.) Graphic view of matched sequences

Distribution of Blast Hits on the Query Sequence



Sequence Comparison Methods (SCM)

3.) List of matched sequences

▼ Descriptions

Sequences producing significant alignments:	Score (Bits)	E Value	
gb|ACF41164.1| granulocyte colony stimulating factor [Homo sa...	351	8e-97	
pdb|1CD9|A Chain A, 2:2 Complex Of G-Csf With Its Receptor >p...	206	3e-53	S
pdb|1RHG|A Chain A, The Structure Of Granulocyte-Colony-Stimu...	206	4e-53	S
ref|NP_757373.1| colony stimulating factor 3 isoform b precur...	205	5e-53	U G
ref|NP_757374.1| colony stimulating factor 3 isoform c [Homo ...	205	7e-53	U G
gb|AAA03056.1| human granulocyte-colony stimulating factor	204	1e-52	G
pdb|1GNC|A Chain A, Structure And Dynamics Of The Human Granu...	201	1e-51	S
ref|NP_000750.1| colony stimulating factor 3 isoform a precur...	200	2e-51	U G
dbj|BAG60399.1| unnamed protein product [Homo sapiens]	114	2e-25	G
dbj|BAG62733.1| unnamed protein product [Homo sapiens]	33.5	0.40	G
dbj|BAG62358.1| unnamed protein product [Homo sapiens]	33.1	0.45	G
ref|NP_000591.1| interleukin 6 precursor [Homo sapiens] >sp|P...	33.1	0.51	U G
pdb|1IL6|A Chain A, Human Interleukin-6, Nmr, Minimized Avera...	33.1	0.52	S
emb|CAA27991.1| 26 kDa protein (aa 32-212) [Homo sapiens]	33.1	0.52	G
pdb|1P9M|B Chain B, Crystal Structure Of The Hexameric Human ...	32.7	0.55	S
pdb|1ALU|A Chain A, Human Interleukin-6	32.7	0.55	S
dbj|BAG60317.1| unnamed protein product [Homo sapiens]	32.7	0.67	G
dbj|BAG62348.1| unnamed protein product [Homo sapiens]	32.3	0.91	G

S — Related Structures

G — Gene Info

U — UniGene Info

Sequence Comparison Methods (SCM)

3.) List of matched sequences

Description → pair-wise alignment

Query cover → %age of input sequence matched

E-value → probability that the matched sequence is not homologous

Max ident → % of sequence identity of the longest fragment

Accession → page with protein description

Sequence Comparison Methods (SCM)

4.) Alignments of matched sequences to Query (download)

4.1) 'Easy' Results: clear homology

Alignments Select All Get selected sequences Distance tree of results Multiple alignment NEW

```
>gb|ACF41164.1|  granulocyte colony stimulating factor [Homo sapiens]
Length=182

 Score =  351 bits (900),  Expect = 8e-97, Method: Compositional matrix adjust.
Identities = 179/179 (100%), Positives = 179/179 (100%), Gaps = 0/179 (0%)

Query  1    CLEQVRKIQDDGAALQEKLCATYKLCHPEELVLLGHSLGIPWAPLSSCPSQALQLAGCLS  60
            CLEQVRKIQDDGAALQEKLCATYKLCHPEELVLLGHSLGIPWAPLSSCPSQALQLAGCLS
Sbjct  1    CLEQVRKIQDDGAALQEKLCATYKLCHPEELVLLGHSLGIPWAPLSSCPSQALQLAGCLS  60

Query  61   QLHSGLFLYQGLLQALEGISPELGPTLDTLQLDVADFATTNWQKNWERNFGGNWAPFVPL  120
            QLHSGLFLYQGLLQALEGISPELGPTLDTLQLDVADFATTNWQKNWERNFGGNWAPFVPL
Sbjct  61   QLHSGLFLYQGLLQALEGISPELGPTLDTLQLDVADFATTNWQKNWERNFGGNWAPFVPL  120

Query  121  XKAPNHQGGLAQLRPGRHFRXISLVFTQERARGEERXGVXVMSYVCSQKYLXKKXIXTL  179
            XKAPNHQGGLAQLRPGRHFRXISLVFTQERARGEERXGVXVMSYVCSQKYLXKKXIXTL
Sbjct  121  XKAPNHQGGLAQLRPGRHFRXISLVFTQERARGEERXGVXVMSYVCSQKYLXKKXIXTL  179
```

Sequence Comparison Methods (SCM)

4.) Alignments of matched sequences to Query

4.1) 'Easy' Results: clear homology

```
>pdb|1RHG|A S Chain A, The Structure Of Granulocyte-Colony-Stimulating Factor
And Its Relationship To Those Of Other Growth Factors
 pdb|1RHG|B S Chain B, The Structure Of Granulocyte-Colony-Stimulating Factor
And Its Relationship To Those Of Other Growth Factors
 pdb|1RHG|C S Chain C, The Structure Of Granulocyte-Colony-Stimulating Factor
And Its Relationship To Those Of Other Growth Factors
 pdb|2D9Q|A S Chain A, Crystal Structure Of The Human Gcsf-Receptor Signaling
Complex
Length=174

 Score =  206 bits (523),  Expect = 4e-53, Method: Compositional matrix adjust.
 Identities = 103/116 (88%), Positives = 104/116 (89%), Gaps = 0/116 (0%)

Query  1    CLEQVRKIQDDGAALQEKLCATYKLCHPEELVLLGHSLGIPWAPLSSCPSQALQLAGCLS  60
            CLEQVRKIQ DGAALQEKLCATYKLCHPEELVLLGHSLGIPWAPLSSCPSQALQLAGCLS
Sbjct  17   CLEQVRKIQGDGAALQEKLCATYKLCHPEELVLLGHSLGIPWAPLSSCPSQALQLAGCLS  76

Query  61   QLHSGLFLYQGLLQALEGISPELGPTLDTLQLDVADFATTNWQKNWERNFGGNWAP  116
            QLHSGLFLYQGLLQALEGISPELGPTLDTLQLDVADFATT WQ+  E        P
Sbjct  77   QLHSGLFLYQGLLQALEGISPELGPTLDTLQLDVADFATTIWQQMEELGMAPALQP  132
```

Sequence Comparison Methods (SCM)

4.) Alignments of matched sequences to Query

4.1) 'Easy' Results: clear homology

```
>dbj|BAG60399.1| unnamed protein product [Homo sapiens]
Length=164

 GENE ID: 1440 CSF3 | colony stimulating factor 3 (granulocyte) [Homo sapiens]
(Over 100 PubMed links)

 Score =  114 bits (284),  Expect = 2e-25, Method: Compositional matrix adjust.
 Identities = 67/116 (57%), Positives = 68/116 (58%), Gaps = 36/116 (31%)

Query  1    CLEQVRKIQDDGAALQEKLCATYKLCHPEELVLLGHSLGIPWAPLSSCPSQALQLAGCLS  60
            CLEQVRKIQ DGAALQEKL                                  AGCLS
Sbjct  43   CLEQVRKIQGDGAALQEKL-----------------------------------AGCLS  66

Query  61   QLHSGLFLYQGLLQALEGISPELGPTLDTLQLDVADFATTNWQKNWERNFGGNWAP  116
            QLHSGLFLYQGLLQALEGISPELGPTLDTLQLDVADFATT WQ+  E        P
Sbjct  67   QLHSGLFLYQGLLQALEGISPELGPTLDTLQLDVADFATTIWQQMEELGMAPALQP  122
```

Sequence Comparison Methods (SCM)

4.) Alignments of matched sequences to Query

4.2) 'Difficult' Results: homology?

```
>dbj|BAG62733.1| unnamed protein product [Homo sapiens]
Length=198

 GENE ID: 3569 IL6 | interleukin 6 (interferon, beta 2) [Homo sapiens]
(Over 100 PubMed links)

 Score = 33.5 bits (75),  Expect = 0.40, Method: Compositional matrix adjust.
 Identities = 19/77 (24%), Positives = 37/77 (48%), Gaps = 1/77 (1%)

Query  3    EQVRKIQDDGAALQEKLCATYKLCHPEELVLLGHSLGIP-WAPLSSCPSQALQLAGCLSQ  61
            +Q+R I D  +AL+++ C    +C   +  L  ++L +P  A    C         CL +
Sbjct  55   KQIRYILDGISALRKETCNKSNMCESSKEALAENNLNLPKMAEKDGCFQSGFNEETCLVK  114

Query  62   LHSGLFLYQGLLQALEG  78
            + +GL  ++  L+ L+
Sbjct  115  IITGLLEFEVYLEYLQN  131
```

Sequence Comparison Methods (SCM)

4.) Alignments of matched sequences to Query

4.2) 'Difficult' Results: homology?

```
>pdb|1ALU|A Chain A, Human Interleukin-6
Length=186

 Score = 32.7 bits (73),  Expect = 0.55, Method: Compositional matrix adjust.
 Identities = 19/77 (24%), Positives = 37/77 (48%), Gaps = 1/77 (1%)

Query  3    EQVRKIQDDGAALQEKLCATYKLCHPEELVLLGHSLGIP-WAPLSSCPSQALQLAGCLSQ  61
            +Q+R I D  +AL+++ C    +C   +  L  ++L +P  A    C         CL +
Sbjct  29   KQIRYILDGISALRKETCNKSNMCESSKEALAENNLNLPKMAEKDGCFQSGFNEETCLVK  88

Query  62   LHSGLFLYQGLLQALEG  78
            + +GL  ++  L+ L+
Sbjct  89   IITGLLEFEVYLEYLQN  105
```

Sequence Comparison Methods (SCM)

Homologous or Not-homologous?

▼ Descriptions

Sequences producing significant alignments:	Score (Bits)	E Value	
gb|ACF41164.1| granulocyte colony stimulating factor [Homo sa...	351	8e-97	
pdb|1CD9|A Chain A, 2:2 Complex Of G-Csf With Its Receptor >p...	206	3e-53	S
pdb|1RHG|A Chain A, The Structure Of Granulocyte-Colony-Stimu...	206	4e-53	S
ref|NP_757373.1| colony stimulating factor 3 isoform b precur...	205	5e-53	U G
ref|NP_757374.1| colony stimulating factor 3 isoform c [Homo ...	205	7e-53	U G
gb|AAA03056.1| human granulocyte-colony stimulating factor	204	1e-52	G
pdb|1GNC|A Chain A, Structure And Dynamics Of The Human Granu...	201	1e-51	S
ref|NP_000750.1| colony stimulating factor 3 isoform a precur...	200	2e-51	U G
dbj|BAG60399.1| unnamed protein product [Homo sapiens]	114	2e-25	G
dbj|BAG62733.1| unnamed protein product [Homo sapiens]	33.5	0.40	G
dbj|BAG62358.1| unnamed protein product [Homo sapiens]	33.1	0.45	G
ref|NP_000591.1| interleukin 6 precursor [Homo sapiens] >sp|P...	33.1	0.51	U G
pdb|1IL6|A Chain A, Human Interleukin-6, Nmr, Minimized Avera...	33.1	0.52	S
emb|CAA27991.1| 26 kDa protein (aa 32-212) [Homo sapiens]	33.1	0.52	G
pdb|1P9M|B Chain B, Crystal Structure Of The Hexameric Human ...	32.7	0.55	S
pdb|1ALU|A Chain A, Human Interleukin-6	32.7	0.55	S
dbj|BAG60317.1| unnamed protein product [Homo sapiens]	32.7	0.67	G
dbj|BAG62348.1| unnamed protein product [Homo sapiens]	32.3	0.91	G

Sequence Comparison Methods (SCM)

Homologous or **Not-homologous?**

1.) **% Sequence Identity (%_ID)**

2.) **Expect value (E-value)**

3.) **Conservation of key-residues**

Sequence Comparison Methods (SCM)

Homologous or Not-homologous?

1.) % Sequence Identity (%_ID)

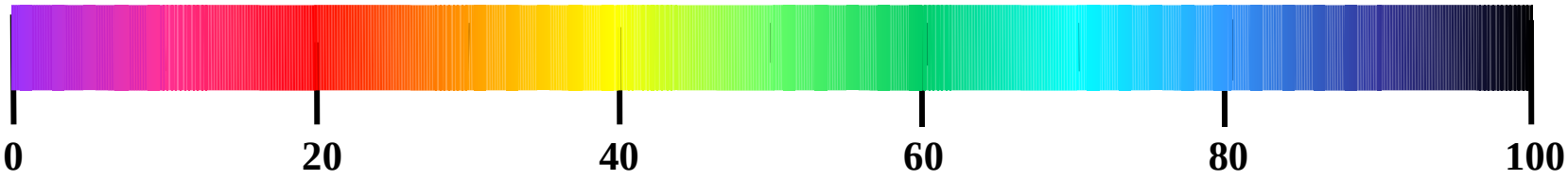
▼ Descriptions

Sequences producing significant alignments:	Score (Bits)	E Value		
gb|ACF41164.1| granulocyte colony stimulating factor [Homo sa...	351	8e-97		100
pdb|1CD9|A Chain A, 2:2 Complex Of G-Csf With Its Receptor >p...	206	3e-53	S	
pdb|1RHG|A Chain A, The Structure Of Granulocyte-Colony-Stimu...	206	4e-53	S	88
ref|NP_757373.1| colony stimulating factor 3 isoform b precur...	205	5e-53	U G	
ref|NP_757374.1| colony stimulating factor 3 isoform c [Homo ...	205	7e-53	U G	
gb|AAA03056.1| human granulocyte-colony stimulating factor	204	1e-52	G	
pdb|1GNC|A Chain A, Structure And Dynamics Of The Human Granu...	201	1e-51	S	
ref|NP_000750.1| colony stimulating factor 3 isoform a precur...	200	2e-51	U G	
dbj|BAG60399.1| unnamed protein product [Homo sapiens]	114	2e-25	G	57
dbj|BAG62733.1| unnamed protein product [Homo sapiens]	33.5	0.40	G	24
dbj|BAG62358.1| unnamed protein product [Homo sapiens]	33.1	0.45	G	
ref|NP_000591.1| interleukin 6 precursor [Homo sapiens] >sp|P...	33.1	0.51	U G	
pdb|1IL6|A Chain A, Human Interleukin-6, Nmr, Minimized Avera...	33.1	0.52	S	
emb|CAA27991.1| 26 kDa protein (aa 32-212) [Homo sapiens]	33.1	0.52	G	
pdb|1P9M|B Chain B, Crystal Structure Of The Hexameric Human ...	32.7	0.55	S	
pdb|1ALU|A Chain A, Human Interleukin-6	32.7	0.55	S	24
dbj|BAG60317.1| unnamed protein product [Homo sapiens]	32.7	0.67	G	
dbj|BAG62348.1| unnamed protein product [Homo sapiens]	32.3	0.91	G	

Sequence Comparison Methods (SCM)

Homologous or Not-homologous?

1.) % Sequence Identity (%_ID)



Random (~ 20%) **'Twilight'** **(Close) Homology (> 40%)**

Length ~ 100 a.a.

Not 'low-complexity'

Homology:

< Length, > %_ID (e.g., 10 a.a.)

> Length, < %_ID

'low complexity' can be 'masked'

Sequence Comparison Methods (SCM)

Homologous or Not-homologous?

2.) Expect value (E-value):

Number of matches (with a certain score)

"expected to be found merely by chance"

▼ Descriptions

Sequences producing significant alignments:	Score (Bits)	E Value	
gb|ACF41164.1| granulocyte colony stimulating factor [Homo sa...	351	8e-97	
pdb|1CD9|A Chain A, 2:2 Complex Of G-Csf With Its Receptor >p...	206	3e-53	S
pdb|1RHG|A Chain A, The Structure Of Granulocyte-Colony-Stimu...	206	4e-53	S
ref|NP_757373.1| colony stimulating factor 3 isoform b precur...	205	5e-53	U G
ref|NP_757374.1| colony stimulating factor 3 isoform c [Homo ...	205	7e-53	U G
gb|AAA03056.1| human granulocyte-colony stimulating factor	204	1e-52	G
pdb|1GNC|A Chain A, Structure And Dynamics Of The Human Granu...	201	1e-51	S
ref|NP_000750.1| colony stimulating factor 3 isoform a precur...	200	2e-51	U G
dbj|BAG60399.1| unnamed protein product [Homo sapiens]	114	2e-25	G
dbj|BAG62733.1| unnamed protein product [Homo sapiens]	33.5	0.40	G
dbj|BAG62358.1| unnamed protein product [Homo sapiens]	33.1	0.45	G
ref|NP_000591.1| interleukin 6 precursor [Homo sapiens] >sp|P...	33.1	0.51	U G
pdb|1IL6|A Chain A, Human Interleukin-6, Nmr, Minimized Avera...	33.1	0.52	S
emb|CAA27991.1| 26 kDa protein (aa 32-212) [Homo sapiens]	33.1	0.52	G
pdb|1P9M|B Chain B, Crystal Structure Of The Hexameric Human ...	32.7	0.55	S
pdb|1ALU|A Chain A, Human Interleukin-6	32.7	0.55	S
dbj|BAG60317.1| unnamed protein product [Homo sapiens]	32.7	0.67	G
dbj|BAG62348.1| unnamed protein product [Homo sapiens]	32.3	0.91	G

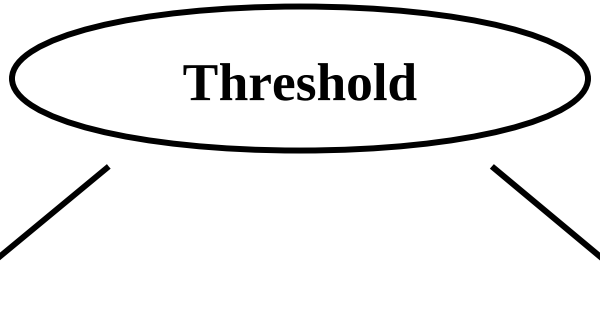
Sequence Comparison Methods (SCM)

Homologous or **Not-homologous?**

2.) Expect value (E-value):

Number of matches (with a certain score)

“expected to be found merely by chance”



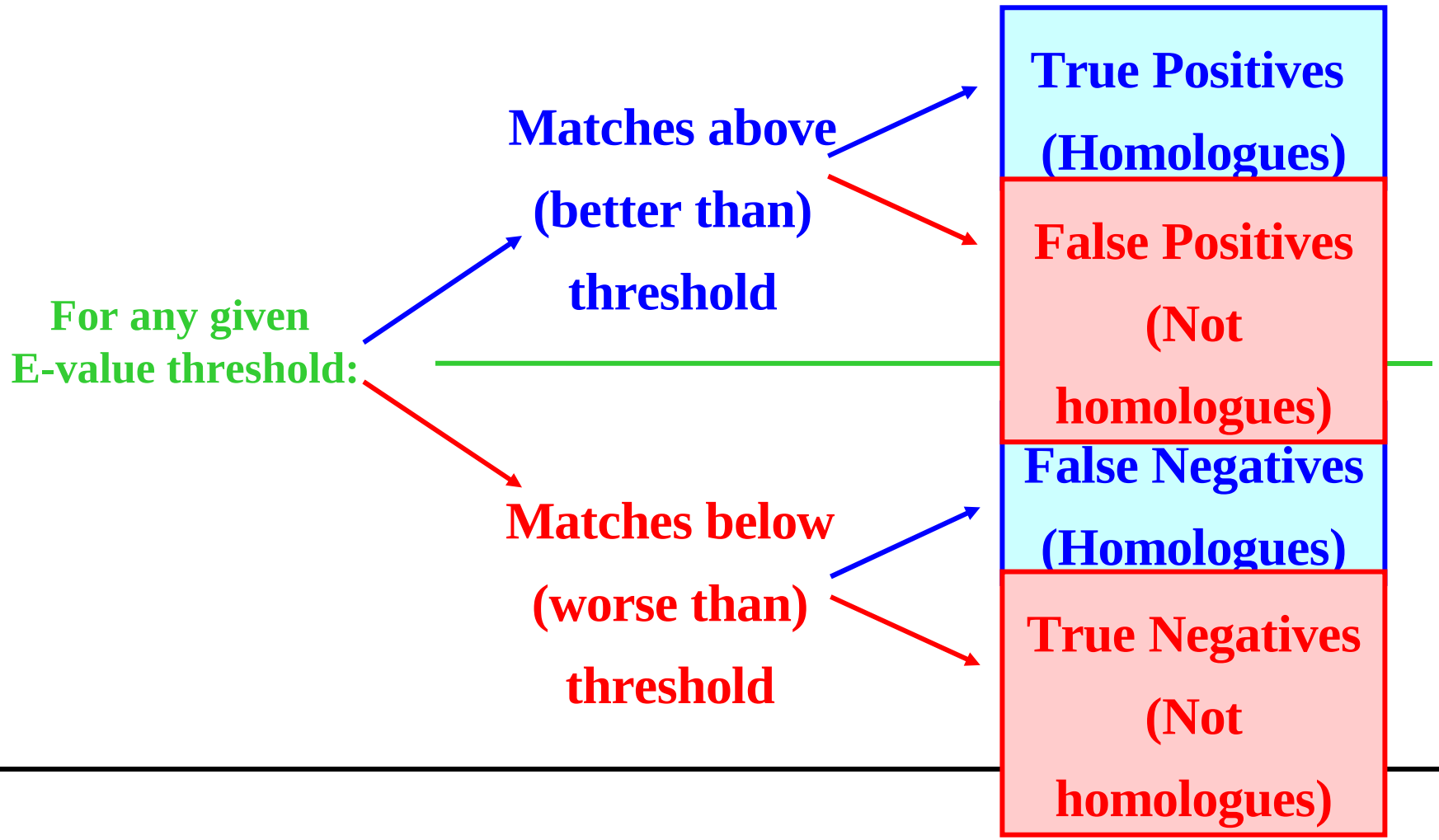
Blast:
10

Psi-Blast:
0.005

Sequence Comparison Methods (SCM)

Homologous or **Not-homologous?**

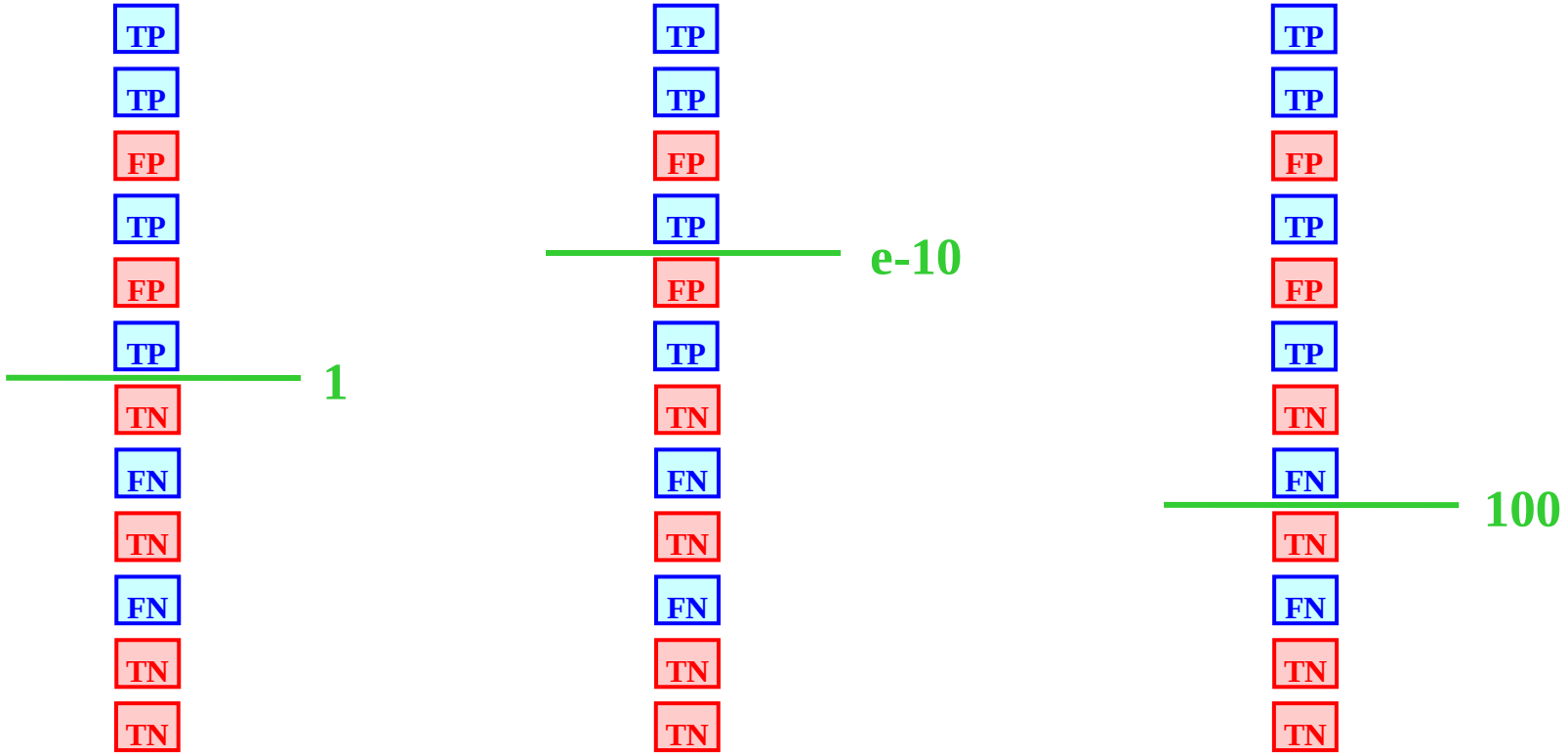
Threshold problem:



Sequence Comparison Methods (SCM)

Homologous or Not-homologous?

2.) Expect value (E-value):



Sequence Comparison Methods (SCM)

Homologous or **Not-homologous?**

1.) % Sequence Identity (%_ID)

2.) Expect value (E-value)

3.) Conservation of key-residues

Protein sequence-structure-function relationships

3D-Structures are more conserved than a.a sequences

1alu & 1bgc

1.1 Å RMSD

14/71 = 20 % Seq. ID

1cnt & 1ax8

0.9 Å RMSD

8/71 = 11 % Seq. ID

1hgu &1lki

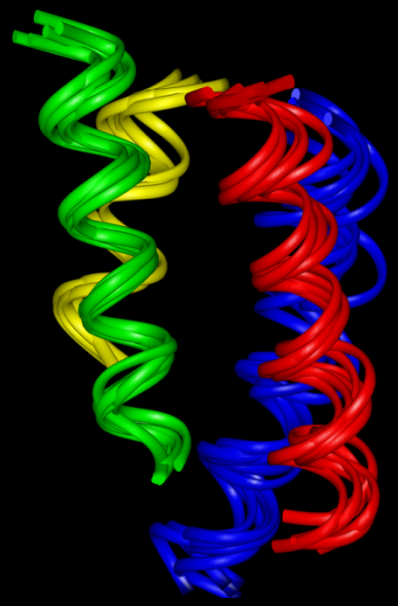
1.7 Å RMSD

9/71 = 13 % Seq. ID

Protein sequence-structure-function relationships

Just a small number of residues ("key-residues") are required to maintain the protein fold

4-alpha-helical cytokines family



common core: 71 a.a.

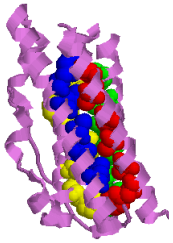
11 highly conserved "key-residues" surrounded by 10 a.a. conserving a generic hydrophobic-neutral character

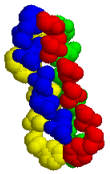
Protein sequence-structure-function relationships

4-helical cytokines family

```
                                -----αA-----                    --β1--                    -----αB-------
                                1*  4     11                    1    6                    1* 3        13
                                |   |     |                     |    |                    |  |        |
1rcb  (1)  ---------hkcdit-LQE-IIKTLNSL-teqktlcte---------LTVTDI-faasknt-------------TE-KETFCRAATVL-rqfyshhekdtrc
1eer  (1)  ----apprlicdsrv-LER-YLLEAKEA-ekittgcaehcsln----EKITVP-dtkvnfyawkrmevgqqav-EV-WQGLALLSEAV-lrgqallvkssqp
1ete  (1)  ----tqdcsfqhspi-SSD-FAVKIREL-sdyllqd-----------YPVTVA-snlqddel----------c-GG-LWRLVLAQRWM-erlktva------
2gmf  (4)  rspspstqpwehvna-IQE-ARRLLNLS-rdtaaemn----------ETVEVI-semfdlqept----------CL-QTRLELYKQGL-rgs----------
1hul  (5)  ---------iptsal-VKE-TLALLSTH-rtllianet---------LRIPVP-vhknh------------------QLCTEEIFQGI-gtlesqtvqg---
3ink  (6)  -------stkktqlq-LEH-LLLDLQMI-lnginnyknpkltrmlt-FKFYMP-kkat-----------------E-LKHLQCLEEEL-kpleevlnlaqsk
1scf  (11) --------------------NVKDVTKL-vanlpkd-----------YMITLK-yvpgmdvlpshc--------WI-SEMVVQLSDSL-tdlldkfs-nise

                                ___ _______                                           __ _________
1rcb  (1)  ---------997592-066-00900630-392818009---------360211-4549992-------------86-94300000300-9900692691990
1eer  (1)  ----99399036595-059-22930960-49508406970509----590500-32939994099597060 00-40-29004502900-8906443998996
1ete  (1)  ----99904199540-498-05541950-3993999-----------860421-50059294----------0-02-00800100930-6904992------
2gmf  (4)  995677981990370-095-08921992-93886449----------680700-2992759915----------00-43109203901-634----------
1hul  (5)  ---------984531-096-03620692-696057299---------290310-71910------------------11005502900-3507891976---
3ink  (6)  -------98992392-079-01900640-29007499095499069-890920-9918-----------------8-49106003700-9506910961459
1scf  (11) --------------------79720970-3780799-----------770606-309129967150--------00-02004101710-67029959-9999
```





Common Core

Structure similarity

‘Key’-residues

Structure & sequence similarity

Sequence Comparison Methods (SCM)

Pair-wise sequence comparison methods do not recognize "key-residues" for protein structure/function

All positions of the alignment are the same and have the same weight on the computed parameters (i.e., %_ID, E-value, etc.)

```
>pdb|1ALU|A Chain A, Human Interleukin-6
Length=186

 Score = 32.7 bits (73),  Expect = 0.55, Method: Compositional matrix adjust.
 Identities = 19/77 (24%)  Positives = 37/77 (48%), Gaps = 1/77 (1%)

Query  3    EQVRKIQDDGAALQEKLCATYKLCHPEELVLLGHSLGIP-WAPLSSCPSQALQLAGCLSQ  61
            +Q+R I D  +AL+++ C    +C   +  L  ++L +P  A    C         CL +
Sbjct  29   KQIRYILDGISALRKETCNKSNMCESSKEALAENNLNLPKMAEKDGCFQSGFNEETCLVK  88

Query  62   LHSGLFLYQGLLQALEG  78
            + +GL  ++  L+ L+
Sbjct  89   IITGLLEFEVYLEYLQN  105
```

Sequence Comparison Methods (SCM)

Multiple sequence alignments (MSA)

- **More informative than pair-wise alignments**
- **Different positions have different conservation**
- **May allow to recognize “key-residues” for protein structure/function**

- **Input sequences:**
  - **Relatively high number**
  - **Similar enough to produce correct alignments (eliminate ‘outliers’, i.e., < 20 %_ID)**
  - **Different enough to distinguish between conserved and variable positions (make non-redundant, i.e., eliminate > 80 %_ID)**

Sequence Comparison Methods (SCM)

‘High-quality’ MSA

Dps proteins

H.pylori	B	1JI4	QADAIVLFMKVHNFHWNVKGTDFFNVHKATEEIYEEFADMFDDLAERIVQ ILEDYKYL LAK-LQKSIW
H.hepaticus	B		QADAAVFYVKVHNFHWNVKGMDFYPTHKATEEIYEKYADVFDDVAERVLQ ILSDYEYF VGE-LQKAIW
V.cholerae	B	3IQ1	LANYQVFYMNTRGYHWNIQGKEFFELHAKFEEIYTDLQLKIDELAERILT LVDGFSIL IRE-QEKLVW
S.degradans	B		LADSYVLYLKTHNFHWNVTGPMFQTLHNMFMDQYTEAWTALDTIAERIRT LLEGQETL IEV-HEKNAW
L.pneumophila	B		LADTYALYLKTQNYHWHVTGPQFKSLHELFEMQYKELAEAVDQIAERIRI LAKDNMMI VAA-HEKAHW
B.anthracis	B	1JIG	VANWNVLYVKLHNYHWYVTGPHFFTLHEKFEEFYNEAGTYIDELAERILA LVNDYSAL HTT-LEQHVW
B.anthracis	B	1JI5	VADWSVLFTKLHNFHWYVKGPQFFTLHEKFEELYTESATHIDEIAERILA IMKDYEMM YTE-LEKHAW
S.aureus	B	2D5K	VANWTVAYTKLHNFHWYVKGPNFFSLHVKFEELYNEASQYVDELAERILA LSQDFTNI QTS-VDKHNW
S.epidermidis	B		VANWTVAYTKLHNFHWYVKGPNFFSLHTKFEELYNEASQYVDDLAERILA LSKDFSKI QTS-VDKHNW
B.subtilis	B	2CHP	LSNWFLLYSKLHRFHWYVKGPHFFTLHEKFEELYDHAAETVDTIAERLLA LVNDYKQI IEE-VEKQVW
S.pyogenes	B	2WLA	VADLSVAASIVHQVHWYMRGPGFLYLHPKMDELLDSLNANLDEMSERLIT LVEVYLYL KTE-AEKTIW
L.monocytogenes	B	2IY4	VANLNVFTVKIHQIHWYMRGHNFFTLHEKMDDLYSEFGEQMDEVAERLLA LVGTLELL KAS-IDKHIW
O.oeni	B		IADISQLKVNVQQTHWYMRGENFFRLHPLMDEYGDQLSEQLDQIAERLIA LVDQFKYL KDE-TDKNIW
E.coli	B	1F33	VIQFIDLSLITKQAHWNMRGANFIAVHEMLDGFRTALIDHLDTMAERAVQ LADRYAIV SRD-LDKFLW
S.enterica	B		VIQFIDLSLITKQAHWNMRGANFIAVHEMLDGFRTALTDHLDTMAERAVQ LADRYAVV SRD-LDKFLW
B.melitensis	B	3GE4	LAATIDLALITKQAHWNLKGPQFIAVHEMLDGFRAELDDHVDTIAERAVQ LIERYGDV SRS-LDKALW

Bacterioferritins

S.enterica	B		LGNELVAINQYFLHARMFKNWGLTRLNDVEYHESIDEMKHADKYIERILF DLRLEL-E LAD-EEGHID
E.coli	B	2HTN	LGNELVAINQYFLHARMFKNWGLKRLNDVEYHESIDEMKHADRYIERILF DLALEL-D LRD-EEGHID
Y.pestis	B		LGNELVAINQYFLHARMFKNWGLMRLNDKEYHESIDEMKHADKYIERILF DLALEL-S LVD-EEEHID
C.B.pennsylvanicus	B		LSDELVAVNQYFLHSKIFNNWGLERLNKIEYQECVDELDHADLYAKRILF DLSLEF-H LKD-EEKHID
A.vinelandii	B	1SOF	LGNELIAINQYFLHARMYEDWGLEKLGKHEYHESIDEMKHADKLIKRILF DLKLEQ-A LES-EEDHID
M.capsulatus	B		LTNELTAINQYFLHARMFKNWGFGKLNEHEYKESIDEMKHADRLIERILF DLQLEQ-Q LES-EEEHVD
S.alaskensis	B		LKNELTAINQYWLHYRMLDNWGVARLAHFEREESIDEMKHADKLADRILF DLALEE-E LES-EEHHVD
H.baltica	B		LKNELTAINQYFLHSRMLKDWGVSVLAEKEYKESIEEMQHADWLIDRILF DLKLEH-D LEN-EEEHVD
B.melitensis	B		LFLELGAVNQYWLHYRLLNDWGYTRLAKKERESIEEMHHADKLIDRIIF DLKGEY-D LAD-EEGHID
Bradyrhizobium sp.	B		LRSELTAINQYWLHYRLLNNWGLLEMAKVWRKESIEEMEHADKFTDRILF DLAAEI-G MKD-EEHHID
P.aeruginosa	B		LTGELAARDQYFIHSRMYEDWGFSKLYERLNHEMEEETQHADALLRRILL DLKLER-H LADTEEDHAY
R.palustris	B		LRGELTAISQYWLHYRLLANWGLKDMAKVWRKESIEEMEHADLLTDRIIF DLAAEM-G MKD-EEHHID
P.fluorescens	B		LTGELAARDQYFVHSRMYEDWGFTKLYERINHEMEEEAAHADALMRRILM DLRLEY-K LHDTEEDHTY
M.capsulatus	B		LAGELAAIDQYFIHAMMYRDWGFHVLYEHTAHEMQEEQAHASALIRRILF DLGVEH-A LDDTEEDHCL
I.loihiensis	B		LAFELTSIDQYTSHSRQYEDMGLMKLYERINHEIDDERGHADLLIRRILF DLKLEH-N LKDTEEDHAY
M.bovis	B		LTSELTAINQYFLHSKMQDNWGFTELAAHTRAESFDEMRHAEEITDRILL DLAIEY-D VAD-EEEHID

Sequence Comparison Methods (SCM)

Multiple sequence alignments

Cobalt alignment; or:

1) Get sequences to align:

- putative homologs detected from a Blast search (saved as text)

2) Align all sequences to one another

- **ClustalW, T-Coffee: www.ebi.ac.uk -> tools -> sequence analysis**

Sequence Comparison Methods (SCM)

Clustal programs

ClustalW2:

- Input sequences
- Multiple Sequence Alignment Options: Aligned vs. Input
- Output (%-age sequence identity)
- Alignment:
    1. Sequences with very different length
    2. Outliers (<20% sequence identity)
    3. Redundant (>80 % sequence identity)

Sequence Comparison Methods (SCM)

Multiple sequence alignment methods

Edit/Visualize MSA:

- **ClustalX (ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/)**
- **JalView (http://www.jalview.org/download.html)**
- **BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html)**
- **WebLogo (http://weblogo.berkeley.edu/logo.cgi)**

Sequence Comparison Methods (SCM)

‘High-quality’ MSA

At the basis of a number of structure/function prediction methods:

- Homology modelling
- domains
- natively unfolded regions
- TM regions
- solvent accessibility
- secondary structures
- ...

Protein sequence-structure-function relationships

Most proteins are modular

Domains: structural, functional, folding and evolutionary units

(30-700 a.a.; 100 a.a. on average)

Analysis and prediction: domain – not whole protein – level

Tools for prediction of:

- **Domains**
- **Disordered regions**
- **Trans-membrane regions**
- **Secondary structure elements**

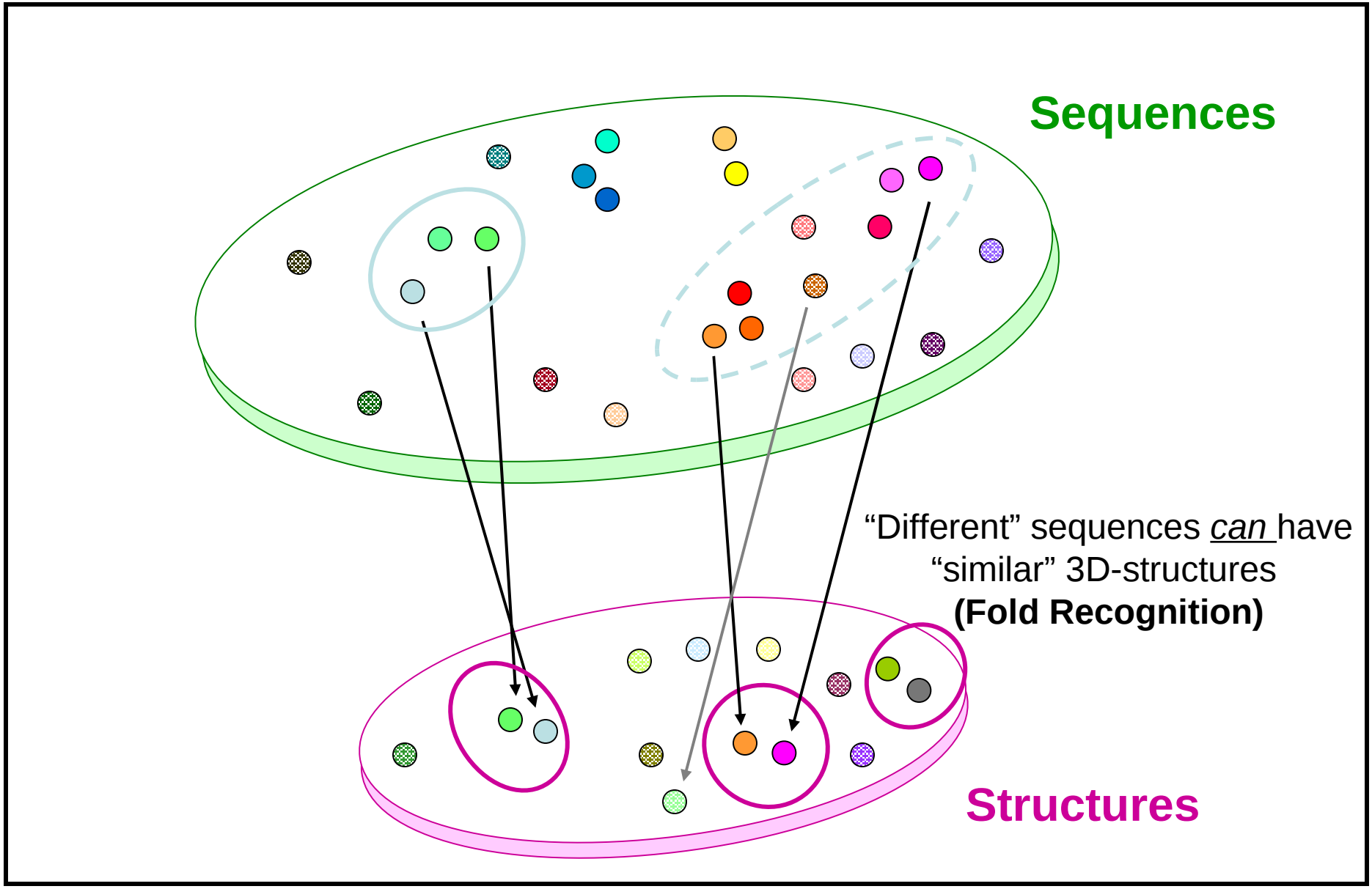
Protein sequence-structure-evolutionary relationships

Structural Classification Of Proteins (SCOP) DB

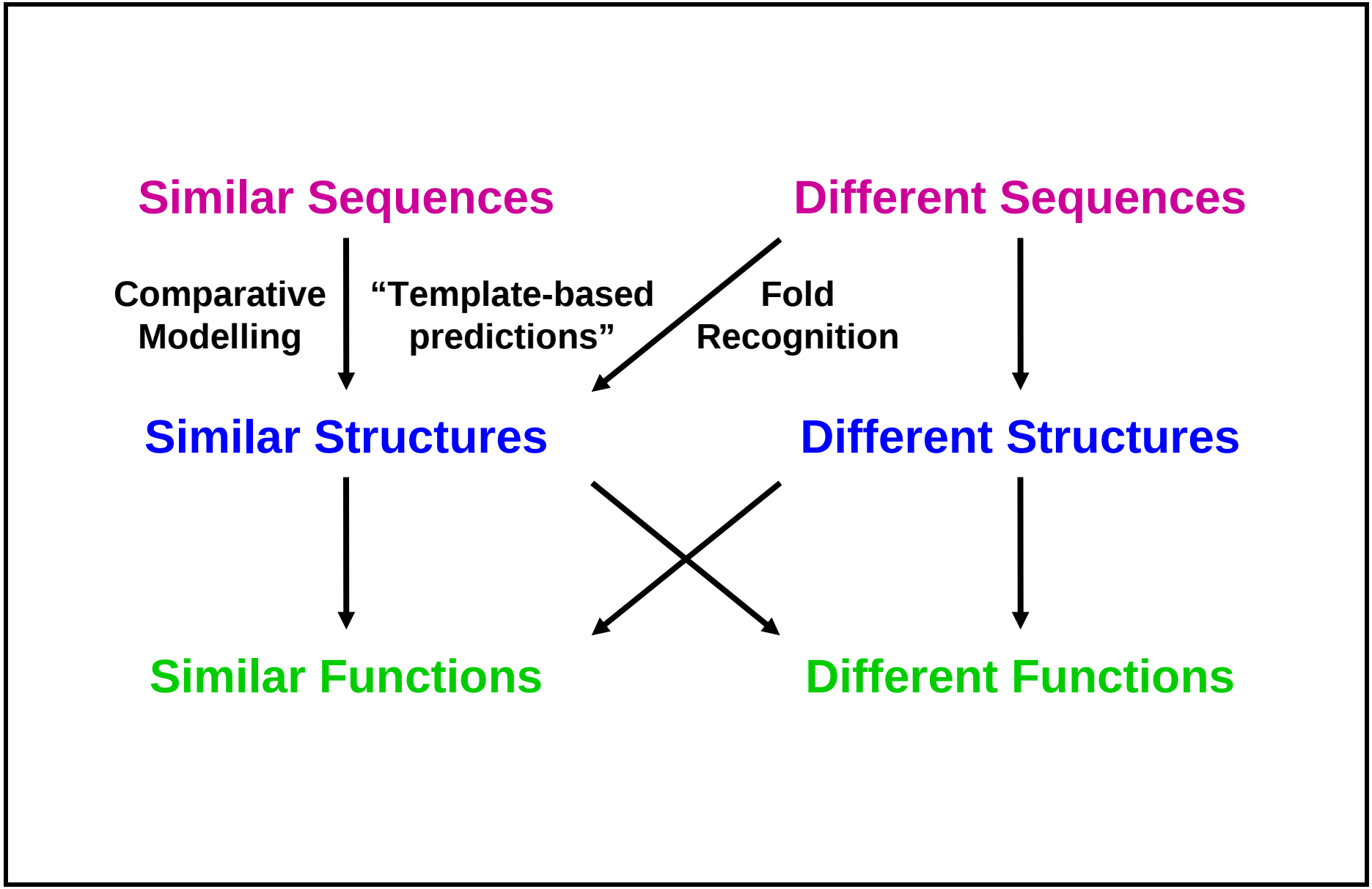
(scop.mrc-lmb.cam.ac.uk/scop/)

DOMAINS:	**Structural, functional, folding and evolutionary protein units**
FAMILY:	**Close evolutionary relationship (structure and sequence similarity)**
SUPERFAMILY:	**Distant evolutionary relationship (structure similarity)**
FOLD:	**Uncertain evolutionary relationship (structural similarity)**
CLASS:	**Secondary structure content (α; β; α/β; α+β; …)**

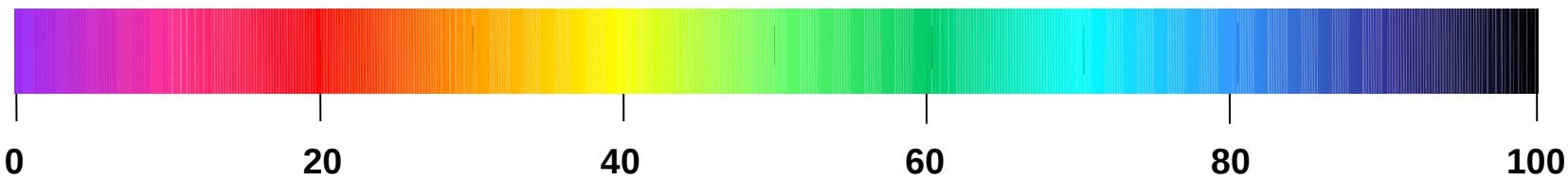
Protein sequence-structure-function relationships



Protein sequence-structure-function relationships



Protein structure prediction methods



%age sequence identity with known structures

····· Random ·····

Homology modelling

Fold recognition

Template-based:
global similarity

Fragment-based

New Fold?

Structure-function relationships

a.a. Sequence
+
3D-Structure

$\xrightarrow{??}$ Function

a.a. Sequence
&
3D-Structure
DBs

$\longrightarrow$ Function $\rightarrow$ Known / Unknown (genomic projects)

Sequence-structure relationships

a.a. Sequence $\xrightarrow{\quad ? \quad}$ **3D-Structure**

"Folding problem"

ab initio
(the "Holy grail")

Physical-Chemical Principles
("nature folds proteins without searching the DBs…")

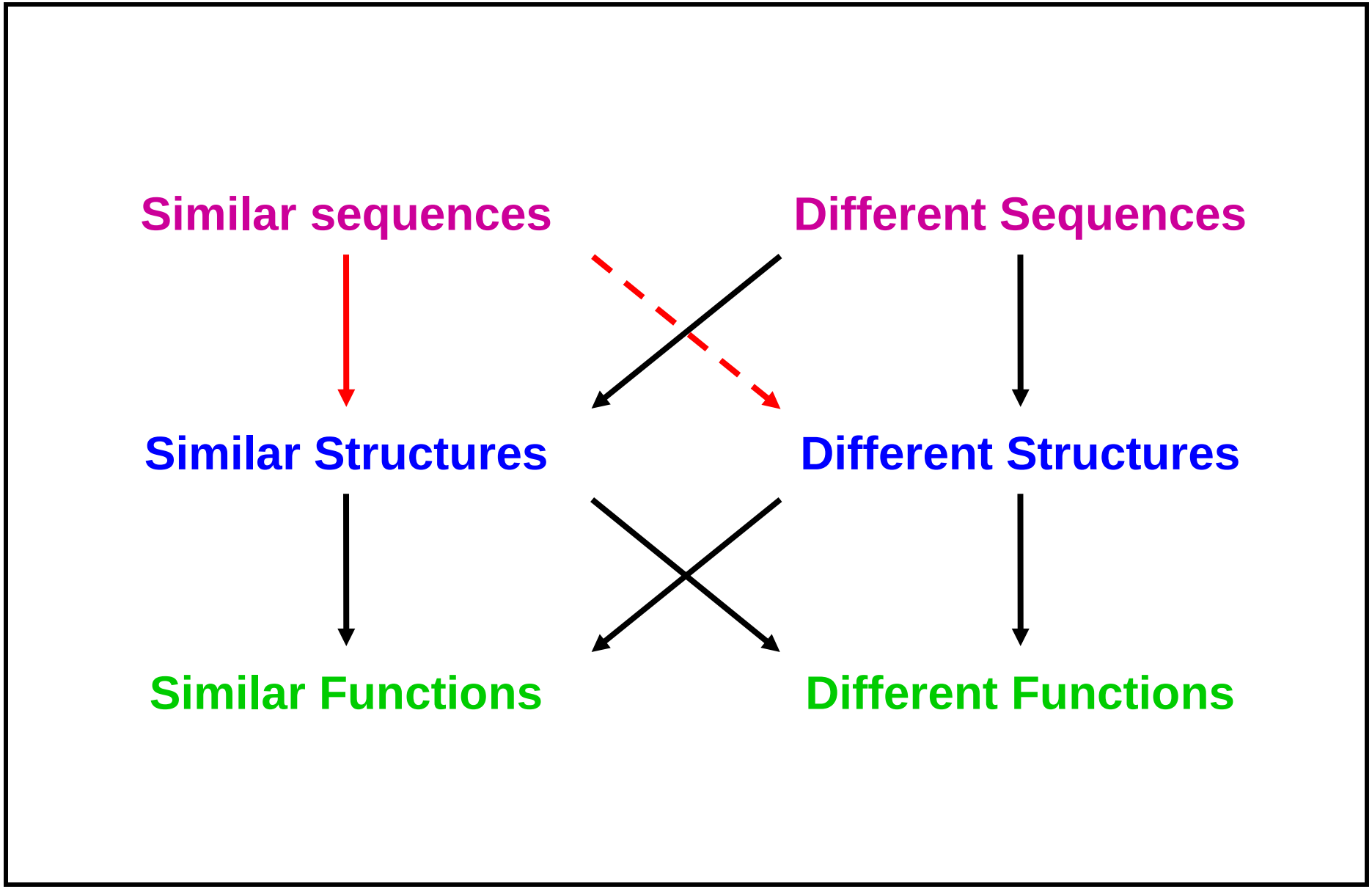
-

evolutionary

Relationships between known sequences and structures (DBs!!!)

+++

Sequence-structure-function relationships



Sequence-structure relationships

a.a. Sequence —**?**→ **3D-Structure**

"Folding problem"

ab initio	Physical-Chemical Principles	-
evolutionary	Relationships between sequences and structures in DBs	**+++**

Global: "Template-based predictions"

- High sequence similarity: Homology modelling ++/+++
- Low sequence similarity: Fold recognition +/++

Local: "New-Fold predictions"

- No sequence similarity: Fragment-based methods -/±

Protein sequence analysis

1) NCBI (UniProt) annotation

2) BLAST

- Conserved domains (CDD) → annotate residues (51-202)
- Similar sequences → download Fasta complete
- Pair-wise alignments → download (text format)
- Multiple alignment (COBALT) → download (text format)